Ab-initio quantum chemistry with neural-network wavefunctions

Jan Hermann[1,†], James Spencer[2,†], Kenny Choo[3,4,†], Antonio Mezzacapo[5], W. M. C. Foulkes[6], David Pfau[2,6,*], Giuseppe Carleo[7,*], and Frank Noé[1,8,9,10,*]

[1]*FU Berlin, Department of Mathematics and Computer Science, Arnimallee 12, 14195 Berlin, Germany*
[2]*DeepMind, 6 Pancras Square, London N1C 4AG, United Kingdom*
[3]*University of Zurich, Department of Physics, Winterthurerstrasse 190, 8057 Zurich, Switzerland*
[4]*IBM Quantum, IBM Research Zurich, Saumerstrasse 4, 8803 Ruschlikon, Switzerland*
[5]*IBM Quantum, Thomas J. Watson Research Center, Yorktown Heights, New York 10598, USA*
[6]*Imperial College London, Department of Physics, South Kensington Campus, London SW7 2AZ, United Kingdom*
[7]*EPFL, Institute of Physics, CH-1015 Lausanne, Switzerland*
[8]*FU Berlin, Department of Physics, Arnimallee 12, 14195 Berlin, Germany*
[9]*Rice University, Department of Chemistry, 6100 Main St, Houston, TX 77005, United States*
[10]*Microsoft Research Cambridge, 21 Station Rd, Cambridge CB1 2FB, United Kingdom*

Abstract Machine learning and specifically deep-learning methods have outperformed human capabilities in many pattern recognition and data processing problems, in game playing, and now also play an increasingly important role in scientific discovery. A key application of machine learning in the molecular sciences is to learn potential energy surfaces or force fields from ab-initio solutions of the electronic Schrödinger equation using datasets obtained with density functional theory, coupled cluster, or other quantum chemistry methods. Here we review a recent and complementary approach: using machine learning to aid the direct solution of quantum chemistry problems from first principles. Specifically, we focus on quantum Monte Carlo (QMC) methods that use neural network ansatz functions in order to solve the electronic Schrödinger equation, both in first and second quantization, computing ground and excited states, and generalizing over multiple nuclear configurations. Compared to existing quantum chemistry methods, these new deep QMC methods have the potential to generate highly accurate solutions of the Schrödinger equation at relatively modest computational cost.

1 Introduction

In the past decade, machine learning (ML) has made inroads into many areas of the physical sciences (Carleo et al. 2019), often outperforming more traditional computational methods (Deringer et al. 2021b; Jumper et al. 2021) or offering entirely new approaches to solve scientific problems (Huang & von Lilienfeld 2020; Noé et al. 2019). Quantum chemistry (QC) has been among the first fields to have been affected by this revolution (Noé et al. 2020; Tkatchenko 2020; von Lilienfeld & Burke 2020). Most applications of ML in QC have been concerned with supervised learning of molecular properties from molecular structure (Dral 2020), either across conformational (Unke et al. 2021) or chemical space (von Lilienfeld et al. 2020), as well as with unsupervised learning for the generation of novel molecules (Bian & Xie 2021). These methods all require a pre-existing dataset of molecules and their properties as an input, typically obtained with standard methods of QC such as density functional theory (Jones 2015) or coupled cluster theory (Bartlett & Musiał 2007). In these scenarios, ML accurately approximates a given method of QC at vastly increased computational efficiency. This approach has been already reviewed in other works cited above. In contrast, the current review focuses on the complementary use of ML as an ab-initio technique in QC, which requires no external data and instead recovers molecular properties from first principles. Here, ML is "integrated" into QC, with the goal of arriving at ab-initio methods with a more favourable accuracy–efficiency trade-off than traditional QC methods.

The goal of computational chemistry is to predict properties of known molecules and to design molecules with desired properties. Most molecular properties are determined by the behaviour of the electrons, so QC methods attempt to approximate the Schrödinger equation for electrons in molecules. Traditionally, QC methods are divided into ab-initio and semi-empirical methods, where the former have no fitted parameters determined from external data, whereas the latter do. Methods that do not use quantum mechanics at all (such as force fields) are called empirical and are typically not considered part of QC, although this view may be changing with the advent of principled and accurate ML-based empirical methods. It is useful to cast these three categories of methods in the light of ML terminology (Fig. 1a).

ML can be roughly divided into supervised, unsupervised, and reinforcement learning. In supervised learning the ML model learns to predict the labels (outputs) of the data (inputs) from a given dataset so as to minimize the difference between the predicted and reference labels. By identifying the inputs with molecular structures and the outputs with molecular properties, all semi-empirical and empirical methods of QC fit into supervised learning, but using mostly relatively simple and physically motivated functional forms rather than the more general and highly flexible functions typical for ML. Vice versa, the many recent successful supervised ML models that predict energies or other molecular properties based on QC training data can be classified as empirical methods (Behler 2021; Deringer et al. 2021a; Musil et al. 2021; Unke et al. 2021). Unsupervised learning is concerned with unlabelled data, and the general task is to learn the underlying probability distribution that would generate a given dataset.

a

Ab-initio
_____
"Self-generative"

no data

Quantum chemistry
_____
Machine learning

Semi-empirical

unlabelled data

Structure generation
_____
Unsupervised

labelled data (input–output)

Empirical
_____
Supervised

b

Accuracy

Deep Quantum Monte Carlo (QMC)

Density functional theory (DFT)

Configuration interaction, quadruples (CISDTQ)

Coupled-cluster (CCSD(T))

Configuration interaction, doubles (CISD)

Møller-Plesset 2nd order (MP2)

Hartree-Fock (HF)

Computational cost

N^3 N^4 N^5 N^6 N^7 N^8 N^9 N^{10} (_N_ electrons)
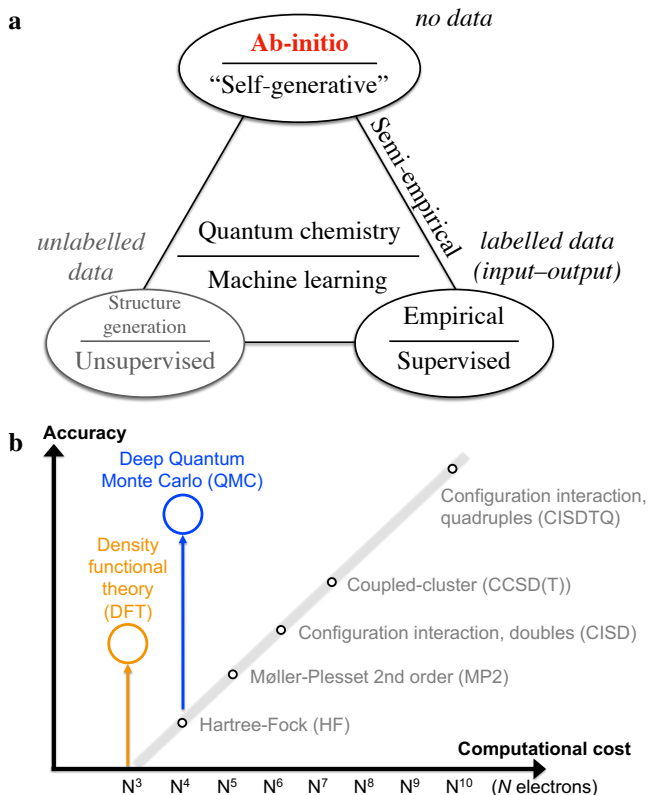
Fig. 1 | Quantum chemistry and machine learning. (**a**) Machine learning disciplines and their dependence on data can be mapped to disciplines in quantum chemistry. This work reviews the use of machine learning in ab-initio quantum chemistry, where the only input to machine learning is the Schrödinger equation itself. This approach uses self-generated data, rather than relying on external data. The closest analogue in machine learning is reinforcement learning with self-play, which substitutes data from an external environment with data generated by the agent, though in many other respects the two approaches are distinct. (**b**) Trade-off between computational efficiency and accuracy in quantum chemistry methods. Accuracy of electronic structure methods against the asymptotic scaling of their computational cost with system size, N. Popular methods, such as density functional theory, are outliers from the general trend.

Examples in chemistry include generative models for structural formulas (Gómez-Bombarelli et al. 2018) as well as full 3D structures of molecules (Hoogeboom et al. 2022; Noé et al. 2019), and in physics the estimation of quantum states from measurements, known as quantum tomography (Torlai et al. 2018). Finally, in reinforcement learning, the ML model (also referred to as an _agent_ is able to interact directly with its environment, rather than to just passively receive data. Here, the aim is for the agent to learn a _policy_ for how to interact with the environment so as to maximize a long-term reward (Sutton & Barto 2018). Reinforcement learning is behind some of the most prominent successes of ML such as playing games at a superhuman level (Mnih et al. 2015; Silver et al. 2016; Tesauro 1994) or the control of plasma in tokamaks (Degrave et al. 2022). In certain settings the agent can self-generate data by treating its own policy as the environment. This is known as _self-play_, and has been the basis for many advances in symmetric games (Heinrich et al. 2015; Silver et al. 2018). Although there are many key differences, this is the branch of ML concep-

tually most similar to ab-initio QC, in the sense that no external data other than the rules of the system or game are required for either.

In the traditional picture, one moves from empirical to ab-initio methods by retaining more of the first-principles physics. Similarly, there is a general trend for ML models in chemistry to encode an increasing amount of molecular physics. This includes physical constraints such as energy conservation (Chmiela et al. 2017), invariance and equivariance of molecular properties with respect to rotation, translation, or exchange of indistinguishable particles (Behler & Parrinello 2007; Schütt et al. 2017), as well as other physical concepts such as many-body expansions (Drautz 2019) or even surrogate quantum-mechanical models (Kirkpatrick et al. 2021; Li et al. 2018; Schütt et al. 2019). Similar considerations can be made for the problem of ab-initio learning of solutions to the electronic Schrödinger equation introduced here and we will discuss different strategies throughout the review.

The Schrödinger equation is an eigenvalue problem that can be equivalently formulated via several variational principles—its solutions, the _eigenstate_ wavefunctions and energies, can be found by searching for stationary points of certain functionals over the space of all physically admissible wavefunctions. Importantly, the ground state of a molecule can be found by minimizing the energy expectation value of a wavefunction. This principle underlies many ab-initio QC methods, and also the methods in this review, as such a variational principle naturally defines a ML problem—the eigenstates (such as the ground state) are represented as a neural network and the parameters of that network are obtained by minimizing the variational electronic energy. The reviewed methods differ in the particular form of the neural-network ansatz used, as described below.

Section 2 briefly reviews the components of electronic structure theory necessary for the development of the ML methods to be discussed later on. The electronic structure problem is mapped to ML in Section 3, which is followed by a review of the ab-initio ML methods for QC formulated in real space and in a discrete basis in Sections 4 and 5, respectively. The review is concluded in Section 6.

2 Electronic structure

2.1 Schrödinger equation

QC aims at finding approximate solutions of the electronic Schrödinger equation that strike a good balance between accuracy and efficiency (Piela 2020) (Fig. 1b). The non-relativistic electronic Schrödinger equation within the Born–Oppenheimer approximation for a given molecule specified by the charges and coordinates of the nuclei, Z_I, $\mathbf{R}_I$, is a second-order differential equation for the wavefunction, $\psi(\mathbf{r}_1, \ldots, \mathbf{r}_N)$, which is a function of the coordinates of N electrons (Fig. 2a):

$$\hat{H}\psi(\mathbf{r}_1, \ldots, \mathbf{r}_N) = E\psi(\mathbf{r}_1, \ldots, \mathbf{r}_N), \quad (1)$$

$$\hat{H} := \sum_i \left(-\frac{1}{2}\nabla^2_{\mathbf{r}_i} - \sum_I \frac{Z_I}{|\mathbf{r}_i - \mathbf{R}_I|} \right) + \sum_{i<j} \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|}. \quad (2)$$

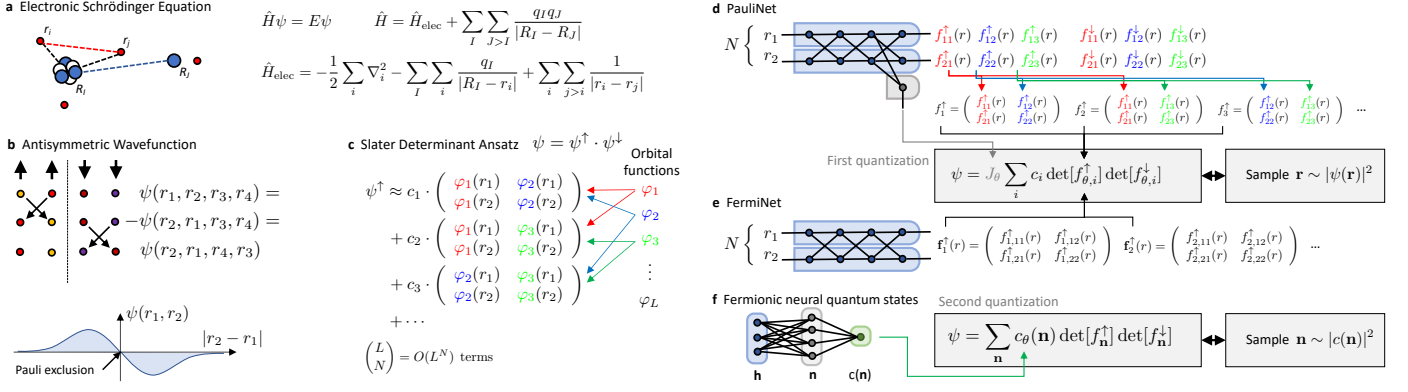An alternative formulation of the Schrödinger equation uses the

Fig. 2 | Electronic structure problem and its neural-network solutions. (**a**) The problem is fully specified by the geometry of a molecule and the electronic Schrödinger equation. (**b**) Only fully antisymmetric wavefunctions are admissible as solutions due to the Pauli exclusion principle and (**c**) these are often represented with Slater determinants. (**d,e**) Solutions formulated in first quantization use antisymmetric neural networks to represent the wavefunction directly in real space. (**f**) Second quantization transfers the antisymmetry to a fixed finite basis, enabling the use of vanilla neural networks.

notion of an *expectation value*,

$$
\begin{aligned}
\langle \hat{H} \rangle_{\psi} &\equiv \frac{\langle \psi | \hat{H} | \psi \rangle}{\langle \psi | \psi \rangle} \\
&= \frac{\int d\mathbf{r}_1 \cdots d\mathbf{r}_N \, \psi(\mathbf{r}_1, \ldots, \mathbf{r}_N) \hat{H} \psi(\mathbf{r}_1, \ldots, \mathbf{r}_N)}{\int d\mathbf{r}_1 \cdots d\mathbf{r}_N |\psi(\mathbf{r}_1, \ldots, \mathbf{r}_N)|^2}.
\end{aligned}
\tag{3}
$$

Instead of solving Eq. (1), the ground-state (lowest-energy) solution can be found by minimizing this energy expectation value with respect to all possible wavefunctions (variational principle),

$$
E = \min_{\psi} \langle \hat{H} \rangle_{\psi}.
\tag{4}
$$

2.2 Antisymmetric wavefunctions

Electrons are fermions, and as such their wavefunction must be antisymmetric with respect to exchange of any two electrons. This cardinal feature of electronic wavefunctions permeates the whole of QC. In general, electrons also possess spin coordinates, $s_i \in \{\uparrow, \downarrow\}$, but the nonrelativistic Hamiltonian does not operate on spin, so the spin coordinate of each electron can be considered fixed. To simplify the presentation here (for full treatment, see Foulkes et al. 2001, Sec. IV.E), we take advantage of the fixed spin coordinates, so the *spatial* wavefunction must be antisymmetric only with respect to the exchange of same-spin electrons, i.e., when $s_i = s_j$ (Fig. 2b),

$$
\psi(\ldots, \mathbf{r}_i, \ldots, \mathbf{r}_j, \ldots) = -\psi(\ldots, \mathbf{r}_j, \ldots, \mathbf{r}_i, \ldots).
\tag{5}
$$

By far the most common way to form antisymmetric wavefunctions in QC is as antisymmetrized products of single-electron functions (orbitals), $\phi_j(\mathbf{r})$. These products can be written as determinants of an $N \times N$ matrix, $\phi_j(\mathbf{r}_i)$, formed by putting N electrons into N orbitals, and are referred to as Slater determinants (Fig. 2c):

$$
D_{\phi}(\mathbf{r}_1, \ldots, \mathbf{r}_N) = \frac{1}{\sqrt{N!}}
\begin{vmatrix}
\phi_1(\mathbf{r}_1) & \phi_1(\mathbf{r}_2) & \cdots & \phi_1(\mathbf{r}_N) \\
\phi_2(\mathbf{r}_1) & \phi_2(\mathbf{r}_2) & \cdots & \phi_2(\mathbf{r}_N) \\
\vdots & \vdots & \ddots & \vdots \\
\phi_N(\mathbf{r}_1) & \phi_N(\mathbf{r}_2) & \cdots & \phi_N(\mathbf{r}_N)
\end{vmatrix}.
\tag{6}
$$

When interpreting $\phi_j(\mathbf{r}_i)$ as the j-the component of a N-dimensional feature vector for the i-th electron (using ML parlance), $\phi(\mathbf{r}_i)$, a Slater determinant is in fact the only antisymmetric function of N feature vectors that is linear in every one of them, making it a natural choice. Alternative antisymmetric forms exist, such as the Pfaffian (Bajdich et al. 2006) or the Vandermonde determinant and its generalizations (Acevedo et al. 2020; Han et al. 2019), but these are far less common and we will not discuss them here.

Slater determinants formed from different orbitals can be further mixed in a linear combination without breaking the antisymmetry (Fig. 2c). In fact, this simple technique is the powerhouse behind all the high-accuracy methods of QC, yet it is also its bane, because the number of Slater determinants required to achieve a given accuracy rises exponentially with the number of atoms in most cases. For fermionic wavefunctions there is no known general approach to effectively reduce the search space from this exponential regime without sacrificing accuracy. However, QC has produced many methods that achieve excellent approximations for specific molecules and materials of practical interest. The cost of these highly accurate methods is generally less than exponential, but nevertheless increases rapidly with system size (Fig. 1b).

2.3 Variational wavefunction methods

An important class of QC methods derives directly from the variational principle (Eq. 4), by assuming a certain wavefunction *ansatz*, $\psi(\cdot; \boldsymbol{\theta})$, parametrized by $\boldsymbol{\theta}$. Minimizing the energy of this ansatz with respect to $\boldsymbol{\theta}$ then always yields an upper bound for the exact ground-state energy,

$$
E = \min_{\psi} \langle \hat{H} \rangle_{\psi} \leq \min_{\boldsymbol{\theta}} \langle \hat{H} \rangle_{\psi(\cdot; \boldsymbol{\theta})}.
\tag{7}
$$

The bound becomes tighter as the expressiveness of the ansatz is improved.

One can distinguish two strategies to construct the ansatzes. First, traditional QC uses relatively simple forms, such that the integral of Eq. (3) can be evaluated analytically, which drastically simplifies the minimization problem (Piela 2020; Szabo & Ostlund 1996). Second, quantum Monte Carlo (QMC) enables

the use of arbitrarily complex ansatzes at the cost of having to do the integral evaluation and minimization stochastically (Becca & Sorella 2017). The latter is a natural framework to incorporate neural networks, and we introduce it in more detail in Section 3.1.

Here we introduce three ansatzes for electronic wavefunctions of the first (traditional) kind, since they serve as scaffolding for the neural-network ansatzes of Sections 4 and 5. We also briefly discuss how they relate to other popular QC methods.

Hartree–Fock Perhaps the simplest nontrivial ansatz in QC is the single Slater determinant of Eq. (6), where the orbitals $\phi_j(\mathbf{r})$ are considered as free parameters. Optimized variationally, this ansatz leads to the so-called Hartree–Fock (HF) method. In practice the orbitals are linearly expanded in a fixed finite one-electron basis, $\varphi_k(\mathbf{r})$, $k = 1, \ldots, K$, with $K \sim N$ in most cases:

$$E_{\mathrm{HF}} = \min_{\phi_j} E[\det \phi_j(\mathbf{r}_i)] \approx \min_{C_{kj}} E\big[\det \textstyle\sum_k C_{kj}\varphi_k(\mathbf{r}_i)\big]. \quad (8)$$

The use of a finite basis set turns the functional optimization problem of Eq. (8) into a computational problem whose cost scales with the fourth power of the number of basis functions, $O(K^4)$, assuming a naive implementation. On its own, the HF ansatz is expressive enough to describe much of chemistry qualitatively, but not always and certainly not quantitatively. However, it can be considered a starting point for most wavefunction-based QC methods.

Density functional theory (DFT) is not such a method, relying instead on an in-principle exact mapping of the ab-initio Hamiltonian (Eq. 2) to a mean-field-like problem, which can be solved exactly with a single Slater determinant (Jones 2015; Teale et al. 2022). However, the variational principle does not hold in DFT because the exchange-correlation contributions to the energy functional are not known exactly and must be approximated in practice. From here on, we will stay within the variational principle and instead focus on increasing the expressiveness of the HF ansatz.

Configuration interaction The HF ansatz can be straightforwardly extended by forming multiple Slater determinants from different sets of orbitals and considering their linear combination (Fig. 2c),

$$\psi(\mathbf{r}_1, \ldots, \mathbf{r}_N) = \sum_p c_p D_{\phi_p}(\mathbf{r}_1, \ldots, \mathbf{r}_N). \quad (9)$$

When the orbitals of each determinant are pooled from a larger superset of (mutually orthogonal) *fixed* orbitals of size $M > N$, and the only free parameters are the linear coefficients of the determinants, the ansatz is called *configuration interaction* (CI). One of the appeals of the CI ansatz is that its Slater determinants can be considered a many-electron antisymmetric basis and labelled using the occupation numbers of the one-electron states. This so-called *second quantized* formalism has many convenient properties for computation (see Box 1). The simplest version of CI, called full CI (FCI), considers all $\binom{M}{N}$ possible Slater determinants and is exact within the chosen finite one-electron basis. In the usual case when $M \sim N$, however, the computational effort scales exponentially with N, which makes FCI applicable only to the smallest molecules. Ways to tackle the exponential scaling include fixed truncation of the CI expansion or its "compression" through analytical means (coupled cluster theory, Bartlett & Musiał 2007; matrix product states, Chan & Sharma 2011), deterministic pruning (selected CI, Huron et al. 1973), or stochastic sampling (FCI-QMC, Booth et al. 2009). Section 5 explores a novel way of "compressing" the CI expansion through neural networks.

Beyond fixed bases The effectiveness of the CI ansatz depends on the choice of the fixed molecular orbitals $\phi_j(\mathbf{r})$ from which
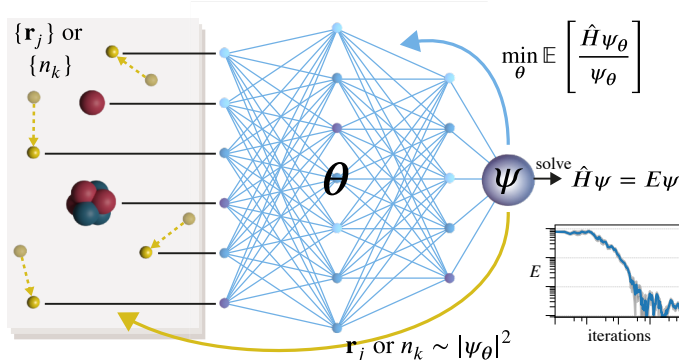
Fig. 3 | Variational Monte Carlo with neural networks. Electron positions, $\mathbf{r}_j$, or orbital occupation numbers, n_k, describe an electron configuration which is an input to the wavefunction, ψ, represented by a neural network parametrized with θ. The wavefunction is used in two ways: first, to sample new electron configurations which provide new input to the neural network (yellow), and second, to evaluate the electronic energy, which is minimized by varying the network parameters (blue).

the Slater determinants $D_{\phi_p}(\mathbf{r}_1, \ldots, \mathbf{r}_N)$ are built. A natural extension of CI allows both the orbitals and the CI expansion coefficients c_p to vary during the variational minimization. Such an ansatz of two stacked linear combinations (Eqs. 8 and 9) is harder to optimize but much more expressive. The most common variant is to consider all $\binom{M'}{N'}$ Slater determinants formed by letting $N' < N$ electrons occupy a space of $M' < M$ orbitals, while the remaining $N - N'$ electrons occupy a fixed set of inactive obitals. This is called the *complete active space self-consistent field* (CASSCF) method (Olsen 2011). Due to the larger variational freedom, a CASSCF ansatz typically requires many fewer determinants than a CI ansatz of comparable accuracy.

But CASSCF and even FCI are still limited by the fixed one-electron basis used to form the molecular orbitals (Eq. 8): FCI is only exact in the complete basis set limit, which in practice cannot be reached for any but the smallest molecular systems. An extension of the CASSCF ansatz would allow not only the one-electron orbitals but also the one-electron basis functions to vary. The stacked structure of such an ansatz would be reminiscent of deep neural networks, and Section 4 explores the culmination of this line of thought by incorporating actual deep neural networks into the ansatz. This removes any *a priori* limitations on the expressiveness. By making each individual determinant maximally expressive, such ansatzes further reduce the number of determinants required to reach a given accuracy.

3 Machine learning for electronic Schrödinger equation

3.1 Mapping quantum mechanics to machine learning

A ML problem and its solution are specified by the model, its inputs and outputs, the data, and the optimization criterion (loss function). In this regard, solving the Schrödinger equation with the variational principle amounts to the following ML problem (Fig. 3). The neural network (Section 3.2) represents a wavefunction, which accepts electron coordinates (first quantization) or occupation numbers (second quantization) as input and outputs the wavefunction value. The loss function is the energy ex-

---

Box 2 | Variational Monte Carlo

Optimization of wavefunctions with neural networks naturally leads to the variational Monte Carlo (VMC) framework. First, Monte Carlo integration of Eq. (3) can handle arbitrarily complicated ansatzes for which analytical integrals are not available. Second, VMC samples these integrals stochastically which naturally combines with the stochastic gradient descent used for optimizing neural networks. In traditional QC, VMC has been used extensively with real-space first-quantized approaches (Foulkes et al. 2001) and more recently in the discrete-basis second-quantized setting (Neuscamman 2013; Sabzevari & Sharma 2018).

The expectation value of any operator, such as the Hamiltonian (Eq. 3), can be written as a Monte Carlo integral over a continuous or discrete basis, $\{|\mathbf{x}\rangle\}$,

$$\langle \hat{H} \rangle_\psi = \int_{\mathbf{x}} \frac{|\langle \mathbf{x}|\psi\rangle|^2}{\langle \psi|\psi\rangle} \frac{\langle \mathbf{x}|\hat{H}|\psi\rangle}{\langle \mathbf{x}|\psi\rangle} = \mathbb{E}_{\mathbf{x}\sim|\langle \mathbf{x}|\psi\rangle|^2} \left[E_{\text{loc}}(\mathbf{x}) \right].$$

Here, the expectation value is obtained as an expected value of a "local" energy E_{loc}, local in the sense that it is defined for every basis element $\mathbf{x}$.

A straightforward and generally applicable way to obtain the samples is Markov-chain Monte Carlo (MCMC). MCMC is an iterative procedure, in which a new sample point, $\mathbf{x}'$, is produced from a current one, $\mathbf{x}$, by making a proposal step with probability $g(\mathbf{x}'|\mathbf{x})$, and then accepting or rejecting the proposal with probability

$$p = \min \left(1, \frac{|\langle \mathbf{x}'|\psi\rangle|^2 g(\mathbf{x}|\mathbf{x}')}{|\langle \mathbf{x}|\psi\rangle|^2 g(\mathbf{x}'|\mathbf{x})} \right).$$

The resulting Markov chain then samples $|\langle \mathbf{x}|\psi\rangle|^2$. Variants of MCMC differ in the construction of the proposal steps and g, and include the simplest Metropolis algorithm ($g(\mathbf{x}'|\mathbf{x}) = g(\mathbf{x}|\mathbf{x}')$) as well as more sophisticated flavours such as Langevin Monte Carlo.

The VMC formula for the expectation value is exact in the limit of infinite sample size, $\mathcal{N} \to \infty$, but in practice it incurs a statistical error proportional to $\sqrt{\text{Var}[E_{\text{loc}}]/\mathcal{N}}$. While $1/\sqrt{\mathcal{N}}$ converges slowly with sample size, VMC has the great benefit that the as the ansatz converges to the exact eigenstates, the local energy converges to a constant (the exact energy), and as such its variance vanishes and so does the statistical sampling error.

---

pectation value corresponding to this wavefunction. The inputs are sampled from the probability distribution given by the square of the wavefunction represented by the current neural network, and the Hamiltonian operator is used to obtain an estimate of the loss function from the samples. The parameters of the network, and thus the wavefunction, are then modified to minimize the loss function. Except for the representation of the wavefunction as a network, this is the regular variational Monte Carlo (VMC) framework (Box 2). The optimization methods used (Box 3) are

Table 1 | Dictionary of electronic structure and machine learning.

Electronic structure	Machine learning
Wavefunction	Probability distribution
Natural orbital	Marginal distribution
Stochastic reconfiguration	Natural gradient descent
Hartree–Fock	Mean-field variational Bayes
Diffusion Monte Carlo	Particle filtering; Sequential Monte Carlo

also fairly conventional, although adapted to a neural network context. This straightforward correspondence between the Schrödinger equation and ML led to the introduction of similar concepts on both sides, albeit known under different names (Table 1).

The applicability of deep learning for quantum-mechanical calculations was first realized and exploited by Carleo & Troyer (2017) for the case of spin lattices in one and two dimensions. Their approach, known as Neural Quantum States (NQS), has since been applied to many different quantum systems (Adams et al. 2021; Astrakhantsev et al. 2021; Nomura et al. 2017; Saito 2017). In essence, this review is concerned with the extension of this approach to electrons in molecules.

3.2 Deep learning

The standard practice in ab-initio QC today is in some ways analogous to the state of computer vision before the rise of deep learning. Prior to 2012, the best pipelines for large-scale image recognition consisted of a combination of hand-designed features and simple ML models (Perronnin et al. 2010). A single deep convolutional neural network trained end-to-end was able to cut the recognition error in half relative to these systems (Krizhevsky et al. 2012), and since then deep neural networks have dominated computer vision research.

In ab-initio QC, ground-state solutions to the Schrödinger equation are usually represented by a wavefunction ansatz with a relatively simple functional form, and parameters are usually fit through a mix of procedures (fixed-point iteration, variational optimization) rather than a unified end-to-end estimation of all parameters simultaneously. The development of deep QMC methods is driven by the hope that the use of neural networks will significantly increase the expressiveness of wavefunction ansatzes, enabling large leaps in accuracy as in image recognition. To appreciate how and why deep neural networks can be usefully applied in QC, a brief review of their application in artificial intelligence is necessary. For a thorough review of the history of deep learning, see Schmidhuber (2015), and for a review of the fundamental concepts in deep learning, see LeCun et al. (2015).

Neural networks date back to the very beginning of computer science (McCulloch & Pitts 1943), and their modern form originates with the single perceptron "unit" (Rosenblatt 1958), which produces as output a non-linear function of the sum of a constant, known as the bias, and a linear combination of its inputs. The non-linear function rises from zero to one as its input increases, mimicking the activation function of a biological neuron. When many such units are assembled in parallel to form a "layer," and several layers are computed serially, taking the output from one

---

Box 3 | Optimizing neural-network ansatzes

Up to the statistical error, the VMC expectation value for the energy (Box 2) obeys the variational principle (Eq. 4). VMC exploits this by varying a parametric wavefunction ansatz ψ_{θ} so as to minimize the energy. For a sufficiently expressive ansatz, the variational energy will eventually approximate the ground state energy of Eq. (1) and the ansatz will approximate the ground state wavefunction Ψ.

The most straightforward optimization method is gradient descent, where the parameters are iteratively updated as

$$\theta \leftarrow \theta - \eta \frac{\partial E(\theta)}{\partial \theta}$$

with learning rate $\eta > 0$. The energy gradient is given by

$$\frac{\partial E}{\partial \theta_k} = \langle \hat{O}_k^* \hat{H} \rangle - \langle \hat{H} \rangle \langle \hat{O}_k^* \rangle,$$

where

$$\hat{O}_k(\mathbf{x}) = \frac{\partial \ln \psi(\mathbf{x}; \theta)}{\partial \theta_k}$$

is an operator representing the logarithmic derivatives of the wavefunction. This gradient can be efficiently estimated using Monte Carlo integration (Box 2).

In some cases the optimization can be sped up and made more stable with higher-order methods, such as the stochastic reconfiguration (SR) scheme (Sorella 1998). SR takes the correlation between individual variational parameters into account by introducing the quantum geometric tensor S:

$$S_{kk'} = \langle O_k^* O_{k'} \rangle - \langle O_k^* \rangle \langle O_{k'} \rangle.$$

The update rule is then modified to

$$\theta \leftarrow \theta - \eta \mathbf{S}^{-1} \frac{\partial E(\theta)}{\partial \theta}.$$

The SR scheme approximates an imaginary-time evolution where each iteration tries to best approximate the state $e^{-\eta \hat{H}} |\psi\rangle$. SR is similar to the natural gradient descent algorithm (Amari 1998) that is well-known in the ML community, and $\mathbf{S}$ can be interpreted as a quantum generalization of the Fisher information matrix (Ay et al. 2017). In some cases, it is convenient to approximate the quantum geometric tensor $\mathbf{S}$ using the Kronecker-factored approximate curvature (KFAC) approach (Martens & Grosse 2015).

---

layer as the input to the next, the resulting multi-layer perceptron (MLP) can, in theory, represent any smooth function to arbitrary accuracy given enough units (Hornik et al. 1989). However, actually fitting or *learning* a set of parameters that matches any given function is different matter. A form of gradient descent utilizing derivatives computed using backpropagation, or reverse-mode automatic differentiation (Linnainmaa 1970, 1976; Werbos 1974), was found to be effective for training neural networks (Rumelhart et al. 1986). This led to a wave of enthusiasm

for neural networks, which eventually faded as several issues were discovered, such as the infamous "vanishing gradients" and getting stuck in local minima.

Several factors were instrumental in rehabilitating neural networks under the banner of "deep learning": a combination of algorithmic advances (Glorot & Bengio 2010) and the use of modern GPU hardware (Hooker 2020) made the computations much faster, and the resulting ability to train larger networks made issues with local minima less severe (Choromanska et al. 2015; Dauphin et al. 2014). Furthermore, deep neural networks with the help of stochastic gradient descent can be applied straightforwardly and efficiently to large datasets, unlike other ML models (Bottou & Bousquet 2008, 2011). Finally, empirical successes like winning the ImageNet Large Scale Visual Recognition Challenge (Russakovsky et al. 2015) helped legitimize deep learning research and generate excitement among researchers.

Today, the barrier to entry for developing and training deep neural networks is quite low, thanks to a mature ecosystem of software libraries for numerical computing with automatic differentiation and hardware accelerators (Abadi et al. 2016; Bradbury et al. 2018; Paszke et al. 2017). However, actually achieving good performance from a deep learning model still requires some finesse and application of various heuristics. It is safe to say that a significant amount of the practice of deep learning remains more art than science. The good news is that once effective heuristics for a particular problem domain have been developed, these same heuristics can often be applied with little modification to other problems in the same domain.

3.3 Neural network architectures

The starting point for most neural networks is the multi-layer perceptron (MLP), formed as a composition of L layers,

$$\begin{aligned} \text{MLP}(\mathbf{x}) &= f^L \circ f^{L-1} \circ \cdots \circ f^1(\mathbf{x}), \\ f^\ell(\mathbf{z}) &= f\left(\mathbf{W}^\ell \mathbf{z} + \mathbf{b}^\ell\right), \end{aligned} \quad (10)$$

where f is some non-linear *activation* function, and $\mathbf{W}^\ell$ and $\mathbf{b}^\ell$ are the matrices of weights and vectors of biases to learn. While a vanilla MLP is capable of representing arbitrary functions, the real power of neural networks comes from more sophisticated architectures. Many of these architectures are designed to encode some particular *invariance* or *equivariance*—that is, when the input to the network is transformed in a particular way, the output should either be unchanged or should transform in a corresponding way. For instance, the weights in a layer of a convolutional neural network (ConvNet) (LeCun et al. 1998) are restricted to be a discrete convolution operator, which constrains each layer to be translation-equivariant, a natural constraint for image recognition, and also dramatically reduces the number of possible weights in a layer.

Equivariance to permutation is another frequently useful property, and one that is especially important in real-space approaches to representing electronic wavefunctions (see Section 4). A simple permutation-equivariant layer first proposed by Shawe-Taylor (1989) can be constructed by applying the same transformation to each input and summing the results. More sophisticated permutation-equivariant layers are used by models like the Transformer (Vaswani et al. 2017) or SchNet (Schütt et al. 2018). Many

of these equivariant layers can be unified in a conceptual framework based around the language of geometry and group theory, wherein the choice of transformation to be equivariant to leads naturally to recipes for constructing the appropriate neural network layers (Bronstein et al. 2021).

Another class of neural network architectures, which have been influential as wavefunction ansatzes, are restricted Boltzmann machines (RBMs) (Hinton & Salakhutdinov 2006). These were originally developed for unsupervised learning, but in the VMC setting considered here they lead to a simple deterministic expression for the log probability that closely resembles a one-layer MLP. Despite their early popularity, RBMs have been largely eclipsed in the AI community by other methods for unsupervised learning, such as variational autoencoders (Kingma & Welling 2013), generative adversarial networks (Goodfellow et al. 2014), normalizing flows (Rezende & Mohamed 2015), autoregressive models (van den Oord et al. 2016a,b), and diffusion models (Sohl-Dickstein et al. 2015). In fact, some of these newer models have started to have an impact as neural network wavefunction ansatzes for spin systems. Examples are deep autoregressive quantum states (Sharir et al. 2020), convolutional neural networks (Choo et al. 2019), recurrent neural networks (Hibat-Allah et al. 2020), and normalizing flows (Xie et al. 2022).

4 Electrons in first quantization

One approach to studying the electronic problem with deep learning is to work with parameterized many-body wavefunctions in first quantization, $\psi(\mathbf{r}; \boldsymbol{\theta})$. Here $\mathbf{r}$ stands for the N-tuple of electron coordinates, $\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N$, and sampling is realized over electronic positions $\mathbf{r}$ (Box 2). The antisymmetry constraint (Eq. 5) must be imposed in ψ to avoid collapsing onto a lower-energy bosonic state. A commonly adopted form is $\psi(\mathbf{r}; \boldsymbol{\theta}) = S(\mathbf{r}; \boldsymbol{\theta}) \times A(\mathbf{r}; \boldsymbol{\theta})$, where the first factor is symmetric (or "bosonic") under exchange of electron coordinates and the second factor carries the necessary antisymmetry. The simplest and most common approach is to build the antisymmetric part of the wavefunctions using Slater determinants (Eq. 6). As discussed in Section 2, single Slater determinants with fixed orbitals have limited expressiveness and many such determinants need to be combined to achieve high accuracy. A natural generalization of a sum of fixed-orbital Slater determinants is the commonly-used Slater–Jastrow wavefunction

$$\psi(\mathbf{r}; \boldsymbol{\theta}) = e^{J(\{\mathbf{r}\}; \boldsymbol{\theta})} \sum_k c_k \begin{vmatrix} \phi_1^k(\mathbf{r}_1; \boldsymbol{\theta}) & \cdots & \phi_1^k(\mathbf{r}_N; \boldsymbol{\theta}) \\ \vdots & \ddots & \vdots \\ \phi_N^k(\mathbf{r}_1; \boldsymbol{\theta}) & \cdots & \phi_N^k(\mathbf{r}_N; \boldsymbol{\theta}) \end{vmatrix} \quad (11)$$

where the Jastrow factor, $J(\{\mathbf{r}\}; \boldsymbol{\theta})$ constitutes the symmetric ("bosonic") part of the state and typically contains one- and two-body (and in many cases higher-order) parameterized correlations. The set notation, $\{\mathbf{r}\} \equiv \{\mathbf{r}_1, \ldots, \mathbf{r}_2\}$, indicates that J does not depend on the order of the electron coordinates. The determinants in Eq. (11) are typically replaced with the product of spin-up and spin-down determinants (Foulkes et al. 2001). Separating the up- and down-spin determinants improves computational efficiency, simplifies the implementation, and makes it easier to handle the electron-electron cusps, while leaving expectation values of spin-independent operators unchanged.
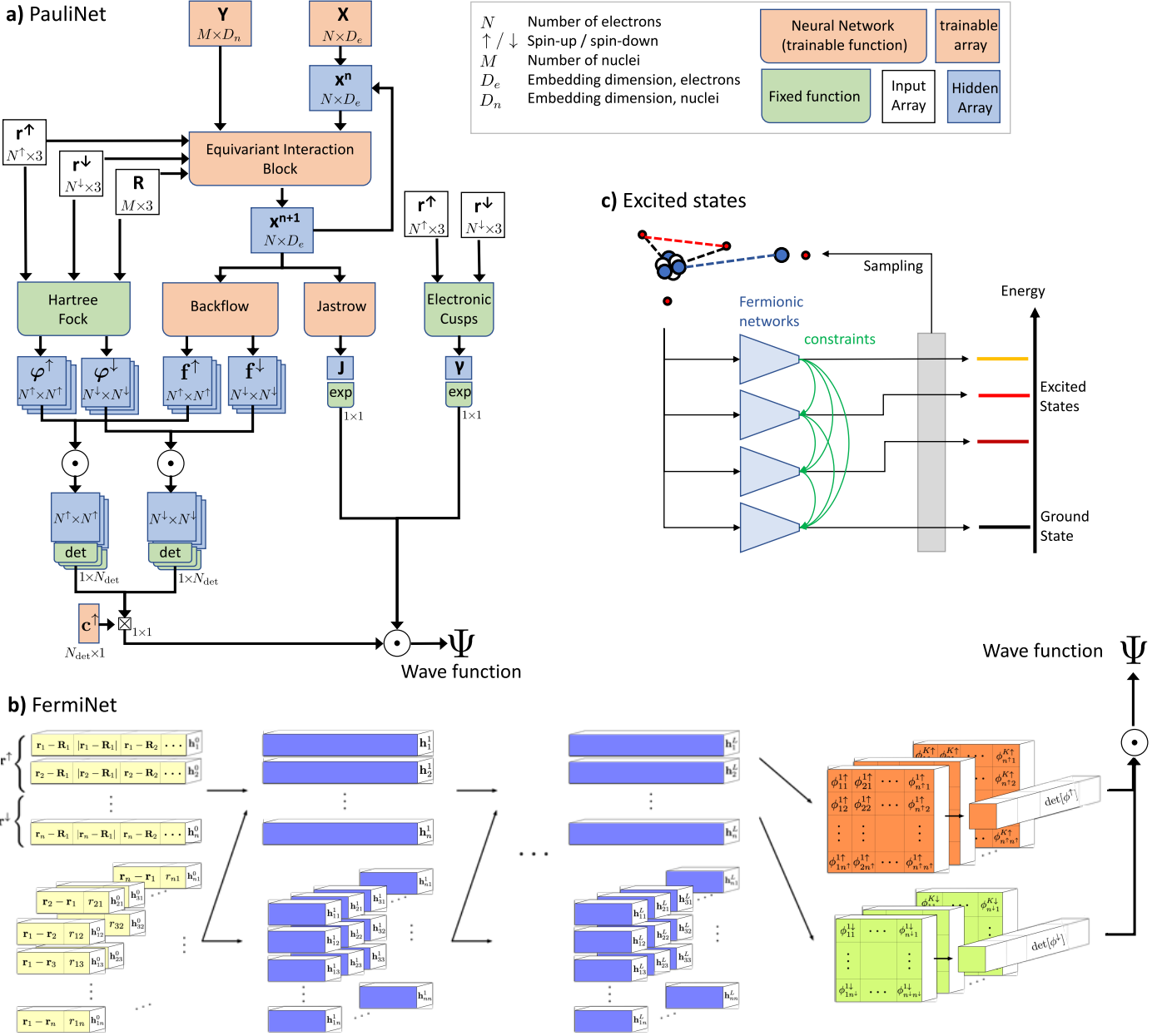
Fig. 4 | Neural-network architectures for selected real-space wavefunctions. (**a**) Original PauliNet architecture from (Hermann et al. 2020). (**b**) Original FermiNet architecture from (Pfau et al. 2020). Both architectures have been modified and extended by various contributions mentioned in this review. (**c**) Approach to computing exciting states in (Entwistle et al. 2022).

More flexible parametric forms can be obtained leveraging the approximation power of artificial neural networks. In the following, we discuss neural-network-based strategies to parameterize these forms.

4.1 Discrete space

The first applications of neural networks to electronic systems were for electrons moving in discretized space, as realized, for example, in the 2D Hubbard model of strongly-interacting electrons. In the following, for simplicity, we discuss the case of N spinless electrons in M lattice sites, and denote with $l(\mathbf{r}) \in [1, M]$ the discrete lattice index corresponding to electron position $\mathbf{r}$. The extension to the spinful case will be considered more in detail

when discussing continuous space later on. The symmetric part $S(\mathbf{r}; \boldsymbol{\theta})$ can be readily parameterized with a strategy closely related to NQS for spins:

$$S(\mathbf{r}; \boldsymbol{\theta}) = g(n(\mathbf{r}); \boldsymbol{\theta}), \tag{12}$$

where $n(\mathbf{r})$ is the unique occupation number representation corresponding to the electronic positions $\mathbf{r}$ and g represent a generic function which could be represented by a neural network. Since the occupation numbers $n(\mathbf{r})$ are invariant under permutation of the electron positions, $g(n(\mathbf{r}))$ is also symmetric under exchange. Any of the NN architectures also adopted for spin systems (Carleo & Troyer 2017) or lattice bosons (Saito 2017) can be used to represent the symmetric part S. Early works on the Hubbard model adopted positive-definite RBM-based parameteriza-

tions of $S(\mathbf{r}; \boldsymbol{\theta})$ (Nomura et al. 2017), while more recent works have adopted deep-network parameterizations allowing for sign changes (Stokes et al. 2020).

The simplest parameterization for the antisymmetric part, $A(\mathbf{r}; \boldsymbol{\theta})$, is again a Slater determinant

$$A(\mathbf{r}; \boldsymbol{\theta}) = \begin{vmatrix} \phi_1(\mathbf{r}_1; \boldsymbol{\theta}) & \cdots & \phi_1(\mathbf{r}_N; \boldsymbol{\theta}) \\ \vdots & \ddots & \vdots \\ \phi_N(\mathbf{r}_1; \boldsymbol{\theta}) & \cdots & \phi_N(\mathbf{r}_N; \boldsymbol{\theta}) \end{vmatrix}, \quad (13)$$

where the matrix $\Phi \in \mathbb{C}^{N \times M}$ of discrete orbitals $\phi_i(\mathbf{r}_j) = \Phi_{i,l(\mathbf{r}_j)}$ holds the variational parameters to be optimized. This approach, however, has the important drawback of not providing enough variational flexibility, since it effectively fixes the anti-symmetric part to a mean-field reference solution.

Neural backflow A significant improvement is obtained by considering a many-body backflow transformation of the orbitals (Feynman & Cohen 1956; Kwon et al. 1993). In this variational form, the matrix of one-electron orbitals Φ is promoted to a parameterized many-electron function depending on all the occupation numbers:

$$\tilde{\Phi}_{ij}(\boldsymbol{\theta}) = \Phi_{ij}(\boldsymbol{\theta}) + \Delta_{ij}(n(\mathbf{r}); \boldsymbol{\theta}), \quad (14)$$

where Δ is a correction to the single-particle orbitals Φ. In physics-inspired parameterizations, Δ is typically taken to be a simple function of the electronic occupation numbers (Tocchio et al. 2008). The neural backflow method (Luo & Clark 2019) instead introduced a flexible parameterization of the backflow orbitals based on artificial neural networks. In this case, Δ is parameterized with a MLP taking as inputs the electronic occupation numbers and outputting a many-body correction to the matrix Φ. This approach allows the orbitals to dynamically change depending on the positions of the electrons, thus allowing one to include genuinely many-body correlations in the antisymmetric part of the wavefunction.

Constrained hidden fermions Neural backflow transformations are not the only way to introduce flexible parameterizations of the antisymmetric part of the wavefunction. The constrained hidden fermion formalism builds on the idea of introducing a set of $\tilde{N}$ auxiliary fermionic particles, with positions $\mathbf{q}$, and living on $\tilde{M}$ lattice sites. These auxiliary particles are used to effectively mediate correlations among the physical degrees of freedom (Robledo Moreno et al. 2022). Calling $\tilde{A}(\mathbf{r}, \mathbf{q}; \boldsymbol{\theta})$ a Slater determinant for the extended (physical+hidden) system, the resulting antisymmetric form for the physical system is given by

$$A(\mathbf{r}; \boldsymbol{\theta}) = \tilde{A}(\mathbf{r}, F(\mathbf{r}; \boldsymbol{\theta})). \quad (15)$$

In this expression, F is a function, parameterized by a neural network, mapping the physical positions to the hidden ones. This approach has been shown to improve systematically over the neural backflow form for the 2D Hubbard model (Robledo Moreno et al. 2022).

4.2 Continuous space

We now focus on describing the important case of first-quantized electrons in continuous space, directly corresponding to the electronic Schrödinger equation. As in the discrete-space case, the Slater–Jastrow form may be improved in a matter suitable for use with neural quantum states by adding a backflow transformation, in which the one-electron orbitals $\phi_i(\mathbf{r}_j; \boldsymbol{\theta})$ are replaced by many-electron functions $\tilde{\phi}_i(\mathbf{r}_j, \{\mathbf{r}\}; \boldsymbol{\theta})$. The backflow transformation can either modify the orbitals directly via a multiplicative and/or additive term:

$$\tilde{\phi}_i(\mathbf{r}_j, \{\mathbf{r}\}; \boldsymbol{\theta}) = \phi_i(\mathbf{r}_j) f_i^{\otimes}(\mathbf{r}_j, \{\mathbf{r}\}; \boldsymbol{\theta}) + f_i^{\oplus}(\mathbf{r}_j, \{\mathbf{r}\}; \boldsymbol{\theta}), \quad (16)$$

or act as a quasiparticle transformation of the electron coordinates:

$$\tilde{\phi}_i(\mathbf{r}_j, \{\mathbf{r}\}; \boldsymbol{\theta}) = \phi_i(\mathbf{r}_j + \boldsymbol{\xi}(\{\mathbf{r}\}; \boldsymbol{\theta})), \quad (17)$$

where the paramterized functions, $f_i^{\otimes}, f_i^{\oplus}, \boldsymbol{\xi}$, are invariant to permutations of $\{\mathbf{r}\}$, and $\boldsymbol{\xi}(\{\mathbf{r}\}; \boldsymbol{\theta})$ is a three-component vector that modifies $\mathbf{r}_j$. If we consider a determinant of orbitals of this form,

$$\begin{vmatrix} \phi_1(\mathbf{r}_1; \{\mathbf{r}\}) & \cdots & \phi_1(\mathbf{r}_N; \{\mathbf{r}\}) \\ \vdots & \ddots & \vdots \\ \phi_N(\mathbf{r}_1; \{\mathbf{r}\}) & \cdots & \phi_N(\mathbf{r}_N; \{\mathbf{r}\}) \end{vmatrix}, \quad (18)$$

then we see that orbitals with backflow transformations are just one example of a broader class of functions: in order for the determinant to be antisymmetric, the matrix with elements $\Phi_{ij} = \phi_i(\mathbf{r}_j; \{\mathbf{r}\})$ must be permutation-equivariant; that is, exchanging electrons k and l also exchanges columns k and l. While traditional Slater–Jastrow–backflow wavefunctions have had considerable success, they also have limitations due to the choice of fixed functional forms. The goal, therefore, is to come up with more flexible permutation-equivariant functions. Here we highlight several approaches that share this common theme.

Iterative backflow Taddei et al. (2015) introduced a form of backflow that applied Eq. (17) repeatedly in an interative fashion. Such an ansatz is formally equivalent to expressing the backflow as a deep neural network (Ruggeri et al. 2018), albeit with artificial restriction on the dimensionality of the hidden layers. The iterative backflow was used for studying the ^{3}He and ^{4}He liquids

DeepWF The DeepWF (Han et al. 2019) approach uses an ansatz similar to a Slater–Jastrow wavefunction but with a simpler antisymmetric term:

$$\psi(\mathbf{r}) = S(\{\mathbf{r}\}, R) A^{\uparrow}(\mathbf{r}^{\uparrow}) A^{\downarrow}(\mathbf{r}^{\downarrow}). \quad (19)$$

The learned symmetric function S is similar to a Jastrow factor and ensures that the wavefunction captures the electron-nuclear and electron-electron cusp conditions. The antisymmetric factors A^{σ} are constructed from the Vandermonde-like determinant of an explicitly antisymmetric two-body function, $A^{\sigma} = \prod_{1 \le i \le j \le N} (a(\mathbf{r}_i, \mathbf{r}_j, r_{ij}) - a(\mathbf{r}_j, \mathbf{r}_i, r_{ij}))$. The two-body antisymmetric function is entirely learned. Such a functional form can be evaluated in $\mathcal{O}(N^2)$ operations, compared to $\mathcal{O}(N^3)$ for a determinant. However, the use of a simplified antisymmetric function is also likely to limit the accuracy achieved: DeepWF obtains only
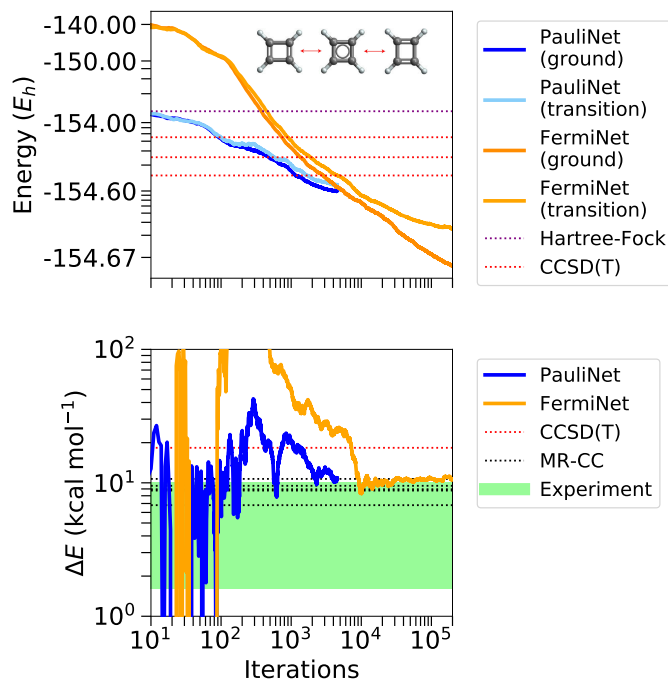
Fig. 5 | Automerization of cyclobutadiene with neural-network ansatzes. Both PauliNet and FermiNet predict relative energies within the range of experimental values and agree with multireference coupled cluster. The PauliNet converges more quickly, while the FermiNet reaches lower total energy. Figure modified from Spencer et al. (2020).

43.6% of the correlation energy for the beryllium atom and does not even reach HF accuracy for the boron atom. The PauliNet and FermiNet approaches described below do much better. Vanilla PauliNet obtained 99.94% and 97.3% of the correlation energies for the beryllium and boron atoms, and FermiNet 99.97% and 99.83%, respectively. Furthermore, FermiNet and PauliNet both substantially surpass conventional Slater-Jastrow-backflow (SJB) wavefunctions on first-row atoms, for which nearly exact benchmark values exist.

PauliNet PauliNet (Hermann et al. 2020) builds upon HF or CASSCF orbitals as a physically meaningful baseline and takes a neural network approach to the SJB wavefunction in order to correct this baseline towards a high-accuracy solution (Fig. 4a). Cusp conditions are explicitly met via the inclusion of cusp correction terms in the wavefunction (Ma et al. 2005). A graph-convolutional block based on SchNet (Schütt et al. 2018) is used to create a permutation-equivariant latent space representation depending on the many-electron configuration. This embedding is then passed into separate deep neural networks that learn the Jastrow factor and a (cuspless) backflow transformation. Hermann et al. (2020) introduced PauliNet with a purely multiplicative backflow as shown in Fig. 4a; Schätzle et al. (2021) generalized this to a multiplicative and additive backflow as shown in Eq. (16). PauliNet is optimized with a fixed number of Slater determinants. Most of the results reported in Hermann et al. (2020) and Schätzle et al. (2021) were obtained with around 10 determinants.

FermiNet FermiNet (Pfau et al. 2020) takes a more minimalist (or machine-learning maximalist) approach and attempts to train a neural network to represent the entire wavefunction (Fig. 4b). FermiNet uses two parallel networks, describing one- and two-electron features respectively. The inputs to each layer in the one-electron stream are permutation-equivariant functions of the activations from the previous layers of the one- and two-electron streams. The final layer projects the latent space into the required number of orbitals, from which determinants can be formed and evaluated. As with PauliNet, the final wavefunction is a sum over a number of determinants. For most of the results reported in Pfau et al. (2020), 16 determinants were used. FermiNet builds up a rich description of electron-electron interactions from the permutation-equivariant mixing of information describing one-and two-electron features. In particular, the electron-nuclear and electron-electron cusps in the wavefunction are represented accurately, despite not being encoded explicitly. Whereas PauliNet is usually trained with the ADAM optimizer, FermiNet training was found to be substantially improved when employing the KFAC optimizer.

While both PauliNet and FermiNet exceed the accuracy of conventional SJB wavefunctions on small systems, there are important tradeoffs between the two models. Results from both on the automerization of cyclobutadiene can be seen in Fig. 5. The FermiNet is typically trained with a larger number of parameters than the PauliNet, requiring more iterations and more computation per iteration to converge, but it typically converges to a lower absolute energy. Recently, Gerard et al. (2022) proposed a hybrid ansatz which uses neural network layers similar to the SchNet and PauliNet in a FermiNet-like architecture. This hybrid ansatz was found to reach even lower absolute energies than the FermiNet on systems like benzene and the potassium atom.

Potential energy surfaces Typically one optimises a wavefunction at a specific geometry but this quickly becomes prohibitively expensive for exploring the high-dimensional potential energy surface of even relatively small molecules. Scherbela et al. (2022) developed a training methodology that allows weight sharing between (simplified) PauliNet architectures targeting different geometries. By switching the geometry being trained at each epoch, they showed that the computational cost for training across a set of geometries can be improved by an order of magnitude without affecting the accuracy of the final energies, with 95% of network parameters shared across all geometries. This implies that the network is learning features of electron correlation in general rather than fitting to a specific geometry. They also demonstrated that a wavefunction for a larger molecule could be initialised from a wavefunction for a smaller molecule and could then be fine-tuned in a relatively short optimization stage. Pretraining neural network wavefunctions from smaller systems has also been shown to dramatically accelerate convergence for Kagome lattice models (Yang et al. 2020a).

In a similar vein, Gao & Günnemann (2022a,b) demonstrated that a meta-learning approach, where a graph neural network is used to parameterize a wavefunction model, can accurately represent the wavefunctions for multiple geometries, enabling a fully quantum-mechanical potential energy surface to be represented in a single model. Their approach used a FermiNet-like wavefunc-

tion model, but the meta-learning concept is directly applicable to other wavefunction representations, assuming the wavefunction form is sufficiently flexible.

Periodic systems There has also been progress on using first-quantized neural network architectures in periodic systems, such as interacting quantum gases in low dimension (Pescia et al. 2022), the electron gas (Cassella et al. 2022; Li et al. 2022b; Wilson et al. 2022), and for small cells of solids such as lithium hydride and graphene (Li et al. 2022b). Again, sufficiently expressive networks at the VMC level have been found capable of rivalling or surpassing the accuracy of fixed-node diffusion Monte Carlo calculations using conventional Slater-Jastrow-backflow trial wavefunctions.

4.3 Extensions

Pseudopotentials The electronic structure of heavy atoms, especially transition metals, is complicated and challenging for all QC methods. The difficulty is compounded by the high computational cost of variational Monte Carlo methods, which scale roughly as $\mathcal{O}(Z^5)$ (Hammond et al. 1987), where Z is the nuclear charge. Whilst the core electrons contribute heavily to the total energy, energy differences are largely determined by the behaviour of the valence electrons. The core electrons can therefore be removed and the effective nuclear charge reduced by the use of pseudopotentials. The use of pseudopotentials is common in many methods, including density functional theory and conventional variational Monte Carlo. Li et al. (2022a) demonstrate that effective core potentials can be readily combined with FermiNet and achieve accuracy comparable to CCSDT(Q) extrapolated to the complete basis set limit for first-row transition metal atoms. The computational time per iteration was reduced by 43% (17%) for the scandium (zinc) atom using an argon core. Again, this approach is not restricted to FermiNet. Pseudopotentials can be used with any first-quantized neural network wavefunction.

Diffusion Monte Carlo (DMC) Projector methods such as DMC (Needs et al. 2020) and auxiliary-field Monte Carlo (Shi & Zhang 2021) go beyond VMC by using stochastic algorithms to sample the ground state without requiring its wavefunction to be represented as a known function or network. DMC is in principle exact but, for many-fermion systems, relies in practice on the fixed-node approximation, in which collapse to the bosonic ground state is avoided by imposing the sign structure of the trial wavefunction on the DMC wavefunction. A DMC simulation therefore samples (stochastically) the lowest energy state with the same sign structure as the trial wavefunction. The improvements that result from applying DMC to conventional Slater-Jastrow-backflow trial functions optimized using VMC methods are substantial, explaining why DMC is so often used to provide improved estimates of the ground-state wavefunction and energy. Wilson et al. (2021) combined DMC with a FermiNet trial wavefunction. For first-row atoms, DMC captured much of the remaining correlation energy (94% of the difference between the VMC energy and the exact energy in the case of the nitrogen atom). However, Wilson et al. (2021) used a simplified FermiNet that gave VMC energies higher than those reported by Pfau et

al. (2020), which were already within 1mH of exact results for all first-row atoms. Given evidence that the mean-field equivalent of PauliNet can essentially match HF in the complete basis set limit (Schätzle et al. 2021), it is possible that the remaining error in PauliNet and FermiNet wavefunctions is dominated by errors in the nodal surface, which are rarely sampled regions during optimisation. If this is the case, diffusion Monte Carlo with the fixed node approximation may not produce substantially lower energies. On the other hand, since neural network wavefunctions routinely capture over 90% of the correlation energy at the VMC level, the need to perform expensive diffusion Monte Carlo calculations is greatly reduced. More recently, Ren et al. (2022) showed that DMC can capture roughly half of the remaining correlation energy for the atoms Li-Ar, when using a very small FermiNet-based architecture. Whilst it is possible to achieve energies within chemical accuracy using FermiNet at the VMC level, these calculations model the case for larger systems where converging the energy with respect to network size might not be feasible. Ren et al. (2022) went on to demonstrate that DMC using FermiNet trial wavefunctions noticeably reduces the energy for larger systems. In the case of the benzene dimer, the reduction was 50mH.

Excited States Our discussion so far, and most VMC calculations, have focused on ground state properties. However, excited states are of critical importance to understand the behaviour of materials. Fortunately, recent algorithmic developments by multiple groups have demonstrated that the calculation of excited states using VMC methods is feasible and can achieve an acceptable trade-off in accuracy and cost. Here we highlight three such approaches utilizing conventional VMC wavefunctions. One approach is the state-averaged VMC method (Dash et al. 2019; Schautz & Filippi 2004), in which the average energy over multiple states is minimised and individual states are projected out via diagonalization within the basis of excited states. Similar techniques are used with other quantum chemistry methods. Zhao & Neuscamman (2016) instead minimized a different objective function, such that the state with energy closest to a desired energy target is obtained. Pathak et al. (2021) suggested a simple alternative, where a state is forced to be (approximately) orthogonal to all lower energy states via a penalty term. These techniques can be readily applied to VMC using neural-network wavefunctions and, in particular, penalty function approaches have recently been explored. As with ground-state calculations, the flexibility of the wavefunction ansatz to represent the desired state is critical. Entwistle et al. (2022) demonstrated that the PauliNet architecture combined with a penalty function can represent the lowest few excited states of molecules up to the size of benzene (Fig. 4c). Relatedly, Choo et al. (2018) demonstrated that NQS on lattice models can obtain the lowest-energy state of any given Abelian symmetry by performing what is essentially a ground-state simulation in that symmetry sector, and multiple states of the same symmetry using a penalty function. However, the most accurate and efficient way to obtain excited states within VMC, irrespective of wavefunction ansatz, remains an open question (Cuzzocrea et al. 2020).
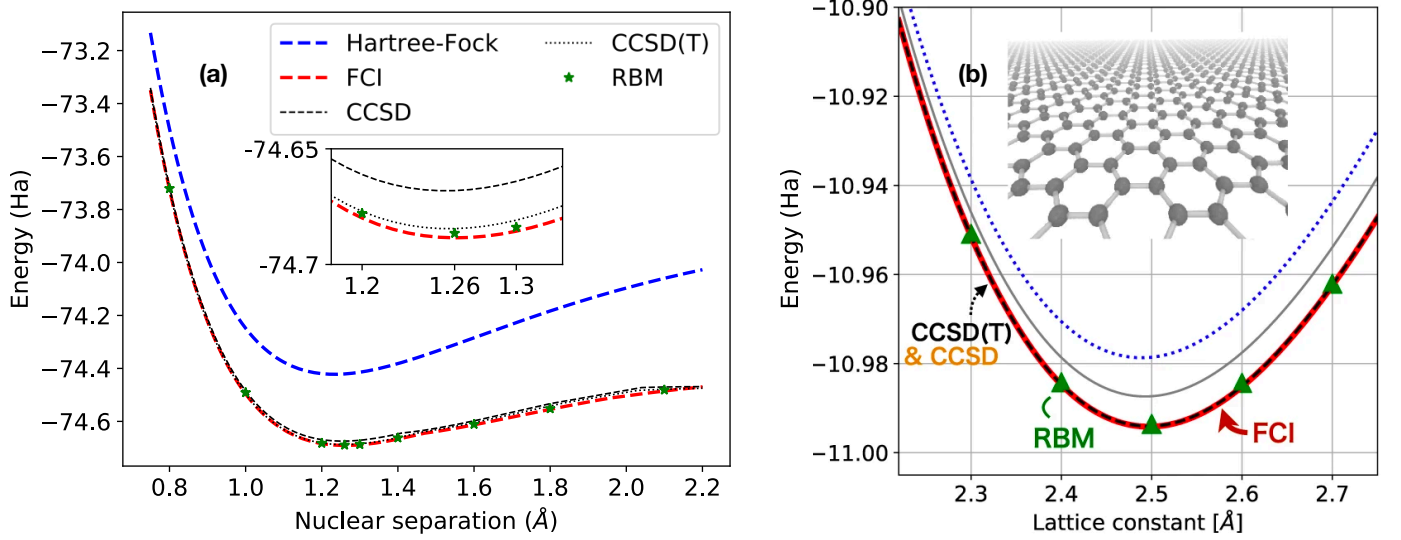
Fig. 6 | Electronic energies for molecules and solids in second quantization (a) Dissociation curve for C_2 molecule in the STO-3G basis. The green stars show results for a restricted Boltzmann machine which represents the electrons in discrete space. Figure taken from (Choo et al. 2020). (b) Graphene on a honeycomb lattice solved using the cc-pVDZ basis set. Figure taken from (Yoshioka et al. 2021).

5 Electrons in second quantization

Instead of working directly with the infinite-dimensional Hilbert space corresponding to the real-space Hamiltonian of Eq. (2), it is common practice in QC to use a finite basis set. By choosing a set of electronic basis functions $\{\varphi_1(\mathbf{r}), \varphi_2(\mathbf{r}), \dots\}$, we can define a set of second-quantised operators $\hat{c}_i^\dagger$ ($\hat{c}_i$) which create (annihilate) an electron in the i-th basis function, and which satisfy the canonical anticommutation relations $\{\hat{c}_i^\dagger, \hat{c}_j\} = \delta_{ij}$. These operators then act on the second-quantized wavefunction $\psi_{n_1 n_2 \dots}$, which encodes amplitudes for different occupations of the orbitals (Box 1). Projecting the real-space Hamiltonian onto this set of orbitals then yields the corresponding discretized Hamiltonian,

$$\hat{H} = \sum_{ij} t_{ij} \, \hat{c}_i^\dagger \hat{c}_j + \sum_{ijkl} u_{ijkl} \, \hat{c}_i^\dagger \hat{c}_j^\dagger \hat{c}_l \, \hat{c}_k, \qquad (20)$$

where

$$t_{ij} = \int \varphi_i^*(\mathbf{r}) \left(-\frac{1}{2}\nabla^2 - \sum_I \frac{Z_I}{|\mathbf{r} - \mathbf{R}_I|} \right) \varphi_j(\mathbf{r}) \, d\mathbf{r}, \qquad (21)$$

$$u_{ijkl} = \iint \varphi_i^*(\mathbf{r}) \varphi_j^*(\mathbf{r}') \frac{1}{|\mathbf{r} - \mathbf{r}'|} \varphi_k(\mathbf{r}) \varphi_l(\mathbf{r}') \, d\mathbf{r} \, d\mathbf{r}', \qquad (22)$$

are matrix elements of the one- and two-electron terms in the real-space Hamiltonian of Eq. (2). For simple basis functions such as Gaussians or plane waves, the matrix elements can be evaluated analytically. This Hamiltonian serves as the starting point for the methods described in this section.

5.1 Fermionic neural quantum states

Instead of working directly with the occupation-number representation of the wavefunction (Box 1), it is also possible to map occupation numbers $n_k \in \{0, 1\}$ onto degrees of freedom $\sigma_k^z \in \{\downarrow, \uparrow\}$ of spin-1/2 particles, such that empty orbitals map to down spins and occupied orbitals to up spins. This mapping makes it possible to leverage NQS and other methods for solving quantum spin

systems. The same duality allows the creation and annihilation operators appearing in the electronic Hamiltonian (Eq. 20) to be written in terms of spin operators. This can be achieved, for example, with the Jordan–Wigner mapping (Jordan & Wigner 1928), that transforms annihilation and creation operators into, respectively, lowering and raising spin operators $\hat{\sigma}_j^\pm = (\hat{\sigma}_j^x \pm i\hat{\sigma}_j^y)/2$. This mapping is not unique, however, and there exist more recent alternatives, such as parity or Bravyi–Kitaev encodings (Bravyi & Kitaev 2002), both of which have been developed in the context of quantum simulations.

Regardless of the choice of spin encoding, the final outcome is a spin Hamiltonian with the general form

$$\hat{H}_q = \sum_{p=1}^{r} h_p \hat{\Xi}_p, \qquad (23)$$

defined as a linear combination with real coefficients h_p of $\hat{\Xi}_p$, which are N-fold tensor products of single-qubit Pauli operators and the identity: $\hat{I}, \hat{\sigma}^x, \hat{\sigma}^y, \hat{\sigma}^z$.

The ground state of the spin Hamiltonian in Eq. (23) can be approximated using a spin-based NQS representation based on complex-valued RBMs (Carleo & Troyer 2017). For a system of N spins, the many-body amplitude corresponding to a state in the σ^z basis, i.e., $\boldsymbol{\sigma} = (\sigma_1^z \dots \sigma_N^z)$, takes the compact form

$$\psi(\boldsymbol{\sigma}; \boldsymbol{\theta}) = e^{\sum_i a_i \sigma_i^z} \prod_{j=1}^{M} 2\cosh\left(b_j + \sum_i^N W_{ij} \sigma_i^z \right), \qquad (24)$$

with parameters $\boldsymbol{\theta} = (a_i, b_j, W_{ij})$. This ansatz can be optimised with VMC techniques (Box 3), typically relying on the stochastic reconfiguration (Sorella 1998) approach. A number of works have adopted this approach and achieved competitive variational results for small basis sets (Choo et al. 2020; Yang et al. 2020b), even in conjunction with quantum computers (Iouchtchenko et al. 2022; Torlai et al. 2020). In Fig. 6 (a), we show the dissociation curve of C_2, in the STO-3G basis, using the RBM as described above (Choo et al. 2020).

Solids The second-quantization framework also allows one to treat solids, using as a basis the Bloch orbitals obtained by solving the crystalline HF equations (Del Re et al. 1967). Creation and annihiliation operators, $\hat{c}_{i\mathbf{k}}^{\dagger}$ and $\hat{c}_{i\mathbf{k}}$, for electrons in band i with crystal momentum k are introduced, and the resulting Hamiltonian is similar to Eq. (20), with the noticeable difference that the one- and two-body matrix elements now depend on the crystal momenta: $t_{ij} \rightarrow t_{ij}^{\mathbf{k}}$ and $u_{ijkl} \rightarrow u_{ijkl}^{\mathbf{k}_1\mathbf{k}_2\mathbf{k}_3\mathbf{k}_4}$, with the four momenta appearing in the two-body integrals satisfying the conservation of the total crystal momentum. Using Gaussian-based atomic functions as the single-particle basis and RBM wavefunctions to represent the many-body state, (Yoshioka et al. 2021) applied this approach to study the electronic structure of solids. In Fig. 6 (b), we show the computed ground-state energies for graphene crystals as a function of the lattice constant.

Exact Sampling Fermionic NQS are typically sampled using the MCMC approach commonly adopted in VMC (Box 2). However, the mixing rate of the MCMC algorithm is known to be slow in some cases, such as close to phase transitions, and MCMC simulations can suffer from critical slowing down. A way to circumvent this limitation is to introduce model wavefunctions explicitly designed to allow exact sampling of their square modulus, thus avoiding the need to use MCMC. One such family are autoregressive neural network wavefunctions (Sharir et al. 2020), a complex-valued generalization of the autoregressive models commonly adopted in deep learning. Such networks represent normalized wavefunctions and allow one to directly obtain perfectly uncorrelated samples; this is useful as the wavefunction distribution for many QC problems can be highly multi-modal. The exact sampling approach was applied to QC hamiltonians in a recent work by Barrett et al. (2022). Optimizations in the way Hamiltonian matrix elements and the corresponding Monte Carlo estimators are computed have made it possible to treat much larger systems than were accessible in the early applications of Choo et al. (2020). Specifically, Zhao et al. (2022)) obtain competitive variational energies, improving on the CCSD energies of molecules in minimal basis sets. Results for up to around 50 electrons in 80 orbitals (Na_2CO_3 at equilibrium) have been obtained at relatively modest computational cost.

5.2 ML-assisted selected CI

For many QC problems, although the dimension of the Hilbert space grows exponentially with system size, the number of relevant configurations in the ground state typically remains sparse. This suggests that by efficiently selecting the relevant configurations and then diagonalising the Hamiltonian on the reduced subspace, one can achieve highly accurate results. This set of approaches is also known as selected CI (Giner et al. 2013; Holmes et al. 2016; Huron et al. 1973; Sharma et al. 2017). Different flavours of selected CI vary in the way relevant configurations are selected.

One well-known approach is called Monte Carlo CI (MCCI) (Greer 1998) and can be briefly summarised as follows:
1. Start from a finite set of configurations $S_i = \{|x\rangle\}$
2. By considering single or double excitations starting from configurations in S_i, construct an expanded set S_i'.

3. Construct the Hamiltonian $\hat{H}_i$ for the expanded set S_i' and diagonalise to obtain the wavefunction coefficients for the configurations in the set.
4. Discard the configurations whose coefficient is less than a given threshold $c_{\min}$. The remaining configurations then form a new set of configurations S_{i+1}.
5. Repeat until convergence.

ML techniques can be used to improve selection of the configuration set. One such approach is to perform supervised learning (Coe 2018; Glielmo et al. 2020), where a neural network is trained to predict the wavefunction coefficients using the data from the MCCI method, i.e., the wavefunction coefficients of the configurations in the set S_i'. After training, the network can be queried or sampled to select the configurations with the largest coefficients. In other words, the network is used to bootstrap and predict the coefficients of configurations not yet seen in the data set. It was shown in Coe (2018) that such an approach converges faster than the vanilla MCCI method.

The task of selecting configurations for selected CI can also be cast as a reinforcement-learning task where the state is the current set of configurations and an agent is trained to perform actions on the set to iteratively modify the configurations with the aim of minimising the variational energy. This approach was applied in Goings et al. (2021) to achieve near-FCI accuracy for small molecules in a small basis set.

6 Challenges and outlook

Ab-initio QC with neural-network wavefunctions has only just emerged as a viable path to highly accurate electronic-structure methods, yet it already competes with established approaches that have been developed for decades. We imagine that it may become the methodology with the best trade-off between efficiency and accuracy for systems with up to one to two hundred electrons and a nontrivial electronic structure. Before that can happen, however, several challenges must be addressed. All the methods are currently in a development stage and only limited benchmarking is available. As such, it is not yet clear whether the excellent accuracy seen so far will be maintained across a broader range of chemical systems, or how rapidly the accuracy will degrade with system size. Related to this is our incomplete understanding of what limits the accuracy of neural-network ansatzes, and how their success or failure is related to physical phenomena such as strong correlation. Since the underlying electronic problem is exponentially hard but the algorithms are polynomial, they must be limited in accuracy in some ways. It is not currently clear, however, whether the limitations seen to date are caused by the restricted expressiveness of the neural networks or by difficulties in optimization or both. For instance, while it has been proven that a single generalized Slater determinant is in principle sufficient to represent any antisymmetric function (Hutter 2020), it might not be possible to parametrize it with a polynomially scaling neural network or train it within a polynomially scaling time.

Apart from these fundamental issues, there are many practical challenges. While the scaling of variational QMC with system size is favourable, the prefactor due to the neural networks is large. Until very recently, this limited applications to systems no larger than the benzene molecule (42 electrons), which is three

to four times below our envisaged applicability range, although results for a 108-electron simulation cell of solid LiH have now been reported (Li et al. 2022b). The prefactor can be reduced by integrating traditional QC techniques such as pseudopotentials (Li et al. 2022a), developing more efficient neural-network architectures, or using ML techniques such as pre-training and transfer learning. Specific to the discrete-basis second-quantized approaches is the issue of basis-set convergence, where sufficiently large basis sets may increase the prefactor by up to three orders of magnitude compared to minimal basis sets. Another challenge is related to the stochastic optimization, which produces noise in the converged energies that is especially amplified when calculating small energy differences.

We are, however, optimistic that many of these challenges can be addressed and can be addressed quickly, thanks to the relative simplicity of the framework based on variational QMC and of neural networks compared to traditional QC approaches. Indeed, this simplicity has already enabled rapid development of multiple extensions to the first single-point ground-state calculations on molecules, including transferable wavefunctions, excited states, and formulations for periodic systems, all originating from multiple independent research groups.

First-quantized approaches such as FermiNet, PauliNet, and their successor architectures already match essentially exact benchmark results to within chemical accuracy for small systems. Yet these networks are just a small subset of possible architectures for representing antisymmetric wavefunctions, and it is unlikely that the optimal ones were found on the first attempt, so we expect that significant innovation lies ahead. We believe that ab-initio methods based on neural-network wavefunctions will become an integral part of the QC toolbox that enables straightforward electronic-structure calculations of complex molecular systems.

References

Abadi, M. et al. (2016). "TensorFlow: A System for Large-Scale Machine Learning." In *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation* (USENIX Association), 265–283. URL: https://www.usenix.org/conference/osdi16/technical-sessions/presentation/abadi.

Acevedo, A., M. Curry, S. H. Joshi, B. Leroux & N. Malaya (2020). "Vandermonde Wave Function Ansatz for Improved Variational Monte Carlo." In *2020 IEEE/ACM Fourth Workshop on Deep Learning on Supercomputers (DLS)* (IEEE), 40–47. DOI: 10/gm7822.

Adams, C., G. Carleo, A. Lovato & N. Rocco (2021). "Variational Monte Carlo Calculations of A≤4 Nuclei with an Artificial Neural-Network Correlator Ansatz." *Phys. Rev. Lett.* 127, 022502. DOI: 10/gpz4r3.

Amari, S. (1998). "Natural Gradient Works Efficiently in Learning." *Neural Comput.* 10, 251. DOI: 10/dtsdwt.

Astrakhantsev, N., T. Westerhout, A. Tiwari, K. Choo, A. Chen, M. H. Fischer, G. Carleo & T. Neupert (2021). "Broken-Symmetry Ground States of the Heisenberg Model on the Pyrochlore Lattice." *Phys. Rev. X* 11, 041021. DOI: 10/gpz4r2.

Ay, N., J. Jost, H. V. Lê & L. Schwachhöfer (2017). *Information Geometry* (Springer). DOI: 10/h83m.

Bajdich, M., L. Mitas, G. Drobný, L. K. Wagner & K. E. Schmidt (2006). "Pfaffian Pairing Wave Functions in Electronic-Structure Quantum Monte Carlo Simulations." *Phys. Rev. Lett.* 96, 130201. DOI: 10/dms82t.

Barrett, T. D., A. Malyshev & A. I. Lvovsky (2022). "Autoregressive Neural-Network Wavefunctions for Ab Initio Quantum Chemistry." *Nat. Mach. Intell.* 4, 351. DOI: 10/gp6r33.

Bartlett, R. J. & M. Musiał (2007). "Coupled-Cluster Theory in Quantum Chemistry." *Rev. Mod. Phys.* 79, 291. DOI: 10/fpfwfb.

Becca, F. & S. Sorella (2017). *Quantum Monte Carlo Approaches for Correlated Systems* 1st ed. (Cambridge Univ. Press). DOI: 10/h83n.

Behler, J. (2021). "Four Generations of High-Dimensional Neural Network Potentials." *Chem. Rev.* 121, 10037. DOI: 10/gjxh9z.

Behler, J. & M. Parrinello (2007). "Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces." *Phys. Rev. Lett.* 98, 146401. DOI: 10/c7kbsq.

Bian, Y. & X.-Q. Xie (2021). "Generative Chemistry: Drug Discovery with Deep Learning Generative Models." *J. Mol. Model.* 27, 71. DOI: 10/gkfx6d.

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.

Bottou, L. & O. Bousquet (2008). "Learning Using Large Datasets." In *Mining Massive Data Sets for Security* (IOS Press), 15–26. DOI: 10/hr97.

— (2011). "The Tradeoffs of Large-Scale Learning." In *Optimization for Machine Learning* (MIT Press), 351–368. URL: https://mitpress.mit.edu/9780262016469/optimization-for-machine-learning/.

Bradbury, J., R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin & S. Wanderman-Milne (2018). *JAX: Composable Transformations of Python+NumPy Programs.* URL: http://github.com/google/jax.

Bravyi, S. & A. Kitaev (2002). "Fermionic Quantum Computation." *Ann. Phys.* 298, 210. DOI: 10/cx53kk.

Bronstein, M. M., J. Bruna, T. Cohen & P. Veličković (2021). "Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges." arXiv: 2104.13478.

Carleo, G., I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto & L. Zdeborová (2019). "Machine Learning and the Physical Sciences." *Rev. Mod. Phys.* 91, 045002. DOI: 10/ggd5qv.

Carleo, G. & M. Troyer (2017). "Solving the Quantum Many-Body Problem with Artificial Neural Networks." *Science* 355, 602. DOI: 10/f9qd53.

Cassella, G., H. Sutterud, S. Azadi, N. D. Drummond, D. Pfau, J. S. Spencer & W. M. C. Foulkes (2022). "Discovering Quantum Phase Transitions with Fermionic Neural Networks." arXiv: 2202.05183.

Chan, G. K.-L. & S. Sharma (2011). "The Density Matrix Renormalization Group in Quantum Chemistry." *Annu. Rev. Phys. Chem.* 62, 465. DOI: 10/bkvrzh.

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., G. Carleo, N. Regnault & T. Neupert (2018). "Symmetries and Many-Body Excitations with Neural-Network Quantum States." *Phys. Rev. Lett.* 121, 167204. DOI: 10/gfgtrg.

Choo, K., A. Mezzacapo & G. Carleo (2020). "Fermionic Neural-Network States for Ab-Initio Electronic Structure." *Nat. Commun.* 11, 2368. DOI: 10/ghmzr5.

Choo, K., T. Neupert & G. Carleo (2019). "Two-Dimensional Frustrated J_1-J_2 Model Studied with Neural Network Quantum States." *Phys. Rev. B* 100, 125124. DOI: 10/ghmzrr.

Choromanska, A., M. Henaff, M. Mathieu, G. B. Arous & Y. LeCun (2015). "The Loss Surfaces of Multilayer Networks." In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics* (PMLR), 192–204. URL: https://proceedings.mlr.press/v38/choromanska15.html.

Coe, J. P. (2018). "Machine Learning Configuration Interaction." *J. Chem. Theory Comput.* 14, 5739. DOI: 10/gfpc2m.

Cuzzocrea, A., A. Scemama, W. J. Briels, S. Moroni & C. Filippi (2020). "Variational Principles in Quantum Monte Carlo: The Troubled Story of Variance Minimization." *J. Chem. Theory Comput.* 16, 4203. DOI: 10/gqpvn2.

Dash, M., J. Feldt, S. Moroni, A. Scemama & C. Filippi (2019). "Excited States with Selected Configuration Interaction-Quantum Monte Carlo: Chemically Accurate Excitation Energies and Geometries." *J. Chem. Theory Comput.* 15, 4896. DOI: 10/gqpvn4.

Dauphin, Y., R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli & Y. Bengio (2014). "Identifying and Attacking the Saddle Point Problem in High-Dimensional Non-Convex Optimization." In *Advances in Neural Information Processing Systems 27* (Curran Associates), 2933–2941. URL: https://papers.nips.cc/paper/2014/hash/17e23e50bedc63b4095e3d8204ce063b-Abstract.html.

Degrave, J. et al. (2022). "Magnetic Control of Tokamak Plasmas through Deep Reinforcement Learning." *Nature* 602, 414. DOI: 10/gpg9dn.

Del Re, G., J. Ladik & G. Biczó (1967). "Self-Consistent-Field Tight-Binding Treatment of Polymers. I. Infinite Three-Dimensional Case." *Phys. Rev.* 155, 997. DOI: 10/czgzkm.

Deringer, V. L., A. P. Bartók, N. Bernstein, D. M. Wilkins, M. Ceriotti & G. Csányi (2021a). "Gaussian Process Regression for Materials and Molecules." *Chem. Rev.* 121, 10073. DOI: 10/gmkffv.

Deringer, V. L., N. Bernstein, G. Csányi, C. Ben Mahmoud, M. Ceriotti, M. Wilson, D. A. Drabold & S. R. Elliott (2021b). "Origins of Structural and Electronic Transitions in Disordered Silicon." *Nature* **589**, 59. DOI: 10/ghsb84.

Dral, P. O. (2020). "Quantum Chemistry in the Age of Machine Learning." *J. Phys. Chem. Lett.* **11**, 2336. DOI: 10/gg2bcq.

Drautz, R. (2019). "Atomic Cluster Expansion for Accurate and Transferable Interatomic Potentials." *Phys. Rev. B* **99**, 014104. DOI: 10/ghdhfv.

Entwistle, M., Z. Schätzle, P. A. Erdman, J. Hermann & F. Noé (2022). "Electronic Excited States in Deep Variational Monte Carlo." arXiv: 2203.09472.

Feynman, R. P. & M. Cohen (1956). "Energy Spectrum of the Excitations in Liquid Helium." *Phys. Rev.* **102**, 1189. DOI: 10/c77rcn.

Foulkes, W. M. C., L. Mitas, R. J. Needs & G. Rajagopal (2001). "Quantum Monte Carlo Simulations of Solids." *Rev. Mod. Phys.* **73**, 33. DOI: 10/dmgh76.

Gao, N. & S. Günnemann (2022a). "Ab-Initio Potential Energy Surfaces by Pairing GNNs with Neural Wave Functions." In *International Conference on Learning Representations (ICLR 2022)* (OpenReview). URL: https://openreview.net/forum?id=apv504XsysP.

— (2022b). "Sampling-Free Inference for Ab-Initio Potential Energy Surface Networks." arXiv: 2205.14962.

Gerard, L., M. Scherbela, P. Marquetand & P. Grohs (2022). "Gold-Standard Solutions to the Schrödinger Equation Using Deep Learning: How Much Physics Do We Need?" arXiv: 2205.09438.

Giner, E., A. Scemama & M. Caffarel (2013). "Using Perturbatively Selected Configuration Interaction in Quantum Monte Carlo Calculations." *Can. J. Chem.* **91**, 879. DOI: 10/gffmj3.

Glielmo, A., Y. Rath, G. Csányi, A. De Vita & G. H. Booth (2020). "Gaussian Process States: A Data-Driven Representation of Quantum Many-Body Physics." *Phys. Rev. X* **10**, 041026. DOI: 10/gmjhdq.

Glorot, X. & Y. Bengio (2010). "Understanding the Difficulty of Training Deep Feedforward Neural Networks." In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics* (PMLR), 249–256. URL: https://proceedings.mlr.press/v9/glorot10a.html.

Goings, J. J., H. Hu, C. Yang & X. Li (2021). "Reinforcement Learning Configuration Interaction." *J. Chem. Theory Comput.* **17**, 5482. DOI: 10/gpz4rz.

Gómez-Bombarelli, R., J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams & A. Aspuru-Guzik (2018). "Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules." *ACS Cent. Sci.* **4**, 268. DOI: 10/ch82.

Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville & Y. Bengio (2014). "Generative Adversarial Nets." In *Advances in Neural Information Processing Systems 27* (Curran Associates), 2672–2680. URL: https://papers.nips.cc/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html.

Greer, J. (1998). "Monte Carlo Configuration Interaction." *J. Comput. Phys.* **146**, 181. DOI: 10/bckb4c.

Hammond, B. L., P. J. Reynolds & W. A. Lester (1987). "Valence Quantum Monte Carlo with Ab Initio Effective Core Potentials." *J. Chem. Phys.* **87**, 1130. DOI: 10/b59sdb.

Han, J., L. Zhang & W. E (2019). "Solving Many-Electron Schrödinger Equation Using Deep Neural Networks." *J. Comput. Phys.* **399**, 108929. DOI: 10/gmg56w.

Heinrich, J., M. Lanctot & D. Silver (2015). "Fictitious Self-Play in Extensive-Form Games." In *Proceedings of the 32nd International Conference on Machine Learning* (PMLR), 805–813. URL: https://proceedings.mlr.press/v37/heinrich15.html.

Hermann, J., Z. Schätzle & F. Noé (2020). "Deep-Neural-Network Solution of the Electronic Schrödinger Equation." *Nat. Chem.* **12**, 891. DOI: 10/ghcm5p.

Hibat-Allah, M., M. Ganahl, L. E. Hayward, R. G. Melko & J. Carrasquilla (2020). "Recurrent Neural Network Wave Functions." **2**, 023358. DOI: 10/ghmzrq.

Hinton, G. E. & R. R. Salakhutdinov (2006). "Reducing the Dimensionality of Data with Neural Networks." *Science* **313**, 504. DOI: 10/b7vgmj.

Holmes, A. A., N. M. Tubman & C. Umrigar (2016). "Heat-Bath Configuration Interaction: An Efficient Selected Configuration Interaction Algorithm Inspired by Heat-Bath Sampling." *J. Chem. Theory Comput.* **12**, 3674. DOI: 10/f82q4p.

Hoogeboom, E., V. G. Satorras, C. Vignac & M. Welling (2022). "Equivariant Diffusion for Molecule Generation in 3D." arXiv: 2203.17003.

Hooker, S. (2020). "The Hardware Lottery." arXiv: 2009.06489.

Hornik, K., M. Stinchcombe & H. White (1989). "Multilayer Feedforward Networks Are Universal Approximators." *Neural Netw.* **2**, 359. DOI: 10/c4wch5.

Huang, B. & O. A. von Lilienfeld (2020). "Quantum Machine Learning Using Atom-in-Molecule-Based Fragments Selected on the Fly." *Nat. Chem.* **12**, 945. DOI: 10/ghqj49.

Huron, B., J. P. Malrieu & P. Rancurel (1973). "Iterative Perturbation Calculations of Ground and Excited State Energies from Multiconfigurational Zeroth-order Wavefunctions." *J. Chem. Phys.* **58**, 5745. DOI: 10/czvzff.

Hutter, M. (2020). "On Representing (Anti)Symmetric Functions." arXiv: 2007.15298.

Iouchtchenko, D., J. F. Gonthier, A. Perdomo-Ortiz & R. G. Melko (2022). "Neural Network Enhanced Measurement Efficiency for Molecular Groundstates." arXiv: 2206.15449.

Jones, R. O. (2015). "Density Functional Theory: Its Origins, Rise to Prominence, and Future." *Rev. Mod. Phys.* **87**, 897. DOI: 10/f7phd9.

Jordan, P. & E. Wigner (1928). "Über Das Paulische Äquivalenzverbot." *Z. Phys.* **47**, 631. DOI: 10/d4dxmd.

Jumper, J., R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al. (2021). "Highly Accurate Protein Structure Prediction with AlphaFold." *Nature* **596**, 583. DOI: 10/gk7nfp.

Kingma, D. P. & M. Welling (2013). "Auto-Encoding Variational Bayes." arXiv: 1312.6114.

Kirkpatrick, J. et al. (2021). "Pushing the Frontiers of Density Functionals by Solving the Fractional Electron Problem." *Science* **374**, 1385. DOI: 10/gnqwfr.

Krizhevsky, A., I. Sutskever & G. E. Hinton (2012). "Imagenet Classification with Deep Convolutional Neural Networks." In *Advances in Neural Information Processing Systems 25* (Curran Associates), 1097–1105. URL: https://papers.nips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html.

Kwon, Y., D. M. Ceperley & R. M. Martin (1993). "Effects of Three-Body and Backflow Correlations in the Two-Dimensional Electron Gas." *Phys. Rev. B* **48**, 12037. DOI: 10/bnjx23.

LeCun, Y., Y. Bengio & G. Hinton (2015). "Deep Learning." *Nature* **521**, 436. DOI: 10/bmqp.

LeCun, Y., L. Bottou, Y. Bengio & P. Haffner (1998). "Gradient-Based Learning Applied to Document Recognition." *Proc. IEEE* **86**, 2278. DOI: 10/d89c25.

Li, H., C. Collins, M. Tanha, G. J. Gordon & D. J. Yaron (2018). "A Density Functional Tight Binding Layer for Deep Learning of Chemical Hamiltonians." *J. Chem. Theory Comput.* **14**, 5764. DOI: 10/gfqk9f.

Li, X., C. Fan, W. Ren & J. Chen (2022a). "Fermionic Neural Network with Effective Core Potential." *Phys. Rev. Research* **4**, 013021. DOI: 10/gpz4rx.

Li, X., Z. Li & J. Chen (2022b). "Ab Initio Calculation of Real Solids via Neural Network Ansatz." arXiv: 2203.15472v2.

Linnainmaa, S. (1970). "The Representation of the Cumulative Rounding Error of an Algorithm as a Taylor Expansion of the Local Rounding Errors." MA thesis. University of Helsinki.

— (1976). "Taylor Expansion of the Accumulated Rounding Error." *BIT Numer. Math.* **16**, 146. DOI: 10/cv4d63.

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.

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.

Martens, J. & R. Grosse (2015). "Optimizing Neural Networks with Kronecker-Factored Approximate Curvature." In *Proceedings of the 32nd International Conference on Machine Learning* (PMLR), 2408–2417. URL: https://proceedings.mlr.press/v37/martens15.html.

McCulloch, W. S. & W. Pitts (1943). "A Logical Calculus of the Ideas Immanent in Nervous Activity." *Bull. Math. Biophys.* **5**, 115. DOI: 10/djsbj6.

Mnih, V., K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al. (2015). "Human-Level Control through Deep Reinforcement Learning." *Nature* **518**, 529. DOI: 10/gc3h75.

Musil, F., A. Grisafi, A. P. Bartók, C. Ortner, G. Csányi & M. Ceriotti (2021). "Physics-Inspired Structural Representations for Molecules and Materials." *Chem. Rev.* **121**, 9759. DOI: 10/gmmn9f.

Needs, R. J., M. D. Towler, N. D. Drummond, P. López Ríos & J. R. Trail (2020). "Variational and Diffusion Quantum Monte Carlo Calculations with the CASINO Code." *J. Chem. Phys.* **152**, 154106. DOI: 10/gm783w.

Neuscamman, E. (2013). "The Jastrow Antisymmetric Geminal Power in Hilbert Space: Theory, Benchmarking, and Application to a Novel Transition State." *J. Chem. Phys.* **139**, 194105. DOI: 10/gkr59c.

Noé, F., S. Olsson, J. Köhler & H. Wu (2019). "Boltzmann Generators: Sampling Equilibrium States of Many-Body Systems with Deep Learning." *Science* **365**, eaaw1147. DOI: 10/gf7qsx.

Noé, F., A. Tkatchenko, K.-R. Müller & C. Clementi (2020). "Machine Learning for Molecular Simulation." *Annu. Rev. Phys. Chem.* **71**, 361. DOI: 10/ggz88h.

Nomura, Y., A. S. Darmawan, Y. Yamaji & M. Imada (2017). "Restricted Boltzmann Machine Learning for Solving Strongly Correlated Quantum Systems." *Phys. Rev. B* **96**, 205152. DOI: 10/gfxfv6.

Olsen, J. (2011). "The CASSCF Method: A Perspective and Commentary: CASSCF Method." *Int. J. Quantum Chem.* **111**, 3267. DOI: 10/cw4pgx.

Paszke, A., S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga & A. Lerer (2017). "Automatic Differentiation in PyTorch." URL: https://openreview.net/forum?id=BJJsrmfCZ.

Pathak, S., B. Busemeyer, J. N. B. Rodrigues & L. K. Wagner (2021). "Excited States in Variational Monte Carlo Using a Penalty Method." *J. Chem. Phys.* **154**, 034101. DOI: 10/gm9xw2.

Perronnin, F., Y. Liu, J. Sánchez & H. Poirier (2010). "Large-Scale Image Retrieval with Compressed Fisher Vectors." In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (IEEE), 3384–3391. DOI: 10/bzns35.

Pescia, G., J. Han, A. Lovato, J. Lu & G. Carleo (2022). "Neural-Network Quantum States for Periodic Systems in Continuous Space." *Phys. Rev. Res.* **4**, 023138. DOI: 10/gqpvnw.

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. (2020). *Ideas of Quantum Chemistry* 3rd ed. (Elsevier). DOI: 10/h83r.

Ren, W., W. Fu & J. Chen (2022). "Towards the Ground State of Molecules via Diffusion Monte Carlo on Neural Networks." arXiv: 2204.13903.

Rezende, D. & S. Mohamed (2015). "Variational Inference with Normalizing Flows." In *Proceedings of the 32nd International Conference on Machine Learning* (PMLR), 1530–1538. URL: https://proceedings.mlr.press/v37/rezende15.html.

Robledo Moreno, J., G. Carleo, A. Georges & J. Stokes (2022). "Fermionic Wave Functions from Neural-Network Constrained Hidden States." *Proc. Natl. Acad. Sci.* **119**, e2122059119. DOI: 10/gqpvnx.

Rosenblatt, F. (1958). "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain." *Psychol. Rev.* **65**, 386. DOI: 10/fg6wr5.

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/gdg64n.

Rumelhart, D. E., G. E. Hinton & R. J. Williams (1986). "Learning Representations by Back-Propagating Errors." *Nature* **323**, 533. DOI: 10/cvjdpk.

Russakovsky, O., J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. (2015). "Imagenet Large Scale Visual Recognition Challenge." *Int. J. Comput. Vis.* **115**, 211. DOI: 10/gcgk7w.

Sabzevari, I. & S. Sharma (2018). "Improved Speed and Scaling in Orbital Space Variational Monte Carlo." *J. Chem. Theory Comput.* **14**, 6276. DOI: 10/gqpjjf.

Saito, H. (2017). "Solving the Bose–Hubbard Model with Machine Learning." *J. Phys. Soc. Jpn.* **86**, 093001. DOI: 10/gbv224.

Schätzle, Z., J. Hermann & F. Noé (2021). "Convergence to the Fixed-Node Limit in Deep Variational Monte Carlo." *J. Chem. Phys.* **154**, 124108. DOI: 10/gm7xvm.

Schautz, F. & C. Filippi (2004). "Optimized Jastrow–Slater Wave Functions for Ground and Excited States: Application to the Lowest States of Ethene." *J. Chem. Phys.* **120**, 10931. DOI: 10/cn3tqr.

Scherbela, M., R. Reisenhofer, L. Gerard, P. Marquetand & P. Grohs (2022). "Solving the Electronic Schrödinger Equation for Multiple Nuclear Geometries with Weight-Sharing Deep Neural Networks." *Nat. Comput. Sci.* **2**, 331. DOI: 10/gqpvzj.

Schmidhuber, J. (2015). "Deep Learning in Neural Networks: An Overview." *Neural Netw.* **61**, 85. DOI: 10/f6v78n.

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.

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., F. Arbabzadah, S. Chmiela, K. R. Müller & A. Tkatchenko (2017). "Quantum-Chemical Insights from Deep Tensor Neural Networks." *Nat. Commun.* **8**, 13890. DOI: 10/f9kzdv.

Sharir, O., Y. Levine, N. Wies, G. Carleo & A. Shashua (2020). "Deep Autoregressive Models for the Efficient Variational Simulation of Many-Body Quantum Systems." *Phys. Rev. Lett.* **124**, 020503. DOI: 10/ggw9km.

Sharma, S., A. A. Holmes, G. Jeanmairet, A. Alavi & C. J. Umrigar (2017). "Semistochastic Heat-Bath Configuration Interaction Method: Selected Configuration Interaction with Semistochastic Perturbation Theory." *J. Chem. Theory Comput.* **13**, 1595. DOI: 10/f9wgq3.

Shawe-Taylor, J. (1989). "Building Symmetries into Feedforward Networks." In *1989 First IEE International Conference on Artificial Neural Networks* (IET), 158–162. URL: https://ieeexplore.ieee.org/document/51951.

Shi, H. & S. Zhang (2021). "Some Recent Developments in Auxiliary-Field Quantum Monte Carlo for Real Materials." *J. Chem. Phys.* **154**, 024107. DOI: 10/gptz6w.

Silver, D., A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. (2016). "Mastering the Game of Go with Deep Neural Networks and Tree Search." *Nature* **529**, 484. DOI: 10/f77tw6.

Silver, D. et al. (2018). "A General Reinforcement Learning Algorithm That Masters Chess, Shogi, and Go through Self-Play." *Science* **362**, 1140. DOI: 10/cxq3.

Sohl-Dickstein, J., E. Weiss, N. Maheswaranathan & S. Ganguli (2015). "Deep Unsupervised Learning Using Nonequilibrium Thermodynamics." In *Proceedings of the 32nd International Conference on Machine Learning* (PMLR), 2256–2265. URL: https://proceedings.mlr.press/v37/sohl-dickstein15.html.

Sorella, S. (1998). "Green Function Monte Carlo with Stochastic Reconfiguration." *Phys. Rev. Lett.* **80**, 4558. DOI: 10/dh43x9.

Spencer, J. S., D. Pfau, A. Botev & W. M. C. Foulkes (2020). "Better, Faster Fermionic Neural Networks." arXiv: 2011.07125.

Stokes, J., J. Izaac, N. Killoran & G. Carleo (2020). "Quantum Natural Gradient." *Quantum* **4**, 269. DOI: 10/ghpb9k.

Sutton, R. S. & A. G. Barto (2018). *Reinforcement Learning: An Introduction* (MIT press).

Szabo, A. & N. S. Ostlund (1996). *Modern Quantum Chemistry* (Dover Publications).

Taddei, M., M. Ruggeri, S. Moroni & M. Holzmann (2015). "Iterative Backflow Renormalization Procedure for Many-Body Ground-State Wave Functions of Strongly Interacting Normal Fermi Liquids." *Phys. Rev. B* **91**, 115106. DOI: 10/gm782w.

Teale, A. M. et al. (2022). "DFT Exchange: Sharing Perspectives on the Workhorse of Quantum Chemistry and Materials Science." *Phys. Chem. Chem. Phys.* DOI: 10/gqphzw.

Tesauro, G. (1994). "TD-Gammon, a Self-Teaching Backgammon Program, Achieves Master-Level Play." *Neural Comput.* **6**, 215. DOI: 10/bqsszt.

Tkatchenko, A. (2020). "Machine Learning for Chemical Discovery." *Nat. Commun.* **11**, 4125. DOI: 10/gg89ph.

Tocchio, L. F., F. Becca, A. Parola & S. Sorella (2008). "Role of Backflow Correlations for the Nonmagnetic Phase of the *t–t*' Hubbard Model." *Phys. Rev. B* **78**, 041101. DOI: 10/fqzg2f.

Torlai, G., G. Mazzola, G. Carleo & A. Mezzacapo (2020). "Precise Measurement of Quantum Observables with Neural-Network Estimators." *Phys. Rev. Res.* **2**, 022060. DOI: 10/ghmzsh.

Torlai, G., G. Mazzola, J. Carrasquilla, M. Troyer, R. Melko & G. Carleo (2018). "Neural-Network Quantum State Tomography." *Nat. Phys.* **14**, 447. DOI: 10/gcz7jj.

Unke, O. T., S. Chmiela, H. E. Sauceda, M. Gastegger, I. Poltavsky, K. T. Schütt, A. Tkatchenko & K.-R. Müller (2021). "Machine Learning Force Fields." *Chem. Rev.* **121**, 10142. DOI: 10/gj7n5p.

Van den Oord, A., S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior & K. Kavukcuoglu (2016a). "WaveNet: A Generative Model for Raw Audio." arXiv: 1609.03499.

Van den Oord, A., N. Kalchbrenner, O. Vinyals, L. Espeholt, A. Graves & K. Kavukcuoglu (2016b). "Conditional Image Generation with PixelCNN Decoders." In *Advances in Neural Information Processing Systems 29* (Curran Associates), 4797–4805. URL: https://papers.nips.cc/paper/2016/hash/b1301141feffabac455e1f90a7de2054-Abstract.html.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser & I. Polosukhin (2017). "Attention Is All You Need." In *Advances in Neural Information Processing Systems 30* (Curran Associates), 5998–6008. URL: https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Von Lilienfeld, O. A. & K. Burke (2020). "Retrospective on a Decade of Machine Learning for Chemical Discovery." *Nat. Commun.* **11**, 4895. DOI: 10/ghqj5d.

Von Lilienfeld, O. A., K.-R. Müller & A. Tkatchenko (2020). "Exploring Chemical Compound Space with Quantum-Based Machine Learning." *Nat. Rev. Chem.* **4**, 347. DOI: 10/ghdztf.

Werbos, P. (1974). "Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences." PhD thesis. Harvard University.

Wilson, M., N. Gao, F. Wudarski, E. Rieffel & N. M. Tubman (2021). "Simulations of State-of-the-Art Fermionic Neural Network Wave Functions with Diffusion Monte Carlo." arXiv: 2103.12570.

Wilson, M., S. Moroni, M. Holzmann, N. Gao, F. Wudarski, T. Vegge & A. Bhowmik (2022). "Wave Function Ansatz (but Periodic) Networks and the Homogeneous Electron Gas." arXiv: 2202.04622.

Xie, H., L. Z. null & L. Wang (2022). "Ab-Initio Study of Interacting Fermions at Finite Temperature with Neural Canonical Transformation." *J. Mach. Learn.* **1**, 38. DOI: 10/gqpv2j.

Yang, L., W. Hu & L. Li (2020a). "Scalable Variational Monte Carlo with Graph Neural Ansatz." arXiv: 2011.12453.

Yang, P.-J., M. Sugiyama, K. Tsuda & T. Yanai (2020b). "Artificial Neural Networks Applied as Molecular Wave Function Solvers." *J. Chem. Theory Comput.* **16**, 3513. DOI: 10/gm782z.

Yoshioka, N., W. Mizukami & F. Nori (2021). "Solving Quasiparticle Band Spectra of Real Solids Using Neural-Network Quantum States." *Commun. Phys.* **4**, 1. DOI: 10/gj6tnr.

Zhao, L. & E. Neuscamman (2016). "An Efficient Variational Principle for the Direct Optimization of Excited States." *J. Chem. Theory Comput.* **12**, 3436. DOI: 10/gqpvn3.

Zhao, T., J. Stokes & S. Veerapaneni (2022). "Scalable Neural Quantum States Architecture for Quantum Chemistry." arXiv: 2204.13903.

Acknowledgements