

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 describes fundamental properties of molecules and materials, but can only be solved analytically for the hydrogen atom. The numerically exact full configuration-interaction method is exponentially expensive in the number of electrons. Quantum Monte Carlo is a possible way out: it scales well to large molecules, can be parallelized, and its accuracy has, as yet, only been limited by the flexibility of the used wave function ansatz. Here we propose PauliNet, a deep-learning wave function ansatz that achieves nearly exact solutions of the electronic Schrödinger equation. 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 comparable state-of-the-art VMC ansatzes for atoms, diatomic molecules and a strongly-correlated hydrogen chain by a margin and is yet computationally efficient. We anticipate that thanks to the favourable scaling with system size, this method may become a new leading method for highly accurate electronic-structure calculations on medium-sized molecular systems.

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 (Piel, 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 as N^7 or worse with the number of electrons, N , making them unworkable for most molecules or materials of practical interest. Computationally less demanding methods, such as the density functional theory, can scale to larger molecules, but at the price of limited accuracy. Fundamentally, high-accuracy methods scale unfavorably with N because the many-body electronic problem is *non-polynomial hard* due to the Pauli exclusion principle (Troyer and Wiese, 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 matrix 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 (Figure 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, as the determinant expansion must recover all many-body

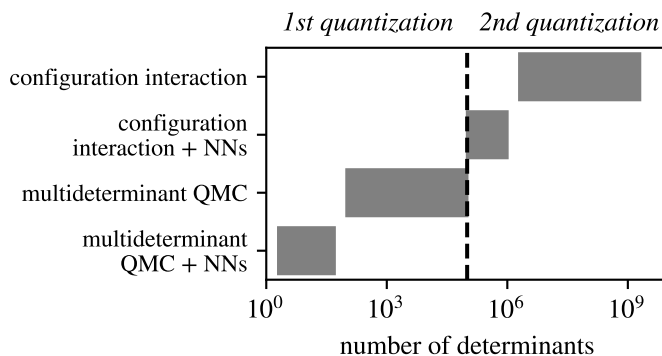


Figure 1 | Typical numbers of Slater determinants required for high-accuracy quantum chemistry methods. The multi-determinant quantum Monte Carlo (QMC) in combination with neural networks (NNs) is the approach developed in this work. The numbers correspond to highly accurate calculations on atomic systems with at most a few tens of electrons.

interactions that are missing in individual determinants (Shavitt and 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. (2019) 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 (Foulkes et al., 2001; Needs et al., 2010; Austin et al., 2012) solve the electronic problem in real space (first quantization), and treat a large part of the correlation in the electronic motion explicitly, which greatly reduces the number of required determinants. Standard QMC variants are still practical for system with hundreds of electrons, such as supramolecular complexes (Ambrosetti et al., 2014) and molecular crystals (Zen et al., 2018). Even though only a single determi-

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

nant can be typically afforded for these larger systems, QMC typically outperforms other electronic-structure methods. 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 the real-space QMC techniques are unable to improve. A key development facilitating the progress in this work is the real-space *backflow* technique (López Ríos et al., 2006). The idea of backflow is to transform the electron positions into pseudoparticles, each of which depends on all electron positions, and then inserting these pseudoparticles into the Slater determinants, leading to an improved nodal surface. While the traditional backflow does not reach the accuracy of the large determinant expansions and does not generalize well to larger systems, Luo and 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 (Behler and Parrinello, 2007; Rupp et al., 2012; Chmiela et al., 2017; Bartók et al., 2017; Smith et al., 2017; Schütt et al., 2018; Faber et al., 2018; 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 highly correlated wave functions with neural networks and their unsupervised training via the variational principle, first proposed by Carleo and 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 wave function complexity of many-body systems. The initial attempts on lattice systems were later generalized to bosons in real space (Saito, 2018), and even electrons in real space (Han et al., 2018), 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 even 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 a multireference Hartree–Fock method as a baseline. These physically motivated choices are essential in order to obtain a method that is not only highly accurate, but also converges robustly, and still maintains relatively high computational efficiency. On a small set of test systems, we show that our neural network ansatz requires only a few determinants to surpass the accuracy of state-of-the-art single-determinant wave function ansatzes by a margin, and approaches the accuracy of multi-determinant QMC methods that use orders of magnitude more determinants. As a result, our method has the asymptotic

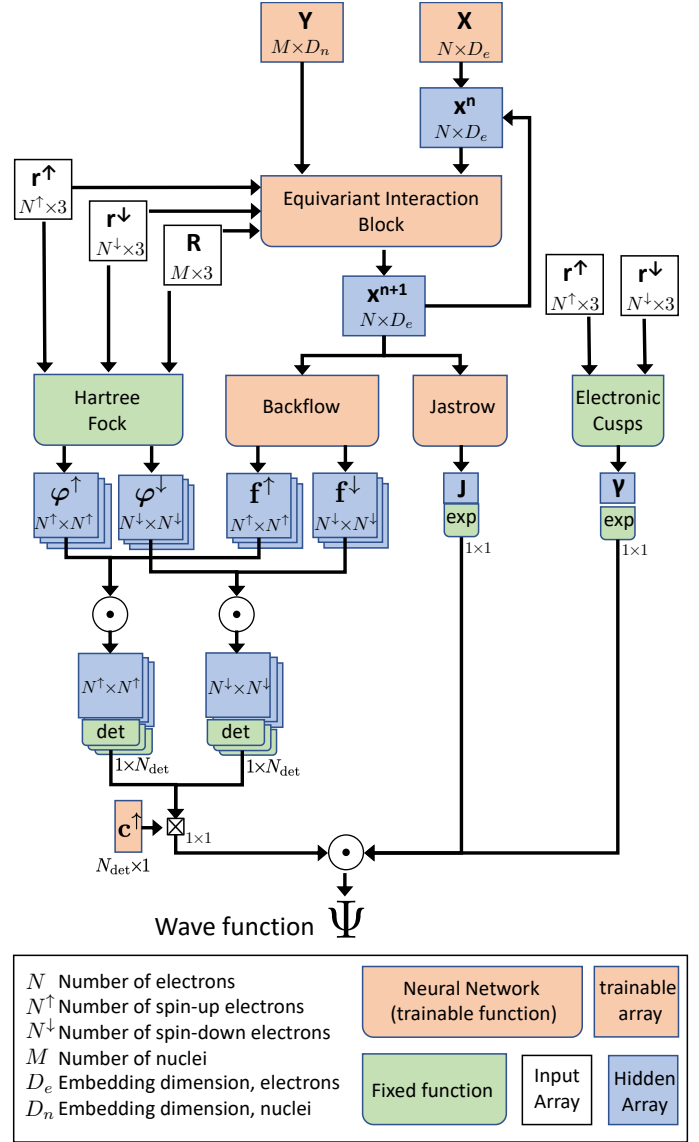


Figure 2 | Information flow in the PauliNet ansatz architecture.

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 highly accurate methods.

The parallel work of Pfau et al. (2019) follows the same basic idea as ours, but differs in several significant aspects. 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. Perhaps as a result of that, their ansatz achieves somewhat higher accuracy at the cost of being about two orders of magnitude more computationally expensive than ours.

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, multi-

determinant expansion, Jastrow factor, backflow transformation, 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 is of the multi-determinant 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 (Figure 2),

$$\begin{aligned}\psi_\theta(\mathbf{r}) &= e^{\gamma(\mathbf{r})+J_\theta(\mathbf{r})} \sum_p c_p \det[\tilde{\varphi}_{\theta,\mu_{pi}}^\dagger(\mathbf{r})] \det[\tilde{\varphi}_{\theta,\mu_{pi}}^\downarrow(\mathbf{r})] \\ \tilde{\varphi}_{\mu i}(\mathbf{r}) &= \varphi_\mu(\mathbf{r}_i) f_{\theta,\mu i}(\mathbf{r})\end{aligned}\quad (1)$$

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:

Antisymmetry 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.

Hartree–Fock baseline To ensure a good starting point for the variational optimization problem, we exploit the approximate Hartree–Fock (HF) quantum chemistry 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 the 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.

Asymptotic behavior 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 and 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 cusplless, 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 with the commonly used form, $e^{J(\mathbf{r})}$, 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 nonlocal coupled coordinates of pseudoelectrons that enter the one-electron molecular orbitals 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 to the aim of 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 convolution DNN that represents each particle with a vector in a high-dimensional abstract feature space, \mathbf{x}_i , which is first initialized by a trainable vector shared by all particles of the same type, and then iteratively refined by interactions with other particles through real-space trainable convolutions, χ_θ , which encode the inter-particle distances and are equivariant with respect to particle exchange,

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

The SchNet architecture and its modifications for PauliNet are described in detail in Methods. After a fixed number of iterations of SchNet, 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 the 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 solution with few determinants

We train PauliNet via the variational principle, minimizing 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 for DeepWF (Han et al., 2018), in particular the hydrogen molecule H_2 , lithium hydride (LiH), beryllium (Be), boron (B), and the linear hydrogen chain H_{10} (Table 1). For these systems, PauliNet recovers between 97% and 99.9% of the electron correlation energy after training for tens of minutes to few hours on a single GTX 1080 Ti graphics processing unit. We compare these results to the standard single-determinant (SD) variational (VMC) and diffusion Monte Carlo (DMC) methods, which, unlike their multi-determinant variants, have the same asymptotic scaling as our approach. For the studied

Table 1 | Ground-state energies of five test systems obtained by four different methods.

system	correlation energy			E/E_h		
	SD-VMC	SD-DMC	DeepWF	PauliNet ^g	exact	
H ₂			98.4%	99.99%	−1.17447 (2)	−1.17447 ^a
LiH	91.5% ^b	99.7% ^b	^c	99.3%	−8.0700 (2)	−8.070548 ^d
Be	61.6% ^e	89.2% ^e	43.6%	99.94%	−14.6674 (3)	−14.66736 ^e
B	60.0% ^e	88.3% ^e	^c	97.3%	−24.6506 (11)	−24.65391 ^e
H ₁₀			63.8%	98.0%	−5.6602 (7)	−5.6655 ^f

^aKolos and Wolniewicz (1965). ^bCasalegno et al. (2003). ^cFor LiH and B, DeepWF does not reach the accuracy of the HF method. ^dCencek and Rychlewski (2000). ^eBrown et al. (2007). ^fMotta et al. (2017). ^gBold digits match the exact energy.

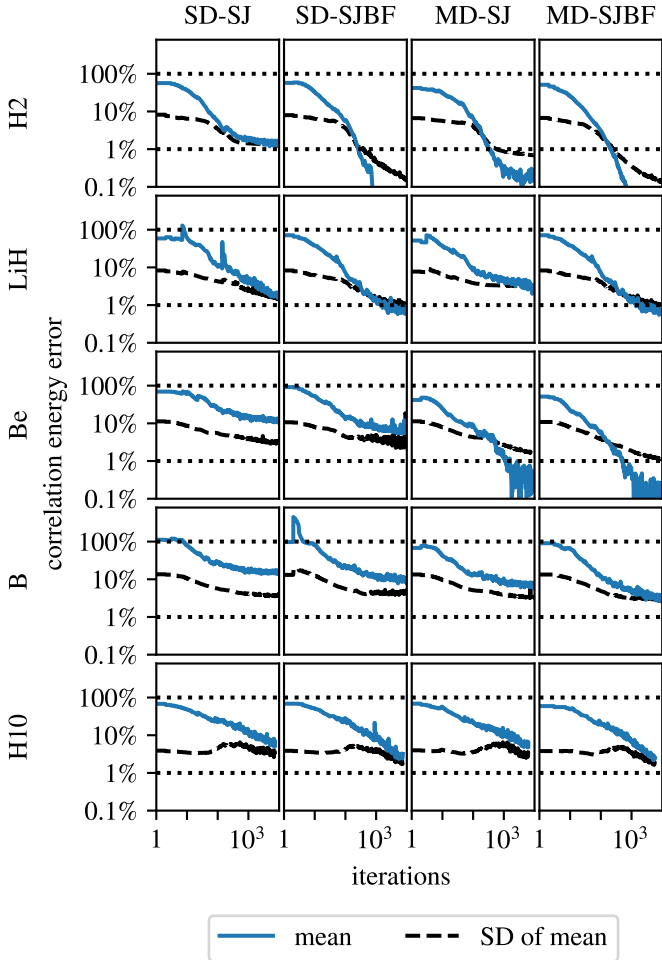


Figure 3 | Learning curves of PauliNet for five molecular systems. The correlation energy error is calculated from the approximate ground-state energy, E_0 , the HF energy, E_{HF} , and the exact ground-state energy, E_0^{exact} , as $1 - (E_0 - E_{HF}) / (E_0^{exact} - E_{HF})$. The mean of the error and its estimated standard deviation ($\sigma / \sqrt{N_b}$) over a single batch of size N_b are plotted for each iteration of the Adam optimization. Optimizations of four different wave function ansatzes are shown, which are combinations of a single or multiple Slater determinants (SD/MD), a Jastrow factor (SJ), and a backflow (BF). In the MD variants, all determinants in a normalized CASSCF expansion with a linear coefficient above 0.01 are used—H₂: 2 determinants, LiH: 6, Be: 4, B: 3, H₁₀: 36.

and SD-DMC recovers 88% to 99.7%, well below the accuracy of PauliNet. The DeepWF ansatz does not even reach the baseline HF accuracy for some systems.

Figure 3 highlights two crucial aspects of our results. First, the error in the correlation energy decreases almost monotonously as the training progresses from the initial HF baseline level to the reported final values. The learning curves in fact never plateau, demonstrating the high expressiveness of our neural network ansatz, and indicating that even higher accuracy could be reached by investing more computational resources. Second, we compare the 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.

Only a few determinants are sufficient to substantially reduce the correlation energy error compared to the single-determinant case: 6 or less determinants for all atoms and diatomic molecules, and 36 determinants for H₁₀. Similar findings were made for FermiNet, indicating that deep learning can indeed 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., 2019).

The total electronic energy is a global measure of the quality of the wave function, but does not necessarily portray whether the wave function is locally correct. Especially with the use of DNNs one could doubt whether the wave function behaves correctly in regions that have not been sampled in the training data. Figure 4 shows that PauliNet generalizes well to such regions. Even though the chosen electron configuration has a very low probability ($|\psi|^2$ orders of magnitude lower than for the most likely configurations), the DNN learned to modify the HF wave function in such a way that the local energy shifts towards the exact value along the whole scan, except for a small region where the probability goes to zero due to the Pauli exclusion principle. The plot also illustrates the correct representation of the nuclear and electronic cusps, as the local energy of the PauliNet wave function does not diverge around the coalescence points.

2.4 Capturing strong correlation

Unlike the atoms and diatomics, the linear hydrogen chain H₁₀ exhibits *strong correlation*, which describes a situation where the single-determinant description of the HF method is insufficient, and the correlation energy constitutes a significant part of the electronic energy (Motta et al., 2017). Without a multi-

systems, SD-VMC recovers 60% to 92% of the correlation energy,

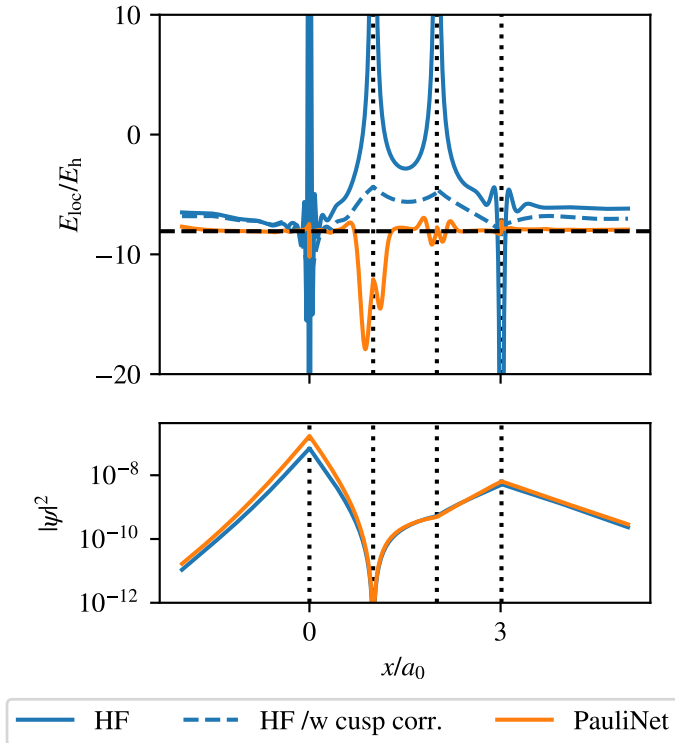


Figure 4 | One-dimensional scan of the wave function and its local energy in LiH. Four vertical dashed lines indicate positions of fixed particles on the x -axis, from left the Li nucleus, a spin-up electron, a spin-down electron, and the H nucleus. The second spin-down electron is positioned randomly off the axis, while the second spin-up electron is moved along the x -axis.

determinant expansion, strong correlation poses a hard test for a Jastrow factor, because distinguishing the high-energy electronic configurations is a complicated function of the positions of all the electrons with respect to each other and to the nuclei, not just of the electron distances (Goetz and Neuscamman, 2017). Already H_2 exhibits strong correlation when the two atoms are dissociated, and PauliNet is able to recover essentially 100% of the correlation energy along the whole dissociation curve (Figure 5a), even in the stripped-down single-determinant form. For H_{10} , we recover 98% (90%) of the exact correlation energy in the equilibrium (stretched) geometry (Figure 5b) with the full multi-determinant form. The decrease of the captured fraction of the correlation energy with larger atom separations illustrates the increased difficulty of the problem due to the strong correlation. Even with a single determinant, our deep Jastrow factor is able to recover more than 90% of the electron correlation for the equilibrium geometry, demonstrating its flexibility.

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 de-

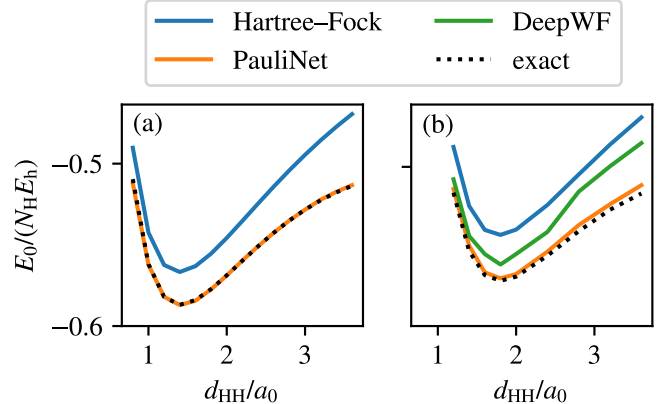


Figure 5 | Electronic energy as a function of distance between hydrogen atoms in H_2 (a) and H_{10} (b). The exact values are calculated with FCI@cc-pvQZ for H_2 and taken from Motta et al. (2017) for H_{10} .

terminant evaluation, and additional N for the evaluation of the kinetic energy) 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 facilitates 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, in standard QMC, the wild oscillations of the local energy close to heavy nuclei mandates the use of pseudopotentials for heavier elements such as transition metals. The flexibility of DNNs could sidestep this necessity by smoothing out these oscillations, which cannot be done with standard wave-function ansatzes.

In classical quantum chemistry methods, strong correlation is usually treated by using large multi-determinant 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 (Goetz and Neuscamman, 2017). In contrast, DNNs are capable of learning strong correlation between electrons without any specialized adaptation. We demonstrate convergence to very high accuracy with single or 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., 2019). This class of methods has the advantage of eliminating much of the com-

plexity 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.

In parallel, Pfau et al. (2019) developed FermiNet, a different deep-learning architecture to solve the electronic Schrödinger equation. Both methods achieve nearly 100% of the correlation energy for the systems tested, illustrating the versatility of deep QMC approaches. The residual error of FermiNet is even smaller than that of PauliNet, but at the expense of computational cost that is approximately two orders of magnitude greater. Lacking learning curves similar to our Figure 3 for FermiNet prevents any detailed comparison of the computational efficiency at this point. However, the results reported by Pfau et al. (2019) were obtained with about 30 times more electronic samples for training, using eight state-of-the-art graphics processing units in parallel instead of one in our case. Establishing a shared protocol for evaluating the accuracy per computational effort for DNN wave function ansatzes will be an important task in order to judge efficiency and usefulness in the future.

Already a brief comparison of the two approaches 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,

$$E_0 = \min_{\psi} E[\psi] \leq \min_{\theta} E[\psi_{\theta}],$$

$$E[\psi] = \int d\mathbf{r} \psi(\mathbf{r}) \hat{H} \psi(\mathbf{r}) \quad (6)$$

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 sampling electron positions with a standard Langevin Monte Carlo approach (Umrigar et al., 1993) every 100 training iterations using the current model wave function ψ_{θ} . 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 and Ba, 2017; Loshchilov and 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),

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

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

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), which is computationally costly and numerically unstable. 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.

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, $\mathbf{r}_i = \mathbf{R}_J$, 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 construction by the DNNs, 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 electronic 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,

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

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

$$\mathbf{x}_i^{(n+1)} := \mathbf{x}_i^{(n)} + \sum_{\pm} \mathbf{g}_{\theta}^{(n,\pm)}(\mathbf{z}_i^{(n,\pm)}) + \mathbf{g}_{\theta}^{(n,n)}(\mathbf{z}_i^{(n,n)}) \quad (11)$$

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 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 electrons and nuclei are particles of 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 and Muwly, 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 in Pytorch. 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 2).

References

A. Ambrosetti, D. Alfè, R. A. DiStasio, Jr., and A. Tkatchenko. Hard Numbers for Large Molecules: Toward Exact Energetics for Supramolecular Systems. *J. Phys. Chem. Lett.*, 5:849–855, 2014. doi:10.1021/jz402663k.

B. M. Austin, D. Y. Zubarev, and W. A. Lester. Quantum Monte Carlo and Related Approaches. *Chem. Rev.*, 112:263–288, 2012. doi:10.1021/cr2001564.

A. P. Bartók, S. De, C. Poelking, N. Bernstein, J. R. Kermode, G. Csányi, and M. Ceriotti. Machine learning unifies the modeling of materials and molecules. *Sci. Adv.*, 3:e1701816, 2017. doi:10.1126/sciadv.1701816.

J. Behler and M. Parrinello. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.*, 98:146401, 2007. doi:10.1103/PhysRevLett.98.146401.

G. H. Booth, A. J. W. Thom, and A. Alavi. Fermion Monte Carlo without fixed nodes: A game of life, death, and annihilation in Slater determinant space. *J. Chem. Phys.*, 131:054106, 2009. doi:10.1063/1.3193710.

Table 2 | 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
Determinant cutoff	1×10^{-2}
Learning rate	0.01
Batch size	10 000
Number of walkers	2000
Number of training steps	7000
Optimizer	AdamW
Learning rate decay period t_0	200
Clipping window q	5
Resampling period	100
Number of discarded sampling steps	50
Number of decorrelation sampling steps	1
Target acceptance	57%

M. D. Brown, J. R. Trail, P. López Ríos, and R. J. Needs. Energies of the first row atoms from quantum Monte Carlo. *J. Chem. Phys.*, 126:224110, 2007. doi:10.1063/1.2743972.

G. Carleo and M. Troyer. Solving the Quantum Many-Body Problem with Artificial Neural Networks. *Science*, 355:602–606, 2017. doi:10.1126/science.aag2302.

M. Casalegno, M. Mella, and A. M. Rappe. Computing accurate forces in quantum Monte Carlo using Pulay’s corrections and energy minimization. *J. Chem. Phys.*, 118:7193, 2003. doi:10.1063/1.1562605.

W. Cencek and J. Rychlewski. Benchmark calculations for He_2^+ and LiH molecules using explicitly correlated Gaussian functions. *Chem. Phys. Lett.*, 320:549–552, 2000. doi:10.1016/S0009-2614(00)00303-1.

D. Ceperley, G. V. Chester, and M. H. Kalos. Monte Carlo simulation of a many-fermion study. *Phys. Rev. B*, 16:3081–3099, 1977. doi:10.1103/PhysRevB.16.3081.

S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt, and K.-R. Müller. Machine learning of accurate energy-conserving molecular force fields. *Sci. Adv.*, 3:e1603015, 2017. doi:10.1126/sciadv.1603015.

K. Choo, A. Mezzacapo, and G. Carleo. Fermionic neural-network states for ab-initio electronic structure. 2019. URL <http://arxiv.org/abs/1909.12852>.

F. A. Faber, A. S. Christensen, B. Huang, and O. A. von Lilienfeld. Alchemical and structural distribution based representation for universal quantum machine learning. *The Journal of Chemical Physics*, 148:241717, 2018. doi:10.1063/1.5020710.

W. M. C. Foulkes, L. Mitás, R. J. Needs, and G. Rajagopal. Quantum Monte Carlo simulations of solids. *Rev. Mod. Phys.*, 73:33–83, 2001. doi:10.1103/RevModPhys.73.33.

B. V. D. Goetz and E. Neuscamman. Suppressing Ionic Terms with Number-Counting Jastrow Factors in Real Space. *J. Chem. Theory Comput.*, 13:2035–2042, 2017. doi:10.1021/acs.jctc.7b00158.

A. Grisafi, A. Fabrizio, B. Meyer, D. M. Wilkins, C. Corminboeuf, and M. Ceriotti. Transferable Machine-Learning Model of the Electron Density. *ACS Cent. Sci.*, 5:57–64, 2019. doi:10.1021/acscentsci.8b00551.

J. Han, L. Zhang, and W. E. Solving Many-Electron Schrödinger Equation Using Deep Neural Networks. 2018. URL <http://arxiv.org/abs/1807.07014>.

T. Kato. On the eigenfunctions of many-particle systems in quantum mechanics. *Commun. Pure Appl. Math.*, 10:151–177, 1957. doi:10.1002/cpa.3160100201.

D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. 2017.

URL <http://arxiv.org/abs/1412.6980>.

- W. Kołos and L. Wolniewicz. Potential-Energy Curves for the $X^1\Sigma^+$, $b^3\Sigma^+$, and $C^1\Pi$ States of the Hydrogen Molecule. *J. Chem. Phys.*, 43:2429–2441, 1965. doi:10.1063/1.1697142.
- P. López Ríos, A. Ma, N. D. Drummond, M. D. Towler, and R. J. Needs. Inhomogeneous backflow transformations in quantum Monte Carlo calculations. *Phys. Rev. E*, 74:066701, 2006. doi:10.1103/PhysRevE.74.066701.
- I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, New Orleans, 2019. URL <https://openreview.net/forum?id=Bkg6RiCqY7>.
- D. Luo and B. K. Clark. Backflow Transformations via Neural Networks for Quantum Many-Body Wave Functions. *Phys. Rev. Lett.*, 122:226401, 2019. doi:10.1103/PhysRevLett.122.226401.
- A. Ma, M. D. Towler, N. D. Drummond, and R. J. Needs. Scheme for adding electron–nucleus cusps to Gaussian orbitals. *J. Chem. Phys.*, 122:224322, 2005. doi:10.1063/1.1940588.
- M. A. Morales, J. McMinis, B. K. Clark, J. Kim, and G. E. Scuseria. Multideterminant Wave Functions in Quantum Monte Carlo. *J. Chem. Theory Comput.*, 8:2181–2188, 2012. doi:10.1021/ct3003404.
- M. Motta, D. M. Ceperley, G. K.-L. Chan, J. A. Gomez, E. Gull, S. Guo, C. A. Jiménez-Hoyos, T. N. Lan, J. Li, F. Ma, A. J. Millis, N. V. Prokof'ev, U. Ray, G. E. Scuseria, S. Sorella, E. M. Stoudenmire, Q. Sun, I. S. Tupitsyn, S. R. White, D. Zgid, S. Zhang, and Simons Collaboration on the Many-Electron Problem. 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, 2017. doi:10.1103/PhysRevX.7.031059.
- R. J. Needs, M. D. Towler, N. D. Drummond, and P. L. Ríos. Continuum variational and diffusion quantum Monte Carlo calculations. *J. Phys. Condens. Matter*, 22:023201, 2010. doi:10.1088/0953-8984/22/2/023201.
- D. Pfau, J. S. Spencer, A. G. d. G. Matthews, and W. M. C. Foulkes. Ab-Initio Solution of the Many-Electron Schrödinger Equation with Deep Neural Networks. 2019. URL <http://arxiv.org/abs/1909.02487>.
- L. Piela. *Ideas of Quantum Chemistry*. Elsevier, Amsterdam, 2nd edition, 2014. ISBN 978-0-444-59436-5.
- M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.*, 108:058301, 2012. doi:10.1103/PhysRevLett.108.058301.
- H. Saito. Method to Solve Quantum Few-Body Problems with Artificial Neural Networks. *J. Phys. Soc. Jpn.*, 87:074002, 2018. doi:10.7566/JPSJ.87.074002.
- K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller. SchNet – A deep learning architecture for molecules and materials. *J. Chem. Phys.*, 148:241722, 2018. doi:10.1063/1.5019779.
- K. T. Schütt, M. Gastegger, A. Tkatchenko, K.-R. Müller, and R. J. Maurer. Unifying machine learning and quantum chemistry with a deep neural network for molecular wavefunctions. *Nat. Commun.*, 10:5024, 2019. doi:10.1038/s41467-019-12875-2.
- I. Shavitt and R. J. Bartlett. *Many-Body Methods in Chemistry and Physics*. Cambridge Molecular Science. Cambridge University Press, Cambridge, 2009. ISBN 978-0-521-81832-2.
- J. S. Smith, O. Isayev, and A. E. Roitberg. ANI-1: An extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.*, 8: 3192–3203, 2017. doi:10.1039/C6SC05720A.
- Q. Sun, T. C. Berkelbach, N. S. Blunt, G. H. Booth, S. Guo, Z. Li, J. Liu, J. D. McClain, E. R. Sayfutyarova, S. Sharma, S. Wouters, and G. K.-L. Chan. PySCF: The Python-based simulations of chemistry framework. *WIREs Comput. Mol. Sci.*, 8:e1340, 2018. doi:10.1002/wcms.1340.
- A. J. W. Thom. Stochastic Coupled Cluster Theory. *Phys. Rev. Lett.*, 105:263004, 2010. doi:10.1103/PhysRevLett.105.263004.
- M. Troyer and U.-J. Wiese. Computational Complexity and Fundamental Limitations to Fermionic Quantum Monte Carlo Simulations. *Phys. Rev. Lett.*, 94: 170201, 2005. doi:10.1103/PhysRevLett.94.170201.
- C. J. Umrigar, M. P. Nightingale, and K. J. Runge. A diffusion Monte Carlo algorithm with very small time-step errors. *J. Chem. Phys.*, 99:2865–2890, 1993. doi:10.1063/1.465195.
- O. T. Unke and M. Meuwly. PhysNet: A Neural Network for Predicting Energies, Forces, Dipole Moments, and Partial Charges. *J. Chem. Theory Comput.*, 15: 3678–3693, 2019. doi:10.1021/acs.jctc.9b00181.
- M. Welborn, L. Cheng, and T. F. Miller. Transferability in Machine Learning for Electronic Structure via the Molecular Orbital Basis. *J. Chem. Theory Comput.*, 14:4772–4779, 2018. doi:10.1021/acs.jctc.8b00636.
- A. Zen, J. G. Brandenburg, J. Klimeš, A. Tkatchenko, D. Alfè, and A. Michaelides. Fast and accurate quantum Monte Carlo for molecular crystals. *Proc. Natl. Acad. Sci.*, 115:1724–1729, 2018. doi:10.1073/pnas.1715434115.

Acknowledgments

We are grateful to the following scientists for inspiring discussions: Cecilia Clementi (Rice, FU Berlin), Jens Eisert (FU Berlin), Gustavo Scuseria (Rice), Jörg Neugebauer (MPIE), and Hao Wu (Tongji). 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. would like to thank K.-R. Müller for support and acknowledge funding from TU Berlin (Project No. 10032745).