Accurate quantum Monte Carlo forces for machine-learned force fields: Ethanol as a benchmark

E. Slootman,¹ I. Poltavsky,² R. Shinde,¹ J. Cocomello,¹ S. Moroni,^{3,*} A. Tkatchenko,^{2,†} and C. Filippi^{1,‡}

¹Computational Chemical Physics, MESA⁺ Institute for Nanotechnology,

University of Twente, PO Box 217, 7500 AE Enschede, The Netherlands

²Physics and Materials Science Research Unit, University of Luxembourg, L-1511 Luxembourg, Luxembourg

³CNR-IOM DEMOCRITOS, Istituto Officina dei Materiali,

and SISSA Scuola Internazionale Superiore di Studi Avanzati, Via Bonomea 265, I-34136 Trieste, Italy

(Dated: April 16, 2024)

Quantum Monte Carlo (QMC) is a powerful method to calculate accurate energies and forces for molecular systems. In this work, we demonstrate how we can obtain accurate QMC forces for the fluxional ethanol molecule at room temperature by using either multi-determinant Jastrow-Slater wave functions in variational Monte Carlo or just a single determinant in diffusion Monte Carlo. The excellent performance of our protocols is assessed against high-level coupled cluster calculations on a diverse set of representative configurations of the system. Finally, we train machine-learning force fields on the QMC forces and compare them to models trained on coupled cluster reference data, showing that a force field based on the diffusion Monte Carlo forces with a single determinant can faithfully reproduce coupled cluster power spectra in molecular dynamics simulations.

I. INTRODUCTION

Accurate forces are crucial to perform geometry relaxation and molecular dynamics (MD) simulations. Classical force fields, which are widely used for such purpose. are often parameterized to reproduce quantum chemical data obtained with approaches such as coupled cluster (CC) or density functional theory (DFT). Unfortunately, these force fields cannot always easily capture effects which are fundamentally quantum mechanical. Moreover, their accuracy is intrinsically limited by the predefined functional form, which is in general unknown. For instance, for a system as simple as ethanol at room temperature, MD trajectories based on classical force fields like AMBER [1] cannot faithfully explore the potentialenergy surface. Consequently, the resulting dynamics does not correctly sample the statistical occupational weights of the hydroxyl rotor group [2].

Machine-learning (ML) force fields [3] enable performing long MD simulations of *ab initio* quality without the need for expensive quantum chemical calculations at every time step, given a sufficient amount of training data. These ML models are often trained on DFT energies and forces. [2, 4–19] Unfortunately, such a procedure can be unreliable due to the use of approximate functionals as, for instance, whenever additional corrections for DFT must be introduced to capture dispersion interactions. Then, the accuracy of the DFT reference data must be assessed against highly correlated methods such as coupled cluster (CC) approaches. The most accurate flavors of coupled cluster are however computationally demanding and therefore limited to relatively small molecules. Quantum Monte Carlo (QMC) calculations can be instrumental in generating the needed reference data for accurate machine-learning potentials. Although QMC is computationally expensive, it provides highly accurate energies and forces, and scales favorably with system size also when forces are computed [20–22]. Calculating atomic forces in quantum Monte Carlo has been an active field of research and different algorithms and approximations have been put forward for this purpose [23– 30]. The use of QMC to construct machine-learning force fields is a relatively new field that has seen applications in the description of high-pressure hydrogen [31–33] and in molecular systems [34, 35]. Recently, the effect of the statistical noise on the resulting potentials has also been investigated [36].

Here, we show how QMC can yield forces as accurate as those computed with the "golden standard" of quantum chemistry, CCSD(T), over a large set of configurations of the fluxional ethanol molecule at room temperature. In particular, competitive accuracy can be obtained either in variational Monte Carlo (VMC) using multi-determinant wave functions or in diffusion Monte Carlo (DMC) with the affordable variational-driftdiffusion approximation [27, 28] and just a single determinant. Since ethanol is characterized by weak intramolecular interactions, we also compare our results with DFT calculations treating dispersion interactions with the Tkatchenko-Scheffler (TS) [37] or the many-body dispersion (MBD) [38] approaches. Finally, we demonstrate the very good performance of the ML potentials trained on QMC forces using the sGDML model [5] on unseen test datasets as well as by reproducing the power spectra obtained from MD simulations with CCSD(T) models.

The manuscript is organized as follows. The algorithms to compute the QMC forces and the choice of wave function are described in Sec. II and the computational details given in Sec. III. The QMC results and the performance of corresponding ML potentials are dis-

^{*} moroni@democritos.it

 $^{^{\}dagger}$ alexandre.tkatchenko@uni.lu

[‡] c.filippi@utwente.nl

cussed in Sec. IV. We conclude in Sec. V.

II. METHOD

A. QMC forces

In QMC [39–41], the energy is computed as

$$E = \int d\mathbf{R} E_L(\mathbf{R}) P(\mathbf{R}) \equiv \langle E_L \rangle_P, \qquad (1)$$

where **R** is the coordinates of the electrons, $E_L(\mathbf{R}) = \mathcal{H}\Psi(\mathbf{R})/\Psi(\mathbf{R})$ is the local energy for a given trial wave function $\Psi(\mathbf{R})$, and $P(\mathbf{R})$ is the probability distribution sampled in the QMC run. In VMC, this is equal to $P_{\text{VMC}}(\mathbf{R}) = |\Psi(\mathbf{R})|^2 / \int d\mathbf{R} |\Psi(\mathbf{R})|^2$ and, in DMC, $P_{\text{DMC}}(\mathbf{R}) = \Phi(\mathbf{R})\Psi(\mathbf{R}) / \int d\mathbf{R} \Phi(\mathbf{R})\Psi(\mathbf{R})$ where $\Phi(\mathbf{R})$ is the fixed-node solution.

The nuclear forces are obtained by taking the derivative of the energy with respect to the nuclear coordinates,

$$F = -\nabla_{\alpha} E$$
(2)
= $-\langle \nabla_{\alpha} E_L(\mathbf{R}) + (E_L(\mathbf{R}) - E) \nabla_{\alpha} \ln P(\mathbf{R}) \rangle_P.$

While the derivative of the distribution function in VMC can be readily performed to compute forces, the distribution function in DMC is not known in closed form but is sampled via a stochastic implementation of the power method through the repeated application of the importance sampled Green function $\mathcal{G}(\mathbf{R}', \mathbf{R}) = \Psi(\mathbf{R}')\langle \mathbf{R}' | \exp[-\tau(\mathcal{H} - E_{\mathrm{T}})] | \mathbf{R} \rangle / \Psi(\mathbf{R})$ with τ the timestep and E_{T} an energy shift. Therefore, once equilibrium is reached, P_{DMC} is given by

$$P_{\text{DMC}}(\mathbf{R}_n) = \int d\mathbf{R}_{n-1} \dots d\mathbf{R}_{n-k} \qquad (3)$$
$$\times \prod_{i=n-k}^{n-1} \mathcal{G}(\mathbf{R}_{i+1}, \mathbf{R}_i) P_{\text{DMC}}(\mathbf{R}_{n-k}),$$

where n is the last iteration. The nuclear force in DMC can then be rewritten as

$$F_{\text{DMC}} = - \langle \nabla_{\alpha} E_L(\mathbf{R}_n) + [E_L(\mathbf{R}_n) - E] \qquad (4)$$
$$\times \sum_{i=n-k_{\text{hist}}}^{n-1} \nabla_{\alpha} \ln \mathcal{G}(\mathbf{R}_{i+1}, \mathbf{R}_i) \rangle_{P_{\text{DMC}}}.$$

where k_{hist} has to be larger than the correlation time between E_L and $\nabla_{\alpha} \ln \mathcal{G}$ along the random walk [28]. The importance-sampled Green function must be approximated and, in the limit of small time-steps, becomes

$$\mathcal{G}(\mathbf{R}',\mathbf{R}) = \frac{e^{-[\mathbf{R}'-\mathbf{R}-V(\mathbf{R})\tau]^2/2\tau}e^{S(\mathbf{R}',\mathbf{R})}}{(2\pi\tau)^{3N/2}},\qquad(5)$$

where $V(\mathbf{R}) = \nabla \Psi(\mathbf{R}) / \Psi(\mathbf{R})$ and $S(\mathbf{R}', \mathbf{R}) = \tau \{ E_{\mathrm{T}} - [E_{\mathrm{L}}(\mathbf{R}') + E_{\mathrm{L}}(\mathbf{R})]/2 \}$. Modified expressions of V and S

are used in actual calculations [42] and the bias due to the short-time approximation can be removed by extrapolating the results at zero time-step.

While it is possible to compute forces in DMC which are fully compatible with the derivative of the fixed-node DMC energy at any given time-step [29], the derivative of the drift-diffusion part of the Green function introduces larger fluctuations in the force estimator. Therefore, we consider here an estimator of the DMC force in the so-called variational drift-diffusion (VD) approximation [27, 28], which only includes the derivative of the branching factor and approximates the derivative of the drift-diffusion contribution by the VMC estimator,

$$F_{\rm VD} = -\langle \nabla_{\alpha} E_L(\mathbf{R}_n) + [E_L(\mathbf{R}_n) - E] \qquad (6)$$

$$\times [\nabla_{\alpha} P_{\rm VMC}(\mathbf{R}_n) + \sum_{i=n-k_{\rm hist}}^{n-1} \nabla_{\alpha} S(\mathbf{R}_{i+1}, \mathbf{R}_i)] \rangle_{P_{\rm DMC}}$$

Intuitively, this approximation can be derived by regarding the random walk in the standard DMC algorithm (drift and diffuse a walker, accept or reject the move, and reweight by the branching factor) as simply reweighting the VMC distribution by the branching factor.

Computationally, the VD approximation comes at no additional cost since the energy derivatives required for the sum have already been calculated at earlier timesteps. Furthermore, as shown in our calculations, the statistical fluctuations in the VD forces are nearly the same as those obtained when computing the even simpler approximate DMC force introduced by Reynolds *et al.* (RE) [23], which computes the VMC force estimator on the DMC distribution,

$$F_{\rm RE} = -\langle \nabla_{\alpha} E_L(\mathbf{R}) + [E_L(\mathbf{R}) - E] \qquad (7)$$
$$\times \ \nabla_{\alpha} \ln P_{\rm VMC}(\mathbf{R}) \rangle_{P_{\rm DMC}} .$$

This approximation can be partially corrected by considering the generalized hybrid estimator, $F_{\text{RE-hybrid}} = 2F_{\text{RE}} - F_{\text{VMC}}$ at the cost of increased statistical fluctuations [25].

In general, in addition to the explicit dependence of the energy on the nuclear coordinates through the potential and the trial wave function when an atom-centered basis set is used, there is an implicit dependence through the variational parameters, p_i . Consequently, the force acquires an additional term, namely,

$$F = -\frac{\partial E}{\partial \alpha} - \sum_{i} \frac{\partial E}{\partial p_{i}} \frac{\partial p_{i}}{\partial \alpha}, \qquad (8)$$

where the second term vanishes if the energy is optimal with respect to the parameter variations. Since we fully optimize the wave function in energy minimization at the VMC level, this additional contribution is equal to zero and our VMC forces are fully consistent with the corresponding energy. In DMC, neglecting this term leads in principle to a bias in the corresponding forces, which has however been shown to be quite small if the wave function is fully optimized in VMC, or partially optimized but of sufficient quality like when a multi-determinant expansion is employed [28].

Finally, all force estimators described above obey a zero-variance principle in the limit of the trail wave function and its derivatives being exact but, for an approximate trial function, display an infinite variance. In VMC, to cure this problem, we employ a guiding wave function which differs from the trial function close to the nodes and is finite at the nodes [43], where we use $d = |\nabla \Psi/\Psi|$ as a measure of the distance from the nodes. While it is possible to adapt this regularization to the computation of DMC forces, this has the downside of promoting walkers close to the nodes. Therefore, in the computation of DMC forces, we adopt instead the reguralization scheme from Ref. 44, where the force estimator is simply multiplied by a function $f_{\epsilon}(x) = 9x^2 - 15x^4 + 7x^6$ if $x = d/\epsilon < 1$ and ϵ is chosen sufficiently small to have a negligible bias.

B. Trial wave function

We employ so-called Jastrow-Slater wave functions of the form

$$\Psi = \mathcal{J} \sum_{i} c_i \mathcal{D}_i, \qquad (9)$$

where \mathcal{D}_i are Slater determinants of single-particle orbitals and \mathcal{J} is the Jastrow correlation factor, which contains electron-electron and electron-nucleus correlation terms [45]. All wave function parameters (Jastrow, orbital, and linear coefficients) are fully optimized in energy minimization at the VMC level.

The determinantal component is here either a single determinant or a multi-determinant expansion generated in an automatic manner with the configuration interaction using a perturbative selection made iteratively (CIPSI) approach [46]. Starting from a wave function expanded on a set of determinants in a given space S,

$$\Psi^{\text{CIPSI}} = \sum_{\mathcal{D}_i \in S} c_i \mathcal{D}_i \,, \tag{10}$$

the approach builds expansions by iteratively selecting determinants based on their second-order perturbation (PT2) energy contribution obtained via the Epstein-Nesbet partitioning of the Hamiltonian [47, 48],

$$\delta E_{\gamma}^{(2)} = \frac{|\langle \gamma | \mathcal{H} | \Psi^{\text{CIPSI}} \rangle|^2}{\langle \Psi^{\text{CIPSI}} | \mathcal{H} | \Psi^{\text{CIPSI}} \rangle - \langle \gamma | \mathcal{H} | \gamma \rangle}, \qquad (11)$$

where $|\gamma\rangle$ denotes a determinant outside the current CI space that is connected to S by \mathcal{H} . The total PT2 energy contribution, $E^{(\text{PT2})}$, goes to zero as the expansion approaches the full CI (FCI) limit.

We are here interested in computing forces on different structural configurations and want to achieve a balanced CIPSI description of the determinantal component of the QMC wave function across the ground-state potential energy surface of ethanol. As a measure of the quality of a CIPSI wave function, we use its PT2 energy contribution, which represents an approximate estimate of the error of the expansion with respect to FCI. Therefore, given the chosen expansion and its energy PT2 correction for an arbitrary reference configuration, we generate expansions for the other configurations by matching the reference $E^{(\rm PT2)}$. In general, the procedure will result in expansions of different length at the different geometries.

III. COMPUTATIONAL DETAILS

The QMC calculations are carried out with the CHAMP code [49]. We employ scalar-relativistic energy consistent Hartree-Fock pseudopotentials and the correlated-consistent Gaussian basis sets specifically constructed for these pseudopotentials [50, 51]. For most calculations, we use the cc-pVTZ basis set and perform convergence tests with the cc-pVQZ basis set. As shown in Table S1 for a representative configuration and a single-determinant wave function, the use of a cc-pVTZ basis yields VMC forces which are converged with respect to the basis set.

All wave function parameters (Jastrow, orbital, and CI coefficients) are optimized by minimizing the energy in VMC using the stochastic reconfiguration method [52] in a low-memory implementation [53]. To cure the diverging variance of the force estimator, we employ a node cutoff parameter ϵ of 0.1 a.u. in VMC and 0.05 a.u. in DMC. In the DMC calculations, we treat the pseudopotentials beyond the locality approximation using the T-move algorithm [54] and employ a time-step of 0.005 a.u. which ensures converged VD forces as shown in Fig. S5. A value of 900 is used for k_{hist} (Eq. (6)) and the dependence of the VD forces on this parameter is illustrated in Fig. S6. In the regularization procedure [44], our choice of 0.05 a.u. for ϵ yields a negligible bias compared to the statistical error as shown in Section S3.

We perform the HF calculations with the program GAMESS(US) [55] and generate the CIPSI wave functions with Quantum Package [56] using the same pseudopotential and basis sets as in QMC. The interface of both these codes with CHAMP uses the TREXIO library [57]. The Psi4 package [58] is employed for the all-electron coupled cluster calculations with Dunning's correlated consistent basis sets [59].

The machine learning models for ethanol use 100 training and 100 validation configurations (thereafter referred as set A), obtained by clustering a set of 2000 representative configurations (set B) down to 200 (set A), based on their geometry and energy. To this aim, we first split the 2000 configurations into 40 clusters based on their geometry, using the agglomerative clustering algorithm and, then, split each cluster into 5 clusters based on energy using the k-means method. Afterward, configurations closest to the centroids of each cluster are selected to form the training set (see Ref. 60 for more details). The initial 2000 configurations (set B) were extracted from a long MD trajectory based on DFT calculations with the PBE-TS functional [4] by sampling them according to the energy distribution in this trajectory. For this purpose, we use the implementation from the symmetric gradient domain machine learning (sGDML) software package [2]: the energies of the configurations are histogrammed and a number of configurations proportional to the height of the histogram is then selected randomly within each bin. A second set of 2000 configurations (set C) is clustered from the complete MD trajectory according to the procedure followed for set A. Therefore, in contrast to set B, set C equally represents different possible molecular geometries and energy states irrespective of their statistical probability in the reference dataset.

The sGDML models are trained on set A with energy and forces computed with different *ab initio* methods, namely, QMC (i.e. VMC, RE, RE-hybrid, and VD DMC), CCSD(T) with the cc-pVTZ and cc-pVQZ basis sets, and DFT PBE-TS and PBE0-MBD. For the DFT and CCSD(T)/cc-pVTZ, sGDML models are also trained on the larger set B. The error of the obtained force fields is analyzed using the open-source FFAST software package [61].

The classical MD simulations are carried out with a time-step of 0.2 fs at a temperature of 300 K employing a Langevin thermostat with a time constant of 100 fs, using the i-PI package [62]. The total duration of the MD trajectories is 0.6 ns.

IV. RESULTS AND DISCUSSION

We demonstrate the performance of QMC forces on the fluxional ethanol molecule, characterized by intramolecular dispersion forces between the the hydroxyl and methyl rotors, namely, between the lone pairs of the oxygen and the partially positive charges of the hydrogen atoms. We compute here the QMC forces with different algorithms and wave functions, and discuss the impact of these choices on the corresponding ML models constructed with the sGDML framework for which ethanol is particularly suitable given its many symmetries. We also compare the QMC results to those obtained with two DFT functionals, namely, PBE-TS and PBE0-MBD.

As reference, we calculate the CCSD(T) forces also with a cc-pVQZ basis on set A, while, on the larger set B, we only perform the CCSD(T) calculations with the smaller cc-pVTZ basis set. We discuss the basis set convergence of the CC results below and in Section S2.

A. Quality of the forces

We begin our investigation by analyzing in detail the behavior of the various methods on seven representa-



FIG. 1. a) Mean absolute deviation (MAD, kcal/mol/Å) of the forces computed with different methods, compared to CCSD(T)/cc-pVQZ for 7 representative configurations of room-temperature ethanol; b) average MAD for each method; c) zoomed-in version of the MADs with their statistical errors.

tive configurations (selected to represent a variety of molecular geometries and energies within subset B) of ethanol at room temperature. In Fig. 1, we show the mean absolute deviation (MAD) of the forces with respect to CCSD(T)/cc-pVQZ for each configuration as well as the average MAD over the seven configurations. Of the methods investigated here, PBE-TS is the least accurate, while the use of PBE0-MBD yields significantly higher accuracy for this system. Moreover, PBE0-MBD demonstrates a relatively small dependence of the MAD upon the specific configuration.

VMC forces with a one-determinant wave function display a significant error that lies between the two DFT methods. Using the CIPSI procedure to go beyond a single determinant, we construct two expansions for each configuration, matching two different values of the total PT2 energy correction to ensure a consistent quality of the wave function across different geometries. More specifically, for configuration 2, we generate two expansions of about 500 and 2500 determinants, yielding a PT2 correction of -0.676 and -0.639 a.u., respectively, and use these two energy values as target in the CIPSI generation at the other configurations. The number of determinants in the resulting expansions ranges between 309-995 and 2098-3484, respectively. Further information on the convergence of the QMC results as a function of determinantal number is given in Section S5.

The results obtained with the CIPSI-based fullyoptimized Jastrow-Slater wave functions are shown in Fig. 1 and denoted for simplicity as "500 det" and "2500 det". At the VMC level, the 500-det wave function yields a big improvement on the one-determinant forces, surpassing the PBE0-MBD results. Further enlargening the expansion with the use of the 2500-det wave functions improves only marginally the accuracy. The relative flatness of the 500-det and 2500-det VMC lines for different geometries is a clear indication of the success of the PT2matching construction in yielding determinantal expansions of comparable quality when employed in a Jastrow-Slater wave function.

When carrying out DMC calculations on these VMCoptimized wave functions, we find that the VD forces perform very well already in the one-determinant case. On the contrary, the RE forces show some improvement over VMC but do not beat the accuracy of DFT/PBE0-MBD. Correcting these forces via the RE-hybrid estimator brings the forces close to the VD ones at the expense of larger statistical fluctuations (see Fig. 1c). The use of DMC-VD in combination with the multi-determinant wave functions shows in general no further, significant improvement compared to the one-determinant VD case: The VMC and DMC-VD forces for the multi-determinant wave functions and the DMC-VD forces for the onedeterminant wave function, have roughly the same MAD with respect to CCSD(T).

The one-determinant DMC-VD case for configuration 3 is clearly an outlier, displaying a larger deviation from the reference. This can be explained by inspecting the geometry of the molecular configuration (shown in Fig. S1), which is quite distorted with an angle of the methyl group characteristic of a region near a barrier in the potential energy surface. The wave function of such a configuration has therefore a more correlated character and must include multiple determinants to be accurately described. In fact, the MAD in DMC-VD for configuration 3 reduces when enlargening the expansion from one to 995 and further to 3484 determinants. he accuracy of the QMC calculations has been pushed to a level where the remaining discrepancy of the forces with respect to the all-electron CCSD(T)/cc-pVQZ results can be attributed to the use of pseudopotentials in QMC, and/or to residual basis set errors in CCSD(T), as further elaborated in Section S2.

To verify the robustness of these findings over a larger dataset, VMC and DMC calculations are performed with a one-determinant fully-optimized Jastrow-Slater wave function on 200 representative configurations (set A) of



FIG. 2. Mean absolute deviation (kcal/mol/Å) of the QMC and DFT forces with respect to CCSD(T)/cc-pVQZ for 200 configurations (set A) of room-temperature ethanol.

room-temperature ethanol. The quality of the QMC results is again assessed against CCSD(T)/cc-pVQZ and also compared to the outcome of the DFT calculations. The MADs of all configurations with respect to coupled cluster are plotted in Fig. 2 and the average MAD values reported in Table I.

The results show the same pattern as observed for the 7 configurations of Fig. 1, corroborating the findings above. In particular, VD-DMC lowers the errors and their spread compared to the VMC and the RE forces, and shows once again that, for this system, the one-determinant DMC VD forces are very accurate. The use of RE-hybrid offer a relatively large improvement on the RE forces. However, since it comes with a statistical error more than twice as large, there is no real use case for this method.

Method	MAD
VMC 1 det	3.055(2)
DMC RE 1 det	1.920(4)
RE-hybrid	1.046(8)
DMC VD 1 det	0.899(4)
PBE-TS	5.181
PBE0-MBD	1.431

TABLE I. Average mean absolute deviation (kcal/mol/Å) of the different methods with respect to CCSD(T)/cc-pVQZ over 200 configurations (set A) of ethanol. The statistical error on the last digit is indicated in brackets.

B. Effect on Machine-Learning Force Fields

With the forces computed with the different methods on the 200 configurations of set A (Fig. 2), we generate ML force fields using the sGDML model, using half of the data as training and the other half as validation points. For CCSD(T)/cc-pVTZ, PBE-TS, and PBE0-MBD, we





FIG. 3. Vibrational spectra of ethanol at room temperature computed with the various ML models (blue) compared to the CCSD(T)/cc-pVQZ (black) vibrational spectrum.

show in Section S6 that constructing the ML force fields on a larger set of configurations does not affect the relative quality of the models.

∕ datasot	$A \subset B$	$A \subset B$	B	C
madal	200	200	2000	2000
moder	Q	Т	Т	Т
VMC	3.2	3.4	3.4	3.4
RE	2.2	2.4	2.5	2.6
RE Hybrid	1.5	1.8	2.0	2.2
VD	1.2	1.4	1.6	1.8
PBE-TS	5.3	5.3	5.3	5.3
PBE0-MBD	1.7	1.9	2.0	2.1
CCSD(T)/cc-pVTZ	1.0	0.7	1.2	1.4
CCSD(T)/cc-pVQZ	0.7	1.0	1.4	1.5

TABLE II. Mean absolute deviation (kcal/mol/Å) of the forces obtained from the ML models on different datasets (A, B, C) against CCSD(T)/cc-pVXZ forces with X=T, Q. These values are calculated with the FFAST software [61].

The performance of the ML models is assessed on three sets of configurations $(A \subset B, B, \text{ and } C)$ by computing the MAD of the ML forces with respect to the CCSD(T) values calculated with either the cc-pVTZ (sets A, B, and C) or the larger cc-pVQZ basis set (set A). Using coupled cluster with the smaller basis as reference on set A leaves the ordering of the MADs unchanged, justifying the use of cc-pVTZ to evaluate the CCSD(T) reference on the larger B and C sets as shown in Table II. For all datasets, we find that the quality of the ML models nicely follows the quality of the underlying *ab initio* forces as depicted in Fig. 2. Not surprisingly, CCSD(T) displays the smallest MAD since the reference values are computed using the same method. Note that the mean absolute errors of the ML models on the validation sets are about 1.2–1.3 kcal/mol/Å (see Table S5). A difference of the same magnitude between the force field predictions and the reference data is therefore not significant for practical applications.

Importantly, we test the ML models on a dataset of 2000 configuration (set C), which is totally independent of the datasets (A and, in Section S6, B) used to generate the force fields. This test further confirms that the relative performance of the ML models follows the accuracy of the *ab initio* forces. Also on this dataset, we find that the model based on DMC-VD forces yields a smaller MAD than the ones constructed with VMC, other DMC approximations, and DFT.

Finally, to further analyze the behavior of the force field models, we compute the vibrational spectra from the velocity autocorrelation functions in classical MD simulations at room temperature. These can lead the system to regions of the potential energy surface which are not well sampled in the testing datasets. The spectra are shown in Fig. 3 and compared to the one obtained with the model trained on CCSD(T)/cc-pVQZ. As regards QMC, we observe again a gradual increase of accuracy moving from VMC, to DMC-RE, and, finally, to DMC VD. This is clearly visible in the overall shift of the spectrum and, in particular, of the C-H vibrational peaks around 3000 cm^{-1} , which are clearly overestimated by the VMC model. We note that also DMC-Hybrid and PBE0-MBD perform rather similarly to DMC-VD, while PBE-TS model underestimates the vibrational frequencies.

V. CONCLUSION

We have investigated the use of different algorithms and wave functions for the calculation of forces in QMC for ethanol at room temperature. For this system, a multi-determinant wave function in VMC is found to yield forces of comparable quality to those obtained with a single-determinant wave function and the DMC-VD approach. In both cases, the forces are in excellent agreement with the CCSD(T) values on a representative set of configurations. Employing the generalized hybrid estimator of the RE-Hybrid method also leads to accurate forces but is of less practical use due to the larger statistical error. Finally, we demonstrated the ability to train accurate machine-learning force fields using QMC. In particular, the sGDML model trained on single-determinant DMC-VD forces is shown to faithfully reproduce the vibrational spectrum of ethanol at room temperature obtained in molecular dynamics sim-

- D. Case, H. Aktulga, K. Belfon, I. Ben-Shalom, J. Berryman, S. Brozell, D. Cerutti, T. Cheatham, G. Cisneros, V. Cruzeiro, T. Darden, N. Forouzesh, G. Giambaşu, T. Giese, M. Gilson, H. Gohlke, A. Goetz, J. Harris, S. Izadi, S. Izmailov, K. Kasavajhala, M. Kaymak, E. King, A. Kovalenko, T. Kurtzman, T. Lee, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, M. Machado, V. Man, M. Manathunga, K. Merz, Y. Miao, O. Mikhailovskii, G. Monard, H. Nguyen, K. O'Hearn, A. Onufriev, F. Pan, S. Pantano, R. Qi, A. Rahnamoun, D. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, A. Shajan, J. Shen, C. Simmerling, N. Skrynnikov, J. Smith, J. Swails, R. Walker, J. Wang, J. Wang, H. Wei, X. Wu, Y. Wu, Y. Xiong, Y. Xue, D. York, S. Zhao, Q. Zhu, and P. Kollman, Amber 2023 (The Amber Project, 2023).
- [2] S. Chmiela, H. E. Sauceda, K.-R. Müller, and A. Tkatchenko, Towards exact molecular dynamics simulations with machine-learned force fields, Nat. Commun. 9, 3887 (2018).
- [3] O. T. Unke, S. Chmiela, H. E. Sauceda, M. Gastegger, I. Poltavsky, K. T. Schütt, A. Tkatchenko, and K.-R. Müller, Machine Learning Force Fields, Chem. Rev. 121, 10142 (2021).
- [4] S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt, and K.-R. Müller, Machine learning of accurate energy-conserving molecular force fields, Sci. Adv. 3, e1603015 (2017).
- [5] S. Chmiela, H. E. Sauceda, I. Poltavsky, K.-R. Müller, and A. Tkatchenko, sGDML: Constructing accurate and data efficient molecular force fields using machine learning, Comput. Phys. Commun. 240, 38 (2019).
- [6] S. Chmiela, V. Vassilev-Galindo, O. T. Unke, A. Kabylda, H. E. Sauceda, A. Tkatchenko, and K.-R. Müller, Accurate global machine learning force fields for molecules with hundreds of atoms, Sci. Adv. 9,

ulations with the CCSD(T)-based model. These findings unveil the potential that QMC methods offer in providing forces as reference data for machine-learning force fields, being as accurate as coupled cluster calculations and yet computationally applicable to large molecular systems.

ACKNOWLEDGMENTS

E.S., R.S., S.M. and C.F. acknowledge partial support by the European Centre of Excellence in Exascale Computing TREX — Targeting Real Chemical Accuracy at the Exascale. This project has received funding in part from the European Union's Horizon 2020 — Research and Innovation program — under grant agreement no. 952165. I.P. and A.T. acknowledge funding in whole, or in part, by the Luxembourg National Research Fund (FNR), grants reference C19/MS/13718694/QML-FLEX and INTER/MERA22/16521502/PHANTASTIC. The calculations were carried out on the Dutch national supercomputer Snellius with the support of SURF Cooperative.

eadf0873 (2023).

- [7] S. Batzner, A. Musaelian, L. Sun, M. Geiger, J. P. Mailoa, M. Kornbluth, N. Molinari, T. E. Smidt, and B. Kozinsky, E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials, Nat. Commun. 13, 2453 (2022).
- [8] Z. Qiao, M. Welborn, A. Anandkumar, F. R. Manby, and T. F. Miller, III, OrbNet: Deep learning for quantum chemistry using symmetry-adapted atomic-orbital features, J. Chem. Phys. **153**, 124111 (2020).
- [9] O. T. Unke and M. Meuwly, PhysNet: A Neural Network for Predicting Energies, Forces, Dipole Moments, and Partial Charges, J. Chem. Theory Comput. 15, 3678 (2019).
- [10] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, and A. Tkatchenko, Quantum-chemical insights from deep tensor neural networks, Nat. Commun. 8, 13890 (2017).
- [11] J. S. Smith, O. Isayev, and A. E. Roitberg, ANI-1: An extensible neural network potential with DFT accuracy at force field computational cost, Chem. Sci. 8, 3192 (2017).
- [12] T. Frank, O. Unke, and K.-R. Müller, So3krates: Equivariant attention for interactions on arbitrary lengthscales in molecular systems, Adv. Neural. Inf. Process. Syst. 35, 29400 (2022).
- [13] O. T. Unke, S. Chmiela, M. Gastegger, K. T. Schütt, H. E. Sauceda, and K.-R. Müller, SpookyNet: Learning force fields with electronic degrees of freedom and nonlocal effects, Nat. Commun. **12**, 7273 (2021).
- [14] T. W. Ko, J. A. Finkler, S. Goedecker, and J. Behler, A fourth-generation high-dimensional neural network potential with accurate electrostatics including non-local charge transfer, Nat. Commun. 12, 398 (2021).
- [15] J. Gasteiger, F. Becker, and S. Günnemann, Gemnet: Universal directional graph neural networks for

molecules, Adv. Neural. Inf. Process. Syst. **34**, 6790 (2021).

- [16] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons, Phys. Rev. Lett. **104**, 136403 (2010).
- [17] V. L. Deringer, A. P. Bartók, N. Bernstein, D. M. Wilkins, M. Ceriotti, and G. Csányi, Gaussian Process Regression for Materials and Molecules, Chem. Rev. 121, 10073 (2021).
- [18] A. S. Christensen, L. A. Bratholm, F. A. Faber, and O. Anatole von Lilienfeld, FCHL revisited: Faster and more accurate quantum machine learning, J. Chem. Phys. 152, 044107 (2020).
- [19] J. Behler and M. Parrinello, Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces, Phys. Rev. Lett. 98, 146401 (2007).
- [20] S. Sorella and L. Capriotti, Algorithmic differentiation and the calculation of forces by quantum monte carlo, J. Chem. Phys 133, 234111 (2010), https://doi.org/10.1063/1.3516208.
- [21] C. Filippi, R. Assaraf, and S. Moroni, Simple formalism for efficient derivatives and multi-determinant expansions in quantum monte carlo, J. Chem. Phys. 144, 194105 (2016), https://doi.org/10.1063/1.4948778.
- [22] R. Assaraf, S. Moroni, and C. Filippi, Optimizing the energy with quantum monte carlo: A lower numerical scaling for jastrow-slater expansions, J. Chem. Theory Comput. 13, 5273 (2017), https://doi.org/10.1021/acs.jctc.7b00648.
- [23] P. J. Reynolds, R. N. Barnett, B. L. Hammond, R. Grimes, and W. A. Lester, Quantum chemistry by quantum monte carlo: Beyond ground-state energy calculations, Int. J. Quantum Chem. 29, 589 (1985).
- [24] A. Badinski, P. D. Haynes, J. R. Trail, and R. J. Needs, Methods for calculating forces within quantum Monte Carlo simulations, J. Phys.: Condens. Matter 22, 074202 (2010).
- [25] R. Assaraf and M. Caffarel, Zero-variance zero-bias principle for observables in quantum Monte Carlo: Application to forces, J. Chem. Theory Comput. **119**, 10536 (2003).
- [26] S. Chiesa, D. M. Ceperley, and S. Zhang, Accurate, Efficient, and Simple Forces Computed with Quantum Monte Carlo Methods, Phys. Rev. Lett. 94, 036404 (2005).
- [27] C. Filippi and C. J. Umrigar, Correlated sampling in quantum Monte Carlo: A route to forces, Phys. Rev. B 61, R16291 (2000).
- [28] S. Moroni, S. Saccani, and C. Filippi, Practical Schemes for Accurate Forces in Quantum Monte Carlo, J. Chem. Theory Comput. 10, 4823 (2014).
- [29] J. Van Rhijn, C. Filippi, S. De Palo, and S. Moroni, Energy Derivatives in Real-Space Diffusion Monte Carlo, J. Chem. Theory Comput. 18, 118 (2022).
- [30] K. Nakano, M. Casula, and G. Tenti, Unbiased and affordable atomic forces in ab initio Variational Monte Carlo (2023), arXiv:2312.17608.
- [31] A. Tirelli, G. Tenti, K. Nakano, and S. Sorella, Highpressure hydrogen by machine learning and quantum Monte Carlo, Phys. Rev. B 106, L041105 (2022).
- [32] H. Niu, Y. Yang, S. Jensen, M. Holzmann, C. Pierleoni, and D. M. Ceperley, Stable Solid Molecular Hydrogen above 900 K from a Machine-Learned Potential Trained

with Diffusion Quantum Monte Carlo, Phys. Rev. Lett. **130**, 076102 (2023).

- [33] G. Tenti, A. Tirelli, K. Nakano, M. Casula, and S. Sorella, Principal deuterium Hugoniot via Quantum Monte Carlo and Δ-learning (2023), arXiv:2301.03570.
- [34] C. Huang and B. M. Rubenstein, Machine Learning Diffusion Monte Carlo Forces, J. Phys. Chem. A 127, 339 (2023).
- [35] B. Huang, O. A. Von Lilienfeld, J. T. Krogel, and A. Benali, Toward DMC Accuracy Across Chemical Space with Scalable Δ-QML, J. Chem. Theory Comput. **19**, 1711 (2023).
- [36] D. M. Ceperley, S. Jensen, Y. Yang, H. Niu, C. Pierleoni, and M. Holzmann, Training models using forces computed by stochastic electronic structure methods, Electron. Struct. 6, 015011 (2024).
- [37] A. Tkatchenko and M. Scheffler, Accurate Molecular Van Der Waals Interactions from Ground-State Electron Density and Free-Atom Reference Data, Phys. Rev. Lett. 102, 073005 (2009).
- [38] A. Tkatchenko, R. A. DiStasio, R. Car, and M. Scheffler, Accurate and Efficient Method for Many-Body van der Waals Interactions, Phys. Rev. Lett. 108, 236402 (2012).
- [39] W. M. C. Foulkes, L. Mitas, R. J. Needs, and G. Rajagopal, Quantum monte carlo simulations of solids, Rev. Mod. Phys. **73**, 33 (2001).
- [40] A. Lüchow, Quantum monte carlo methods, Wiley Interdiscip. Rev.: Comput. Mol. Sci. 1, 388 (2011).
- [41] B. M. Austin, D. Y. Zubarev, and W. A. Lester, Quantum monte carlo and related approaches, Chem. Rev. 112, 263 (2012).
- [42] C. J. Umrigar, M. P. Nightingale, and K. J. Runge, A diffusion Monte Carlo algorithm with very small timestep errors, J. Chem. Phys. 99, 2865 (1993).
- [43] C. Attaccalite and S. Sorella, Stable Liquid Hydrogen at High Pressure by a Novel *Ab Initio* Molecular-Dynamics Calculation, Phys. Rev. Lett. **100**, 114501 (2008).
- [44] S. Pathak and L. K. Wagner, A light weight regularization for wave function parameter gradients in quantum Monte Carlo, AIP Adv. 10, 085213 (2020).
- [45] As Jastrow factor, we use the exponential of the sum of two fifth-order polynomials of the electron-nuclear and the electron-electron distances, respectively, and rescale the inter-particle distances as $R = (1 - \exp(-\kappa r))/\kappa$ with κ set to 0.6 a.u. We employ different electron-nucleus Jastrow factors to describe the correlation of an electron with C, O, and H. The total number of free parameters to be optimized in the Jastrow factor is 17 for the systems considered here.
- [46] B. Huron, J. P. Malrieu, and P. Rancurel, Iterative perturbation calculations of ground and excited state energies from multiconfigurational zeroth-order wavefunctions, J. Chem. Phys. 58, 5745 (1973).
- [47] P. S. Epstein, The stark effect from the point of view of schrödinger's quantum theory, Phys. Rev. 28, 695 (1926).
- [48] R. Nesbet, Configuration interaction in orbital theories, Proc. R. Soc. London, Ser. A 230, 312 (1955).
- [49] CHAMP is a quantum Monte Carlo program package written by C. Filippi, S. Moroni and C. J. Umrigar with significant contributions by R. Shinde, N. Renaud, V. Azizi, E. Landinez, S. Shepard and E. Slootman, https://github.com/filippi-claudia/champ (accessed on 15-04-2024).

- [50] M. Burkatzki, C. Filippi, and M. Dolg, Energy-consistent pseudopotentials for quantum Monte Carlo calculations, J. Chem. Phys. **126**, 234105 (2007).
- [51] For the hydrogen atom, we use a more accurate BFD pseudopotential and basis set. Dolg, M.; Filippi, C., private communication.
- [52] S. Sorella, M. Casula, and D. Rocca, Weak binding between two aromatic rings: Feeling the van der Waals attraction by quantum Monte Carlo methods, J. Chem. Phys **127**, 014105 (2007).
- [53] E. Neuscamman, C. J. Umrigar, and G. K.-L. Chan, Optimizing large parameter sets in variational quantum Monte Carlo, Phys. Rev. B 85, 045103 (2012).
- [54] M. Casula, Beyond the locality approximation in the standard diffusion Monte Carlo method, Phys. Rev. B 74, 161102 (2006).
- [55] M. W. Schmidt, K. K. Baldridge, J. A. Boatz, S. T. Elbert, M. S. Gordon, J. H. Jensen, S. Koseki, N. Matsunaga, K. A. Nguyen, S. Su, T. L. Windus, M. Dupuis, and J. A. Montgomery Jr, General atomic and molecular electronic structure system, J. Comput. Chem. 14, 1347 (1993).
- [56] Y. Garniron, T. Applencourt, K. Gasperich, A. Benali, A. Ferté, J. Paquier, B. Pradines, R. Assaraf, P. Reinhardt, J. Toulouse, P. Barbaresco, N. Renon, G. David, J.-P. Malrieu, M. Véril, M. Caffarel, P.-F. Loos, E. Giner, and A. Scemama, Quantum Package 2.0: An Open-Source Determinant-Driven Suite of Programs, J. Chem. Theory Comput. 15, 3591 (2019).
- [57] E. Posenitskiy, V. G. Chilkuri, A. Ammar, M. Hapka, K. Pernal, R. Shinde, E. J. Landinez Borda, C. Filippi, K. Nakano, O. Kohulák, S. Sorella, P. de Oliveira Castro, W. Jalby, P. L. Ríos, A. Alavi, and A. Scemama, TREXIO: A file format and library for quantum chem-

istry, J. Chem. Phys. 158, 174801 (2023).

- [58] D. G. A. Smith, L. A. Burns, A. C. Simmonett, R. M. Parrish, M. C. Schieber, R. Galvelis, P. Kraus, H. Kruse, R. Di Remigio, A. Alenaizan, A. M. James, S. Lehtola, J. P. Misiewicz, M. Scheurer, R. A. Shaw, J. B. Schriber, Y. Xie, Z. L. Glick, D. A. Sirianni, J. S. O'Brien, J. M. Waldrop, A. Kumar, E. G. Hohenstein, B. P. Pritchard, B. R. Brooks, H. F. Schaefer, III, A. Y. Sokolov, K. Patkowski, A. E. DePrince, III, U. Bozkaya, R. A. King, F. A. Evangelista, J. M. Turney, T. D. Crawford, and C. D. Sherrill, PSI4 1.4: Open-source software for high-throughput quantum chemistry, J. Chem. Phys. 152, 184108 (2020).
- [59] T. H. Dunning, Gaussian basis sets for use in correlated molecular calculations. i. the atoms boron through neon and hydrogen, J. Chem. Phys. **90**, 1007 (1989).
- [60] G. Fonseca, I. Poltavsky, V. Vassilev-Galindo, and A. Tkatchenko, Improving molecular force fields across configurational space by combining supervised and unsupervised machine learning, J. Chem. Phys. 154, 124102 (2021).
- [61] G. Fonseca, I. Poltavsky, and A. Tkatchenko, Force Field Analysis Software and Tools (FFAST): Assessing Machine Learning Force Fields under the Microscope, J. Chem. Theory Comput. **19**, 8706 (2023).
- [62] V. Kapil, M. Rossi, O. Marsalek, R. Petraglia, Y. Litman, T. Spura, B. Cheng, A. Cuzzocrea, R. H. Meißner, D. M. Wilkins, B. A. Helfrecht, P. Juda, S. P. Bienvenue, W. Fang, J. Kessler, I. Poltavsky, S. Vandenbrande, J. Wieme, C. Corminboeuf, T. D. Kühne, D. E. Manolopoulos, T. E. Markland, J. O. Richardson, A. Tkatchenko, G. A. Tribello, V. Van Speybroeck, and M. Ceriotti, I-PI 2.0: A universal force engine for advanced molecular simulations, Comput. Phys. Commun. 236, 214 (2019).