Electronic excited states in deep variational Monte Carlo

M. T. Entwistle[1, *] Z. Schätzle[1, *] P. A. Erdman[1] J. Hermann[1, †] and F. Noé[2, 1, 3, 4, ‡]

[1]*FU Berlin, Department of Mathematics and Computer Science, Arnimallee 12, 14195 Berlin, Germany*
[2]*Microsoft Research AI4Science*
[3]*FU Berlin, Department of Physics, Arnimallee 14, 14195 Berlin, Germany*
[4]*Rice University, Department of Chemistry, Houston, Texas 77005, USA*
(Dated: January 18, 2023)

Obtaining accurate ground and low-lying excited states of electronic systems is crucial in a multitude of important applications. One *ab initio* method for solving the Schrödinger equation that scales favorably for large systems is variational quantum Monte Carlo (QMC). The recently introduced deep QMC approach uses ansatzes represented by deep neural networks and generates nearly exact ground-state solutions for molecules containing up to a few dozen electrons, with the potential to scale to much larger systems where other highly accurate methods are not feasible. In this paper, we extend one such ansatz (PauliNet) to compute electronic excited states. We demonstrate our method on various small atoms and molecules and consistently achieve high accuracy for low-lying states. To highlight the method's potential, we compute the first excited state of the much larger benzene molecule, as well as the conical intersection of ethylene, with PauliNet matching results of more expensive high-level methods.

I. INTRODUCTION

The fundamental challenge of quantum chemistry, solid-state physics and many areas of computational materials science is to obtain solutions to the electronic Schrödinger equation for a given system, which in principle provide complete access to its chemical properties. The ground and low-lying excited states typically determine the behavior of a system and are therefore of the most interest in many applications. Understanding and being able to describe excited-state processes [1], including a wide variety of important spectroscopy methods such as fluorescence, photoionization and optical absorption of molecules and solids, is key to the successful design of new materials.

Unfortunately, the Schrödinger equation cannot be solved exactly except in the simplest cases, such as one-dimensional toy systems or a single hydrogen atom. Accordingly, many approximate numerical methods have been developed which provide solutions at varying degrees of accuracy. Time-dependent density functional theory [2, 3] (TDDFT) is the most popular method due to its computational efficiency, but has well known limitations [4–9]. Higher-accuracy methods have a computational cost that scales rapidly with system size — the well established full configuration interaction [10] (FCI) and coupled cluster [11] (CC) techniques scale $\sim \mathcal{O}(\exp(N))$ [12] and $\sim \mathcal{O}(N^{5-10})$ [13] respectively, where N is the number of electrons, thereby severely limiting their practical use. There is thus a huge need for *ab initio* methods that scale more favorably with system size, allowing the modeling of practically relevant molecules and materials.

Quantum Monte Carlo (QMC) techniques offer a route forward with their favorable scaling ($\mathcal{O}(N^{3-4})$) and therefore dominate high-accuracy calculations where other methods are too expensive [14, 15]. A state-of-the-art QMC calculation typically involves the construction of a multi-determinant baseline wavefunction through standard electronic-structure methods, which is augmented with a Jastrow factor to efficiently incorporate electron correlation, and then optimized through variational QMC (VMC) to obtain a trial wavefunction. This is then used within fixed-node diffusion QMC (DMC) to obtain a final electronic energy. The fixed-node approximation is used to avoid exponential scaling, with the drawback that the nodal surface of the trial wavefunction cannot be modified, which limits the accuracy of the DMC result [16]. A more expressive baseline wavefunction can improve upon this but traditional DMC often needs thousands to hundreds of thousands of determinants to reach convergence [17]. Additionally, DMC only provides the final energy, restricting the calculation of other electronic properties [18]. Both of these limitations can, in principle, be resolved at the VMC level, with the accuracy of VMC constrained only by the flexibility of the trainable wavefunction ansatz. So far, these techniques have mostly been developed for ground-state calculations, with different extensions proposed to address excited states [14, 19–28].

Recently, the new *ab initio* approach of deep VMC methods has been introduced [29–32] and subsequently further extended and improved [33–35]. In particular, PauliNet [29] and FermiNet [30] were the first methods to demonstrate that highly accurate ground-state results for molecules could be obtained using deep VMC with a lower computational complexity and using orders of magnitude fewer Slater determinants typically employed in other methods that achieve similar accuracy.

In the same spirit as Carleo and Troyer proposed for optimizing quantum states in lattice models [36], VMC

* M. T. Entwistle and Z. Schätzle contributed equally to this work.
† science@jan.hermann.name
‡ franknoe@microsoft.com

is used in order to train a neural network model that represents the many-body wavefunction in an unsupervised fashion, i.e. in contrast to other quantum machine learning approaches the only input to the method is the Hamiltonian, and training data are generated on the fly by sampling from the current wavefunction model and minimizing the variational energy. In both PauliNet and FermiNet deep antisymmetric neural networks are used to represent the fermionic wavefunction in the real space of electron coordinates.

Recently, there has been much interest in developing deep learning methods for excited states [37]. In this paper, we extend PauliNet towards the *ab initio* computation of electronic excited states (see the "Methods" section for details). The input is again only the Hamiltonian of the quantum system. By employing a simple energy minimization and numerical orthogonalization procedure, we are able to obtain the lowest excited-state wavefunctions of a given system. The excited-state optimization makes use of a penalty method that minimizes the overlap between the n-th excited state and the lower-lying states in the spectrum. Optimization methods that introduce additional constraints have been used in the context of VMC before [28] and provide a simple way to obtain orthogonal states without explicit enforcement in the wavefunction ansatzes. Combining these techniques with the expressiveness of neural network ansatzes yields highly accurate approximations to excited states with direct access to the wavefunctions for the evaluation of electronic observables. Neural network-based methods have targeted low-lying excited states of one-dimensional lattice models [27], but have not been applied to first-principles systems.

We demonstrate our method on a variety of small- and medium-sized molecules, where we consistently achieve highly accurate total energies, outperforming traditional quantum chemistry methods. We also compute excitation energies, transition dipole moments and oscillator strengths, the main ground-to-excited transition properties, with the latter two known to be more sensitive to errors in the underlying wavefunctions than energies. In all test systems, we find PauliNet closely matches high-order CC and experimental results. Next, we show that our method can be applied in a straightforward manner to much larger molecules, using the example of benzene where we match significantly more expensive high-level electronic-structure methods. Finally, we demonstrate that PauliNet can be used to compute excited-state potential energy surfaces by modeling an avoided crossing and conical intersection of ethylene, a highly multi-referential problem.

II. RESULTS

A. Nearly exact solutions for small atoms and molecules

To demonstrate our method we start by applying it to a range of small atoms and molecules. We optimize the lowest-lying excited states and compute their vertical excitation energies for the ground-state equilibrium geometry (see Supplementary Information Table I), with each PauliNet wavefunction containing a maximum of 10 determinants. In all systems, we obtain highly accurate total energies and estimates of the first few excitation energies competitive with high-accuracy quantum chemistry methods.
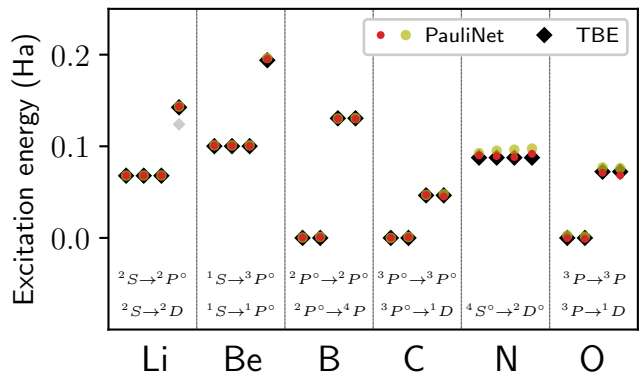

FIG. 1. **Deep VMC obtains highly accurate excited states for single elements.** PauliNet results for the excitation energies (with (red) and without (yellow) variance matching (see the "Methods" section for details)) are compared to the theoretical best estimates (TBE) taken from the NIST database [38]. Multiple PauliNet ansatzes with identical energies correspond to orthogonal degenerate states. For the TBE we have depicted four excitations per atom, taking account of the degeneracies. For all atoms we find the first excited state with high accuracy. For B, C and O the ground state is threefold degenerate. For these systems we choose one of the three states to compute excitation energies, resulting in transitions with a relative energy of zero. For Li and Be a further excitation energy is found. While we obtain the second excited state for Be, in Li we miss out intermediate states and instead find the transition from the ground state to the ^2D state. This can be related to the generic CASSCF initialization of the ansatzes. (The numerical data can be found in Supplementary Information Table II.)

In Fig. 1 the excitation energies of the lowest states are shown for several atoms. For all the atoms the excitation energies are obtained within 4 mHa of the theoretical best estimates (TBE) [38]. Due to the high degree of symmetry the atoms exhibit degeneracies, that is, multiple orthogonal states can be found with the same energy. Being subject to the orthogonalization constraint, PauliNet approximates all orthogonal states of an energy level individually, which is observed by attaining multiple results at the same energy level. The multiplicity of
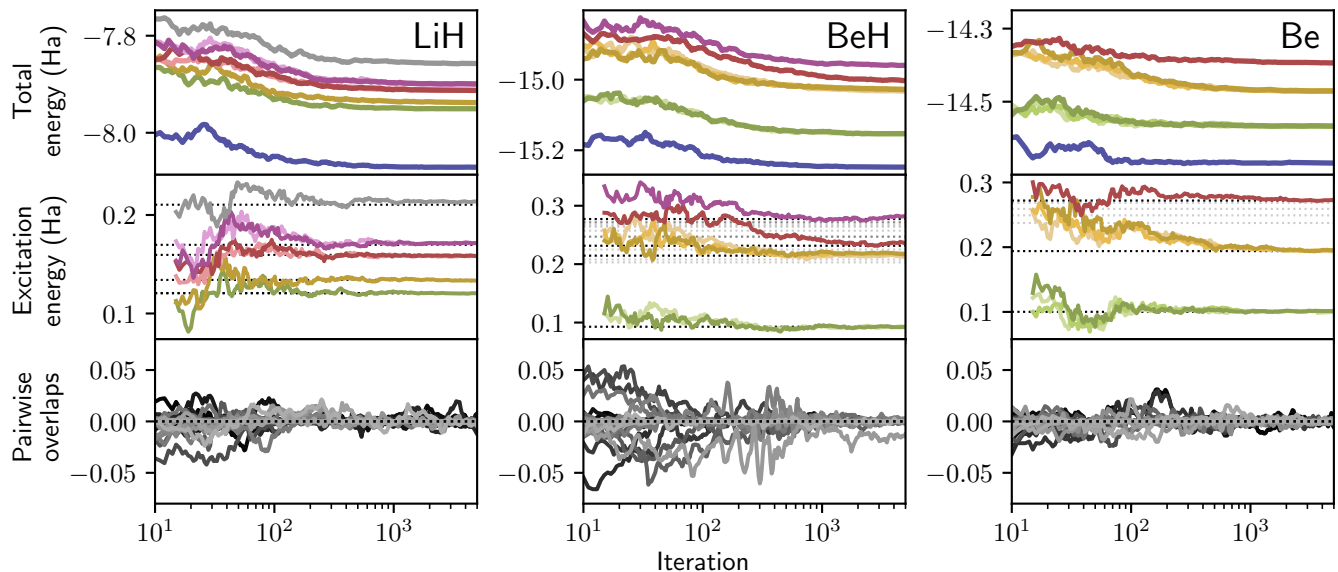
FIG. 2. **Optimizing low-lying excited states for small molecules.** Several excited states of LiH, BeH and Be are approximated. The convergence of the total energies (upper row), excitation energies (middle row) and the pairwise overlaps between the wavefunctions (bottom row) are shown. For degenerate states multiple ansatzes attain the same energy. Dotted horizontal lines are excitation energies from FCI calculations and other highly accurate references [38–41]. Due to the initialization from the CASSCF baseline the wavefunctions start with a small overlap, which is retained throughout the optimization. (The numerical data can be found in Supplementary Information Table II.)

the exact solution can be obtained theoretically by considering the electronic configurations of the atoms and is reproduced within our experiments.

We then compute a larger number of excited states for LiH, BeH and Be. In each experiment, we optimize eight ansatzes in parallel. In Fig. 2 we illustrate the training process by plotting the convergence of the total energies and excitation energies. Additionally, we plot the training estimates of the pairwise overlaps of the wavefunctions, which remain small throughout the optimization process. We confirm that the final overlaps are near-zero by exhaustively sampling the trained wavefunctions, thereby obtaining well-converged Monte Carlo estimates (see Supplementary Information Table VI). Based on the degeneracies we find a total of five (LiH), four (BeH) and three (Be) distinct excitation energies, respectively. The excitation energies match those from reference values, and in particular, we find that for all systems studied here we reliably obtain the first excited state, and apart from one case also the second excited state. However, especially for clusters of higher-lying excited states with similar energies, we typically do not find all members of the cluster. In these cases, which states are found depends on the initialization of our ansatzes, as well as the total number of states that are being sought. To give a transparent picture of the capabilities of our method, in this work we have refrained from optimizing the CASSCF baseline in order to find all possible excitations.

B. Highly accurate wavefunctions: transition dipole moments and oscillator strengths

Total energies and vertical excitation energies are the primary focus when benchmarking excited-state methods as they are readily available from many theoretical models and provide a good initial guess of a particular method's accuracy. However, they provide only a partial characterization of the electronic states, and while a method in question may give accurate energies, other quantities of key importance may be inaccurate [42–44].

Transition dipole moments (TDM) and oscillator strengths are two principal ground-to-excited transition properties and are of great interest. TDMs determine how polarized electromagnetic radiation will interact with a system due to its distribution of charge, and therefore determine transition rates and probabilities of induced state changes. In the electric dipole approximation, the TDM between two states i and j is given by

$$\mathbf{d}_{ij} = \langle \psi_i | \hat{\boldsymbol{\mu}} | \psi_j \rangle, \tag{1}$$

where $\hat{\boldsymbol{\mu}} = \sum_k q \hat{\mathbf{r}}_k$ is the sum over the position operator of each particle weighted by its charge, with $q = -e$ for electronic systems. We obtain the expectation value by Monte Carlo sampling according to Eq. (15). While the TDM is important for understanding a number of processes, including optical spectra, it is a complex vector quantity and not an experimental observable by itself. The closely related oscillator strength is what is inferred

through experiment and is given by

$$f_{ij} = \frac{2}{3}\Delta E d_{ij}^2, \qquad (2)$$

where ΔE is the excitation energy between states i and j, and d_{ij}^2 is the dipole strength. It is known that, in addition to being more basis-set sensitive, d_{ij} and f_{ij} are both highly dependent on the quality of the trial wavefunctions [45] and represent a more rigorous test for *ab initio* methods than just energies.

Recently, transition energies and oscillator strengths for a variety of small molecules have been computed using high-order CC calculations, systematically extrapolating to the complete basis set (CBS) limit, and comparing to experimental results where possible, in order to supply a comprehensive set of theoretical benchmarks [46, 47]. In that spirit, we now use these results to benchmark the accuracy of oscillator strengths computed using PauliNet. Furthermore, we also compare to multireference CC (MR-CC) results where possible [48]. We compute the first few electronic states for five molecules (BH, CH^+, H_2O, NH_3, CO), such that we obtain the first non-zero oscillator strength (within the dipole approximation) for each. All calculations [49] are performed at the same ground-state equilibrium geometries as refs. 46 and 47 (see Supplementary Information Table I) and using the same number of determinants (≤ 10) as in Section II A.

Our results for all systems are shown in Fig 3. First, we compute the amount of correlation energy recovered in the ground state, and find PauliNet matches high-order CC methods (panel a). Second, we compute the excitation energy for each transition and find this to be close to the TBE, on par with CC and much more consistent than TDDFT where the accuracy depends on the molecule and on the exact TDDFT method used (panel b). Finally, we compare the oscillator strengths (for the $0 \rightarrow 2$ transition) in panel c. Even high-order methods such as CC and MR-CC can produce a spectrum of results depending on the expansion and basis set used, with this exacerbated in cheaper methods such as TDDFT (see the example of CO). In all systems, PauliNet compares well with experimental results, demonstrating the quality of deep VMC wavefunctions with just a minimal number of determinants.

C.   Application to larger molecules

The previous two sections showed that we achieve highly accurate results across a range of small systems. While this is encouraging, traditional high-accuracy methods that are better established are readily available for such small systems. In this section, to demonstrate the potential of excited PauliNet, we show that it can be applied in a straightforward manner to significantly larger molecules. For this objective, we choose the example of the benzene molecule (panel a of Fig. 4). Studies
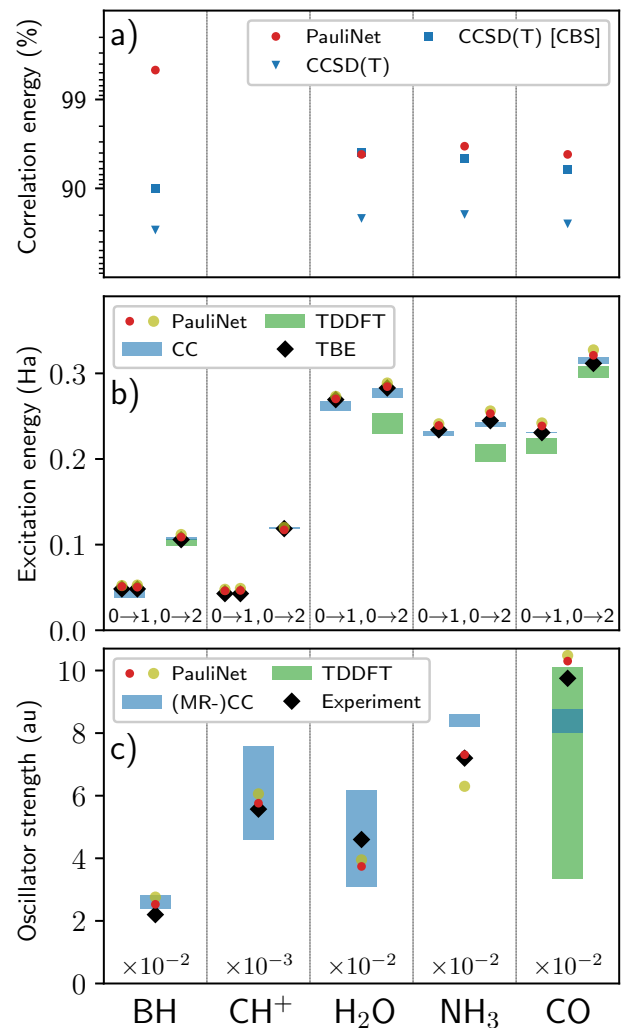

FIG. 3. **Deep VMC obtains highly accurate excited-state energies and wavefunctions for small molecules.** **a** PauliNet recovers the same amount of correlation energy as high-order CC methods [38]. (CH^+: No better reference energy to compare with.) **b** Lowest triplet ($0\rightarrow1$) and singlet ($0\rightarrow2$) excitation energies obtained using PauliNet (with (red) and without (yellow) variance matching), CC and TDDFT, with the TBE given. (BH and CH^+ exhibit degeneracy for the triplet state; CC is CCSD or higher, except for the triplet state of BH which includes CC2.) **c** Oscillator strengths computed for the $0\rightarrow2$ transitions. PauliNet compares well to experiment in all systems and matches the accuracy of (MR-)CC results, demonstrating the quality of few-determinant PauliNet wavefunctions. (We have omitted a factor of two linked to degeneracy in BH and CO.) Refs: exact correlation energies [38, 50, 51]; excitation energies from CC [46–48, 52–56], TDDFT [55–58] and TBE [46, 47, 52, 59, 60]; oscillator strengths from (MR-)CC [46–48, 61, 62], TDDFT [63] and experiment [64–68]. (The numerical data can be found in Supplementary Information Table III.)

of its electronic structure and other properties are plentiful due to its importance in bio and organic chemistry, and with 42 electrons all-electron calculations will be extremely demanding or even intractable for a high-level description of its electronic states, depending on the theory level used.

Using a PauliNet ansatz with just 10 determinants, the same as in the much smaller systems, and slightly deeper neural networks (see Supplementary Information Table VII) we obtain very good total energies for the ground state and first excited state (upper left of Fig. 4). We note the better accuracy than high-level CC calculations, with this signifying highly accurate wavefunctions that can be used to compute other observables, as demonstrated in the previous section. The computed excitation energy is also shown (right of Fig. 4), with PauliNet compared against several experimental and theoretical results. The lower experimental result [69] (dashed black line) quantifies an adiabatic excitation energy, i.e. the energy difference between the ground state and the excited state at the corresponding relaxed geometries. This quantity is corrected to obtain the vertical excitation energy [28] (solid black line), which omits nuclear relaxation and vibrational effects. As our calculations are performed at the ground-state equilibrium geometry, we are targeting the vertical excitation energy, and therefore consider this corrected experimental result to be closer to the ground truth. We find this to be slightly underestimated by high-order methods (CC, DMC), and slightly overestimated by PauliNet. In other systems (panel b of Fig 3) we notice a similar trend when comparing to the TBE.

PauliNet formally scales as $\mathcal{O}(N^4)$ with the number of electrons N, and in practice we observe a scaling behavior $\mathcal{O}(N^3)$ for the systems investigated so far, which is related to quadratic scaling of the neural network with an extra factor from the evaluation of the local energy. As PauliNet is currently implemented in a research code, which is not optimized for production purposes, the computational time will have a large prefactor which makes it computationally unfavorable to, e.g. CC methods for small molecules. However, its very favorable scaling in N compared to $\mathcal{O}(N^{5-10})$ of high-level electronic-structure methods dominates for larger molecules, and this is clearly visible in benzene. For instance, ref. 70 used several state-of-the-art methods to obtain accurate benzene ground-state energies, with calculations run on several CPU types in a highly parallel manner (see Supporting Information of ref. 70 for details). PauliNet was run on a single RTX 3090 GPU at a fraction of the number of node hours. Although PauliNet is the computationally cheapest method in this comparison, it provides a significantly better (variational) ground-state energy than all methods (~ 0.48 Ha lower). As all methods compared in Fig. 4 provide similar excitation energies, these cannot be used to group the methods into more or less accurate, but overall this data indicates that PauliNet and deep VMC methods in general have a very favor-
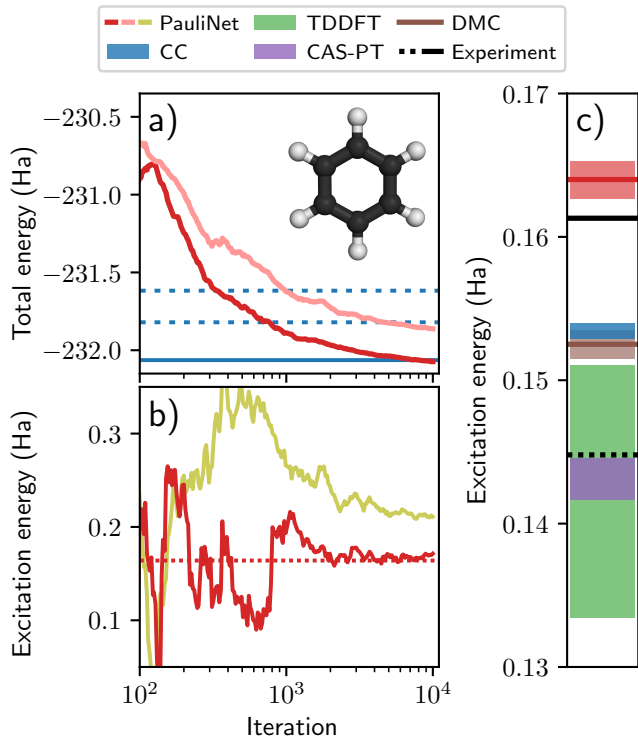

FIG. 4. **Calculating the two lowest electronic states of the benzene molecule.** Inset: Benzene structure. **a** Convergence of the total energies of the ground state (red) and excited state (light red) with training. Total energies of the ground state from CCSD(T) in the frozen-core approximation with the aug-cc-pVnZ basis set ($n = $ D, T) (dashed blue), and full CCSD(T) at the CBS limit (solid blue) are shown [38]. **b** Convergence of the excitation energy with training (with (red) and without (yellow) variance matching). **c** Excitation energy computed using PauliNet, TDDFT [58], CC [71], DMC [28], CAS-PT ([72] and calculations in openMolcas [73]) and Experiment [28, 69] (adiabatic (dashed black) and vertical (solid black) excitation energies). (The numerical data can be found in Supplementary Information Table IV.)

able cost/accuracy trade-off for molecules of the size of benzene and beyond.

D. Multi-reference application: conical intersections

Molecular configurations that produce electronic states with similar energies are fundamental in photochemical applications. Such configurations can lead to several states mixing, meaning they are all necessary for an accurate description of a particular process. Conical intersections are produced when two states become degenerate and require the computation of excited-state potential energy surfaces. The modeling of energy surfaces near degeneracies is inherently multi-reference with significant electronic correlation and is thus a challenging application for electronic-structure methods.

As a final application of excited PauliNet, we compute ground- and excited-state potential energies for ethylene ($H_2C=CH_2$) as a function of its torsion and pyramidalization angles (see inset of Fig. 5). Twisting around the C=C bond raises the energy of the ground state while lowering that of the first-excited singlet state, giving rise to an avoided crossing at a torsion angle τ of 90°. From this twisted structure, the energy gap between the two states is further reduced through the pyramidalization of one of the CH_2 groups, leading to a conical intersection. These potential energy curves, whose modeling is often too challenging for single-reference methods [74–76], have been characterized using multi-reference configuration interaction (MR-CI) methods [77] which we use for comparison.

We choose the same ground-state (planar) geometry as ref. 77 (optimized using a small CAS and the aug-cc-pVDZ basis set; see Supplementary Information Table I) and find the excitation energy between the ground state and first-excited singlet state to be within a few mHa of the MR-CI results. As we vary τ, while keeping all other geometric parameters fixed, we find the energy curves to be well reproduced by PauliNet, with an avoided crossing at $\tau = 90°$ (panel a of Fig. 5; curves symmetric about $\tau = 90°$). Single-reference methods, such as TDDFT (see figure), often overestimate the energy of the ground state at $\tau = 90°$ (barrier) and produce an unphysical cusp.

Next, we take the same twisted structure ($\tau = 90°$) as ref. 77 (optimized using a small CAS and the aug-cc-pVDZ basis set; see Supplementary Information Table I) and vary the pyramidalization angle ϕ, while keeping all other geometric parameters fixed. While there is a small discrepancy between PauliNet and the MR-CI results (panel b of Fig. 5), the trend of the energy curves is well described, including the correct minimum of the excited-state curve ($\sim 70°$) and the conical intersection (PauliNet: $\phi \sim 100°$; MR-CI: $\phi \sim 96°$). We note that many single-reference methods are unable to even qualitatively describe the conical intersection, instead predicting spurious features [76].

III. DISCUSSION

We have introduced an approach to compute highly accurate excited-state solutions of the electronic Schrödinger equation for molecules by using deep neural networks that are trained in an unsupervised manner with variational Monte Carlo. We have employed the PauliNet architecture [29] to approximate the ground- and excited-state wavefunctions, however other architectures such as FermiNet [30] or second quantization approaches [31] could also be employed, with suitable modifications. As our approach to find excited states only constrains the excited-state wavefunctions, the ability to compute highly accurate and variational absolute ground-state energies is unchanged. In addition, we demonstrate for a number of small molecules containing
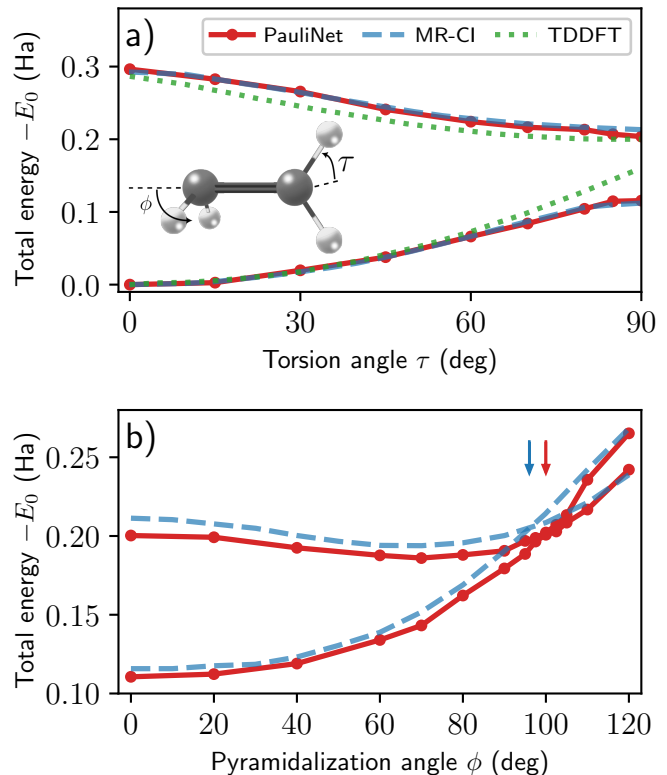

FIG. 5. **Modeling a conical intersection of ethylene.** Inset: Ethylene structure. **a** Total energies (relative to the ground state of the planar geometry E_0) of the ground state and first-excited singlet state as a function of torsion angle τ, with MR-CI [77] and TDDFT [75] results also plotted for comparison. TDDFT overestimates the barrier (the ground state at $\tau = 90°$) and produces an unphysical cusp, while the MR-CI results which predict an avoided crossing are well reproduced by PauliNet. **b** Same as above but as a function of pyramidalization angle ϕ ($\tau = 90°$), with the degeneracy of the two states producing a conical intersection. The arrows denote the conical intersection, with PauliNet ($\phi \sim 100°$) closely matching the MR-CI result ($\phi \sim 96°$). Note: The geometric parameters (bond lengths and angles) vary slightly between the torsion and pyramidalization experiments (see ref. 77). (The numerical data can be found in Supplementary Information Table V.)

up to 42 electrons, that excited PauliNet can reliably find the first excitation energies with an accuracy that is on par with high-level electronic-structure methods, whereas cheaper methods such as TDDFT are less consistent in approximating these energies. The accuracy of the excited-state wavefunctions is underlined by an accurate match of oscillator strengths, which depend on the transition dipole moment, a quantity that is more sensitive to the exact form of the wavefunction than the energy. For benzene (42 electrons), PauliNet already requires significantly less computational time than higher-order methods, and this advantage will only improve for larger molecules. Formally, a single PauliNet is an $\mathcal{O}(N^4)$ method for N electrons, due to the computational cost

of the Hartree-Fock or CASSCF baseline, however, in practice we empirically observe an $\mathcal{O}(N^3)$ dependency for the system sizes tested, as discussed above. In addition, for excited-state calculations n PauliNet replicas are used which gives rise to $\mathcal{O}(nN^3)+\mathcal{O}(n^2N^2)$, with the latter term arising from the pairwise overlaps and having a much smaller prefactor than the former.

Notably, almost identical excited PauliNet architectures are used across the systems shown in this paper – up to minor modifications such as the budget of Slater determinants and the total number of excited states requested, and a deeper network for benzene to adapt for a potentially more complex wavefunction. Whereas a skilled quantum chemist can usually tune and specialize an existing electronic-structure method to give very high-accuracy results for a given molecule, our aim is the exact opposite: to provide a method that, by leveraging machine learning tools, is as automated as possible and will work over a wide range of Hamiltonians provided.

We have demonstrated that we can compute ground- and excited-state potential energy surfaces with the example of ethylene where we model an avoided crossing and conical intersection. Here, where single-reference methods often fail, PauliNet performs well against multi-reference CI results. By combining the present approach with recent and ongoing extensions of PauliNet [34] and FermiNet [35] that variationally compute entire potential energy surfaces, both highly accurate ground- and excited-state energy surfaces are now accessible with deep VMC methods. Future work will investigate the application of PauliNet to other interesting processes where molecular dynamics interacts with excited states.

One of the limitations of the current approach is that it appears difficult to reliably find all excited states up to a given desired number, especially in cases where several excited states have similar energies. This is a complex problem which depends on the Hartree-Fock/CASSCF initialization, on the total number of states requested, on the learning algorithm, and the expressiveness of the architecture and will be studied in more detail elsewhere. However, the first excited state could be reliably found for all molecules studied here, and apart from one exception also the second excited state. This, in combination with the high numerical accuracy and the favorable computational cost, makes deep VMC a promising method to compute both ground- and excited-state properties for small- and medium-sized molecules with dozens or even low hundreds of electrons.

METHODS

PauliNet ansatz

At the heart of our approach is the PauliNet ansatz, introduced in ref. 29 and further refined in ref. 78, a multi-determinant Slater-Jastrow-backflow type trial wavefunction that is parametrized by highly expressive deep neural networks:

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

$$(3)$$

$$\tilde{\varphi}_{\boldsymbol{\theta},\mu i}(\mathbf{r}) = \varphi_\mu(\mathbf{r}_i) f^{(\mathrm{m})}_{\boldsymbol{\theta},\mu i}(\mathbf{r}) + f^{(\mathrm{a})}_{\boldsymbol{\theta},\mu i}(\mathbf{r}),$$

$$(4)$$

where $\mathbf{r} = (\mathbf{r}_1, ..., \mathbf{r}_N)$ is the $3N$-dimensional real space of electron coordinates. The structure of our ansatz ensures that the correct physics is encoded: the wavefunction obeys exact asymptotic behavior through the fixed electronic cusps γ, and is antisymmetric with respect to the exchange of like-spin electrons through the use of generalized Slater determinants, guaranteeing the Pauli exclusion principle is obeyed.

The expressiveness of PauliNet is contained in the Jastrow factor $J_{\boldsymbol{\theta}}$ and backflow $\mathbf{f}_{\boldsymbol{\theta}}$, which introduce many-body correlation, and are both represented through deep neural networks (denoted by trainable parameters $\boldsymbol{\theta}$). $J_{\boldsymbol{\theta}}$ and $\mathbf{f}_{\boldsymbol{\theta}}$ are constructed in ways that preserve the antisymmetry of the fermionic wavefunction with respect to exchanging like-spin electrons, as well as its cusp behavior. The Jastrow factor is an exchange-symmetric function, and captures complex correlation effects through augmenting the Slater-determinant baseline, but is incapable of modifying the nodal surface of the determinant expansion. Changes to the nodal surface are possible through the backflow, which acts on the single-electron orbitals φ_μ directly, transforming them into permutation-equivariant many-electron orbitals $\tilde{\varphi}_\mu$. $\mathbf{f}_{\boldsymbol{\theta}}$ is split into multiplicative (m) and additive (a) components (Eq. (4)), and is designed to be equivariant under the exchange of like-spin electrons.

Ground-state optimization

Like traditional VMC methods, PauliNet is based on the variational principle, which guarantees that the energy expectation value of a trial wavefunction $\psi_{\boldsymbol{\theta}}$ is an upper bound to the true ground-state energy:

$$E_0 = \min_{\psi}\langle\psi|\hat{H}|\psi\rangle \leq \min_{\boldsymbol{\theta}}\langle\psi_{\boldsymbol{\theta}}|\hat{H}|\psi_{\boldsymbol{\theta}}\rangle. \qquad (5)$$

For a given system, a standard quantum chemistry method (Hartree-Fock (HF) for a single determinant; complete active space self-consistent field (CASSCF) for multiple determinants) is performed, with the solution supplemented by the analytically-known cusp conditions, thus producing a reasonable baseline wavefunction. We

then optimize the PauliNet ansatz by minimizing the total electronic energy (serving directly as the loss), following the standard VMC trick of evaluating it as an expectation value of the local energy, $E_{\mathrm{loc}}(\mathbf{r}) = \hat{H}\psi(\mathbf{r})/\psi(\mathbf{r})$, over the probability distribution $|\psi_{\boldsymbol{\theta}}|^2$:

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{r}\sim|\psi_{\boldsymbol{\theta}}|^2}\left[E_{\mathrm{loc}}[\psi_{\boldsymbol{\theta}}](\mathbf{r})\right]. \qquad (6)$$

This means that, in practice, we alternate between sampling electron positions generated using a Langevin algorithm with the probability of the trial wavefunction serving as the target distribution, and optimizing the trial wavefunction parameters using stochastic gradient descent. For further details, see ref. 29.

Computing excited states

We now introduce the central idea of this paper: a deep VMC method to compute the ground and low-lying excited states of a given electronic system. While we employ PauliNet to represent the individual wavefunctions, the method can also employ FermiNet or other real-space wavefunction representations with suitable modifications.

In a similar spirit to the ground-state optimization process, we first obtain a reasonable baseline for each state by performing a minimal state-averaged CASSCF calculation. This optimizes the energy average for all states in question and yields a single set of orbitals to construct each multi-determinant wavefunction, which in turn are supplemented by the analytically-known cusp conditions. We fix the number of determinants in our ansatz by cutting off the CASSCF expansion based on the absolute values of their determinant coefficients. The choice of the CASSCF baseline ensures that the PauliNet ansatzes for the different excited states are close to orthogonal upon initialization. In contrast to the ground-state calculation, the optimization of excited states requires a more nuanced choice of the active space. In principle, we must ensure that the solutions contain determinants with orbitals of the necessary rotational symmetries (the Jastrow factor and backflow correction are rotationally-symmetric modifications of the orbitals) and spin configurations (the choice of the number of spin-up and spin-down electrons does impose restrictions on the states that may be attained by our ansatz). For most systems studied in this paper, a generic choice of the active space was sufficient (see Supplementary Table VIII) and we have not studied the dependence on the CAS initialization in more depth. As shown in previous studies the quality of the orbitals has only a minor effect on the training and does not change the final energy [78]. If, however, the initialization is not accounted for and the baseline solutions provide a qualitatively wrong spectrum of excited states our ansatzes may be trapped in local minima and miss intermediate excited states (see Fig. 2), even though we keep the Slater-determinant coefficients

c_p and linear coefficients $c_{\mu k}$ of the single-electron orbitals $\varphi_\mu(\mathbf{r}_i) = \sum_k c_{\mu k}\phi_k(\mathbf{r}_i)$ trainable.

Our objective is to calculate the lowest n eigenstates of a given system, that is, find the set of orthogonal states that minimizes the energy expectation value. We approach this challenge by introducing a penalty term to the energy loss function (Eq. (6)) and optimizing the joint loss for n PauliNet instances:

$$\mathcal{L}(\boldsymbol{\theta}) = \underbrace{\sum_i \mathbb{E}_i\big[E_{\mathrm{loc}}[\psi_{\boldsymbol{\theta},i}](\mathbf{r})\big]}_{\text{energy minimization}} + \alpha \underbrace{\sum_{i>j}\left(\frac{1}{1-|S_{ij}|}-1\right)}_{\text{overlap penalty}},$$

(7)

where $\mathbb{E}_i = \mathbb{E}_{\mathbf{r}\sim|\psi_{\boldsymbol{\theta},i}|^2}$ and S_{ij} is the pairwise overlap between states i and j. The functional form of the overlap penalty is chosen to diverge when two states collapse and behave linearly when states are close to orthogonal (see the next section for details). This allows states to overlap during the optimization procedure, while preventing their collapse and eventually driving them to orthogonality when they have settled in a local minimum of the energy. The hyperparameter α weights the two loss terms and can be increased throughout the training to strengthen the orthogonality condition when approaching the final wavefunctions. For a sufficiently large α the true minimum of the loss function corresponds to the sum of the energies of the lowest-lying excited states with these states having no overlap. Thus, optimizing the penalized loss function (Eq. (7)) leads to an unbiased convergence towards the lowest-lying excited states (see below). In practice a small α is typically sufficient, making a robust choice possible.

To stabilize the training and reduce the computational cost we detach gradients in such a way that we only consider the overlap with the lower-lying states respectively, that is, the ground state is subject to unconstrained energy minimization and the n-th excited state introduces n pairwise penalty terms. We compute the overlap of the unnormalized states i and j as the geometric mean of the two Monte Carlo estimates, obtained over distributions $|\psi_{\boldsymbol{\theta},i}|^2$ and $|\psi_{\boldsymbol{\theta},j}|^2$, respectively:

$$S_{ij} = \mathrm{sgn}\left(\mathbb{E}_i\left[\frac{\psi_{\boldsymbol{\theta},j}(\mathbf{r})}{\psi_{\boldsymbol{\theta},i}(\mathbf{r})}\right]\right) \times \sqrt{\mathbb{E}_i\left[\frac{\psi_{\boldsymbol{\theta},j}(\mathbf{r})}{\psi_{\boldsymbol{\theta},i}(\mathbf{r})}\right] \mathbb{E}_j\left[\frac{\psi_{\boldsymbol{\theta},i}(\mathbf{r})}{\psi_{\boldsymbol{\theta},j}(\mathbf{r})}\right]}.$$

(8)

The sign of the overlap can be obtained from either of the two estimators, which match in the limit of infinite sampling. If the overlap is close to zero and the signs of the two estimates differ due to statistical noise of the sampling, we consider the states to be orthogonal. Similar to the energy loss, the gradient [79] of the pairwise overlap can be formulated such that it depends on the first derivative of the log wavefunction with respect to the parameters only (see below for details).

Finally, we note that different states may be modeled at different levels of quality, which can lead to erroneous

excitation energies. In order to improve the error cancellation of our ansatzes we employ a variance-matching technique. As the variance of the energy σ^2 can be considered a metric of how close a wavefunction is to a true eigenstate, variance-matching procedures can be useful tools [23, 80, 81]. Here, we utilize a simple scheme: for single-state quantities such as total energies, we evaluate all wavefunctions at the end of training. For multi-state quantities, such as excitation energies or transition dipole moments, we match states of a similar variance. That is, if final $\psi_{\boldsymbol{\theta},i}$ has a lower variance than final $\psi_{\boldsymbol{\theta},j}$, we take $\psi_{\boldsymbol{\theta},i}$ at an earlier point in training. This simply involves computing σ^2 of the training energies and applying exponential moving average at each iteration to monitor convergence (see below for details). We find this procedure typically improves the final results.

Loss function and overlap penalty

There are a number of choices of possible loss functions for the optimization of excited states in quantum Monte Carlo [22, 28, 82]. In order to assess the feasibility of excited-state optimization with deep neural network ansatzes in variational Monte Carlo we conducted a range of experiments with different types of optimization objectives. Our empirical findings showed that employing a penalty method is the conceptually most straightforward approach and gives stable results when combining it with our implementation of PauliNet. Initially, we started with an overlap penalty term similar to Pathak et al[28]. However, we found that our optimization could still collapse even if we chose a sufficiently large prefactor (α) and the training could not recover. We therefore switched to an alternative penalty term (Eq. (7)) which diverges upon a collapse of the states. The effect of our penalty term can be illustrated by considering the loss for a two-state system with the exact ground state $|\psi_0\rangle$ and a linear combination of the ground and first excited state $|\psi_1\rangle$ (see Fig. 6):

$$|\psi_\epsilon\rangle = \sqrt{1-\epsilon}|\psi_1\rangle + \sqrt{\epsilon}|\psi_0\rangle.$$

(9)

The overlap and the energy can be obtained as

$$\langle\psi_0|\psi_\epsilon\rangle = \sqrt{\epsilon}, \quad \langle\psi_\epsilon|H|\psi_\epsilon\rangle = (1-\epsilon)E_1 + \epsilon E_0.$$

(10)

In the vicinity of the orthogonal solution, the Taylor expansion of the penalty term is

$$\frac{1}{1-|S|}-1 = |S| + |S|^2 + |S|^3 + ..., \text{ at } |S| = 0, \quad (11)$$

that is, the overlap penalty behaves linearly to first order. This gives rise to a penalty that is locally stable for any prefactor, lower bounded by the S^2 penalty term, and diverges if states collapse. For a large enough α parameter the global optimum of the total loss is at zero overlap, that is, the optimization method is incentivized to find exactly orthogonal states without mixing.
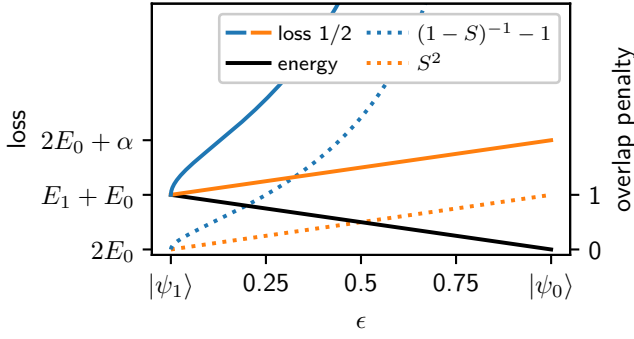
FIG. 6. **Sketch of the loss function.** This figure illustrates the behavior of our loss function for a two-state system. The ground state is kept fixed and the second state is considered to be a linear combination of the ground state and first excited state (Eq. (9)). The scales are to be understood in arbitrary units, as they depend on the choice of the hyperparameters and the energies of the system under investigation.

In practice, for the batch sizes used in our calculations, we have not observed a bias due to the non-linear nature of the penalty when applied to sampled expectation values of the overlap. However, it is expected that this is no longer the case in the limit of small batches. In order to elucidate how our loss function behaves in this regard, we compute the two lowest states of LiH using a range of different batch sizes (see Fig. 7). We find the optimization procedure to be robust for the large batch sizes that we typically employ (≥ 2000), with the excitation energy within 1 mHa of the exact, and the pairwise overlap remaining small throughout training (panel c). For smaller batch sizes, we observe a larger degree of statistical noise in the pairwise overlap, which leads to a less reliable approximation for the excited state and the corresponding excitation energy (panel b).

Gradient of the loss function

In order to differentiate the loss function we explicitly formulate the gradient. We consider the general case of a mixed observable:

$$O_{ij} = \frac{1}{N_i N_j} \int d^3\mathbf{r} \; \psi_{\boldsymbol{\theta},i}(\mathbf{r}) \big[\hat{O}\psi_{\boldsymbol{\theta},j}(\mathbf{r}) \big], \quad (12)$$

$$= \frac{N_i}{N_j} \mathbb{E}_i \left[\frac{\hat{O}\psi_{\boldsymbol{\theta},j}(\mathbf{r})}{\psi_{\boldsymbol{\theta},i}(\mathbf{r})} \right], \quad (13)$$

where N_i, N_j are the norms of the wavefunctions and $\mathbb{E}_i = \mathbb{E}_{\mathbf{r}\sim|\psi_{\boldsymbol{\theta},i}|^2}$. By the property of Hermitian matrices, $O_{ij} = O_{ji}$, we derive an expression that does not depend
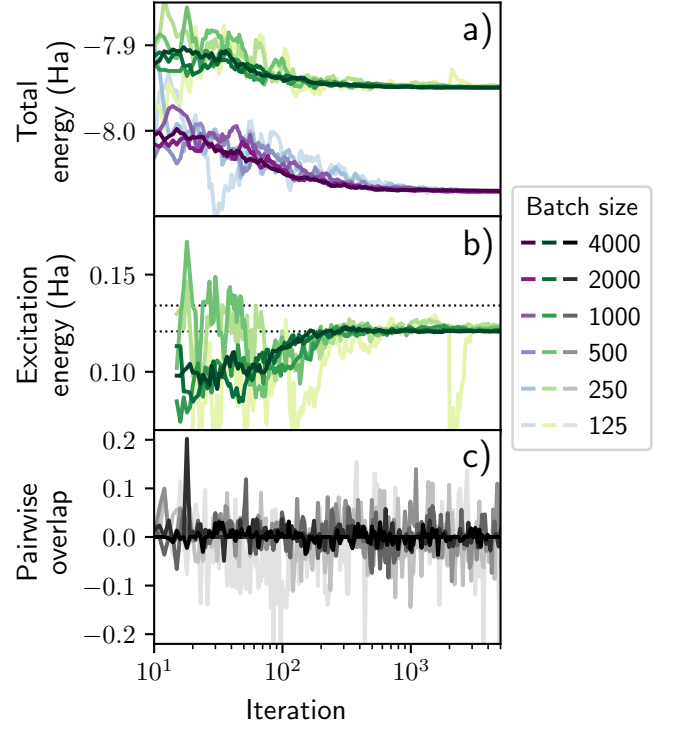

FIG. 7. **Behavior of the loss function with batch size.** The ground state and first excited state of LiH are approximated. The convergence of the total energies **(a)**, excitation energy **(b)** and the pairwise overlap between the wavefunctions **(c)** are shown. Dotted horizontal lines are excitation energies from highly accurate references [39, 40]. While the optimization works well for the large batches that we typically employ in our calculations (≥ 2000), this becomes less reliable, at least for the excited state, in the limit of smaller batch sizes. Note: Darker corresponds to a larger batch size in each respective color.

on the wavefunction norms:

$$O_{ij} = \sqrt{\frac{N_i}{N_j}\mathbb{E}_i\left[\frac{\hat{O}\psi_{\boldsymbol{\theta},j}(\mathbf{r})}{\psi_{\boldsymbol{\theta},i}(\mathbf{r})}\right]} \sqrt{\frac{N_j}{N_i}\mathbb{E}_j\left[\frac{\hat{O}\psi_{\boldsymbol{\theta},i}(\mathbf{r})}{\psi_{\boldsymbol{\theta},j}(\mathbf{r})}\right]},$$

$$(14)$$

$$= \operatorname{sgn}\left(\mathbb{E}_i\left[\frac{\hat{O}\psi_{\boldsymbol{\theta},j}(\mathbf{r})}{\psi_{\boldsymbol{\theta},i}(\mathbf{r})}\right]\right)$$

$$\times \sqrt{\mathbb{E}_i\left[\frac{\hat{O}\psi_{\boldsymbol{\theta},j}(\mathbf{r})}{\psi_{\boldsymbol{\theta},i}(\mathbf{r})}\right]\mathbb{E}_j\left[\frac{\hat{O}\psi_{\boldsymbol{\theta},i}(\mathbf{r})}{\psi_{\boldsymbol{\theta},j}(\mathbf{r})}\right]}. \quad (15)$$

This expression reduces to the pairwise overlaps (Eq. (8)) upon setting $\hat{O} = \mathrm{Id}$. The derivative of this term can be expressed as

$$\partial O_{ij} = \frac{1}{O_{ij}}\Bigg\{ \mathbb{E}_i\left[\left(\frac{\hat{O}\psi_{\boldsymbol{\theta},j}(\mathbf{r})}{\psi_{\boldsymbol{\theta},i}(\mathbf{r})} - \mathbb{E}_i\left[\frac{\hat{O}\psi_{\boldsymbol{\theta},j}(\mathbf{r})}{\psi_{\boldsymbol{\theta},i}(\mathbf{r})}\right]\right) \right.$$

$$\left. \partial \ln|\psi_{\boldsymbol{\theta},i}(\mathbf{r})| \right] \times \mathbb{E}_j\left[\frac{\hat{O}\psi_{\boldsymbol{\theta},i}(\mathbf{r})}{\psi_{\boldsymbol{\theta},j}(\mathbf{r})}\right] + (i \Longleftrightarrow j)\Bigg\},$$

$$(16)$$

where $(i \Longleftrightarrow j)$ is an additional term with the two indices interchanged.

By considering the Hamiltonian operator $\hat{H}$ and setting $i = j$ we recover the gradient of the energy loss [29]:

$$\partial E_{ii} = 2\mathbb{E}_i\left[\left(\frac{\hat{H}\psi_{\boldsymbol{\theta},i}(\mathbf{r})}{\psi_{\boldsymbol{\theta},i}(\mathbf{r})} - \mathbb{E}_i\left[\frac{\hat{H}\psi_{\boldsymbol{\theta},i}(\mathbf{r})}{\psi_{\boldsymbol{\theta},i}(\mathbf{r})}\right]\right)\partial\ln|\psi_{\boldsymbol{\theta},i}(\mathbf{r})|\right].$$ (17)

Variance matching

As far as relative energies are concerned most computational chemistry methods rely heavily on the cancellation of error. While quantum Monte Carlo methods using neural network-based trial wavefunctions provide highly accurate total energies, the flexibility of these ansatzes is difficult to control which can lead to varying qualities of approximations for different states. In order to account for potential imbalances we utilize the variance of the wavefunctions as a measure of the quality of the approximation (zero-variance principle) and employ a variance-matching scheme. Variance-matching techniques as well as variance extrapolation have typically been applied by optimizing a family of ansatzes and comparing variances across the optimized wavefunctions [81]. Instead of training multiple ansatzes we checkpoint wavefunctions during the training and compute excitation energies by rewinding the ground state to match the variance of the excited state as depicted in Fig. 8. The mean and variance of each wavefunction are computed over the batch dimension at each step in training and smoothed with an exponential walking average. For the final estimation of excitation energies, the respective wavefunctions are then sampled exhaustively as in the usual evaluation process. While the variance matching hardly impacts the excitation energies for small systems, for larger and harder to optimize systems, such as benzene, it becomes increasingly relevant.

Spin treatment

PauliNet encodes only the spatial part of the wavefunction and its like-spin antisymmetry explicitly [14], while the spin part, which guarantees the opposite-spin antisymmetry, is only implicit. Every spin-assigned spatial ansatz such as PauliNet is always an eigenstate of $\mathcal{S}_z$ with an eigenvalue of $M = \frac{1}{2}(N_\uparrow - N_\downarrow)$, but it may not be an eigenstate of $\mathcal{S}^2$. The spatial part of eigenstates of $\mathcal{S}^2$ is characterized by specific sets of permutational symmetries involving opposite-spin electrons [83]. PauliNet does not enforce these symmetries but instead attempts to learn them through the variational principle because eigenstates of the Hamiltonian are also eigenstates of $\mathcal{S}^2$. Therefore, we do not, in general, control the spin of the eigenstates found in the optimization procedure — they

are simply found in the order of increasing energy, independent of spin. The spin of a found eigenstate can be obtained in principle by Monte Carlo sampling [84]. Whether a particular spin state is found in practice may be influenced by the spin of the CASSCF baseline wavefunction, which we therefore report in Supplementary Table VIII. In special cases, we may wish to target a specific spin state (e.g., see the section "Multi-reference application: conical intersections"), and for that, we can take advantage of the orbital-assigned backflow of PauliNet. Combined with the freezing of the determinant coefficients, this ensures that PauliNet remains in the same spin state as the CASSCF baseline wavefunction.
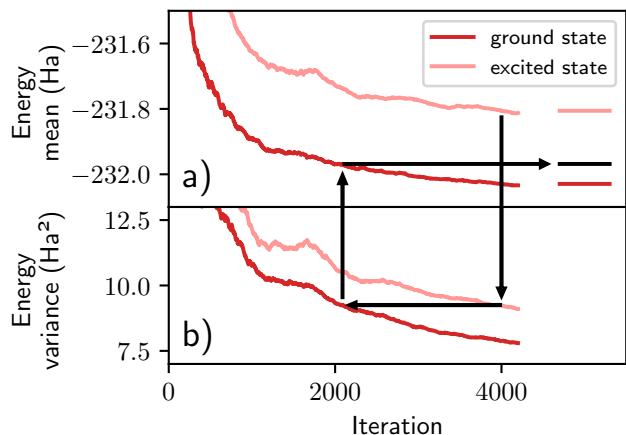
FIG. 8. **Sketch of the variance-matching procedure.** The excitation energy of the benzene calculation at step 4000 is obtained for illustration purposes. The variance (b) of the excited state is higher than that of the ground state and is therefore matched with the variance of the ground state at a previous iteration. The excitation energy is computed by comparing the mean energies (a) at the respective iterations. This acts to reduce the excitation energy and is found to improve the results in all of our experiments.

DATA AVAILABILITY

CODE AVAILABILITY

[1] R. Lindh and L. González, *Quantum Chemistry and Dynamics of Excited States: Methods and Applications* (John Wiley & Sons, 2020).

[2] W. Kohn and L. J. Sham, Phys. Rev. **140**, A1133 (1965).

[3] E. Runge and E. K. U. Gross, Phys. Rev. Lett. **52**, 997 (1984).

[4] P. Elliott, J. I. Fuks, A. Rubio, and N. T. Maitra, Phys. Rev. Lett. **109**, 266404 (2012).

[5] J. I. Fuks, K. Luo, E. D. Sandoval, and N. T. Maitra, Phys. Rev. Lett. **114**, 183002 (2015).

[6] Y. Suzuki, L. Lacombe, K. Watanabe, and N. T. Maitra, Phys. Rev. Lett. **119**, 263401 (2017).

[7] N. Singh, P. Elliott, T. Nautiyal, J. K. Dewhurst, and S. Sharma, Phys. Rev. B **99**, 035151 (2019).

[8] N. T. Maitra, Journal of Physics: Condensed Matter **29**, 423001 (2017).

[9] C. A. Ullrich and I. V. Tokatly, Phys. Rev. B **73**, 235102 (2006).

[10] P. G. Szalay, T. Müller, G. Gidofalvi, H. Lischka, and R. Shepard, Chemical Reviews **112**, 108 (2012).

[11] K. Sneskov and O. Christiansen, WIREs Computational Molecular Science **2**, 566 (2011).

[12] FCI scales exponentially, while truncated CI scales polynomially.

[13] The scaling of CC depends on the particular method used: CC2 $\mathcal{O}(N^5)$, CCSD $\mathcal{O}(N^6)$, CCSD(T) $\mathcal{O}(N^7)$, CC3 $\mathcal{O}(N^7)$, CCSDT $\mathcal{O}(N^8)$, CCSDT(Q) $\mathcal{O}(N^9)$, CCSDTQ $\mathcal{O}(N^{10})$.

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

[15] K. T. Williams, Y. Yao, J. Li, L. Chen, H. Shi, M. Motta, C. Niu, U. Ray, S. Guo, R. J. Anderson, J. Li, L. N. Tran, C.-N. Yeh, B. Mussard, S. Sharma, F. Bruneval, M. van Schilfgaarde, G. H. Booth, G. K.-L. Chan, S. Zhang, E. Gull, D. Zgid, A. Millis, C. J. Umrigar, and L. K. Wagner (Simons Collaboration on the Many-Electron Problem), Phys. Rev. X **10**, 011041 (2020).

[16] M. A. Morales, J. McMinis, B. K. Clark, J. Kim, and G. E. Scuseria, Journal of Chemical Theory and Computation **8**, 2181 (2012).

[17] A. Benali, K. Gasperich, K. D. Jordan, T. Applencourt, Y. Luo, M. C. Bennett, J. T. Krogel, L. Shulenburger, P. R. C. Kent, P.-F. Loos, A. Scemama, and M. Caffarel, The Journal of Chemical Physics **153**, 184111 (2020).

[18] B. M. Austin, D. Y. Zubarev, and W. A. Lester, Chemical Reviews **112**, 263 (2012).

[19] D. M. Ceperley and B. Bernu, The Journal of Chemical Physics **89**, 6316 (1988).

[20] N. S. Blunt, S. D. Smart, G. H. Booth, and A. Alavi, The Journal of Chemical Physics **143**, 134117 (2015).

[21] R. Send, O. Valsson, and C. Filippi, Journal of Chemical Theory and Computation **7**, 444 (2011).

[22] M. Dash, J. Feldt, S. Moroni, A. Scemama, and C. Filippi, Journal of Chemical Theory and Computation **15**, 4896 (2019).

[23] S. D. Pineda Flores and E. Neuscamman, The Journal of Physical Chemistry A **123**, 1487 (2019).

[24] L. Zhao and E. Neuscamman, Journal of Chemical Theory and Computation **12**, 3436 (2016).

[25] J. A. R. Shea and E. Neuscamman, Journal of Chemical Theory and Computation **13**, 6078 (2017).

[26] N. S. Blunt and E. Neuscamman, Journal of Chemical Theory and Computation **15**, 178 (2019).

[27] K. Choo, G. Carleo, N. Regnault, and T. Neupert, Phys. Rev. Lett. **121**, 167204 (2018).

[28] S. Pathak, B. Busemeyer, J. N. B. Rodrigues, and L. K. Wagner, The Journal of Chemical Physics **154**, 034101 (2021).

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

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

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

[32] J. Han, L. Zhang, and W. E, J. Comput. Phys. **399**, 108929 (2019).

[33] J. S. Spencer, D. Pfau, A. Botev, and W. M. C. Foulkes, arXiv:2011.07125 [physics.comp-ph] (2021).

[34] M. Scherbela, R. Reisenhofer, L. Gerard, P. Marquetand, and P. Grohs, arXiv:2105.08351v2 [physics.comp-ph] (2021).

[35] N. Gao and S. Günnemann, arXiv:2110.05064v2 [cs.LG] (2021).

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

[37] J. Westermayr and P. Marquetand, Chemical Reviews **121**, 9873 (2021).

[38] R. D. J. III, NIST Computational Chemistry Comparison and Benchmark Database, NIST Standard Reference Database Number 101 (2021).

[39] A. Bande, H. Nakashima, and H. Nakatsuji, Chemical Physics Letters **496**, 347 (2010).

[40] P. Jasik, J. E. Sienkiewicz, J. Domsta, and N. E. Henriksen, Phys. Chem. Chem. Phys. **19**, 19777 (2017).

[41] J. Pitarch-Ruiz, J. Sánchez-Marín, and A. M. Velasco, Journal of Computational Chemistry **29**, 523 (2008).

[42] E. Brémond, M. Savarese, C. Adamo, and D. Jacquemin, Journal of Chemical Theory and Computation **14**, 3715 (2018).

[43] A. Tajti and P. G. Szalay, Journal of Chemical Theory and Computation **15**, 5523 (2019).

[44] A. Tajti, L. Tulipán, and P. G. Szalay, Journal of Chemical Theory and Computation **16**, 468 (2020).

[45] R. Crossley, Physica Scripta **T8**, 117 (1984).

[46] P.-F. Loos, A. Scemama, A. Blondel, Y. Garniron, M. Caffarel, and D. Jacquemin, Journal of Chemical Theory and Computation **14**, 4360 (2018), pMID: 29966098.

[47] A. Chrayteh, A. Blondel, P.-F. Loos, and D. Jacquemin, Journal of Chemical Theory and Computation **17**, 416 (2021), pMID: 33256412.

[48] D. Bhattacharya, N. Vaval, and S. Pal, The Journal of Chemical Physics **138**, 094108 (2013).

[49] CH^+ was not included in the CC calculations in refs. 46 and 47. We instead compare to (MR-)CC results in ref. 48, using the same ground-state equilibrium geometry, which was obtained in a split-valence basis augmented with diffuse and polarization functions. See refs. 48 and 60 for more details.

[50] D. P. O'eill and P. M. W. Gill, Molecular Physics **103**, 763 (2005).

[51] E. Giner, D. Traore, B. Pradines, and J. Toulouse, The Journal of Chemical Physics **155**, 044109 (2021).

[52] H. Larsen, K. Hald, J. Olsen, and P. Jørgensen, The Journal of Chemical Physics **115**, 3015 (2001).

[53] K. Kowalski and P. Piecuch, Chemical Physics Letters **347**, 237 (2001).

[54] K. Kowalski and P. Piecuch, The Journal of Chemical Physics **115**, 643 (2001).

[55] P. Cronstrand, B. Jansik, D. Jonsson, Y. Luo, and H. Ågren, The Journal of Chemical Physics **121**, 9239 (2004).

[56] P. Sałek, T. Helgaker, O. Vahtras, H. Ågren, D. Jonsson, and J. Gauss, Molecular Physics **103**, 439 (2005).

[57] F. Liu, Z. Gan, Y. Shao, C.-P. Hsu, A. Dreuw, M. Head-Gordon, B. T. Miller, B. R. Brooks, J.-G. Yu, T. R. Furlani, and J. Kong, Molecular Physics **108**, 2791 (2010).

[58] C. Adamo, G. E. Scuseria, and V. Barone, The Journal of Chemical Physics **111**, 2889 (1999).

[59] Z. Biglari, A. Shayesteh, and A. Maghari, Computational and Theoretical Chemistry **1047**, 22 (2014).

[60] J. Olsen, A. M. De Merás, H. J. A. Jensen, and P. Jørgensen, Chemical Physics Letters **154**, 380 (1989).

[61] M. Barysz, Theoretica chimica acta **90**, 257 (1995).

[62] J. R. Lane, V. Vaida, and H. G. Kjaergaard, The Journal of Chemical Physics **128**, 034302 (2008).

[63] Y. Tawada, T. Tsuneda, S. Yanagisawa, T. Yanai, and K. Hirao, The Journal of Chemical Physics **120**, 8425 (2004).

[64] C. H. Douglass, H. H. Nelson, and J. K. Rice, The Journal of Chemical Physics **90**, 6940 (1989).

[65] B. Mahan and A. O'Keefe, Astrophysical Journal **248** (1981).

[66] P. A. Thorn, M. J. Brunger, P. J. O. Teubner, N. Diakomichalis, T. Maddern, M. A. Bolorizadeh, W. R. Newell, H. Kato, M. Hoshino, H. Tanaka, H. Cho, and Y.-K. Kim, The Journal of Chemical Physics **126**, 064306 (2007).

[67] T. Chen, Y. W. Liu, X. J. Du, Y. C. Xu, and L. F. Zhu, The Journal of Chemical Physics **150**, 064311 (2019).

[68] X. Kang, Y. W. Liu, L. Q. Xu, D. D. Ni, K. Yang, N. Hiraoka, K. D. Tsuei, and L. F. Zhu, The Astrophysical Journal **807**, 96 (2015).

[69] J. P. Doering, The Journal of Chemical Physics **51**, 2866 (1969).

[70] J. J. Eriksen, T. A. Anderson, J. E. Deustua, K. Ghanem, D. Hait, M. R. Hoffmann, S. Lee, D. S. Levine, I. Magoulas, J. Shen, N. M. Tubman, K. B. Whaley, E. Xu, Y. Yao, N. Zhang, A. Alavi, G. K.-L. Chan, M. Head-Gordon, W. Liu, P. Piecuch, S. Sharma, S. L. Ten-no, C. J. Umrigar, and J. Gauss, The Journal of Physical Chemistry Letters **11**, 8922 (2020), pMID: 33022176.

[71] P.-F. Loos, F. Lipparini, M. Boggio-Pasqua, A. Scemama, and D. Jacquemin, Journal of Chemical Theory and Computation **16**, 1711 (2020), pMID: 31986042.

[72] J. Lorentzon, P.-Å. Malmqvist, M. Fülscher, and B. O. Roos, Theoretica chimica acta **91**, 91 (1995).

[73] I. Fdez. Galván, M. Vacher, A. Alavi, C. Angeli, F. Aquilante, J. Autschbach, J. J. Bao, S. I. Bokarev, N. A. Bogdanov, R. K. Carlson, *et al.*, Journal of chemical theory and computation **15**, 5925 (2019).

[74] A. I. Krylov, Chemical Physics Letters **350**, 522 (2001).

[75] M. Mališ and S. Luber, Journal of Chemical Theory and Computation **16**, 4071 (2020), pMID: 32602729.

[76] M. Barbatti and R. Crespo-Otero, in *Density-Functional Methods for Excited States* (Springer International Publishing, 2014) pp. 415–444.

[77] M. Barbatti, J. Paier, and H. Lischka, The Journal of Chemical Physics **121**, 11614 (2004).

[78] Z. Schätzle, J. Hermann, and F. Noé, The Journal of Chemical Physics **154**, 124108 (2021).

[79] We employ gradient clipping to stabilize the training.

[80] L. Otis, I. Craig, and E. Neuscamman, The Journal of Chemical Physics **153**, 234105 (2020).

[81] P. J. Robinson, S. D. Pineda Flores, and E. Neuscamman, The Journal of Chemical Physics **147**, 164114 (2017).

[82] S. M. Garner and E. Neuscamman, The Journal of Chemical Physics **153**, 144108 (2020).

[83] R. Pauncz, *Spin Eigenfunctions* (Springer New York, NY).

[84] C.-J. Huang, C. Filippi, and C. J. Umrigar, The Journal of Chemical Physics **108**, 8838 (1998).

ACKNOWLEDGMENTS

AUTHOR CONTRIBUTIONS

M.T.E, Z.S, J.H and F.N designed the research. M.T.E, Z.S, P.A.E, J. H and F.N developed the method. M.T.E, Z.S and J.H wrote the computer code. M.T.E and Z.S carried out the numerical calculations. M.T.E, Z.S, P.A.E, J.H and F.N analyzed the data. M.T. E, Z.S, J.H and F.N wrote the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary material is available below.

Correspondence and requests for materials should be addressed to J. Hermann or F. Noé.

Appendix A: Geometries of test systems

All calculations presented in Sections II A – C of the main paper were done at the ground-state equilibrium geometries. The geometries for all molecules are listed in Supplementary Table I. For the ethylene structures presented in Section II D, we list the planar (ground-state equilibrium) and twisted ($\tau = 90$) geometries.

Supplementary Table I. Ground-state equilibrium geometries of the test systems.

Molecule	Atom	Position (Å)	Molecule	Atom	Position (Å)
LiH	Li	(0.000, 0.000, 0.000)	BeH	Be	(0.000000, 0.000000, 0.000000)
	H	(1.595, 0.000, 0.000)		H	(1.326903, 0.000000, 0.000000)
BH	B	(0.000000, 0.000000, 0.000000)	CH^+	C	(0.00000, 0.00000, 0.00000)
	H	(0.000000, 0.000000, 1.222874)		H	(1.13092, 0.00000, 0.00000)
H_2O	O	(0.000000, 0.000000, -0.069903)	CO	C	(0.000000, 0.000000, -0.661165)
	H	(0.000000, 0.757532, 0.518435)		O	(0.000000, 0.000000, 0.472379)
	H	(0.000000, -0.757532, 0.518435)			
NH_3	N	(0.067759, -0.000000, 0.000000)	C_2H_4 (planar)	C	(-0.675000, 0.000000, 0.000000)
	H	(-0.313823, 0.468746, -0.811891)		C	(0.675000, 0.000000, 0.000000)
	H	(-0.313823, -0.937491, -0.000000)		H	(-1.242900, 0.000000, -0.930370)
	H	(-0.313823, 0.468746, 0.811891)		H	(-1.242900, 0.000000, 0.930370)
				H	(1.242900, 0.000000, -0.930370)
				H	(1.242900, 0.000000, 0.930370)
C_2H_4 (twisted)	C	(-0.688500, 0.000000, 0.000000)	C_6H_6	C	(0.000000, 1.396792, 0.000000)
	C	(0.688500, 0.000000, 0.000000)		C	(0.000000, -1.396792, 0.000000)
	H	(-1.307207, 0.000000, -0.915547)		C	(1.209657, 0.698396, 0.000000)
	H	(-1.307207, 0.000000, 0.915547)		C	(-1.209657, -0.698396, 0.000000)
	H	(1.307207, -0.915547, 0.000000)		C	(-1.209657, 0.698396, 0.000000)
	H	(1.307207, 0.915547, 0.000000)		C	(1.209657, -0.698396, 0.000000)
				H	(0.000000, 2.484212, 0.000000)
				H	(2.151390, 1.242106, 0.000000)
				H	(-2.151390, -1.242106, 0.000000)
				H	(-2.151390, 1.242106, 0.000000)
				H	(2.151390, -1.242106, 0.000000)
				H	(0.000000, -2.484212, 0.000000)

Appendix B: Energies for small atoms and molecules

We tabulate the total energies and vertical excitation energies that were obtained for the small systems, as presented in Section II A of the main paper. These are listed in Supplementary Table II.

Supplementary Table II. Total energies and vertical excitation energies (both in Ha) for the small atoms in Section II A.

System	E_n	Correlation energy (%)	ΔE	ΔE (σ^2 matching)
Li	-7.47800(4)	99.9(1)[a]	—	—
	-7.40998(5)	—	0.06801(7)	0.06807(7)
	-7.40988(7)	—	0.06812(8)	0.06811(9)
	-7.40977(7)	—	0.06823(8)	0.06824(9)
	-7.335(4)	—	0.143(4)	0.143(4)
Be	-14.66736(7)	100.0(1)[a]	—	—
	-14.5663(6)	—	0.1011(6)	0.1009(6)
	-14.5664(1)	—	0.1010(1)	0.1007(2)
	-14.5664(1)	—	0.1010(1)	0.1004(2)
	-14.4710(2)	—	0.1964(2)	0.1952(2)
B	-24.6509(2)	97.6(1)[a]	—	—
	-24.6506(2)	—	0.0004(3)	0.0005(3)
	-24.6498(2)	—	0.0011(3)	0.0008(3)
	-24.5206(2)	—	0.1303(2)	0.1300(3)
	-24.5202(2)	—	0.1307(3)	0.1300(3)
C	-37.8388(3)	96.0(2)[a]	—	—
	-37.8388(2)	—	-0.0001(4)	0.0001(4)
	-37.8378(2)	—	0.0010(4)	0.0008(4)
	-37.7918(2)	—	0.0470(4)	0.0470(4)
	-37.7911(3)	—	0.0477(4)	0.0451(4)
N	-54.5836(3)	97.0(2)[a]	—	—
	-54.4911(4)	—	0.0925(5)	0.0900(5)
	-54.4885(4)	—	0.0952(5)	0.0893(6)
	-54.4874(4)	—	0.0963(6)	0.0883(6)
	-54.4861(4)	—	0.0975(5)	0.0918(6)
O	-75.0532(5)	94.6(2)[a]	—	—
	-75.0506(5)	—	0.0027(7)	-0.0003(8)
	-75.0519(5)	—	0.0013(7)	-0.0011(7)
	-74.9763(6)	—	0.0769(8)	0.0711(8)
	-74.9771(6)	—	0.0761(8)	0.0703(9)

[a] Exact energy and HF energy at the CBS limit [50]

Supplementary Table II (continued). Total energies and vertical excitation energies (both in Ha) for the many-state calculations in Section II A.

System	E_n	Correlation energy (%)	ΔE	ΔE (σ^2 matching)
LiH	-8.0695(1)	98.9(1)[a]	—	—
	-7.9497(1)	—	0.1198(1)	0.1199(1)
	-7.9348(1)	—	0.1348(2)	0.1347(2)
	-7.9112(1)	—	0.1584(2)	0.1573(2)
	-7.9114(1)	—	0.1582(1)	0.1574(2)
	-7.8978(1)	—	0.1717(2)	0.1710(2)
	-7.8977(1)	—	0.1718(2)	0.1709(2)
	-7.8549(1)	—	0.2147(2)	0.2130(2)
BeH	-15.2452(4)	98.3(4)[a]	—	—
	-15.1516(2)	—	0.0936(4)	0.0916(3)
	-15.1507(2)	—	0.0946(5)	0.0914(3)
	-15.0286(3)	—	0.2166(5)	0.2132(4)
	-15.0259(2)	—	0.2193(5)	0.2156(4)
	-15.0249(2)	—	0.2203(5)	0.2138(17)
	-14.9951(4)	—	0.2500(5)	0.2379(5)
	-14.9528(3)	—	0.2923(5)	0.2793(16)
Be	-14.6667(1)	99.3(1)[a]	—	—
	-14.5655(2)	—	0.1012(2)	0.1011(2)
	-14.5654(1)	—	0.1013(2)	0.1007(2)
	-14.5651(2)	—	0.1016(2)	0.1009(3)
	-14.4663(3)	—	0.2004(4)	0.1967(10)
	-14.4671(3)	—	0.1996(3)	0.1976(5)
	-14.4665(3)	—	0.2002(3)	0.1982(5)
	-14.3915(2)	—	0.2752(2)	0.2732(4)

[a] Exact energy and HF energy at the CBS limit [50]

Appendix C: Energies and oscillator strengths for intermediate systems

We tabulate the total energies, vertical excitation energies and oscillator strengths that were obtained for the intermediate systems, as presented in Section II B of the main paper. These are listed in Supplementary Table III.

Supplementary Table III. Total energies (in Ha), vertical excitation energies (in Ha) and oscillator strengths (in au) for the systems in Section II B.

Molecule	E_n	Correlation energy (%)	ΔE	ΔE (σ^2 matching)	f	f (σ^2 matching)
BH	-25.2883(2)	99.5(1)[a]	—	—	—	—
	-25.2361(2)	—	0.0522(3)	0.0511(3)	—	—
	-25.2356(2)	—	0.0527(3)	0.0503(3)	—	—
	-25.1765(2)	—	0.1118(3)	0.1088(3)	0.0276(1)	0.0253(1)
CH$^+$	-38.0863(2)	—	—	—	—	—
	-38.0385(2)	—	0.0478(3)	0.0460(3)	—	—
	-38.0375(3)	—	0.0488(4)	0.0465(4)	—	—
	-37.9664(3)	—	0.1199(4)	0.1172(4)	0.00606(4)	0.00576(3)
H$_2$O	-76.4230(5)	95.9(1)[b]	—	—	—	—
	-76.1499(6)	—	0.2731(8)	0.2702(8)	—	—
	-76.1343(6)	—	0.2887(8)	0.2845(9)	0.0395(3)	0.0374(3)
NH$_3$	-56.5533(3)	96.6(1)[b]	—	—	—	—
	-56.3122(4)	—	0.2411(5)	0.2391(6)	—	—
	-56.2972(4)	—	0.2561(5)	0.2531(6)	0.0630(4)	0.0731(5)
CO	-113.3039(6)	95.9(1)[b]	—	—	—	—
	-113.0618(7)	—	0.2421(9)	0.2386(9)	—	—
	-112.9765(7)	—	0.3274(9)	0.3211(9)	0.1048(5)	0.1030(5)

[a] Exact energy [51] and HF energy at the CBS limit [38]
[b] Exact energy and HF energy at the CBS limit [50]

Appendix D: Energies for benzene

We tabulate the total energies and vertical excitation energies that were obtained for benzene, as presented in Section II C of the main paper. These are listed in Supplementary Table IV.

Supplementary Table IV. Total energies and vertical excitation energies (both in Ha) for benzene in Section II C.

E_0	E_1	ΔE	ΔE (σ^2 matching)
-232.0675(11)	-231.8628(9)	0.2047(14)	0.1637(13)

Appendix E: Energies for ethylene

We tabulate the total energies and vertical excitation energies that were obtained for all ethylene structures, as presented in Section II D of the main paper. These are listed in Supplementary Table V.

Supplementary Table V. Total energies and vertical excitation energies (both in Ha) for ethylene in Section II D.

τ ($\phi = 0$)	E_n	ΔE	ΔE (σ^2 matching)	ϕ ($\tau = 90°$)	E_n	ΔE	ΔE (σ^2 matching)
0	-78.5650(4)	—	—	0	-78.4477(4)	—	—
	-78.2617(5)	0.3033(6)	0.2964(7)		-78.3578(5)	0.0899(6)	0.0897(7)
15	-78.5609(4)	—	—	20	-78.4471(4)	—	—
	-78.2755(5)	0.2854(6)	0.2799(7)		-78.3589(5)	0.0882(6)	0.0869(7)
30	-78.5461(4)	—	—	40	-78.4391(4)	—	—
	-78.2925(5)	0.2536(6)	0.2459(7)		-78.3661(5)	0.0730(6)	0.0735(6)
45	-78.5254(4)	—	—	60	-78.4267(5)	—	—
	-78.3171(5)	0.2083(6)	0.2032(7)		-78.3704(5)	0.0563(7)	0.0537(7)
60	-78.4952(4)	—	—	70	-78.4161(5)	—	—
	-78.3339(5)	0.1613(6)	0.1580(7)		-78.3721(5)	0.0440(7)	0.0428(7)
70	-78.4771(4)	—	—	80	-78.3992(4)	—	—
	-78.3416(5)	0.1355(6)	0.1324(7)		-78.3701(5)	0.0291(6)	0.0258(7)
80	-78.4540(5)	—	—	90	-78.3828(4)	—	—
	-78.3449(5)	0.1091(7)	0.1089(7)		-78.3675(5)	0.0153(6)	0.0112(7)
85	-78.4483(4)	—	—	95	-78.3697(4)	—	—
	-78.3512(5)	0.0971(6)	0.0922(7)		-78.3612(5)	0.0085(6)	0.0083(7)
90	-78.4425(5)	—	—	97.5	-78.3634(5)	—	—
	-78.3531(5)	0.0894(7)	0.0881(7)		-78.3593(5)	0.0041(7)	0.0023(7)
				100	-78.3587(5)	—	—
					-78.3559(5)	0.0028(7)	0.0014(7)
				102.5	-78.3552(5)	—	—
					-78.3529(5)	0.0023(7)	0.0042(7)
				105	-78.3509(5)	—	—
					-78.3448(5)	0.0061(7)	0.0048(7)
				110	-78.3449(5)	—	—
					-78.3224(5)	0.0225(7)	0.0189(7)
				120	-78.3284(5)	—	—
					-78.2929(5)	0.0355(7)	0.0231(7)

Appendix F: Overlaps of trained wavefunctions

For each system we compute the pairwise overlaps between all trained wavefunctions and tabulate the mean. These are listed in Supplementary Table VI.

Supplementary Table VI. Mean pairwise overlaps for all systems.

System	$\overline{|S_{ij}|}$	System	$\overline{|S_{ij}|}$	System	$\overline{|S_{ij}|}$	System	$\overline{|S_{ij}|}$
Li	0.0010(6)	CH$^+$	0.0018(7)	C$_2$H$_4$ ($\tau = 80$)	0.002(2)	C$_2$H$_4$ ($\phi = 95$)	0.002(1)
Be	0.004(1)	H$_2$O	0.007(3)	C$_2$H$_4$ ($\tau = 85$)	0.001(1)	C$_2$H$_4$ ($\phi = 97.5$)	0.003(2)
B	0.0023(9)	NH$_3$	0.014(3)	C$_2$H$_4$ ($\tau = 90$)	0.008(2)	C$_2$H$_4$ ($\phi = 100$)	0.002(3)
C	0.0030(4)	CO	0.003(1)	C$_2$H$_4$ ($\phi = 0$)	0.009(1)	C$_2$H$_4$ ($\phi = 102.5$)	0.007(1)
N	0.0040(5)	C$_2$H$_4$ ($\tau = 0$)	0.011(3)	C$_2$H$_4$ ($\phi = 20$)	0.002(2)	C$_2$H$_4$ ($\phi = 105$)	0.007(2)
O	0.0035(4)	C$_2$H$_4$ ($\tau = 15$)	0.007(3)	C$_2$H$_4$ ($\phi = 40$)	0.003(2)	C$_2$H$_4$ ($\phi = 110$)	0.009(1)
LiH	0.0047(9)	C$_2$H$_4$ ($\tau = 30$)	0.004(3)	C$_2$H$_4$ ($\phi = 60$)	0.001(1)	C$_2$H$_4$ ($\phi = 120$)	0.002(3)
BeH	0.0068(5)	C$_2$H$_4$ ($\tau = 45$)	0.003(2)	C$_2$H$_4$ ($\phi = 70$)	0.002(2)	C$_6$H$_6$	0.001(4)
Be (many states)	0.0035(5)	C$_2$H$_4$ ($\tau = 60$)	0.008(2)	C$_2$H$_4$ ($\phi = 80$)	0.002(3)		
BH	0.0033(9)	C$_2$H$_4$ ($\tau = 70$)	0.002(2)	C$_2$H$_4$ ($\phi = 90$)	0.006(2)		

Appendix G: Hyperparameters

We tabulate the hyperparameters that were used in all calculations. These are listed in Supplementary Table VII. We also tabulate the active spaces and spin configurations (S^2, $M = (N_\uparrow - N_\downarrow)/2$ (eigenvalue of $\mathcal{S}_z$)) used to generate the CASSCF baselines. These are listed in Supplementary Table VIII.

Supplementary Table VII. Hyperparameters used in calculations. (See ref. 29 for more details.)

Hyperparameter	Value	Hyperparameter	Value
One-electron basis	6-311G	Maximum number of determinants	10
Dimension of $\mathbf{e}$ (# distance features)	32	Dimension of $\mathbf{x}_i$ (embedding dimension)	128
Dimension of $\mathbf{z}_i$ (kernel dimension)	64	Number of layers in $\mathbf{w}_{\boldsymbol{\theta}}$	1
Number of layers in $\mathbf{h}_{\boldsymbol{\theta}}$	2	Number of layers in $\mathbf{g}_{\boldsymbol{\theta}}$	2
Number of interaction layers L:		Number of layers in $\eta_{\boldsymbol{\theta}}$:	
small/intermediate systems & ethylene	4	small/intermediate systems & ethylene	3
benzene	5	benzene	5
Number of layers in $\kappa_{\boldsymbol{\theta}}$:		Number of walkers:	
small/intermediate systems & ethylene	3	small/intermediate systems & ethylene	1000
benzene	5	benzene	400
Batch size	2000	Number of equilibration steps	500
Number of training steps:		Optimizer	AdamW
small systems	5000/10000	Learning rate scheduler	CyclicLR
intermediate systems	20000	Minimum/maximum learning rate:	
ethylene	10000	small/intermediate systems & ethylene	0.0001/0.005
benzene	10000	benzene	0.0001/0.001
Cyclic frequency	1000	Clipping window q	5
Minimum/maximum α	0.25/1.0	Epoch size	10
Resampling frequency	3	Number of decorrelation sampling steps:	
Target acceptance	57%	small/intermediate systems & ethylene	5
		benzene	15

Supplementary Table VIII. The active spaces, specifying N electrons across P orbitals, and spin configurations, used to generate the CASSCF baselines for all systems.

System	CAS(P, N)	Spin(S^2, M)	System	CAS(P, N)	Spin(S^2, M)
Li	$(5, 3)$	$(\frac{3}{4}/\frac{3}{4}/\frac{3}{4}/\frac{3}{4}/\frac{3}{4}, 1)$	Be (many states)	$(6, 4)$	$(0/2/2/2/0/0/0/2, 1)$
Be	$(5, 4)$	$(0/2/2/2/0, 0)$	BH	$(5, 2)$	$(0/2/2/0, 0)$
B	$(5, 5)$	$(\frac{3}{4}/\frac{3}{4}/\frac{3}{4}/\frac{15}{4}/\frac{15}{4}, 1)$	CH$^+$	$(5, 2)$	$(0/2/2/0, 0)$
C	$(5, 6)$	$(2/2/2/0/0, 0)$	H$_2$O	$(5, 2)$	$(0/2/0, 0)$
N	$(5, 7)$	$(\frac{15}{4}/\frac{3}{4}/\frac{3}{4}/\frac{3}{4}/\frac{3}{4}, 1)$	NH$_3$	$(7, 2)$	$(0/2/0, 0)$
O	$(5, 8)$	$(2/2/2/0/0, 0)$	CO	$(10, 10)$	$(0/2/0, 0)$
LiH	$(12, 4)$	$(0/2/0/2/2/0/0/2, 0)$	C$_2$H$_4$	$(4, 4)$	$(0/0, 0)$
BeH	$(12, 5)$	$(\frac{3}{4}/\frac{3}{4}/\frac{3}{4}/\frac{3}{4}/\frac{15}{4}/\frac{15}{4}/\frac{3}{4}/\frac{3}{4}, 1)$	C$_6$H$_6$	$(4, 4)$	$(0/2, 0)$