Sorting Out Quantum Monte Carlo

Jack Richter-Powell [1 2] **Luca Thiede** [2 3] **Alán Aspuru-Guzik** [2 3] **David Duvenaud** [2 3]

Abstract

Molecular modeling at the quantum level requires choosing a parameterization of the wavefunction that both respects the required particle symmetries, and is scalable to systems of many particles. For the simulation of fermions, valid parameterizations must be antisymmetric with respect to the exchange of particles. Typically, antisymmetry is enforced by leveraging the anti-symmetry of determinants with respect to the exchange of matrix rows, but this involves computing a full determinant each time the wavefunction is evaluated. Instead, we introduce a new antisymmetrization layer derived from sorting, the *sortlet*, which scales as $O(N \log N)$ with regards to the number of particles – in contrast to $O(N^3)$ for the determinant. We show numerically that applying this anti-symmeterization layer on top of an attention based neural-network backbone yields a flexible wavefunction parameterization capable of reaching chemical accuracy when approximating the ground state of first-row atoms and small molecules.

1. Introduction

1.1. Overview

Quantum Monte Carlo (QMC) methods are a class of algorithm that aim to model the wavefunction for a system of quantum particles, typically an atom or molecule. In its simplest form, Variational Quantum Monte Carlo (VQMC) is a QMC method that aims to estimate the lowest-energy state of a system via the variational principle. This is done by minimizing the Rayleigh quotient of the system's Hamiltonian over a parametric family of wavefunctions – commonly referred to as the *wavefunction ansatz*. Variational principles from functional analysis yield that under mild assumptions, the minimizer approaches the true ground state wavefunction as the parametric family increases in expressiveness.

The first method resembling modern VQMC was proposed by London and Heilter in the late 1920s [9], where they attempted to calculate the ground state of the diatomic hydrogen molecule. Fermi and then later Kalos [21] converted the problem into a Monte Carlo Sampling one. The advent of the practical Slater determinant ansatz in the 1950s and the growth in available computational power since has allowed QMC algorithms to become one of the benchmark frameworks for deriving the properties of chemical systems *in silico*. Often, QMC is used to benchmark other non-sampling methods, such as Coupled Cluster methods [2] and Density Functional Theory (DFT).

Advances in automatic differentiation complemented by empirical ML experience have recently produced new types of wavefunction ansatz, assembled around deep neural network backbones. Neural networks are particularly attractive in this setting, due to their favourable scalability as dimension increases. Electron configuration space, $\mathbb{R}^{3N}$, grows exponentially in the number of electrons N, rendering this ability critical. Hybrid methods, seeking to combine Slater determinants with neural orbitals or Jastrow factors, have recently shown promise on systems comprised of a large number of particles – even when relatively few Slater determinants are employed [34]. Their performance is especially notable when contrasted against existing Hartree-Fock implementations with an equivalent number of determinants.

Despite these recent successes, there is a good reason to look beyond Slater determinants when designing ansatz – determinants are relatively computationally expensive. Evaluating a determinant-based ansatz scales $O(N^3)$ in the number of electrons, devolving to $O(N^4)$ when evaluating the local energy (due to the Laplacian). Estimating the energy is required at every step of nearly every QMC algorithm (variational or otherwise), so this quickly becomes a bottleneck if we look to scale QMC beyond small molecules to problems of practical interest, where N could be on the order of thousands or even tens of thousands of electrons.

[1]Massachusetts Institute of Technology, Cambridge MA, USA [2]Vector Institute, Toronto, Canada [3]University of Toronto, Toronto, Canada. Correspondence to: Jack Richter-Powell <jricpow@mit.edu>.

AI4Science Workshop at the 41st International Conference on Machine Learning, Vienna, Austria, 2024.

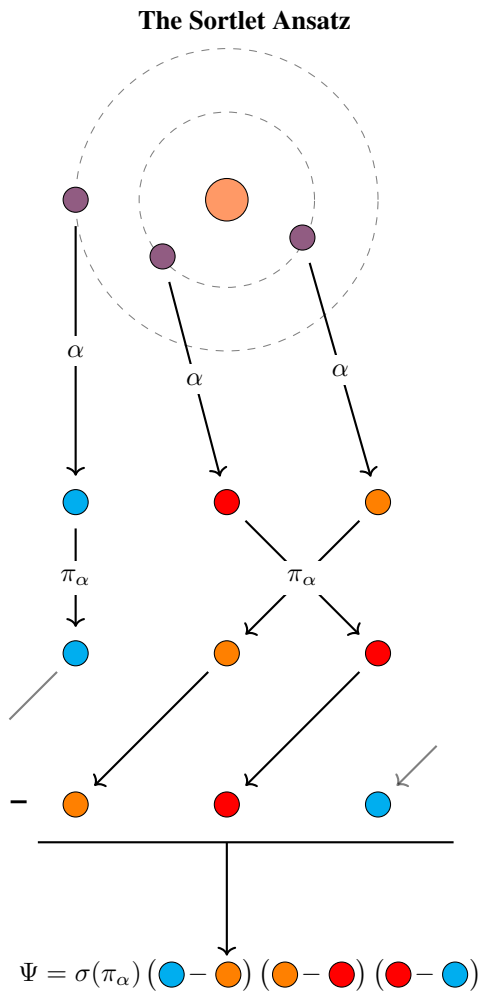


Figure 1: Geometric illustration of the Sortlet ansatz construction given in Equation (4). Here π_α is the permutation that sorts the output of $\alpha(r)$.

1.2. Our Contribution

Seeking to reduce this complexity, we introduce a novel antisymmetrization operation, the *sortlet*, and apply it within the VQMC framework to model ground states of various atoms and molecules. Crucially, our operation enjoys an improved $O(N \log N)$ complexity, which comes from replacing the determinant with a cheaper alternative: sorting. Sorting is known to be universal for systems of 1d electrons [40; 20], and has been alluded to previously as a method for designing ansatz (see Appendix B of [34] and [24]). Our contribution is twofold – we show both that a natural extension of sorting for electrons in 3d space exists, as well as that this ansatz can achieve chemical accuracy, at least on small systems, when paired with a sufficiently flexible functional backbone such as a modern attention based neural network.

Prior work [31; 1] explored the usage of Vandermonde determinants which scale $O(N^2)$, but both were unable to achieve the high degree of accuracy required for applications in quantum chemistry. In Section 5.2 we present a topological obstruction to a single continuous Vandermonde determinant learning the exact ground state, based on the known geometry of the wavefunction's nodal surface.

The construction of a universal, sub-cubic-time continuous representation of ground-state wavefunctions remains an open problem. However, we show that, like the vandermonde determinant, our sortlet can represent the ground state exactly if we allow discontinuous parameterizations. We also highlight a key benefit of the lower scaling complexity of the sortlet – allowing more terms in the wavefunction ansatz with the same (asymptotic) computational budget. Classical QMC techniques have relied on larger expansions to help mitigate topological inconsistencies between the parametric ansatz and the true ground state [3; 5]. In the context of VQMC, this might partially explain why our method is able to achieve higher accuracy than previous Vandermonde constructions [31; 1].

Numerically, we demonstrate that combining our sortlet antisymmetrization layer with the PsiFormer [41] attention backbone is sufficient to achieve chemical accuracy on a handful of atoms and small molecules, as well as reproduce the potential energy surface of H_4 rectangles as one bond length is scanned. While wavefunction ansatz built from alternatives to determinants have existed in the QMC literature for some time, to the best of our knowledge, this is the first work to demonstrate chemical accuracy with something other than a determinant – even if only on small molecular systems at the current stage. Flexibility of the attention neural network backbone of [41] offers another partial explanation for why our ansatz has proven more accurate than other full determinant alternatives in the past. That said, at the current state our results are far from competitive with those of neural network ansatz with full determinants [34; 38; 41]. In fairness to our approach, the results presented in that body of work are the outcome of a long series of incremental improvements – far beyond the scope of the initial proof-of-concept described in this paper. Our accuracy on small systems, we believe, is a evidence that the sortlet ansatz is a promising direction for further study, and that with similar investments in software engineering, our method could become competitive on more difficult benchmarks.

1.3. Where did the determinants come from anyway?

Determinants appear in nearly all QMC ansatz as a simple mathematical way to satisfy the generalized Pauli Exclusion Principle: any valid wavefunction must be antisymmetric under exchange (transposition) of any two electrons with the same spin. Intuitively, this follows from the idea that

quantum particles such as electrons are indistinguishable, so changing their order (which has no physical meaning) cannot change the state of the system. Explaining the appearance of the -1 factor is more involved – representing the fundamental difference between Fermions and Bosons (see [17]) – but for our purposes, we just assume any valid wavefunction for Fermions needs to respect the antisymmetry equation:

$$\Psi(r_1^\uparrow, \cdots, r_M^\uparrow, r_1^\downarrow, \cdots, r_i^\downarrow, \cdots, r_j^\downarrow, \cdots) = -\Psi(r_1^\uparrow, \cdots, r_M^\uparrow, r_1^\downarrow, \cdots, r_j^\downarrow, \cdots, r_i^\downarrow, \cdots) \tag{AS}$$

Alternation of the determinant under the exchange of any two rows or columns has been conventionally employed to satisfy this property. Originally, the Slater determinants typically consisted of parameterizing a matrix function $\Phi : \mathbb{R}^{3N} \longrightarrow \mathbb{R}^{N\times N}$ via a collection of N single electron orbitals $\Phi_{ij} = \phi_i(r_j)$, each of which was only a function of a single r_j. Upon interchanging the electron positions, the resulting matrix Φ'_{ij} is exactly Φ_{ij} with two rows swapped, flipping the sign of $\Psi := \det[\Phi_{ij}]$.

Though this approach is undeniably tidy, it suffers limited expressive power due to its inability to accurately model electronic correlation, which involves the interactions of all N-particles at a time rather than treating particles as mean-field orbitals. Common remedies include either the addition of parametric Jastrow factors which are multiplicative against the determinant, i.e $\Psi = e^{J(r)} \det[\Phi_{ij}(r)]$, or backflow transformations [10], which makes each orbital $\phi_i(q_j)$ dependent on all electron positions through so called pseudo-coordinates $q_j = r_j + \varepsilon^i r_i$. Modern neural network approaches such as FermiNet, PsiFormer or PauliNet can be seen as a more flexible generalization of backflow, since they opt to parameterize the matrix Φ_{ij} as the output of a deep neural network, but in a way that the Φ_{ij} depend symmetrically on all electron positions except for r_j, which is allowed non-symmetric influence. Filling out Φ_{ij} this way preserves the antisymmetry of the Slater determinant, and provides sufficient flexibility to capture electronic correlation. Large neural networks have allowed this approach to scale to systems of around 40 electrons with state-of-the-art results [41].

2. Variational Quantum Monte Carlo

The following is a brief summary of the core aspects VQMC, but it is by no means exhaustive. Those experienced with the framework will likely find it redundant, and those completely new will probably find it incomplete. For the latter group, we recommend either [4] or [25], which provide much more exhaustive treatments of the relevant background.

2.1. The Born-Oppenheimer Hamiltonian

While our wavefunction ansatz is general enough to describe any system of Fermions, in this paper we will focus on quantum chemical systems comprised of electrons. Similar to [34; 38] we will work in the Born-Oppenheimer approximation, where the molecular nuclei are treated as fixed, and the Hamiltonian governing the forces acting on the electrons is given by

$$\hat{H} = -\frac{\nabla^2}{2} + \left[\sum_{i>j} \frac{1}{|r_i - r_j|} - \sum_{i,I} \frac{Z_I}{|r_i - R_I|} + \sum_{I>J} \frac{Z_I Z_J}{|R_I - R_J|} \right]$$

$$\text{where} \quad \boxed{\begin{array}{c} r_i\text{: electron positions} \\ R_i\text{: nuclei positions} \\ Z_I\text{: nuclear charges} \end{array}} \tag{BO}$$

The terms inside the square brackets are often written simply as $V(r)$, since they represent the potential energy of the electrons. Analogously, $-\nabla^2$ is the kinetic energy operator. For a given system $\hat{H}$, the ground state Ψ_g is nothing more than the eigenfunction corresponding to the smallest eigenvalue E_1 of $\hat{H}$.

2.2. The Variational Principle

The *variational* aspect of VQMC comes from the following proposition:

Proposition 1. *Variational Principle*
Let Ψ_1 be minimum eigenvector of the Hermitian operator $\hat{H}$, that is, the eigenfunction associated to the smallest eigenvalue E_1. Then

$$\Psi_1 = \underset{\Psi}{\operatorname{argmin}} \frac{\left\langle \Psi | \hat{H} | \Psi \right\rangle}{\langle \Psi | \Psi \rangle} \tag{VAR}$$

The term inside argmin is often called the *Rayleigh quotient* and so denoted by $R(\Psi)$. Proving this proposition amounts to hardly more than a little bookkeeping, see Appendix A.1

2.3. The Local Energy and its Vanishing Variance

The variational energy $E = R(\Psi_\theta)$, for any parameteric configuration of θ can in principle be computed from the inner product in (VAR). Since we are limited to approximating the energy with a finite average in practice, this is suboptimal – the variance will be prohibitive, and we will need a tremendous number of samples to achieve a reasonable estimate. A much better approach is to instead define the local energy (at any point $x \in \mathbb{R}^{3N}$)

$$E_{\text{loc}}(\Psi_\theta)(x) = \frac{[\hat{H}\Psi_\theta](x)}{\Psi_\theta(x)} \qquad \mathbb{E}_{x\sim\Psi_\theta^2}[E_{\text{loc}}(x)] = E \tag{1}$$

where $x \sim \Psi_\theta^2$ is taken to mean that the x are drawn from the normalized distribution, $\frac{\Psi_\theta^2}{\langle \Psi_\theta | \Psi_\theta \rangle}$. Reformulating the en-

ergy approximation in terms of E_{loc} instead of the Rayleigh quotient has two immediate advantages : (1) we lower the variance of our estimator drastically by using importance sampling (drawing $x \sim \Psi_\theta^2$) and (2) we don't need to estimate the normalization constant of Ψ_θ^2, since the local energy is invariant to scalar multiplication of Ψ_θ. VQMC also benefits from a third seemingly magical property: as we update θ to minimize E, the variance *vanishes*. To see why, consider the following:

Proposition 1. *Let Ψ_1 be the eigenfunction associated to the ground state energy E_1 (as before), then*

$$E_{loc}(\Psi_1)(x) = E_1 \qquad \forall x \in \mathbb{R}^{3N} \tag{2}$$

i.e the local energy is constant at optimality.

Proof. Equation (2) is a direct consequence of the variational principle; Ψ_1 is an eigenvector of $\hat{H}$, thus $\hat{H}\Psi_1 = E_1\Psi_1$, which yields $E_{\text{loc}}(\Psi_1)(x) = \frac{\hat{H}\Psi_1(x)}{\Psi_1(x)} = \frac{E_1\Psi_1(x)}{\Psi_1(x)} = E_1$ □

2.4. Derivatives of the Energy Functional

Performing any sort of first order optimization over the parameters of our ansatz Ψ_θ to minimize its energy, denoted $E(\theta)$, will also require its gradients. Following the preceding section, we will estimate $E(\theta)$ via the following Monte-Carlo estimate

$$E(\theta) \approx \bar{E}(\theta) = \sum_{i=1}^{N} E_{\text{loc}}(\Psi_\theta)(x_i) \qquad x_i \sim \Psi_\theta^2 \tag{3}$$

Any MCMC method (such as Random-Walk Metropolis Hastings for example) can be used to draw the x_i, which we will detail further in the next section. Approximating the energy via (3), while superior for the reasons laid out above, comes at the price of making the computation of its gradients more involved than naively applying automatic differentiation to each term in the finite sum. Differentiating $\frac{\partial E}{\partial \theta}$ correctly requires also accounting for the influence of the parameters on the distribution Ψ_θ^2 which the x_i are drawn from. Like in many other other works [34; 41; 4; 1; 38; 13], we use the gradient estimator

$$\nabla E(\theta) = \frac{2}{n}\sum_{i=1}^{n}\Bigg[E_{\text{loc}}(x_i)\nabla \log \Psi_\theta(x_i) - \Big(\frac{1}{n}\sum_{j=1}^{n}[E_{\text{loc}}(x_j)]\Big)\nabla \log \Psi_\theta(x_i)\Bigg]$$

where $x_i \sim \Psi_\theta^2$ are drawn using MHMC, but we offer a full derivation in Appendix A.2. Notably, this can be seen as an instance of the REINFORCE estimator widely used in reinforcement learning [16].

3. The Sortlet Ansatz

3.1. The Sortlet Ansatz

We are ready to define what we coin the *Sortlet Ansatz*. First, let's define what we call a *sortlet*, denoted by Ψ_α:

$$\Psi_\alpha(r) = \sigma(\pi_\alpha)\prod_{i=1}^{N}(\alpha_{i+1}(r) - \alpha_i(r)) \tag{4}$$

where

- $\alpha : \mathbb{R}^{N\times 3} \longrightarrow \mathbb{R}^N$, permutation equivariant, and reindexed such that $\alpha_i < \alpha_{i+1} : 1 \leq i \leq N-1$, and $\alpha_{N+1} = \alpha_1$. **This is the neural network backbone.**
- π_α is the permutation on N letters which re-indexes (sorts) the tuple $\alpha(r)$
- $\sigma(\pi_\alpha)$ is the parity, equivalently number of transpositions in π_α, mod 2.

Analogously to other VQMC ansatz, we will employ an expansion of sortlets to serve as our sortlet ansatz in VQMC. Specifically, our Ψ is expanded as

$$\Psi = \exp[-J_\beta]\sum_{i=1}^{K}\left[\Psi_{\alpha_i}\sum_{I} e^{-\gamma_I \sum_j |r_j - R_I|}\right] \tag{5}$$

where J_β is the same as in [38; 41]

As we will show in Section 3.2, with discontinuous α we can exactly represent the ground state wavefunction with a single sortlet.

Complexity: The sortlet for a given α can be computed in time $O(N \log N)$. Taking $K = O(N)$ and with the assumption $\alpha(r)$ can be evaluated in time $O(N^2)$, we have that Ψ can be evaluated in time $O(N^2 \log N)$, significantly faster than a determinant.

3.2. Properties of the Sortlet Ansatz

Proposition 2. *The Sortlet Ansatz satisfies the generalized Pauli exclusion principle – it is anti-symmetric to the exchange of electrons with the same spin.*

Proposition 3. *Each Ψ_α is continuously once-differentiable on $\mathbb{R}^{3N}$, meaning $\frac{\partial \Psi_\alpha}{\partial r^i}$ is continuous for all i.*

We defer the proof to Appendix A.4 and Appendix A.3 respectively.

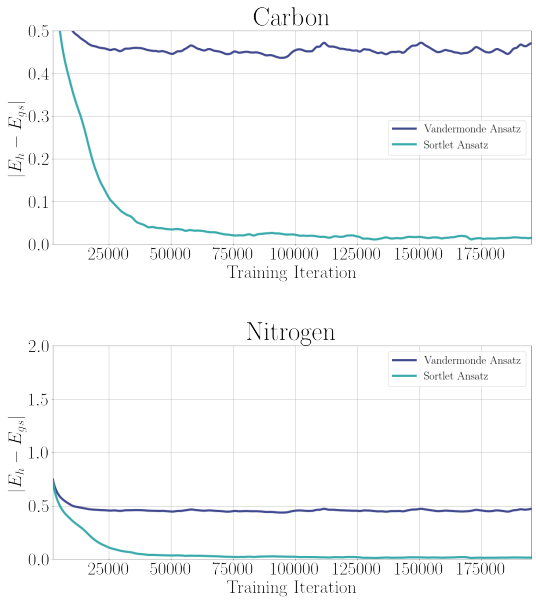


Figure 2: Comparison of Vandermonde and Sortlet Ansatz

4. Numerical Experiments

4.1. Evaluation on first row atoms and small molecules

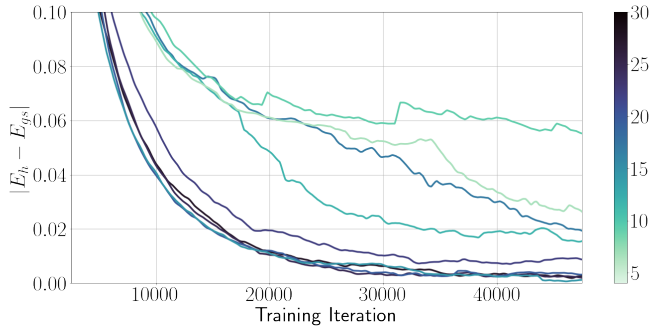
Section 6 collects the results of running QMC with our Sortlet ansatz on the ground state for a small collection of first row atoms and small molecules. We see that up to and including Boron, we are able to recover the ground state experimental energy to chemical accuracy. For Carbon and beyond, while we no longer achieve chemical accuracy, our method still outperforms the previous $O(N^2)$ method of [31].

4.2. Performance as a function of sortlets on Boron

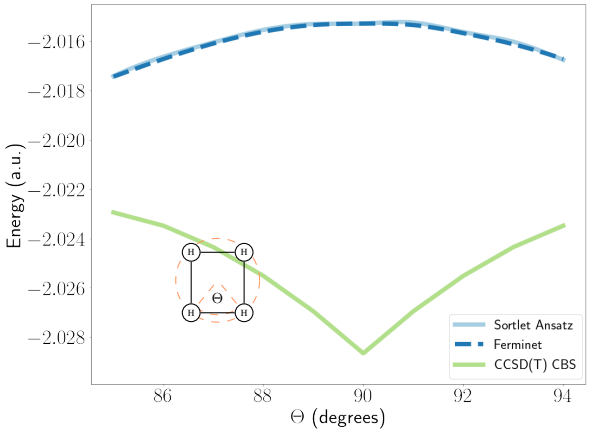
In Figure 3a, we show an ablation study comparing performance in terms of absolute difference to the empirical ground state values after $50,000$ optimization steps, against the number of sortlets on Boron. While all except one of the runs terminate at less than 20 mHa to the ground state, we see that those with $K > 16$ reach a slightly lower error on average, but achieve that value significantly faster than those with $K < 16$. Size of the network is kept constant for all runs (with the exception of the linear output layer), the only variable explored in this study is the number of terms K.

4.3. Reproduction of H_4 potential energy curve

In Figure 3b we reproduced the H_4 experiment from [34], where the geometry of the H_4 rectangle was linearly adjusted by varying the bond length of the bottom two hydrogen atoms. In agreement with FermiNet, we were able to reproduce the curve showing a maximum at $\Theta = 90\,\text{deg}$, not a minimum like the benchmark Coupled Cluster calculations.



(a) Comparing the number of Sortlets K (color) against the error to the empirical ground state energy.



(b) Reproduction of H_4 potential energy surface from the FermiNet paper.

Figure 3: Boron Sortlet Ablation and H_4 rectangle.

4.4. Direct Comparison to Vandermonde Ansatz

In Section 3.2 we directly compare training curves of vandermonde ansatz vs those of the Sortlet ansatz *with the same number of terms in their expansion*. We see that even with effectively more computational resources (since each term is $O(N^2)$), the Vandermonde ansatz is still unable to outperform the Sortlet.

5. Comparison with other Pairwise Constructions and Universality

In this section, we show a weak form of universality of the sortlet and that of the Vandermonde construction put forward in [1; 31]. We show that a slight modification of Theorem 2 in [31] also yields universality for the sortlet, but similar to the argument in the Appendix of [34], requires discontinuous functions. Theoretical universality so remains an open question, but we finish by discussing the implications of the nodal structure on experimental results.

5.1. Non-smooth universality

A nearly identical construction to Theorem 2 in [31] recovers a similar type of universality for the sortlet:

Proposition 4. *For a given ground state wavefunction Ψ_g, if we allow α to be discontinuous, setting*

$$\alpha_j = \pi^*(r,j)\left(\frac{\Psi^g\left[\pi^*(r)r\right]}{(N-1)}\right)^{1/N} \tag{6}$$

yields that $\Psi_\alpha = \Psi_g$. Here $\pi^*(r)$ is the permutation on N letters that maps $r \in \mathbb{R}^{3N}$ back to the fundamental domain (see [7]), $\pi^*(r,j) \in [1,N]$ is the index the jth electron maps to.

Proof. The proof is immediate from the definition of Ψ_α in Equation (4)

$$\Psi_\alpha = \sigma(\pi^*)\prod_{i=1}^{N}[\alpha_{i+1}-\alpha_i] \tag{7}$$

$$= \sigma(\pi^*)\left(\prod_{i=1}^{N-1}\left[\frac{|\Psi_g|}{(N-1)}\right]^{1/n}\right)(N-1)\left[\frac{|\Psi_g|}{(N-1)}\right]^{1/n} \tag{8}$$

$$= \sigma(\pi^*)|\Psi_g| \tag{9}$$

□

5.2. Comparison with the Vandermonde Ansatz

In [18; 1] and later [31], a similar type of pairwise antisymmetric ansatz was proposed, which takes the form

$$\psi'_{\text{pair}}(r) = \prod_{i<j}\left[\phi_B(r_i,\{r\}) - \phi_B(r_j,\{r\})\right] \tag{10}$$

where the second argument to ϕ_B is denoted as such to indicate the function is invariant to permutations of the electron positions, except for r_i. This expression is also equivalent to a Vandermonde matrix determinant (see Equation 1 [31]), which is why we refer to it simply as the Vandermonde Ansatz. Pang et al claim to have proved this form is universal for ground states, but unfortunately a simple argument shows this is not quite accurate. Ground states for small atoms (Li, Be) have been proven to possess exactly two nodal domains [26] – open sets $\mathcal{N} \subseteq \mathbb{R}^{3N}$ where the wavefunction is non-zero. So to represent them exactly, the ansatz must also have two nodal domains. Below we prove that for ψ'_{pair} this is not the case – regardless of the parameterization of ϕ_B, ψ'_{pair} always has at least 4 nodal domains.

Proposition 5. *ψ'_{pair}, as defined in (10) has $N^{\uparrow}! + N^{\downarrow}!$, nodal domains. The ground state is proven to have only 2 for Li, Be (see [5]) so this implies one Vandermonde determinant is insufficient to accurately represent the ground state.*

For the proof and discussion see Appendix A.5

5.3. Advantages of the Sortlet Ansatz against the Vandermonde Ansatz

To be clear, a single sortlet suffers from the same issue of too many nodal domains proven for the Vandermonde determinant in Proposition 5. Nonetheless, an advantage of the Sortlet is that, due to the $O(N\log N)$ scaling of each term, we can use an expansion with a linear number of sortlets, $K = O(N)$, while retaining log-quadratic complexity $O(N^2\log N)$. To achieve the same scaling, a Vandermonde ansatz would be limited to a logarithmic number of terms. Increasing the number of terms is known to be crucial in classical QMC methods, where thousands or even millions of terms can be employed to refine the energy [3]. For Hartree-Fock wavefunctions, increasing the number of terms is known to reduce the number of nodal domains [27; 26; 5].

Learned neural QMC ansatz such as [34; 38] have mostly bucked this trend, opting to fix the number of terms to a relatively small constant ($K \approx 16$), but increase network size instead. While we are not able to match the flexibility of the determinant PsiFormer, which is likely universal with a constant number of determinants, as seen in Figure 3a, mitigating the nodal inaccuracy of our sortlet by increasing K does seem to increase the speed of convergence.

6. Related Work

Neural Networks as representations for wavefunctions were applied to discrete quantum systems in [6; 8; 28; 30; 24] and to continuous bosonic systems in [35]. However, the first models considering fermionic continuous-space systems were DeepWF [18], Ferminet [34], and Paulinet [19]. These approaches were furthered with the introduction of the Psiformer model [42] and works that pushed the framework forward by fine-tuning the techniques and devising new specialized architectures [13; 22; 23; 32] and training procedures [29]. Other directions tried to construct architectures that are easy to sample [40; 43] by leveraging

normalizing flows, and meta-learning approaches where the wavefunction is conditioned on the Hamiltonian to transfer to different systems [12; 36; 37; 11]. Sorting as a method of guaranteeing antisymmetry was recently proposed in [40], but limited to the setting where electron positions are restricted to one-dimension.

References

[1] A. Acevedo, M. Curry, S. H. Joshi, B. Leroux, and N. Malaya. Vandermonde Wave Function Ansatz for Improved Variational Monte Carlo. In *2020 IEEE/ACM Fourth Workshop on Deep Learning on Supercomputers (DLS)*, pages 40–47, Atlanta, GA, USA, Nov. 2020. IEEE. ISBN 978-1-66542-245-1. doi:10.1109/DLS51937.2020.00010. URL https://ieeexplore.ieee.org/document/9297114/.

[2] Y. S. Al-Hamdani, P. R. Nagy, A. Zen, D. Barton, M. Kállay, J. G. Brandenburg, and A. Tkatchenko. Interactions between large molecules pose a puzzle for reference quantum mechanical methods. *Nature Communications*, 24 June 2021. ISSN 2041-1723. doi:10.1038/s41467-021-24119-3.

[3] A. Ammar, E. Giner, and A. Scemama. Optimization of large determinant expansions in quantum monte carlo. *Journal of Chemical Theory and Computation*, 18(9):5325–5336, 2022. doi:10.1021/acs.jctc.2c00556. URL https://doi.org/10.1021/acs.jctc.2c00556. PMID: 35997484.

[4] F. Becca and S. Sorella. *Quantum Monte Carlo Approaches for Correlated Systems*. Cambridge University Press, 2017. doi:10.1017/9781316417041.

[5] D. Bressanini. Implications of the two nodal domains conjecture for ground state fermionic wave functions. *Physical Review B*, 86(11): 115120, Sept. 2012. ISSN 1098-0121, 1550-235X. doi:10.1103/PhysRevB.86.115120. URL https://link.aps.org/doi/10.1103/PhysRevB.86.115120.

[6] G. Carleo and M. Troyer. Solving the quantum many-body problem with artificial neural networks. *Science*, 355(6325):602–606, 2017.

[7] D. M. Ceperley. Fermion nodes. *Journal of Statistical Physics*, 63(5):1237–1267, June 1991. ISSN 1572-9613. doi:10.1007/BF01030009. URL https://doi.org/10.1007/BF01030009.

[8] K. Choo, G. Carleo, N. Regnault, and T. Neupert. Symmetries and many-body excitations with neural-network quantum states. *Physical review letters*, 121 (16):167204, 2018.

[9] S. Dushman. Elements of the quantum theory. x. the hydrogen molecule. part i. heitler-london theory. *Journal of Chemical Education*, 13(6):287, 1936. doi:10.1021/ed013p287. URL https://doi.org/10.1021/ed013p287.

[10] R. P. Feynman and M. Cohen. Energy Spectrum of the Excitations in Liquid Helium. *Physical Review*, 102(5):1189–1204, June 1956. ISSN 0031-899X. doi:10.1103/PhysRev.102.1189. URL https://link.aps.org/doi/10.1103/PhysRev.102.1189.

[11] N. Gao and S. Günnemann. Ab-initio potential energy surfaces by pairing gnns with neural wave functions. *arXiv preprint arXiv:2110.05064*, 2021.

[12] N. Gao and S. Günnemann. Generalizing neural wave functions. *arXiv preprint arXiv:2302.04168*, 2023.

[13] 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? *Advances in Neural Information Processing Systems*, 35:10282–10294, 2022.

[14] 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?, 2022.

[15] W. A. Glauser, W. R. Brown, W. A. Lester, D. Bressanini, B. L. Hammond, and M. L. Koszykowski. Random-walk approach to mapping nodal regions of N -body wave functions: Ground-state Hartree–Fock wave functions for Li–C. *The Journal of Chemical Physics*, 97(12):9200–9215, Dec. 1992. ISSN 0021-9606, 1089-7690. doi:10.1063/1.463296. URL https://pubs.aip.org/aip/jcp/article/97/12/9200-9215/794560.

[16] P. W. Glynn. Likelihood ratio gradient estimation for stochastic systems. *Commun. ACM*, 33(10):75–84, oct 1990. ISSN 0001-0782. doi:10.1145/84537.84552. URL https://doi.org/10.1145/84537.84552.

[17] D. J. Griffiths and D. F. Schroeter. *Introduction to Quantum Mechanics*. Cambridge University Press, 3 edition, 2018. doi:10.1017/9781316995433.

[18] J. Han, L. Zhang, and E. Weinan. Solving many-electron schrödinger equation using deep neural networks. *Journal of Computational Physics*, 399:108929, 2019.

[19] J. Hermann, Z. Schätzle, and F. Noé. Deep-neural-network solution of the electronic schrödinger equation. *Nature Chemistry*, 12(10):891–897, 2020.

[20] M. Hutter. On Representing (Anti)Symmetric Functions, July 2020. URL http://arxiv.org/abs/2007.15298. arXiv:2007.15298 [quant-ph].

[21] M. H. Kalos. Monte carlo calculations of the ground state of three- and four-body nuclei. *Phys. Rev.*, 128:1791–1795, Nov 1962. doi:10.1103/PhysRev.128.1791. URL https://link.aps.org/doi/10.1103/PhysRev.128.1791.

[22] 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 preprint arXiv:2305.08831*, 2023.

[23] W. T. Lou, H. Sutterud, G. Cassella, W. Foulkes, J. Knolle, D. Pfau, and J. S. Spencer. Neural wave functions for superfluids. *arXiv preprint arXiv:2305.06989*, 2023.

[24] D. Luo and B. K. Clark. Backflow Transformations via Neural Networks for Quantum Many-Body Wave-Functions. *Physical Review Letters*, 122(22):226401, June 2019. ISSN 0031-9007, 1079-7114. doi:10.1103/PhysRevLett.122.226401. URL http://arxiv.org/abs/1807.10770. arXiv:1807.10770 [cond-mat, physics:physics].

[25] R. M. Martin, L. Reining, and D. M. Ceperley. *Interacting Electrons: Theory and Computational Approaches*. Cambridge University Press, 2016. doi:10.1017/CBO9781139050807.

[26] L. Mitas. Structure of fermion nodes and nodal cells. *Phys. Rev. Lett.*, 96:240402, Jun 2006. doi:10.1103/PhysRevLett.96.240402. URL https://link.aps.org/doi/10.1103/PhysRevLett.96.240402.

[27] L. Mitas. Fermion nodes and nodal cells of noninteracting and interacting fermions, 2006.

[28] A. Nagy and V. Savona. Variational quantum monte carlo method with a neural-network ansatz for open quantum systems. *Physical review letters*, 122(25): 250501, 2019.

[29] K. Neklyudov, J. Nys, L. Thiede, J. Carrasquilla, Q. Liu, M. Welling, and A. Makhzani. Wasserstein quantum monte carlo: A novel approach for solving the quantum many-body schr\" odinger equation. *arXiv preprint arXiv:2307.07050*, 2023.

[30] Y. Nomura, A. S. Darmawan, Y. Yamaji, and M. Imada. Restricted boltzmann machine learning for solving strongly correlated quantum systems. *Physical Review B*, 96(20):205152, 2017.

[31] T. Pang, S. Yan, and M. Lin. $O(N^2)$ Universal Antisymmetry in Fermionic Neural Networks, June 2022. URL http://arxiv.org/abs/2205.13205. arXiv:2205.13205 [physics, physics:quant-ph].

[32] G. Pescia, J. Nys, J. Kim, A. Lovato, and G. Carleo. Message-passing neural quantum states for the homogeneous electron gas. *arXiv preprint arXiv:2305.07240*, 2023.

[33] N. C. H. Peter T. A. Galek and W. A. Lester. Quantum monte carlo studies on small molecules. *Molecular Physics*, 104(19):3069–3085, 2006. doi:10.1080/00268970600962071. URL https://doi.org/10.1080/00268970600962071.

[34] D. Pfau, J. S. Spencer, A. G. d. G. Matthews, and W. M. C. Foulkes. Ab-Initio Solution of the Many-Electron Schr\"odinger Equation with Deep Neural Networks. *Physical Review Research*, 2(3):033429, Sept. 2020. ISSN 2643-1564. doi:10.1103/PhysRevResearch.2.033429. URL http://arxiv.org/abs/1909.02487. arXiv:1909.02487 [physics].

[35] H. Saito. Method to solve quantum few-body problems with artificial neural networks. *Journal of the Physical Society of Japan*, 87(7):074002, 2018.

[36] 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. *Nature Computational Science*, 2(5):331–341, 2022.

[37] M. Scherbela, L. Gerard, and P. Grohs. Towards a foundation model for neural network wavefunctions. *arXiv preprint arXiv:2303.09949*, 2023.

[38] Z. Schätzle, P. B. Szabó, M. Mezera, J. Hermann, and F. Noé. DeepQMC: An open-source software suite for variational optimization of deep-learning molecular wave functions. *The Journal of Chemical Physics*, 159(9):094108, 09 2023. ISSN 0021-9606. doi:10.1063/5.0157512. URL https://doi.org/10.1063/5.0157512.

[39] P. Seth, P. L. Ríos, and R. J. Needs. Quantum Monte Carlo study of the first-row atoms and ions. *The Journal of Chemical Physics*, 134(8):084105, 02 2011. ISSN 0021-9606. doi:10.1063/1.3554625. URL https://doi.org/10.1063/1.3554625.

[40] L. Thiede, C. Sun, and A. Aspuru-Guzik. Waveflow: Enforcing boundary conditions in smooth normalizing flows with application to fermionic wave functions, 2023.

[41] I. von Glehn, J. S. Spencer, and D. Pfau. A Self-Attention Ansatz for Ab-initio Quantum Chemistry, Apr. 2023. URL http://arxiv.org/abs/2211.13672. arXiv:2211.13672 [physics].

[42] I. von Glehn, J. S. Spencer, and D. Pfau. A Self-Attention Ansatz for Ab-initio Quantum Chemistry, Apr. 2023. URL http://arxiv.org/abs/2211.13672. arXiv:2211.13672 [physics].

[43] H. Xie, L. Zhang, and L. Wang. Ab-initio study of interacting fermions at finite temperature with neural canonical transformation. *arXiv preprint arXiv:2105.08644*, 2021.

Molecule	Exp. Energy	$O(N^2)$ Ansatz Sortlet Energy (Ours)	Vandermonde Energy[31]	$O(N^3)$ Ansatz Non-learned QMC[33; 39]	FermiNet[34]
Li	-7.4780	-7.477(8)	-7.4782	-7.4780	-7.4779
LiH	-8.0705	-8.070(3)	-	-8.070	-8.7050
Be	-14.6673	-14.667(1)	-14.6673	-14.6671	-14.6673
Li2	-14.9947	-14.994(5)	-	-14.9555	-14.9947
B	-24.6539	-24.652(7)	24.5602	-24.6533	-24.6537
C	-37.8450	-37.83(1)	-37.3531	-37.8437	-37.8447
N	-54.5892	-54.01(8)	-53.1855	-54.5873	-54.5888
CH4	-	-40.20(1)	-	-40.4416	-40.5140

Figure 4: Results from applying the Sortlet ansatz to a selection of atoms and small molecules. Energy values are all given in Hartree (Ha) atomic units. Chemical accuracy is defined to be within 1.5 mHa of the experimental values, green denotes values within this tolerance, red outside. Uncertainty is included in the parenthesis around the last digit – see Appendix B.

A. Assorted Proofs

A.1. Proof of the Variational Principle

Proof. Since multiplying Ψ by any complex constant $c \in \mathbb{C}$ leaves the value of $R(\Psi)$ unchanged, we can, without loss of generality, assume that $\langle \Psi | \Psi \rangle = 1$. Taking the orthonormal basis Ψ_n from the application of the spectral theorem to $\hat{H}$, we can expand

$$\Psi = \sum_{n=0}^{\infty} \alpha_n \Psi_n$$

but since $\langle \Psi_i | \Psi_j \rangle = \delta_{ij}$ (1 if $i = j$, 0 otherwise), plugging this expansion into R yields that

$$\begin{aligned} R(\Psi) &= R\left(\sum_{n=0}^{\infty} \alpha_n \Psi_n\right) \\ &= \left\langle \sum_{n=0}^{\infty} \alpha_n \Psi_n \left| \hat{H} \right| \sum_{m=0}^{\infty} \alpha_m \Psi_m \right\rangle \\ &= \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} \alpha_n \alpha_m^* \left\langle \Psi_n \left| \hat{H} \right| \Psi_m \right\rangle \\ &= \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} \alpha_n \alpha_m^* E_m \underbrace{\langle \Psi_n | \Psi_m \rangle}_{\delta_{mn}} = \sum_{n=0}^{\infty} \alpha_n^2 E_n \end{aligned}$$

Since we assumed $\langle \Psi | \Psi \rangle = 1$, it must hold $\sum_n \alpha_n^2 = 1$. This yields that the minimum of R is achieved precisely when

$$\alpha_0 = 1, \alpha_n = 0 : \forall n \geq 1$$

which implies $\Psi = \Psi_0$. □

A.2. Derivation of the energy gradient

First, let's rewrite the expectation of E_{loc} in bra-ket notation

$$\mathbb{E}_{x \sim \Psi_\theta^2}[E_{\text{loc}}(x)] = \frac{\langle \Psi_\theta | E_{\text{loc}} | \Psi_\theta \rangle}{\langle \Psi_\theta | \Psi_\theta \rangle} \tag{11}$$

Differentiating by θ yields

$$\frac{\partial}{\partial \theta} \left[\frac{\langle \Psi_\theta | E_{\text{loc}} | \Psi_\theta \rangle}{\langle \Psi_\theta | \Psi_\theta \rangle} \right] = \frac{\partial \langle \Psi_\theta | E_{\text{loc}} | \Psi_\theta \rangle}{\partial \theta} \frac{1}{\langle \Psi_\theta | \Psi_\theta \rangle} + \langle \Psi_\theta | E_{\text{loc}} | \Psi_\theta \rangle \frac{\partial}{\partial \theta} \frac{1}{\langle \Psi_\theta | \Psi_\theta \rangle} \tag{12}$$

For the derivative in the first term, since $\hat{H}$ is Hermitian, we have:

$$\frac{\partial \langle \Psi_\theta | E_{\text{loc}} | \Psi_\theta \rangle}{\partial \theta} = \frac{\partial \left\langle \Psi_\theta | \hat{H} | \Psi_\theta \right\rangle}{\partial \theta} \tag{13}$$

$$= 2 \left\langle \Psi_\theta | \hat{H} | \frac{\partial \Psi_\theta}{\partial \theta} \right\rangle = 2 \left\langle \Psi_\theta | \hat{H} | \frac{\partial \log \Psi_\theta}{\partial \theta} \Psi_\theta \right\rangle \tag{14}$$

$$= 2 \left\langle \Psi_\theta | \hat{E}_{\text{loc}} | \frac{\partial \log \Psi_\theta}{\partial \theta} \Psi_\theta \right\rangle \tag{15}$$

As for the derivative in the second term, some manipulations yield:

$$\frac{\partial}{\partial \theta} \frac{1}{\langle \Psi_\theta | \Psi_\theta \rangle} = -\frac{2}{\langle \Psi_\theta | \Psi_\theta \rangle^2} \left\langle \frac{\partial \Psi_\theta}{\partial \theta} | \Psi_\theta \right\rangle \tag{16}$$

$$= -\frac{2}{\langle \Psi_\theta | \Psi_\theta \rangle^2} \left\langle \frac{\partial \log \Psi_\theta}{\partial \theta} \Psi_\theta | \Psi_\theta \right\rangle \tag{17}$$

Substituting these expressions into (12), we find

$$\begin{aligned}
\frac{\partial}{\partial \theta} \left[\frac{\langle \Psi_\theta | E_{\text{loc}} | \Psi_\theta \rangle}{\langle \Psi_\theta | \Psi_\theta \rangle} \right] &= \frac{\partial \langle \Psi_\theta | E_{\text{loc}} | \Psi_\theta \rangle}{\partial \theta} \frac{1}{\langle \Psi_\theta | \Psi_\theta \rangle} + \langle \Psi_\theta | E_{\text{loc}} | \Psi_\theta \rangle \frac{\partial}{\partial \theta} \frac{1}{\langle \Psi_\theta | \Psi_\theta \rangle} \\
&= \left[2 \left\langle \Psi_\theta | \hat{E}_{\text{loc}} | \frac{\partial \log \Psi_\theta}{\partial \theta} \Psi_\theta \right\rangle \right] \frac{1}{\langle \Psi_\theta | \Psi_\theta \rangle} \\
&- \langle \Psi_\theta | E_{\text{loc}} | \Psi_\theta \rangle \left[\frac{2}{\langle \Psi_\theta | \Psi_\theta \rangle^2} \left\langle \frac{\partial \log \Psi_\theta}{\partial \theta} \Psi_\theta | \Psi_\theta \right\rangle \right] \\
&= 2 \frac{\left\langle \Psi_\theta | \hat{E}_{\text{loc}} | \frac{\partial \log \Psi_\theta}{\partial \theta} \Psi_\theta \right\rangle}{\langle \Psi_\theta | \Psi_\theta \rangle} - 2 \frac{\langle \Psi_\theta | E_{\text{loc}} | \Psi_\theta \rangle}{\langle \Psi_\theta | \Psi_\theta \rangle} \frac{\left\langle \frac{\partial \log \Psi_\theta}{\partial \theta} \Psi_\theta | \Psi_\theta \right\rangle}{\langle \Psi_\theta | \Psi_\theta \rangle}
\end{aligned}$$

At first, this might not appear like much of an improvement. Recognizing, however, that every term here can be written as an expectation against Ψ_θ^2, makes clear its usefulness

$$\boxed{\nabla E(\theta) = 2 \mathop{\mathbb{E}}_{x \sim \Psi_\theta^2} \left[E_{\text{loc}}(x) \nabla \log \Psi_\theta(x) - 2 \left(\mathop{\mathbb{E}}_{y \sim \Psi_\theta^2} [E_{\text{loc}}(y)] \right) \nabla \log \Psi_\theta(x) \right]} \tag{18}$$

In practice, we will estimate (**??**) using a finite collection of samples drawn from the normalized density $\frac{\Psi_\theta^2}{\langle \Psi_\theta | \Psi_\theta \rangle}$.

$$\nabla E(\theta) = 2 \frac{1}{n} \sum_{i=1}^{n} \left[E_{\text{loc}}(x_i) \nabla \log \Psi_\theta(x_i) - 2 \left(\frac{1}{n} \sum_{i=1}^{n} [E_{\text{loc}}(x_i)] \right) \nabla \log \Psi_\theta(x_i) \right] \tag{19}$$

Again, this can thankfully be achieved through standard MHMC techniques without having to estimate the normalization constant.

Remark: Isn't that estimate biased?
Mathematically, it is true that since we are reusing the same batch of x_i to estimate both expectations in (18) via (**??**), there will be bias introduced – the covariance between the two portions of the second term will be non-zero. Vanishing variance of the energy, however, will save us again here, since we can upper bound the covariance using Cauchy-Schwarz as

$$X = \frac{1}{n} \sum_{i=1}^{n} [E_{\text{loc}}(x_i)] \qquad Y = \frac{1}{n} \sum_{i=1}^{n} [\nabla \log \Psi_\theta(x_i)] \qquad \text{cov}(X, Y) \leq \sqrt{\text{Var}(X) \text{Var}(Y)} \tag{20}$$

and we know $\text{Var}(X) \longrightarrow 0$ as our estimate of the ground state improves. So for practical purposes, this correlation is negligible.

A.3. Proof that the sortlet is once-differatiable

Proof. By assumption, α is itself assumed smooth (C^∞), so since σ is constant except for points in the set

$$D = \left\{ r \in \mathbb{R}^{3N} : \exists i \in \{0..N\} \text{ such that } \alpha_i(r) = \alpha_{i+1}(r) \right\} \tag{21}$$

Ψ_α is also smooth on $\mathbb{R}^{3N} \backslash D$. Now for $r \in D$, continuity of Ψ_α at r is straightforward since $\prod_{i=1}^{N} (\alpha_{i+1}(r) - \alpha_i(r)) = 0$, which exactly cancels the discontinuity created by $\sigma(r)$.

Once-Differentiability

As for differentiability, we claim that the derivative is given by

$$\frac{\partial \Psi_\alpha}{\partial r^i}(x) = \left(\frac{\partial \alpha_{k+1}}{\partial r^i} - \frac{\partial \alpha_k}{\partial r^i} \right) \prod_{j=1, j \neq k}^{N} (\alpha_{j+1}(r) - \alpha_j(r)) \tag{22}$$

where k satisfies $\alpha_{k+1} - \alpha_k = 0$. By smoothness, we know this expression is correct on $\mathbb{R}^d \backslash D$, so the claim is that it extends, as one might expect, to D as well. To see this, let's consider two cases:

1. k **is non-unique.** Then it's easy to see that $\frac{\partial \Psi_\alpha}{\partial r^i} = 0$, since the resulting sum from applying the product rule always has a zero in each term, canceling any discontinuity potentially resulting from $\sigma(\pi_\alpha)$. To put this more rigorously, the limit from any direction $\lim_{r' \longrightarrow r} \frac{\partial \Psi_\alpha}{\partial r^i}(r') = 0$.

2. k **is unique**. Handling the case where k is unique requires a little more delicate manipulation. Consider that locally, for any $q \in B_\varepsilon(r)$

$$\psi_\alpha(q) = \sigma(\pi_\alpha) \left(\alpha_{k+1}(q) - \alpha_k(q) \right) \prod_{j=1, j \neq k}^{N} (\alpha_{j+1}(q) - \alpha_j(q)) \tag{23}$$

$$= \left(\widehat{\alpha}_{k+1}(q) - \widehat{\alpha}_k(q) \right) \prod_{j=1, j \neq k}^{N} (\alpha_{j+1}(q) - \alpha_j(q)) \tag{24}$$

where $\widehat{\alpha}_k$ and $\widehat{\alpha}_{k+1}$ are the same outputs of α, evaluated locally, but ordered according to a fixed $\pi_{\hat{\alpha}}(r)$. The exact $\varepsilon > 0$ can be computed by taking a minimum over all ϵ_i for $i \neq k$, defined to each satisfy

$$(\alpha_{i+1}(y) - \alpha_i(y)) > 0 : \forall y \in B_{\varepsilon_i}(x) \tag{25}$$

This means that locally, Ψ_α is smooth, since the sorted order of the other α_j does not change inside this neighbourhood.

□

A.4. Proof that the sortlet is anti-symmetric

Proof. The product $\prod_{i=1}^{N} [\alpha_{i+1} - \alpha_i]$ is invariant by design since the terms are always sorted before taking the pairwise difference, negating any permutation. $\sigma(\pi_\alpha)$ flips sign upon each odd permutation, since σ is multiplicative, and undoing the permutation then sorting the original array is equivalent to sorting the permutated array, i.e if τ is any transposition of two electrons with the same spin, $\pi_{\alpha(r)} \tau^{-1} = \pi_{\alpha(\tau r)}$ and so

$$\sigma(\pi_{\alpha(\tau r)}) = \sigma(\pi_\alpha) \sigma(\tau^{-1}) = -\sigma(\pi_\alpha) \tag{26}$$

since any transposition has parity -1. □

Note that to avoid over-constraining the ansatz by also making it anti-symmetric to transposition of opposite spin electrons – something the true wavefunction is **not** required to satisfy – we use the same trick as in [41], attaching a spin ± 1 term to each input electron coordinate r_i as $(r_i, \pm 1)$, which breaks equivariance of the output $\alpha(r)$ for those pairs.

A.5. Proof of Proposition 5

Proof. Let $r \in \mathbb{R}^{3N}$ be an arbitrary configuration, and let $\pi \in S(N^{\uparrow}) \cup S(N^{\downarrow})$ be a permutation on either the spin up or spin down electrons. We claim then that r and πr must always lie in separate nodal domains. To see this, let (i, j) be any indices of two electrons with the same spin who are transposed by π. Note that the product comprising ψ'_{pair} contains a pairwise term $[\phi_B(r_i, \{r\}) - \phi_B(r_j, \{r\})]$. Without loss of generality, asume $\phi_B(r_i, \{r\}) > \phi_B(r_j, \{r\})$. But for $\hat{r} = \pi r$ we must have

$$[\phi_B(r_i, \{r\}) - \phi_B(r_j, \{r\})] = -[\phi_B(\hat{r}_i, \{\hat{r}\}) - \phi_B(\hat{r}_j, \{\hat{r}\})] \tag{27}$$

(because $\hat{r}_i = r_j$ and $\hat{r}_j = r_i$, and ϕ_B is otherwise invariant to πr). Applying the intermediate value theorem (since ϕ_B is assumed continuous), it then follows for any path

$$\gamma(t) : [0, 1] \longrightarrow \mathbb{R}^{3N} \qquad \gamma(0) = r \qquad \gamma(1) = \hat{r} \tag{28}$$

connecting $r, \hat{r}$ must satisfy $[\phi_B(\gamma(t')_i, \{\gamma(t')\}) - \phi_B(\gamma(t')_j, \{\gamma(t')\})] = 0$ for some $t' \in (0, 1)$. Since $\psi'_{\text{pair}} = \prod_{i<j} [\phi_B(r_i, \{r\}) - \phi_B(r_j, \{r\})]$, zero in any $\phi_B(r_i, \{r\}) - \phi_B(r_j, \{r\})$ zeros Ψ. Since $\pi \in S(N^{\uparrow}) \cup S(N^{\downarrow})$ was arbitrary, so there are exactly $N^{\uparrow}! + N^{\downarrow}!$ nodal domains.

□

Remark: Since π was arbitrary, this also prohibits the sort of triple exchanges which in the exact ground state, connect same sign regions and collapse the number of nodal cells to 2 [27]. Additionally, although Beryllium is the largest system for which the two nodal domain conjecture is proven, it's frequently conjectured through the literature to be true universally, again see [7; 27; 5]. It's also known that π^* is not constant across nodal domains as claimed in [31] – regions were π^* is constant are referred to as permutation domains, and each nodal domain contains multiple permutation cells (see Figure 3 in [15]).

B. Experimental Details

All results in Figure 3a, Figure 3b and Section 6 were produced by modifying the FermiNet codebase to use our sortlet instead of the determinant. For the values in Figure 3b we optimized the wavefunction for 20,000 iterations, but for the others in Section 6, we ran each for 100,000 iterations. For the smaller systems, Li, LiH, Be, H_4 we used a batch size of 512. For those with more electrons, B, C, N we used between 2048 and 4096. We changed the initialization envelope parameters pursuant to [14], but we found that initializing to the suggested Z/row value too unstable with our ansatz, so we used 2 instead for Boron and Carbon. All training runs were completed using 2 A6000 GPUs, and ran for approximately 12 hours on average. The energy values except for N, CH_4 in Section 6 were computed by averaging over 10,000 separate estimates, with 500 mcmc equilibriation steps in between. For $N, CH4$, due to time constraints, the values were estimated using the last 5000 iterations of training. In both cases, the $3 - \sigma$ confidence interval reflected by the parenthesis (x) around a digit where estimated using the Gaussian formula from the central limit theorem as $3\sigma = 3\frac{\hat{\sigma}}{\sqrt{n}}$. Molecules below C in Section 6 used $K = 16$ determinants, but those above used $K = 32$.