The Journal of Chemical Physics

Ab initio molecular dynamics with noisy forces: Validating the quantum Monte Carlo approach with benchmark calculations of molecular vibrational properties

Cite as: J. Chem. Phys. **141**, 194112 (2014); https://doi.org/10.1063/1.4901430
Submitted: 27 July 2014 . Accepted: 30 October 2014 . Published Online: 19 November 2014

Ye Luo, Andrea Zen, and Sandro Sorella

AIP Publishing

Ab initio molecular dynamics with noisy forces: Validating the quantum Monte Carlo approach with benchmark calculations of molecular vibrational properties

Ye Luo,[1,a)] Andrea Zen,[2,b)] and Sandro Sorella[1,c)]
[1]*International School for Advanced Studies (SISSA), and CRS Democritos, CNR-INFM, Via Bonomea 265, I-34136 Trieste, Italy*
[2]*Dipartimento di Fisica, Università di Roma "La Sapienza," Piazzale Aldo Moro 2, I-00185 Rome, Italy*

(Received 27 July 2014; accepted 30 October 2014; published online 19 November 2014)

We present a systematic study of a recently developed *ab initio* simulation scheme based on molecular dynamics and quantum Monte Carlo. In this approach, a damped Langevin molecular dynamics is employed by using a statistical evaluation of the forces acting on each atom by means of quantum Monte Carlo. This allows the use of an highly correlated wave function parametrized by several variational parameters and describing quite accurately the Born-Oppenheimer energy surface, as long as these parameters are determined at the minimum energy condition. However, in a statistical method both the minimization method and the evaluation of the atomic forces are affected by the statistical noise. In this work, we study systematically the accuracy and reliability of this scheme by targeting the vibrational frequencies of simple molecules such as the water monomer, hydrogen sulfide, sulfur dioxide, ammonia, and phosphine. We show that all sources of systematic errors can be controlled and reliable frequencies can be obtained with a reasonable computational effort. This work provides convincing evidence that this molecular dynamics scheme can be safely applied also to realistic systems containing several atoms. © *2014 AIP Publishing LLC*. [http://dx.doi.org/10.1063/1.4901430]

I. INTRODUCTION

In the last decade, much progress has been done for the simulation of electronic systems by quantum Monte Carlo (QMC), namely, by a fully *ab initio* approach aimed to solve in a stochastic way the Schrödinger equation, with an appropriate and consistent description of the electron correlation. Only a few years ago, a very general and robust method of optimization[1] was introduced, that has opened the possibility to determine by QMC a variational wave function containing up to 10 000 parameters.[2,3] This progress is particularly remarkable, as the variational Monte Carlo (VMC) method was introduced in the early 1960s[4] and, until a few years ago, only a few tens of parameters were optimized within the VMC approach. Another recent and important development in QMC was the solution[5–7] of the infinite variance problem occurring in the straightforward calculation of nuclear forces in the simplest variational Monte Carlo scheme.[5,8] Moreover, thanks to the algorithmic differentiation,[7] the cost of computing all the force components in a system containing several atoms, can be afforded with a computational time at most a factor four larger than the one corresponding to the energy. This progress has led to several works, where structural optimization and highly accurate evaluations of the equilibrium configurations as well as related properties were possible even for quite large systems containing several atoms.[9–13]

Despite this remarkable progress, we notice that *ab initio* molecular dynamics (MD) simulation based on quantum Monte Carlo remains so far at a very early stage, as only a few simulations on liquid hydrogen[8,14–16] are known. Instead, within the density functional theory (DFT) community, MD simulations in the Born-Oppenheimer (BO) approximation, are quite well established, due to almost three decades of achievements from the pioneering work of Roberto Car and Michele Parrinello.[17] Indeed, DFT-based MD simulations are routinely used to study several properties of condensed matter systems at ambient conditions up to extremely high pressures and temperatures,[18–26] and represents nowadays a quite reliable tool to predict new materials, sometimes more effective or at least much cheaper than experiments.

The application of quantum Monte Carlo for *ab initio* simulation of bulk materials or large chemical compounds remains difficult not only because of the heavy computational cost, but also, in our opinion, due to the theoretical difficulties in applying the Newton's equations of motion when the forces are given with a statistical uncertainty. For instance, the basic law of energy conservation cannot be met at all, when the forces are not exactly given at each step. In this context, it is worth mentioning that Ceperley and Dewing have introduced the penalty method[27] that does not rely on any dynamics, and therefore is not affected by this problem. In their method the canonical distribution is directly sampled without using forces, while the statistical uncertainty in the knowledge of the energy is compensated by rejecting the proposed moves more frequently than in the standard Metropolis algorithm. Unfortunately this method is very expensive, especially in the low temperature regime, because of too many rejected moves, and so far applications have been limited to hydrogen with up to 54 protons in this regime.[28–30]

[a)]Electronic mail: xw111luoye@gmail.com
[b)]Electronic mail: zen.andrea.x@gmail.com
[c)]Electronic mail: sorella@sissa.it

Generally speaking it is clear that, when the computational cost for the calculation of the nuclear forces is comparable to that of the energy, MD should be more efficient, because with the same cost all the atoms are moved in a statistically relevant region of the phase space, without any rejection. For instance in DFT, where the forces are obtained almost for free by applying the Hellmann-Feynman theorem, MD is a common practice to sample the canonical distribution, and, to our knowledge, only hybrid methods based on Monte Carlo and MD[31] can be competitive.

As we have already emphasized, at present QMC allows us to compute forces in an efficient way, and we believe it is now important to study systematically how reliable is the standard MD by means of the QMC evaluation of forces. In particular, we want to understand:

- how the noise in the forces affects the reliability of the dynamics,
- how the systematic error due to the discretization in time affects the calculation in presence of noisy forces,
- how well the Born-Oppenheimer constraint is satisfied, namely, how accurately it is possible to evolve the electronic wave function following the minimum energy condition. Indeed in a statistical method, the variational parameters cannot be optimized with machine precision accuracy and the departure of the wave function from its minimum energy may represent an important bias requiring a careful study.

In order to answer to the above issues, we show in this paper the performances of QMC-based approaches (including MD simulations) by benchmark calculations of structural and vibrational properties of small molecules. These properties are often of interest in Chemistry and Materials Science because they help the interpretation of experiments, for instance, of infrared and Raman spectroscopy.[32]

In *ab initio* approaches, the vibrational properties are usually obtained within the Born-Oppenheimer approximation, that separates the electronic and nuclear degrees of freedom. Thus, their evaluation relies on the properties of the potential energy surface (PES) in the neighborhood of the structural minimum of the molecule.[32,33] The simplest approach is to assume that the PES in the neighborhood of the minimum is well characterized in harmonic approximation, so the frequencies are obtained from the diagonalization of the mass-weighted Hessian matrix,[32] which is calculated by performing static *ab initio* computations in the minimum of the PES, or in its neighborhood. This approach neglects the anharmonicity of the PES, so *ad hoc* scaling factors[34,35] have to be introduced in order to compare with the experimental frequencies. The most accurate approaches[36–47] go beyond the harmonic approximation, for instance, taking the force fields of the PES around the configurational minimum up to the fourth order expansion and using the second order vibrational perturbation theory (PT2).[44–47] Other *ab initio* approaches are based on *ab initio* molecular dynamics simulations, which directly includes finite temperature nuclear motions, and the Infrared and Raman spectra can be directly obtained from the Fourier transform of dipole and polarizability autocorrelation functions,[33,49] as obtained by the Linear Response Theory[48] and the Fermi Golden Rule.

In this work, we have evaluated the structural and vibrational properties by using and comparing three different methods: (i) a simple fitting method with the Hessian, and in a few cases with higher order derivatives of the PES that are estimated by a careful fit of independent measurements of energy and forces. These quantities are calculated over a set of molecular configurations arranged on a grid[9,47] around the equilibrium structure of the molecule; (ii) a fitting method with finite temperature molecular dynamics which is similar as (i) by using in the fitting samples of energies and forces at various molecular configurations. However, these configurations are generated automatically by a QMC-based MD simulation, at a given temperature T, with noisy forces;[8,15] (iii) a covariance matrix method by using time averaged correlations in a QMC-based MD simulation.

It is clear that, if the QMC-based MD simulation is consistent and the BO constraint is satisfied correctly, all different methods should provide consistent results, provided all sources of systematic errors can be removed in a controlled way, in order to converge to unbiased evaluations of the geometrical and vibrational properties.

In this work, we show that the method (ii) provides very accurate results with an efficiency comparable with the standard method (i), whereas the method (iii) is computationally very demanding and is used therefore here only for testing the MD, as emphasized above. The method (ii), that we are proposing, is in our opinion better than the standard one (i) because it can be easily and systematically extended to complex systems containing several atoms. In such cases, it is very difficult to work with the standard method, because it relies on a careful choice of the grid of atomic positions that are used to fit the PES.[47] This method is difficult to be generalized to very complicated systems, and in particular the grid cannot be generated by a black box tool, as it depends instead on the user's choice. Instead we propose here the much more general and flexible method (ii), allowing a systematic and robust evaluation of harmonic frequencies. In this technique, only a single parameter has to be tuned, namely, the target temperature of the MD simulation.

The paper is organized as follows: in Sec. II, we introduce the molecular dynamics scheme with noisy forces evaluated by QMC; in Sec. III, the three approaches of evaluating vibrational properties are explained in detail; in Sec. IV, we describe the wave functions and the basis sets we use for all the molecules; whereas the discussion of all sources of systematic errors related to the present QMC dynamics is given in Sec. V; Sec. VI contains our results on several molecules with some discussion; finally in Sec. VII we draw our conclusions.

II. MD WITH NOISY FORCES

Our *ab initio* MD simulations are performed via variational quantum Monte Carlo (VMC) by employing `TurboRVB` QMC package.[50] A second order Langevin dynamics[51] (SLD) is used in the sampling of the ionic configurations within a ground state Born-Oppenheimer approach.

Ionic forces are computed with finite and small variance by algorithmic differentiation,[7] which allows feasible simulations of a large number of atoms. Moreover the statistical noise, corresponding to the forces, is used to drive the dynamics at finite temperature by means of an appropriate generalized Langevin dynamics.[8] A similar approach has been proposed in Refs. 52 and 53, where a SLD algorithm has been devised also at the DFT level. In this work, we adopt a different numerical integration scheme for the SLD which allows us to use large time steps, even in presence of large friction matrices. For reasons of clarity and completeness, we present in this section the method introduced in the original paper of Attaccalite and Sorella,[8] with more details in the derivations, whereas the more advanced techniques, that can be straightforwardly derived following the same analysis, are described in Appendix B.

Let us consider solving the set of differential equations of the SLD,

$$\dot{\boldsymbol{v}} = -\boldsymbol{\gamma}(\boldsymbol{R}) \cdot \boldsymbol{v} + \boldsymbol{f}(\boldsymbol{R}) + \boldsymbol{\eta}(t), \tag{1}$$

$$\dot{\boldsymbol{R}} = \boldsymbol{v}, \tag{2}$$

$$\begin{aligned} \langle \boldsymbol{\eta}(t) \rangle &= 0, \\ \left\langle \eta_i(t) \eta_j(t') \right\rangle &= \alpha_{ij}(\boldsymbol{R})\, \delta(t - t'), \end{aligned} \tag{3}$$

where $\boldsymbol{R}$, $\boldsymbol{v}$, $\boldsymbol{f}$, $\boldsymbol{\eta}$ are the $3N$-dimensional vectors made by the positions, the velocities, the deterministic, and the stochastic forces of the N nuclei, respectively, and the indices i, j run over all the $3N$ nuclear coordinates. The symbol $\langle \cdots \rangle$ indicates the average over the ensemble of possible realizations, and it is used to define properties of the stochastic force $\boldsymbol{\eta}$, which are determined by the fluctuation-dissipation theorem, namely, its instantaneous correlation $\boldsymbol{\alpha}$ is given by

$$\boldsymbol{\alpha}(\boldsymbol{R}) = 2T\, \boldsymbol{\gamma}(\boldsymbol{R}), \tag{4}$$

where $T \equiv 1/\beta$ is the temperature[63] and both $\boldsymbol{\gamma}(\boldsymbol{R})$ and $\boldsymbol{\alpha}(\boldsymbol{R})$ are $3N$-dimensional square matrices, implicitly depending on the atomic positions.

Notice that in the above equations we have assumed that all the masses of the particles are set to unit values in atomic Rydberg units, namely, twice the electronic mass $2m_e$ is one in our conventions. In the following we will always use unit masses, because, in order to sample the canonical distribution the actual values of the masses are immaterial. In order to match the usual atomic units, for instance in the hydrogen case already studied in Refs. 8 and 15, the time units have to be scaled by the square root of the ratio between the proton mass and twice the electron mass ($\sqrt{m_p/2m_e} \sim 30.3$). In a polyatomic molecule — like water — the inverse mass of each different atom multiples the force components $\boldsymbol{f}(\boldsymbol{R})$ in the commonly adopted Langevin equations. However, also in this case, it is possible to reduce back to the case studied, by a further appropriate scaling of the length of each particle (distinguishable in classical dynamics). Thus our formulation is quite general up to an appropriate scaling of time and lengths,[64] and therefore can be also used to study the physical Newtonian dynamics — e.g., necessary to compute the diffusion constant in liquid water — with $\boldsymbol{\gamma} \to 0$ and physical masses.

In Eq. (4), one of the two matrices is arbitrary and we can choose

$$\boldsymbol{\alpha}(\boldsymbol{R}) = \alpha_0 \mathbf{I} + \Delta_0 \boldsymbol{\alpha}^{\mathrm{QMC}}(\boldsymbol{R}), \tag{5}$$

$$\boldsymbol{\gamma}(\boldsymbol{R}) = \frac{\boldsymbol{\alpha}(\boldsymbol{R})}{2T,} \tag{6}$$

where $\mathbf{I}$ is the identity matrix, α_0 and Δ_0 are two constants that should be suitably defined in order to minimize the autocorrelation time and therefore the efficiency of the sampling, and the $3N$-dimensional matrix $\boldsymbol{\alpha}^{\mathrm{QMC}}(\boldsymbol{R})$ is the variance-covariance matrix of the nuclear forces $\boldsymbol{f}(\boldsymbol{R})$ evaluated by QMC at the nuclear configuration $\boldsymbol{R}$, and it is defined as

$$\alpha_{ij}^{\mathrm{QMC}}(\boldsymbol{R}) = \langle (f_i(\boldsymbol{R}) - \langle f_i(\boldsymbol{R}) \rangle)(f_j(\boldsymbol{R}) - \langle f_j(\boldsymbol{R}) \rangle) \rangle, \tag{7}$$

where $\langle \cdots \rangle$ refers to the average over the QMC sampling. In practice, $\boldsymbol{\alpha}^{\mathrm{QMC}}$ is computed as $\boldsymbol{C}_{\delta}^{s}$ in Eq. (40) (see Sec. III C for more details).

We now assume only that in the time interval

$$t_n - \tau/2 < t < t_n + \tau/2,$$

n indexing the time steps $t_n = n \times \tau$, the positions $\boldsymbol{R}$ are changing very little and, within a good approximation, we can neglect the $\boldsymbol{R}$ dependence in the RHS of Eq. (1), and indicate $R(t_n) = R_n$. The second equation (2) can be integrated easily once the value of a velocity is known at a given time

$$\boldsymbol{R}_{n+1} - \boldsymbol{R}_n \simeq \tau\, \boldsymbol{v}(t), \tag{8}$$

where $t_n \le t \le t_{n+1}$. A better way to integrate the equation is given in Appendix B. For the time being, we assume the above simple form, and for a better accuracy it is useful to consider that the velocities $\boldsymbol{v}_n$ are computed at half-integer times $t_n - \tau/2$,

$$\boldsymbol{v}_n \equiv \boldsymbol{v}(t_n - \tau/2) \tag{9}$$

and the quantities that are functions of $\boldsymbol{R}$ in Eq. (1) are calculated in $\boldsymbol{R}_n$,

$$\boldsymbol{f}_n \equiv \boldsymbol{f}(\boldsymbol{R}_n), \tag{10}$$

$$\boldsymbol{\gamma}_n \equiv \boldsymbol{\gamma}(\boldsymbol{R}_n)\,. \tag{11}$$

Once in this small time integration interval the values of $f(\boldsymbol{R}) = \boldsymbol{f}_n$ and $\gamma(\boldsymbol{R}) = \boldsymbol{\gamma}_n$ are assumed constant, the solution to Eq. (1) is given in a closed form in Eq. (A7) of Appendix A, with the initial time $\bar{t}$ and the final one t arbitrary

$$\begin{aligned} \boldsymbol{v}(t) = {} & \exp[\boldsymbol{\gamma}_n(\bar{t} - t)]\boldsymbol{v}_n \\ & + \int_{\bar{t}}^{t} \exp[\gamma_n(t' - t)][\boldsymbol{f}_n + \boldsymbol{\eta}(t')]dt'\,. \end{aligned} \tag{12}$$

In this way, after substituting the initial and final time with $t_n \mp \tau/2$ a Markov chain of the following form is obtained:

$$\boldsymbol{v}_{n+1} = e^{-\boldsymbol{\gamma}_n \tau} \boldsymbol{v}_n + \boldsymbol{\Gamma}_n \cdot (\boldsymbol{f}_n + \tilde{\boldsymbol{\eta}}), \tag{13}$$

$$\boldsymbol{R}_{n+1} = \boldsymbol{R}_n + \tau v_{n+1}, \tag{14}$$

namely, we have singled out the "noisy" corrections to the force components in Eq. (13) ($\boldsymbol{f}_n + \tilde{\boldsymbol{\eta}}$) by defining the following quantities:

$$\boldsymbol{\Gamma}_n = \boldsymbol{\gamma}_n^{-1}(\mathbf{I} - e^{-\boldsymbol{\gamma}_n \tau}), \tag{15}$$

$$\tilde{\boldsymbol{\eta}} = \frac{\boldsymbol{\gamma}_n}{2\sinh(\boldsymbol{\gamma}_n \tau/2)} \int_{t_n-\tau/2}^{t_n+\tau/2} e^{\boldsymbol{\gamma}_n (t-t_n)} \boldsymbol{\eta}(t)\, dt\,. \tag{16}$$

By using that $[\boldsymbol{\alpha}, \boldsymbol{\gamma}] = 0$ and a little algebra, the correlator defining the discrete (time integrated) noise can be computed by substitution of Eqs. (16) in (3) and is given by the following $3N \times 3N$ matrix:

$$\langle \tilde{\eta}_i \tilde{\eta}_j \rangle = \bar{\boldsymbol{\alpha}} = T \boldsymbol{\gamma}_n^2 \coth(\boldsymbol{\gamma}_n \tau/2)\,. \tag{17}$$

The simulation temperature T appearing in the above expression is an input parameter of the dynamics. If the discretization of the SLD is accurate enough this temperature should be related to the mean square velocities measured during the dynamics ($\langle v_i^2 \rangle = T/2$ for each Cartesian component). In the following, and in particular in Sec. V A, we refer to this quantity as the "effective temperature" T_{mes}, as $T_{\text{mes}} - T$ can be used to judge the quality of the approximations in discretizing the SLD.

As discussed also in Ref. 8 (see also Sec. III C), all the QMC force evaluations $\boldsymbol{f}$ are affected by an intrinsic stochastic noise, that usually determines an effective temperature higher than the target one. This problem can be avoided, by means of the *noise correction* introduced in Ref. 8. Indeed, we can follow the correct dynamics by adding to the QMC noise of the force the external noise $\tilde{\boldsymbol{\eta}}_{\text{ext}}$ so that the total noise $\tilde{\boldsymbol{\eta}}$ satisfies the correct expressions in Eq. (17). In this way, we have to subtract the $3N \times 3N$ QMC correlation of the forces $\boldsymbol{\alpha}^{\text{QMC}}$ from the above described correlation matrix $\bar{\boldsymbol{\alpha}}$ and obtain that

$$\bar{\boldsymbol{\alpha}}_{\text{ext}} = \bar{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^{\text{QMC}} \tag{18}$$

is the true external noise, we have to add to the force components during the dynamics. Indeed the correlation matrix $\boldsymbol{\alpha}^{\text{QMC}}$ can be independently evaluated during the dynamics and the computation of Eq. (7) is possible with some statistical error. In this way, it is possible to take into account that QMC forces are affected by a correlated noise, and obtain an, in principle, unbiased simulation following the correct SLD.

It can be shown, by a simple numerical calculation, that the resulting matrix $\bar{\boldsymbol{\alpha}}_{\text{ext}}$ is indeed positive definite provided $\Delta_0 > \tau$, so that $\bar{\boldsymbol{\alpha}}_{\text{ext}}$ is a well defined correlation for an external noise. In the present work, we have discovered after several tests, that the value of Δ_0, optimizing the efficiency of the calculation, is not necessarily the minimum one, i.e., $\Delta_0 = \tau$. Indeed much larger time steps and better performances are possible if $\Delta_0 \gg \tau$. In order to understand this behavior, it is important to realize that the covariance matrix $\boldsymbol{\alpha}^{\text{QMC}}$ obtained with QMC is empirically proportional to the dynamical matrix (see Fig. 1). Therefore with a finite and large Δ_0, the high energy modes with high frequency vibrations can be systematically damped, and this clearly allows a faster propagation with larger time step τ.

We have already shown in Ref. 15 that the present integration scheme of the Langevin equations is much better than the Euler integration method. In this work, we also show that the present dynamics is also much more convenient within QMC because we can use a friction matrix proportional to the mentioned QMC covariance matrix ($\Delta_0 > 0$). To this purpose we have implemented exactly the same dynamics within the `Quantum ESPRESSO`[54] package, and showed in Fig. 2 that, within the DFT dynamics, only quite smaller time steps τ are

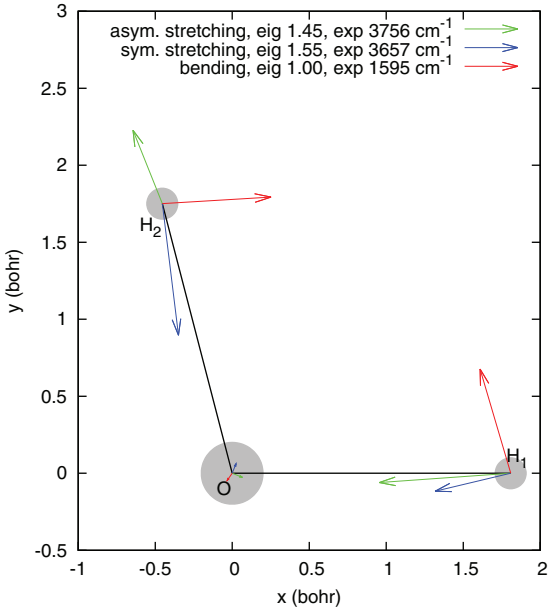

FIG. 1. Eigenvectors of the 3×3 correlation matrix $\boldsymbol{\alpha}^{\text{QMC}}$ which is constructed by $f^x_{\text{H}_1}$, $f^x_{\text{H}_2}$, and $f^y_{\text{H}_2}$ (x_{H_1}, x_{H_2}, and y_{H_2} are chosen as internal coordinates). These eigenvectors correspond to the three vibrational modes of the water monomer: bending (red), symmetrical (blue), and asymmetrical (green) stretching. The smaller eigenvalue of $\boldsymbol{\alpha}^{\text{QMC}}$ corresponds to the lowest frequency vibrational mode. The eigenvalues in the plot are all rescaled by the lowest eigenvalue.

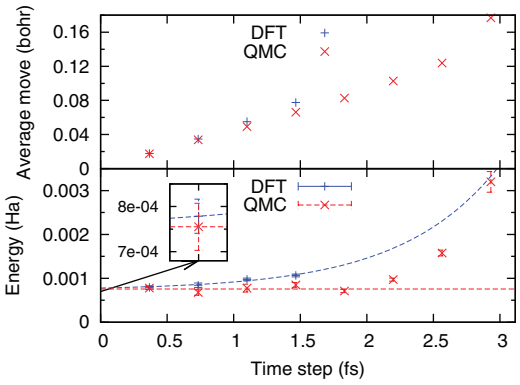

FIG. 2. Convergence of the internal potential energy and average ion displacements as a function of the time step τ for the water dimer, obtained with MD at 50 K. The offset energy (minimum of the PES) values of DFT and QMC calculations are -34.410806 and $-34.50405(4)$ Ha. The same friction $\alpha_0 = 0.03$ a.u. is used for both. Our simulation with DFT becomes unstable when $\tau \geq 1.8$ fs. Instead the QMC dynamics is always stable in the range studied because the friction matrix in this case contains also an important non-diagonal contribution proportional to QMC covariance matrix (see text and Fig. 1). In the top panel, we compare the average distance that ions experience at each step in QMC and DFT dynamics. The increased stability achieved by using this covariance matrix in the friction is therefore obtained with an almost negligible slowing down of the QMC dynamics.

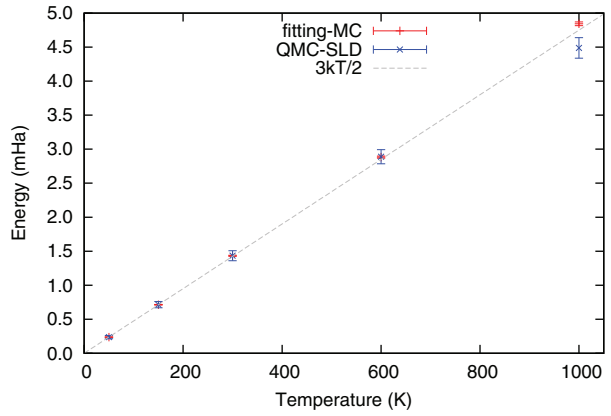

FIG. 3. The internal energy of the water monomer in SLD simulations with QMC and MC simulations with the fitted potential at various temperatures. Since this potential expanded up to the fourth order is fitted with forces, the zeroth order coefficient is set to zero and the QMC internal energy, for comparison, is shifted by $-17.24909(3)$ Ha which is the lowest possible variational energy with $\mathcal{N}_{\rm QMC} = 20480$ and $n_{\rm opt} = 20$. In these calculations, for simplicity, only coordinates $x_{\rm H_1}$, $x_{\rm H_2}$, and $y_{\rm H_2}$ are free. Therefore, the internal energy should be $\frac{3}{2}kT$ if Harmonic approximation is assumed.

possible, just because in this case $\boldsymbol{\alpha}^{\rm QMC} = 0$, and it is not possible to damp the too high frequency vibrations.

In order to confirm that our SLD scheme samples the exact partition function $Z = \int dx \exp(-V(x)/kT)$, we have also done a conventional Monte Carlo simulation with the fitted potential for the water monomer. In Fig. 3, both simulations show a perfect agreement in the internal energy at various temperatures consistent with the harmonic approximation up to 1000 K.

III. CALCULATION OF VIBRATIONAL PROPERTIES

The standard method of calculating vibrational modes is well known in the literature.[55] In this section, we summarize the main formulas and introduce the notations adopted in the rest of the paper.

Within the Born-Oppenheimer approximation, the full Hamiltonian is separated into electronic and nuclear parts and only the latter is related to the calculation of vibrational modes. The nuclear Hamiltonian $\hat{H}$ is a summation of kinetic energy $\hat{T}$ and potential energy $\hat{V}$. Given a molecule with N atoms, $\hat{H}$ can be expressed in terms of $3N$-dimensional Cartesian coordinates $\boldsymbol{R}$,

$$\hat{H} = \hat{T} + \hat{V} = -\frac{1}{2}\sum_{\xi}^{N}\frac{1}{M_\xi}\nabla_\xi^2 + V(\boldsymbol{R}), \tag{19}$$

where M_ξ is the mass of the atom ξ. Since the potential energy $V(\boldsymbol{R})$ is generally assumed to be invariant under the translation and rotation of molecules, it is independent of both the molecule's center of mass and orientation. Therefore, it can be written in terms of $3N-6$ (or $3N-5$ for linear molecules) internal coordinates $\bar{\boldsymbol{s}}$.

Since the molecule is usually assumed to be semi-rigid, its potential energy can be simply recast in terms of the displacement from the equilibrium structure $\boldsymbol{R}_0$, corresponding to the (local) minimum of the PES. The Cartesian displacement is $\boldsymbol{X} = \boldsymbol{R} - \boldsymbol{R}_0$. Similarly, the displacement in the internal coordinates is $\boldsymbol{s} = \bar{\boldsymbol{s}} - \bar{\boldsymbol{s}}_0$.

In general, the mapping between Cartesian coordinates and internal coordinates is curvilinear. With the Taylor expansion, $\boldsymbol{s}$ becomes

$$s_i = B_i^a X_a + \frac{1}{2!}B_i^{ab}X_aX_b + \frac{1}{3!}B_i^{abc}X_aX_bX_c + \cdots, \tag{20}$$

where $i = 1, \ldots, 3N-6$ labels the internal coordinates and $a, b, c = 1, \ldots, 3N$ label the Cartesian coordinates. The Einstein summation notation of repeated indices is assumed hereafter. The coefficients in the series are the derivatives with respect to the Cartesian displacement: $B_i^a = \partial s_i/\partial X_a$, $B_i^{ab} = \partial^2 s_i/(\partial X_a \partial X_b)$ and so on and so forth. The coefficients B_i^a in the linear term define the so-called Wilson $\boldsymbol{B}$ matrix.

Hence, the potential energy can be expanded around the equilibrium structure in terms of internal coordinate displacements as

$$V(\boldsymbol{s}) = F^0 + F^i s_i + \frac{1}{2!}F^{ij}s_is_j + \frac{1}{3!}F^{ijk}s_is_js_k + \cdots, \tag{21}$$

where the coefficients in the expansion are defined as $F^0 \equiv V$, $F^i \equiv \partial V/\partial s_i$, $F^{ij} \equiv \partial^2 V/(\partial s_i \partial s_j)$, etc., calculated at $s = 0$. Clearly, all the coefficients $F^i = 0$. Since F^0 is an irrelevant offset for all the vibrational modes, we ignore it by putting $V(\boldsymbol{s}) - F^0$ instead of $V(\boldsymbol{s})$. In the following paper, V2 and V4 are used to indicate the potential energy surface expanded up to the second and the fourth orders.

In the standard method of calculating vibrational modes within the harmonic approximation, only the leading terms, i.e., the quadratic ones, are kept in both the potential and kinetic energies, while all the rest are neglected

$$V_{\rm har}(\boldsymbol{s}) = \frac{1}{2}F^{ij}s_is_j = \frac{1}{2}\boldsymbol{s}^\dagger\boldsymbol{F}\boldsymbol{s}, \tag{22}$$

$$T_{\rm har}(\dot{\boldsymbol{s}}) = \frac{1}{2}(\boldsymbol{G}^{-1})^{ij}\dot{s}_i\dot{s}_j = \frac{1}{2}\dot{\boldsymbol{s}}^\dagger\boldsymbol{G}^{-1}\dot{\boldsymbol{s}}, \tag{23}$$

where $\dot{\boldsymbol{s}}$ is the time derivative of $\boldsymbol{s}$, and the symbol $\dagger$ indicates the transpose. Meanwhile, the $(3N-6)\times(3N-6)$ matrix $\boldsymbol{G}$ is calculated as

$$G_{ij} = \sum_{\xi}^{N}\sum_{\alpha}^{3}\frac{1}{M_\xi}B_i^{\xi,\alpha}B_j^{\xi,\alpha}, \tag{24}$$

where $B_i^{\xi,\alpha}$ are the same linear terms defined in Eq. (20), upon replacement of the index a with the pair (ξ, α), indicating more explicitly the component α corresponding to the atom ξ.

By introducing $3N-6$ normal coordinates $\boldsymbol{q}$, the potential and kinetic energies are recast as

$$V_{\rm har}(\boldsymbol{q}) = \frac{1}{2}\sum_{i}^{3N-6}\lambda_i q_i^2 \tag{25}$$

$$T_{\rm har}(\dot{\boldsymbol{q}}) = \frac{1}{2}\sum_{i}^{3N-6}\dot{q}_i^2\,, \tag{26}$$

where $\lambda_i = \omega_i^2$ are the harmonic force constants corresponding to harmonic frequencies ω_i.[65] Assuming the transformation between internal coordinates and normal coordinates $\boldsymbol{s} = \boldsymbol{L}\boldsymbol{q}$, we replace $\boldsymbol{s}$ in Eqs. (22) and (23) and compare them with Eqs. (25) and (26). The final relations are written in matrix form as

$$\boldsymbol{L}^\dagger \boldsymbol{F}\boldsymbol{L} = \boldsymbol{\Phi}, \quad \boldsymbol{L}^\dagger \boldsymbol{G}^{-1}\boldsymbol{L} = \mathbf{I}\,, \tag{27}$$

where $\boldsymbol{\Phi}$ is a diagonal matrix with λ_i on the diagonal and $\mathbf{I}$ is a $3N-6$ dimensional identity matrix. With some very simple algebra, Eqs. (27) turns into $\boldsymbol{G}\boldsymbol{F}\boldsymbol{L} = \boldsymbol{L}\boldsymbol{\Phi}$ which represents a standard generalized eigenvalue problem, where λ_r are the corresponding eigenvalues. This approach is also called Wilson's GF method.[55]

Ab initio methods can be used to calculate the Hessian matrix $\boldsymbol{F}$ in the potential energy, so that the application of the GF method is possible. In the standard method, it is necessary to perform a very accurate structural optimization of the molecule, and then to calculate the derivatives for the optimized geometry using analytic or finite-difference methods. A very tight structural optimization is computationally very demanding for QMC, thus alternative methods specifically engineered for stochastic-error affected approaches are preferable, as discussed in Ref. 47 and summarized in Sec. III A. We propose here other two possible approaches, described in Secs. III B and III C.

We have reported also some results, labeled as fundamental frequencies, coming from second order perturbation theory (PT2), that uses also the third and fourth order derivatives of $V(\boldsymbol{s})$, in order to take into account of the anharmonicity of the PES. The use and implementation of PT2 in presence of error affected PES have been widely discussed in Ref. 47, and we remand to this reference.

A. Simple fitting method

The conventional way to obtain the Hessian of V is to fit the parametrized Hessian matrix $\boldsymbol{F}$ with energies or forces computed at the chosen grid points of the $3N$ multidimensional space defined by the nuclear positions. In each of the $3N$ directions, at least 3 points are needed in the neighborhood of the equilibrium position in order to fit the Hessian. Obviously, this requires a tight (gradient $<10^{-5}$ a.u., for the harmonic approach) or very tight (gradient $<10^{-7}$ a.u., for PT2) structure optimization criteria[47] which can be easily achieved by self-consistent iterations in DFT or other deterministic methods.

However, these criteria are not feasible for QMC since all the energies and forces calculated by QMC are error-affected. The stochastic error $\sigma_{\rm QMC}$ is inversely proportional to the square root of the number of QMC samples $\mathcal{N}_{\rm QMC}$. Thus, in order to have an error 10 times smaller, 100 times more expensive calculation is required. For this reason, the QMC stochastic errors are never pushed to very small values, especially for vibrational property calculations. Typically, the errors are $\sigma_E \sim 10^{-4}$ a.u. for energy and $\sigma_F \sim 10^{-3}$ a.u. for each force component. In brief, both the PES and equilibrium structure are very much affected by the stochastic noise.

Zen *et al.*[47] proposed a multidimensional fitting scheme of the PES of a molecule in proximity of its equilibrium configuration, by using a function $V_{\boldsymbol{k}}(\boldsymbol{R})$ that is parametrized by parameters $\boldsymbol{k}$, in order to obtain the accurate Hessian and equilibrium structure. In particular, a data set $\boldsymbol{D}_F$ containing $\mathcal{N}_m$ samples $\{\boldsymbol{R}_m, \boldsymbol{f}_m, \boldsymbol{C}_m\}_m$, $m = 1, \ldots, \mathcal{N}_m$ – where the QMC force $\boldsymbol{f}_m$ and its $3N$-dimensional covariance matrix $\boldsymbol{C}_m$ (see definition Eq. (7) and discussion in Sec. III C) are calculated at the atomic configuration $\boldsymbol{R}_m$ – is used to determine the parameters $\boldsymbol{k}$ that provide the best fit of the PES via the function $V_{\boldsymbol{k}}(\boldsymbol{R})$. They showed that the fitting using QMC forces brings smaller stochastic error than the fitting with energies. So we stick to forces for the fitting. Moreover, for our calculations with the water molecule we choose the "mesh-5" (see definition in Ref. 47), which consists of 59 independent grid point calculations.

The fitting with forces of the PES, against the data set $\boldsymbol{D}_F$, is achieved by maximizing the likelihood function

$$\mathcal{L}(\boldsymbol{k}|\boldsymbol{D}_F) = \prod_m^{\mathcal{N}_m} \frac{e^{-\frac{1}{2}\sum_{a,b}^{3N}(\boldsymbol{C}_m{}^{-1})_{ab}\Delta\mathcal{F}_m^a(\boldsymbol{k})\Delta\mathcal{F}_m^b(\boldsymbol{k})}}{(2\pi)^{3N/2}\sqrt{\det(\boldsymbol{C}_m)}}, \tag{28}$$

where $\Delta\mathcal{F}_m^a(\boldsymbol{k})$ is defined as

$$\Delta\mathcal{F}_m^a(\boldsymbol{k}) = \mathcal{F}^a(\boldsymbol{R}_m, \boldsymbol{k}) - f_m^a\,,$$

namely, the difference between the QMC force f_m^a of component a and the corresponding value of the parametrized force $\mathcal{F}^a(\boldsymbol{R}_m, \boldsymbol{k})$, which is given by

$$\mathcal{F}^a(\boldsymbol{R}_m, \boldsymbol{k}) = -\left.\frac{\partial V_{\boldsymbol{k}}(\boldsymbol{R})}{\partial R^a}\right|_{\boldsymbol{R}_m}. \tag{29}$$

The problem of maximizing $\mathcal{L}(\boldsymbol{k}|D_F)$ is equivalent to minimize the function

$$\sum_m^{\mathcal{N}_m}\sum_{a,b}^{3N}(\boldsymbol{C}_m{}^{-1})_{ab}\Delta\mathcal{F}_m^a(\tilde{\boldsymbol{k}})\Delta\mathcal{F}_m^b(\tilde{\boldsymbol{k}}) \tag{30}$$

and, as discussed in Ref. 47, in the case that we can neglect the covariance between QMC force evaluations of the different components (i.e., we can assume $\boldsymbol{C}_m$ diagonal, and the diagonal elements $(\boldsymbol{C}_m)_{aa} = (\sigma_m^a)^2$ are the variance of the QMC force evaluations of component a), the previous expression corresponds to the chi-squared-function

$$\chi_F^2 = \sum_m^{\mathcal{N}_m}\sum_a^{3N}\left(\frac{\mathcal{F}^a(\boldsymbol{R}_m, \boldsymbol{k}) - f_m^a}{\sigma_m^a}\right)^2. \tag{31}$$

We can quantify the quality of the fit by using the reduced-chi-squared function (goodness of fit),

$$\chi_{\rm red}^2 = \frac{\chi_F^2}{(3N \times \mathcal{N}_m - N_k - 1)}\,, \tag{32}$$

where $3N$ is the number of force components, $\mathcal{N}_m$ is the number of molecular configurations considered, and N_k is the number of fitted parameters. According to statistical theory, the closer $\chi_{\rm red}^2 \simeq 1$, the better the fit is.

B. Fitting method with finite temperature molecular dynamics

In the simple fitting method described in Sec. III A, the choice of the grid points, where energy and forces are evaluated, is crucial for accessing accurate vibrational frequencies. A good mesh should span a region neither too small, in order to be less affected by the stochastic noise, nor too large, to avoid strong anharmonicity (which cannot be well described by simple parametrization of the PES in a truncated Taylor expansion around the minimum). In order to reduce the systematic error, the best mesh should be expanded along the directions of the normal coordinates which are however known only after the fitting. An efficient compromise is to use internal coordinates based on certain conventional rules. After the region and expansion direction of the mesh is given, the density of the grid points should be also chosen properly. Too sparse mesh limits the accuracy while too dense mesh wastes computation. Since we did only once the evaluation of energy and forces on each grid point in the simple fitting method, good optimization of the wavefunction and accurate calculation of energy and forces are both necessary and therefore expensive. Usually, for each step during the optimization, relatively small statistics is used. In this way, the accuracy of the energy and forces evaluated in the last iteration of the optimization does not meet the necessary precision for the fitting. So a further much longer run at fixed optimal values of the variational parameters is required to compute the energy and forces precisely. This has also the drawback that the error in the optimization has to be negligible compared with the requested statistical error, a condition that is difficult to control systematically.

With finite temperature molecular dynamics, things are instead much easier. By controlling only the temperature, a proper mesh is automatically decided by the trajectory of the moving ions. Since the dynamics follows the normal modes, the mesh has already been expanded around the best directions. The density and range of the mesh is directly tuned by the temperature. Since ions move very little for each iteration of MD, a heavy optimization is no longer necessary because the electronic wave function obtained in the previous iteration of the dynamics is a very good starting point for the current iteration, once only the positions of the atomic localized orbitals are consistently updated. The possibility to expand our electronic variational wave function in terms of localized atomic orbitals is actually a very remarkable advantage. Indeed, after a few optimization steps, the wave function is usually converged within given statistical errors, and, as we will see later, the error in the optimization can be systematically controlled. Apart from the cheap optimization, energy and forces also require much less accuracy. In Table I, the error bars of the energy and forces during the dynamics are about 40 and 60 times larger than those of the simple fitting. The values obtained in the last step of the optimization are already sufficient and thus, a substantial amount of computation is saved. The fitting procedure of the sampled configurations coming from finite temperature molecular dynamics is exactly the same described in Sec. III A, and we can use Eqs. (28)–(31).

TABLE I. Specifications of the fitting with manually chosen grid and MD.

Name	χ^2_{red}	Grid points	$\sigma(E)$ (Ha)	$\sigma(F)$ (a.u.)	Cpuh BG/Q
Simple fitting V4	4.444	59	9.0×10^{-5}	1.5×10^{-4}	22.2k
MD fit V2 50 K	1.007	13784	3.5×10^{-3}	6.2×10^{-3}	24.0k
MD fit V4 1000 K	1.021	13784	3.7×10^{-3}	6.3×10^{-3}	23.5k

In our tests on the water monomer, the anharmonic effects are quite strong if the MD is performed at high temperature. We have systematically studied the effect of the temperature in Fig. 4, where it is clear that anharmonic effects can be neglected only below 50 K, namely, when the temperature corresponds to a frequency 50 times smaller than the lowest frequency of the system ($\simeq$ 2300 K). This criterion cannot be easily extended to larger systems as the smallest frequency significantly drops, and a calculation at too small temperatures cannot provide enough information for the fit, yielding large statistical errors for the frequencies. For this reason it is important to include in the fitting also the cubic and quartic terms, and, as it is also shown in Fig. 7, it is really remarkable that we can obtain a very reliable and converged estimate of the frequencies even at 1000 K.

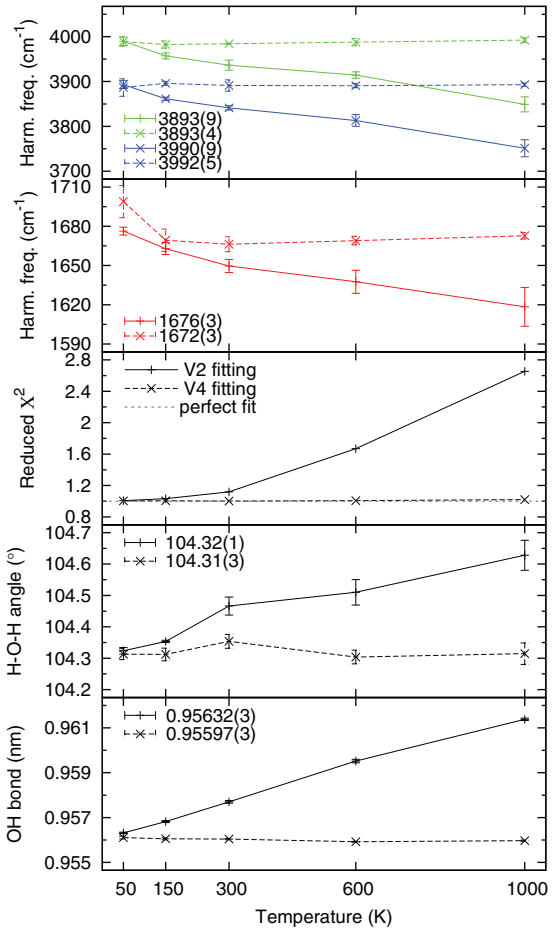

FIG. 4. Harmonic vibrational frequencies and equilibrium geometry of water monomer obtained by fitting V2 (solid line) and V4 (dashed line) as a function of temperature. The values reported in the keys are obtained at 50 K for V2 and 1000 K for V4.

In order to improve further the fit, we generate more statistical samples by taking advantage of the molecular symmetry. For example, H_2O, H_2S, and SO_2 have the C_{2v} symmetry, while NH_3 and PH_3 have the C_{3v} symmetry. By simply swapping the positions and forces of each pair of H or O atoms in the molecules, we obtain once or twice more samples, without an extra computational effort. This procedure does not change much the frequencies and equilibrium geometry, as well as their statistical errors, but allows to enforce the symmetry of a molecule, namely, recovering all equal X–H bond lengths and degenerate frequencies, if related by the mentioned symmetries. We have used a similar method also in the simple fitting (Subsection III A), but in that case the purpose was mainly to reduce the number of points in the grid and to save computational resources.

C. Covariance matrix method

The previous two methods of evaluating vibrational frequencies give very accurate results but require an explicit parametrization of the PES in the neighborhood of the minimum, as well as the reduction of the number of parameters by using the symmetries of the molecule. To avoid this human overhead, we introduce another way of computing vibrational frequencies based on the evaluation of few appropriate covariance matrices, described in the following.

By employing a simple Gaussian integral over the statistical weight $\exp(-V_{\text{har}}(\boldsymbol{s})/kT)$ where $V_{\text{har}}(\boldsymbol{s})$ is defined in Eq. (22), we easily obtain the relation

$$\boldsymbol{C}_s = kT\,\boldsymbol{F}^{-1}, \tag{33}$$

where $\boldsymbol{C}_s$ is the covariance matrix of the internal coordinates $\boldsymbol{s}$ according to the definition

$$\boldsymbol{C}_s = \langle(\boldsymbol{s} - \langle\boldsymbol{s}\rangle)(\boldsymbol{s} - \langle\boldsymbol{s}\rangle)^\dagger\rangle, \tag{34}$$

where $\langle\cdots\rangle$ refers to the ensemble average, while in practice it is computed as the time average along the trajectory of MD. Therefore the matrix $\boldsymbol{F}$, necessary in the Wilson's GF method, can be obtained by computing $\boldsymbol{C}_s$ with a simple algebraic inversion and a scaling by kT.

In a more direct method, the information of the matrix $\boldsymbol{F}$ can be obtained by computing the covariance matrix of the forces. Indeed, the forces $\boldsymbol{f}(\boldsymbol{s})$ defined as

$$f_i(\boldsymbol{s}) = -\frac{\partial V(\boldsymbol{s})}{\partial s_i} \tag{35}$$

have a very simple form if we can reliably work in harmonic approximation, namely, $V(\boldsymbol{s}) \simeq V_{\text{har}}(\boldsymbol{s})$, and are

$$f_i(\boldsymbol{s}) \simeq -\boldsymbol{F}\boldsymbol{s}\,. \tag{36}$$

Therefore, the covariance matrix of the forces is more simply related to the matrix $\boldsymbol{F}$ as

$$\boldsymbol{C}_f = kT\,\boldsymbol{F}, \tag{37}$$

where $\boldsymbol{C}_f$ is the covariance matrix of $\boldsymbol{f}$ similar to Eq. (34).

However, in QMC the forces are noisy and correlated since they are evaluated with the same Markov chains of finite length $\mathcal{N}_{\text{QMC}}$, namely,

$$\boldsymbol{f}_{\text{noisy}} \equiv \langle \boldsymbol{f}_{\text{local}}\rangle = \boldsymbol{f}_{\text{exact}} + \boldsymbol{\delta}, \tag{38}$$

where $\boldsymbol{f}_{\text{local}}$ is the local force evaluated in each QMC sample and $\boldsymbol{\delta}$ is the statistical error associated with the QMC evaluation of the force. In order to obtain accurate frequencies, it is necessary to remove this bias for calculating the covariance matrix of forces, which we have done in the following way. The thermal average which is used to compute the covariance matrix can be divided into two steps – the average of all electronic realizations generated by quantum Monte Carlo at fixed ionic configuration $\boldsymbol{s}$ and the average of all the ionic configurations obtained during the Langevin dynamics. In the first step, it is necessary to accumulate the covariance of the exact force components which however are known only with some statistical error ($\boldsymbol{f}_{\text{noisy}}$). Therefore we can write $\boldsymbol{f}_{\text{exact}} = \boldsymbol{f}_{\text{noisy}} - \boldsymbol{\delta}$, where the error $\boldsymbol{\delta}$ depends on the quantum Monte Carlo statistics and vanishes only for $\mathcal{N}_{\text{QMC}} \to \infty$. It follows therefore that

$$\langle \boldsymbol{f}_{\text{exact}} \boldsymbol{f}^\dagger_{\text{exact}}\rangle^{\boldsymbol{s}} \approx \langle \boldsymbol{f}_{\text{noisy}} \boldsymbol{f}^\dagger{}_{\text{noisy}}\rangle^{\boldsymbol{s}} - \boldsymbol{C}^{\boldsymbol{s}}_{\boldsymbol{\delta}}, \tag{39}$$

where the superscript $\boldsymbol{s}$ refers to the restriction of a given ionic configuration $\boldsymbol{s}$, whereas $\boldsymbol{C}^{\boldsymbol{s}}_{\boldsymbol{\delta}}$ is the covariance of the noise, that can be in turn estimated by standard statistical methods using the finite number $\mathcal{N}_{\text{QMC}}$ of Monte Carlo electronic samples used for the given ionic configuration $\boldsymbol{s}$, namely,

$$\boldsymbol{C}^{\boldsymbol{s}}_{\boldsymbol{\delta}} \approx \frac{1}{\mathcal{N}_{\text{QMC}}(\mathcal{N}_{\text{QMC}} - 1)} \times \sum_{j=1}^{\mathcal{N}_{\text{QMC}}} \left(\boldsymbol{f}^{j,\boldsymbol{s}}_{\text{local}} - \boldsymbol{f}^{\boldsymbol{s}}_{\text{noisy}}\right)\left(\boldsymbol{f}^{j,\boldsymbol{s}\dagger}_{\text{local}} - \boldsymbol{f}^{\boldsymbol{s}\dagger}_{\text{noisy}}\right), \tag{40}$$

where $\boldsymbol{f}^{j,\boldsymbol{s}}_{\text{local}}$ are the force components corresponding to an independent electronic QMC sample j. This matrix can be more conveniently evaluated with the Jackknife technique and the reweighting method[8,9] to ensure a finite variance calculation of the forces. With a finite number $\mathcal{N}_{\text{QMC}}$ of independent samples, $\boldsymbol{C}^{\boldsymbol{s}}_{\boldsymbol{\delta}}$ scales as $1/\mathcal{N}_{\text{QMC}}$ and hence, the frequencies have a corresponding correction proportional to $\boldsymbol{C}^{\boldsymbol{s}}_{\boldsymbol{\delta}}$ and therefore the bias scales as $\frac{1}{\mathcal{N}_{\text{QMC}}}$ if the proposed noise correction is not applied. We will show the clear advantage to use this noise correction scheme in Sec. V C.

IV. VARIATIONAL WAVE FUNCTIONS

In all the following calculations, we have used a variational wave function of a standard Jastrow-Slater form expanded on a localized basis set. According to our previous work[9] for the water monomer, we use $(4s,5p,1d)$ primitive basis with 4 hybrid orbitals on oxygen (in the shorthand notation, O:$(4s,5p,1d)$/{4}) and H:$(3s,1p)$/{1} in the determinant part, whereas the Jastrow is expanded as a two-body part $\frac{1}{2b}(1 - e^{-br})$ with one body rescaled and a three-body part with O:$(3s,2p,1d)$/{2} and H:$(2s,2p)$/{2} on hydrogen. The exponents of the primitive basis in both the determinant and Jastrow parts are optimized at the equilibrium configuration and kept fixed during the dynamics. This choice is also made for all the other molecules. Its VMC ground state energy at the equilibrium geometry is $-17.24927(3)$. During the dynamics, the exponents of the primitive basis in both determinant

TABLE II. Total energies and geometries of the structural minimum of H_2S, SO_2, NH_3, and PH_3 molecules. Potential V4 is fitted for H_2S and SO_2, while only V2 are fitted for NH_3 and PH_3.

		Equilibrium geom.		GS energy
Name	Basis set	Bond (nm)	Angle (°)	Hatree
H_2S		S–H	H-S-H	
JHF_nooptZ	S:(7*s*,8*p*,1*d*)/{4} H:(3*p*,1*d*)/{1}	1.33216(6)	92.00(3)	– 11.40043(2)
JHF	S:(5*s*,4*p*,1*d*)/{4} H:(3*p*,1*d*)/{1}	1.33180(5)	92.42(2)	– 11.40902(2)
JAGP	S:(5*s*,4*p*,1*d*)/{5} H:(3*p*,1*d*)/{3}	1.33237(5)	92.36(2)	– 11.41154(2)
CCSD(T)	aug-cc-pVTZ	1.3419	92.299	...
Exp.	...	1.328	92.2	...
SO_2		S–O	O-S-O	
JHF	S:(6*s*,6*p*,1*d*)/{6} O:(6*s*,7*p*,1*d*)/{5}	1.4180(2)	119.91(6)	– 42.27474(8)
JAGP	S:(6*s*,6*p*,1*d*)/{6} O:(6*s*,7*p*,1*d*)/{5}	1.4193(2)	120.06(6)	– 42.28255(7)
CCSD(T)	aug-cc-pVTZ	1.4553	118.367	...
Exp.	...	1.432	119.5	...
NH_3		N–H	H-N-H	
JHF	N:(6*s*,6*p*,1*d*)/{4} H:(3*p*,1*d*)/{1}	1.00886(3)	107.01(1)	– 11.74967(2)
JAGP	N:(4*s*,4*p*,1*d*)/{4} H:(3*p*,1*d*)/{1}	1.01014(3)	106.53(2)	– 11.7512(3)
CCSD(T)	aug-cc-pVQZ	1.0128	106.541	...
Exp.	...	1.012	106.67	...
PH_3		P–H	H-P-H	
JHF	P:(6s,7p,1d)/{4} H:(3p,1d)/{1}	1.40925(7)	93.72(1)	– 8.34788(1)
JAGP	P:(6s,7p,1d)/{4} H:(3p,1d)/{1}	1.41067(5)	93.580(9)	– 8.34899(1)
CCSD(T)	cc-pVTZ	1.4186	93.501	...
Exp.	...	1.421	93.3	...

and three body Jastrow are all fixed and only 232 parameters in total are optimized on the fly. This basis is much more compact compared with the ones used in our previous work.[9] We have indeed verified that a larger basis does not improve much the inter-atomic description but decreases only the total energy. On the other hand, too many parameters make the optimization part too heavy and inefficient during the dynamics. Hence, we have to choose a compromise between accuracy and efficiency due to the available computational resources. Despite this limitation, we are generally working close to the Complete Basis Set limit as long as relevant chemical properties are concerned, thanks also to the rapid convergence in the basis set obtained within explicitly correlated wavefunctions, satisfying for instance all the electron-electron and electron-ion cusp conditions even with a finite basis set.

For H_2S, SO_2, NH_3, and PH_3 molecules, the basis sets used for the determinant part of the wave function are listed in Table II. The Jastrow has the same two body part as H_2O and its three-body part consists of (3*s*,2*p*,1*d*) on N/O/P/S and (2*s*,2*p*) on H.

In our calculation, energy-consistent pseudopotentials (ECP) of Burkatzki *et al.*[56] are used to replace the core electrons of N, O, P, and S atoms in order to have a reduced computational cost. Helium core is used for both N and O and Neon core is used for both P and S.

V. CONTROL OF ALL SOURCES OF SYSTEMATIC ERRORS IN QMC DYNAMICS

In the standard electronic structure calculation based on molecular dynamics, there are essentially three systematic errors to take into consideration: the time step τ used to integrate the SLD equations of motions, the accuracy in satisfying the Born-Oppenheimer approximation and the total time of simulation t_{TOT}. In the following, we consider simple small molecules such that the simulation time is much larger than any reasonable correlation time of the system, so that this error can be safely neglected for simplicity. In QMC, we have to take into account also that, at each step of the discretized dynamics in Eqs. (13), only a statistical evaluation of the forces f_i with a finite number of samples $\mathcal{N}_{\mathrm{QMC}}$ is possible. This yields a statistical error $\propto \frac{1}{\sqrt{\mathcal{N}_{\mathrm{QMC}}}}$ that can be decreased very slowly with the computational time $\propto \mathcal{N}_{\mathrm{QMC}}$.

A. Time step error

As far as the time step τ error, this is simple to control, because unbiased solutions of the SLD equations of motion can be obtained by reducing τ to a sufficiently small value, within any reasonable integration scheme. In QMC, we can perform the limit of $\tau \to 0$ for instance at fixed $\mathcal{N}_{\mathrm{QMC}}$. As long as there is no other source of bias (see Subsection V B and V C) other than a finite $\mathcal{N}_{\mathrm{QMC}}$, we expect to have unbiased results for $\tau \to 0$ even within QMC, as explained in the following. After the time integration of the SLD equations in a small time interval τ, the statistical noise associated with the forces is multiplied by the integration time τ in Eq. (13), that is negligible compared to the stochastic part $\propto \sqrt{2T\tau}$ used to keep the temperature within the given target. In this way, the systematic QMC error is expected to vanish *linearly* in τ and for $\tau \to 0$ the exact canonical distribution can be sampled,

$$\exp(-V(\boldsymbol{R})/kT), \tag{41}$$

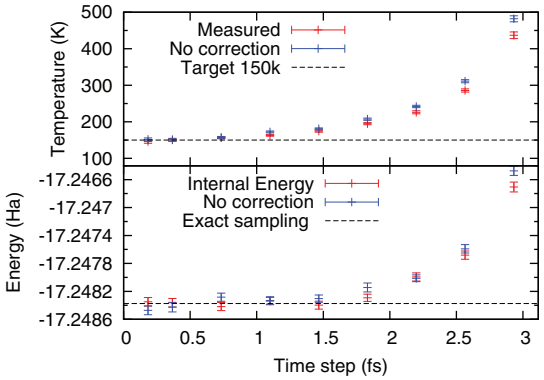

FIG. 5. The measured temperature and internal potential energy versus the time step used in the dynamics of water monomer. The target temperature is set at 150 K, the friction is 0.3 a.u. and $\Delta_0 = 8.0$. The minimum of the PES is $-17.24909(3)$ Ha. At each step of MD we perform $n_{\rm opt} \simeq 10$ steps of optimization, where all energy derivatives are estimated with $\mathcal{N}_{\rm QMC} = 20480$ samples generated by the Metropolis algorithm with 80 proposed attempts for each new sample (acceptance rate $\simeq 50\%$). The same plot, without using the noise correction (see Eq. (18)) is also shown. The dashed lines indicate the "exact" results for the average temperature and the internal energy, the latter obtained by sampling exactly the fitted potential.

where $V(\boldsymbol{R})$ is the BO-energy surface corresponding to a variational wavefunction $\psi_{\alpha,R}$ defined by several variational parameters α for given atomic positions

$$V(\boldsymbol{R}) = \mathrm{Min}_{\alpha} \frac{\langle \Psi_{\alpha,\boldsymbol{R}} | H_{\boldsymbol{R}} | \Psi_{\alpha,\boldsymbol{R}} \rangle}{\langle \Psi_{\alpha,\boldsymbol{R}} | \Psi_{\alpha,\boldsymbol{R}} \rangle} . \tag{42}$$

This error can be made in principle smaller by the "noise correction" scheme that was introduced in a previous work.[8] In practice, as it is shown in Fig. 5, the convergence in τ looks very well behaved and a reasonable accuracy is obtained also by using quite large time steps. In this case, the mentioned noise correction scheme does not lead to a meaningful improvement probably because the use of a finite large $\Delta_0 = 8$ makes our dynamics more stable and less sensitive to the stochastic noise.

B. Error in sampling the BO energy surface $n_{\rm opt} \to \infty$

In the previous estimate of the error in τ we have to assume that, given the atomic positions, the energy derivatives of the BO energy surface $V(\boldsymbol{R})$ can be computed statistically, but without systematic bias. This means that the variational parameters α are exactly at the minimum energy condition that defines $V(\boldsymbol{R})$ in Eq. (42), and only in this case the forces are unbiased. Unfortunately, this condition is never met in a statistical optimization of the variational parameters and some approximation has to be done in practice. In the following we introduce the control parameter $n_{\rm opt}$. Each run of MD is obtained by performing several thousand iterations of the SLD discretized with a time interval τ. For each step of MD, we perform $n_{\rm opt}$ optimization steps of the electronic wave function with the linear method introduced in Refs. 1 and 3. For $n_{\rm opt} \to \infty$ and fixed $\mathcal{N}_{\rm QMC}$, the optimized wavefunction converges to an approximate minimum of the BO energy surface where the energy derivatives, namely, the atomic forces, differ at most by $\frac{1}{\sqrt{\mathcal{N}_{\rm QMC}}}$ from the exact BO ones. Therefore, we have found that it is convenient to study

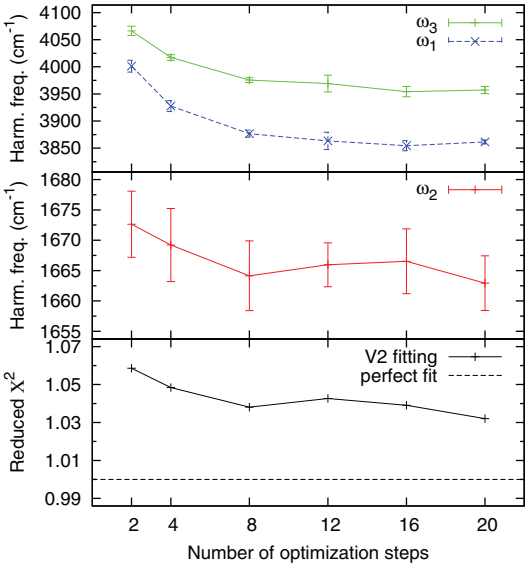

FIG. 6. Harmonic vibrational frequencies of water monomer obtained by fitting V2 with samples generated by MD simulation at 150 K as a function of QMC optimization steps $n_{\rm opt}$.

long, well equilibrated MD simulations at fixed statistical accuracy (i.e., $\mathcal{N}_{\rm QMC}$ fixed) and given τ, by increasing $n_{\rm opt}$ in a systematic way. In the optimization method, we have used a given tolerance $\epsilon = 0.001$ in the inversion of the ill conditioned overlap matrix S corresponding to the chosen set of atomic orbitals used in the Jastrow and the determinantal part of our wave function. As now well established, the knowledge of this matrix S is extremely useful for an efficient optimization scheme (see Refs. 3 and 57). Moreover, for the sake of a stable and systematic optimization technique we have also attenuated the wave function change predicted by the linear method[1] by 50%. As it is shown in Figs. 6 and 7, the finite $n_{\rm opt}$ error is probably the most important one in QMC, because several optimization steps are necessary to achieve converged

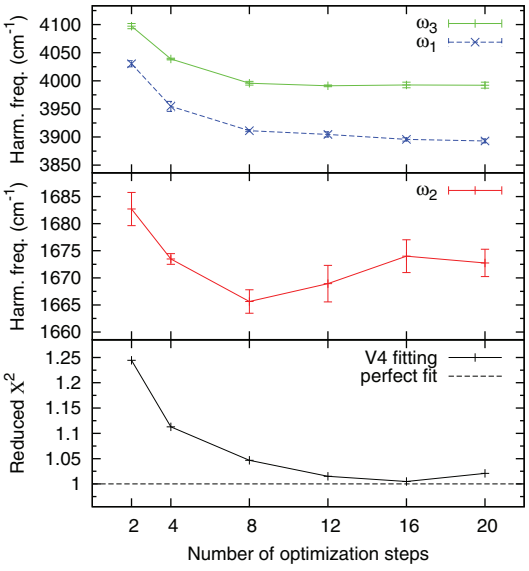

FIG. 7. Harmonic vibrational frequencies of water monomer obtained by fitting V4 with samples generated by MD simulation at 1000 K as a function of QMC optimization steps $n_{\rm opt}$.

frequencies, especially the high frequency ones. Notice that, in this plot we use the fitting method, and the systematic error in τ, as defined in Subsection III A, is not present. Despite the slow convergence in $n_{\rm opt}$, it is quite evident that, by using $n_{\rm opt} \geq 10$ the simulations are still affordable and the error bars of the frequencies are quite small with reasonable computational resources, even for large $n_{\rm opt}$. In these plots, the error bars have been evaluated by the standard Jackknife technique.

C. Residual QMC error $\mathcal{N}_{\rm QMC} \to \infty$

Once all the above sources of error have been controlled, we are still left with the Monte Carlo statistical noise, namely, the fact that we have to work with a finite number of samples $\mathcal{N}_{\rm QMC}$ for each iteration of the dynamics. Among the various techniques considered in this work, this error affects mostly the method described in Sec. III C.

As we have mentioned in Sec. III B, this systematic error is affecting the evaluation of the energy derivatives by an error of order $\frac{1}{\sqrt{\mathcal{N}_{\rm QMC}}}$. This means that the variational parameters α have a typical error of this magnitude $\frac{1}{\sqrt{\mathcal{N}_{\rm QMC}}}$ during the MD simulation. However, since the energy at the minimum is affected quadratically by the error in the variational parameters, we can expect that all the frequency estimates, based only on energy expectation values, show a much smaller error inversely proportional to the number of sampling $\mathcal{N}_{\rm QMC}$. This is readily seen in Fig. 8 where the calculation of frequencies is seen to converge linearly in $\frac{1}{\mathcal{N}_{\rm QMC}}$. In this calculation it is also simple to identify the most important source of error, that is due to the stochastic estimation of the covariance matrix of the forces. Once we correct this source of bias described in Sec. III C, we see that this error is almost negligible (see Fig. 9).

Indeed, the other two fitting methods are also affected by the statistical correlation of the force components due to

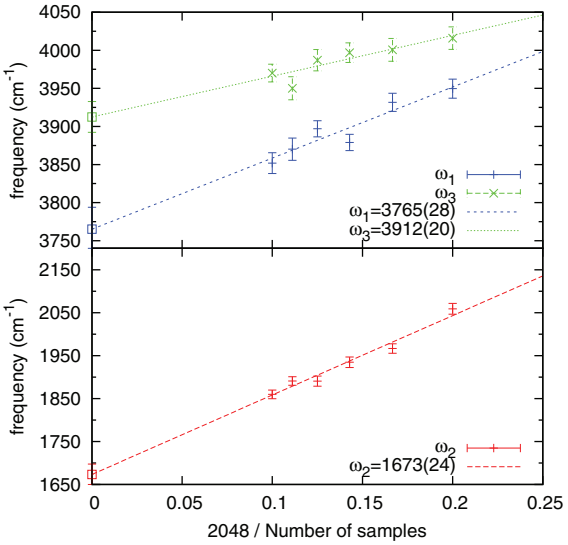

FIG. 8. Vibrational frequencies of the water monomer obtained with the force-force covariance matrix without noise correction as a function of $1/\mathcal{N}_{\rm QMC}$ from MD simulation at 300 K. The prefactor 2048 on the slope of the linear extrapolation does not change the intercept for the limit $1/\mathcal{N}_{\rm QMC} \to 0$.

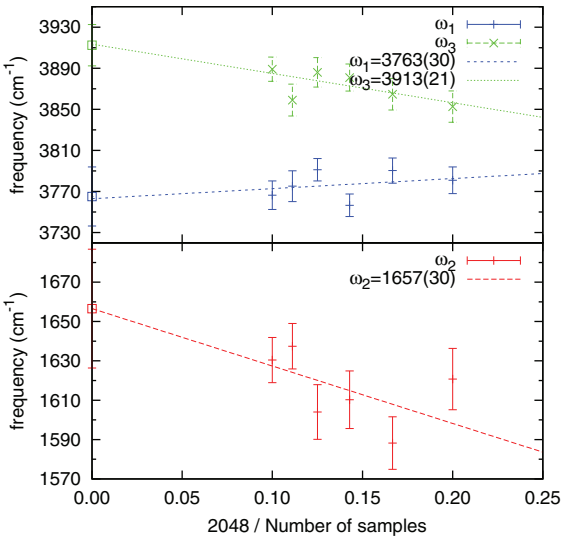

FIG. 9. Vibrational frequencies of water monomer obtained with force-force covariance matrix plus noise correction as a function of $1/\mathcal{N}_{\rm QMC}$ from MD simulation at 300 K.

the finite QMC samples $\mathcal{N}_{\rm QMC}$. A better control of this bias is obtained by minimizing the function given in Eq. (30) instead of the one corresponding to Eq. (31) in Sec. III A. In Table III, the two sets of frequencies labeled with "cov" are done with the force covariance matrix, and differ very little from the simpler ones where different force components are assumed to be uncorrelated. This means that this approximation is almost correct in practice.

Finally, in order to quantify more clearly the statistical error, we have also performed a calculation with $\mathcal{N}_{\rm QMC}$ doubled and we have not been able to measure a sizable departure from the measured frequencies (see Table III, the harmonic frequencies at 150 K, "db" indicates double $\mathcal{N}_{\rm QMC}$).

In this work, we have not studied how these different errors will affect larger systems, but it is clear that all the sources of errors, that we have described in detail in this section, can be systematically reduced by changing three parameters $\tau \to 0, n_{\rm opt}, \mathcal{N}_{\rm QMC} \to \infty$ in the same way we have done for the smaller system simulations.

VI. RESULTS

In this section, we summarize our final results on the water monomer in Table III, as well as the calculation of frequencies for several systems of chemical interest within our Jastrow-Slater ansatz (see Tables II and IV).

As far as the water monomer is concerned, the $\chi^2_{\rm red}$ ($\chi^2_{\rm red} \sim 1$) obtained by fitting samples generated by finite temperature molecular dynamics is much smaller and therefore superior to the one obtained in the standard approach ($\chi^2_{\rm red} > 4$) with a fixed grid. In the data for the geometry, the H-O-H angle, a very sensitive parameter, has been improved significantly (0.11° closer to the experimental value) and the O–H bond length, a rather insensitive parameter, remains very good. The two sets of harmonic frequencies for the V_2 and V_4 fitting are both consistent within the error bars. According to our experience, anharmonic fit including the V_4 term is preferred and more robust thanks to its more accurate

TABLE III. Equilibrium geometries and vibrational frequencies of the water monomer obtained by all the methods described in Sec. III.

			Equilibrium geom.		Harmonic freq. (cm^{-1})			Fundamental freq. (cm^{-1})		
Method	Temp. (K)	χ^2_{red}	OH bond (nm)	H-O-H angle (°)	ω_2	ω_1	ω_3	ω_2	ω_1	ω_3
Simple fitting V4	...	4.444	0.9565(1)	104.21(2)	1672(3)	3897(7)	3990(4)	1614(1)	3698(5)	3771(3)
MD fit V2	50	1.007	0.95632(3)	104.32(1)	1676(3)	3893(9)	3890(9)	...	...	...
MD fit V2 cov	50	0.672	0.95634(3)	104.32(1)	1677(3)	3892(9)	3989(8)	...	...	...
MD fit V4	150	1.005	0.95605(3)	104.31(2)	1669(8)	3895(11)	3982(12)	1614(90)	3738(178)	3628(210)
MD fit V4 db	150	1.002	0.95596(3)	104.34(2)	1666(6)	3892(5)	3986(10)	1670(38)	3697(99)	3661(174)
MD fit V4	1000	1.021	0.95597(3)	104.31(3)	1673(3)	3893(3)	3992(4)	1615(2)	3689(5)	3754(5)
MD fit V4 cov	1000	0.681	0.95598(2)	104.31(4)	1672(2)	3892(4)	3993(3)	...	...	...
Covariance matrix	300	...	...	...	1673(24)	3765(28)	3912(20)	...	...	...
Experiment[58]	...	...	0.95721(30)	104.522(50)	1648.47	3832.17	3942.53	1594.59	3656.65	3755.79

parametrization of PES. Furthermore, for the same amount of computation time, the frequencies obtained by V4 fitting have statistical errors about one half smaller than those of the V2 fitting and even smaller than the ones of the simple fitting. In addition, Figs. 6 and 7, show a very nice feature – the lower the frequency is, the smaller the corresponding error is. On the other hand, the calculation based on the covariance matrix method shows that this technique requires much more statistics than all the other methods because it requires a simulation much longer than the correlation time, that in turn can be extremely large at low temperatures.

The other four molecules we considered are divided into two groups – AB_2 and AB_3. In the first group, we use the same parametrization of the Hessian as in H_2O because this type of nonlinear AB_2 molecules is very similar to H_2O. The other group consists of two non-planar AB_3 molecules. The Hessian matrices used in both groups are simplified by using the molecular symmetries in order to improve the accuracy. We choose both JHF and JAGP types of wave function to compute the vibrational frequencies and compare them with CCSD(T) and experimental data from NIST database.[59,60]

H_2S molecule has 8 valence electrons which is exactly the same as the water monomer. In Table II, the equilibrium geometry obtained with JHF wavefunction but without the optimized exponents (JHF_nooptZ) gives the worst values. After optimizing the exponents, both geometry and frequencies

TABLE IV. Vibrational frequencies of H_2S, SO_2, NH_3, and PH_3. The exponents of the primitive basis in both the determinant and Jastrow parts are optimized at the equilibrium configurations and kept fixed during the dynamics.

Name	T(K)	χ^2_{red}	Type	Freq. (cm^{-1})			
H_2S				A_1	A_1	B_2	
JHF_nooptZ	1000	1.077	Harm.	1248(5)	2774(6)	2789(6)	
			Fund.	1217(4)	2652(4)	2656(7)	
JHF	1000	1.018	Harm.	1246(1)	2756(3)	2772(4)	
			Fund.	1215(1)	2639(2)	2654(4)	
JAGP	1000	1.049	Harm.	1235(3)	2752(3)	2767(4)	
			Fund.	1206(2)	2630(1)	2643(2)	
CCSD(T)	...	...	Harm.	1206	2711	2727	
Expt.	...	...	...	1183.0	2615.0	2626.0	
SO_2				A_1	A_1	B_2	
JHF	800	1.024	Harm.	570(4)	1214(8)	1441(11)	
			Fund.	563(3)	1211(5)	1429(8)	
JAGP	800	1.006	Harm.	559(2)	1204(8)	1445(13)	
			Fund.	557(2)	1193(7)	1426(10)	
CCSD(T)	...	...	Harm.	506	1136	1332	
Expt.	...	...	...	517.7	1151.4	1361.8	
NH_3				A_1	E	A_1	E
JHF	50	1.006	Harm.	1064(3)	1712(4)	3523(8)	3651(5)
JAGP	50	1.021	Harm.	1098(7)	1709(6)	3523(7)	3640(7)
CCSD(T)	...	...	Harm.	1159	1673	3476	3607
Expt.	...	...	...	950.0	1627.0	3337.0	3444.0
PH_3				A_1	E	A_1	E
JHF	50	1.025	Harm.	1048(6)	1181(3)	2445(8)	2461(13)
JAGP	50	1.022	Harm.	1045(4)	1178(3)	2431(8)	2437(4)
CCSD(T)	...	...	Harm.	1018	1142	2412	2421
Expt.	...	...	...	992.0	1118.0	2323.0	2328.0

are improved, in agreement with the conclusions of Ref. 9. For this reason, we have optimized all the exponents at equilibrium positions for all the remaining molecules studied. By using the JAGP wavefunction, the H-S-H angle of the equilibrium geometry is further improved as compared with experiments and the fundamental frequencies lower by $\sim$10 cm^{-1} for each mode in the best calculation reported in Table IV. Compared with the experimental data, its RMS difference from fundamental frequencies is only 19 cm^{-1}.

SO_2 molecule has 18 valence electrons and requires a calculation much heavier than H_2S. Similar to H_2S, the use of JAGP wave function provides better results than those obtained with JHF wave function. Its RMS difference of fundamental frequencies from experimental values is 50 cm^{-1}. Even though CCSD(T) frequencies are much closer to the experiments, we should notice that they are harmonic frequencies rather than fundamental ones and the equilibrium geometry of CCSD(T) ($+0.023$ nm for S–O bond and -1.1° for O-S-O angle) is much worse than our values (-0.013 nm for S–O bond and $+0.5^\circ$ for O-S-O angle).

Both NH_3 and PH_3 have 8 valence electrons. Since we do not include the anharmonic correction for the fit, we compare their harmonic frequencies with the corresponding ones calculated with CCSD(T). For both molecules, the frequencies obtained with JAGP wavefunction are again better than those corresponding to JHF wavefunctions. Both NH_3 and PH_3 have equilibrium geometries very close (<0.01 nm for N/P–H bond and $<0.1^\circ$ for H-N/P-H angle) to those obtained by CCSD(T). The RMS difference of harmonic frequencies from CCSD(T) values are 46 cm^{-1} and 26 cm^{-1} for NH_3 and PH_3, respectively.

VII. CONCLUSIONS

In this work, we have studied the performance of a recently developed molecular dynamics scheme based on quantum Monte Carlo. We have considered particularly simple systems by targeting the vibrational properties of simple molecules, that are well studied and understood with well established quantum chemistry methods. In this way, we have been able to identify and systematically control all possible sources of systematic error which may affect this molecular dynamics. The main conclusion of this work is that the statistical error (the finite number of samples $\mathcal{N}_{\mathrm{QMC}}$ used for each iteration of MD) and the time discretization error due to finite τ can be easily pushed to negligible values. On the other hand, we have found that the most difficult bias comes from the requirement to satisfy the BO constraint along the dynamics. We have found that it is important to employ a sufficiently large number $n_{\mathrm{opt}} \gtrsim 10$ of energy optimization for each step of molecular dynamics, in order to satisfactorily fulfill the BO constraint. Since the computational time is proportional to n_{opt}, in the present scheme this is probably the most difficult bias to remove. Despite this difficulty, the calculation remains still feasible and can be extended to large systems by using massively parallel supercomputers.

Our work is also relevant to establish vibrational frequencies in complex electronic systems. Among the three methods that we have used for evaluating vibrational properties, the fitting method with samples generated by finite temperature molecular dynamics gives the best results for the same amount of computation cost. Compared with the standard fitting procedure of Ref. 47, it is easier and more systematic to set up and use, and yields better distributions of the configurations around the equilibrium structure, thus improving the quality of fit as well as an equilibrium geometry closer to the experiment. Even though our method based on the force-force correlations is the most direct and simplest approach, it usually requires much more statistics. All methods, apart from the one containing the anharmonic corrections, have favorable scaling with the system size, and are in principle very promising because can be extended to large systems, as well as generalized to the calculation of phonons in solids. However we have seen that, in order to neglect anharmonic effects, we have to work with so small temperatures that it is already very difficult to simulate a slightly larger system (such as the water dimer). On the other hand, we expect that the method which includes in the fitting also the anharmonic corrections, should work also for larger systems, despite the difficulty to represent the V4 term with a number of parameters scaling with the fourth power of the number of atoms. Also for this reason this method is very difficult to implement in practice for large systems, and therefore we have limited our study to molecules containing at most four atoms.

A very interesting feature that we have noted in the estimation of vibrational frequencies by QMC is that the small frequencies are much less biased by the systematic errors in our tests. This is really promising because small frequencies are often more interesting as they characterize the intermolecular interactions, whereas the high frequency modes are determined by the well understood intra-molecular properties. Moreover, we have systematically found that the use of the JAGP wave function in place of the more commonly-adopted Jastrow-Slater paradigm, improves significantly the calculation of both equilibrium structures and vibrational frequencies, basically without extra computational effort.

As well-known QMC scales very well with system size and, once the problem of including anharmonic effects will be solved at least in an approximate way, say by self-consistent harmonic approximation,[61] the computation of vibrational frequencies in large systems will be possible with a reasonable cost. In addition, we have shown that the present molecular dynamics can be extended to large systems already at present,[15,62] provided the little systematic errors and especially the BO constraint are under control in the way we have carefully described in this work.

ACKNOWLEDGMENTS

We acknowledge R. Henning for his suggestion to use the position covariance matrix to compute harmonic frequencies.

APPENDIX A: INTEGRATION OF SLD EQUATIONS

In this Appendix, we sketch how to integrate exactly the differential Eq. (1),

$$\dot{\boldsymbol{v}}(s) = -\boldsymbol{\gamma}(\boldsymbol{R}) \cdot \boldsymbol{v}(s) + \boldsymbol{f}(\boldsymbol{R}) + \boldsymbol{\eta}(s) \qquad \text{(A1)}$$

in an arbitrary interval $\bar{t} \le s \le t$ within the assumption that the vector $\boldsymbol{R}$ is not changing much during the integration interval and that therefore it can be considered independent of s. In this Appendix, in order to avoid confusion, we indicate by s the generic time defining the SLD dynamics, whereas with $\bar{t}$ and t, the initial and final time of the integration, respectively, so that the initial condition reads

$$\boldsymbol{v}(\bar{t}) = \bar{\boldsymbol{v}}\,. \tag{A2}$$

As well known this kind of equations can be solved in terms of the simple exponential solution $\boldsymbol{v}(s) = \exp\left[\boldsymbol{\gamma}(\boldsymbol{R})(\bar{t}-s)\right]\bar{\boldsymbol{v}}$, valid in absence of the external force and the noise (i.e., $\boldsymbol{f}(\boldsymbol{R}) + \boldsymbol{\eta}(s) = 0$). We search therefore a solution of the form

$$\boldsymbol{v}(s) = \exp[-\boldsymbol{\gamma}(\boldsymbol{R})s]\boldsymbol{y}(s)\,. \tag{A3}$$

By replacing the above equation in Eq. (A1), we easily obtain that

$$\dot{\boldsymbol{y}}(s) = \exp[\boldsymbol{\gamma}(\boldsymbol{R})s][\boldsymbol{f}(\boldsymbol{R}) + \boldsymbol{\eta}(s)] \tag{A4}$$

with the initial condition given by inverting Eq. (A3) for $s = \bar{t}$,

$$\boldsymbol{y}(\bar{t}) = \exp[\boldsymbol{\gamma}(\boldsymbol{R})\bar{t}]\bar{\boldsymbol{v}}\,. \tag{A5}$$

Equation (A4) can be integrated immediately from $\bar{t}$ to t, because its RHS is a known function of s,

$$\boldsymbol{y}(s) = \boldsymbol{y}(\bar{t}) + \int_{\bar{t}}^{s} \exp[\boldsymbol{\gamma}(\boldsymbol{R})t'][\boldsymbol{f}(\boldsymbol{R}) + \boldsymbol{\eta}(t')]dt'\,. \tag{A6}$$

We now go back to the original ansatz (Eq. (A3)), and by replacing the initial condition (Eq. (A5)) in the above equation, we obtain the final solution

$$\boldsymbol{v}(t) = \exp[\boldsymbol{\gamma}(\boldsymbol{R})(\bar{t}-t)]\boldsymbol{v}(\bar{t}) + \int_{\bar{t}}^{t} \exp[\boldsymbol{\gamma}(\boldsymbol{R})(t'-t)][\boldsymbol{f}(\boldsymbol{R}) + \boldsymbol{\eta}(t')]dt'\,. \tag{A7}$$

APPENDIX B: BETTER INTEGRATION SCHEME

In this Appendix, we describe how to avoid the approximation in Eq. (8) to integrate Eq. (2), with a more involved method, that was already introduced in Ref. 15. However, we have noted that in the proposed integrator it is not necessary to compute the velocities at half-integer times because we perform the integration of Eq. (2) in an exact unbiased way. In the following, we describe this derivation and obtain expression very similar to the ones introduced in Ref. 15, with the main difference that here we use integer time both for velocities and positions

$$\boldsymbol{v}_n \equiv \boldsymbol{v}(t_n), \tag{B1}$$

$$\boldsymbol{R}_n \equiv \boldsymbol{R}(t_n)\,. \tag{B2}$$

Having the general expression of the velocity by Eq. (A7), we can use to integrate Eqs. (1) and (2) in the interval $t_n \le s \le t_{n+1}$ and obtain, with a little involved algebra, just a bit more than the original scheme[8] described in Sec. II,

$$\boldsymbol{v}_{n+1} = e^{-\boldsymbol{\gamma}_n\tau}\boldsymbol{v}_n + \boldsymbol{\Gamma}_n \cdot (\boldsymbol{f}_n + \tilde{\boldsymbol{\eta}}), \tag{B3}$$

$$\boldsymbol{R}_{n+1} = \boldsymbol{R}_n + \boldsymbol{\Gamma}_n \cdot \boldsymbol{v}_n + \boldsymbol{\Theta}_n \cdot (\boldsymbol{f}_n + \tilde{\tilde{\boldsymbol{\eta}}}), \tag{B4}$$

where we have introduced the following matrices, mainly to single out, as before, the actual noisy terms $\tilde{\boldsymbol{\eta}}$ and $\tilde{\tilde{\boldsymbol{\eta}}}$, that have to be added to the force components in the above Markov iterations for the velocities and coordinates, respectively,

$$\boldsymbol{\Gamma}_n = \boldsymbol{\gamma}_n^{-1}(\mathbf{I} - e^{-\boldsymbol{\gamma}_n\tau}), \tag{B5}$$

$$\boldsymbol{\Theta}_n = \boldsymbol{\gamma}_n^{-1}(\tau\mathbf{I} - \boldsymbol{\Gamma}_n)\,, \tag{B6}$$

$$\tilde{\boldsymbol{\eta}} = \boldsymbol{\Gamma}_n^{-1}e^{-\boldsymbol{\gamma}_n\tau}\int_{t_n}^{t_{n+1}} e^{\boldsymbol{\gamma}_n(t-t_n)}\boldsymbol{\eta}(t)\,dt = \boldsymbol{\Gamma}_n^{-1}e^{-\boldsymbol{\gamma}_n\tau}\int_{0}^{\tau} e^{\boldsymbol{\gamma}_n t}\boldsymbol{\eta}(t)\,dt\,, \tag{B7}$$

$$\tilde{\tilde{\boldsymbol{\eta}}} = \boldsymbol{\Theta}_n^{-1}\int_{t_n}^{t_{n+1}} dt \int_{t_n}^{t} dt' e^{\boldsymbol{\gamma}_n(t'-t)}\boldsymbol{\eta}(t') = \boldsymbol{\Theta}_n^{-1}\int_{0}^{\tau} dt e^{-\boldsymbol{\gamma}_n t}\int_{0}^{t} dt' e^{\boldsymbol{\gamma}_n t'}\boldsymbol{\eta}(t')\,. \tag{B8}$$

In order to define the Markov process, it is enough to compute the correlation of the previously mentioned noisy terms, which we define as follows:

$$\langle\tilde{\boldsymbol{\eta}}_i\tilde{\boldsymbol{\eta}}_j\rangle \equiv \bar{\boldsymbol{\alpha}}^{\mathbf{1,1}}{}_{ij}, \tag{B9}$$

$$\langle\tilde{\tilde{\boldsymbol{\eta}}}_i\tilde{\tilde{\boldsymbol{\eta}}}_j\rangle \equiv \bar{\boldsymbol{\alpha}}^{\mathbf{2,2}}{}_{ij}, \tag{B10}$$

$$\langle\tilde{\boldsymbol{\eta}}_i\tilde{\tilde{\boldsymbol{\eta}}}_j\rangle \equiv \bar{\boldsymbol{\alpha}}^{\mathbf{1,2}}{}_{ij}, \tag{B11}$$

$$\langle\tilde{\tilde{\boldsymbol{\eta}}}_i\tilde{\boldsymbol{\eta}}_j\rangle \equiv \bar{\boldsymbol{\alpha}}^{\mathbf{2,1}}{}_{ij} = \bar{\boldsymbol{\alpha}}^{\mathbf{1,2}}{}_{ij}. \tag{B12}$$

Then a straightforward integration in time, by using that the assumed correlation is given by Eq. (3) and that, by Eq. (4) the corresponding matrix $\bar{\boldsymbol{\alpha}} = 2T\boldsymbol{\gamma}$, we obtain

$$\bar{\boldsymbol{\alpha}}^{\mathbf{1,1}} = T\boldsymbol{\gamma}_n^2\coth(\boldsymbol{\gamma}_n\tau/2),$$
$$\bar{\boldsymbol{\alpha}}^{\mathbf{2,2}} = T(2\boldsymbol{\Theta}_n - \boldsymbol{\Gamma}_n^2)\cdot\boldsymbol{\Theta}_n^{-2},$$
$$\bar{\boldsymbol{\alpha}}^{\mathbf{1,2}} = T\boldsymbol{\gamma}_n\boldsymbol{\Gamma}_n\cdot\boldsymbol{\Theta}_n^{-1}.$$

The above Markov process can be straightforwardly implemented, as well as the very similar one described in Ref. 15. However, we have tested that all methods, including the simplest one described in Sec. II, behave equally well, with comparable performances, probably because a too high accurate integration scheme is not necessary for the available accuracy, possible at present with QMC.

As discussed also in Ref. 8 (see also Sec. III C), all the QMC force evaluations $\boldsymbol{f}$ are affected by an intrinsic stochastic noise, which usually determines an effective temperature

higher than the target one. This problem can be avoided, by generalizing the method of the *noise correction* described in Sec. II to this specific case. Indeed, we can follow the correct dynamics by adding to the QMC noise of the forces the two external noises $\tilde{\boldsymbol{\eta}}_{\text{ext}}$ and $\tilde{\tilde{\boldsymbol{\eta}}}_{\text{ext}}$ so that the total noises $\tilde{\boldsymbol{\eta}}$ and $\tilde{\tilde{\boldsymbol{\eta}}}$ satisfy the correct expressions in Eqs. (B9)–(B12). In this way, we have to subtract the $3N \times 3N$ QMC correlation of the forces $\boldsymbol{\alpha}^{\text{QMC}}$ to each of the four submatrices, namely,

$$\bar{\boldsymbol{\alpha}}^{a,b}_{\text{ext}} = \bar{\boldsymbol{\alpha}}^{a,b} - \boldsymbol{\alpha}^{\text{QMC}}, \tag{B13}$$

is the true external noise we have to add to the system, to take into account that QMC forces contain already a correlated noise, that is independently evaluated during the dynamics. It can be shown that the resulting matrix $\bar{\boldsymbol{\alpha}}_{\text{ext}}$ is indeed positive definite provided Δ_0 is large enough, so that $\bar{\boldsymbol{\alpha}}_{\text{ext}}$ is a well defined correlation for an external noise.

[1] C. J. Umrigar, J. Toulouse, C. Filippi, S. Sorella, and H. Rhenning, Phys. Rev. Lett. **98**, 110201 (2007).
[2] E. Neuscamman, C. J. Umrigar, and G. K.-L. Chan, Phys. Rev. B **85**, 045103 (2012).
[3] S. Sorella, M. Casula, and D. Rocca, J. Chem. Phys. **127**, 014105 (2007).
[4] W. McMillan, Phys. Rev. **138**, A442 (1965).
[5] R. Assaraf and M. Caffarel, J. Chem. Phys. **113**, 4028 (2000).
[6] R. Assaraf and M. Caffarel, J. Chem. Phys. **119**, 10536 (2003).
[7] S. Sorella and L. Capriotti, J. Chem. Phys. **133**, 234111 (2010).
[8] C. Attaccalite and S. Sorella, Phys. Rev. Lett. **100**, 114501 (2008).
[9] A. Zen, Y. Luo, S. Sorella, and L. Guidoni, J. Chem. Theory Comput. **9**, 4332 (2013).
[10] E. Coccia, D. Varsano, and L. Guidoni, J. Chem. Theory Comput. **9**, 8 (2013).
[11] M. Barborini and L. Guidoni, J. Chem. Phys. **137**, 224309 (2012).
[12] E. Coccia, D. Varsano, and L. Guidoni, J. Chem. Theory Comput. **10**, 501 (2014).
[13] R. Guareschi and C. Filippi, J. Chem. Theory Comput. **9**, 5513 (2013).
[14] J. C. Grossmann and L. Mitas, Phys. Rev. Lett. **94**, 056403 (2005).
[15] G. Mazzola, S. Yunoki, and S. Sorella, Nat. Commun. **5**, 3847 (2014).
[16] F. A. Reboredo and J. Kim, J. Chem. Phys. **140**, 074103 (2014).
[17] R. Car and M. Parrinello, Phys. Rev. Lett. **55**, 2471 (1985).
[18] A. Laio, S. Bernard, G. Chiarotti, S. Scandolo, and E. Tosatti, Science **287**, 1027 (2000).
[19] S. Scandolo, Proc. Natl. Acad. Sci. U.S.A. **100**, 3051 (2003).
[20] C. Cavazzoni, Science **283**, 44 (1999).
[21] G. Kresse and J. Hafner, Phys. Rev. B **47**, 558 (1993).
[22] G. Kresse and J. Hafner, Phys. Rev. B **49**, 14251 (1994).
[23] U. Rothlisberger and W. Andreoni, J. Chem. Phys. **94**, 8129 (1991).
[24] J. VandeVondele, F. Mohamed, M. Krack, J. Hutter, M. Sprik, and M. Parrinello, J. Chem. Phys. **122**, 014515 (2005).
[25] D. Alfe, M. J. Gillan, and G. D. Price, Earth Planet. Sci. Lett. **195**, 91 (2002).
[26] M. Bernasconi, P. Silvestrelli, and M. Parrinello, Phys. Rev. Lett. **81**, 1235 (1998).
[27] D. Ceperley and M. Dewing, J. Chem. Phys. **110**, 9812 (1999).
[28] C. Pierleoni and D. M. Ceperley, Lect. Notes Phys. **703**, 641 (2006).
[29] M. A. Morales, C. Pierleoni, E. Schwegler, and D. M. Ceperley, Proc. Natl. Acad. Sci. U.S.A. **107**, 12799 (2010).
[30] M. A. Morales, C. Pierleoni, and D. M. Ceperley, Phys. Rev. E **81**, 021202 (2010).
[31] S. Duane, A. Kennedy, B. Pendleton, and D. Roweth, Phys. Lett. B **195**, 216 (1987).
[32] E. B. Wilson, J. C. Decius, and P. C. Cross, *Molecular Vibrations: The Theory of Infrared and Raman Vibrational Spectra* (Dover Publications, 1955).
[33] M.-P. Gaigeot, M. Martinez, and R. Vuilleumier, Mol. Phys. **105**, 2857 (2007).
[34] B. A. Hess, L. J. Schaad, P. Carsky, and R. Zahradnik, Chem. Rev. **86**, 709 (1986).
[35] A. P. Scott and L. Radom, J. Phys. Chem. **100**, 16502 (1996).
[36] S. Carter, S. J. Culik, and J. M. Bowman, J. Chem. Phys. **107**, 10458 (1997).
[37] J. Koput, S. Carter, and N. C. Handy, J. Chem. Phys. **115**, 8345 (2001).
[38] P. Cassam-Chenai and J. Lievin, Int. J. Quantum Chem. **93**, 245 (2003).
[39] J. M. Bowman, Acc. Chem. Res. **19**, 202 (1986).
[40] R. B. Gerber and M. A. Ratner, Adv. Chem. Phys. **70**, 97 (1988).
[41] G. M. Chaban, J. O. Jung, and R. B. Gerber, J. Chem. Phys. **111**, 1823 (1999).
[42] G. Czako, T. Furtenbacher, A. G. Csaszar, and V. Szalay, Mol. Phys. **102**, 2411 (2004).
[43] E. Matyus, G. Czako, B. T. Sutcliffe, and A. G. Csaszar, J. Chem. Phys. **127**, 084102 (2007).
[44] W. Schneider and W. Thiel, Chem. Phys. Lett. **157**, 367 (1989).
[45] D. A. Clabo, W. D. Allen, R. B. Remington, Y. Yamaguchi, and H. F. Schaefer, Chem. Phys. **123**, 187 (1988).
[46] V. Barone, J. Chem. Phys. **122**, 014108 (2005).
[47] A. Zen, D. Zhelyazov, and L. Guidoni, J. Chem. Theory Comput. **8**, 4204 (2012).
[48] R. Kubo, M. Toda, and N. Hashitsume, *Statistical Physics II*, Springer Series in Solid-State Sciences Vol. 31 (Springer, Berlin, 1985).
[49] M.-P. Gaigeot and M. Sprik, J. Phys. Chem. B **107**, 10344 (2003).
[50] S. Sorella, TurboRVB quantum Monte Carlo package see http://people.sissa.it/sorella/web/index.html (accessed May 2013).
[51] T. Schlick, *Molecular Modeling and Simulation: An Interdisciplinary Guide* (Springer, 2002).
[52] T. Kühne, M. Krack, F. Mohamed, and M. Parrinello, Phys. Rev. Lett. **98**, 066401 (2007).
[53] F. Krajewski and M. Parrinello, Phys. Rev. B **73**, 041105 (2006).
[54] P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G. L. Chiarotti, M. Cococcioni, I. Dabo *et al.*, J. Phys. Cond. Matt. **21**, 395502 (2009).
[55] E. B. Wilson, J. Chem. Phys. **9**, 76 (1941).
[56] M. Burkatzki, C. Filippi, and M. Dolg, J. Chem. Phys. **126**, 234105 (2007).
[57] S. Sorella, "Many-Electron Approaches in Physics, Chemistry and Mathematics," edited by Walker Bach and Luigi Delle Site (Springer, Berlin, 2013), pp. 207–236.
[58] W. S. Benedict, N. Gailar, and E. K. Plyler, J. Chem. Phys. **24**, 1139 (1956).
[59] See http://webbook.nist.gov/chemistry for Nist chemistry webbook.
[60] See http://cccbdb.nist.gov/ for computational chemistry comparison and benchmark database.
[61] I. Errea, M. Calandra, and F. Mauri, Phys. Rev. B **89**, 064302 (2014).
[62] A. Zen, Y. Luo, G. Mazzola, L. Guidoni, and S. Sorella, "Ab-initio molecular dynamics simulation of liquid water by Quantum Monte Carlo," (unpublished).

[63] In this section, the Boltzmann constant k is conventionally set to one for simplicity.
[64] In an electronic system containing atoms with different masses, we can scale each length x_i corresponding to a mass M_i by $x_i = \sqrt{\frac{1}{M_i}} x'_i$. After a little algebra, one obtains the Langevin equation with unit masses given in Eq. (2) and a scaled friction matrix $\gamma'_{i,j} = \gamma_{i,j} / \sqrt{M_i M_j}$ with the fluctuation dissipation theorem written in the same form as in Eq. (4).
[65] We adopt in the paper the conventional choice, that comes from spectroscopic conventions, to report the wavenumber $\bar{\nu}$, expressed in cm^{-1}, corresponding to the frequency ω. They are related by the equation $\omega = 2\pi c \bar{\nu}$, being c the speed of light in a vacuum.