Communication: Toward an improved control of the fixed-node error in quantum Monte Carlo: The case of the water molecule

Michel Caffarel, Thomas Applencourt, Emmanuel Giner, and Anthony Scemama

Communication: Toward an improved control of the fixed-node error in quantum Monte Carlo: The case of the water molecule

Michel Caffarel,[1] Thomas Applencourt,[1] Emmanuel Giner,[2] and Anthony Scemama[1]
[1]*Laboratoire de Chimie et Physique Quantique, CNRS-Université de Toulouse, Toulouse, France*
[2]*Dipartimento di Scienze Chimiche e Farmaceutiche, Universit degli Studi di Ferrara, Ferrara, Italy*

All-electron Fixed-node Diffusion Monte Carlo calculations for the nonrelativistic ground-state energy of the water molecule at equilibrium geometry are presented. The determinantal part of the trial wavefunction is obtained from a selected Configuration Interaction calculation [Configuration Interaction using a Perturbative Selection done Iteratively (CIPSI) method] including up to about 1.4×10^6 of determinants. Calculations are made using the cc-pCVnZ family of basis sets, with $n = 2$ to 5. In contrast with most quantum Monte Carlo works no re-optimization of the determinantal part in presence of a Jastrow is performed. For the largest cc-pCV5Z basis set the lowest upper bound for the ground-state energy reported so far of $-76.437\,44(18)$ is obtained. The fixed-node energy is found to decrease regularly as a function of the cardinal number n and the Complete Basis Set limit associated with *exact nodes* is easily extracted. The resulting energy of $-76.438\,94(12)$ — in perfect agreement with the best experimentally derived value — is the most accurate theoretical estimate reported so far. We emphasize that employing selected configuration interaction nodes of increasing quality in a given family of basis sets may represent a simple, deterministic, reproducible, and systematic way of controlling the fixed-node error in diffusion Monte Carlo. *Published by AIP Publishing.* [http://dx.doi.org/10.1063/1.4947093]

The only uncontrolled source of error[1] in quantum Monte Carlo (QMC) methods is the fixed-node approximation introduced to suppress the wild fluctuations of the sign of the wavefunction (fermion sign problem). Although the fixed-node error is small (typically, a few percents of the correlation energy), and many fixed-node QMC calculations of impressive accuracy have been realized, the error can still be too large in some applications, particularly in the important case of the computation of (very) small energy differences.

A major challenge for QMC is thus to set up a strategy of construction of trial wavefunctions having "good" nodes and, even more importantly, to propose a *systematic* way of improving such nodes. In practice, a standard strategy consists in introducing trial wavefunctions of the best possible quality and then to optimize their parameters in a preliminary Variational Monte Carlo (VMC) step through minimization of the variational energy or its variance.[2] Many functional forms for the trial function Ψ_T have been explored in the literature, the most popular one being the Jastrow Slater form[3]

$$\Psi_T = e^J \sum_i c_i D_i, \qquad (1)$$

combining a Jastrow prefactor e^J containing explicit electronic correlations and a short multi-determinantal expansion (typically, a few thousands of determinants) describing the multireference character of the wavefunction (static correlation effects).

Very recently some of us have proposed to keep the standard Jastrow Slater form for the trial wavefunction but to rely on the more conventional Configuration Interaction (CI) expansions of quantum chemistry for the multideterminantal

part. No stochastic re-optimization of the CI expansion is performed, so that "pure CI" nodes are employed.[4–6] The rationale behind this proposal is to search for a better control of the fixed-node error by exploiting the unique properties of CI wavefunctions. Indeed, CI approaches provide a simple, deterministic, and systematic way to build wavefunctions of controllable quality. In a given one-particle basis set, the wavefunction is improved by increasing the number of determinants, up to the Full CI (FCI) limit. Then, by increasing the basis set, the wavefunction can be further improved, up to the complete basis set (CBS) limit where the exact solution of the electronic Schrödinger equation is reached. CI nodes, defined as the zeroes of the CI expansions, are also expected to display such a systematic improvement. The main difficulty is of course the exponential growth of the space of determinants with respect to the number of electrons and orbitals. However, this severe exponential increase can be dramatically attenuated by considering *Selected* CI (SCI) approaches designed to keep only the most important determinants. In practice, we have proposed to make use of the CIPSI method (Configuration Interaction using a Perturbative Selection done Iteratively),[7,8] one of the numerous variants of SCI proposed in the literature (see, e.g., Refs. 7–15). In this approach the multideterminant expansion is built iteratively by selecting determinants according to the importance of their second-order perturbational contribution to the total energy. As illustrated by a number of applications, CIPSI represents a very efficient way of approaching the FCI limit using only a *tiny* fraction of the total FCI space (see, for example, a recent all-electron FCI-converged CIPSI calculation for $CuCl_2$ involving 25 electrons and 36 active orbitals for a FCI

space including 10^{18} determinants[16]). This remarkable result is actually common to all variants of SCI approaches, including the FCI-QMC approach of Alavi *et al.*,[17,18] which can be considered as a stochastic version of SCI. In practice, the main difficulty in using lengthy multideterminant expansions in QMC is the expensive cost of evaluating at each Monte Carlo step the first and second derivatives of the trial wavefunction (drift vector and local energy). However, efficient algorithms have been proposed to perform such calculations.[19–21] Here, we shall use our recently introduced algorithm allowing to perform converged diffusion Monte Carlo (DMC) calculations using multideterminant expansions including up to a few millions of determinants for a system like the water molecule.[22]

A remarkable property systematically observed so far in our first DMC applications using large CIPSI expansions[4–6] is that, except for a possible transient regime at small number of determinants,[23] the fixed-node error associated with CIPSI nodes decreases monotonically, both as a function of the number of determinants and of the basis set size, leading to the possibility of a control of the fixed-node error. Such a result is known not to be systematically true for a general CI expansion (see, e.g., Ref. 24). However, its validity here could be attributed to the fact that determinants are selected in a hierarchical way (the most important ones first), so that the wavefunction quality increases step by step, and so the quality of nodes.

In this Communication all-electron DMC/CIPSI calculations for the water molecule at equilibrium geometry using the cc-pCVnZ family of basis sets with n ranging from 2–5 and large multideterminant expansions including up to 1 423 377 determinants are presented. The lowest (upper bound) fixed-node energy reported so far of $-76.437\,44(18)$ is obtained. Performing the Complete Basis set (CBS) limit by extrapolating fixed-node energies as a function of the cardinal number n of the basis set a value of $-76.438\,94(12)$ for the total energy associated with exact nodes is obtained, in full agreement with the best known estimate of -76.4389.[25]

CIPSI expansion. The multideterminant CIPSI expansion is built by selecting iteratively the most important determinants of the FCI expansion. In short (for more details, see Ref. 4), at iteration n the multideterminant expansion Ψ_D is written as the sum of the $N^{(n)}$ previously selected determinants (thus, defining the reference space at this iteration)

$$\Psi_D^{(n)} = \sum_{i=1}^{N^{(n)}} c_i^{(n)} D_i, \qquad (2)$$

with energy $E_0^{(n)} = \frac{\langle \Psi_D^{(n)} | H | \Psi_D^{(n)} \rangle}{\langle \Psi_D^{(n)} | \Psi_D^{(n)} \rangle}$. Then, one determinant (or a group of determinants) D_j not belonging to the reference space and corresponding to the greatest second-order energy change (or close to it within some threshold),

$$\delta E = -\frac{\langle \Psi_D^{(n)} | H | D_j \rangle^2}{\langle D_j | H | D_j \rangle - E_0^{(n)}}, \qquad (3)$$

is (are) selected and added to the reference space. At iteration $(n + 1)$ the new expansion $\Psi_D^{(n+1)}$ and energy $E_0^{(n+1)}$ are obtained by diagonalizing the Hamiltonian matrix within the new set of selected determinants. The iterative process is started with the Hartree-Fock determinant or a short expansion and is stopped when a target number of determinants is reached. In what follows the variational energy associated with the final CI expansion will be denoted as E_0^{var}.

Water molecule. In this study we present benchmark calculations for the non-relativistic ground-state energy of the water molecule at equilibrium geometry, $R_{OH} = 0.9572$ Å and $\theta_{OH} = 104.52°$.

CIPSI results. All configuration interaction calculations have been carried out using our perturbatively selected CI program QUANTUM PACKAGE (downloadable at Ref. 26). Standard Dunning type correlation-consistent polarized core-valence basis sets cc-pCVnZ with n going from 2–5 are employed. CIPSI calculations have been performed using natural orbitals issued from the diagonalization of the one-body density matrix obtained in a preliminary CIPSI run. For each basis set, the selected CI expansion has been stopped for one million determinants, except for the largest cc-pCV5Z basis sets for which two million determinants were considered. Results are presented in Table I and compared to the recent benchmark CI calculations of Almora-Dìaz including up to sextuple excitations.[27] As we shall see below, truncated versions of these one- and two million-determinant CIPSI expansions will actually be used in DMC, results are thus presented for these shorter expansions. A remarkable point is the high efficiency of CIPSI in obtaining accurate CI expansion with a small number of determinants. For the cc-pCVDZ basis set, the variational energy obtained with the 172 256 determinants used in DMC is different from the FCI value of Almora-Dìaz by only 0.7 mhartree. For the other basis sets, the differences remain small, that is 1.8, 1.8, and 2.5 mhartree for the cc-pCVTZ, cc-pCVQZ, and cc-pCV5Z basis sets, respectively.

Fixed-node DMC (FN-DMC) results. All-electron DMC calculations have been realized using our general-purpose QMC program QMC=CHEM (downloadable at Ref. 28).

TABLE I.   Number of determinants and corresponding variational energies for CIPSI expansions used in DMC for each cc-pCVnZ (n = 2 to 5) basis set. Last column: Deviations of the variational energy to the best FCI estimates of Almora-Dìaz.[27] Energies in atomic units.

Basis set	FCI size	# dets used in DMC	E_0^{var}	FCI, Almora-Dìaz[27]	Deviation
cc-pCVDZ	$\sim 10^{10}$	172 256	$-76.282\,136$	$-76.282\,865$	0.0007
cc-pCVTZ	$\sim 2 \cdot 10^{14}$	640 426	$-76.388\,287$	$-76.390\,158$	0.0018
cc-pCVQZ	$\sim 2 \cdot 10^{17}$	666 927	$-76.419\,324$	$-76.421\,148$	0.0018
cc-pCV5Z	$\sim 7 \cdot 10^{19}$	1 423 377	$-76.428\,550$	$-76.431\,105$	0.0025

A minimal Jastrow prefactor taking care of the electron-electron cusp condition is employed and molecular orbitals are slightly modified at very short electron-nucleus distances to impose exact electron-nucleus cusp conditions. The time step used, $\tau = 2.10^{-4}$ a.u., has been chosen small enough to make the finite time step error not observable with statistical fluctuations.

To accelerate DMC calculations and not to use the full one- and two million-determinant expansion of the initial CIPSI calculations we have employed the improved truncation scheme described in Ref. 22. In short, the approach consists in writing the CI expansion as

$$\sum_{i=1}^{N_{\text{det}}^{\uparrow}} \sum_{j=1}^{N_{\text{det}}^{\downarrow}} C_{ij} D_i^{\uparrow}(\mathbf{R}_{\uparrow}) D_j^{\downarrow}(\mathbf{R}_{\downarrow}).$$

For each σ-determinant D_k^{σ} ($\sigma = \uparrow, \downarrow$), the contribution to the norm of the wavefunction is given either by $\sum_{l=1}^{N_{\text{det}}^{\downarrow}} C_{kl}^2$ for $\sigma = \uparrow$ or $\sum_{l=1}^{N_{\text{det}}^{\uparrow}} C_{lk}^2$ for $\sigma = \downarrow$. If this contribution is below a given threshold ϵ_{σ} the σ-determinant D_k^{σ} is discarded (C_{kl} or $C_{lk} = 0 \; \forall l$, for the $\uparrow$- or $\downarrow$-sector, respectively). Here, N_{det}^{σ} denotes the number of *different* σ-determinants in the CI expansion. Such numbers being usually much smaller than the total number of products of determinants $D(\mathbf{R}) = D^{\uparrow} D^{\downarrow}$, the gain in computational cost can be important (see, Table 5 of Ref. 22). Here, we chose to truncate the expansion by taking $\epsilon_{\uparrow} = \epsilon_{\downarrow} = 10^{-8}$, except for the cc-pCV5Z basis where a value of 10^{-9} has been used. Values for ϵ_{σ} have been chosen small enough to get converged fixed-node energies as a function of the number of selected determinants within statistical errors. In other words, nodes employed in this work are expected to be close to FCI nodes. The final numbers of selected determinants used are given in Table I.

The efficiency of our algorithm for computing large multideterminant expansions can be quantified by measuring the ratio of CPU times needed to realize one Monte Carlo step using either the full expansion or only the single HF determinant. Such ratios are presented in Table II for each basis set. Fixed-node DMC energies (in atomic units) obtained with CIPSI nodes for the various basis sets are also given in Table II and plotted in Fig. 1 as a function of the inverse of the cardinal number $n = 2$ to 5. The horizontal line is the best estimate of the total nonrelativistic energy reported in the literature, see Ref. 25. For comparison, we have also reported the best estimates of the FCI energies of Almora-Dìaz. Quite remarkably both sets of points display a very similar overall behavior. In particular, the values converge smoothly to the
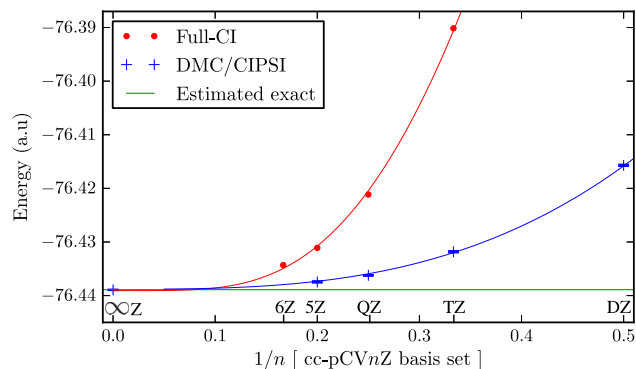

FIG. 1. CBS extrapolation of FCI and DMC/CIPSI energies. Error bars on DMC data are plotted but almost imperceptible.

same CBS limit as a function of the cardinal number n with a typical inverse third power law. Using a simple two-parameter fitting function, $E_0(n) = E_0(\text{CBS}) + an^{-3}$, the CBS limit for DMC results gives an extrapolated value of $-76.438\,94(12)$. The error bar has been estimated by reproducing the fit over a large statistical ensemble of independent data drawn according to their respective error bars. No correlation between data being considered, the error value should be considered as rather conservative. Note that the energy of $-76.437\,44 \pm 0.000\,18$ obtained with the cc-pCV5Z nodes is the lowest upper bound reported so far in DMC or any other approach. Regarding computational aspects, calculation of each FN-DMC energies of Table II were performed using 800 cores on the Curie machine (TGCC/CEA/Genci) during about 15 h. The cost of deterministic CIPSI calculations to build the trial wavefunctions is marginal. Roughly speaking, the cost is similar to that needed for making CISD calculations with the same basis sets.

In Table III a selection of the best (lowest) values reported in the literature for the total energy of the water molecule is presented. Using DMC, the lowest value published so far is that of Clark *et al.* of $-76.4368(4)$. Here, using the nodes of the CIPSI/cc-pCV5Z expansion an improved value of $-76.437\,44(18)$ is obtained. The lowest upper bound reached using a post-Hartree Fock correlated approach is that of Almora-Dìaz of -76.4343, a value significantly higher than DMC values. Finally, the best (non-variational) estimates are those obtained by performing CBS extrapolation. At FCI level the most accurate one is that of Bytautas and Ruedenberg,[32]

TABLE II. All-electron DMC energies (in a.u.) obtained with CIPSI nodes for each basis set. Second column: Increase of CPU time due to the use of the large multideterminant expansion.

Basis set[Ndets]	$T_{\text{CPU}}(\text{Ndets})/T_{\text{CPU}}(\text{1det})$	E_0^{DMC}
cc-pCVDZ[172 256]	~101	$-76.415\,71(20)$
cc-pCVTZ[640 426]	~185	$-76.431\,82(19)$
cc-pCVQZ[666 927]	~128	$-76.436\,22(14)$
cc-pCV5Z[1 423 377]	~235	$-76.437\,44(18)$

TABLE III. Comparison of nonrelativistic ground-state total energies of water obtained with the most accurate theoretical methods. Energies in a.u.

Clark *et al.*,[20] DMC (upper bound)	$-76.436\,8(4)$
This work, DMC (upper bound)	$-76.437\,44(18)$
Almora-Dìaz,[27] CISDTQQnSx (upper bound)	$-76.434\,3$
Helgaker *et al.*,[29] R12-CCSD(T)	$-76.439(2)$
Muller and Kutzelnigg,[30] R12-CCSD(T)	$-76.437\,3$
Almora-Dìaz,[27] FCI + CBS	$-76.438\,6(9)$
Halkier *et al.*,[31] CCSD(T) + CBS	$-76.438\,6$
Bytautas and Ruedenberg,[32] FCI + CBS	$-76.439\,0(4)$
This work, DMC + CBS	$-76.438\,94(12)$
Experimentally derived estimate[25]	$-76.438\,9$

$E_0 = -76.4390(4)$. Here, our value of $-76.438\,94(12)$ is, to the best of our knowledge, the most accurate value reported so far. In both cases the best experimentally derived estimate of -76.4389 is recovered within error bars.

In this study we have performed DMC calculations using nodes of multideterminant CI expansions obtained through a perturbative selection of the most important determinants (selected CI). In contrast with most QMC works, no-reoptimization of nodes in presence of a Jastrow prefactor has been performed. For each basis set of the cc-pCVnZ family ($n = 2$–5), CIPSI nodes obtained are of near-Full-CI quality. As a result of the deterministic construction of nodes using CI expansions, the total fixed-node energy is found to be a smoothly decreasing function of the cardinal number of the basis set with a typical inverse third power law. The Complete Basis Set (CBS) limit leading to the total energy associated with *exact nodes* is then easy to perform. From a general perspective, we emphasize that employing selected CI nodes of increasing quality in a given family of basis sets may represent a simple, deterministic, reproducible, and systematic way of controlling the fixed-node error in DMC.

[1] Here, by uncontrolled it is meant that there does not exist a systematic way of arbitrarily reducing the fixed-node error.

[2] C. J. Umrigar, J. Toulouse, C. Filippi, S. Sorella, and R. G. Hennig, Phys. Rev. Lett. **98**, 110201 (2007).

[3] K. E. Schmidt and J. W. Moskowitz, J. Chem. Phys. **93**, 4172 (1990).

[4] E. Giner, A. Scemama, and M. Caffarel, Can. J. Chem. **91**, 879 (2013).

[5] A. Scemama, T. Applencourt, E. Giner, and M. Caffarel, J. Chem. Phys. **141**, 244110 (2014).

[6] E. Giner, A. Scemama, and M. Caffarel, J. Chem. Phys. **142**, 044115 (2015).

[7] B. Huron, J. P. Malrieu, and P. Rancurel, J. Chem. Phys. **58**, 5745 (1973).

[8] S. Evangelisti, J. P. Daudey, and J. P. Malrieu, Chem. Phys. **75**, 91 (1983).

[9] C. F. Bender and E. R. Davidson, Phys. Rev. **183**, 23 (1969).

[10] R. J. Buenker and S. D. Peyerimhoff, Theor. Chim. Acta **35**, 33 (1974).

[11] R. J. Buenker and S. D. Peyerimhoff, Theor. Chim. Acta **39**, 217 (1975).

[12] R. J. Buenker, S. D. Peyerimhoff, and W. Butscher, Mol. Phys. **35**, 771 (1978).

[13] P. J. Bruna, D. S. Peyerimhoff, and R. J. Buenker, Chem. Phys. Lett. **72**, 278 (1980).

[14] R. J. Buenker, S. D. Peyerimhoff, and P. J. Bruna, *Computational Theoretical Organic Chemistry* (Reidel, Dordrecht, 1981), p. 55.

[15] R. J. Harrison, J. Chem. Phys. **94**, 5021 (1991).

[16] M. Caffarel, E. Giner, A. Scemama, and A. Ramírez-Solís, J. Chem. Theory Comput. **10**, 5286 (2014).

[17] G. H. Booth, A. J. W. Thom, and A. Alavi, J. Chem. Phys. **131**, 054106 (2009).

[18] D. Cleland, G. H. Booth, and A. Alavi, J. Chem. Phys. **132**, 041103 (2010).

[19] P. K. V. V. Nukala and P. R. C. Kent, J. Chem. Phys. **130**, 204105 (2009).

[20] B. K. Clark, M. A. Morales, J. McMinis, J. Kim, and G. E. Scuseria, J. Chem. Phys. **135**, 244105 (2011).

[21] G. L. Weerasinghe, P. López-Ríos, and R. J. Needs, Phys. Rev. E **89**, 023304 (2014).

[22] A. Scemama, T. Applencourt, E. Giner, and M. Caffarel, "Quantum Monte Carlo with very large multideterminant wavefunctions," J. Comput. Chem. (in press).

[23] An increase in the fixed-node energy may be sometimes observed at small number of determinants, (say, less than a few thousands) when large basis sets and/or canonical orbitals are used. This transient behavior has been found to systematically disappear when natural orbitals are used and/or larger expansions are considered.

[24] H. J. Flad, M. Caffarel, and A. Savin, *Recent Advances in Quantum Monte Carlo Methods* (World Scientific Publishing, 1997).

[25] W. Klopper, Mol. Phys. **99**, 481 (2001).

[26] A. Scemama, E. Giner, T. Applencourt, G. David, and M. Caffarel (2015). "Quantum Package," Zenodo. http://dx.doi.org/10.5281/zenodo.30304.

[27] C. X. Almora-Díaz, J. Chem. Phys. **140**, 184302 (2014).

[28] A. Scemama, E. Giner, T. Applencourt, and M. Caffarel, Qmc=chem, https://github.com/scemama/qmcchem.

[29] T. Helgaker, W. Klopper, H. Koch, and J. Noga, J. Chem. Phys. **106**, 9639 (1997).

[30] H. Muller and W. Kutzelnigg, Mol. Phys. **92**, 535 (1997).

[31] A. Halkier, T. Helgaker, P. Jorgensen, W. Klopper, H. Koh, J. Olsen, and A. K. Wilson, Chem. Phys. Lett. **286**, 243 (1998).

[32] L. Bytautas and K. Ruedenberg, J. Chem. Phys. **124**, 174304 (2006).