Neural Wave Functions for Superfluids

Wan Tong Lou,[1] Halvard Sutterud,[1] Gino Cassella,[1] W. M. C. Foulkes,[1] Johannes Knolle,[1,2,3] David Pfau,[4,1] and James S. Spencer[4]

[1]*Department of Physics, Imperial College London, South Kensington Campus, London SW7 2AZ, United Kingdom*
[2]*Department of Physics TQM, Technische Universität München, James-Franck-Straße 1, D-85748 Garching, Germany*
[3]*Munich Center for Quantum Science and Technology (MCQST), 80799 Munich, Germany*
[4]*Google DeepMind, 6 Pancras Square, London N1C 4AG, United Kingdom*

(Received 21 May 2023; accepted 19 April 2024; published 22 May 2024)

Understanding superfluidity remains a major goal of condensed matter physics. Here, we tackle this challenge utilizing the recently developed fermionic neural network (FermiNet) wave function *Ansatz* [D. Pfau *et al.*, Phys. Rev. Res. **2**, 033429 (2020).] for variational Monte Carlo calculations. We study the unitary Fermi gas, a system with strong, short-range, two-body interactions known to possess a superfluid ground state but difficult to describe quantitatively. We demonstrate key limitations of the FermiNet *Ansatz* in studying the unitary Fermi gas and propose a simple modification based on the idea of an antisymmetric geminal power singlet (AGPs) wave function. The new AGPs FermiNet outperforms the original FermiNet significantly in paired systems, giving results which are more accurate than fixed-node diffusion Monte Carlo and are consistent with experiment. We prove mathematically that the new *Ansatz*, which differs from the original *Ansatz* only by the method of antisymmetrization, is a strict generalization of the original FermiNet architecture, despite the use of fewer parameters. Our approach shares several advantages with the original FermiNet: The use of a neural network removes the need for an underlying basis set; sand the flexibility of the network yields extremely accurate results within a variational quantum Monte Carlo framework that provides access to unbiased estimates of arbitrary ground-state expectation values. We discuss how the method can be extended to study other superfluid.

DOI: 10.1103/PhysRevX.14.021030

Subject Areas: Computational Physics, Condensed Matter Physics, Superfluidity

I. INTRODUCTION

The unitary Fermi gas (UFG) is a paradigmatic example of a strongly interacting system of two-component fermions that possesses a superfluid ground state and lies in the crossover region between a Bardeen-Cooper-Schrieffer (BCS) superconductor and a Bose-Einstein condensate [1,2]. The effective range of the interaction is zero and the s-wave scattering length diverges (the "unitarity limit"), so the UFG has no intrinsic length scale. The only remaining length is the inverse of the Fermi wave vector $1/k_F$, on which all thermodynamic quantities depend. For example, regardless of the particle density, the ground-state energy per particle of a unitary Fermi gas can be written as

$$E = \xi E_{\mathrm{FG}} = \xi \frac{3}{5}\frac{\hbar^2 k_F^2}{2m}, \tag{1}$$

where E_{FG} is the energy per particle of a noninteracting Fermi gas of the same density. The dimensionless constant ξ is known as the Bertsch parameter [3].

Because of the universality of the UFG model, it can be used to describe many real physical systems at different scales, such as the neutron matter in the inner crust of a neutron star [4] or the quantum criticality of an s-wave atomic superfluid [5,6]. The size of the pairs in the UFG is comparable to the interparticle spacing, which is also a feature of many high-T_c superconductors [7–9]. As a result, the UFG has been studied extensively [10]. Although the UFG is an idealized model, it can be accurately realized in the laboratory using ultracold atomic gases in which the interactions have been tuned by using an external magnetic field to drive the system across a Feshbach resonance [11].

The UFG has been studied for decades, but it remains difficult to calculate its ground-state properties accurately using analytic methods. Mean-field treatments such as BCS

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

theory [12] give good results for systems with weak interactions, but there is no guarantee of success in the strongly interacting regime. As a result, various quantum Monte Carlo (QMC) methods [13,14] have been used to simulate the properties of the UFG to high accuracy at zero and finite temperature. Methods used include variational Monte Carlo (VMC), fixed-node diffusion Monte Carlo (FNDMC), fixed-node Green function Monte Carlo, auxiliary field Monte Carlo (AFMC), and diagrammatic Monte Carlo [15–25]. However, a full quantitative description remains an open and challenging problem.

Recent advances in machine learning algorithms and the growing availability of inexpensive GPU-based computational resources have allowed neural-network-based approaches to permeate many areas of computational physics, including lattice [26–29] and continuum [30–33] QMC simulations. Here, we employ a neural-network *Ansatz* within a VMC approach to study the unitary Fermi gas. The *Ansatz* we use, the fermionic neural network (FermiNet) [30], gives very accurate results for atoms and molecules [30,34–36] and has recently been applied to periodic solids and the homogeneous electron gas (HEG) with comparable success [37]. In the case of the HEG, the variational optimization of the FermiNet *Ansatz* discovered the quantum phase transition between the Fermi liquid and Wigner crystal ground states without external guidance [38]. In contrast, previous approaches required different *Ansätze* to be used for the two different phases. The FermiNet has not previously been applied to fermionic superfluids such as the UFG.

The paper is organized as follows. Section II describes the architecture of the FermiNet. We find that the original FermiNet *Ansatz* is insufficient to capture the two-particle correlations of superfluids. Although a FermiNet wave function with one determinant and a sufficiently large neural network is, in principle, able to represent any fermionic state [30], it is often advantageous to use a network of a fixed size and a small linear combination of FermiNet determinants. In the case of the unitary Fermi gas, however, we find that the number of block-diagonal determinants required to describe the ground state accurately scales exponentially with the system size. This is the first example in which the FermiNet has been seen to fail both quantitatively and qualitatively and suggests that the FermiNet wave function may not be able to represent arbitrary fermionic wave functions in practice. To remedy the problem, we utilize the neural-network part of the FermiNet architecture to build a different type of wave function based on the idea of an antisymmetric geminal power singlet wave function (AGPs) [29,39–42], which we discuss in detail in Sec. III. This leads to substantial improvements, even though the neural-network part of the wave function remains unchanged. The implementation of the AGPs wave function using the FermiNet, as well as its relation to the original block-diagonal multideterminant FermiNet, are discussed in Sec. IV. Our computational results are presented in Sec. V, followed by a summary and discussion in Sec. VI. The Appendixes A–F include detailed explanations and derivations of important formulas as well as implementation and training details.

II. FERMINET

The fermionic neural network, or FermiNet [30], is a neural network that can be used to approximate the ground-state wave function of any system of interacting fermions. The inputs to the network are the positions $\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N$ and spin coordinates $\sigma_1, \sigma_2, \ldots, \sigma_N$ of the N particles, and the output is the value of the wave function $\Psi(\mathbf{r}_1, \sigma_1, \mathbf{r}_2, \sigma_2, \ldots, \mathbf{r}_N, \sigma_N)$ corresponding to those inputs. The network is trained using the VMC method [13]: The weights and biases that define the network are varied at each training iteration to minimize the energy expectation value according to the variational principle. If the network is flexible enough, the approximate wave function obtained after training may be very close to the true ground state. The FermiNet provides a more general and accurate alternative to the conventional Slater-Jastrow (SJ) and Slater-Jastrow-backflow *Ansätze* that have been used in most VMC and FNDMC calculations to date and may improve VMC and FNDMC results for strongly correlated systems.

In conventional SJ *Ansätze*, the antisymmetry of the N-electron wave function is represented using Slater determinants, which are antisymmetrized products of single-particle orbitals. For simulations of solids, it is common to use one determinant only; for molecules, a linear combination of determinants is usually employed. In both cases, the presence of determinants guarantees that the wave function has the correct exchange antisymmetry. To improve the representation of electronic correlations, especially the correlations that chemists call "dynamic," the determinants are multiplied by a totally symmetric nonnegative function of the electron coordinates known as a Jastrow factor. This acts to decrease the value of the wave function as pairs of electrons approach each other, reducing the total Coulomb repulsion energy.

If the Hamiltonian is independent of spin and all the single-particle orbitals are eigenfunctions of total S_z, one can assign spins to the electrons and every Slater determinant can be factored into a product of spin-up and spin-down Slater determinants [13,43]. The wave function is no longer antisymmetric under the exchange of electrons of opposite spin, but expectation values of spin-independent operators are unaltered. Including a spin-assigned Jastrow factor expressed in the form e^J, a one-determinant SJ *Ansatz* becomes

$$\Psi_{\mathrm{SJ}}(\{\mathbf{r}^{\uparrow}\},\{\mathbf{r}^{\downarrow}\}) = e^{J(\{\mathbf{r}^{\uparrow}\},\{\mathbf{r}^{\downarrow}\})} \det\left[\phi_i^{\uparrow}(\mathbf{r}_j^{\uparrow})\right] \det\left[\phi_i^{\downarrow}(\mathbf{r}_j^{\downarrow})\right], \quad (2)$$

where $\{\mathbf{r}^{\uparrow}\}$ and $\{\mathbf{r}^{\downarrow}\}$ are the sets of position coordinates of the $N^{\uparrow}$ electrons assigned to be spin up and the $N^{\downarrow}$ electrons assigned to be spin down, respectively.

One can improve the SJ *Ansatz* by transforming the electron coordinates as

$$\mathbf{r}_j^{\alpha} \to \mathbf{x}_j^{\alpha} = \mathbf{r}_j^{\alpha} + \sum_{\substack{i=1 \\ (i\neq j)}}^{N^{\alpha}} \eta_{\parallel}(r_{ij}^{\alpha\alpha})(\mathbf{r}_i^{\alpha} - \mathbf{r}_j^{\alpha}) + \sum_{i=1}^{N^{\bar{\alpha}}} \eta_{\nparallel}(r_{ij}^{\bar{\alpha}\alpha})(\mathbf{r}_i^{\bar{\alpha}} - \mathbf{r}_j^{\alpha}), \tag{3}$$

where α and $\bar{\alpha}$ are the two possible spin components of an electron, $r_{ij}^{\beta\alpha} = |\mathbf{r}_i^{\beta} - \mathbf{r}_j^{\alpha}|$, and $\eta_{\parallel}(r)$ and $\eta_{\nparallel}(r)$ are parametrized functions of a single distance argument. The coordinate-transformed SJ *Ansatz* is called a Slater-Jastrow-backflow (SJB) wave function, and the new coordinates are called *quasiparticle coordinates*. Note that the quasiparticle coordinate $\mathbf{x}_j^{\alpha}$ is invariant under the exchange of any two position vectors in $\{\mathbf{r}_{/j}^{\alpha}\} = \{\mathbf{r}_1^{\alpha}, \ldots, \mathbf{r}_{j-1}^{\alpha}, \mathbf{r}_{j+1}^{\alpha}, \ldots, \mathbf{r}_{N^{\alpha}}^{\alpha}\}$ or in $\{\mathbf{r}^{\bar{\alpha}}\} = \{\mathbf{r}_1^{\bar{\alpha}}, \ldots, \mathbf{r}_{N^{\bar{\alpha}}}^{\bar{\alpha}}\}$ [13].

The backflow transformation replaces every single-particle orbital $\phi_i^{\alpha}(\mathbf{r}_j^{\alpha})$ by a transformed orbital $\phi_i^{\alpha}(\mathbf{x}_j^{\alpha})$, which depends on the position of every electron in the system. Exchanging the coordinates of any two spin-parallel electrons still exchanges two rows of the Slater determinant, so the antisymmetry is preserved. The downside is that moving one electron now changes every element of the Slater matrix, preventing the use of efficient rank-1 update formulas and increasing the cost of reevaluating the determinant by a factor of N. Despite the extra cost, however, the enrichment of the description of correlations between electrons makes SJB wave functions significantly better than SJ wave functions, and they are frequently used in VMC and FNDMC simulations.

The FermiNet [30] takes the idea of permutation equivariant backflow much further, replacing the orbitals $\phi_i^{\alpha}(\mathbf{r}_j^{\alpha})$ entirely by neural networks. The orbitals represented by these networks differ from SJB orbitals, because they are not functions of a single three-dimensional vector $\mathbf{x}_j^{\alpha}$ but depend in a very general way on $\mathbf{r}_j^{\alpha}$ and all the elements of the sets $\{\mathbf{r}_{/j}^{\alpha}\}$ and $\{\mathbf{r}^{\bar{\alpha}}\}$. They are best written as $\phi_i^{\alpha}(\mathbf{r}_j^{\alpha}; \{\mathbf{r}_{/j}^{\alpha}\}; \{\mathbf{r}^{\bar{\alpha}}\})$. The exchange antisymmetry is maintained, because $\phi_i^{\alpha}(\mathbf{r}_j^{\alpha}; \{\mathbf{r}_{/j}^{\alpha}\}; \{\mathbf{r}^{\bar{\alpha}}\})$ is totally symmetric on exchange of any pair of coordinates in $\{\mathbf{r}_{/j}^{\alpha}\}$ or $\{\mathbf{r}^{\bar{\alpha}}\}$. Furthermore, because they are represented as neural networks, the FermiNet orbitals need not be expanded in terms of an explicit basis set, widening the class of functions they can represent [44]. In order to build functions with the correct exchange symmetry properties, a carefully constructed neural-network architecture is used, which is described below.

The FermiNet architecture consists of two parts: the one-electron stream, which takes electron-nucleus separation vectors $\mathbf{r}_i^{\alpha} - \mathbf{R}_I$ and distances $|\mathbf{r}_i^{\alpha} - \mathbf{R}_I|$ as inputs, and the two-electron stream, which takes electron-electron separations $\mathbf{r}_i^{\alpha} - \mathbf{r}_j^{\beta}$ and distances $|\mathbf{r}_i^{\alpha} - \mathbf{r}_j^{\beta}|$ as inputs, with $i, j \in \{1, 2, \ldots, N^{\alpha}\}$ and $\alpha, \beta \in \{\uparrow, \downarrow\}$. The inputs to the one-electron stream are concatenated to form one input vector for each electron, and the inputs to the two-electron stream are concatenated to form one input vector for each pair of electrons:

$$\mathbf{h}_i^{0\alpha} = (\mathbf{r}_i^{\alpha} - \mathbf{R}_I, |\mathbf{r}_i^{\alpha} - \mathbf{R}_I| \ \forall \ I), \tag{4}$$

$$\mathbf{h}_{ij}^{0\alpha\beta} = (\mathbf{r}_i^{\alpha} - \mathbf{r}_j^{\beta}, |\mathbf{r}_i^{\alpha} - \mathbf{r}_j^{\beta}|), \tag{5}$$

where the superscript 0 means that the vectors are the inputs to the first layer of the network. The distances between particles are passed into the network to help it to model the wave function cusps, i.e., the discontinuities in the derivatives of the wave function when two electrons or an electron and a nucleus coincide. These discontinuities create divergences in the kinetic energy that exactly cancel the divergences in the potential energy as pairs of charged particles approach each other [30].

Each electron stream consists of several layers. At each layer $l \in \{0, \ldots, L-1\}$, the outputs $\mathbf{h}_i^{l\alpha}$ and $\mathbf{h}_{ij}^{l\alpha\beta}$ from the streams are averaged and concatenated in the following way:

$$\begin{aligned} \mathbf{f}_i^{l\alpha} &= (\mathbf{h}_i^{l\alpha}, \mathbf{g}^{l\uparrow}, \mathbf{g}^{l\downarrow}, \mathbf{g}_i^{l\alpha\uparrow}, \mathbf{g}_i^{l\alpha\downarrow}), \\ \mathbf{g}^{l\uparrow} &= \frac{1}{N^{\uparrow}} \sum_{j=1}^{N^{\uparrow}} \mathbf{h}_j^{l\uparrow}, \qquad \mathbf{g}^{l\downarrow} = \frac{1}{N^{\downarrow}} \sum_{j=1}^{N^{\downarrow}} \mathbf{h}_j^{l\downarrow}, \\ \mathbf{g}_i^{l\alpha\uparrow} &= \frac{1}{N^{\uparrow}} \sum_{j=1}^{N^{\uparrow}} \mathbf{h}_{ij}^{l\alpha\uparrow}, \qquad \mathbf{g}_i^{l\alpha\downarrow} = \frac{1}{N^{\downarrow}} \sum_{j=1}^{N^{\downarrow}} \mathbf{h}_{ij}^{l\alpha\downarrow}. \end{aligned} \tag{6}$$

The concatenated one-electron vectors are then passed into the next layer, as are the two-electron vectors:

$$\mathbf{h}_i^{(l+1)\alpha} = \tanh(\mathbf{V}^l \mathbf{f}_i^{l\alpha} + \mathbf{b}^l) + \mathbf{h}_i^{l\alpha}, \tag{7}$$

$$\mathbf{h}_{ij}^{(l+1)\alpha\beta} = \tanh\left(\mathbf{W}^l \mathbf{h}_{ij}^{l\alpha\beta} + \mathbf{c}^l\right) + \mathbf{h}_{ij}^{l\alpha\beta}, \tag{8}$$

where $\mathbf{V}^l$ and $\mathbf{W}^l$ are matrices, $\mathbf{b}^l$ and $\mathbf{c}^l$ are vectors, and all of them are optimizable. We denote the number of hidden units in each layer in the one-electron stream by n_l such that $\mathbf{h}_i^{l\alpha} \in \mathbb{R}^{n_l}$, $l \in \{0, 1, 2, \ldots, L\}$ [45]. The outputs from the final layer L of the one-electron streams are used to build the many-particle FermiNet orbitals:

$$\phi_i^{k\alpha}(\mathbf{r}_j^{\alpha}; \{\mathbf{r}_{/j}^{\alpha}\}; \{\mathbf{r}^{\bar{\alpha}}\}) = (\mathbf{w}_i^{k\alpha} \cdot \mathbf{h}_j^{L\alpha} + g_i^{k\alpha}) \chi_i^{k\alpha}(\mathbf{r}_j^{\alpha}), \tag{9}$$

where $\mathbf{w}_i^{k\alpha}$ is an optimizable vector and $g_i^{k\alpha}$ an optimizable scalar. The $\chi_i^{k\alpha}(\mathbf{r}_j^\alpha)$ factor is an envelope function to ensure that the wave function satisfies the relevant boundary conditions. For example, in a system which requires the wave function to tend to zero as $|\mathbf{r}_j^\alpha - \mathbf{R}_m| \to \infty$, exponential envelopes are used:

$$\chi_i^{k\alpha}(\mathbf{r}_j^\alpha) = \sum_m \pi_{im}^{k\alpha} \exp(-\sigma_{im}^{k\alpha}|\mathbf{r}_j^\alpha - \mathbf{R}_m|), \tag{10}$$

where $\pi_{im}^{k\alpha}$ and $\sigma_{im}^{k\alpha}$ are variational parameters. No attempt is made to ensure that the FermiNet orbitals are normalized or orthogonal to each other.

As mentioned earlier, FermiNet orbitals are not functions of one electron position $\mathbf{r}_i^\alpha$ only but also depend on the positions of all the other electrons in the system in an appropriately permutation-invariant way. No Jastrow factor is needed, as the electron-electron correlations are included in the network. The full wave function is, thus, a block-diagonal determinant of the FermiNet orbitals $\phi_i^{k\alpha}(\mathbf{r}_j^\alpha; \{\mathbf{r}_{/j}^\alpha\}; \{\mathbf{r}^{\bar{\alpha}}\})$. Multiple determinants may also be used, in which case the wave function is a weighted linear combination:

$$\begin{aligned}\Psi^D_{\text{Slater FermiNet}}&\left(\mathbf{r}_1^\uparrow, \ldots, \mathbf{r}_{N^\uparrow}^\uparrow, \mathbf{r}_1^\downarrow, \ldots, \mathbf{r}_{N^\downarrow}^\downarrow\right) \\ &= \sum_k^D \omega_k \det\left[\phi_i^{k\uparrow}(\mathbf{r}_j^\uparrow; \{\mathbf{r}_{/j}^\uparrow\}; \{\mathbf{r}^\downarrow\})\right] \\ &\quad \times \det\left[\phi_i^{k\downarrow}(\mathbf{r}_j^\downarrow; \{\mathbf{r}_{/j}^\downarrow\}; \{\mathbf{r}^\uparrow\})\right],\end{aligned} \tag{11}$$

where the superscript D specifies the number of determinants of FermiNet orbitals in the linear combination of determinants that makes up the full wave function and the "Slater FermiNet" subscript serves to specify this specific wave function *Ansatz* and is discussed in more detail below. In practice, the weights ω_k are absorbed into the orbitals, which are not normalized.

The VMC method is then applied to the FermiNet *Ansatz* and the parameters of the network are optimized using a second-order method known as the Kronecker-factored approximate curvature algorithm [46]. The aim is to minimize the expectation value of the Hamiltonian $\langle H \rangle$, which acts as our loss function. For a more detailed explanation of the FermiNet architecture, see Pfau *et al.* [30] and the discussion of the improved JAX implementation [47] in Spencer *et al.* [36].

The FermiNet architecture can be extended to study the periodic system [37,38,48]. Consider the basis $\{\mathbf{a_1}, \mathbf{a_2}, \mathbf{a_3}\}$ of the Bravais lattice generated by repeating the finite simulation cell periodically. Any position vector may be written as $\mathbf{r} = s_1\mathbf{a_1} + s_2\mathbf{a_2} + s_3\mathbf{a_3}$. To ensure that the FermiNet represents a periodic function, the position coordinates s_i are replaced in the FermiNet inputs by pairs of periodic functions: $s_i \to [\sin(2\pi s_i), \cos(2\pi s_i)]$. Thus, if any electron is moved by any simulation-cell Bravais lattice vector, the inputs to the network are unchanged. It follows that the output, the value of the wave function, is also unchanged. A periodic envelope function is used to improve the speed of convergence [38]:

$$\chi_i^{k\alpha}(\mathbf{r}_j^\alpha) = \sum_m [\nu_{im}^{k\alpha} \cos(\mathbf{k}_m \cdot \mathbf{r}_j^\alpha) + \mu_{im}^{k\alpha} \sin(\mathbf{k}_m \cdot \mathbf{r}_j^\alpha)], \tag{12}$$

where the $\mathbf{k}_m$ are simulation-cell reciprocal lattice vectors up to the Fermi wave vector of the noninteracting Fermi gas. This specific way of adapting the FermiNet to periodic systems was proposed by Cassella *et al.* [38], although other similar methods exist [37,48].

The FermiNet has been used to study only systems of electrons interacting via Coulomb forces to date but can easily be adapted to systems of other spin-1/2 particles simply by changing the Hamiltonian. Here, we use the periodic FermiNet *Ansatz* to approximate the ground state of the UFG Hamiltonian in a cubic box subject to periodic boundary conditions. Since there are no atomic nuclei and the wave function has no electron-nuclear cusps, the inputs to the one-electron streams are simpler than shown in Eq. (4), containing only the particle coordinates [49]: $\mathbf{h}_i^{0\alpha} = \mathbf{r}_i^\alpha$, with respect to an origin placed at one corner of the simulation cell. A detailed discussion of translational symmetry of the wave function can be found in Appendix E.

As is demonstrated below, the Slater FermiNet is sufficient to learn the superfluid ground state for small systems but fails for large systems. Hence, we propose a modification to the method of building orbitals. The motivation for this modification comes from earlier work using antisymmetrized products of two-particle orbitals known as antisymmetrized geminal power (AGP) wave functions [15,29,39,40,42,50,51]. We describe the antisymmetrized geminal power singlet (AGPs) wave function in the next section.

The authors of Ref. [30] used the term "FermiNet wave function" to refer to all wave functions constructed using a FermiNet neural network. Now that we are going to use almost the same neural network to generate AGPs-based pairing wave functions in addition, more precise terminology is required. Wave functions of the type introduced in Ref. [30], which contain many-particle generalizations of the one-particle orbitals that appear in Slater determinants, are called one-determinant or multideterminant Slater FermiNets. Wave functions built using determinants of many-particle generalizations of pairing functions are called one-determinant or multideterminant AGPs FermiNets. Since every AGPs FermiNet determinant is built using one pairing function or "geminal," we also refer to one-geminal or multigeminal AGPs FermiNets.

III. ANTISYMMETRIZED GEMINAL POWER WAVE FUNCTION

The FermiNet and other *Ansätze* that expand the ground state as a linear combination of Slater determinants give very accurate results for many molecules and solids but may still fail to capture strong two-particle correlations in superfluids. An alternative starting point, which is better at capturing two-particle correlations, is the AGP wave function [40–42,52]. This uses an antisymmetrized product of two-particle functions known as pairing orbitals or geminals instead of an antisymmetrized product of single-particle orbitals.

Although one can build a general AGP wave function with pairings between arbitrary particles, the UFG Hamiltonian contains only interactions between particles of opposite spin. It is, therefore, sufficient to consider pairing orbitals involving particles of opposite spin only. In this case, the wave function is called an AGPs. The rest of this section summarizes the main features of the AGPs *Ansatz* and explains how the FermiNet architecture can be modified to produce many-particle generalizations of AGPs pairing orbitals. Detailed discussions of AGPs wave functions, including derivations of the equations, can be found in Refs. [29,40–42,52] and Appendixes A–F.

A. AGP singlet wave functions

It is helpful to start by considering an unpolarized system with an even number ($N = 2p$) of particles and total spin $S_z = 0$. An AGPs wave function for such a system is constructed using a singlet pairing function of the form

$$\Phi(\mathbf{r}_i,\sigma_i;\mathbf{r}_j,\sigma_j) = \varphi(\mathbf{r}_i,\mathbf{r}_j) \times \langle\sigma_i\sigma_j|\frac{1}{\sqrt{2}}(|\uparrow\downarrow\rangle - |\downarrow\uparrow\rangle), \quad (13)$$

where $\varphi(\mathbf{r}_i,\mathbf{r}_j)$ is a symmetric function of its arguments. We work with spin-assigned wave functions, so we set the spins of particles $1, 2, \ldots, p$ to $\uparrow$ and the spins of particles $p+1, p+2, \ldots, 2p$ to $\downarrow$. If, for example, $i \le p$ and $j > p$, so that particle i is spin up and particle j is spin down, the spin-assigned pairing function is

$$\Phi(\mathbf{r}_i,\uparrow;\mathbf{r}_j,\downarrow) = \varphi(\mathbf{r}_i,\mathbf{r}_j)/\sqrt{2}. \quad (14)$$

The spin-assigned singlet pairing function is equal to zero if the spins of particles i and j are the same.

The spin-assigned AGPs wave function is a determinant of spatial pairing functions [15,39]:

$$\Psi_{\mathrm{AGPs}}(\mathbf{r}_1^\uparrow,\ldots,\mathbf{r}_p^\uparrow,\mathbf{r}_1^\downarrow,\ldots,\mathbf{r}_p^\downarrow) = \det[\varphi(\mathbf{r}_i^\uparrow,\mathbf{r}_j^\downarrow)]. \quad (15)$$

Like all spin-assigned wave functions, it depends on position coordinates only. For convenience, we change the particle labeling scheme: i and j now run from 1 to p, and arrow superscripts are added to distinguish up-spin from down-spin particles. Note that the AGPs wave function coincides with the BCS wave function projected onto a fixed particle number subspace (see Ref. [39] and Appendix A 1). It is, therefore, suitable for describing singlet-paired systems, including s-wave superfluids.

B. AGPs with unpaired states

We can generalize the spin-assigned AGPs wave function to allow for unpaired particles. Consider a system with $N = 2p + u + d$ particles, where p is the number of pairs, u is the number of unpaired spin-up particles, and d is the number of unpaired spin-down particles. The total number of spin-up particles is $p + u$, and the total number of spin-down particles is $p + d$. The AGPs wave function can be written as a determinant of pairing functions and single-particle orbitals [39,42,52] as shown in Eq. (16):

$$\Psi(1,2,\ldots,2p+u+d) = \det\begin{pmatrix} \varphi(\mathbf{r}_1^\uparrow,\mathbf{r}_1^\downarrow) & \cdots & \varphi(\mathbf{r}_1^\uparrow,\mathbf{r}_{p+d}^\downarrow) & \phi_1^\uparrow(\mathbf{r}_1^\uparrow) & \cdots & \phi_u^\uparrow(\mathbf{r}_1^\uparrow) \\ \varphi(\mathbf{r}_2^\uparrow,\mathbf{r}_1^\downarrow) & \cdots & \varphi(\mathbf{r}_2^\uparrow,\mathbf{r}_{p+d}^\downarrow) & \phi_1^\uparrow(\mathbf{r}_2^\uparrow) & \cdots & \phi_u^\uparrow(\mathbf{r}_2^\uparrow) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \varphi(\mathbf{r}_{p+u}^\uparrow,\mathbf{r}_1^\downarrow) & \cdots & \varphi(\mathbf{r}_{p+u}^\uparrow,\mathbf{r}_{p+d}^\downarrow) & \phi_1^\uparrow(\mathbf{r}_{p+u}^\uparrow) & \cdots & \phi_u^\uparrow(\mathbf{r}_{p+u}^\uparrow) \\ \phi_1^\downarrow(\mathbf{r}_1^\downarrow) & \cdots & \phi_1^\downarrow(\mathbf{r}_{p+d}^\downarrow) & 0 & \cdots & 0 \\ \phi_2^\downarrow(\mathbf{r}_1^\downarrow) & \cdots & \phi_2^\downarrow(\mathbf{r}_{p+d}^\downarrow) & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \phi_d^\downarrow(\mathbf{r}_1^\downarrow) & \cdots & \phi_d^\downarrow(\mathbf{r}_{p+d}^\downarrow) & 0 & \cdots & 0 \end{pmatrix}, \quad (16)$$

where $\varphi(\mathbf{r}_i^{\uparrow}, \mathbf{r}_j^{\downarrow})$ is an arbitrary singlet pairing function and $\phi_i^{\sigma_i}(\mathbf{r}_j^{\sigma_i})$ are arbitrary single-particle functions. For the UFG considered in this paper, we need only the case where $u = 1$ and $d = 0$ or vice versa. This represents a fully paired $2p$-particle system to which one particle has been added.

IV. AGP SINGLET FERMINET

Having discussed the form of the AGPs wave function, we now discuss how it can be implemented using FermiNet. In the original Slater FermiNet architecture, the outputs of the one-electron stream are used to build FermiNet orbitals $\phi_i^{k\alpha}(\mathbf{r}_j^{\alpha}; \{\mathbf{r}_{/j}^{\alpha}\}; \{\mathbf{r}^{\bar{\alpha}}\})$. The full many-particle wave function is a weighted sum of terms, each of which is the product of one up-spin and one down-spin determinant of the FermiNet orbital matrices, as shown in Eq. (11).

To build a many-particle pairing function using the neural-network part of FermiNet, one can make use of its outputs $\mathbf{h}_i^{L\alpha}$ from the last layer L of the one-electron stream. Instead of using these outputs to build FermiNet orbitals, as in Eq. (9), they can be used to build FermiNet pairing orbitals, also known as FermiNet geminals:

$$\varphi^k(\mathbf{r}_i^{\alpha}, \mathbf{r}_j^{\bar{\alpha}}; \{\mathbf{r}_{/i}^{\alpha}\}; \{\mathbf{r}_{/j}^{\bar{\alpha}}\}) = [\mathbf{w}^k \cdot (\mathbf{h}_i^{L\alpha} \odot \mathbf{h}_j^{L\bar{\alpha}}) + g^k]\chi^k(\mathbf{r}_i^{\alpha})\chi^k(\mathbf{r}_j^{\bar{\alpha}}), \quad (17)$$

where $\chi^k(\mathbf{r})$ are the envelope functions, $\mathbf{w}^k$ are vectors, g^k is a scalar, and $\odot$ denotes the elementwise product. Note that the same FermiNet geminal is used for all pairs of particles, so the envelope functions in Eq. (17) do not require the particle and spin indices that appear in the envelope functions of the FermiNet orbitals defined in Eq. (9). This construction generates a many-particle pairing function between particles $\mathbf{r}_i^{\alpha}$ and $\mathbf{r}_j^{\bar{\alpha}}$, retaining the permutation-invariant property possessed by FermiNet orbitals. Depending on the number of FermiNet geminals generated, the wave function can be written as one or a weighted sum of multiple determinants of FermiNet geminals:

$$\Psi^D_{\text{AGPs FermiNet}}\left(\mathbf{r}_1^{\uparrow}, \ldots, \mathbf{r}_{N^{\downarrow}}^{\downarrow}\right) = \sum_k^D \omega_k \det\left[\varphi^k(\mathbf{r}_i^{\alpha}, \mathbf{r}_j^{\bar{\alpha}}; \{\mathbf{r}_{/i}^{\alpha}\}; \{\mathbf{r}_{/j}^{\bar{\alpha}}\})\right], \quad (18)$$

where the superscript D in $\Psi^D_{\text{AGPs FermiNet}}$ specifies the number of determinants (and, thus, FermiNet geminals) appearing in the linear combination that makes up the wave function. This is analogous to a weighted sum of conventional single-determinant AGPs wave functions of the type defined in Eq. (15), but the replacement of the two-particle pairing orbitals by FermiNet geminals that depend on the positions of all the particles makes it much more general.

Although using the outputs from the one-electron stream is sufficient to build an AGPs, one can also include the outputs from the two-electron stream:

$$\varphi^k(\mathbf{r}_i^{\alpha}, \mathbf{r}_j^{\bar{\alpha}}; \{\mathbf{r}_{/i}^{\alpha}\}; \{\mathbf{r}_{/j}^{\bar{\alpha}}\}) = [\mathbf{w}_1^k \cdot (\mathbf{h}_i^{L\alpha} \odot \mathbf{h}_j^{L\bar{\alpha}}) + \mathbf{w}_2^k \cdot (\mathbf{h}_{ij}^{L\alpha\bar{\alpha}} \odot \mathbf{h}_{ji}^{L\bar{\alpha}\alpha}) + g^k]\chi^k(\mathbf{r}_i^{\alpha})\chi^k(\mathbf{r}_j^{\bar{\alpha}}), \quad (19)$$

where $\mathbf{w}_1^k$ and $\mathbf{w}_2^k$ are vectors.

Note that Eqs. (17) and (19) are two possible ways of building a many-particle pairing function. There are many other ways, and they are all valid as long as the appropriate symmetries are preserved. An alternative method is given by Xie *et al.* [53].

The benefit of building AGPs-like wave functions using the FermiNet is that the many-particle pairing function $\varphi(\mathbf{r}_i^{\alpha}, \mathbf{r}_j^{\bar{\alpha}}; \{\mathbf{r}_{/i}^{\alpha}\}; \{\mathbf{r}_{/j}^{\bar{\alpha}}\})$ now depends not only on $\mathbf{r}_i^{\alpha}$ and $\mathbf{r}_j^{\bar{\alpha}}$, but also on the positions of the other particles in the system [54]. Correlations between the singlet pair and the other particles can, thus, be captured. In a similar way, the original Slater FermiNet replaces Hartree-Fock-like single-particle orbitals $\phi_i^{\alpha}(\mathbf{r}_j^{\alpha})$ by many-particle orbitals (FermiNet orbitals) $\phi_i^{\alpha}(\mathbf{r}_j^{\alpha}; \{\mathbf{r}_{/j}^{\alpha}\}; \{\mathbf{r}^{\bar{\alpha}}\})$, helping to capture correlations between the particle at $\mathbf{r}_j^{\alpha}$ and all other particles.

A. Relations between the Slater FermiNet and the AGPs FermiNet

Next, we clarify the relation between the Slater FermiNet with block-diagonal determinants and the AGPs FermiNet, showing that the AGPs FermiNet is the more general of the two. A FermiNet geminal with a two-particle stream term is even more general than a FermiNet geminal without, so it is sufficient for this purpose to omit the two-particle stream term. We also neglect the envelope functions and the bias term g^k, which is set to 0 in all results presented here. The use of envelope functions circumvents numerical difficulties in finite systems and can speed up the network optimization but does not affect the generality of the *Ansatz*.

Let us first define a many-particle pairing function in the following way:

$$\varphi(\mathbf{r}_i^{\uparrow}, \mathbf{r}_j^{\downarrow}; \{\mathbf{r}_{/i}^{\uparrow}\}; \{\mathbf{r}_{/j}^{\downarrow}\}) = \sum_{kl} W_{kl} h_i^{L\uparrow(k)} h_j^{L\downarrow(l)}, \quad (20)$$

where $h_i^{L\alpha(k)} = [\mathbf{h}_i^{L\alpha}]_k$ are the outputs from the final layer of the one-electron stream for particle i of spin α. As we explain below, Eq. (20) is equivalent to the simpler FermiNet geminal described above:

$$\varphi(\mathbf{r}_i^{\uparrow}, \mathbf{r}_j^{\downarrow}; \{\mathbf{r}_{/i}^{\uparrow}\}; \{\mathbf{r}_{/j}^{\downarrow}\}) = \mathbf{w} \cdot (\mathbf{h}_i^{L\uparrow} \odot \mathbf{h}_j^{L\downarrow}). \quad (21)$$

We choose to write the many-particle pairing function in the form of Eq. (20) only because this makes it easier to relate to FermiNet orbitals. Since Eqs. (20) and (21) are equivalent, the choice does not affect the conclusions of the argument. In the rest of this section, for the sake of simplicity, we omit the sets $\{\mathbf{r}_{/i}^{\uparrow}\}$ and $\{\mathbf{r}_{/j}^{\downarrow}\}$ from the arguments of the many-particle pairing functions and orbitals.

To explain the equivalence of Eqs. (20) and (21), it is helpful to represent the matrix W_{kl} as its singular-value decomposition:

$$W_{kl} = \sum_{\gamma=1}^{n_L} \sigma_\gamma U_{\gamma k} V_{\gamma l}, \quad (22)$$

where $U \in \mathbb{R}^{n_L \times n_L}$ and $V \in \mathbb{R}^{n_L \times n_L}$ are orthogonal matrices and n_L is the size of the vector $\mathbf{h}_i^{L\alpha}$ output by the final layer L of the one-electron stream. This is also known as the number of hidden units in layer L. The many-particle pairing function in Eq. (20) becomes

$$\varphi(\mathbf{r}_i^{\uparrow}, \mathbf{r}_j^{\downarrow}) = \sum_{kl} W_{kl} h_i^{L\uparrow(k)} h_j^{L\downarrow(l)} \quad (23)$$

$$= \sum_{\gamma} \sigma_\gamma \left(\sum_k U_{\gamma k} h_i^{L\uparrow(k)} \right) \left(\sum_l V_{\gamma l} h_j^{L\uparrow(l)} \right) \quad (24)$$

$$= \sum_{\gamma} \sigma_\gamma (U \mathbf{h}_i^{L\uparrow})_\gamma (V \mathbf{h}_j^{L\downarrow})_\gamma. \quad (25)$$

Given the universal approximation theorem [44] and the fact that every layer of the network contains an arbitrary linear transformation, it is reasonable to assume that the functions $\mathbf{h}_i^{L\alpha}$ and $O\mathbf{h}_i^{L\alpha}$, where O is U or V, have the same variational freedom and information content. In other words, we assume that any network capable of representing $\mathbf{h}_i^{L\alpha}$ can also represent $O\mathbf{h}_i^{L\alpha}$, since this is merely a rotation of the vectors $\mathbf{h}_i^{L\alpha}$ in the last layer. We, thus, define $\tilde{\mathbf{h}}_i^{L\uparrow} = U\mathbf{h}_i^{L\uparrow}$ and $\tilde{\mathbf{h}}_i^{L\downarrow} = V\mathbf{h}_i^{L\downarrow}$, such that the many-particle pairing function becomes

$$\varphi(\mathbf{r}_i^{\uparrow}, \mathbf{r}_j^{\downarrow}) = \sum_{\gamma} \sigma_\gamma (U \mathbf{h}_i^{L\uparrow})_\gamma (V \mathbf{h}_j^{L\downarrow})_\gamma$$
$$= \sum_{\gamma} \sigma_\gamma (\tilde{\mathbf{h}}_i^{L\uparrow} \odot \tilde{\mathbf{h}}_j^{L\downarrow})_\gamma, \quad (26)$$

which is equivalent to Eq. (21).

To relate the AGPs FermiNet and the Slater FermiNet, we expand an AGPs determinant constructed using the many-particle pairing function from Eq. (20) as a sum of block-diagonal determinants of FermiNet orbitals. It will be sufficient to consider matrices $W \in \mathbb{R}^{n_L \times n_L}$ of rank M, with $p \le M \le n_L$. We can decompose any such matrix using rank factorization:

$$W_{kl} = \sum_{\gamma=1}^{M} F_{\gamma k} G_{\gamma l}, \quad (27)$$

where F and G are matrices in $\mathbb{R}^{M \times n_L}$. Equation (20) then becomes

$$\varphi(\mathbf{r}_i^{\uparrow}, \mathbf{r}_j^{\downarrow}) = \sum_{\gamma=1}^{M} \left(\sum_{k=1}^{n_L} F_{\gamma k} h_i^{L\uparrow(k)} \right) \left(\sum_{l=1}^{n_L} G_{\gamma l} h_j^{L\downarrow(l)} \right)$$
$$= \sum_{\gamma=1}^{M} \phi_\gamma^{\uparrow}(\mathbf{r}_i^{\uparrow}) \phi_\gamma^{\downarrow}(\mathbf{r}_j^{\downarrow}), \quad (28)$$

where the last line defines the functions $\phi_\gamma^{\uparrow}$ and $\phi_\gamma^{\downarrow}$. In the case when $M = p$, where $p = N/2$ is the number of pairs in the system, the determinant of the many-particle pairing function can be written as a block-diagonal determinant of FermiNet orbitals:

$$\det\left[\varphi(\mathbf{r}_i^{\uparrow}, \mathbf{r}_j^{\downarrow})\right] = \det\left[\sum_{\gamma=1}^{p} \phi_\gamma^{\uparrow}(\mathbf{r}_i^{\uparrow}) \phi_\gamma^{\downarrow}(\mathbf{r}_j^{\downarrow})\right]$$
$$= \det\left[(\phi^{\uparrow T} \phi^{\downarrow})_{ij}\right]$$
$$= \det\left[\phi_i^{\uparrow}(\mathbf{r}_j^{\uparrow})\right] \det\left[\phi_i^{\downarrow}(\mathbf{r}_j^{\downarrow})\right], \quad (29)$$

where $[\phi^\alpha]_{ij} = \phi_i^\alpha(\mathbf{r}_j^\alpha)$ are matrices in $\mathbb{R}^{M \times p}$ with $M = p$. The product of two $p \times p$ determinants can be written as the determinant of a single $2p \times 2p$ matrix, with the $p \times p$ spin-up and spin-down blocks on the diagonal. Therefore, a single-geminal AGPs FermiNet wave function constructed using the many-particle pairing function from Eq. (20) with a rank-p matrix W_{kl} is equivalent to a $2p \times 2p$ block-diagonal determinant of FermiNet orbitals. The equivalence is already well known [51] for AGPs wave functions constructed using conventional two-particle orbitals.

Now consider the more general case where $p \le M \le n_L$. The Cauchy-Binet formula states that

$$\det\left[\varphi(\mathbf{r}_i^{\uparrow},\mathbf{r}_j^{\downarrow})\right]=\det\left[\sum_{\gamma=1}^{M}\phi_\gamma^{\uparrow}(\mathbf{r}_i^{\uparrow})\phi_\gamma^{\downarrow}(\mathbf{r}_j^{\downarrow})\right]$$
$$=\sum_{1\le j_1<j_2<\cdots<j_p\le M}\det\begin{pmatrix}\phi_{j_1}^{\uparrow}(\mathbf{r}_1^{\uparrow}) & \phi_{j_1}^{\uparrow}(\mathbf{r}_2^{\uparrow}) & \cdots & \phi_{j_1}^{\uparrow}(\mathbf{r}_p^{\uparrow})\\ \phi_{j_2}^{\uparrow}(\mathbf{r}_1^{\uparrow}) & \phi_{j_2}^{\uparrow}(\mathbf{r}_2^{\uparrow}) & \cdots & \phi_{j_2}^{\uparrow}(\mathbf{r}_p^{\uparrow})\\ \vdots & \vdots & \ddots & \vdots\\ \phi_{j_p}^{\uparrow}(\mathbf{r}_1^{\uparrow}) & \phi_{j_p}^{\uparrow}(\mathbf{r}_2^{\uparrow}) & \cdots & \phi_{j_p}^{\uparrow}(\mathbf{r}_p^{\uparrow})\end{pmatrix}\det\begin{pmatrix}\phi_{j_1}^{\downarrow}(\mathbf{r}_1^{\downarrow}) & \phi_{j_1}^{\downarrow}(\mathbf{r}_2^{\downarrow}) & \cdots & \phi_{j_1}^{\downarrow}(\mathbf{r}_p^{\downarrow})\\ \phi_{j_2}^{\downarrow}(\mathbf{r}_1^{\downarrow}) & \phi_{j_2}^{\downarrow}(\mathbf{r}_2^{\downarrow}) & \cdots & \phi_{j_2}^{\downarrow}(\mathbf{r}_p^{\downarrow})\\ \vdots & \vdots & \ddots & \vdots\\ \phi_{j_p}^{\downarrow}(\mathbf{r}_1^{\downarrow}) & \phi_{j_p}^{\downarrow}(\mathbf{r}_2^{\downarrow}) & \cdots & \phi_{j_p}^{\downarrow}(\mathbf{r}_p^{\downarrow})\end{pmatrix},\tag{30}$$

where the sum is over all $\binom{M}{p}$ distinct choices of p rows from the two $M\times p$ matrices $\phi_\gamma^{\uparrow}(\mathbf{r}_i^{\uparrow})$ and $\phi_\gamma^{\downarrow}(\mathbf{r}_i^{\downarrow})$. The products of the determinants of the two $p\times p$ matrices associated with each such choice are summed to reproduce the AGPs. This is similar to the linear combination of multiple block-diagonal determinants of FermiNet orbitals without weights given by Eq. (11) and in the original FermiNet paper [30,55].

Note that the intermediate layers, i.e., the one- and two-electron streams, are identical in the Slater FermiNet and the AGPs FermiNet. The only modifications are made at the orbital shaping layer, or, equivalently, the method of antisymmetrization has changed. Thus, the representational power of the intermediate layers of the AGPs FermiNet remains the same as for the Slater FermiNet. Thus, it must be the method of antisymmetrization that limits the performance of the Slater FermiNet when applied to the UFG.

We have shown that a single AGPs determinant constructed using the many-particle pairing function from Eq. (20) with a matrix W_{kl} of rank greater than p contains multiple block-diagonal determinants of FermiNet orbitals. If the rank of W_{kl} is equal to p, the AGPs is equivalent to a single block-diagonal FermiNet determinant. Conversely, any single-determinant FermiNet wave function can be written as an AGPs of rank p. Therefore, the AGPs FermiNet provides a more powerful *Ansatz* with fewer variational parameters than the Slater FermiNet, since the former contains the latter.

It is worth mentioning another advantage of using the FermiNet to build geminals. By generating more sets of independent parameters $\mathbf{w}_i^k$ in Eq. (19), one can easily construct an arbitrary number N_{det} of FermiNet geminals $\varphi^k(\mathbf{r}_i^{\alpha},\mathbf{r}_j^{\bar{\alpha}};\{\mathbf{r}_{/i}^{\alpha}\};\{\mathbf{r}_{/j}^{\bar{\alpha}}\})$ with $k\in\{1,2,\ldots,N_{\mathrm{det}}\}$, all without the use of a basis set. This allows one to use weighted sums of AGPs determinants as trial wave functions, similar to the weighted sum of conventional FermiNet determinants seen in Eq. (11).

B. AGPs FermiNet with unpaired states

To extend the AGPs FermiNet to systems with unpaired states, such as an odd number of particles system, we use FermiNet geminals and orbitals to replace both the pairing orbitals and the single-particle orbitals in Eq. (16). In this work, we consider systems with equal numbers of up-spin and down-spin particles, which are assumed to be fully paired, and systems containing one additional unpaired particle, which may have spin up or spin down. For example, the AGPs FermiNet with an extra spin-up particle is given by

$$\Psi_{\mathrm{FermiNet}}^{\mathrm{AGPs}}\left(\mathbf{r}_1^{\uparrow},\ldots,\mathbf{r}_{p+1}^{\uparrow},\mathbf{r}_1^{\downarrow},\ldots,\mathbf{r}_p^{\downarrow}\right)$$
$$=\sum_k\omega_k\det\begin{pmatrix}\varphi^k(\mathbf{r}_1^{\uparrow},\mathbf{r}_1^{\downarrow};\{\mathbf{r}_{/1}^{\uparrow}\};\{\mathbf{r}_{/1}^{\downarrow}\}) & \cdots & \varphi^k(\mathbf{r}_1^{\uparrow},\mathbf{r}_p^{\downarrow};\{\mathbf{r}_{/1}^{\uparrow}\};\{\mathbf{r}_{/p}^{\downarrow}\}) & \phi_1^{k\uparrow}(\mathbf{r}_1^{\uparrow};\{\mathbf{r}_{/1}^{\uparrow}\};\{\mathbf{r}^{\downarrow}\})\\ \varphi^k(\mathbf{r}_2^{\uparrow},\mathbf{r}_1^{\downarrow};\{\mathbf{r}_{/2}^{\uparrow}\};\{\mathbf{r}_{/1}^{\downarrow}\}) & \cdots & \varphi^k(\mathbf{r}_2^{\uparrow},\mathbf{r}_p^{\downarrow};\{\mathbf{r}_{/2}^{\uparrow}\};\{\mathbf{r}_{/p}^{\downarrow}\}) & \phi_1^{k\uparrow}(\mathbf{r}_2^{\uparrow};\{\mathbf{r}_{/2}^{\uparrow}\};\{\mathbf{r}^{\downarrow}\})\\ \vdots & \ddots & \vdots & \vdots\\ \varphi^k(\mathbf{r}_{p+1}^{\uparrow},\mathbf{r}_1^{\downarrow};\{\mathbf{r}_{/p+1}^{\uparrow}\};\{\mathbf{r}_{/1}^{\downarrow}\}) & \cdots & \varphi^k(\mathbf{r}_{p+1}^{\uparrow},\mathbf{r}_p^{\downarrow};\{\mathbf{r}_{/p+1}^{\uparrow}\};\{\mathbf{r}_{/p}^{\downarrow}\}) & \phi_1^{k\uparrow}(\mathbf{r}_{p+1}^{\uparrow};\{\mathbf{r}_{/p+1}^{\uparrow}\};\{\mathbf{r}^{\downarrow}\})\end{pmatrix},\tag{31}$$

where $\varphi^k(\mathbf{r}_i^\uparrow, \mathbf{r}_j^\downarrow; \{\mathbf{r}_{/i}^\uparrow\}; \{\mathbf{r}_{/j}^\downarrow\})$ can be defined as either Eq. (17) or Eq. (19) and $\phi_i^{k\uparrow}(\mathbf{r}_j^\uparrow; \{\mathbf{r}_{/j}^\uparrow\}; \{\mathbf{r}^\downarrow\})$ is defined in Eq. (9). In practice, we generate the required number of FermiNet geminal and orbital parameters in batch at the orbital projection layer. For example, one FermiNet geminal and one FermiNet orbital are generated per determinant for a system with one extra spin-up particle.

V. RESULTS

The power of the AGPs FermiNet *Ansatz* may be demonstrated by studying the UFG. The Hamiltonian is

$$\hat{H} = -\frac{1}{2}\sum_i^N \nabla_i^2 + \sum_{ij}^{N/2} U(\mathbf{r}_i^\uparrow - \mathbf{r}_j^\downarrow), \tag{32}$$

where

$$U(\mathbf{r}) = -\frac{2v_0\mu^2}{\cosh^2(\mu r)} \tag{33}$$

is the modified Pöschl-Teller potential, which is widely used in variational and diffusion QMC simulations [15–22]. It would be preferable to use a delta function interaction with an infinite s-wave scattering length, but it is difficult to simulate systems with deltalike potentials using QMC methods. Thus, a finite but short-ranged interaction is typically used. The s-wave scattering length of the Pöschl-Teller potential diverges when $v_0 = 1$. By changing the value of μ at fixed $v_0 = 1$, it is possible to vary the effective range of the interaction, $r_e = 2/\mu$, while holding the s-wave scattering length infinite.

We choose to study a system with density parameter $r_s = 1$, where r_s, the radius of a sphere that contains one particle on average, provides a convenient measure of the interparticle distance. Throughout this work, we employ the dimensionless system based on Hartree atomic units: The unit of length is the Bohr radius a_0, and the unit of energy is the Hartree. To ensure that the range of the interaction is small compared with the interparticle separation, we set $\mu = 12$ ($r_e = 1/6$), keeping $v_0 = 1$ to ensure that the scattering length remains infinite [56]. We also simulate the system with $k_F = 1$ [equivalent to $r_s = (9\pi/4)^{1/3} \approx 1.92$] and $\mu = 12$ to compare with the FNDMC result from Forbes, Gandolfi, and Gezerlis [20].

We use both the Slater FermiNet with multiple block-diagonal determinants and the AGPs FermiNet with multiple geminals to study the unitary Fermi gas from $N = 4$ to $N = 38$ particles in a cubic box subject to periodic boundary conditions, as well as AGPs FermiNet on the $N = 66$ system [57]. The same network size, number of determinants, and number of training iterations are used for both *Ansätze*. The FermiNet orbitals are given by Eq. (9) without the bias term. The FermiNet geminal used for systems containing from 4 to 28 particles is the one defined in Eq. (17). Unless otherwise stated, all calculations used a linear combination of 32 determinants or 32 geminals. Including contributions from the two-electron stream improves the optimization rate and can achieve a slightly lower variational energy in larger systems, so Eq. (19) is used for systems of $N \geq 29$. The inclusion of plane-wave envelopes as defined in Eq. (12) also improves the optimization rate. For molecular and electron gas systems, we find that the bias term in the FermiNet orbital projection [Eq. (9)] does not affect the accuracy or optimization of the model. We, hence, set the bias term, $g_i^{k\alpha}$ and g^k as appropriate, to zero for all calculations presented here.

A comparison of the ground-state energy expectation values given by the two *Ansätze* is shown in Fig. 1(a). The Slater FermiNet, which consists of a linear combination of block-diagonal determinants of FermiNet orbitals, performs well when the number of particles N is smaller than around 10, but the AGPs FermiNet is much superior in larger systems. It is clear that the Slater FermiNet *Ansatz* has difficulties learning the ground states of large paired systems [60].

In systems containing an odd number of particles, one must be left unpaired. This raises the energy a little and explains the zigzag shape in Fig. 1(a). The odd-even staggering is lost for larger systems with the Slater FermiNet *Ansatz*, indicating the absence of pair formation [15,61]. The Slater FermiNet fails to learn the superfluid state. For the AGPs FermiNet, by contrast, the amplitude of the odd-even zigzag remains constant, superposed on the linear increase with N expected of any extensive quantity.

Another comparison between the two *Ansätze* is shown in Fig. 1(b), which depicts the ratio of the interacting and noninteracting energies per particle, known as the Bertsch parameter [3] and defined in Eq. (1), as a function of N. All FermiNet energies are variational and the noninteracting energies are exact, so the AGPs FermiNet, for which the Bertsch parameter is lower by up to around 30%, is much the better of the two *Ansätze*.

We next compare our results with the state-of-the-art FNDMC results of Forbes, Gandolfi, and Gezerlis [20], shown in Fig. 2 for the case $k_F = 1$ and $\mu = 12$. The AGPs FermiNet achieves a lower energy per particle than FNDMC for all system sizes except for $N = 4$ and $N = 6$. The dependence of the Bertsch parameter on system size is also smoother when calculated with the AGPs FermiNet [62]. A full training curve of $N = 66$ with comparison to the FNDMC energy can be found in Fig. 8 in Appendix F.

The pairing gap may be found using the approximation formula [14,61]

$$\Delta = (-1)^N \left[E(N+1) - \frac{1}{2}[E(N) + E(N+2)]\right], \tag{34}$$

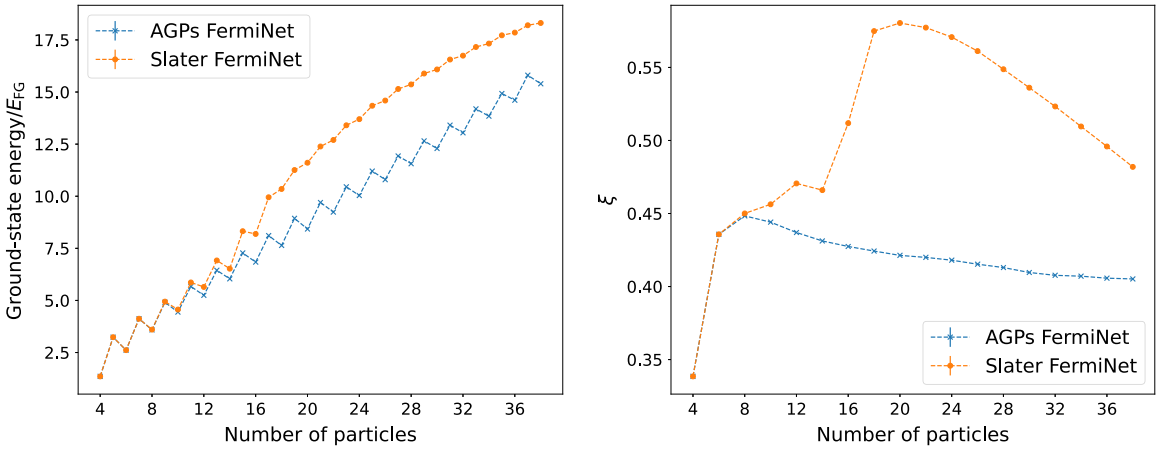

FIG. 1. Comparison between results obtained using the AGPs FermiNet and the Slater FermiNet for different numbers of particles N, with $r_s = 1$ and $\mu = 12$. All simulations use 32 determinants, 300 000 optimization steps, and the same hyperparameters, which are detailed in the Appendixes A–F. (a) The total energy of the UFG simulation cell, measured in units of the free Fermi gas energy E_{FG}. The Slater FermiNet Ansatz begins to fail when $N \gtrapprox 10$. (b) The Bertsch parameter ξ (the ratio of the interacting and noninteracting ground-state energies per particle) as a function of the number of particles N.

where N is the total number of particles in the box. The results from $N = 4$ to $N = 36$ are shown in Fig. 3. Also shown is the thermodynamic ($N \to \infty$) limit of the BCS pairing gap including Gorkov's polarization correction [63]:

$$\Delta_{\mathrm{BCS}} = \frac{8}{e^2}\frac{\hbar^2 k_F^2}{2m}\exp\left(\frac{\pi}{2k_F a}\right), \tag{35}$$

$$\Delta_{\mathrm{Gorkov}} = \frac{1}{(4e)^{1/3}}\Delta_{\mathrm{BCS}}. \tag{36}$$

Here, a is the scattering length of the interaction, which is infinite in the UFG. In this limit, $\Delta_{\mathrm{BCS}} = 1.804E_{\mathrm{FG}}$ and $\Delta_{\mathrm{Gorkov}} = 0.815E_{\mathrm{FG}}$, where $E_{\mathrm{FG}} = \frac{3}{5}(\hbar^2 k_F^2/2m)$ is the average energy per particle of an unpolarized noninteracting Fermi gas and e is Euler's number [64]. The UFG is a

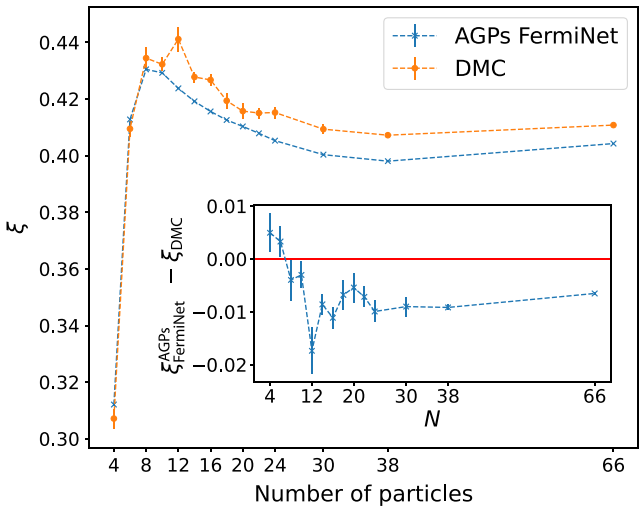

FIG. 2. Comparison of the system-size-dependent values of the Bertsch parameter, ξ, as calculated using the AGPs FermiNet and FNDMC, with $k_F = 1$ and $\mu = 12$. According to the variational principle, lower values are better. The error bars on the AGPs FermiNet results are smaller than the sizes of the crosses. Inset: difference between the AGPs and FNDMC values of the Bertsch parameter. The errors in the inset are obtained by adding the standard errors of the FNDMC and AGPs FermiNet results in quadrature. The latter are obtained by computing the standard error of the Markov chain Monte Carlo-averaged Bertsch parameter accumulated over 50 000 inference steps.

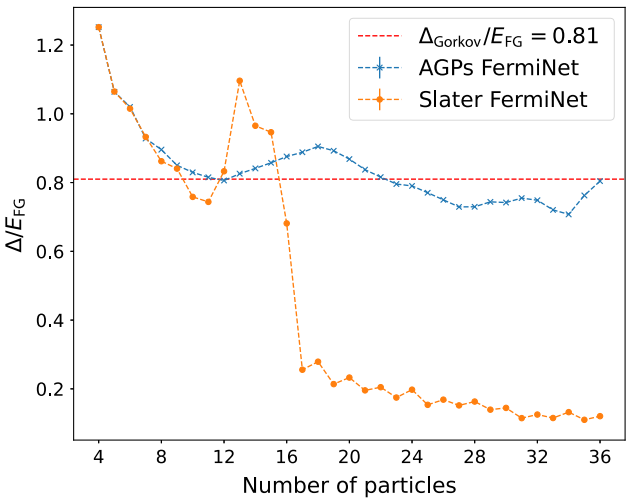

FIG. 3. Pairing gaps calculated with the Slater FermiNet and the AGPs FermiNet for different numbers of particles N, with $r_s = 1$ and $\mu = 12$.

strongly coupled system, so the BCS and Gorkov estimates of the gap need not be accurate.

The striking collapse of the pairing gap with increasing system size shows that the Slater FermiNet *Ansatz* struggles to describe paired states in systems of more than ten particles. The AGPs FermiNet *Ansatz* behaves much better, although the oscillations with system size suggest that significant finite-size errors remain even for the largest systems simulated.

Another signature of fermionic superfluidity is the presence of off-diagonal long-ranged order in the two-body density matrix (TBDM):

$$\rho^{(2)}_{\uparrow\downarrow}(\mathbf{r}_1,\mathbf{r}_2;\mathbf{r}'_1,\mathbf{r}'_2) = \langle \hat{\psi}^\dagger_\uparrow(\mathbf{r}_1)\hat{\psi}^\dagger_\downarrow(\mathbf{r}_2)\hat{\psi}_\downarrow(\mathbf{r}'_2)\hat{\psi}_\uparrow(\mathbf{r}'_1)\rangle, \quad (37)$$

the largest eigenvalue of which diverges as the number of particles N tends to infinity [65]. The superfluid condensate fraction c may be obtained by evaluating [66]

$$c = \lim_{r\to\infty} \frac{\Omega^2}{N_\uparrow} \rho^{(2)\mathrm{TR}}_{\uparrow\downarrow}(r), \quad (38)$$

where Ω is the volume of the simulation cell, $N_\uparrow$ is the number of spin-up particles, and $\rho^{(2)\mathrm{TR}}_{\uparrow\downarrow}(r)$ is the rotational and translational average of the TBDM:

$$\rho^{(2)\mathrm{TR}}_{\uparrow\downarrow}(r) = \frac{1}{4\pi r^2\Omega^2}\int \rho^{(2)}_{\uparrow\downarrow}(\mathbf{r}_1,\mathbf{r}_2;\mathbf{r}_1+\mathbf{r}',\mathbf{r}_2+\mathbf{r}') \times \delta(|\mathbf{r}'|-r)d\mathbf{r}_1 d\mathbf{r}_2 d\mathbf{r}'. \quad (39)$$

The one-body density matrix, by contrast, tends to zero in the $r\to\infty$ limit [65]. A full discussion of the methods used to evaluate the condensate fraction in QMC simulations can be found in the Appendixes A–F and the CASINO manual [66].

After fully training both the Slater FermiNet and the AGPs FermiNet for the $N = 38$ particle system, we use the resulting neural wave functions to compute the quantity $(\Omega^2/N_\uparrow)\rho^{(2)\mathrm{TR}}_{\uparrow\downarrow}(r)$. The results are shown in Fig. 4, which provide further evidence that the Slater FermiNet fails to converge to the superfluid ground state; the quantity $(\Omega^2/N_\uparrow)\rho^{(2)\mathrm{TR}}_{\uparrow\downarrow}(r)$ appears to be approaching zero in the large pair-separation limit, implying that the condensate fraction is also zero. The same quantity for the AGPs FermiNet approaches a finite value, which we estimate to be roughly $c = 0.44(1)$ using the eight data points with separations $r/r_s \geq 2.0$. This value is consistent with previous estimations from experiments and the most recent AFMC value from Ref. [21] (Table I).

In addition, we also estimate the condensate fractions for the $N = 66$ UFG at a fixed $\mu = 12$ with two different densities: $r_s = 1$ ($k_F r_e = 0.32$) and $k_F = 1$ ($k_F r_e = 0.17$), respectively. We compute the quantity $(\Omega^2/N_\uparrow)\rho^{(2)\mathrm{TR}}_{\uparrow\downarrow}(r)$ at

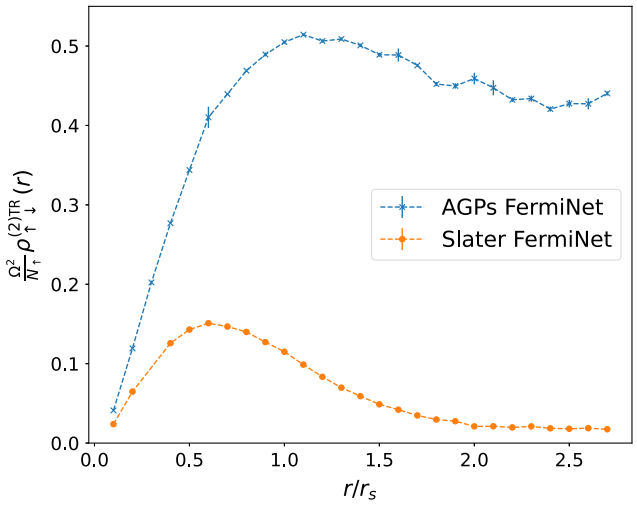

FIG. 4. Comparison of the TBDM estimators calculated using the AGPs FermiNet and the Slater FermiNet with $N = 38$, $r_s = 1$, and $\mu = 12$. The error bars show the standard error of the TBDM estimator, accumulated over 2000 inference steps. Most of the error bars are so small that they are obscured the symbols.

five sequentially spaced separations r near $r = L/2$, where the quantity has approached its asymptotic value. We then take the average of the five data points to get estimated values of the condensate fraction. [71]. Our estimate of the condensate fraction for $N = 66$ is $c = 0.42(1)$ at $k_F r_e = 0.32$ and $c = 0.52(1)$ at $k_F r_e = 0.17$, which are both consistent with the experiments. The results are summarized in Table I.

Although VMC methods are generally considered to be less accurate than FNDMC methods, an important advantage of VMC methods is that almost any expectation value, including any reduced density matrix, may be estimated without bias. The same is not true of FNDMC simulations,

TABLE I. Estimates of the superfluid condensate fraction at unitarity using various methods. The quantity $k_F r_e$ is a dimensionless number, indicating the deviation of the simulated system from a perfect UFG with zero-range interaction.

Method	Value
Our estimate for $N = 38$ at $k_F r_e = 0.32$	0.44(1)
Our estimate for $N = 66$ at $k_F r_e = 0.32$	0.42(1)
Our estimate for $N = 66$ at $k_F r_e = 0.17$	0.52(1)
FNDMC for $N = 38$ at $k_F r_e = 0.03$ [67]	0.61(2)
FNDMC for $N = 66$ at $k_F r_e = 0.03$ [67]	0.57(2)
FNDMC for $N = 128$ with VMC extrapolation at $k_F r_e = 0.32$ [17]	0.51
FNDMC with $k_F r_e \to 0$ extrapolation for $N = 66$ [68]	0.56(1)
AFMC with $k_F r_e \to 0$ extrapolation for $N = 66$ [21]	0.43(2)
Experiment [69]	0.46(7)
Experiment [70]	0.47(7)

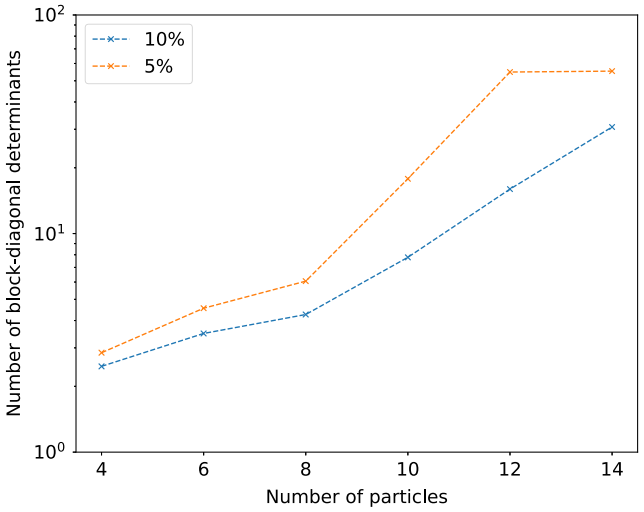

FIG. 5. Number of block-diagonal determinants required for the multideterminant Slater FermiNet energy to fall within 5% and 10% of the energy obtained using a single-determinant or single-geminal AGPs FermiNet.

which sample the wave function instead of its square modulus and produce biased "one-sided" estimates of the expectation values of operators that do not commute with the Hamiltonian [13]. Thus, there are very few unbiased and accurate first-principles calculations of the condensate fraction. Our approach, having both the advantages of VMC and surpassing the accuracy of DMC, provides solutions to these problems and a more accurate way to estimate general expectation values.

Finally, we study how the number of block-diagonal determinants required to achieve a given accuracy scales with the number of particles in the system. We choose six even-particle systems from $N = 4$ to $N = 14$ and compare the energies obtained using linear combinations of multiple block-diagonal FermiNet Slater determinants against energies obtained using a *single-determinant* (and, thus, single-geminal) FermiNet AGPs wave function. All other hyperparameters are as given in Table II (Appendix B 1). In Fig. 5, we show that the number of block-diagonal determinants required to achieve a given percentage accuracy increases approximately exponentially with the number of particles. Plots for each individual system, along with a more detailed discussion, can be found in Appendix C. These results suggest that multideterminant Slater FermiNet wave functions constructed using a neural network of fixed size are incapable of describing the ground state of the UFG accurately unless the number of block-diagonal determinants rises exponentially with system size. Hence, in practice, the AGPs FermiNet is required for studying paired systems.

VI. DISCUSSION

In this work, we used neural wave functions to study the benchmark superfluid system known as the UFG [72]. We showed that the Slater FermiNet *Ansatz* has difficulties in describing paired systems with strong, short-ranged attractive interactions between particles of opposite spin. Hence, we proposed a way to improve the variational *Ansätze* by using determinants of FermiNet geminals, similar to an AGPs or a BCS wave function. We showed mathematically that the Slater FermiNet is a limiting case of the AGPs FermiNet despite the use of fewer parameters in the latter. It follows that any FermiNet wave function can, in principle, be written as an AGPs FermiNet wave function.

We compared the total energies and energies per particle of the UFG as calculated using the Slater FermiNet and the AGPs FermiNet. The former fails to produce a paired state when the number of particles, N, is greater than around 10, while the AGPs FermiNet works very well.

As the UFG has a superfluid ground state, we computed the pairing gap and condensate fraction for the $N = 38$ system and compared estimates made with the Slater FermiNet and the AGPs FermiNet. There is a clear qualitative difference between the pairing gap obtained using the AGPs FermiNet and the Slater FermiNet, with the latter approaching zero as the number of particles N increases. Calculations of the superfluid condensate fraction show a similar behavior: The AGPs FermiNet gives an accurate finite result, while the value obtained using the Slater FermiNet tends to zero in the limit of large system size. Although the AGPs pairing gap shows significant finite-size errors, it lies close to the mean-field BCS result with Gorkov-Melik-Barkhudarov corrections [63]. Taken together, these results show that the Slater FermiNet is unable to represent large systems with superfluid ground states. The AGPs FermiNet is much more suitable for studying paired systems such as the UFG.

To demonstrate the success of the AGPs FermiNet, we also compared our calculated total energies with state-of-the-art fixed-node diffusion QMC energies obtained using a Jastrow-BCS *Ansatz* [20]. For all systems with more than a few particles, the AGPs FermiNet achieves lower (i.e., better) variational energies than FNDMC using the same model interaction and system parameters.

The inability of the Slater FermiNet *Ansatz* to accurately describe the UFG ground state comes as a surprise, because the original FermiNet paper [30] argued that any many-body fermionic wave function can be represented as a single determinant of FermiNet orbitals. However, the mathematical argument relies on the construction of FermiNet orbitals with unphysical discontinuities. Whether or not any wave function can be represented as a single determinant of FermiNet orbitals of the type used in practice, which are differentiable everywhere except at electron-electron and electron-nuclear coalescence points, remains an open question. Another limitation is that the architecture of the FermiNet neural network, which is rather simple, may not be able to represent an arbitrary many-electron FermiNet orbital. Even if a single-determinant Slater FermiNet wave

function is general in principle, there is no guarantee that it is equally easy to represent all wave functions. It may be that producing an accurate representation of a paired wave function requires the width and number of layers in the neural network to increase rapidly with system size. Furthermore, if a network of fixed size is used, it may be necessary to increase the number of Slater FermiNet determinants rapidly as the system size increases. The observation that the Slater FermiNet works well when $N \lessapprox 10$ but that the quality of the results degrades rapidly for larger systems, along with the scaling study presented in the final part of Sec. V, suggests that this is, in fact, the case.

The AGPs FermiNet introduced in this paper shares many of the strengths of the Slater FermiNet. In particular, there is no need to construct and optimize a new basis set for every new system or particle type. If the AGPs FermiNet proves equally successful in other paired systems, it may now be relatively easy to investigate the importance of pairing in molecules, electron-positron systems, electron-hole liquids, and other s-wave superfluids. Another strength of the AGPs FermiNet is the ease with which it is possible to optimize linear combinations of determinants of FermiNet pairing orbitals, such as the one in Eq. (31). This is much more difficult to accomplish with conventional wave functions based on explicit two-electron pairing orbitals or pairing orbitals represented as outer products of single-particle orbitals or basis functions. Just as the many-particle orbitals in a Slater FermiNet radically generalize single-particle orbitals by incorporating electron-electron terms in a permutation-equivariant fashion, so the pairing functions in an AGPs FermiNet generalize BCS-style pairing functions by incorporating the effects of the remaining electrons in a permutation-equivariant fashion.

The AGPs FermiNet introduced here has a straightforward Pfaffian extension and can, thus, be applied to non-s-wave and triplet pairing. Therefore, we expect it to become a powerful tool for understanding strongly correlated non-s-wave superfluid and superconducting systems such as helium-3 or high-T_c and p-wave superconductors. Finally, our approach is not limited to the FermiNet neural network and can be readily adapted to use more recent architectures such as the Psiformer [74], GLOBE and MOON [75], and DeepErwin [76].

ACKNOWLEDGMENTS

We thank Stefano Gandolfi and Michael M. Forbes for providing the diffusion Monte Carlo data. We gratefully acknowledge the European Union's PRACE program for the award of computing resources on the JUWELS Booster supercomputer in Jülich; the HPC RIVR consortium and EuroHPC JU for resources on the Vega high performance computing system at IZUM, the Institute of Information Science in Maribor; and the United Kingdom Engineering and Physical Sciences Research Council for resources on the Baskerville Tier 2 HPC service. Baskerville was funded by the EPSRC and UKRI through the World Class Labs scheme (EP/T022221/1) and the Digital Research Infrastructure program (EP/W032244/1) and is operated by Advanced Research Computing at the University of Birmingham. We also gratefully acknowledge the Gauss Centre for Supercomputing e.V. for funding this project by providing computing time through the John von Neumann Institute for Computing (NIC) on the GCS Supercomputer JUWELS at Jülich Supercomputing Centre (JSC). J. K. is part of the Munich Quantum Valley, which is supported by the Bavarian state government with funds from the Hightech Agenda Bayern Plus. W. T. L. is supported by an Imperial College President's Ph.D. Scholarship; H. S. is supported by the Aker Scholarship; and G. C. is supported by the United Kingdom Engineering and Physical Sciences Research Council (EP/T51780X/1). We also acknowledge the support of the Imperial-TUM flagship partnership.

APPENDIX A: ANTISYMMETRIC GEMINAL POWER WAVE FUNCTION

This section discusses the relation between the AGPs wave function and the BCS wave function.

1. Fixed-particle number BCS ground state

The AGPs wave function may be obtained by projecting the BCS ground-state wave function [12] into a subspace of fixed particle number [39]. We start with the BCS ground state:

$$\begin{aligned}|\Psi_{\mathrm{BCS}}\rangle &= \prod_{\mathbf{k}}(u_{\mathbf{k}} + v_{\mathbf{k}}\hat{c}^{\dagger}_{\mathbf{k},\uparrow}\hat{c}^{\dagger}_{-\mathbf{k},\downarrow})|0\rangle \\ &= \left(\prod_{\mathbf{k}} u_{\mathbf{k}}\right) e^{\sum_{\mathbf{k}} \varphi_{\mathbf{k}}\hat{c}^{\dagger}_{\mathbf{k},\uparrow}\hat{c}^{\dagger}_{-\mathbf{k},\downarrow}}|0\rangle,\end{aligned}$$

where $\varphi_{\mathbf{k}} = (v_{\mathbf{k}}/u_{\mathbf{k}})$. Ignoring global phase factors and coefficients, the fixed-particle BCS wave function can be written as

$$|\Psi_{\mathrm{PBCS}}\rangle = \left(\sum_{\mathbf{k}} \varphi_{\mathbf{k}}\hat{c}^{\dagger}_{\mathbf{k},\uparrow}\hat{c}^{\dagger}_{-\mathbf{k},\downarrow}\right)^{p}|0\rangle, \tag{A1}$$

where $p = N/2$ is the number of pairs in the system and N is the number of electrons. After Fourier transforming this wave function, we get the real space wave function [39]:

$$\Psi_{\mathrm{PBCS}} = \mathcal{A}[\Phi(1,1')\Phi(2,2')\ldots\Phi(p,p')], \tag{A2}$$

where $\mathcal{A}$ is the antisymmetrizer, the wave function corresponding to $\Phi(i,i')$ is

$$\Phi(i,i') \equiv \Phi(\mathbf{r}_i,\sigma_i;\mathbf{r}_{i'}\sigma_{i'}) = \varphi(\mathbf{r}_i,\mathbf{r}_{i'}) \times \langle\sigma_i\sigma_{i'}|\frac{1}{\sqrt{2}}(|\uparrow\downarrow\rangle - |\downarrow\uparrow\rangle), \quad \text{(A3)}$$

and $\varphi(\mathbf{r}_i,\mathbf{r}_{i'}) = \sum_{\mathbf{k}} \varphi_{\mathbf{k}} e^{i\mathbf{k}\cdot(\mathbf{r}_i-\mathbf{r}_{i'})}$ is the Fourier transform of $\varphi_{\mathbf{k}}$.

As mentioned in the main text, Eq. (A2) can be written as a determinant of the pairing function $\varphi(\mathbf{r}_i,\mathbf{r}_{i'})$ after spin assignment [39,40]:

$$\Psi_{\mathrm{PBCS}} = \det\left[\varphi(\mathbf{r}_i^{\uparrow},\mathbf{r}_j^{\downarrow})\right]. \quad \text{(A4)}$$

APPENDIX B: EXPERIMENTAL SETUP

In this section, we report the FermiNet setup and hyperparameters used in this work.

1. FermiNet

The periodic version of the FermiNet implemented by Cassella *et al.* [38], which can be found in the FermiNet repository [77], is used as the basis for the AGPs code. The small modifications required to support AGPs wave functions are made using the JAX Python library [47]. For the majority of our calculations, four NVIDIA A100 GPUs are used. For systems with $N > 30$ particles, we use four nodes with a total of 16 A100 GPUs to speed up the calculations. A JAX implementation of the Kronecker-factored approximate curvature (KFAC) gradient descent algorithm [46,78] is used for optimization. The initial parameters of the network are initialized using Xavier (random) initialization [79], and the positions of the particles are initialized uniformly in the simulation box. We do not observe significant run-to-run variation in the final energy of a given system as a function of the random initialization. The FermiNet hyperparameters are shown in Table II and the network sizes in Table III. All training runs use 3×10^5 iterations to ensure convergence, except for the $N = 66$ system (see Appendix F). When evaluating expectation values with an optimized wave function, 5×10^4 inference steps are used.

TABLE II. Hyperparameters used in all simulations. "std. dev." stands for standard deviation.

Parameter	Value
Batch size	4096
Training iterations	3×10^5
Pretraining iterations	None
Learning rate	$(2 \times 10^4 + t)^{-1}$
Local energy clipping	5.0
KFAC momentum	0
KFAC covariance moving average decay	0.95
KFAC norm constraint	1×10^{-3}
KFAC damping	1×10^{-3}
Markov chain Monte Carlo proposal std. dev. (per dimension)	0.02
Markov chain Monte Carlo steps between parameter updates	10

TABLE III. Network sizes and number of determinants used in all simulations. The corresponding mathematical symbols mentioned in the main text of the paper, where available, are also listed.

Parameter	Symbol	Value
One-electron stream network size	n_l	512
Two-electron stream network size	⋯	64
Number of network layers	L	4
Number of determinants	D	32

2. Estimation of the two-body density matrix

The TBDM in first quantized notation can be written as

$$\rho^{(2)}_{\alpha\beta}(\mathbf{r}_1,\mathbf{r}_2;\mathbf{r}_1',\mathbf{r}_2') = N_\alpha(N_\beta - \delta_{\alpha\beta})\frac{\int |\Psi(\mathbf{R})|^2 \frac{\Psi(\mathbf{r}_1',\mathbf{r}_2')}{\Psi(\mathbf{r}_1,\mathbf{r}_2)} d\mathbf{r}_3 \ldots d\mathbf{r}_N}{\int |\Psi(\mathbf{R})|^2 d\mathbf{R}}, \quad \text{(B1)}$$

where α and β denote the spin or particle species. The superfluid condensate fraction in a finite and periodic system is defined as

$$c = \frac{\Omega^2}{N_\alpha}\lim_{r\to\infty}\rho^{(2)\mathrm{TR}}_{\alpha\beta}(r), \quad \text{(B2)}$$

where Ω is the volume of the simulation cell, N_α is the number of particles with spin α, and $\rho^{(2)\mathrm{TR}}_{\alpha\beta}(r)$ is the translational and rotational average of the TBDM given in Eq. (39).

The one-body density matrix (OBDM) is expected to tend to zero as $r \to \infty$. However, because of finite-size effects, the OBDM is not necessarily zero within our simulation cell. We, therefore, use an improved estimator in Eq. (B3) that removes the one-body contribution explicitly [66]:

$$\rho^{(2)}_{\alpha\beta}(\mathbf{r}_1,\mathbf{r}_2;\mathbf{r}_1',\mathbf{r}_2') = N_\alpha(N_\beta - \delta_{\alpha\beta}) \times \frac{\int |\Psi(\mathbf{R})|^2 \left[\frac{\Psi(\mathbf{r}_1',\mathbf{r}_2')}{\Psi(\mathbf{r}_1,\mathbf{r}_1)} - \frac{\Psi(\mathbf{r}_1',\mathbf{r}_2)}{\Psi(\mathbf{r}_1,\mathbf{r}_2)}\frac{\Psi(\mathbf{r}_1,\mathbf{r}_2')}{\Psi(\mathbf{r}_1,\mathbf{r}_2)}\right] d\mathbf{r}_3 \ldots d\mathbf{r}_N}{\int |\Psi(\mathbf{R})|^2 d\mathbf{R}}. \quad \text{(B3)}$$

This quantity can then be estimated using Monte Carlo sampling.

APPENDIX C: HOW MANY BLOCK-DIAGONAL DETERMINANTS DOES THE SLATER FERMINET NEED TO ACHIEVE THE SAME ACCURACY AS THE AGPS FERMINET WITH ONE DETERMINANT?

In the main text, we demonstrate that the Slater FermiNet with 32 block-diagonal determinants is able to capture superfluidity in small systems but fails at larger systems. Therefore, we are interested in the scaling of the original block-diagonal determinant FermiNet wave function with respect to the system size; i.e., how many block-diagonal determinants do we need in order for the Slater FermiNet to converge to the ground state at each system size?

To answer this question, first we set the AGPs FermiNet with one determinant energies as baselines and plot the percentage difference of the Slater FermiNet from the baseline against the number of block-diagonal determinants used in the Slater FermiNet wave function, repeated at different system sizes from 4 to 14 even-particle systems. The results are shown in Fig. 6.

As the results suggest, it becomes more difficult for the Slater FermiNet to get close to the AGPs FermiNet baseline as the number of determinant increases, especially for larger systems. This is due to the limited performance of the optimizer as the number of determinant increases. Hence, due to the constraints on time and resources, it is not feasible to continue to increase the number of determinants until the Slater FermiNet achieves the same accuracy as the AGPs FermiNet. Instead, we decide to set two thresholds for the percentage difference between the two results. Here, we use 5% and 10% as the thresholds, plotted as the two horizontal lines in Fig. 6. By plotting the x intercept of the curves with the two threshold lines against system size, we can determine the relationship between the two, as shown in Fig. 5. As the y axis in Fig. 5 is in logarithmic scale, a roughly linear relationship suggests an exponential scaling of the number of block-diagonal determinants as system size increases.

The result indicates that, in theory, the Slater FermiNet is capable of converging to the ground state given the number of block-diagonal determinants is sufficiently large. In practice, the number of block-diagonal determinants required to learn the ground state increases exponentially as the system size gets bigger and the number gets inaccessible rapidly.

APPENDIX D: DENSE DETERMINANT

Previous works [32,38,74,80–82] have suggested that the use of dense determinants

$$\Psi^{\text{Dense}}_{\text{FermiNet}} = \det\begin{pmatrix} \phi_i^{\uparrow}(\mathbf{r}_j^{\uparrow};\{\mathbf{r}_{/j}^{\uparrow}\};\{\mathbf{r}^{\downarrow}\}) & \phi_i^{\uparrow}(\mathbf{r}_j^{\downarrow};\{\mathbf{r}_{/j}^{\downarrow}\};\{\mathbf{r}^{\uparrow}\}) \\ \phi_i^{\downarrow}(\mathbf{r}_j^{\uparrow};\{\mathbf{r}_{/j}^{\uparrow}\};\{\mathbf{r}^{\downarrow}\}) & \phi_i^{\downarrow}(\mathbf{r}_j^{\downarrow};\{\mathbf{r}_{/j}^{\downarrow}\};\{\mathbf{r}^{\uparrow}\}) \end{pmatrix}, \tag{D1}$$

in contrast to block-diagonal diagonal determinants with FermiNet in Eq. (11), provides a slight gain in accuracy in

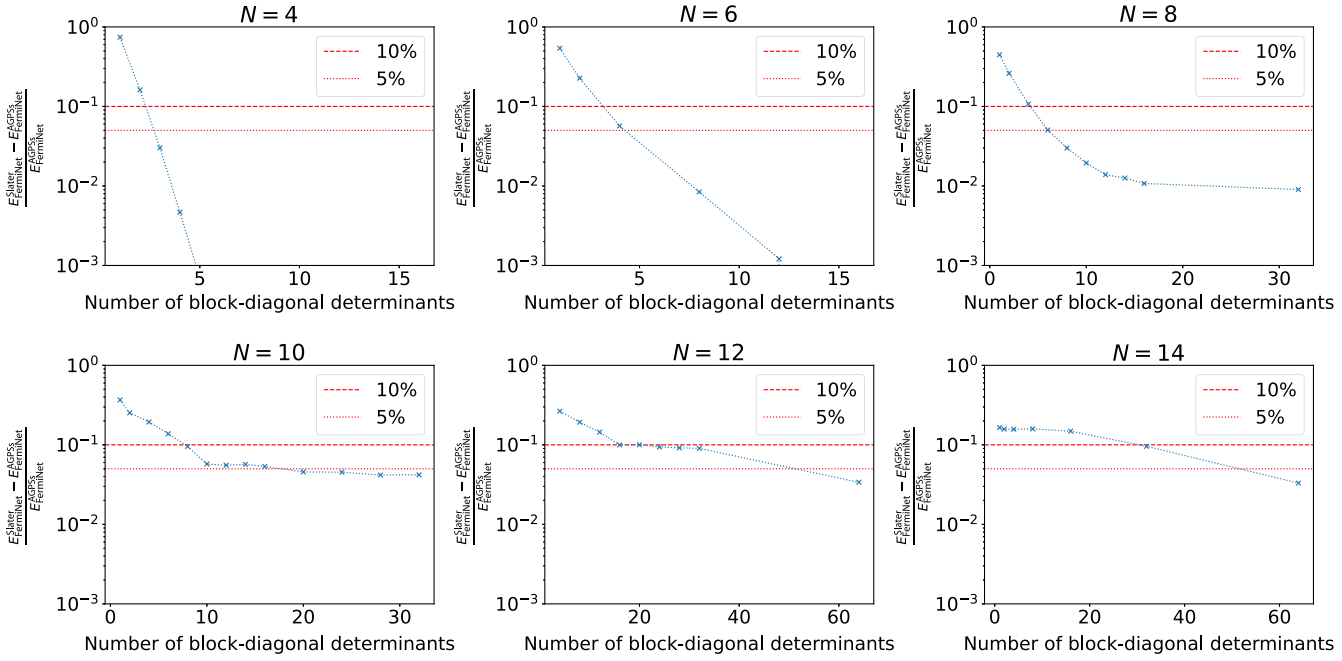

FIG. 6. Percentage difference between the Slater FermiNet with various number of block-diagonal determinants and the AGPs FermiNet with one determinant at different system sizes, from 4 to 14 particles.

TABLE IV. Total energy of the UFG with 36–38 particles and their corresponding pairing gap using different wave functions.

	Total energy (E_{FG})			
Type of wave function	$N = 36$	$N = 37$	$N = 38$	Pairing gap (E_{FG})
Block	17.8523(5)	18.2032(5)	18.3131(6)	0.121(2)
Dense	17.7027(5)	18.1601(5)	18.2570(5)	0.180(2)
AGPs	14.6059(4)	15.8060(5)	15.3975(5)	0.804(1)

various infinite and periodic systems. In this subsection, we present a small set of calculations of the UFG with dense determinants and compare the results with the block-diagonal FermiNet and the AGPs FermiNet. The results are presented in Table IV.

Although the use of dense determinants in the UFG does provide lower energies comparing to the block-diagonal determinants, they are still significantly higher than the AGPs FermiNet energies. In addition, the qualitative behaviors, such as the absence of odd-even staggering, are still similar to the block-diagonal FermiNet, which are qualitatively different from the AGPs FermiNet.

APPENDIX E: TRANSLATIONAL SYMMETRY OF THE AGPS FERMINET

In this section, we discuss the consequences of setting an origin in the simulation cell.

As mentioned in the main text, in all of our simulations, an origin is set at one corner of the simulation cell. The particle coordinates $\mathbf{r}_i^\alpha$, which are used as inputs to the one-electron stream network, are taken with respect to this origin. Thus, the AGPs FermiNet *Ansatz* does not impose translational invariance with respect to simultaneous translations of all particle coordinates.

As the model unitary Fermi gas system has a translationally invariant ground state, it is natural to use a translationally invariant *Ansatz* to study it. While it is possible to neglect the one-electron stream network and use only the two-electron stream network, which is translationally invariant by construction, in the AGPs FermiNet simulations, we find that doing so gives significantly worse energies than the standard AGPs FermiNet. This is due to the limitation on the expressivity of the two-electron stream network, which, despite being translationally invariant, is not sufficient to represent the ground state of the unitary

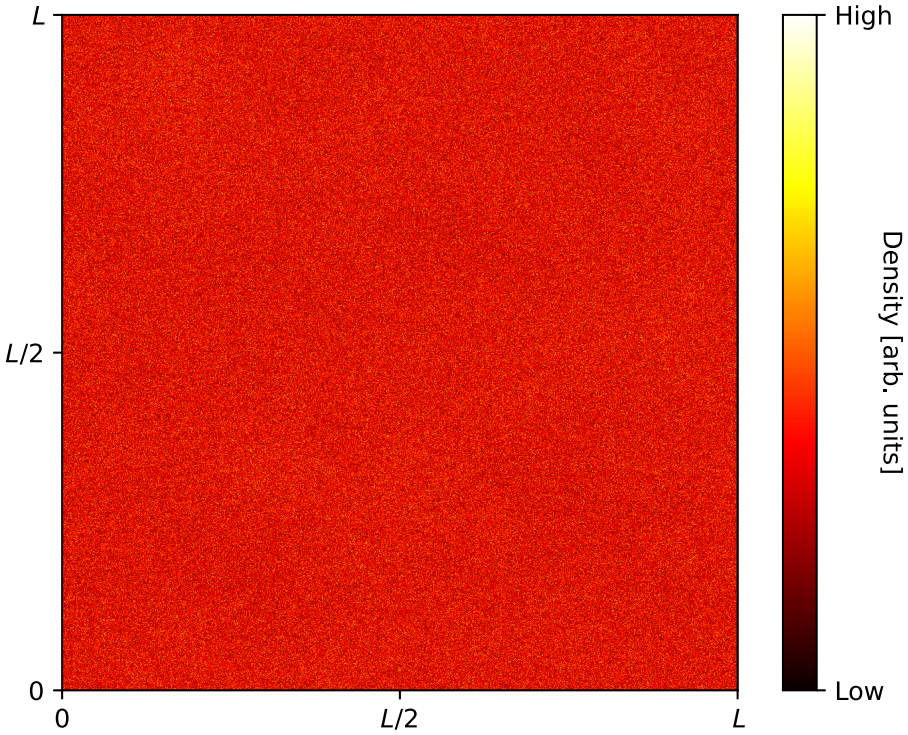

FIG. 7. A 2D projection of the $N = 14$ unitary Fermi gas density in the simulation cell from sampling the converged AGPs FermiNet wave function. L is the length of the simulation cell.

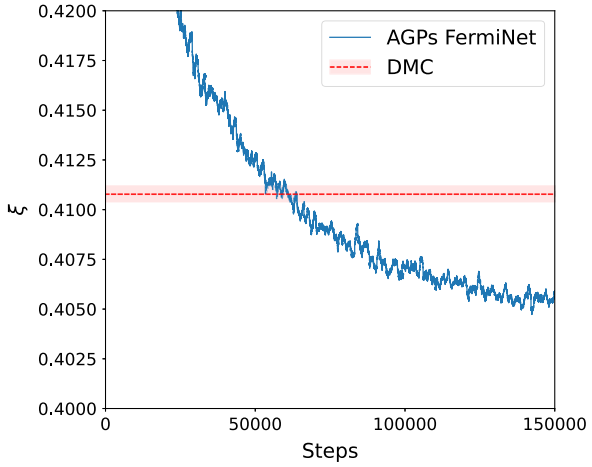

FIG. 8. Training curve of the 66-particle UFG with 32 determinants and using the geminals in Eq. (19). The red dashed line is the FNDMC result from Ref. [20] under the same density and interaction width.

Fermi gas. This finding is also consistent with the findings in the previous study of the homogeneous electron gas by Cassella *et al.* [38]. Hence, we include the one-electron stream networks in our *Ansatz* when studying the unitary Fermi gas.

Despite an explicit origin being embedded into the one-electron stream network in the AGPs FermiNet *Ansatz* when studying the unitary Fermi gas, we conclude that this does not create any obvious bias in the converged wave function. Evidence supporting this is shown in Fig. 7, where a 2D-projected density of the $N = 14$ unitary Fermi gas, obtained using the converged AGPs FermiNet wave function, is plotted. There is no obvious structure in the density plot, and it is uniform across the whole simulation cell. Hence, we believe the *Ansatz* is capable of representing translationally invariant wave functions, even though this is not explicitly imposed.

APPENDIX F: TRAINING CURVES OF THE 66-PARTICLE UFG

We present the training curves of the $N = 66$ UFG in this section. Because of limited computational resources, we train the 66-particle UFG system only up to 150 000 steps, which is half the number of steps we use to train all other systems in the paper (see Table II). Therefore, even though our $N = 66$ results are able to outperform the FNDMC results by a similar amount as the smaller systems, it is not fully converged as demonstrated in the training curve in Fig. 8. We emphasize that the energy we obtain is still a variational upper bound on the ground-state energy.

[1] A. J. Leggett, *Diatomic molecules and Cooper pairs*, in *Modern Trends in the Theory of Condensed Matter*, edited by A. Pekalski and J. A. Przystawa (Springer, Berlin, 1980), pp. 13–27.
[2] P. Nozières and S. Schmitt-Rink, *Bose condensation in an attractive fermion gas: From weak to strong coupling superconductivity*, J. Low Temp. Phys. **59**, 195 (1985).
[3] T. Papenbrock and G. F. Bertsch, *Pairing in low-density Fermi gases*, Phys. Rev. C **59**, 2052 (1999).
[4] S. Gandolfi, A. Gezerlis, and J. Carlson, *Neutron matter from low to high density*, Annu. Rev. Nucl. Part. Sci. **65**, 303 (2015).
[5] Y. Nishida and D. T. Son, *ϵ expansion for a Fermi gas at infinite scattering length*, Phys. Rev. Lett. **97**, 050403 (2006).
[6] P. Nikolić and S. Sachdev, *Renormalization-group fixed points, universal phase diagram, and 1/n expansion for quantum liquids with interactions near the unitarity limit*, Phys. Rev. A **75**, 033608 (2007).
[7] M. Randeria, J.-M. Duan, and L.-Y. Shieh, *Superconductivity in a two-dimensional Fermi gas: Evolution from Cooper pairing to Bose condensation*, Phys. Rev. B **41**, 327 (1990).
[8] M. Randeria, *Pre-pairing for condensation*, Nat. Phys. **6**, 561 (2010).
[9] G. C. Strinati, P. Pieri, G. Röpke, P. Schuck, and M. Urban, *The BCS-BEC crossover: From ultra-cold Fermi gases to nuclear systems*, Phys. Rep. **738**, 1 (2018).
[10] S. Giorgini, L. P. Pitaevskii, and S. Stringari, *Theory of ultracold atomic Fermi gases*, Rev. Mod. Phys. **80**, 1215 (2008).
[11] C. Chin, R. Grimm, P. Julienne, and E. Tiesinga, *Feshbach resonances in ultracold gases*, Rev. Mod. Phys. **82**, 1225 (2010).
[12] J. Bardeen, L. N. Cooper, and J. R. Schrieffer, *Microscopic theory of superconductivity*, Phys. Rev. **106**, 162 (1957).
[13] W. M. C. Foulkes, L. Mitas, R. J. Needs, and G. Rajagopal, *Quantum Monte Carlo simulations of solids*, Rev. Mod. Phys. **73**, 33 (2001).
[14] J. Carlson, S. Gandolfi, and A. Gezerlis, *Superfluid pairing in neutrons and cold atoms*, in *Fifty Years of Nuclear BCS* (World Scientific, Singapore, 2013), pp. 348–359.
[15] J. Carlson, S.-Y. Chang, V. R. Pandharipande, and K. E. Schmidt, *Superfluid Fermi gases with large scattering length*, Phys. Rev. Lett. **91**, 050401 (2003).
[16] S. Y. Chang, V. R. Pandharipande, J. Carlson, and K. E. Schmidt, *Quantum Monte Carlo studies of superfluid Fermi gases*, Phys. Rev. A **70**, 043602 (2004).
[17] A. J. Morris, P. López Ríos, and R. J. Needs, *Ultracold atoms at unitarity within quantum Monte Carlo methods*, Phys. Rev. A **81**, 033619 (2010).
[18] S. Y. Chang and V. R. Pandharipande, *Ground-state properties of Fermi gases in the strongly interacting regime*, Phys. Rev. Lett. **95**, 080402 (2005).
[19] A. Gezerlis and J. Carlson, *Low-density neutron matter*, Phys. Rev. C **81**, 025803 (2010).
[20] M. M. Forbes, S. Gandolfi, and A. Gezerlis, *Resonantly interacting fermions in a box*, Phys. Rev. Lett. **106**, 235303 (2011).
[21] R. He, N. Li, B.-N. Lu, and D. Lee, *Superfluid condensate fraction and pairing wave function of the unitary Fermi gas*, Phys. Rev. A **101**, 063615 (2020).

[22] Y.-H. Song, Y. Kim, N. Li, B.-N. Lu, R. He, and D. Lee, *Quantum many-body calculations using body-centered cubic lattices*, Phys. Rev. C **104**, 044304 (2021).

[23] O. Goulko and M. Wingate, *Numerical study of the unitary Fermi gas across the superfluid transition*, Phys. Rev. A **93**, 053604 (2016).

[24] S. Jensen, C. N. Gilbreth, and Y. Alhassid, *Contact in the unitary Fermi gas across the superfluid phase transition*, Phys. Rev. Lett. **125**, 043402 (2020).

[25] L. Pisani, P. Pieri, and G. C. Strinati, *Spatial emergence of off-diagonal long-range order throughout the BCS-BEC crossover*, Phys. Rev. B **105**, 054505 (2022).

[26] G. Carleo and M. Troyer, *Solving the quantum many-body problem with artificial neural networks*, Science **355**, 602 (2017).

[27] K. Choo, T. Neupert, and G. Carleo, *Two-dimensional frustrated $J_1 - J_2$ model studied with neural network quantum states*, Phys. Rev. B **100**, 125124 (2019).

[28] M. Hibat-Allah, M. Ganahl, L. E. Hayward, R. G. Melko, and J. Carrasquilla, *Recurrent neural network wave functions*, Phys. Rev. Res. **2**, 023358 (2020).

[29] D. Luo and B. K. Clark, *Backflow transformations via neural networks for quantum many-body wave functions*, Phys. Rev. Lett. **122**, 226401 (2019).

[30] 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).

[31] J. Hermann, Z. Schätzle, and F. Noé, *Deep-neural-network solution of the electronic Schrödinger equation*, Nat. Chem. **12**, 891 (2020).

[32] L. Gerard, M. Scherbela, P. Marquetand, and P. Grohs, *Gold-standard solutions to the Schrödinger equation using deep learning: How much physics do we need?*, in *Advances in Neural Information Processing Systems*, Vol. 35, edited by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Curran Associates, Inc., Red Hook, NY, 2022), pp. 10282–10294.

[33] G. Pescia, J. Han, A. Lovato, J. Lu, and G. Carleo, *Neural-network quantum states for periodic systems in continuous space*, Phys. Rev. Res. **4**, 023138 (2022).

[34] X. Li, C. Fan, W. Ren, and J. Chen, *Fermionic neural network with effective core potential*, Phys. Rev. Res. **4**, 013021 (2022).

[35] M. Wilson, N. Gao, F. Wudarski, E. Rieffel, and N. M. Tubman, *Simulations of state-of-the-art fermionic neural network wave functions with diffusion Monte Carlo*, arXiv:2103.12570.

[36] J. S. Spencer, D. Pfau, A. Botev, and W. M. C. Foulkes, *Better, faster fermionic neural networks*, arXiv:2011.07125.

[37] X. Li, Z. Li, and J. Chen, *Ab initio calculation of real solids via neural network ansatz*, Nat. Commun. **13**, 7895 (2022).

[38] G. Cassella, H. Sutterud, S. Azadi, N. D. Drummond, D. Pfau, J. S. Spencer, and W. M. C. Foulkes, *Discovering quantum phase transitions with fermionic neural networks*, Phys. Rev. Lett. **130**, 036401 (2023).

[39] J. Bouchaud, A. Georges, and C. Lhuillier, *Pair wave functions for strongly correlated fermions and their determinantal representation*, J. Phys. (Paris) **49**, 553 (1988).

[40] M. Casula and S. Sorella, *Geminal wavefunctions with Jastrow correlation: A first application to atoms*, J. Chem. Phys. **119**, 6500 (2003).

[41] M. Bajdich, *Generalized pairing wave functions and nodal properties for electronic structure quantum Monte Carlo*, arXiv:0712.3066.

[42] C. Genovese, T. Shirakawa, K. Nakano, and S. Sorella, *General correlated geminal ansatz for electronic structure calculations: Exploiting Pfaffians in place of determinants*, J. Chem. Theory Comput. **16**, 6114 (2020).

[43] J. Toulouse and C. J. Umrigar, *Optimization of quantum Monte Carlo wave functions by energy minimization*, J. Chem. Phys. **126**, 084102 (2007).

[44] K. Hornik, M. Stinchcombe, and H. White, *Multilayer feedforward networks are universal approximators*, Neural Netw. **2**, 359 (1989).

[45] Although n_l can be different in different layers, in practice, we use an equal number of hidden units in each layer except for the input layer, where n_0 is fixed by the number of particles in the system.

[46] J. Martens and R. Grosse, *Optimizing neural networks with Kronecker-factored approximate curvature*, in *Proceedings of the 32nd International Conference on Machine Learning*, Vol. 37, edited by F. Bach and D. Blei (PMLR, Lille, France, 2015), pp. 2408–2417.

[47] J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang, *JAX: Composable transformations of* Python + NumPy *programs*, https://github.com/google/jax.

[48] M. Wilson, S. Moroni, M. Holzmann, N. Gao, F. Wudarski, T. Vegge, and A. Bhowmik, *Neural network ansatz for periodic wave functions and the homogeneous electron gas*, Phys. Rev. B **107**, 235139 (2023).

[49] Although, in principle, the particle coordinates $\mathbf{r}_i^\alpha$ are not needed in a periodic system, we have observed that including them often improves convergence and allows the variational *Ansatz* to achieve a lower energy. Hence, we include them in all of our calculations.

[50] M. Casula, C. Attaccalite, and S. Sorella, *Correlated geminal wave function for molecules: An efficient resonating valence bond approach*, J. Chem. Phys. **121**, 7110 (2004).

[51] M. Marchi, S. Azadi, M. Casula, and S. Sorella, *Resonating valence bond wave function with molecular orbitals: Application to first-row molecules*, J. Chem. Phys. **131**, 154116 (2009).

[52] M. Bajdich, L. Mitas, G. Drobný, L. K. Wagner, and K. E. Schmidt, *Pfaffian pairing wave functions in electronic-structure quantum Monte Carlo simulations*, Phys. Rev. Lett. **96**, 130201 (2006).

[53] H. Xie, Z.-H. Li, H. Wang, L. Zhang, and L. Wang, *Deep variational free energy approach to dense hydrogen*, Phys. Rev. Lett. **131**, 126501 (2023).

[54] Note that the idea of an AGPs or BCS FermiNet is very similar to the previous work by Luo and Clark [29], where they used the neural-network backflow wave function implemented on top of a Bogoliubov–de Gennes or BCS wave function to study lattice systems.

[55] To get a rough estimate of the number of terms in the sum, we can take the number of hidden units in the one-electron stream to be $M = n_L \sim 200$ and the number of pairs to be $p \sim 20$; this number is approximately 10^{27}.

[56] Although the range of the potential is finite, which introduces a second length scale even at unitarity, most of the results in the literature have been obtained with the same value of μ. This allows us to compare our results with theirs. To obtain exact results, one has to study the limit as $\mu \to \infty$, but converging the simulations becomes more difficult as μ increases.

[57] As previous works [17,21,58,59] have shown, the ground-state energy of the $N = 66$ particle UFG is the closest to the ground-state energy in the thermodynamic limit even among larger systems. This can also be seen by comparing the ratio of E_{FG} and the energy per particle from summing over the energies occupied by the particles in the discretized k space, where this ratio is just 0.5% away from unity for $N = 66$.

[58] J. Carlson, S. Gandolfi, K. E. Schmidt, and S. Zhang, *Auxiliary-field quantum Monte Carlo method for strongly paired fermions*, Phys. Rev. A **84**, 061602(R) (2011).

[59] M. M. Forbes, S. Gandolfi, and A. Gezerlis, *Effective-range dependence of resonantly interacting fermions*, Phys. Rev. A **86**, 053603 (2012).

[60] All calculations presented in this section are performed using the Slater FermiNet with *block-diagonal determinants*. A small set of calculations are performed with the Slater FermiNet with *dense determinants*, which are presented in Appendix D. In short, we do found slight improvements over the block-diagonal Slater FermiNet results, but it is still qualitatively wrong and is also much worse than the AGPs FermiNet results.

[61] G. Palkanoglou, F. K. Diakonos, and A. Gezerlis, *From odd-even staggering to the pairing gap in neutron matter*, Phys. Rev. C **102**, 064324 (2020).

[62] Limited by the scope of this work, we do not proceed to carry out a similar set of calculations done by Forbes, Gandolfi, and Gezerlis [20] for extrapolating the results to $\mu \to \infty$.

[63] L. P. Gorkov and T. K. Melik-Barkhudarov, Sov. Phys. JETP **13**, 1018 (1961)https://www.mit.edu/~levitov/jetp/gorkov_melik-barkhudarov1961.pdf.

[64] Although writing the pairing gap in units of the Fermi energy would be more suitable as it is a one-body quantity, we choose to express it in units of E_{FG} instead, because (i) we want to be consistent with the units used in the paper by Carlson *et al.* [15], which also used E_{FG} as the energy unit for the pairing gap, and (ii) to keep the units consistent with other quantities in this work.

[65] 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).

[66] R. J. Needs, M. D. Towler, N. D. Drummond, P. López Ríos, and J. R. Trail, *Variational and diffusion quantum Monte Carlo calculations with the CASINO code*, J. Chem. Phys. **152**, 154106 (2020).

[67] G. E. Astrakharchik, J. Boronat, J. Casulleras, and S. Giorgini, *Momentum distribution and condensate fraction of a fermion gas in the BCS-BEC crossover*, Phys. Rev. Lett. **95**, 230405 (2005).

[68] X. Li, J. c. v. Kolorenč, and L. Mitas, *Atomic Fermi gas in the unitary limit by quantum Monte Carlo methods: Effects of the interaction range*, Phys. Rev. A **84**, 023615 (2011).

[69] M. W. Zwierlein, C. H. Schunck, C. A. Stan, S. M. F. Raupach, and W. Ketterle, *Formation dynamics of a fermion pair condensate*, Phys. Rev. Lett. **94**, 180401 (2005).

[70] W. J. Kwon, G. D. Pace, R. Panza, M. Inguscio, W. Zwerger, M. Zaccanti, F. Scazza, and G. Roati, *Strongly correlated superfluid order parameters from dc Josephson supercurrents*, Science **369**, 84 (2020).

[71] Because of limited resources, we compute only five data points at far separations instead of the whole range of separations from $r = 0$ to $r = L/2$ for $N = 66$.

[72] Recently, a closely related preprint on neural networks and the Fermi gas appeared [73], proposing a similar extension to continuous-space neural-network *Ansätze*.

[73] J. Kim, G. Pescia, B. Fore, J. Nys, G. Carleo, S. Gandolfi, M. Hjorth-Jensen, and A. Lovato, *Neural-network quantum states for ultra-cold Fermi gases*, arXiv:2305.08831.

[74] I. von Glehn, J. S. Spencer, and D. Pfau, *A self-attention ansatz for ab-initio quantum chemistry*, arXiv:2211.13672.

[75] N. Gao and S. Günnemann, *Generalizing neural wave functions*, in *Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA*, Proceedings of Machine Learning Research Vol. 202, edited by A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett (PMLR, 2023), pp. 10708–10726.

[76] M. Scherbela, R. Reisenhofer, L. Gerard, P. Marquetand, and P. Grohs, *Solving the electronic Schrödinger equation for multiple nuclear geometries with weight-sharing deep neural networks*, Nat. Comput. Sci. **2**, 331 (2022).

[77] J. S. Spencer and D. Pfau, contributors, FermiNet, http://github.com/deepmind/ferminet.

[78] A. Botev and J. Martens, *KFAC-JAX*, http://github.com/deepmind/kfac-jax.

[79] X. Glorot and Y. Bengio, *Understanding the difficulty of training deep feedforward neural networks*, in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research Vol. 9, edited by Y. W. Teh and M. Titterington (PMLR, Chia Laguna Resort, Sardinia, Italy, 2010), pp. 249–256.

[80] N. Gao and S. Gunnemann, *Sampling-free inference for ab-initio potential energy surface networks*, arXiv:2205.14962.

[81] W. Ren, W. Fu, X. Wu, and J. Chen, *Towards the ground state of molecules via diffusion Monte Carlo on neural networks*, Nat. Commun. **14**, 1860 (2023).

[82] J. Lin, G. Goldshlager, and L. Lin, *Explicitly antisymmetrized neural network layers for variational Monte Carlo simulation*, J. Comput. Phys. **474**, 111765 (2023).