

Deep neural network solution of the electronic Schrödinger equation

Jan Hermann^{1,2,*}, Zeno Schätzle¹, and Frank Noé^{1,3,4,*}

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*

Abstract The electronic Schrödinger equation can only be solved analytically for the hydrogen atom, and the numerically exact full configuration-interaction method is exponentially expensive in the number of electrons. Quantum Monte Carlo methods are a possible way out: they scale well for large molecules, can be parallelized, and their accuracy has, as yet, been only limited by the flexibility of the wave function ansatz used. Here we propose PauliNet, a deep-learning wave function ansatz that achieves nearly exact solutions of the electronic Schrödinger equation for molecules with up to 30 electrons. PauliNet has a multireference Hartree–Fock solution built in as a baseline, incorporates the physics of valid wave functions, and is trained using variational quantum Monte Carlo (VMC). PauliNet outperforms previous state-of-the-art VMC ansatzes for atoms, diatomic molecules and a strongly correlated linear H_{10} , and matches the accuracy of highly specialized quantum chemistry methods on the transition-state energy of cyclobutadiene, while being computationally efficient.

1 Introduction

A solution of the time-independent electronic Schrödinger equation of a given atomic system provides, in principle, full access to its chemical properties. This equation can be solved analytically only for an isolated hydrogen atom, but solid-state physics and quantum chemistry have been remarkably successful in developing numerical approximation methods (Piela 2014). For small molecules containing up to a few tens of electrons, methods based on the configuration interaction and the closely related coupled cluster approaches or the multideterminant quantum Monte Carlo (QMC) can reach impressive accuracy of up to six significant digits in the total electronic energy (Morales et al. 2012).

Unfortunately, the computational cost of such high-accuracy methods increases with a high power of the number of electrons, N , making them impractical for most relevant molecules or materials. Computationally less demanding methods, such as the density functional theory (DFT), can scale to larger molecules, but at the price of limited accuracy. Fundamentally, high-accuracy methods scale unfavorably with N because the dimension of the solution space of the Schrödinger equation for a many-body problem scales exponentially. A distinct but related problem that appears in some approaches is the so-called *sign problem* originating from the Pauli exclusion principle (Troyer & Wieser 2005). The trade-off between accuracy and computational cost is apparent when considering that most quantum chemistry methods represent electronic wave functions by linear combinations of Slater determinants. A Slater matrix is constructed by selecting N out of $M > N$ molecular orbitals, and assigning N electrons to them, resulting in a combinatorial growth of all possible matrices with system size.

The Slater determinants have different roles in different quantum chemistry methods (Fig. 1). In the configuration interaction and coupled cluster approaches, the electronic problem is solved entirely in the basis of the determinants (second quantization), and as such their number in typical applications is the largest,

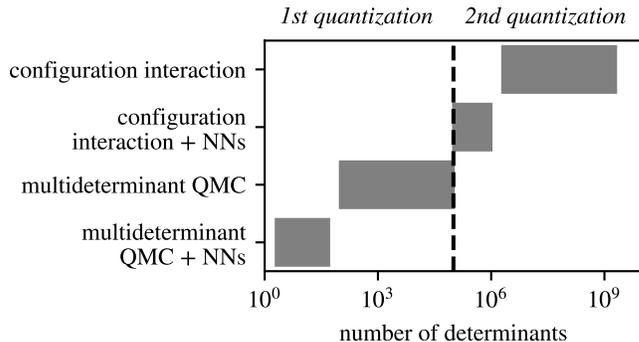


Figure 1 | Combinatorial explosion of the number of Slater determinants in quantum chemistry. Both configuration interaction and multideterminant quantum Monte Carlo (QMC) approaches suffer from the rapid scaling of the number of Slater determinants with system size, which in both cases can be reduced with neural networks (NNs). The multideterminant QMC combined with neural networks is the approach developed in this work. The plot shows typical numbers of Slater determinants used by high-accuracy quantum chemistry methods in state-of-the-art calculations on atomic systems with at most a few tens of electrons.

as the determinant expansion must recover all many-body interactions that are missing in individual determinants (Shavitt & Bartlett 2009). Stochastic methods that sample over vast determinant spaces have been developed (Booth et al. 2009; Thom 2010), but the underlying scaling trap persists nevertheless. Recent work by Choo et al. (2020) suggests that the number of required determinants can be reduced with the use of neural networks, but whether this would also reduce the scaling issue has yet to be demonstrated.

Conventional QMC methods (Austin et al. 2012; Foulkes et al. 2001; Needs et al. 2010) solve the electronic problem in real space (first quantization), and treat a large portion of the correlation in the electronic motion explicitly, which greatly reduces the number of required determinants. Standard QMC variants are still practical for systems with hundreds of electrons, such as supramolec-

*Emails: jan.hermann@fu-berlin.de, frank.noe@fu-berlin.de

ular complexes (Ambrosetti et al. 2014) and molecular crystals (Zen et al. 2018). Even though only a single determinant can be typically used for such large systems, QMC still outperforms other electronic-structure methods applicable to such systems, such as DFT or the random-phase approximation. However, for small systems, where second-quantized approaches are applicable, standard QMC methods need to use at least hundreds of determinants to be competitive, and this number increases rapidly with N . The large number of determinants is necessary to accurately represent the *nodal surface* of the wave function, which is otherwise difficult to improve with standard real-space QMC methods. A key development enabling the progress in the present work is the real-space *backflow* technique (López Ríos et al. 2006). The idea of the backflow is to transform the electrons into pseudoparticles, the position of each of which depends on the positions of all the electrons, and this many-body mixing then leads to an improved nodal surface (Feynman & Cohen 1956; Schmidt et al. 1979). While the traditional backflow does not reach the accuracy of the large determinant expansions and does not generalize well to larger systems, Luo & Clark (2019) recently showed that representing the backflow with a neural network is a powerful generalization.

Machine learning has had significant impact on quantum chemistry, especially in the case of supervised learning and prediction of electronic energies (Bartók et al. 2017; Behler & Parrinello 2007; Chmiela et al. 2017; Faber et al. 2018; Rupp et al. 2012; Schütt et al. 2018; Smith 2017; Welborn et al. 2018), electron densities (Grisafi et al. 2019), and molecular orbitals (Schütt et al. 2019). This approach entirely avoids the solution of the Schrödinger equation, at the price of requiring datasets of preexisting solutions, obtained for instance by density functional theory or the coupled cluster method.

In contrast, the direct representation of correlated wave functions with neural networks and their unsupervised training via the variational principle, first proposed by Carleo & Troyer (2017) for discrete spin lattice systems, is an ab-initio approach that requires no preexisting data and has no fundamental limits to its accuracy. It is motivated by the fact that neural networks are universal function approximators, and could therefore provide more efficient means for approximating the exponentially scaling complexity of many-body quantum systems. The initial attempts on lattice systems were later generalized to bosons in real space (Ruggeri et al. 2018; Saito 2018), and even electrons in real space (Han et al. 2019), but the latter approach does not use a wave function ansatz in the form of a Slater determinant, and perhaps for that reason does not reach the accuracy of the baseline Hartree–Fock method for some systems.

In this work, we develop PauliNet, a deep learning QMC approach that replaces existing ad-hoc functional forms used in the standard Jastrow factor and backflow transformation with more powerful deep neural network representations. Besides the sheer gain in expressive power, our neural network architecture is specifically designed to encode the physics of valid wave functions and incorporates the multireference Hartree–Fock method as a baseline. These physically motivated choices are essential to obtain a method that is not only highly accurate, but also converges robustly, while maintaining computational efficiency. Using several test systems, we demonstrate that our neural network ansatz significantly outperforms the accuracy of state-of-the-art

wave function ansatzes using a similar number of determinants. Thanks to the trainable backflow ansatz, high accuracy can be obtained with orders of magnitude fewer determinants compared to traditional QMC methods. Our method has the asymptotic scaling of N^4 , and we expect that it will be feasible to apply it to much larger systems than is currently possible with existing high-accuracy methods. We demonstrate this with an accurate calculation of the transition-state energy of the 28-electron cyclobutadiene molecule, which was previously achievable only with highly specialized methods.

The parallel work of Pfau et al. (2020) follows the same basic idea as ours, but differs in one important aspect. Their architecture does not encode any physical knowledge about wave functions besides the essential antisymmetry, which is compensated by a much larger number of optimized parameters. This difference likely leads to the higher computational cost per iteration. In addition, their architecture is trained substantially longer and as a consequence reaches higher accuracy for some systems.

2 Results

2.1 Deep neural network electronic wave function ansatz

At the core of our deep learning approach to the electronic Schrödinger equation is a wave function ansatz, dubbed PauliNet, which incorporates both the well-established essential physics of electronic wave functions—Slater determinants, multideterminant expansion, Jastrow factor, backflow transformation, and cusp conditions—as well as deep neural networks capable of encoding the complex features of the electronic motion in heterogeneous molecular systems. Our proposed trial wave function, $\psi_{\theta}(\mathbf{r})$, $\mathbf{r} = (\mathbf{r}_1, \dots, \mathbf{r}_N)$, is of the multideterminant Slater–Jastrow–backflow type (Brown et al. 2007), where both the Jastrow factor, J , and the backflow, \mathbf{f} , are represented by deep neural networks (DNNs) with trainable parameters θ (Fig. 2),

$$\psi_{\theta}(\mathbf{r}) = e^{\gamma(\mathbf{r})+J_{\theta}(\mathbf{r})} \sum_p c_p \det[\tilde{\varphi}_{\theta,\mu p}^{\uparrow}(\mathbf{r})] \det[\tilde{\varphi}_{\theta,\mu p}^{\downarrow}(\mathbf{r})] \quad (1)$$

$$\tilde{\varphi}_{\mu i}(\mathbf{r}) = \varphi_{\mu}(\mathbf{r}_i) f_{\theta,\mu i}(\mathbf{r})$$

While the expressiveness of PauliNet is contained in the Jastrow factor and backflow DNNs, the physics is encoded by the determinant form, the one-electron molecular orbitals, φ_{μ} , and the electronic cusps, γ , in the following way.

Every valid electronic wave function must be antisymmetric with respect to the exchange of same-spin electrons,

$$\psi(\dots, \mathbf{r}_i, \dots, \mathbf{r}_j, \dots) = -\psi(\dots, \mathbf{r}_j, \dots, \mathbf{r}_i, \dots) \quad (2)$$

As is common in quantum chemistry, we enforce antisymmetry via matrix determinants, as determinants change sign upon exchanging any two rows or columns.

To ensure a good starting point for the variational optimization problem, we exploit the approximate Hartree–Fock (HF) method. Specifically, we use a multireference HF calculation with a small complete active space, and select the most dominant determinants and their orbitals based on the magnitude of their linear coefficient. The HF-optimized one-electron molecular orbitals, $\varphi_{\mu}(\mathbf{r})$, are then used as an input to PauliNet, and are modified during training only by the backflow transformation.

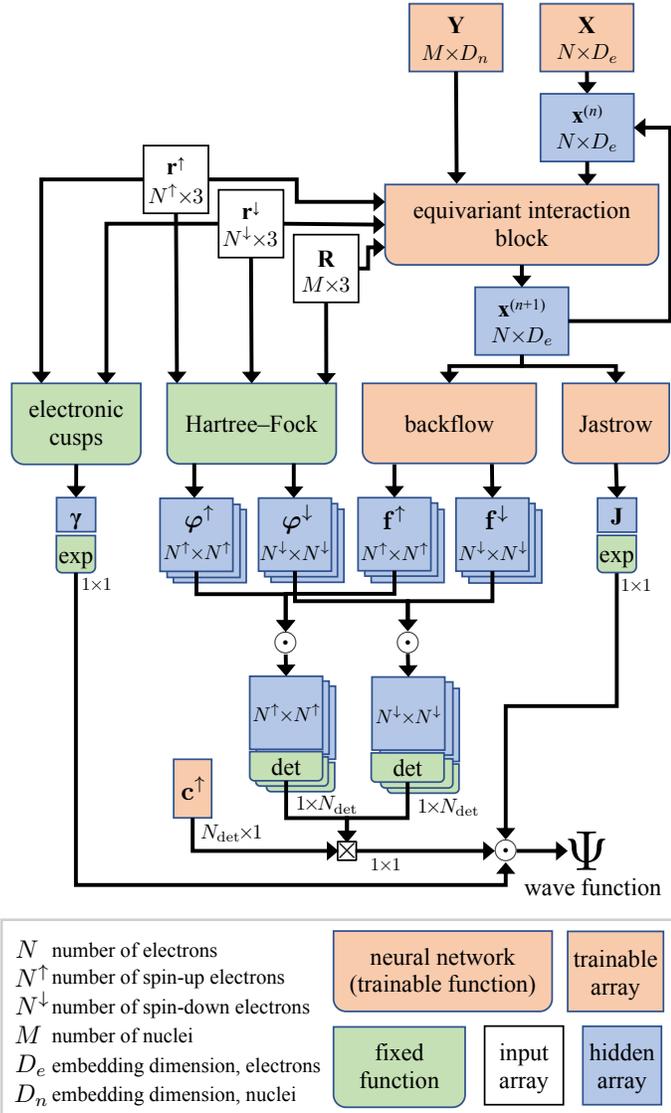


Figure 2 | Architecture of the newly developed PauliNet wave function ansatz. The information flows from the input electron and nuclear coordinates, \mathbf{r} and \mathbf{R} , to the output wave function value, Ψ . Modeling the wave function via Jastrow and backflow functions is common in quantum Monte Carlo, but here these functions are learned with deep neural networks.

Any ground-state electronic wave function obeys exact asymptotic behavior defined by the *cusp conditions* as electrons approach each other and the nuclei (Kato 1957). We chose to build the cusp conditions directly into the PauliNet functional form as this makes the training more efficient as well as stable by removing divergences from the local electronic energy. We incorporate the nuclear cusps by modifying the molecular orbitals using the technique from Ma et al. (2005), and the electronic cusps by the fixed cusp function, $\gamma(\mathbf{r})$. We ensure that the trainable Jastrow factor and backflow DNNs are cuspless, so as to maintain the enforced cusp behavior (see Methods for details).

2.2 Robust deep Jastrow factor and backflow

PauliNet differs from conventional QMC ansatzes by representing the Jastrow factor and backflow functions with specialized DNNs. To retain the antisymmetry of the wave function, as enforced by the Slater determinants, the Jastrow factor and backflow DNNs are constructed to be invariant and equivariant, respectively, with respect to the exchange of same-spin electrons, \mathcal{P}_{ij} ,

$$J(\mathcal{P}_{ij}\mathbf{r}) = J(\mathbf{r}), \quad \mathcal{P}_{ij}f_{\mu i}(\mathbf{r}) = f_{\mu j}(\mathcal{P}_{ij}\mathbf{r}) \quad (3)$$

The Jastrow factor is a nonnegative totally symmetric function, which can encode complex electron correlations into the wave function, but cannot modify the nodal surface inherited from the determinant expansion.

We found that attempting to express the standard backflow form of coupled electron coordinates with DNNs leads to a difficult optimization problem. Instead, the PauliNet backflow has the form of multiplying the bare one-electron molecular orbitals with many-electron equivariant functions, \mathbf{f} (Eq. 1). In combination with just a few determinants, this presents a powerful representation of the electronic nodal surface.

The requirements of invariance and equivariance with respect to permutation of particles, and the fact that particle interactions are a function of their distances, are closely related constructing DNNs that learn potential energy functions. PauliNet uses an adapted form of one such DNN architecture, called SchNet (Schütt et al. 2018). SchNet is a graph DNN that represents each particle with a vector in a high-dimensional abstract feature space, \mathbf{x}_i , which is iteratively refined by interactions with other particles through real-space trainable convolutions, χ_θ , which encode the inter-particle distances and are invariant with respect to particle exchange,

$$\mathbf{x}_i^{(n+1)} := \mathbf{x}_i^{(n)} + \chi_\theta(\{\mathbf{x}_j^{(n)}, \{|\mathbf{r}_j - \mathbf{r}_k|\}\}) \quad (4)$$

The SchNet architecture and its modifications for PauliNet are described in detail in Methods. After a fixed number of iterations, the final electron representations, $\mathbf{x}_i^{(L)}$, which now encode complex many-body electron correlations, are used as an input to two trainable functions, η_θ , κ_θ , which return the Jastrow factor and backflow, respectively.

$$J := \eta_\theta(\sum_i \mathbf{x}_i^{(L)}), \quad \mathbf{f}_i := \kappa_\theta(\mathbf{x}_i^{(L)}), \quad (5)$$

Since the feature vectors, $\mathbf{x}_i^{(n)}$, are equivariant with respect to electron exchange at each iteration, so are the backflow vectors, \mathbf{f}_i . As a result, the Slater determinants in PauliNet produce an antisymmetric wave function. Furthermore, J is by construction invariant with respect to exchanges of electrons and therefore a symmetric function that maintains this antisymmetry.

2.3 Approaching exact solutions with few determinants

We train PauliNet via the variational principle, minimizing the total electronic energy (variational QMC). The training data are electron configurations that are generated on-the-fly by sampling the electron distribution, $|\psi|^2$ (Methods for details). We first investigate the same systems that were used to test DeepWF (Han et al. 2019), in particular the hydrogen molecule (H_2), lithium

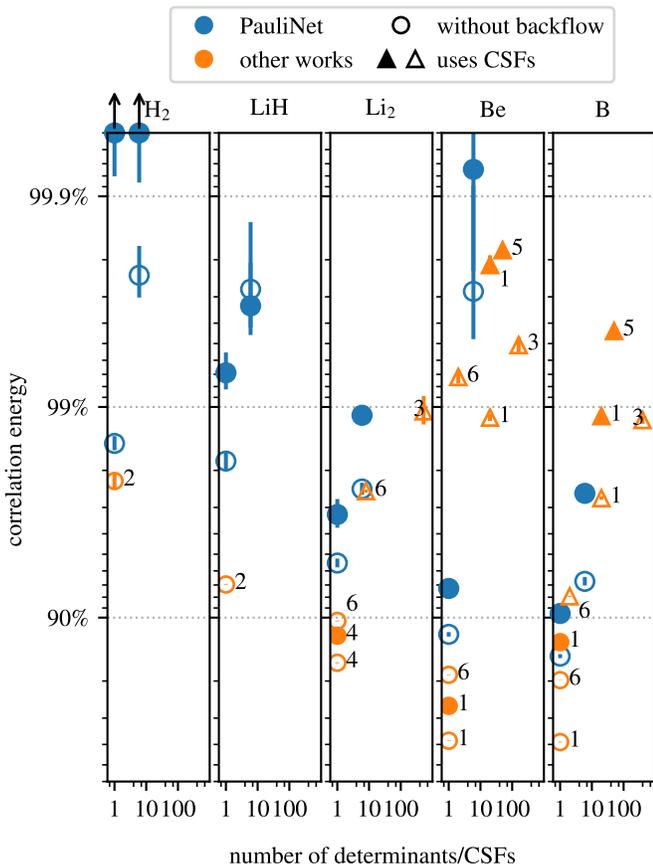


Figure 3 | Performance of PauliNet with one and six determinants on atoms and diatomic molecules. PauliNet recovers 97% to 99.9% of correlation energy with one to two orders of magnitude less determinants than standard variational ansatzes. Four variants of PauliNet are shown, single- and multideterminant as well as with and without backflow. The reference results are taken from (1) Brown et al. (2007), (2) Casalegno et al. (2003), (3) Morales et al. (2012), (4) Ríos & Conduit (2019), (5) Seth et al. (2011), and (6) Toulouse & Umrigar (2008). Each configuration state function (CSF) corresponds to a few to a few dozen determinants, depending on a system and the particular method. The numerical data can be found in Table A1.

hydride (LiH), beryllium (Be), boron (B), and the linear hydrogen chain H_{10} . For the mono- and diatomic systems, PauliNet recovers between 97% and 99.9% of the electron correlation energy (Fig. 3) after training for tens of minutes to a few hours on a single GTX 1080 Ti graphics processing unit (GPU), in all cases far beyond the accuracy of DeepWF. We compare these results to the standard single-determinant (SD) and multideterminant (MD) variational Monte Carlo (VMC) methods with and without backflow. In all cases, PauliNet with 6 determinants is substantially better than all single-determinant and few-determinant ansatzes, and is only surpassed by trial wave functions with tens or hundreds of configuration state functions (CSFs), corresponding to hundreds to thousands of determinants.

Figure 4 highlights two crucial aspects of our method. First, the error in the correlation energy decreases monotonously as the training progresses from the initial HF baseline level to the final reported values. The learning curves are not yet fully plateaued in most cases, demonstrating the high expressiveness of our neural

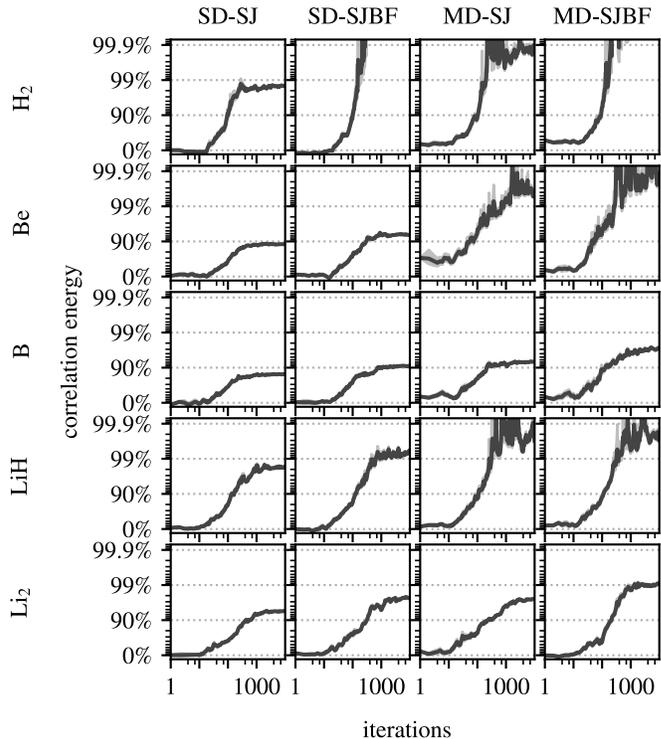


Figure 4 | Roles of backflow and multiple determinants in training of PauliNet ansatz. The four variants are combinations of a single or multiple Slater determinants (SD/MD), a Jastrow factor (SJ), and a backflow (BF). Both the backflow and the use of a few determinants is crucial for reaching high accuracy. Exponential moving average is applied to the energy at each iteration.

ansatz, and indicating that even higher accuracy could be achieved with more computational resources. Second, we compare our full ansatz to variants using only a single determinant and variants without backflow, and find that both these components are important for refining the nodal surface of the HF baseline, and thus reaching high accuracy. The fact that only a few determinants are sufficient to substantially reduce the correlation energy error compared to the single-determinant case indicates that deep learning can be an efficient tool to reduce the large number of determinants sampled in other VMC approaches that directly operate on determinants of fixed orbitals (Booth et al. 2009; Choo et al. 2020). By having a powerful backflow transformation, each additional determinant substantially increases the flexibility of the ansatz.

We further analyze the scaling of PauliNet with the number of determinants and with system size on a set of four homonuclear diatomic molecules from Li_2 to C_2 (Fig. 5). PauliNet reaches high accuracy quickly with increasing but small numbers of determinants. In the regime of a few to a few dozen of determinants, PauliNet surpasses existing variational results and in most cases reaches the accuracy of the corresponding diffusion Monte Carlo (DMC) results. We note that DMC can be implemented straightforwardly for our approach, and is expected to lead to further increase its accuracy.

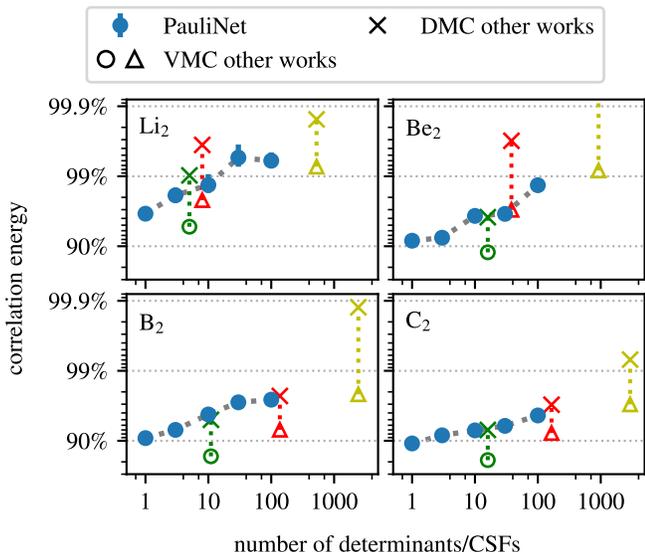


Figure 5 | Convergence of PauliNet with the number of determinants. The accuracy of PauliNet increases with increasing number of determinants as in standard methods, but reaches the accuracy of diffusion Monte Carlo already at the variational level. Comparison is shown to VMC (circle/triangle) and corresponding DMC (cross) results taken from Filippi & Umrigar (1996) (green), Toulouse & Umrigar (2008) (red), and Morales et al. (2012) (yellow). Either the number of determinants (circle) or of CSFs (triangle) is plotted.

2.4 Capturing strong correlation

Unlike the atoms and diatomics, the linear hydrogen chain H_{10} exhibits *strong correlation*, which describes a situation where the single-determinant or even few-determinant description of the HF method is qualitatively insufficient, and the correlation energy constitutes a significant part of the electronic energy (Motta et al. 2017). For H_{10} , we recover 98.41(8) % and 98.4(3) % of the correlation energy in the equilibrium and stretched geometry, respectively, using 16 determinants (Fig. 6). The results are only slightly worse using a single determinant (98.10(9) %/97.5(4) %), but significantly worse when the trainable backflow is also switched off (93.7(2) %/82(2) %). Compared to standard VMC ansatzes that were specifically adjusted for this particular application (Motta et al. 2017), we reach significantly higher accuracy at equilibrium and comparable accuracy at the dissociated limit.

We make three observations based on these results. First, even though the system size and complexity of H_{10} is significantly higher compared to the systems in the previous section, we achieve the same level of accuracy while using the same form of the ansatz. Second, there is no decrease in accuracy in the stretched geometry due to the increased strongly correlated character. Third, adding multiple determinants to the ansatz recovers only a fraction of the correlation energy compared to the backflow transformation, highlighting the central role of the trainable backflow in PauliNet.

2.5 Straightforward generalization to larger molecules

The previous two sections demonstrated the performance of PauliNet on relatively small benchmark systems, for which essentially exact results are already available from well-established

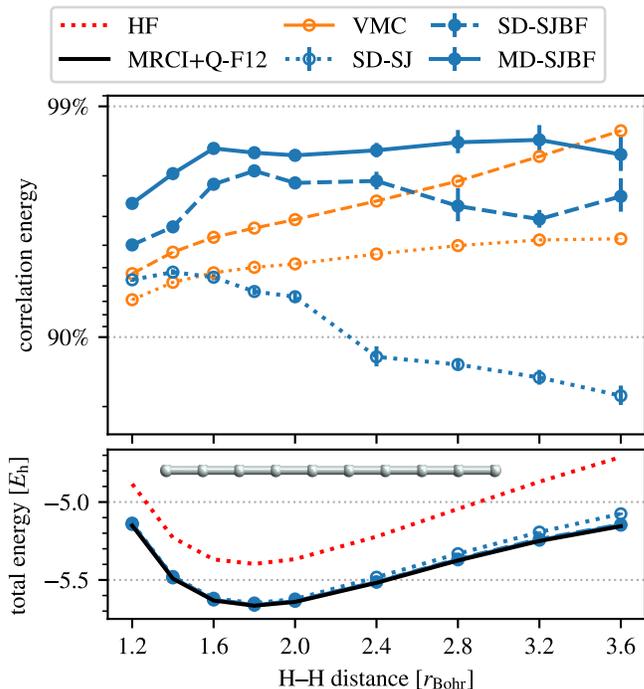


Figure 6 | PauliNet captures strong correlation in H_{10} along the dissociation curve. PauliNet results (blue) with single determinant (SD) or 16 determinants (MD) and with or without backflow (BF) are shown. The backflow plays a much larger role than multiple determinants. PauliNet outperforms highly specialized VMC ansatzes (yellow) of the single-determinant (dotted) and multideterminant geminal (dashed) form by Motta et al. (Motta et al. 2017). The correlation energy is calculated with respect to multireference configuration interaction (MRCI) results, also by Motta et al.

methods. In this section, we show that the same PauliNet ansatz scales straightforwardly to larger molecules with complex electronic structure, for which only highly specialized derivatives of standard quantum chemistry methods could deliver satisfactory results as yet. For this purpose, we chose the automerization of cyclobutadiene (Fig. 7a, 28 electrons), a chemical process that has received considerable attention from both experiment and theory (Lyakh et al. 2012). The experimental estimates of the energy barrier range between 1.6 and 10 kcal/mol, while the standard CCSD(T) method predicts 18 kcal/mol, a two-fold overestimation. The best computational estimates are available from various flavors of the multireference coupled cluster (MR-CC) theory and fall between 7 and 11 kcal/mol, without a decisive answer as to which of the variants is closer to the ground truth.

Using the PauliNet ansatz with 10 determinants and the same hyperparameters as used for the much smaller systems, we obtain all-electron variational energies for the energy minimum and transition states of cyclobutadiene, and thus for the energy barrier. Since the energy barrier is only 0.01 % of the total energy, we use a modified optimization protocol to stabilize the training of the neural network with respect to the inherent stochasticity, in which 10 independent copies are optimized simultaneously, and periodically synchronized such that the five copies with higher energies are discarded, and the rest duplicated.

With this modification, the total energies converge smoothly (Fig. 7b) and by running two independent optimizations with

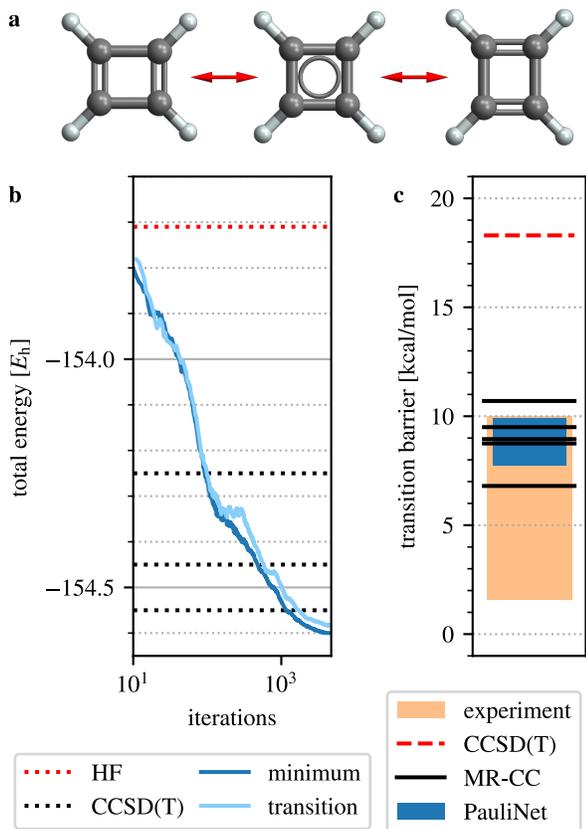


Figure 7 | Calculating the transition barrier of cyclobutadiene automerization. (a) Cyclobutadiene automerization. The transition state has a highly multireferential character. (b) Convergence of the total energy of the energy minimum and transition state with training. Absolute energies of the energy minimum from HF at the complete basis set limit and from CCSD(T) with the cc-pVnZ basis set ($n = D, T, Q$) are shown. (c) Energy barrier obtained by sampling the trained PauliNet wave functions, compared to references taken from Lyakh et al. (2012). Results from five variants of multireference coupled cluster (MR-CC) theory are shown (top to bottom): MR-DI-EOMCCSD, RM RCCSD(T), Mk-MRCCSD(T), MRCISD+Q, and BW-MRCCSD(T).

the synchronization period of 250 and 375 iterations, we obtain the estimates of the energy barrier of 9.9 ± 0.6 and 7.7 ± 0.6 kcal/mol, respectively. The range larger than the respective statistical sampling errors suggests a remaining degree of stochasticity in the optimization, but nevertheless both results are well within the range spanned by the MR-CC methods. We note that compared to the MR-CC methods, many of which require use of system- and state-specific wave function ansatzes, the PauliNet ansatz is constructed in essentially the same way for all the systems studied in this work. The computational cost of the cyclobutadiene optimization is 50 s per iteration on a single GTX 1080 Ti GPU for each of the synchronized optimizations.

3 Discussion

We have designed PauliNet, a DNN representation of electronic wave functions in real space and shown that it can outperform state-of-the-art variational quantum chemistry methods that do not use large determinant expansions. In contrast, our approach

requires only a few determinants, and as a result we anticipate that its computational cost scales asymptotically as N^4 (N^3 for a determinant evaluation, and additional N for the evaluation of the kinetic energy), subject to additional technical details (Foulkes et al. 2001). PauliNet is thus a candidate for a quantum chemistry method that can scale to much larger systems with high accuracy.

Compared to standard functional forms used in QMC, the use of DNNs has several advantages. First, the much higher flexibility of DNNs allows a variational approach to reach or exceed the accuracy of diffusion QMC, which aids the calculation of accurate derived electronic properties beyond the electronic energy. Second, besides encoding more complex many-body correlations between electrons, DNNs have an essentially unlimited flexibility in the spatial degrees of freedom, circumventing the curse of incomplete basis sets of quantum chemistry, which can be removed only with DMC when using standard techniques. Third, the rapid oscillations of the local energy close to heavy nuclei in standard QMC mandate the use of pseudopotentials for heavier elements such as transition metals. The flexibility of DNNs could sidestep this necessity by smoothing out such oscillations.

In classical quantum chemistry methods, strong correlation is usually treated by using large multideterminant expansions, which are computationally demanding and introduce the problem of selecting the proper subset of determinants. Treating strong correlation on the level of Jastrow factors traditionally requires construction of specialized many-body forms (Gasperich et al. 2017; Neuscamman et al. 2012). In contrast, we show that DNNs are capable of learning strong correlation between electrons without any specialized adaptation. Convergence to high accuracy can be achieved with only a few determinants, changing the problem from searching or sampling over exponentially many determinants to letting a deep neural network search over exponentially many functions. Although it is unclear whether this is advantageous in a strict mathematical sense, this is precisely the task which deep neural networks have been demonstrated to be strong at in a variety of real-world applications. Complementary approaches that use variational QMC in a second-quantized form of the electronic problem have also been proposed (Choo et al. 2020). This class of methods has the advantage of eliminating much of the complexity of electronic wave functions (such as the antisymmetry or cusp conditions) from the machine-learning part of the problem, but needs to cope with the ubiquitous limitations of single-particle basis sets.

Already a brief comparison of our approach with that of Pfau et al. (2020) hints at potential improvements of both architectures. The combination of architectural design and optimization methods used in FermiNet with the built-in physical constraints of PauliNet appears to be a promising venue for computationally affordable, scalable, yet highly accurate black-box methods for quantum chemistry. We hope that the introduction of neural networks into the field of electronic QMC opens the possibility to utilize the striking advances in deep learning from the last decade in a new field.

4 Methods

Ansatz optimization We optimize the PauliNet ansatz individually for each atomic system in an unsupervised fashion using the

variational principle for the total electronic energy,

$$E_0 = \min_{\psi} E[\psi] \leq \min_{\theta} E[\psi_{\theta}], \quad (6)$$

$$E[\psi] = \int d\mathbf{r} \psi(\mathbf{r}) \hat{H} \psi(\mathbf{r})$$

Following the standard QMC technique, the energy integral is evaluated as an expected value of the *local energy*, $E_{\text{loc}}[\psi](\mathbf{r}) = \hat{H}\psi(\mathbf{r})/\psi(\mathbf{r})$, over the probability distribution $|\psi^2(\mathbf{r})|$,

$$E[\psi] = \mathbb{E}_{\mathbf{r} \sim |\psi|^2} [E_{\text{loc}}[\psi](\mathbf{r})] \quad (7)$$

We generate training data for PauliNet on-the-fly by periodically alternating training on one hand and sampling electron positions with a standard Langevin Monte Carlo approach on the other (Umrigar et al. 1993). Each sampled electron configuration is used only once in an optimization run. We use a simplified version of the method by Umrigar et al. (1993), in which the radial step proposal is replaced with clipping the step length such that the step size is always shorter than the distance to the nearest nucleus, so the nucleus can never be “overshot”. The initial electron positions for the Markov chain are sampled from Gaussian distributions around the nuclei such that the effective atomic Mulliken charges obtained from the HF method are respected.

To optimize the parameters θ in the Jastrow and backflow neural networks, we use the weighted Adam optimizer (Kingma & Ba 2017; Loshchilov & Hutter 2019) together with the total energy used directly as the loss function. To calculate the stochastic gradient of the loss function over a batch of samples, we use a gradient formula that takes advantage of the fact that the Hamiltonian operator is Hermitian (Ceperley et al. 1977),

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathbf{r} \sim |\psi|^2} [E_{\text{loc}}[\psi_{\theta}](\mathbf{r})]$$

$$\nabla_{\theta} \mathcal{L}(\theta) = 2 \mathbb{E}_{\mathbf{r} \sim |\psi|^2} [(E_{\text{loc}}[\psi_{\theta}](\mathbf{r}) - \mathcal{L}(\theta)) \nabla_{\theta} \ln |\psi_{\theta}|] \quad (8)$$

This expression for the gradient requires calculating only second derivatives of the wave function (for the Laplace operator), whereas direct differentiation would require third derivatives (derivative of the Laplace operator). We smoothly clip the local energy of each sample by a logarithmically growing clipping function outside the window defined as 5 times the mean deviance from the median local energy in a given batch. The learning rate is controlled by a cyclic scheduling policy (Smith 2017).

Cusp conditions Eq. (1) ensures the nuclear cusp conditions via the molecular orbitals $\varphi_{\mu}(\mathbf{r}_i)$. We achieve this by modifying the molecular orbitals using the technique from Ma et al. (2005) with one simplification—we optimize the orbital values at atomic nuclei, $\mathbf{r}_i = \mathbf{R}_I$, via the energy variational principle, rather than fitting them against reference values. The electronic cusp conditions are enforced by $\gamma(\mathbf{r})$,

$$\gamma(\mathbf{r}) := \sum_{i < j} -\frac{c_{ij}}{1 + |\mathbf{r}_i - \mathbf{r}_j|}, \quad (9)$$

where c_{ij} is either $\frac{1}{2}$ or $\frac{1}{4}$ depending on the spins of the two electrons. To preserve the cusp conditions built into φ_{μ} and γ , the Jastrow factor and backflow DNNs must be cusplless,

$$\nabla_{\mathbf{r}_i} J(\mathbf{r}) \Big|_{\mathbf{r}_i = \{\mathbf{r}_k, \mathbf{R}_a\}} = 0, \quad \nabla_{\mathbf{r}_i} f_{\mu i}(\mathbf{r}) \Big|_{\mathbf{r}_i = \{\mathbf{r}_k, \mathbf{R}_a\}} = 0 \quad (10)$$

These conditions are ensured by constructing the DNNs appropriately, as detailed below.

PauliNet extension of SchNet SchNet is an instance of the class of graph convolutional neural networks, and was designed to model the molecular energy as a function of just the nuclear charges and coordinates (Schütt et al. 2018). In PauliNet, we use SchNet to represent electrons in molecular environments by implementing the iteration rule in Eq. (4),

$$\mathbf{z}_i^{(n,\pm)} := \sum_{j \neq i}^{\pm} \mathbf{w}_{\theta}^{(n,\pm)}(\mathbf{e}(|\mathbf{r}_i - \mathbf{r}_j|)) \odot \mathbf{h}_{\theta}^{(n)}(\mathbf{x}_j^{(n)})$$

$$\mathbf{z}_i^{(n,n)} := \sum_I \mathbf{w}_{\theta}^{(n,n)}(\mathbf{e}(|\mathbf{r}_i - \mathbf{R}_I|)) \odot \mathbf{Y}_{\theta,I} \quad (11)$$

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

where “ \odot ” denotes element-wise multiplication, $\mathbf{w}_{\theta}^{(n)}$, $\mathbf{h}_{\theta}^{(n)}$, and $\mathbf{g}_{\theta}^{(n)}$ are trainable functions represented by ordinary fully-connected DNNs, and \mathbf{e} is a radial basis function that featurizes the interatomic distances. The modifications of the original SchNet are as follows.

- (i) Since the wave function is a function of electron coordinates, the iterated feature vectors $\mathbf{x}_i^{(n)}$ represent electrons, not atoms.
- (ii) The messages $\mathbf{z}_i(n)$ received by the electron feature vectors at each iteration are split into three channels, corresponding to same-spin electrons (+), opposite-spin electrons (−), and the nuclei (n). This builds more flexibility into the architecture, and is motivated by the fact that electrons and nuclei are particles of an entirely different type.
- (iii) Each channel has a separate receiving function \mathbf{g}_{θ} , again increasing flexibility without substantially increasing the number of parameters.
- (iv) Each nucleus is represented by a trainable embedding $\mathbf{Y}_{\theta,I}$, which is shared across all iterations and not iteratively updated. In VMC, the wave function is always optimized for a given fixed geometry of the nuclei, so the nuclear embeddings can be assumed to already represent each nucleus with its (fixed) atomic environment, hence the absence of need for their iterative refinement.
- (v) The distance features \mathbf{e} are constructed to be cusplless, as detailed below.

We use a distance featurization inspired by the PhysNet architecture (Unke & Meuwly 2019), with a modified envelope that forces all the Gaussian features and their derivatives to zero at zero distance,

$$e_k(r) := r^2 e^{-r - (r - \mu_k)^2 / \sigma_k^2}, \quad (12)$$

$$\mu_k := r_c q_k^2 \quad (13)$$

$$\sigma_k := \frac{1}{7}(1 + r_c q_k) \quad (14)$$

where q_k equidistantly spans the interval (0, 1) and r_c is a cutoff parameter.

Computational details All reported methods were implemented with Pytorch (Paszke et al. 2019) in the open-source DeepQMC package, which is available on Zenodo (Hermann et al. 2020c) and developed on Github at <https://github.com/>

Table 1 | Hyperparameters used in numerical calculations.

Hyperparameter	Value
One-electron basis	6-311G
Dimension of \mathbf{e}	32
Dimension of \mathbf{x}_i	128
Dimension of \mathbf{z}_i	64
Number of interaction layers L	3
Number of layers in η_θ	3
Number of layers in κ_θ	3
Number of layers in \mathbf{w}_θ	2
Number of layers in \mathbf{h}_θ	1
Number of layers in \mathbf{g}_θ	1
Batch size	10 000
Number of walkers	2000
Number of training steps	10 000
Optimizer	AdamW
Minimum/maximum learning rate	0.0001/0.01
Cyclic frequency	500
Clipping window q	5
Resampling frequency	10
Number of decorrelation sampling steps	10
Target acceptance	57%

deepqmc/deepqmc. The linear coefficients of the HF orbitals φ_μ as well as of the determinants in a multideterminant expansion were calculated with PySCF (Sun et al. 2018) using the “6-311G” Gaussian basis set. The plain fully-connected DNNs that represent the trainable functions in our architecture were chosen such the total number of trainable parameters is around 7×10^4 (see Table 1). The raw data (Hermann et al. 2020a) and scripts (Hermann et al. 2020b) are available on Figshare and Zenodo, respectively.

References

- Ambrosetti, A., D. Alfè, R. A. DiStasio Jr. & A. Tkatchenko (2014). “Hard Numbers for Large Molecules: Toward Exact Energetics for Supramolecular Systems.” *J. Phys. Chem. Lett.* **5**, 849. DOI: 10/f5tt25.
- Austin, B. M., D. Y. Zubarev & W. A. Lester (2012). “Quantum Monte Carlo and Related Approaches.” *Chem. Rev.* **112**, 263. DOI: 10/fzmcf8.
- Bartók, A. P., S. De, C. Poelking, N. Bernstein, J. R. Kermode, G. Csányi & M. Ceriotti (2017). “Machine Learning Unifies the Modeling of Materials and Molecules.” *Sci. Adv.* **3**, e1701816. DOI: 10/gcqj8b.
- Behler, J. & M. Parrinello (2007). “Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces.” *Phys. Rev. Lett.* **98**, 146401. DOI: 10/c7kbsq.
- Booth, G. H., A. J. W. Thom & A. Alavi (2009). “Fermion Monte Carlo without Fixed Nodes: A Game of Life, Death, and Annihilation in Slater Determinant Space.” *J. Chem. Phys.* **131**, 054106. DOI: 10/d3828h.
- Brown, M. D., J. R. Trail, P. López Ríos & R. J. Needs (2007). “Energies of the First Row Atoms from Quantum Monte Carlo.” *J. Chem. Phys.* **126**, 224110. DOI: 10/frnc4n.
- Carleo, G. & M. Troyer (2017). “Solving the Quantum Many-Body Problem with Artificial Neural Networks.” *Science* **355**, 602. DOI: 10/f9qd53.
- Casalegno, M., M. Mella & A. M. Rappe (2003). “Computing Accurate Forces in Quantum Monte Carlo Using Pulay’s Corrections and Energy Minimization.” *J. Chem. Phys.* **118**, 7193. DOI: 10/d4gsp4.
- Ceperley, D., G. V. Chester & M. H. Kalos (1977). “Monte Carlo Simulation of a Many-Fermion Study.” *Phys. Rev. B* **16**, 3081. DOI: 10/bj4jwm.
- Chmiela, S., A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt & K.-R. Müller (2017). “Machine Learning of Accurate Energy-Conserving Molecular Force Fields.” *Sci. Adv.* **3**, e1603015. DOI: 10/f97dkz.
- Choo, K., A. Mezzacapo & G. Carleo (2020). “Fermionic Neural-Network States for Ab-Initio Electronic Structure.” *Nat. Commun.* **11**, 2368. DOI: 10/ghmzr5.
- Faber, F. A., A. S. Christensen, B. Huang & O. A. von Lilienfeld (2018). “Alchemical and Structural Distribution Based Representation for Universal Quantum Machine Learning.” *J. Chem. Phys.* **148**, 241717. DOI: 10/gds5t4.
- Feynman, R. P. & M. Cohen (1956). “Energy Spectrum of the Excitations in Liquid Helium.” *Phys. Rev.* **102**, 1189. DOI: 10/c77rcn.
- Filippi, C. & C. J. Umrigar (1996). “Multiconfiguration Wave Functions for Quantum Monte Carlo Calculations of First-row Diatomic Molecules.” *J. Chem. Phys.* **105**, 213. DOI: 10/d9k3kd.
- Foulkes, W. M. C., L. Mitás, R. J. Needs & G. Rajagopal (2001). “Quantum Monte Carlo Simulations of Solids.” *Rev. Mod. Phys.* **73**, 33. DOI: 10/dmgh76.
- Gasperich, K., M. Deible & K. D. Jordan (2017). “H₄: A Model System for Assessing the Performance of Diffusion Monte Carlo Calculations Using a Single Slater Determinant Trial Function.” *J. Chem. Phys.* **147**, 074106. DOI: 10/gbsx59.
- Grisafi, A., A. Fabrizio, B. Meyer, D. M. Wilkins, C. Corminboeuf & M. Ceriotti (2019). “Transferable Machine-Learning Model of the Electron Density.” *ACS Cent. Sci.* **5**, 57. DOI: 10/gfs72w.
- Han, J., L. Zhang & W. E (2019). “Solving Many-Electron Schrödinger Equation Using Deep Neural Networks.” *J. Comput. Phys.* **399**, 108929. DOI: 10/gmg56w.
- Hermann, J., Z. Schätzle & F. Noé (2020a). Data in HDF5 format. DOI: 10.6084/m9.figshare.12720569.v2.
- (2020b). Code as Git repository. DOI: 10.6084/m9.figshare.12720833.v1.
- (2020c). *DeepQMC*. Code as Git repository. DOI: 10.5281/zenodo.3960827.
- Kato, T. (1957). “On the Eigenfunctions of Many-Particle Systems in Quantum Mechanics.” *Commun. Pure Appl. Math.* **10**, 151. DOI: 10/crp3p9.
- Kingma, D. P. & J. Ba (2017). “Adam: A Method for Stochastic Optimization.” URL: <http://arxiv.org/abs/1412.6980>.
- López Ríos, P., A. Ma, N. D. Drummond, M. D. Towler & R. J. Needs (2006). “Inhomogeneous Backflow Transformations in Quantum Monte Carlo Calculations.” *Phys. Rev. E* **74**, 066701. DOI: 10/b8m23v.
- Loshchilov, I. & F. Hutter (2019). “Decoupled Weight Decay Regularization.” In *International Conference on Learning Representations (ICLR 2019)* (OpenReview). URL: <https://openreview.net/forum?id=Bkg6RiCqY7>.
- Luo, D. & B. K. Clark (2019). “Backflow Transformations via Neural Networks for Quantum Many-Body Wave Functions.” *Phys. Rev. Lett.* **122**, 226401. DOI: 10/ghp332.
- Lyakh, D. I., M. Musiał, V. F. Lotrich & R. J. Bartlett (2012). “Multireference Nature of Chemistry: The Coupled-Cluster View.” *Chem. Rev.* **112**, 182. DOI: 10/fzxmzv.
- Ma, A., M. D. Towler, N. D. Drummond & R. J. Needs (2005). “Scheme for Adding Electron–Nucleus Cusps to Gaussian Orbitals.” *J. Chem. Phys.* **122**, 224322. DOI: 10/ffqp2b.
- Morales, M. A., J. McMinis, B. K. Clark, J. Kim & G. E. Scuseria (2012). “Multi-determinant Wave Functions in Quantum Monte Carlo.” *J. Chem. Theory Comput.* **8**, 2181. DOI: 10/f34p94.
- Motta, M. et al. (2017). “Towards the Solution of the Many-Electron Problem in Real Materials: Equation of State of the Hydrogen Chain with State-of-the-Art Many-Body Methods.” *Phys. Rev. X* **7**, 031059. DOI: 10/gb3br3.
- Needs, R. J., M. D. Towler, N. D. Drummond & P. L. Ríos (2010). “Continuum Variational and Diffusion Quantum Monte Carlo Calculations.” *J. Phys. Condens. Matter* **22**, 023201. DOI: 10/bc5ckv.
- Neuscamman, E., C. J. Umrigar & G. K.-L. Chan (2012). “Optimizing Large Parameter Sets in Variational Quantum Monte Carlo.” *Phys. Rev. B* **85**, 045103. DOI: 10/fzmv6k.
- Paszke, A. et al. (2019). “PyTorch: An Imperative Style, High-Performance Deep Learning Library.” In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)* (Curran Associates), 8026–8037. URL: <http://papers.nips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library>.
- Pfau, D., J. S. Spencer, A. G. D. G. Matthews & W. M. C. Foulkes (2020). “Ab Initio Solution of the Many-Electron Schrödinger Equation with Deep Neural Networks.” *Phys. Rev. Research* **2**, 033429. DOI: 10/ghmf5c.
- Piela, L. (2014). *Ideas of Quantum Chemistry* 2nd ed. (Elsevier). DOI: 10.1016/C2017-0-03971-X.
- Ríos, P. L. & G. J. Conduit (2019). “Tail-Regression Estimator for Heavy-Tailed Distributions of Known Tail Indices and Its Application to Continuum Quantum Monte Carlo Data.” *Phys. Rev. E* **99**, 063312. DOI: 10/gm7wsh.
- Ruggeri, M., S. Moroni & M. Holzmann (2018). “Nonlinear Network Description for Many-Body Quantum Systems in Continuous Space.” *Phys. Rev. Lett.* **120**, 205302. DOI: 10/gdgd64n.

- Rupp, M., A. Tkatchenko, K.-R. Müller & O. A. von Lilienfeld (2012). “Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning.” *Phys. Rev. Lett.* **108**, 058301. DOI: 10/fzzvdb.
- Saito, H. (2018). “Method to Solve Quantum Few-Body Problems with Artificial Neural Networks.” *J. Phys. Soc. Jpn.* **87**, 074002. DOI: 10/ghmz5x.
- Schmidt, P. C., A. Weiss & T. P. Das (1979). “Effect of Crystal Fields and Self-Consistency on Dipole and Quadrupole Polarizabilities of Closed-Shell Ions.” *Phys. Rev. B* **19**, 5525. DOI: 10/dccct3c.
- Schütt, K. T., H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko & K.-R. Müller (2018). “SchNet – A Deep Learning Architecture for Molecules and Materials.” *J. Chem. Phys.* **148**, 241722. DOI: 10/gdtbj6.
- Schütt, K. T., M. Gastegger, A. Tkatchenko, K.-R. Müller & R. J. Maurer (2019). “Unifying Machine Learning and Quantum Chemistry with a Deep Neural Network for Molecular Wavefunctions.” *Nat. Commun.* **10**, 5024. DOI: 10/ggdh49.
- Seth, P., P. L. Ríos & R. J. Needs (2011). “Quantum Monte Carlo Study of the First-Row Atoms and Ions.” *J. Chem. Phys.* **134**, 084105. DOI: 10/d4wp82.
- Shavitt, I. & R. J. Bartlett (2009). *Many-Body Methods in Chemistry and Physics* (Cambridge Univ. Press). DOI: 10.1017/CB09780511596834.
- Smith, L. N. (2017). “Cyclical Learning Rates for Training Neural Networks.” In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)* (IEEE), 464–472. DOI: 10/gfpb29.
- Sun, Q. et al. (2018). “PySCF: The Python-based Simulations of Chemistry Framework.” *WIREs Comput. Mol. Sci.* **8**, e1340. DOI: 10/gfz3nm.
- Thom, A. J. W. (2010). “Stochastic Coupled Cluster Theory.” *Phys. Rev. Lett.* **105**, 263004. DOI: 10/cpktqb.
- Toulouse, J. & C. J. Umrigar (2008). “Full Optimization of Jastrow–Slater Wave Functions with Application to the First-Row Atoms and Homonuclear Diatomic Molecules.” *J. Chem. Phys.* **128**, 174101. DOI: 10/c7355p.
- Troyer, M. & U.-J. Wiese (2005). “Computational Complexity and Fundamental Limitations to Fermionic Quantum Monte Carlo Simulations.” *Phys. Rev. Lett.* **94**, 170201. DOI: 10/b34365.
- Umrigar, C. J., M. P. Nightingale & K. J. Runge (1993). “A Diffusion Monte Carlo Algorithm with Very Small Time-step Errors.” *J. Chem. Phys.* **99**, 2865. DOI: 10/fr75mw.
- Unke, O. T. & M. Meuwly (2019). “PhysNet: A Neural Network for Predicting Energies, Forces, Dipole Moments, and Partial Charges.” *J. Chem. Theory Comput.* **15**, 3678. DOI: 10/gf9x9m.
- Welborn, M., L. Cheng & T. F. Miller (2018). “Transferability in Machine Learning for Electronic Structure via the Molecular Orbital Basis.” *J. Chem. Theory Comput.* **14**, 4772. DOI: 10/gd3wz3.
- Zen, A., J. G. Brandenburg, J. Klimeš, A. Tkatchenko, D. Alfè & A. Michaelides (2018). “Fast and Accurate Quantum Monte Carlo for Molecular Crystals.” *Proc. Natl. Acad. Sci.* **115**, 1724. DOI: 10/gc5mxt.

Acknowledgments

We are grateful to C. Clementi (Rice, FU Berlin), J. Eisert (FU Berlin), G. Scuseria (Rice), J. Neugebauer (MPIE), and H. Wu (Tongji) for inspiring discussions. Funding is acknowledged from the European Commission (ERC CoG 772230 “Scale-Cell”), Deutsche Forschungsgemeinschaft (CRC1114/A04, GRK2433 DAEDALUS/P04), the MATH+ Berlin Mathematics research center (AA1x6, EF1x2). J. H. thanks K.-R. Müller for support and acknowledges funding from TU Berlin.

Table A1 | Variational correlation energy (%) of five test systems obtained with four types of trial wave functions.

ref.	Slater–Jastrow			Slater–Jastrow with backflow		
	SD ^a	MD ^b		SD	MD	
H ₂ PauliNet (Casalegno et al. 2003)	98.5(1)	99.8(1)	6 D	100.02(3)	99.98(3)	6 D
	97.8(2)	–	–	–	–	–
Be PauliNet (Toulouse & Umrigar 2008) (Brown et al. 2007) (Morales et al. 2012) (Seth et al. 2011)	88.0(3)	99.7(2)	6 D	92.7(4)	99.9(1)	6 D
	81.31(5)	99.28(5)	2 CSF	–	–	–
	61.6(1)	98.88(4)	20 CSF	73.8(1)	99.79(2)	20 CSF
	–	99.49(4)	160 CSF	–	–	–
B PauliNet (Toulouse & Umrigar 2008) (Brown et al. 2007) (Morales et al. 2012) (Seth et al. 2011)	84.7(3)	93.3(3)	6 D	90.5(4)	97.4(2)	6 D
	80.19(4)	92.08(4)	2 CSF	–	–	–
	61.0(2)	97.29(5)	20 CSF	86.92(8)	98.89(3)	20 CSF
	–	98.85(5)	396 CSF	–	–	–
LiH PauliNet (Casalegno et al. 2003)	98.1(2)	99.7(1)	6 D	99.3(1)	99.7(1)	6 D
	93.04(2)	–	–	–	–	–
Li ₂ PauliNet (Toulouse & Umrigar 2008) (Morales et al. 2012) (López Ríos et al. 2006)	94.5(2)	97.6(2)	6 D	96.8(5)	98.9(1)	6 D
	89.63(4)	97.49(4)	8 CSF	–	–	–
	–	99.0(2)	526 CSF	–	–	–
	83.62(8)	–	–	87.81(8)	–	–

^aSingle-determinant ansatz.

^bMulti-determinant ansatz. “D” denotes a determinant, “CSF” denotes a configuration state function.