Convergence to the fixed-node limit in deep variational Monte Carlo

Z. Schätzle,[1] J. Hermann,[1, 2] and F. Noé[1, 3, 4, a]

[1]*FU Berlin, Department of Mathematics and Computer Science, Arnimallee 6, 14195 Berlin, Germany*

[2]*TU Berlin, Machine Learning Group, Marchstr. 23, 10587 Berlin, Germany*

[3]*FU Berlin, Department of Physics, Arnimallee 14, 14195 Berlin, Germany*

[4]*Rice University, Department of Chemistry, Houston, TX 77005, USA*

Variational quantum Monte Carlo (QMC) is an ab-initio method for solving the electronic Schrödinger equation that is exact in principle, but limited by the flexibility of the available ansatzes in practice. The recently introduced deep QMC approach, specifically two deep-neural-network ansatzes PauliNet and FermiNet, allows variational QMC to reach the accuracy of diffusion QMC, but little is understood about the convergence behavior of such ansatzes. Here, we analyze how deep variational QMC approaches the fixed-node limit with increasing network size. First, we demonstrate that a deep neural network can overcome the limitations of a small basis set and reach the mean-field complete-basis-set limit. Moving to electron correlation, we then perform an extensive hyperparameter scan of a deep Jastrow factor for LiH and H_4 and find that variational energies at the fixed-node limit can be obtained with a sufficiently large network. Finally, we benchmark mean-field and many-body ansatzes on H_2O, increasing the fraction of recovered fixed-node correlation energy of single-determinant Slater–Jastrow-type ansatzes by half an order of magnitude compared to previous variational QMC results and demonstrate that a single-determinant Slater–Jastrow–backflow version of the ansatz overcomes the fixed-node limitations. This analysis helps understanding the superb accuracy of deep variational ansatzes in comparison to the traditional trial wavefunctions at the respective level of theory, and will guide future improvements of the neural network architectures in deep QMC.

I. INTRODUCTION

The fundamental problem in quantum chemistry is to solve the electronic Schrödinger equation as accurately as possible at a manageable cost. Variational quantum Monte Carlo (variational QMC, or VMC for short) is an ab-initio method based on the stochastic evaluation of quantum expectation values that scales favorably with system size and provides explicit access to the wavefunction[1]. Although exact in principle, VMC strongly depends on the quality of the trial wavefunction, which determines both efficiency and accuracy of the computation and typically constitutes the limiting factor of VMC calculations.

Recently, deep QMC has been introduced. Deep QMC involves a new class of ansatzes that complement traditional trial wavefunctions with the expressiveness of deep neural networks (DNNs). This ab-initio approach is orthogonal to the supervised learning of electronic structure that requires external datasets[2,3]. The use of neural network trial wavefunctions has been pioneered for spin lattice systems[4] and later generalized to molecules in second quantization[5]. The first application to molecules in real space was a proof-of-principle effort, but did not reach the accuracy close to traditional VMC[6]. The DNN architectures PauliNet and FermiNet advanced the real-space deep QMC approach[7,8], increasing the accuracy to state-of-the-art levels and beyond. Demonstrating very high accuracy with far fewer determinants than traditional counterparts, these deep neural network trial wavefunctions provide an alternative to increasing the number of Slater determinants, thus potentially improving the unfavorable scaling with respect to the number of electrons that complicates accurate calculations for large systems. Application of the deep QMC method to many-particle quantum systems other than electrons is also possible[9].

Currently, there is little understanding of why these DNN wavefunctions work well and how their individual components contribute to the approximation of the ground-state wavefunction and energy. Examining their expressive power and measuring their accuracy in comparison to traditional approaches is essential to establish neural-network trial wavefunctions as a standard technique in VMC and to guide further development.

Here, we identify a hierarchy of model ansatzes based on the traditional VMC methodology (Fig. 1) that enables us to distinguish the effects of improving single-particle orbitals and adding correlation in the symmetric part of the wavefunction ansatz. This is of particular interest in the context of discriminating these improvements from reducing the energy by solving the intricate problem of missing many-body effects in the nodal surface.

The trial wavefunctions in QMC are typically constructed by combining a symmetric Jastrow factor with an antisymmetric part that implements the Pauli exclusion principle for fermions by specifying the nodal surface of the ansatz—the hyperplane in the space of electron coordinates, $\mathbf{r} = (\mathbf{r}_1, \ldots, \mathbf{r}_N)$, on which the wavefunction changes sign. Expressing the antisymmetric part as a linear combination of Slater determinants gives rise to the ansatz of the Slater–Jastrow–backflow-type that comprises most VMC ansatzes, including the deep variants PauliNet and FermiNet,

$$\psi(\mathbf{r}) = \underbrace{e^{J(\mathbf{r})}}_{\text{symmetric}} \underbrace{\sum_p c_p \det[\mathbf{A}_p^\uparrow(\mathbf{r})] \det[\mathbf{A}_p^\downarrow(\mathbf{r})]}_{\text{antisymmetric}} \tag{1}$$

The ability of neural networks to represent antisymmetric (wave) functions has been also explored theoretically[10,11].

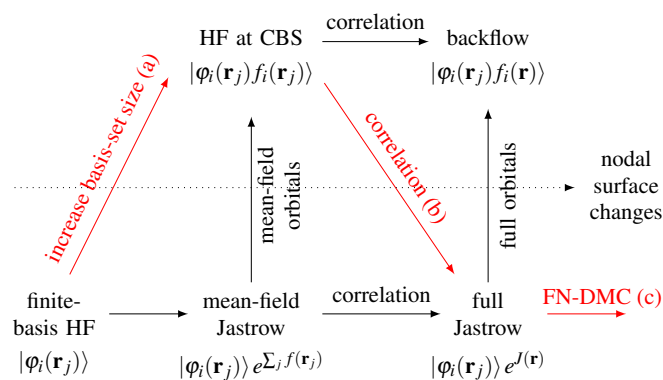[a]Electronic mail: frank.noe@fu-berlin.de

FIG. 1. **Hierarchy of single-determinant ansatzes in QMC.** The starting point of a finite-basis Hartree–Fock (HF) calculation can be extended by a "mean-field" Jastrow factor to improve the one-electron density of the ansatz. From that point, the ansatz can be improved in one of two directions—by modifying the orbitals (bottom–top) or introducing electron correlation (left–right). The red pathway shows a standard approach in traditional QMC and is the path we pursue in our analysis.

Traditionally, Slater determinants are antisymmetrized product states constructed from single-particle molecular orbitals, which are expressed in a one-electron basis set consisting of basis functions ϕ_k,

$$A_{ij} = \varphi_i(\mathbf{r}_j) = \sum_k c_{ik}\phi_k(\mathbf{r}_j) \qquad (2)$$

Employing such basis sets transforms the problem of searching over infinitely many functions into a problem of searching over coefficients in a system of equations, which can be solved by means of linear algebra applying for instance the Hartree–Fock (HF), the multi-configurational self-consistent field (MCSCF), or the full configuration interaction (FCI) method. The projection comes at the cost of introducing the finite-basis-set error, which completely vanishes only in the limit of infinitely many basis function—the complete-basis-set (CBS) limit (Fig. 1a). Finite-basis-set errors are inherent to the second-quantized representation, which nevertheless provides an alternative platform to introduce deep learning to quantum chemistry[12].

The real-space formulation of VMC allows to introduce explicit electron correlation efficiently by modelling many-body interactions with a Jastrow factor (Fig. 1b). The Jastrow factor is a symmetric function of the electron coordinates that traditionally involves an expansion in one, two, and three-body terms[13]. Although strongly improving the ansatz, traditional Jastrow factors do not have sufficient expressiveness to reach high accuracy and an initial VMC calculation is typically followed by a computationally demanding fixed-node diffusion QMC (FN-DMC) simulation (Fig. 1c), which eventually projects out the exact solution for the given nodal surface—the fixed-node limit[14]. DMC is based on the imaginary-time Schrödinger equation and offers yet another entry point for the use of neural networks to represent quantum states[15,16].

The nodal surface of the trial wavefunctions can be improved by increasing the number of determinants or by applying the backflow technique, transforming single-particle orbitals to many-body orbitals under consideration of the symmetry constraints. These are key concepts to efficiently reach very high accuracy with VMC and integral features of deep QMC. Using multiple determinants, applying the backflow technique and modifying the symmetric component of the ansatz at the same time however makes it difficult identify the contributions of each individual part. Benchmarking deep QMC ansatzes in conceptually simpler contexts confirms their correct functionality and helps to achieve a better understanding.

In this paper we take a closer look at how neural networks compensate for errors arising from finite basis sets and demonstrate convergence to the fixed-node limit within the VMC framework by systematically increasing the expressiveness of a deep Jastrow factor. For the sake of disentangling the individual contributions to the overall accuracy, we conduct our analysis mainly with Slater–Jastrow-type trial wavefunctions with an antisymmetric part consisting of a single determinant, that is, with ansatzes possessing a mean-field nodal surface. We compare neural-network variants with traditional functional forms, as well as with DMC results. In particular we investigate the PauliNet, a recently proposed neural-network trial wavefunction[7]. PauliNet combines ideas from conventional trial wavefunctions, such as a symmetric Jastrow factor, a generalized backflow transformation, multideterminant expansions, quantum chemistry baselines, and an explicit implementation of physical constraints of ground-state wavefunctions. Since PauliNet is a powerful instance of the general ansatz in (1), we can obtain traditional types of QMC ansatzes at different levels of the theory by deactivating certain trainable parts of PauliNet. The hierarchy of ansatzes sketched in Fig. 1 maps restricted single-determinant versions of PauliNet and their eventual expressiveness in the context of the traditional single-determinant VMC approach. The incentive of implementing restricted variants of PauliNet is to test the behavior of the ansatz in settings that are well solved by existing methods and investigate the expressiveness of the individual components of PauliNet on well-defined subproblems. These restricted variants however are not intended to be used in order to achieve best accuracy, which is attained when taking advantage of the full flexibility of the PauliNet ansatz as demonstrated previously[7].

The rest of the paper is organized as follows. In Sec. II we review the general PauliNet ansatz and show how different levels of the model hierarchy (Fig.1) can be obtained. In Sec. III we use these instances of PauliNet to investigate several subproblems of the fixed-node limit within the deep QMC approach. First we demonstrate that DNNs can be employed to correct the single-particle orbitals of a HF calculation in a small basis and obtain energies close to the CBS limit. Next we benchmark the deep Jastrow factor. We start by applying it to two node-less test systems, H_2 and He, where results within five significant digits of the exact energy are achieved. Next we conduct an extensive hyperparameter search for two systems with four electrons, LiH and the H_4 rectangle, revealing that the expressiveness of the ansatz can be systematically increased to converge to the fixed-node limit

imposed by the employed antisymmetric ansatz. We further explore the convergence aspect by sampling the dipole moment for the LiH ansatzes and evaluating energy differences for two configurations of the hydrogen rectangle. Thereafter we show the size consistency of the method, examining the optimization of the deep Jastrow factor for systems of non-interacting molecules (H_2-H_2 and LiH-H_2). Finally we test various single-determinant variants of PauliNet in an analysis of the water molecule and compare them to traditional trial wavefunctions. Section IV discusses the results.

II. THEORY AND METHODS

A. PauliNet

The central object of our investigation is PauliNet, a neural-network trial wavefunction of the form in (1). PauliNet extends the traditional multi-determinant Slater–Jastrow–backflow-type trial wavefunctions[17], retaining physically motivated structural features while replacing ad-hoc parameterizations with highly expressive DNNs,

$$\psi_{\boldsymbol{\theta}}(\mathbf{r}) = e^{\gamma(\mathbf{r})+J_{\boldsymbol{\theta}}(\mathbf{r})} \sum_p c_p \det[\tilde{\varphi}^{\uparrow}_{\boldsymbol{\theta},\mu_{pi}}(\mathbf{r})] \det[\tilde{\varphi}^{\downarrow}_{\boldsymbol{\theta},\mu_{pi}}(\mathbf{r})] \quad (3)$$

$$\tilde{\varphi}_{\boldsymbol{\theta},\mu i}(\mathbf{r}) = \varphi_\mu(\mathbf{r}_i) f^{\otimes}_{\boldsymbol{\theta},\mu i}(\mathbf{r}) + f^{\oplus}_{\boldsymbol{\theta},\mu i}(\mathbf{r}) \quad (4)$$

The ansatz consists of a linear combination of Slater determinants of molecular single-particle orbitals φ_μ corrected by a generalized backflow transformation $\mathbf{f}_{\boldsymbol{\theta}}$ and of a Jastrow factor $J_{\boldsymbol{\theta}}$. The DNN components are indicated by the $\boldsymbol{\theta}$ subscript, denoting the trainable parameters of the involved neural networks. The expansion coefficients c_p and the single-particle orbitals are initialized from a preceding standard quantum-chemistry calculation (HF or MCSCF). The analytically known electron–nucleus and electron–electron cusp conditions[18] are enforced within the orbitals φ_μ and as a fixed part γ of the Jastrow factor, respectively. The correct cusps are maintained by designing the remaining trial wavefunction architecture to be cusp-less.

Both backflow transformation and Jastrow factor can introduce many-body correlation and are constructed in such a way that they preserve the anti-symmetry of the trial wavefunction. The Jastrow factor consists of a symmetric function, that is to say it retains the antisymmetry upon being invariant under the exchange of same-spin electrons. This however has the consequence that it scales the wavefunction without altering the nodes of the ansatz. The backflow transformation on the other hand alters the nodal surface by acting on the orbitals directly. Traditionally the backflow correction introduces many-body correlation by assigning quasiparticle coordinates that get streamed through the original orbitals. PauliNet generalizes this concept, based on the observation that equivariance with respect to the exchange of electrons is a sufficient criterion to retain the antisymmetry of the Slater determinant, and considers the backflow correction as a many-body transformation of the orbitals themselves. In fact it has been shown that in principle a single Slater determinant

with generalized orbitals is capable of representing any antisymmetric function, if the many-body orbitals are sufficiently expressive[11]. Both Jastrow factor $J_{\boldsymbol{\theta}}$ and backflow transformation $\mathbf{f}_{\boldsymbol{\theta}}$ are obtained from a joint latent-space representation encoded by a graph-convolutional neural network. The network acts on the rotation- and translation-invariant representation of the system given by the fully-connected graph of distances between all electrons and nuclei. The latent-space many-body representation is designed to be equivariant under the exchange of same-spin electrons, which is used to construct the permutation-equivariant backflow transformation and the permutation-invariant Jastrow factor. Details on the graph-convolutional neural network architecture can be found in Sec. A 1. Combining an expansion in Slater-determinants with Jastrow factor and backflow transformation introduces multiple different ways to model many-body effects, helping to efficiently encode correlation in the ansatz by representing i.e. dynamic correlation explicitly while implementing static correlation with multiple determinants.

The PauliNet ansatz is optimized according to the standard VMC scheme[1] of minimizing its energy expectation value. This is based on the variational principle of quantum mechanics that guarantees the energy expectation value of any trial wavefunction to be lower-bounded by the ground-state energy, as long as the fermionic antisymmetry constraint is implemented,

$$E_0 = \min_{\psi} \langle \psi | \hat{H} | \psi \rangle \leq \min_{\boldsymbol{\theta}} \langle \psi_{\boldsymbol{\theta}} | \hat{H} | \psi_{\boldsymbol{\theta}} \rangle, \quad \psi \in \mathscr{H}^- \quad (5)$$

In VMC this expectation value is approximated by Monte Carlo integration,

$$\langle \psi_{\boldsymbol{\theta}} | \hat{H} | \psi_{\boldsymbol{\theta}} \rangle \approx \frac{1}{M} \sum_{k=1}^{M} \frac{\hat{H}\psi_{\boldsymbol{\theta}}(\mathbf{r}_k)}{\psi_{\boldsymbol{\theta}}(\mathbf{r}_k)}, \quad \mathbf{r}_k \in \mathbb{R}^{3N} \sim |\psi_{\boldsymbol{\theta}}|^2 \quad (6)$$

In practice this gives rise to an alternating scheme of sampling electronic configurations according to the probability density associated with the trial wavefunction with a standard Langevin sampling approach and optimizing the parameters of this wavefunction by following their (stochastic) gradient with respect to estimates of the expectation value over small batches. For further details of the training methodology see ref. 7. Numerical calculations were carried out with the DeepQMC Python package[19], with training hyperparameters as reported in Table VI.

Next we show how to obtain the ansatzes of Fig. 1 from the general PauliNet architecture and introduce the respective optimization problems to be solved.

B. Deep orbital correction

The simplest way to approach the quantum many-body problem is by considering a mean-field theory. The HF method gives the optimal mean-field solution within the space of the employed basis set. A mean-field variant of the PauliNet architecture can be used to account for finite-basis-set errors in the HF baseline, by introducing a real-space cor-

rection to the single-particle orbitals,

$$\psi_\theta(\mathbf{r}) = \det[\tilde{\varphi}_{\theta,\mu}^\uparrow(\mathbf{r}_i)]\det[\tilde{\varphi}_{\theta,\mu}^\downarrow(\mathbf{r}_i)] \tag{7}$$

$$\tilde{\varphi}_{\theta,\mu}(\mathbf{r}_i) = \varphi_\mu(\mathbf{r}_i)f_{\theta,\mu}^\otimes\big(\mathbf{x}_i^{(L)}(\mathbf{r}_i)\big) + f_{\theta,\mu}^\oplus\big(\mathbf{x}_i^{(L)}(\mathbf{r}_i)\big) \tag{8}$$

The functions $f_\theta^\otimes$ and $f_\theta^\oplus$ are implemented by DNNs that generate a multiplicative and an additive correction to the HF orbitals φ_μ, respectively. Combining a multiplicative and an additive correction serves the practical purpose of facilitating the learning process, as the multiplicative correction has a strong effect where the value of the orbital is large, while the additive correction can alter the nodes of the molecular orbital. (In principle, an additive correction only would be a sufficient parameterization.) This approach is a special case of the generalized backflow transformation in (4), in which the backflow correction depends on the position of the i-th electron only. The single-particle representation $\mathbf{x}_i^{(L)}(\mathbf{r}_i)$ can be obtained by a slight modification of the graph-convolutional architecture as described in Section A 2. If Gaussian-type orbitals are used, it is common to correct the missing nuclear cusp at the coalescence points within the orbitals. We employ the cusp correction of Ma *et al.* [20] and construct the DNNs to be cusp-less. Though the DNN could in principle approximate the orbitals from scratch, providing the HF baseline that ensures the correct asymptotics and offers a good initial guess reduces the training cost and makes the training process more robust. In the mean-field theory the HF energy at the CBS limit constitutes a benchmark for the best possible solution to the optimization problem.

C. Deep Jastrow factor

The Slater–Jastrow-type ansatz goes beyond the mean-field theory by introducing explicit electronic correlation. The symmetric Jastrow factor however cannot alter the nodal surface and the single-determinant Slater–Jastrow-type ansatz is therefore a many-body ansatz possessing a mean-field nodal surface,

$$\psi_\theta(\mathbf{r}) = e^{\gamma(\mathbf{r})+J_\theta(\mathbf{r})}\det[\varphi_\mu^\uparrow(\mathbf{r}_i)]\det[\varphi_\mu^\downarrow(\mathbf{r}_i)] \tag{9}$$

The deep Jastrow factor J_θ is obtained from the latent-space many-body representation encoded by the graph-convolutional neural network described in Sec. A 1,

$$J_\theta(\mathbf{r}) := \eta_\theta\big(\textstyle\sum_i \mathbf{x}_i^{(L)}(\mathbf{r})\big) \tag{10}$$

To enforce the symmetry of the Jastrow factor, the permutation-equivariant many-body embeddings $\mathbf{x}_i^{(L)}$ are summed over the electrons to give permutation invariant features. These features serve as input to a fully-connected neural network η_θ, which returns the final Jastrow factor. The process of obtaining the latent-space representation involves multiple smaller components, such as trainable arrays and fully-connected neural networks, whose full specification gives rise to a collection of hyperparameters that influence the expressiveness of the ansatz. A list of the components and the respective hyperparameters can be found in Table V.

Benchmarking Jastrow factors comes with the difficulty of distinguishing errors arising from the nodal surface from those present due to a lack of expressiveness in the Jastrow factor. The optimal energy of a Slater–Jastrow-type trial wavefunction however can be obtained with the FN-DMC algorithm, that gives the exact ground state of the Schrödinger equation under the fixed-node constraint of the antisymmetric part of the ansatz.

D. Mean-field Jastrow factor

We furthermore implement a mean-field Jastrow factor, which constitutes another point in the space of ansatz classes (Fig. 1),

$$J_\theta(\mathbf{r}) := \textstyle\sum_i \eta_\theta\big(\mathbf{x}_i^{(L)}(\mathbf{r}_i)\big) \tag{11}$$

The mean-field Jastrow factor can optimize the one-electron density of the ansatz without modifying the nodal surface or introducing correlation, making its variations a strict subset of the orbital correction. This equips us with an intermediate step in approaching the finite-basis-set limit, that can be used to relate the finite-basis-set error to the fixed-node error of the HF baseline. If the many-body Jastrow factor is used, the mean-field version is not needed, as it is implicitly included in the many-body version.

III. RESULTS

A. Large basis sets are not necessary in DNN ansatzes

We start from a HF baseline obtained in the small 6-31G basis set. Instead of introducing more basis functions, the PauliNet ansatz follows the alternative approach of correcting the orbitals directly in real space. We trained the mean-field variant of the PauliNet ansatz (Sec. II B) on H_2, He, Be, LiH and the hydrogen square H_4. For all five test systems we obtained energies close to the extrapolated CBS limit and recovered at least 97% of the finite-basis-set error (Fig. 2). This shows that the use of a very small basis set for the baseline of PauliNet does not introduce any fundamental limitation to accuracy, because the neural network is able to correct it. We note that such an approach to the CBS limit is practical only within the context of the full PauliNet, not as a standalone technique to replace large basis sets in quantum chemistry.

B. Exact solutions for two electron systems

Next we turn to modelling electron correlation with the deep Jastrow factor (Sec. II C). We start by evaluating the deep Jastrow factor for H_2 and He, two-electron closed-shell systems for which the ground state is node-less (the antisymmetry comes from the spin part of the wavefunction only),
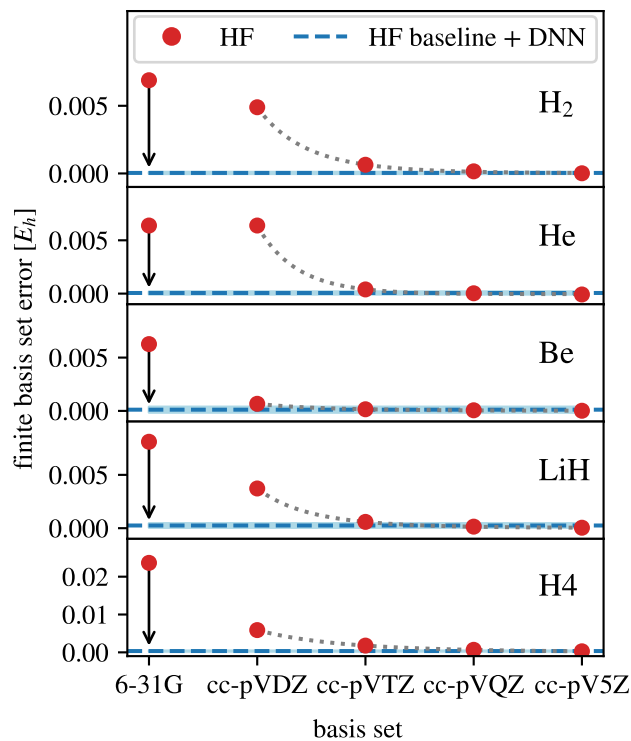
FIG. 2. **Removing the basis-set error of a HF calculation**. The error with respect to the estimated CBS limit of HF calculations with increasingly large basis sets, as well as the result of employing a DNN to correct the single-particle orbitals of a HF calculation in a small basis (6-31G) are shown. The statistical error of the Monte Carlo integration is shown in light blue. The DNN is capable to correct deficiencies arising from the finite basis, producing energies close to the CBS limit.

such that the Jastrow factor is, in principle, sufficient to reproduce exact results. This yields a pure test for the expressiveness of the deep Jastrow factor. The recovered many-body correlation is measured by the correlation energy,

$$\eta = \frac{E_{\mathrm{VMC}} - E_{\mathrm{HF}}}{E_{\mathrm{exact}} - E_{\mathrm{HF}}} \tag{12}$$

TABLE I. **Results for two-electron node-less systems.**

system	deep Jastrow factor	exact energy	η [%]
H_2 ($d = 1.4$)	-1.17446(1)	-1.1744748[21]	99.97(3)
He	2.90372(1)	-2.9037247[23]	99.98(2)

For both systems we obtain energies matching five significant digits of the exact references (Table I). We evaluate the ansatz along the dissociation curve of H_2 (Fig. 3). Deep QMC outperforms FCI even with the large cc-pV5Z basis set, reducing the error in correlation energy by one to two orders of magnitude at compressed geometries and still being more accurate at stretched geometries, where the system exhibits
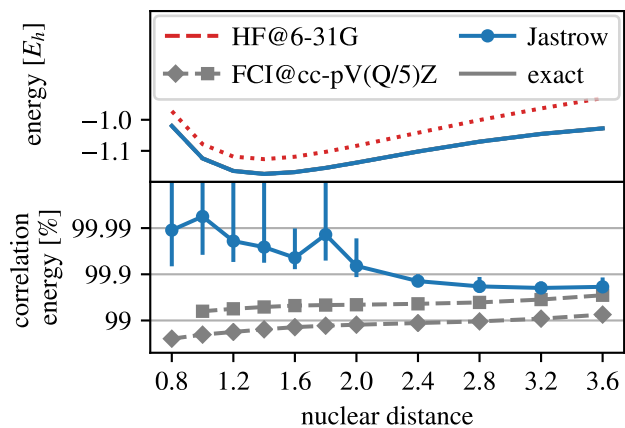

FIG. 3. **Dissociation curve of the hydrogen molecule.** Upper panel shows total energy. Exact results[21] cannot be distinguished from FCI and the deep Jastrow factor. Lower panel shows the percentage of correlation energy recovered. FCI results were obtained with PySCF[22] in the cc-pVQZ basis (orange) and cc-pV5Z basis (green). Deep QMC surpasses the FCI accuracy for the entire dissociation curve.

static correlation and the restricted HF baseline gives qualitatively wrong results (ionic contributions resulting in negative interaction energy). The results demonstrate the difficulty of modelling dynamic correlation in Slater-determinant space when applying purely second-quantized approaches and showcase the advantages of explicitly encoding many-body correlations.

C. Systematically approaching the fixed-node limit

The complexity of modeling correlation increases steeply with the number of particles. We evaluate the performance of the deep Jastrow factor for LiH and the hydrogen rectangle H_4. While these four-electron systems exhibit more intricate interactions, they are computationally lightweight, such that the hyperparameter space of the deep ansatzes can be explored exhaustively. With multiple same-spin electrons, the spatial wavefunction is no longer node-less and the single-determinant Slater–Jastrow ansatz possesses a fixed-node error. Instead of comparing to exact energies, we therefore measure the performance of the Jastrow factor with respect to the fixed-node limit estimated from FN-DMC calculations, and reporting the fixed-node correlation energy,

$$\eta_{\mathrm{FN}} = \frac{E_{\mathrm{VMC}} - E_{\mathrm{HF}}}{E_{\mathrm{DMC}} - E_{\mathrm{HF}}} \tag{13}$$

As the fixed-node correlation energy is defined for ansatzes with an identical nodal surface, the nodes of the FN-DMC benchmark have to be reconstructed. For the mean-field nodal surface this implies starting from a HF computation with the same basis set.

For the H_4 rectangle we performed a scan of all the hyperparameters of the deep Jastrow factor including those of the
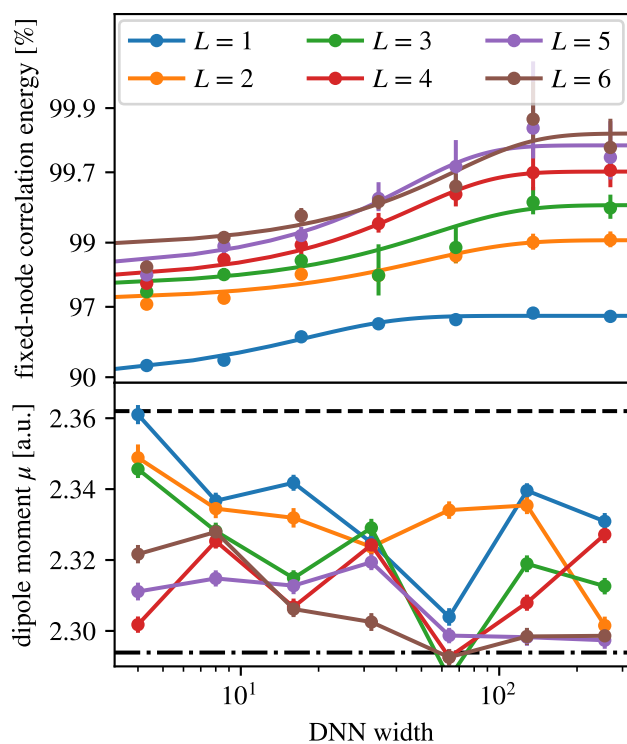
FIG. 4. **Approaching FN-DMC accuracy with deep Jastrow factor.** For LiH increasingly expressive Jastrow factors are trained. The dependence on the dimension the convolutional kernel (DNN width) and the number of interactions (L) in the Jastrow graph-neural-network is shown. The orbitals of the antisymmetric part are expressed in the TZP basis[24]. The upper panel shows the energy of the trial wavefunctions. The most accurate ansatzes give results within the sampling error of the FN-DMC energy of a single-determinant benchmark from Casalegno, Mella, and Rappe [25], indicated by the shaded region at the upper end of the graph. In the lower panel the dipole moment of the trial wavefunctions is evaluated. As a reference the dipole moment of a HF (dashed line) and of an explicitly-correlated coupled-cluster (CCSD(T)-R12) calculation[26] (dashed doted line) are shown. For the models with five and six interactions the dipole moment converges to the results from the coupled-cluster calculation.

graph-convolutional neural network architecture (Table V). The scan was a grid search that involved the training of 864 models, comprising models with all combinations of the hyperparameters in the vicinity of their default values. In order to reduce the dimensionality of the experiment some hyperparameters were merged and varied together. Further details of the scan can be found in the caption of Fig. 8, depicting the energies of all the model instances. The experiment aimed at obtaining a first impression of the hyperparameter space and revealed that by increasing the total number of trainable parameters, the fixed-node limit can be approached. The experiment shows that the energy behaves smoothly with respect to changes in the hyperparameters and there are no strong mutual dependencies between hyperparameters. Several important hyperparameters for systematically scaling the architec-
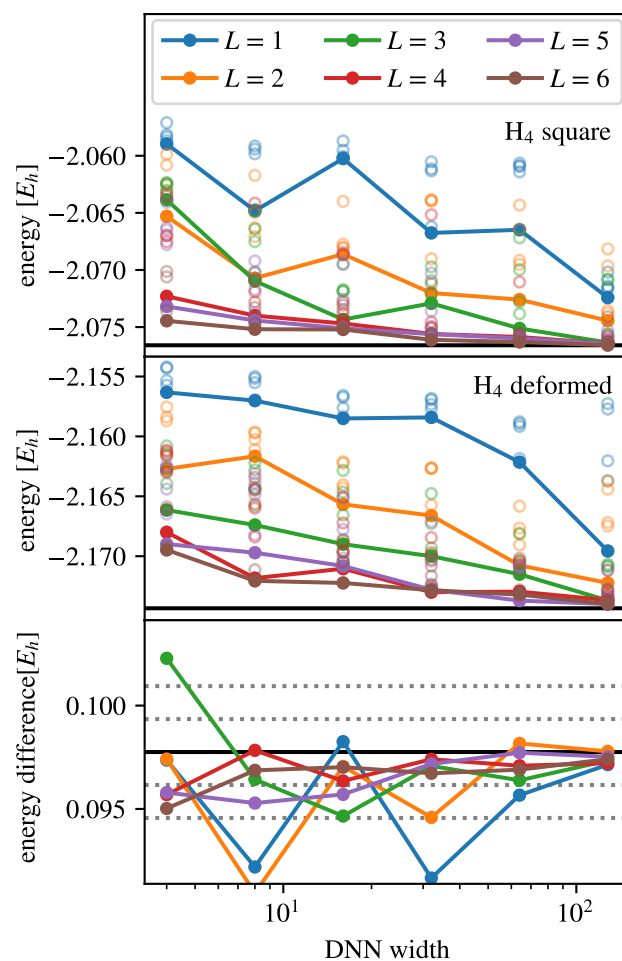

FIG. 5. **Relative energy for hydrogen rectangle.** The convergence of the energy to the DMC reference[27] (black line) with increasing model size for two geometries of the hydrogen rectangle, as well as the relative energy difference is shown. The minimal energy over five independent runs is highlighted in the upper plots and used to compute the relative energy. The dashed lines indicate error margins of 1 and 2 kcal/mol.

ture can be identified, such as the depth of the neural network η_θ from (10), the number of interactions L and the dimension of the convolutional kernel, referring to the dimension of the latent space where the interactions within the graph-convolutional neural network take place (Sec. A 1).

The results were used to perform a thorough investigation of the convergence behavior on LiH, varying a subset of hyperparameters and fixing the remaining hyperparameters at suitable values. We show a systematic convergence to the fixed-node limit with an increasing dimension of the convolutional kernel (DNN width) as well as the number of interactions L (Fig. 4). This is an indication that the deep Jastrow factor can be extended towards completeness in a computationally feasible way. The remaining fluctuations of the fixed-node correlation energy are caused by the stochasticity of the training and the sampling error of evaluating the energy of the wavefunction.

By evaluating the dipole moment of the LiH wavefunctions we go beyond the energy and investigate the convergence of a property that the PauliNet ansatz is not explicitly optimized for. We found that upon converging to the fixed-node limit with increasingly large models the dipole moment approaches the coupled cluster reference (Fig. 4). Even though the energies of the LiH wavefunctions converge consistently, the convergence of the dipole moment is subject to fluctuations, which are particularly strong for the small models and decrease as the fixed-node limit is approached. This can be explained by degenerate energy minima of the ansatz with respect to the parameters. Multiple solutions to the optimization problem can be present if the exact solution is outside the variational subspace. This ambiguity however decreases with increasing expressiveness of the trial wavefunction.

While the accuracy with respect to the total energy is an appropriate measure for expressiveness of the trial wavefunction ansatz, in practice relative energies are most often of interest. The capability of the full PauliNet ansatz in computing relative energies has been previously demonstrated for the cyclobutadiene automerization[7] and the results with the deep Jastrow factor for the node-less H_2 (Fig. 3) provide interaction energies at the level of FCI. Here we want to study how the relative energy converges with an increasing expressiveness of the deep Jastrow factor and demonstrate a cancellation of errors at different geometries. This is a feature that makes relative energy calculations usually more accurate than total energy calculations and is very desirable for any quantum chemistry method. We optimized increasingly expressive versions of the deep Jastrow factor for two geometries of the hydrogen rectangle and determined their relative energy (Fig. 5). In order to reduce the level of stochasticity in the training we performed five independent optimization runs and used the ones with the lowest energy to calculate the relative energies. Both the total and relative energy converge to the DMC reference with an increasing number of trainable parameters of the ansatz. Furthermore the relative energy fluctuates within 2 kcal/mol of the DMC reference for all models with more than two interactions, and is well within 1 kcal/mol for the models with the largest DNN width. This demonstrates that the deep Jastrow factor can achieve similar accuracy for both geometries and exhibits a cancellation of errors. The stochasticity of the optimization however complicates the comparison of individual runs, which will be subject to further investigation. Looking at the energy of the different optimizations though, we found a decrease in stochasticity of the final energy with increasing model size, which is convenient for practical purpose, where typically large models would be used. The difficulty of optimizing small models is well-known in the context of training neural networks, which tends to be improved by increasing the number of trainable parameters[28].
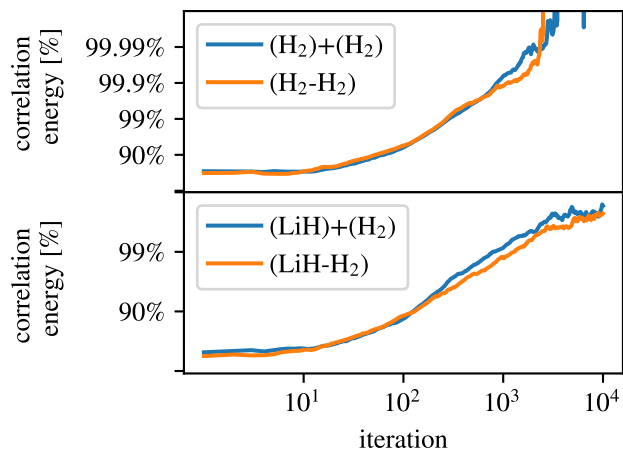

FIG. 6. **Size consistency of deep Jastrow factor optimization.** The figure shows the smoothed training curves of the deep Jastrow factor for two systems of non-interacting molecules. The optimization of an ansatz for the joint system (orange) is compared to the independent optimization of two separate ansatzes for the subsystems respectively (blue).

One of the essential properties of any proper electronic structure method is size consistency. Traditional Jastrow factors are factorizable in the electronic and nuclear coordinates of two infinitely distant subsystems, which leads to exact size consistency for identical copies of a given system, and to approximate size consistency for an assembly of different systems (because optimized parameters are now shared by different systems). In PauliNet the embeddings $\mathbf{x}_i$ for two electrons at two distant subsystems are independent of each other by construction. Although the subsequent nonlinear transformation η_θ applied to the sum of the embeddings breaks exact factorizability, it could be restored by applying the transformation before summing the embeddings, which in numerical experiments does not affect performance. Regardless, in numerical experiments with two systems of non-interacting molecules (H_2-H_2 and LiH-H_2) we show that even the variant of our ansatz which is not exactly factorizable is size-consistent in practice (Table II). For the system composed of two distant hydrogen molecules, both the combined and individual calculations give nearly exact results, 99.99(1)% and 100.00(1)% of the correlation energy, respectively. In the second test with LiH and H_2, 99.65(2)% and 99.68(2)% of the correlation energy is achieved, respectively, which corresponds to the difference of less than 10% of the overall error of PauliNet with respect to the exact energy. The results furthermore show that optimization of the ansatz for the combined system works similarly well as optimizing the separate instances for the respective subsystems (Fig. 6).

D. Application of different levels of theory to H_2O

The results for the small test systems showed that DNNs can be used to converge to the CBS limit within the mean-

TABLE II. **Sampled energies of the size consistency experiment.**

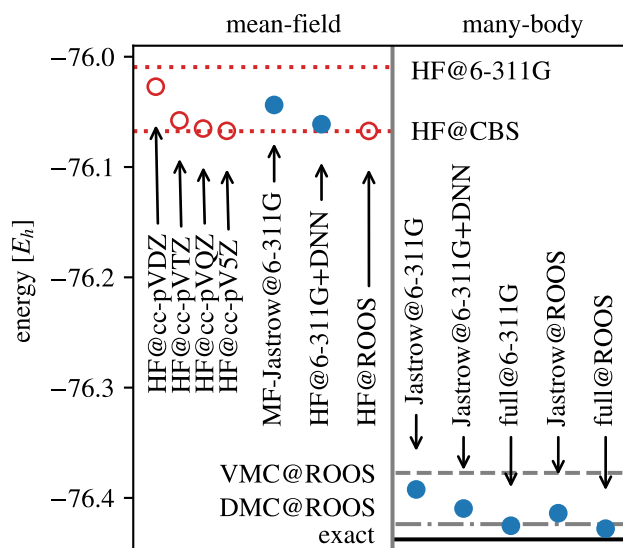| system	combined	individual	exact[21,29]
H_2-H_2	-2.34894(1)	-2.34895(1)	-2.34895
LiH-H_2	-9.24394(7)	-9.24405(7)	-9.24501

FIG. 7. **Single-determinant variants of PauliNet evaluated on the H$_2$O molecule.** All three restricted variants of PauliNet introduced in Methods are compared: the deep orbital correction (6-311G+DNN), the full deep Jastrow factor, and the mean-field (MF) Jastrow factor. Furthermore the energy of the combined ansatz (Jastrow@6-311G+DNN) and single-determinant versions of the full PauliNet (full) are shown. ROOS denotes the Roos-aug-DZ-ANO basis set. VMC and DMC references are taken from Gurtubay and Needs[31]. HF@CBS and exact energy are taken from Rosenberg and Shavitt[33]. The Monte Carlo sampling error is smaller than the size of the markers. The training curves for the ansatzes are shown in Fig. 9.

field theory, and that by adding correlation with a deep Jastrow factor the fixed-node limit can be approached. To investigate how these ansatzes behave for larger systems, we evaluated the respective instances of PauliNet on the water molecule (Fig. 7). These experiments aim at demonstrating that the same ansatzes can be applied to a variety of systems without any modifications and test how much their respective accuracy decreases if the size of the graph-convolutional neural network is kept fixed. For the experiment we chose 4 interactions and a kernel dimension of 256, which is equal to the large models from the H$_4$ and LiH experiments. Due to the convolutional nature of the neural network the number of trainable parameters is mostly independent of the number of electrons, hence it is similar to the previous experiments.

We again start with the mean-field theory and consider the finite-basis-set error. We corrected the HF orbitals in the small 6-311G basis set with the deep orbital correction (Sec. II B), which recovered 90% of the finite-basis-set error. We then estimated how much of the finite-basis-set error amounts to the fixed-node error by applying the mean-field Jastrow factor (Sec. II D), which can recover about half of the finite-basis-set error only. This suggests that upon approaching the finite-basis-set limit the nodal surface is altered significantly.

Next, we investigate single-determinant ansatzes with the full Jastrow factor (Sec. II C). We benchmark the deep Jas-

trow factor with a HF determinant in the Roos augmented double-zeta basis (Roos-aug-DZ-ANO)[34], a basis set that is frequently used for calculations on H$_2$O and gives HF energies at the CBS limit. We compare to VMC and DMC results from the literature, achieving 97.2(1)% of the fixed-node correlation energy and surpassing the accuracy of previous VMC calculations with single-determinant Slater–Jastrow trial wavefunctions by half an order of magnitude (Table III).

In order to study how finite-basis-set errors manifest in the mean-field nodal surface of both the HF and the many-body ansatzes, we computed the energies of the deep Jastrow factor with a HF determinant in a 6-311G basis. The results suggest that finite-basis-set errors of the HF calculations transfers directly to the many-body regime. In particular, the differences between the energies of the mean-field ansatzes match the differences of the respective Slater–Jastrow trial wavefunctions and errors in the energy due to finite-basis-set effects are not altered by the many-body correlation.

We furthermore demonstrate that both methods can be combined, by optimizing a trial wavefunction composed of a deep Jastrow factor and a Slater determinant of orbitals of an imprecise HF baseline that are modified by the orbital correction. The parameters of both Jastrow factor and orbital correction were optimized simultaneously. The HF baseline was computed in the small 6-311G basis set. With this setup we were able to achieve energies close to the fixed-node limit of the optimal mean-field nodal surface. Starting from a minimal baseline we recovered 96.0(2)% of the fixed-node correlation energy with respect to the Roos-aug-DZ-ANO basis.

Finally we show that the full PauliNet ansatz can go beyond the fixed-node approximation and train an instance with the same graph-convolutional architecture as in the previous experiments, but using the full backflow transformation. With this ansatz we obtained a VMC energy of -76.4252(3) and -76.4281(3) for the 6-311G and the Roos-aug-DZ-ANO basis set respectively, amounting to 96.67(8)% and 97.38(8)% of the total correlation energy. This energy is significantly below the single-determinant DMC results (Fig. 7), demonstrating energetically favorable changes in the nodal surface due to the backflow transformation. A comparison to traditional VMC results shows that the single-determinant version of PauliNet strongly improves on single-determinant Slater–Jastrow–backflow (SD-SJB) trial wavefunctions and multi-determinant Slater–Jastrow (MD-SJ) wavefunctions need thousands of determinants to obtain a similar accuracy (Table IV). Here it should be stated that in principle the accuracy of PauliNet can be further improved by increasing the size of the graph-convolutional neural network architecture or introducing multiple determinants. The comparison should therefore not be understood as ultimate, but serves to give an impression of the capabilities of the PauliNet backflow. More exemplary calculations with the full PauliNet ansatz including multi-determinant ansatzes have been carried out previously[7] and a more thorough investigation of the improvements in the nodal-surface as well as a benchmark of the computational complexity will be conducted in future work.

TABLE III. **Benchmarking single-determinant (SD) Slater–Jastrow (SJ) ansatzes on H_2O.**

reference	HF	VMC (SD-SJ)	DMC (SD-SJ)	η_{FN} [%]	basis set
PauliNet	−76.009	−76.3923(7)	–	91.2(2)[a]	6-311G
PauliNet	−76.0612	−76.4096(7)	–	96.0(2)[a]	6-311G+DNN
PauliNet	−76.0672	−76.4139(5)	–	97.2(1)[a]	Roos-aug-DZ-ANO
Clark et al. [30]	–	−76.3938(4)	−76.4236(2)	91.6(1)	Roos-aug-TZ-ANO
Gurtubay and Needs [31]	−76.0672	−76.3773(2)	−76.42376(5)	87.01(6)	Roos-aug-DZ-ANO
Gurtubay et al. [32]	−76.0587	−76.327(1)	−76.42102(4)	73.5(3)	6–311++G(2d,2p)

[a] The fixed-node correlation energy is computed with respect to the reference FN-DMC energy of Gurtubay and Needs [31].

TABLE IV. **Comparison of full PauliNet instance with traditional trial wavefunctions.**

	ansatz	# determinants	VMC
PauliNet	SD-SJB	1	−76.4281(3)
Gurtubay and Needs [31]	SD-SJB	1	−76.4034(2)
Clark et al. [30]	MD-SJ	2316	−76.4259(6)
Clark et al. [30]	MD-SJ	7425	−76.4289(8)

IV. DISCUSSION

We have demonstrated that the choice of the architecture does not introduce fundamental limitations regarding the flexibility of the investigated components of the PauliNet and that a systematic improvement of the accuracy is possible when increasing the number of trainable parameters in a suitable way. For both the deep orbital correction and the deep Jastrow factor, close to exact energies for the corresponding level of theory can be obtained. This highlights the generality and expressiveness of deep QMC—a single ansatz without any problem-specific modifications can be applied to a variety of systems and extended systematically to improve the accuracy without introducing new components to the trial wavefunction architecture. Though the results with the deep orbital correction and the deep Jastrow factor emphasize the potential of the deep QMC approach, the major benefit of deep QMC over FN-DMC calculations remains that it can go beyond the fixed-node approximation by faithfully representing the nodal surface upon introducing many-body correlation at the level of the orbitals. We have outlined this with an exemplary calculation on the water molecule, using a single-determinant instance of the full PauliNet ansatz. The presented analysis paves the way for future investigations on how the full PauliNet ansatz improves the nodes and overcomes the fixed-node limitations.

DATA AVAILABILITY

The data that support the findings of this study are openly available at `http://doi.org/10.6084/m9.figshare.13077158.v3`.

[1] W. M. C. Foulkes, L. Mitas, R. J. Needs, and G. Rajagopal, Rev. Mod. Phys. **73**, 33 (2001).

[2] K. T. Schütt, M. Gastegger, A. Tkatchenko, K.-R. Müller, and R. J. Maurer, Nat. Commun. **10**, 5024 (2019).

[3] M. Gastegger, A. McSloy, M. Luya, K. T. Schütt, and R. J. Maurer, J. Chem. Phys. **153**, 044123 (2020).

[4] G. Carleo and M. Troyer, Science **355**, 602 (2017).

[5] K. Choo, A. Mezzacapo, and G. Carleo, Nature Communications **11**, 2368 (2020).

[6] J. Han, L. Zhang, and W. E, Journal of Computational Physics **399**, 108929 (2019).

[7] J. Hermann, Z. Schätzle, and F. Noé, Nature Chemistry **12**, 891 (2020).

[8] D. Pfau, J. S. Spencer, A. G. D. G. Matthews, and W. M. C. Foulkes, Phys. Rev. Research **2**, 033429 (2020).

[9] C. Adams, G. Carleo, A. Lovato, and N. Rocco, "Variational Monte Carlo calculations of A ≤ 4 nuclei with an artificial neural-network correlator ansatz," (2020), preprint at http://arxiv.org/abs/2007.14282.

[10] J. Han, Y. Li, L. Lin, J. Lu, J. Zhang, and L. Zhang, "Universal approximation of symmetric and anti-symmetric functions," (2019), preprint at http://arxiv.org/abs/1912.01765.

[11] M. Hutter, "On Representing (Anti)Symmetric Functions," (2020).

[12] K. Choo, A. Mezzacapo, and G. Carleo, Nat. Commun. **11**, 2368 (2020).

[13] N. D. Drummond, M. D. Towler, and R. J. Needs, Phys. Rev. B **70**, 235119 (2004).

[14] R. J. Needs, M. D. Towler, N. D. Drummond, and P. L. Ríos, J. Phys.: Condens. Matter **22**, 023201 (2009).

[15] A. Barr, W. Gispen, and A. Lamacraft, in *Proceedings of Machine Learning Research*, Vol. 107 (2020) pp. 635–653.

[16] J. Han, J. Lu, and M. Zhou, "Solving high-dimensional eigenvalue problems using deep neural networks: A diffusion Monte Carlo like approach," (2020), preprint at http://arxiv.org/abs/2002.02600.

[17] P. López Ríos, A. Ma, N. D. Drummond, M. D. Towler, and R. J. Needs, Phys. Rev. E **74**, 066701 (2006).

[18] T. Kato, Communications on Pure and Applied Mathematics **10**, 151 (1957).

[19] J. Hermann, Z. Schätzle, and H. E. Sauceda, "Deepqmc/deepqmc: Deep-QMC 0.3.0," Zenodo (2021).

[20] A. Ma, M. D. Towler, N. D. Drummond, and R. J. Needs, J. Chem. Phys. **122**, 224322 (2005).

[21] W. Kołos and L. Wolniewicz, J. Chem. Phys. **43**, 2429 (1965).

[22] Q. Sun, X. Zhang, S. Banerjee, P. Bao, M. Barbry, N. S. Blunt, N. A. Bogdanov, G. H. Booth, J. Chen, Z.-H. Cui, J. J. Eriksen, Y. Gao, S. Guo, J. Hermann, M. R. Hermes, K. Koh, P. Koval, S. Lehtola, Z. Li, J. Liu,

N. Mardirossian, J. D. McClain, M. Motta, B. Mussard, H. Q. Pham, A. Pulkin, W. Purwanto, P. J. Robinson, E. Ronca, E. R. Sayfutyarova, M. Scheurer, H. F. Schurkus, J. E. T. Smith, C. Sun, S.-N. Sun, S. Upadhyay, L. K. Wagner, X. Wang, A. White, J. D. Whitfield, M. J. Williamson, S. Wouters, J. Yang, J. M. Yu, T. Zhu, T. C. Berkelbach, S. Sharma, A. Y. Sokolov, and G. K.-L. Chan, J. Chem. Phys. **153**, 024109 (2020).

[23] T. Kinoshita, Phys. Rev. **115**, 366 (1959).

[24] B. P. Pritchard, D. Altarawy, B. Didier, T. D. Gibson, and T. L. Windus, J. Chem. Inf. Model. **59**, 4814 (2019).

[25] M. Casalegno, M. Mella, and A. M. Rappe, J. Chem. Phys. **118**, 7193 (2003).

[26] D. Tunega and J. Noga, Theor Chem Acc **100**, 78 (1998).

[27] K. Gasperich, M. Deible, and K. D. Jordan, J. Chem. Phys. **147**, 074106 (2017).

[28] R. Livni, S. Shalev-Shwartz, and O. Shamir, in *Advances in Neural Information Processing Systems*, Vol. 27, edited by Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger (Curran Associates, Inc., 2014) pp. 855–863.

[29] W. Cencek and J. Rychlewski, Chemical Physics Letters **320**, 549 (2000).

[30] B. K. Clark, M. A. Morales, J. McMinis, J. Kim, and G. E. Scuseria, The Journal of Chemical Physics **135**, 244105 (2011).

[31] I. G. Gurtubay and R. J. Needs, J. Chem. Phys. **127**, 124306 (2007).

[32] I. G. Gurtubay, N. D. Drummond, M. D. Towler, and R. J. Needs, J. Chem. Phys. **124**, 024318 (2006).

[33] B. J. Rosenberg and I. Shavitt, The Journal of Chemical Physics **63**, 2162 (1975).

[34] P.-O. Widmark, P.-Å. Malmqvist, and B. O. Roos, Theoret. Chim. Acta **77**, 291 (1990).

[35] P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner, C. Gulcehre, F. Song, A. Ballard, J. Gilmer, G. Dahl, A. Vaswani, K. Allen, C. Nash, V. Langston, C. Dyer, N. Heess, D. Wierstra, P. Kohli, M. Botvinick, O. Vinyals, Y. Li, and R. Pascanu, arXiv:1806.01261 [cs, stat] (2018), arXiv:1806.01261 [cs, stat].

[36] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, arXiv:1704.01212 [cs] (2017), arXiv:1704.01212 [cs].

[37] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller, J. Chem. Phys. **148**, 241722 (2018).

Appendix A

1. Graph-convolutional neural-network architecture

At the core of the PauliNet architecture a graph-convolutional neural network architecture generates a permutation-equivariant latent-space many-body representation of a given electron configuration. The following section gives a short introduction to the network architecture, discussing both the general concept and the particular application in the context of PauliNet.

Graph neural networks are constructed to represent functions on graph domains[35] and have become increasing popular for modeling chemical systems, as they can be designed to comply with symmetries of molecules[36]. The graph-convolutional neural network of PauliNet is a modification of SchNet[37], an architecture developed to predict molecular properties from atom positions upon being trained in a supervised setting, that is to say by repeated exposure to known pairs of input and output data. In SchNet a trainable embedding is assigned to each atom, which serves to give an abstract representation of the atomic properties in a high-dimensional feature space and is successively updated to encode information about the atomic environment. The updates are implemented as convolutions over the graph of atomic distances, which makes the architecture invariant to translation, rotation and equivariant with respect to the exchange of identical atoms. The graph convolutions furthermore implement parameter sharing across the edges, such that the number of network parameters does not depend on the number of interacting entities, hence is constant with system size. The final features are then used to predict the molecular properties respectively.

In quantum chemistry we consider modeling electrons and nuclei. At this level a molecule can be represented as a complete graph, where nodes correspond to electrons and nuclei and the distances of each pair of particles is assigned to the edge between their respective nodes. The graph-convolutional neural network at the core of PauliNet acts on this graph representation of the system. Similar to the SchNet implementation a representation in an abstract feature space is assigned to each node by introducing electronic embeddings $\mathbf{X}_{\boldsymbol{\theta}, s_i}$ and nuclear embeddings $\mathbf{Y}_{\boldsymbol{\theta}, I}$ respectively. The embeddings are trainable arrays that are initialized randomly. As same-spin electrons are indistinguishable they share representations and get initialized with a copy of the same electronic embedding,

$$\mathbf{x}_i^{(0)} := \mathbf{X}_{\boldsymbol{\theta}, s_i} \tag{A1}$$

The electronic embeddings constitute the latent-space representation, which serves to obtain Jastrow factor and backflow transformation in the later process of evaluating the trial wavefunction. In order to encode positional information of each electron with respect to the nuclei as well as electronic many-body correlation into the latent-space representation, the electronic embeddings are updated in an interaction process. Information is transmitted along the edges of the graph, by exchanging messages that take the distances to the nuclei and

the other electrons into account,

$$\mathbf{z}_i^{(n,n)} := \sum_I \mathbf{w}_{\boldsymbol{\theta}}^{(n,n)}\big(\mathbf{e}(|\mathbf{r}_i - \mathbf{R}_I|)\big) \odot \mathbf{Y}_{\boldsymbol{\theta},I}$$
$$\mathbf{z}_i^{(n,\pm)} := \sum_{j \neq i}^{\pm} \mathbf{w}_{\boldsymbol{\theta}}^{(n,\pm)}\big(\mathbf{e}(|\mathbf{r}_i - \mathbf{r}_j|)\big) \odot \mathbf{h}_{\boldsymbol{\theta}}^{(n)}\big(\mathbf{x}_j^{(n)}\big) \tag{A2}$$

Here the functions $\mathbf{w}_{\boldsymbol{\theta}}$ and $\mathbf{h}_{\boldsymbol{\theta}}$ are implemented by fully-connected neural networks, $\mathbf{e}$ represents an expansion of the distances in a basis of Gaussian functions and $\odot$ indicates element-wise multiplication. For each two interacting particles the filter generating function $\mathbf{w}_{\boldsymbol{\theta}}$ generates a mask that is applied to their respective embeddings. Thereby the filter-generating function moderates interactions based on the distance of the particles. By summing up the messages of identical particles their overall contribution is invariant under the exchange of these identical particles. The transformation $\mathbf{h}_{\boldsymbol{\theta}}$ serves to introduce additional flexibility to the architecture, by separating the latent-space representation from the interaction space. The superscripts of the neural networks indicate that different functions are applied at each subsequent interaction and the filter-generating functions for the interactions with spin-up electrons, spin-down electrons and nuclei are different. The distance expansion $\mathbf{e}$ is truncated with an envelope that ensures it to be cusp-less, that is to say that all Gaussian features have a vanishing derivative at zero distance, and imposes a long-range cutoff. The final step in the interaction process is to update to the electronic embeddings

$$\mathbf{x}_i^{(n+1)} := \mathbf{x}_i^{(n)} + \sum_{\pm} \mathbf{g}_{\boldsymbol{\theta}}^{(n,\pm)}\big(\mathbf{z}_i^{(n,\pm)}\big) + \mathbf{g}_{\boldsymbol{\theta}}^{(n,n)}\big(\mathbf{z}_i^{(n,n)}\big) \tag{A3}$$

Therefore the messages are transformed from the interaction space to the embedding space and added to the original embedding. The transformation $\mathbf{g}_{\boldsymbol{\theta}}$ is again implemented by fully-connected neural networks. This interaction process is repeated L times, to successively encode increasingly complex many-body information. The continuous-filter convolutions over the molecular graph and the initialization of electrons with identical embeddings make the architecture equivariant with respect to the exchange of same-spin electrons,

$$\mathscr{P}_{ij}\mathbf{x}_i(\mathbf{r}) = \mathbf{x}_j(\mathscr{P}_{ij}\mathbf{r}) \tag{A4}$$

Overall this gives a latent-space representation that can efficiently encode electronic many-body effects while intrinsically fulfilling the desired permutation equivariance. Information about the hyperparameters of all of the components is collected in Table V.

2. Single-particle variant of the graph-convolutional neural network

A single-particle variant of the graph-convolutional neural-network architecture can be obtained by considering only interactions along edges between electrons and nuclei. That is to say the overall architecture of the network remains identical but the electronic updates $\mathbf{z}_i^{(n,\pm)}$ in (A3) are removed. Given that the convolution over the electronic distances is the only interaction between electrons, the final embeddings do not contain any many-body correlation.

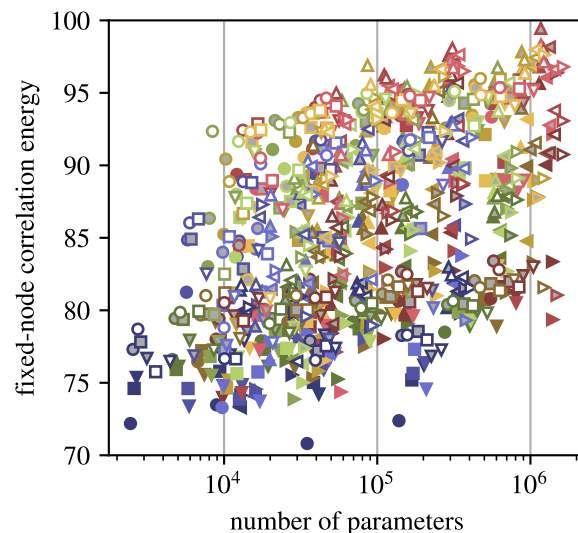| criterion	hyperparameter	values
tone	# layers $\eta_{\boldsymbol{\theta}}$	1, 2, 3
color	# interactions L	1, 2, 3, 4
marker	(# layers $\mathbf{h}_{\boldsymbol{\theta}}$ & $\mathbf{g}_{\boldsymbol{\theta}}$, # layers $\mathbf{w}_{\boldsymbol{\theta}}$)	(1,1), (1,2), (1,3), (2,1), (2,2), (2,3)
fill	# distance features	8, 16, 32
implicit	dim $\mathbf{x}_i$ & kernel	4, 16, 64, 256

FIG. 8. **Hyperparameter scan of the H$_4$ square.** The figure maps models with different combinations of hyperparameters in the space of accuracy with respect to the number of total trainable parameters. The exact set of hyperparameters of each model can be decoded via the accompanying legend. In order to reduce the degrees of freedom the embedding and kernel dimensions were joined and varied together. The models with increasing embedding and kernel dimensions can be distinguished by their number of total parameters, hence no explicit criterion is introduced. The energy of the models is estimated from the final steps of the optimization procedure. The fixed-node correlation energy is obtained with respect to FN-DMC results[27], where a single-determinant trial wavefunction with a basis consisting of s functions from the cc-pV5Z basis set and the p and d functions from the cc-pVTZ basis set was used. Increasing the number of trainable parameters the fixed-node limit can be approached.

TABLE V. **Components of the deep Jastrow factor.**

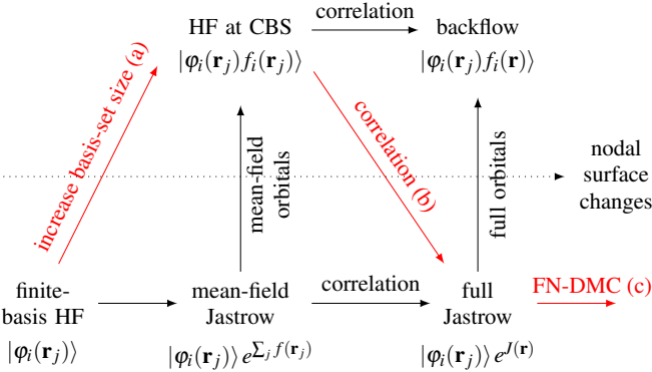| component	type	hyperparameter
$\mathbf{X}_{\theta,s_i}$ – electronic embedding	trainable array	embedding dimension
$\mathbf{Y}_{\theta,I}$ – nuclear embeddings	trainable array	kernel dimension
$\mathbf{e}$ – distance expansion	fixed function	# distance features
$\mathbf{w}_\theta$ – filter generating function	DNN	(# distance features $\rightarrow$ kernel dimension), depth
$\mathbf{h}_\theta$ – transformation embedding to kernel space	DNN	(embedding dimension $\rightarrow$ kernel dimension), depth
$\mathbf{g}_\theta$ – transformation kernel to embedding space	DNN	(kernel dimension $\rightarrow$ embedding dimension), depth
η_θ – Jastrow network	DNN	(embedding dimension $\rightarrow$ 1), depth
full architecture	—	# interactions L
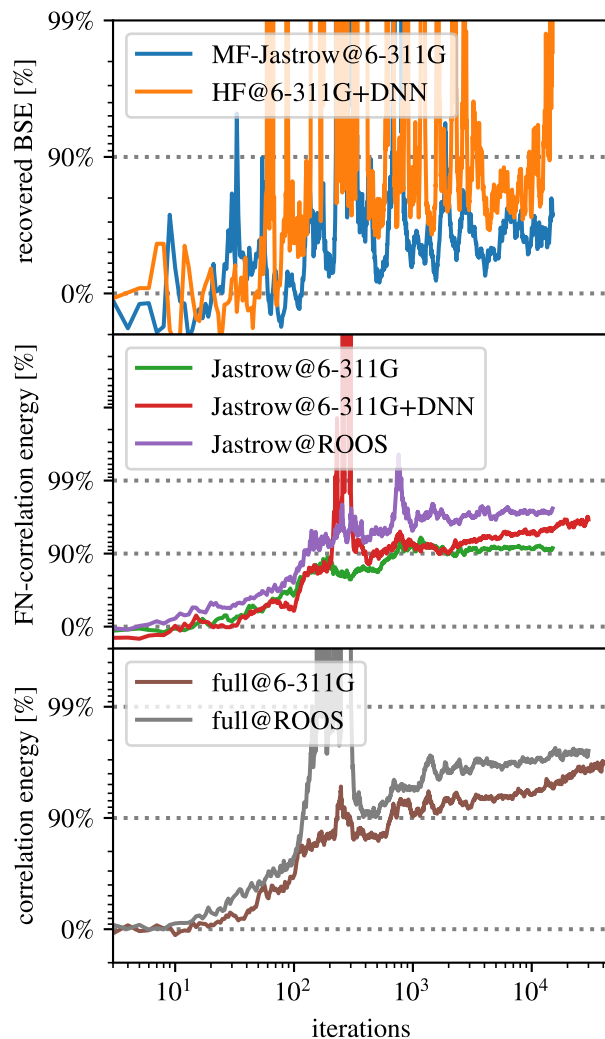
FIG. 9. **Training curve for the H_2O experiment.** Exponential walking averages of the energy along the optimization process are shown for the ansatzes from Fig. 7. Upper panel depicts the percentage of the recovered finite-basis-set error (BSE) for the mean-field ansatzes, given by the energy difference of the HF@6-311G baseline from the complete-basis-set limit estimate. The center panel gives the fixed-node correlation energy of the Slater–Jastrow type versions of PauliNet with respect to the DMC reference in the Roos-aug-DZ-ANO basis set[31]. Lower panel shows the total correlation energy for the full PauliNet ansatzes.
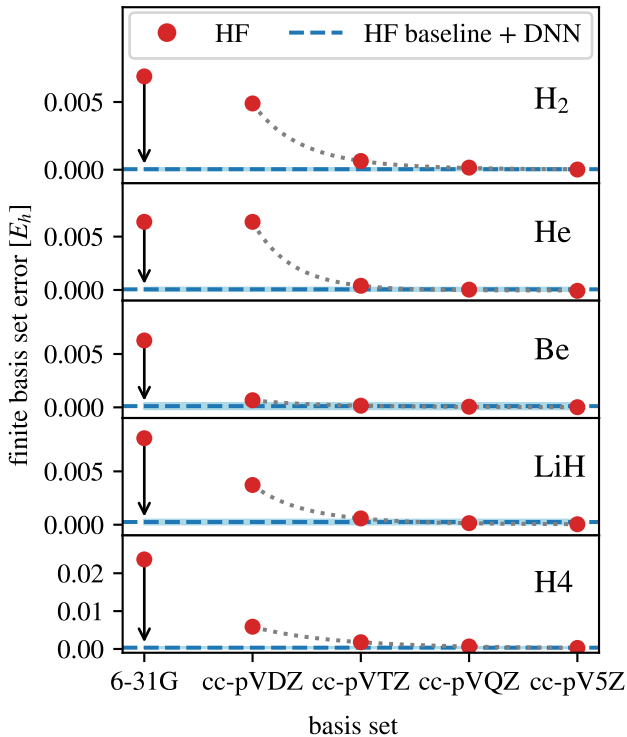
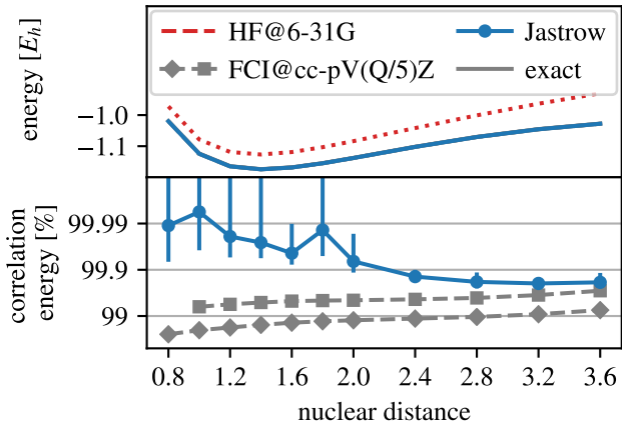TABLE VI. **Hyperparameters used in numerical calculations.**

Hyperparameter	Value
One-electron basis	6-31G
Dimension of $\mathbf{e}$ (# distance features)	16
Dimension of $\mathbf{x}_i$ (embedding dimension)	128
Number of interaction layers L	4
Number of layers in η_θ	3
Number of layers in $\mathbf{w}_\theta$	1
Number of layers in $\mathbf{h}_\theta$	2
Number of layers in $\mathbf{g}_\theta$	2
Batch size	2000
Number of walkers	2000
Number of training steps	H_4: 5000 H_2: 10000 He: 10000 Be: 10000 LiH: 10000 H_2O: see Fig. 9
Optimizer	AdamW
Learning rate scheduler	CyclicLR
Minimum/maximum learning rate	0.0001/0.01
Clipping window q	5
Epoch size	100
Number of decorrelation sampling steps	4
Target acceptance	57%

TABLE VII. **Geometries of test systems.**

Molecule	Atom	position [Å]
LiH	Li	(0.000, 0.000, 0.000)
	H	(1.595, 0.000, 0.000)
H_4 square	H	(-0.635, -0.635, 0.000)
	H	(-0.635, 0.635, 0.000)
	H	(0.635, -0.635, 0.000)
	H	(0.635, 0.635, 0.000)
H_4 deformed	H	(-0.900, -0.635, 0.000)
	H	(-0.900, 0.635, 0.000)
	H	(0.900, -0.635, 0.000)
	H	(0.900, 0.635, 0.000)
H_2O	O	(0.00000, 0.00000, 0.00000)
	H	(0.75695, 0.58588, 0.00000)
	H	(-0.75695, 0.58588, 0.00000)

HF at CBS $\xrightarrow{\text{correlation}}$ backflow

$|\varphi_i(\mathbf{r}_j)f_i(\mathbf{r}_j)\rangle$ $|\varphi_i(\mathbf{r}_j)f_i(\mathbf{r})\rangle$

increase basis-set size (a)

mean-field orbitals

correlation (b)

full orbitals

nodal surface changes

finite-basis HF $\xrightarrow{}$ mean-field Jastrow $\xrightarrow{\text{correlation}}$ full Jastrow $\xrightarrow{}$ FN-DMC (c)

$|\varphi_i(\mathbf{r}_j)\rangle$ $|\varphi_i(\mathbf{r}_j)\rangle e^{\sum_j f(\mathbf{r}_j)}$ $|\varphi_i(\mathbf{r}_j)\rangle e^{J(\mathbf{r})}$

Legend: ● HF   - - - HF baseline + DNN

Y-axis: finite basis set error [E_h]

Panels (top to bottom): H$_2$, He, Be, LiH, H4

X-axis (basis set): 6-31G, cc-pVDZ, cc-pVTZ, cc-pVQZ, cc-pV5Z

mean-field   many-body

HF@6-311G

HF@CBS

HF@cc-pVDZ
HF@cc-pVTZ
HF@cc-pVQZ
HF@cc-pV5Z
MF-Jastrow@6-311G
HF@6-311G+DNN
HF@ROOS

Jastrow@6-311G
Jastrow@6-311G+DNN
full@6-311G
Jastrow@ROOS
full@ROOS

VMC@ROOS
DMC@ROOS
exact

criterion	hyperparameter	values
tone	# layers $\eta_{\boldsymbol{\theta}}$	1, 2, 3
color	# interactions L	1, 2, 3, 4
marker	(# layers $\mathbf{h}_{\boldsymbol{\theta}}$ & $\mathbf{g}_{\boldsymbol{\theta}}$, # layers $\mathbf{w}_{\boldsymbol{\theta}}$)	(1,1), (1,2), (1,3), (2,1), (2,2), (2,3)
fill	# distance features	8, 16, 32
implicit	dim $\mathbf{x}_i$ & kernel	4, 16, 64, 256