RESERVE THIS SPACE

Practical diffusion Monte Carlo simulations for large noncovalent systems

Kenta Hongo[1,*] and Ryo Maezono[1]

[1]School of Information Science, Japan Advanced Institute of Science and Technology (JAIST), Asahidai 1-1, Nomi, Ishikawa 923-1292, Japan

Fixed-node diffusion Monte Carlo (FNDMC) simulations are one of the most promising methods for describing the noncovalent systems to high accuracy within reasonable computational times. The advent of massively parallel computers enables one to apply FNDMC to various noncovalent systems such as supramolecules and molecular crystals. It is, however, to be noted that a reliable description of subtle noncovalent interactions requires a much higher accuracy than that of typical chemical bindings, e.g., the subchemical accuracy of 0.1 kcal/mol for small noncovalent complexes. This is a severe requirement for FNDMC based on stochastic approaches and raises the computational issues of reliable estimates of not only error bar, but also energy itself. Firstly, our recent works on several noncovalent systems are demonstrated. Then we address the issues and propose a new strategy for statistical estimates to meet the subchemical accuracy.

Introduction

Fixed-node diffusion Monte Carlo (FNDMC) is one of the most promising approaches to noncovalent systems among state-of-the-art *ab initio* simulations (*1,2*). Its accuracy is often compatible with the "gold standard" quantum chemistry, CCSD(T)/CBS (coupled cluster including single, double, and

RESERVE THIS SPACE

noniterative triple excitations with complete basis set). This can be demonstrated by typical benchmark systems such as water (*3–6)* and benzene (*3,7,8*) dimers as well as systematic benchmarks with the S22 (*8–10*) and A24 (*11*) data sets. Moreover, it has been expected to be applicable to larger systems, because of its moderate computational scaling of $N^{3\text{-}4}$ (N stands for the number of electrons in system), though FNDMC is a computationally intensive technique requiring a vast number of statistical accumulations (*stats*). Actually, recent advents of massively parallel computers enable FNDMC to treat larger and more realistic systems including bulk solid/liquid water (*12–14*), graphite (*15*) and graphene (*16*) layers, molecular crystals (*17–19*), biomolecules (*20,21*), host-guest complexes (*22,23*), and so on. In addition, FNDMC can be regarded as a useful tool of investigating industrial issues involving microscopic information about molecular interactions. For instance, FNDMC has very recently been used to evaluate Hamaker's constants, which is closely related to control of wettability in solution processes (*24*).

Despite its success, a qualitatively accurate evaluation of molecular interactions is still a serious challenge even for FNDMC. This arises from the fact that molecular interactions are generally complicated combinations of different types of interactions with different energy scales such as hydrogen bonding and dispersion (typically 0.5 ~ 30 kcal/mol). Hence their reliable description requires a much higher accuracy than typical chemical bindings such as covalent, ionic, and metallic bonds. For instance, "subchemical accuracy" of 0.1 kcal/mol is necessary for small complexes (*1,2*), though in larger systems the desired maximum error grows proportionally. This is quite crucial in accurately estimating energies as well as their statistical errors. The latter is always taken into account in FNDMC simulations, while the former is not well recognized, which we shall mainly discuss in this chapter. To perform larger FNDMC simulations than ever, the following two issues are to be addressed: (i) efficient generation of vast number of sampling data points N_t and (ii) energy sampling scheme with less biases.

(i) The subchemical accuracy explicitly indicates a desirable error bar $\sigma \sim$ 0.1 kcal/mol can be obtained from a hundred times or more sampling points N_t than the case of chemical accuracy (1 kcal/mol), following $\sigma^2 \sim N_t$. In FNDMC simulations, the number of total sampling points N_t consists of the Monte Carlo time steps N_s times the random walker populations N_c ($N_t = N_s \times N_c$). In practice, their values are set according to computational resources available. For instance, it has been reported that the CASINO code of FNDMC implementation (*25*) exhibits more than 99 % parallel efficiency on the K computer (*26*) using a "flat MPI" parallelization with fifty thousands cores (*27*). Here the term "flat MPI" means that instead of OpenMP, MPI is used for parallelization between cores of a CPU. In FNDMC simulations with the flat MPI parallelization, configurations (random walkers) are distributed among all the cores communicating with each other via MPI. This communication might give rise to a significant latency in

parallel computation, but the CASINO implementation circumvents this deficiency by using several techniques such as asynchronous communication (*25*). Therefore, without loss of efficiency, FNDMC can treat more N_c in parallel as the number of available cores N_{core} increases. For a fixed σ, the larger N_c means the smaller N_s, leading to a significant saving of computational time, because the time is simply proportional to N_s. As described later, however, it is to be noted that the minimum value exists in choosing N_s so as to get a reliable σ value. In other words, the naive use of the more degree of parallelization does not necessarily imply the more efficient FNDMC simulations, even though more resources would be available with the advent of the next-generation supercomputers.

(ii) So far, FNDMC has been applied mainly to covalent or metallic systems involving the chemical accuracy at the most. For noncovalent systems, however, the subchemical accuracy of 0.1 kcal/mol is crucial for the FNDMC energy estimator. This stems from that actual FNDMC simulations on computer do not realize "true" stochastic processes and may be biased. Although this has not been well recognized so far, for instance, it was reported that FNDMC estimates were biased due to a poor performance of (pseudo) random number generators and/or their seeds adopted (*28,29*). Ideally, bias-free sampling schemes would be needed to capture the very subtle interactions to the subchemical accuracy. In this chapter, we propose a simple sampling scheme useful for both (i) and (ii).

In what follows, we first demonstrate the FNDMC performance of describing several types of noncovalent systems, referring to our recent works including molecular crystals, biomolecules, and precursor molecules in a liquid process. Next, we discuss computational issues arising from the above applications. In particular, we propose a simple but useful strategy appropriate for achieving the subchemical accuracy. Finally, we summarize and make remarks on future FNDMC simulations of large-scale noncovalent systems.

FNDMC applied to noncovalent systems

Cyclohexasilane dimer

Despite its potential, much less has been reported so far on FNDMC applications to an industrially huge demand for better understanding and then controlling the interactions to open up novel technologies. Here we demonstrate our FNDMC framework of evaluating Hamaker (or van der Waals) constants of molecules *a prior*, which are obtained from their long-range asymptotic behaviors of binding curves (*24*). The Hamaker constant plays a crucial role in controlling wettability in liquid processes, but no *ab initio* approach has been developed because a handy but reliable energy solver is required. The FNDMC framework would extensively provide a reliable value of Hamaker constant even

for unknown molecules with no reference data, satisfying the industrial demand. We applied the framework to a practical size of liquid molecule, cyclohexasilane (Si_6H_{12}). Si_6H_{12} is used as a precursor ink to fabricate amorphous silicon semiconductors by using a liquid process (*30*). The liquid process may significantly save running costs and material resources, but currently relies on control of wettability by trial and error. The wettability strongly affects a product quality and can be described in terms of Hamaker constants. Hence development of its reliable evaluation scheme is a first step towards a quantitative control of wettability from microscopic viewpoints.

To accurately evaluate the Hamaker constant for Si_6H_{12}, we attempted to apply the CCSD(T)/cc-pVTZ level to the dimer using the Gaussian09 code (*31*), but it is too large to be tractable even using our Altix UV1000 supercomputer with 512GB memory shared by 64 parallel cores; CCSD(T)/cc-pVDZ was the most accurate method available for the dimer. Therefore, instead of adopting a usual CBS extrapolation with cc-pVDZ and cc-pVTZ, the CCSD(T)/CBS estimate was obtained from a semi-empirical treatment combining CCSD(T)/cc-VDZ with MP2/CBS and MP2/cc-pVDZ (*32*). This has been known to work well for π-π interaction systems. It is not obvious, however, whether or not it also works for the σ-σ interaction in the Si_6H_{12} dimer because little attention to such interactions has been paid so far. In this work, the Hamaker constant is evaluated from a coefficient of R^{-6} term in 6-12 Lennard-Jones potentials, to which *ab initio* binding curves are fit. Consequently, the binding curve should be as accurate as possible not only at its equilibrium, but also at its long-range distances. Relying on FNDMC, we calibrated the performance of the "semi-empirical" CCSD(T)/CBS as well as various DFT approaches. Our FNDMC simulations were done using the CASINO code (*25*), employing B3LYP single Slater determinants as the fixed nodes (denoted FNDMC/B3LYP). As has been reported previously (*33-36*), FNDMC/B3LYP gave the better variational energies than the LDA and PBE nodes. Our FNDMC and CCSD(T)/CBS binding curves are in good agreement with each other, leading to the Hamaker constants of 104 ± 4 [zJ] and 99 [zJ], respectively. From their computational viewpoints, however, FNDMC is more advantageous over CCSD(T). In particular, the FNDMC accuracy generally shows less dependence on basis sets, compared with CCSD(T). In other words, FNDMC converges to its "exact" solution much faster than CCSD(T) with respect to basis sets (*1*).

We also benchmarked the performance of several exchange-correlation (XC) functionals developed recently for describing van der Waals interactions. It is remarkable that M06-2X/B97-D gives good/poor equilibrium properties, but poor/good long-range behaviors. Within the DFT framework, B3LYP-GD3 revealed a fairly good performance on both equilibrium and long-range properties. At short range, however, any XC functionals significantly deviate from FNDMC and CCSD(T)/CBS. This may be attributed to the fact that most DFT approaches cannot generally remove their self-interactions, giving rise to a

poor description of the exchange repulsion there. Although the attraction of DFT is its cost, its accuracy at long range also strongly depends on its functional. There is no *a priori* knowledge of selecting XC functionals appropriate for a certain problem. In this sense, FNDMC is much more useful even for this kind of industrial applications. Finally, we comment on accuracy of our Hamaker constants. Since no reference value was available for the constant, its experimental value was estimated from a linear regression between the constant and the corresponding molecular weight, being reasonably comparable with our numerical results.

B-DNA

In contrast to the above case, biomolecular systems are more challenging for state-of-the-art *ab initio* simulations in both theoretical and computational sense. FNDMC was first applied to the stacked and Watson-Crick bound adenine/thymine (A/T) and cytosine/guanine (C/G) DNA base pair dimers (*8*). Generally, biomolecular structures are preserved by noncovalent interactions among not only their basic building blocks, but also their surroundings. Such interactions drastically change depending on molecules involved. To better understand how DNA stabilizes its structure, the more realistic DNA modeling is Watson-Crick base-pair steps in DNA as shown in Figure 1 (a). We applied FNDMC/B3LYP to the Adenine-Thymine base-pair step (AA:TT) (*20*). Similar to the Si_6H_{12} case, the B3LYP node was variationally better than the other two nodes, i.e., LDA and GGA-PBE. It was found that our FNDMC stacking energy reasonably agrees with the CCSD(T)/CBS one. We also benchmarked various DFT functionals. It is remarkable that some recent XC functionals (CAM-B3LYP/LC-ωPBE) predicted it to be unbound (see Figure 1).

Very recently, we have evaluated the stacking interactions of ten unique DNA base-pair steps in B-DNA (see Figure 1 (a)) using the same FNDMC procedure as our previous study. (*20*) Figure 1 (c) plots the stacking energies for the ten cases evaluated from B3LYP, CCSD(T)/CBS, and FNDMC/B3LYP. B3LYP does not reproduce the stacking, while FNDMC/B3LYP does, even though starting with such a poor wave function. FNDMC gives an overall trend similar to CCSD(T)/CBS, but quantitatively deviates more than 1 kcal/mol from CCSD(T)/CBS for GA:TC and AG:CT. Since experimental values are not available, it is impossible to tell which is more reliable. Here we just make remarks on their methodological issues. Similar to the Si_6H_{12} case, the CCSD(T)/CBS results were obtained using both the BSSE and CBS corrections, but the 6-31G** basis sets (less reliable than cc-pVDZ) were adopted due to the cost (*37*). In contrast, our FNDMC/B3LYP with the VTZ basis set is found to be accurate enough to describe the binding curve, comparing the VTZ and VDZ levels. In this sense, FNDMC has the more advantage over the other correlated methods in quantum chemistry.

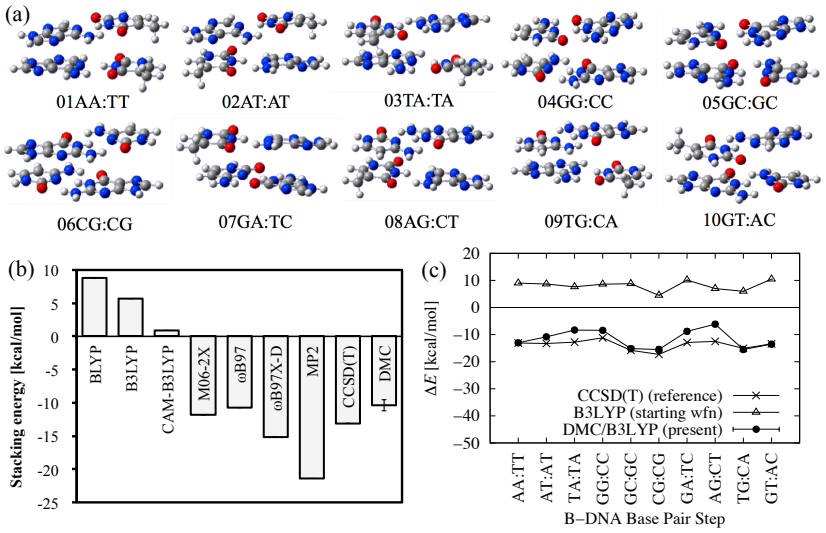

Figure 1. (a) Ten unique Watson-Crick base-pair steps in B-DNA. (b) Stacking energy of AA:TT evaluated from various methods (20). (c) Stacking energies of all the ten cases evaluated from B3LYP/VTZ and FNDMC with the B3LYP/VTZ nodes. For comparison, the corresponding CCSD(T)/CBS (37) results are also plotted. All energies are given in units of kcal/mol.

Molecular crystal polymorphism

Molecular crystal polymorphism is one of the most important issues in theoretical and applied chemistry, and has been investigated by various *ab initio* simulations (*38*). FNDMC had not been an appropriate choice of treating this issue due to its intensive cost, and hence its applicability was limited to benchmark cases such as some ice polymorphs (*39*). Using FNDMC, we have for the first time attempted to investigate the polymorphism of para-diiodobenzene (*p*-DIB), which requires more accuracy and cost because of their π-π stacking interactions involving dispersion (*17-19*).

Our p-DIB study in 2010 (*17*) could conduct only a 1×1×1 simulation cell because only a 128-core machine was available. We thus restricted ourselves to use of semi-empirical scheme due to Kwee, Zhang, and Krakauer (KZK) (*40*) in order to investigate the finite size effects (FSEs) in the *p*-DIB polymorphs. Unlike standard DFT approaches, our FNDMC approach predicted the *p*-DIB

polymorphism correctly. It was not evident, however, that the KZK correction scheme is appropriate for strongly anisotropic systems like *p*-DIB because its parameterization was obtained from isotropic systems using LDA (*40,41*). In order to address this issue, with the help of the K computer, we performed FNDMC simulations with a 1×3×3 simulation cell (1,512 electrons), which was the largest and most expensive simulations, consuming 6.4×10^5 core hours for each polymorph (*19*). We studied in detail the FSEs analyzing several correction schemes, and found the 1×3×3 simulation cell still gives rise to a large error in the total energies for each of the two polymorphs, but a significant error cancellation between them leads to the correct prediction.

For comparison, we also investigated the FSEs within the DFT framework. It was found that 1×3×3 deviates by 0.1 kcal/mol from 2×6×6 that converges to 4×4×12 within 0.01 kcal/mol. This may be regarded as a good measure of the one-body contribution to the FSEs, though not the two-body one. We therefore considered several possible two-body corrections in FNDMC and found their differences between the two polymorphs are less than the total energy difference (i.e., relative stability). To calibrate the two-body contribution more precisely, however, we should have directly dealt with the 2×6×6 cell size at least. This simulation cell includes 12,096 electrons and hence requires 512 (= 8^3) times more cost than the 1×3×3 one (1,512 electrons). Hence its cost would be estimated to be 3.3×10^8 core hours. Realistically, this is intractable for current petascale supercomputers, and can be still challenging even for exascale supercomputers. We shall discuss computational issues in more detail in the next section.

Computational issues

Computational costs

Table I summarizes computational conditions and costs to achieve the chemical accuracy for the above-mentioned systems. Moderate amounts of computational costs and resources were necessary for the isolated molecular systems, while huge amounts of them for the molecular crystal system. Note that, unlike the former, the latter suffers from twice longer queuing time (14 days) than the execution time (7 days), which was submitted to a SMALL job class with N_{core} = 2,048 parallel cores on the K computer. This queuing time arose from that we chopped up a statistical accumulation (*stats*) job with N_s = 6,500 and N_c = 20,480 into 13 sequential *stats* jobs with $N_s/13$ = 500 and N_c = 20,480, each of which is executable within the CPU time limit of 12 hours for the job class on the K computer. This indicates that prior to running large FNDMC simulations on supercomputers having a large number of batch queues, it is very

important in practice to consider their feasibility according to the actual status of machine utilization.

Table I. Computational conditions and costs for single-point FNDMC simulations to achieve the chemical accuracy. In "Timing" column, the queuing time is negligible and hence not shown for the isolated molecular systems, while that is shown in parenthesis for the *p*-DIB with the 1×3×3 simulation cell.

System	N_{elec}	Machine	N_{core}	Core-hour	Timing (real)
$(Si_6H_{12})_2$	72	Altix UV1000	32	1.0×10^2	3 hours
B-DNA	196	Fujitsu CX250	320	1.5×10^4	2 days
p-DIB	1,512	K computer	2048	6.4×10^5	7 days (+14 days)

A simple way of circumventing large amounts of queuing times is just to remarkably increase N_c and reduce N_s, if more computational resources (i.e., N_{core}) are available. Based on our *p*-DIB simulations using the flat MPI parallelization with N_{core} = 2,048, N_c = 20,480, and N_s = 6,500, we may estimate the cost-performance of the N_{core} = 512,000 case on getting the same $N_t = N_c \times N_s$, where N_{core} = 512,000 is almost the largest available on the K computer. Assuming an ideal MPI scaling holds even for N_{core} = 512,000, as reported in Ref. (*27*), we could set N_c = 5,120,000 and N_s = 26, without deceleration at each Monte Carlo step. Thus we could achieve a speedup of 250 times at *stats*, as shown in Figure 2 (a). It is to be noted that this holds only for *stats*, not for the equilibrium procedure (*equil*), if we rely only on the flat MPI parallelization. A converged DMC distribution is achieved by propagating the configurations in a fixed period (the number of equil steps, N_{eq}), which is just determined by how far the initial distribution is from the converged one. This means that N_{eq} is incapable of being decreased by any parallelization. In order to accelerate the *equil* procedure, computations in the propagation should be accelerated. In the flat MPI parallelization, however, all the cores are devoted to the parallelization for configurations, and cannot afford to accelerate any computation in the propagation. Hence the flat MPI parallelization is not able to achieve any speedup at *equil*. Here we consider a computational efficiency in terms of the total CPU times including both *equil* and *stats*. Figure 2 (b) highlights a comparison of the cost between N_{core} = 2,048 and 512,000. In this case, the effective speedup would be only 5 times in spite of exploiting 250 times more resources.

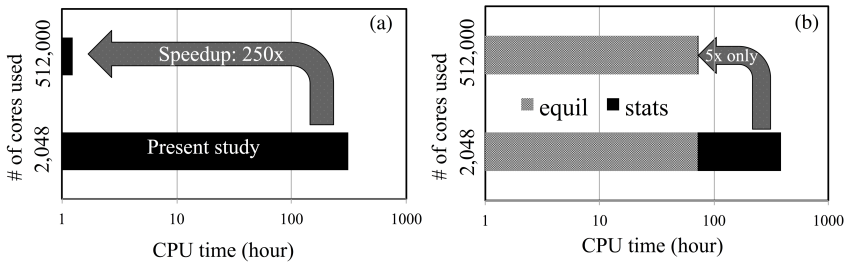

Figure 2. Computational times (CPU-time) for (a) stats and (b) the whole process including equil and stats. In each case, the time for N_{core} = 2,048 *is actually observed in the FNDMC simulations of p-DIB on the K computer, while the one for* N_{core} = 512,000 *is estimated using an extrapolation from* N_{core} = 2,048 *assuming an ideal scaling in parallel efficiency is valid up to* N_{core} = 512,000. *The corresponding speedups are also shown in each case.*

Furthermore, we may claim that it is not necessarily beneficial to use much larger supercomputers only with a flat MPI parallelization for larger systems. To see this, consider the core hours as a figure of merit for the cost-performance in parallel computing. Figure 3 indicates that a percentage of *equil* to total core hours is 18.75% for N_{core} = 2,048, while that is 98.34% for N_{core} = 512,000. Since the computational cost consumed at *equil* is not used at all to obtain the final result, almost all the cost would be in vain for N_{core} = 512,000.

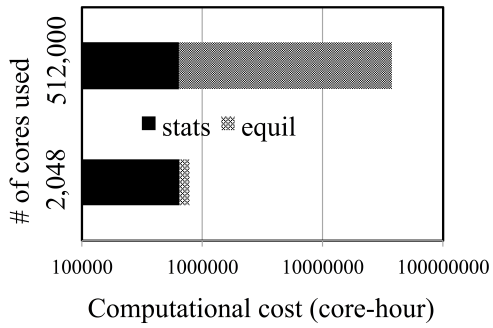

Figure 3. Total computational costs including equil and stats in terms of core hours for N_{core} *= 2,048 (observed) and* N_{core} *= 512,000 (estimated).*

The above consumption at *equil* occurs when using only a flat MPI parallelization because each of the cores is allocated to the task of driving a set of walkers in the FNDMC simulations. A hybrid parallelization with MPI and OpenMP could be one of the possible solutions, but it is not obvious if it would be absolutely in success. Instead, GPGPU would be one of the most promising, because an *equil* procedure does not require numerical precision. Indeed, such implementation has been reported to achieve a remarkable speedup in the FNDMC simulations (*42,43*). The procedure is as follows: First, one performs a preliminary FNDMC simulation with smaller values of N_c and N_s, compared with the corresponding production run. After warming up the walkers and executing *stats*, one record the walkers at each step. The collected population according to the equilibrium distribution is used as an initial one for the production run and then allocated into each of the cores.

Reblocking for an accurate estimate of error bar

Except the above problem, the use of massively parallel computers could be regarded as useful for large-scale FNDMC simulations. This can be mostly true, but there still exist issues to be addressed in order to achieve the subchemical accuracy for both an energy E and its uncertainty (error bar σ). We begin with the issue of σ, and discuss E in the next subsection. A usual FNDMC simulation adopts the so-called reblocking technique to estimate σ properly (*44,45*). This is because sequentially sampled data points generally have a serial correlation with a positive length (e.g. τ_{corr}) and hence a naive estimate of error bar, σ_{raw}, is usually underestimated compared with a true σ. According to statistics, $\sigma^2 = \sigma^2_{\mathrm{raw}} (1 + \tau_{\mathrm{corr}})$, where τ_{corr} (called integrated correlation length) is given as a total sum of autocorrelation functions over all intervals. Although σ can be evaluated from the above relation, the reblocking scheme is used in practice, which is simpler and more convenient than the above relation. The reblocking procedure is as follows: (i) it divides the raw data into contiguous blocks of length B, (ii) averaging the data over each block to generate a new data set (blocked data set), (iii) evaluating a naive error bar σ_B from the blocked data set, (iv) changing B values, plot $\log_2(B)$ versus σ_B ("reblocking plot") and find the peak or plateau to give a desirable σ ($\sigma_{\mathrm{reblock}}$). This is computationally advantageous to on the fly monitor the current σ during simulations. In particular, the CASINO suite of program codes (*4*) has a more sophisticated implementation to evaluate $\sigma_{\mathrm{reblock}}$ without plotting and searching the peak or plateau (*45*).

Here we compare the above two estimates of σ for the p-DIB case with N_s = 6,500, N_c = 20,480, and $\delta\tau$ = 0.001. For those data points, we obtained σ_{raw} = 0.0876 ± 0.007 and τ_{corr} = 109 ± 47, arriving at σ = 0.92 ± 0.20. The reblocking plot in Figure 4 finds the peak appearing at $B = 2^9 = 512$, giving $\sigma_{\mathrm{reblock}}$ = 0.94 ± 0.19. We thus found that the reblocking is practically equivalent to the above relation and each of blocked data sets with B = 512 is statistically independent.

Note that N_s should be much larger than τ_{corr} to properly evaluate σ using the reblocking. In other words, there exists a minimum value of N_s for a reliable σ evaluation by the reblocking.

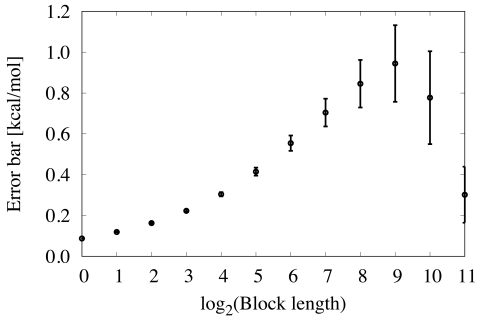

Figure 4. Plot of block length versus error bars evaluated from reblocked data points for the p-DIB case (see text for detailed computational conditions).

Next, we investigate τ_{corr} in more detail, which relates to a choice of N_s. Our production run for *p*-DIB gave $\tau_{corr} = 107 \pm 47$ for $N_c = 20{,}480$ and $N_s = 6{,}500$, while a preliminary one gave $\tau_{corr} = 104 \pm 13$ for $N_c = 1{,}280$ and $N_s = 74{,}000$, both of which achieved the chemical accuracy. Therefore, $\tau_{corr} \approx 100 = 1/\delta\tau$ holds for both the cases, though the latter has a smaller error bar of correlation length because of its larger N_s. Consequently, we may expect the $N_c = 5{,}120{,}000$ case to have almost the same τ_{corr}. Recalling our choice of $N_s = 26$ for $N_c = 5{,}120{,}000$, $N_s < \tau_{corr}$ holds. This indicates that the use of $N_{core} = 512{,}000$ MPI parallelization could achieve the 250 times speedup, but its choice of N_s is too short to properly estimate σ by reblocking. The above is for our case study, but should be kept in mind whenever using supercomputers.

A new sampling strategy

FNDMC evaluates E and σ from random sampling, according to the law of large numbers and the central limit theorem (*46*), respectively. Occasionally, biases in E may arise from an artifact of finite sampling. Actually, it has been reported that some combinations of random numbers and systems give rise to biases more than the chemical accuracy (*28,29*). Obviously, such biases do not satisfy the requirement of the subchemical accuracy, implying that one might draw an incorrect conclusion from such biased E values.

We propose a simple but robust scheme against such biases based on the reproductive property of the normal distribution (*46*). FNDMC usually generates

data points from a single time-series job. Assume the total data points M_{tot} in the single job give the estimates of E and σ. We start by chopping the single job with M_{tot} into statistically independent N_p jobs with $N_s = M_{\mathrm{tot}}/N_p$ steps. We then individually perform the N_p jobs with N_s using different random seeds, thereby generating a data set of $\{E_i, \sigma_i ; i = 1, \cdots, N_p\}$. Finally, we average them following:

$$E_{\mathrm{ave}} = \frac{1}{N_p}\sum_{i=1}^{N_p} E_i \quad \text{and} \quad \sigma_{\mathrm{ave}} = \frac{1}{N_p}\sqrt{\sum_{i=1}^{N_p}\sigma_i^2} \ , \tag{1}$$

where $E_{\mathrm{ave}} = E$ and $\sigma_{\mathrm{ave}} = \sigma$ can be derived from the reproductive property of the normal distribution (*46*). We may call this scheme "*chopped-up stats and job-averaging*" scheme. Note that the scheme assumes the distribution of the data set to be the normal one. This assumption is, however, invalid in some cases (*47*) where the data sets follow stable ones, and more consideration is needed.

We demonstrate the above scheme by adopting the DMC simulation of the ground-state He atom. This was chosen because: (i) The "exact" E of -2.903724 Hartree is available (*48*); (ii) Its wave function is free of the fixed-node bias because it has no nodes; (iii) The time-step bias is negligibly small (less than 0.01 mHartree at $\delta\tau = 0.001$). We first perform $N_p = 16$ jobs with $N_s = 250$ as shown in Figure 5 (a). According to Eq. (1), we average them to get E_{ave} and σ_{ave}, which is denoted 'ave' in Figure 5 (b). It is found E_{ave} is in good agreement with the exact energy within σ_{ave}. For comparison, we also run 16 single jobs with $M_{\mathrm{tot}} = N_s = 4{,}000$ steps, resuming *stats* from the above 16 jobs with $N_s = 250$, as shown in Figure 5 (b). In each figure, N_s is common to each of the single jobs, and one differs from the other only in random seed. We found that artifacts exist in the single job estimates of E and σ depending on their seeds. Let us consider the artifacts in more detail below.

(i) σ: Looking at Figure 5 (a), it seems that seed Nos. 4, 6, and 14 have much larger σ than the other, while seed No. 10 has a smaller σ. Increasing *stats* steps as in Figure 5 (b), such an artifact disappears for Nos. 4 and 6, but still remains for No. 14. During *stats* simulations, one sometimes monitors current σ values to estimate how many steps are additionally needed to achieve a desired accuracy, which is useful for computation plan. If one grabs such an unreliable σ, it is useless to expect the completion time. This calls attention to the risk of using a single job scheme only. On the other hand, Figure 5 (b) shows the job-averaging scheme has a reasonable σ value, comparable with most of the single job values.

(ii) E: In addition to σ, a finite sampling of the single job causes a significant bias larger than the subchemical accuracy for many cases, while the job-averaging gives quite a good estimate of the exact energy. We found biases at Nos. 1, 5, 9, 12, 13, and 14 appear in Figure 5 (a), but disappear in Figure 5 (b). Note that a new bias at No. 3 in Figure 5 (b) emerges, regardless of increasing the steps. This implies that this kind of bias accidentally occurs when

relying on the single job because of the finite sampling. In contrast, the statistically independent averaging can weaken biases, leading to more reliable estimates.

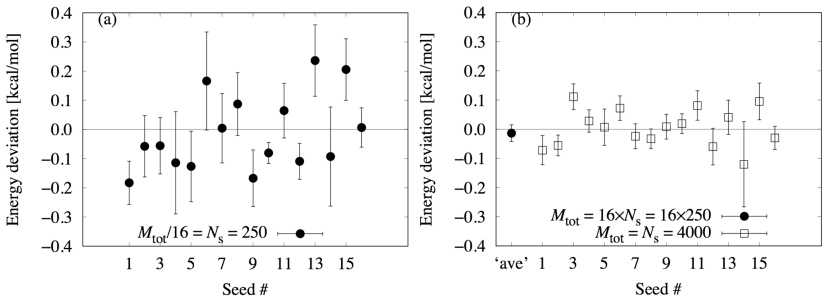

Figure 5. Energy deviations from the exact He energy for (a) N_p = 16 jobs with N_s = 250 using different random seeds, and (b) job-averaging scheme (denoted 'ave') and 16 jobs with M_{tot} = 4,000 (see text in detail). Energies are in units of kcal/mol.

Here we verified the "*chopped-up stats and job-averaging*" scheme provides more reliable estimates of E and σ than the single job scheme. We can conclude that it well satisfies the requirement of the subchemical accuracy, which is crucial for describing the noncovalent interactions. Furthermore, our job-averaging scheme would enable one to carry out large-scale FNDMC simulations without using supercomputers, but with using a large number of small/middle class computers available everywhere. In practice, the latter is more favorable than the former in the sense that the former may impose much more queuing time on users than the latter. Besides that, the queuing time randomly varies and hence it is difficult to prediction when *stats* completes.

Concluding Remarks

We demonstrate our recent FNDMC works on successfully describing noncovalent interactions in both isolated and periodic molecular systems. Evidently, FNDMC exhibits a high ability of tackling noncovalent problems. Nevertheless, a major challenge in FNDMC still lies in describing the noncovalent interactions to the subchemical accuracy of 0.1 kcal/mol even when using state-of-the-art massively parallel computers. That is, the subchemical accuracy requires numerically reliable estimates of not only error bar σ, but also energy E itself.

To achieve σ within 0.1 kcal/mol, the hundred times more sampling points are needed, compared with the requirement of the chemical accuracy of 1.0 kcal/mol. Ideally, exascale supercomputers could solve this issue, but practically, one should keep in mind the following two points: (i) A flat MPI parallelization does not accelerate the *equil* procedure, just giving rise to a huge consumption of computational resources in terms of core hours. (ii) *Stats* steps N_s should be large enough to estimate a reliable σ by using the reblocking technique. The other acceleration techniques such as OpenMP and GPGPU might be useful for circumventing these issues.

Occasionally, a finite sampling gives rise to a biased energy estimate, depending on a system and pseudo random numbers. Although this has not been recognized seriously so far, within the energy scale of the subchemical accuracy, the biased E could make a misprediction theoretically. To avoid this issue, we propose a new scheme named "*chopped-up stats and job-averaging*", where the serial *stats* with long steps is equally divided into a number of statistically independent *stats* with shorter steps and their estimates of E and σ are averaged, based on the reproductive property of the normal distribution. We found our averaging scheme is more stable than the single one. Furthermore, this is applicable to large-scale FNDMC simulations, without using supercomputers, but using a bunch of small/middle-class computers.

Acknowledgements

K.H. is grateful for financial support from a KAKENHI grant (15K21023). The authors also acknowledge the support by the Computational Materials Science Initiative (CMSI/Japan) for the computational resources, Project Nos. hp120086, hp140150, hp150014 at the K computer, and SR16000 (Center for Computational Materials Science of the Institute for Materials Research, Tohoku University/Japan). R.M. is grateful for financial support from MEXT-KAKENHI grants 26287063, 25600156, 22104011 and that from the Asahi glass Foundation. The computation in this work has been partially performed using the facilities of the Center for Information Science in JAIST.

References

1. Dubecký, M.; Mitas, L.; Jurečka, P. Noncovalent Interactions by Quantum Monte Carlo. *Chem. Rev.* **2016**, *116*, 5188-5215.
2. Řezáč, J.; Hobza, P. Benchmark Calculations of Interaction Energies in Noncovalent Complexes and Their Applications. *Chem. Rev.* **2016**, *116*, 5038-5071.

3. Diedrich, C.; Lüchow, A.; Grimme, S. Weak intermolecular interactions calculated with diffusion Monte Carlo. *J. Chem. Phys.* **2005**, *123*, 184106.
4. Benedek, N. A.; Snook, I. K.; Towler, M. D.; Needs, R. J. Quantum Monte Carlo calculations of the dissociation energy of the water dimer. *J. Chem. Phys.* **2006**, *125*, 104302.
5. Gurtubay, I. G.; Needs, R. J. Dissociation energy of the water dimer from quantum Monte Carlo calculations. *J. Chem. Phys.* **2007**, *127*, 124306.
6. Sterpone, F.; Spanu, L.; Ferraro, L.; Sorella, S.; Guidoni, L. Dissecting the Hydrogen Bond: A Quantum Monte Carlo Approach. *J. Chem. Theory Comput.* **2008**, *4*, 1428-1434.
7. Sorella, S.; Casula, M.; Rocca, D. Weak binding between two aromatic rings: Feeling the van der Waals attraction by quantum Monte Carlo methods. *J. Chem. Phys.* **2007**, *127*, 014105.
8. Korth, M.; Lüchow, A.; Grimme, S. Toward the Exact Solution of the Electronic Schrodinger Equation for Noncovalent Molecular Interactions: Worldwide Distributed Quantum Monte Carlo Calculations. *J. Phys. Chem. A* **2008**, *112*, 2104–2109.
9. Dubecký, M.; Jurečka, P.; Derian, R.; Hobza, P.; Otyepka, M.; Mitas, L. Quantum Monte Carlo Methods Describe Noncovalent Interactions with Subchemical Accuracy. *J. Chem. Theory Comput.* **2013**, *9*, 4287-4292.
10. Dubecký, M.; Derian, R.; Jurečka, P.; Mitas, L.; Hobza, P.; Otyepka, M. Quantum Monte Carlo for noncovalent interactions: an efficient protocol attaining benchmark accuracy. *Phys. Chem. Chem. Phys.* **2014**, *16*, 20915-20923.
11. Řezáč, J.; Dubecký, M.; Jurecka, P.; Hobza, P. Extensions and applications of the A24 data set of accurate interaction energies. *Phys. Chem. Chem. Phys.* **2015**, *17*, 19268-19277.
12. Raza, Z.; Alfè, D.; Salzmann, C. G.; Klimes, J.; Michaelides, A.; Slater, B. Proton ordering in cubic ice and hexagonal ice; a potential new ice phase-XIc. *Phys. Chem. Chem. Phys.* **2011**, *13*, 19788-19795.
13. Alfè, D.; Bartók, A. P.; Csányi, G.; Gillan, M. J. Analyzing the errors of DFT approximations for compressed water systems. *J. Chem. Phys.* **2014**, *141*, 014104.
14. Morales, M. A.; Gergely, J. R.; McMinis, J.; McMahon, J. M.; Kim, J.; Ceperley, D. M. Quantum Monte Carlo Benchmark of Exchange-Correlation Functionals for Bulk Water. *J. Chem. Theory Comput.* **2014**, *10*, 2355-2362.
15. Spanu, L.; Sorella, S.; Galli, G. Nature and Strength of Interlayer Binding in Graphite. *Phys. Rev. Lett.* **2009**, *103*, 196401.

16. Mostaani, E.; Drummond, N. D.; Fal'ko, V. I. Quantum Monte Carlo Calculation of the Binding Energy of Bilayer Graphene. *Phys. Rev. Lett.* **2015**, *115*, 115501.
17. Hongo, K.; Watson, M. A.; Sánchez-Carrera, R. S.; Iitaka, T.; Aspuru-Guzik, A. Failure of Conventional Density Functionals for the Prediction of Molecular Crystal Polymorphism: A Quantum Monte Carlo Study. *J. Phys. Chem. Lett.* **2010**, *1*, 1789-1794.
18. Watson, M. A.; Hongo, K.; Iitaka, T.; Aspuru-Guzik, A. A Benchmark Quantum Monte Carlo Study of Molecular Crystal Polymorphism: A Challenging Case for Density-Functional Theory. *In Advances in Quantum Monte Carlo*; Tanaka, S., Rothstein, S. M., Lester, W. A., Jr., Eds.; ACS Symposium Series, No. *1094*; American Chemical Society: Washington, DC, 2012; Chapter 9, pp. 101-117.
19. Hongo, K.; Watson, M. A.; Iitaka, T.; Aspuru-Guzik, A.; Maezono, R. *J. Chem. Theory Comput.* **2015**, *11*, 907-917.
20. Hongo, K.; Cuong, N. T.; Maezono, R. The Importance of Electron Correlation on Stacking Interaction of Adenine-Thymine Base-Pair Step in B-DNA: A Quantum Monte Carlo Study. *J. Chem. Theory Comput.* **2013**, *9,* 1081-1086.
21. Benali, A.; Shulenburger, L.; Romero, N. A.; Kim, J.; von Lilienfeld, O. A. Application of Diffusion Monte Carlo to Materials Dominated by van der Waals Interactions. *J. Chem. Theory Comput.* **2014**, *10*, 3417-3422.
22. Tkatchenko, A.; DiStasio, R. A.; Car, R.; Scheffler, M. Accurate and Efficient Method for Many-Body van der Waals Interactions. *Phys. Rev. Lett.* **2012**, *108*, 236402.
23. Ambrosetti, A.; Alfè, D.; Robert A. DiStasio, J.; Tkatchenko, A. Hard Numbers for Large Molecules: Toward Exact Energetics for Supramolecular Systems. *J. Phys. Chem. Lett.* **2014**, *5*, 849-855.
24. Hongo, K.; Maezono, Diffusion Monte Carlo methods applied to Hamaker Constant evaluations. 2016, arXiv:1605.00580.
25. Needs, R. J.; Towler, M. D.; Drummond, N. D.; Ŕıos, P. L. Continuum variational and diffusion quantum Monte Carlo calculations. *J. Phys.: Condens. Matter* **2010**, *22*, 023201.
26. K computer at RIKEN, Japan. http://www.aics.riken.jp/en/, Accessed: 2016-08-01.
27. Maezono, R.; Towler, M. D. private communication, 2013.
28. Hongo, K.; Maezono, R.; Miura, K. Random number generators tested on quantum Monte Carlo simulations. *J. Comput. Chem.* **2010**, *31*, 2186-2194.
29. Hongo, K.; Maezono, R. Quantum Monte Carlo Simulations with RANLUX Random Number Generator. *Prog. Nucl. Sci. Tec.* **2011**, *2,* 51-55.

30. Shimoda, T.; Matsuki, Y.; Furusawa, M.; Aoki, T.; Yudasaka, I.; Tanaka, H.; Iwasawa, H.; Wang, D.; Miyasaka, M.; Takeuchi, Y. Solution-processed silicon films and transistors. *Nature* **2006**, *440*, 783-786.
31. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. Gaussian 09 Revision D.01. Gaussian Inc. Wallingford CT 2009.
32. Sinnokrot, M. O.; Sherrill, C. D. Highly Accurate Coupled Cluster Potential Energy Curves for the Benzene Dimer: Sandwich, T-Shaped, and Parallel-Displaced Configurations. *J. Phys. Chem. A* **2004**, *108*, 10200-10207.
33. Kolorenč, J.; Hu, S.; Mitas, L. Wave functions for quantum Monte Carlo calculations in solids: Orbitals from density functional theory with hybrid exchange-correlation functionals. *Phys. Rev. B* **2010**, *82*, 115108.
34. Hongo, K.; Maezono, R. A benchmark quantum Monte Carlo study of the ground state chromium dimer. *Int. J. Quantum Chem.* **2012**, *112*, 1243-1255.
35. Hongo, K.; Maezono, R. A Quantum Monte Carlo Study of the Ground State Chromium Dimer In *Advances in Quantum Monte Carlo*; Tanaka, S., Rothstein, S. M., Lester, W. A., Jr., Eds.; ACS Symposium Series, No. *1094*; American Chemical Society: Washington, DC, 2012; Chapter 8, pp. 91-99.
36. Per, M. C.; Walker, K. A.; Russo, S. P. How Important is Orbital Choice in Single-Determinant Diffusion Quantum Monte Carlo Calculations? *J. Chem. Theory Comput.* **2012**, *8*, 2255-2259.
37. Šponer, J.; Jurečka, P.; Marchan, I.; Luque, F. J.; Orozco, M.; Hobza, P. Nature of Base Stacking: Reference Quantum-Chemical Stacking Energies in Ten Unique B-DNA Base-Pair Steps. *Chem. Eur. J.* **2006**, *12*, 2854-2865.

38. Beran, G. J. O. Modeling Polymorphic Molecular Crystals with Electronic Structure Theory. *Chem. Rev.* **2016**, *116*, 5567-5613.
39. Santra, B.; Klimeš, J.; Alfè, D.; Tkatchenko, A.; Slater, B.; Michaelides, A.; Car, R.; Scheffer, M. Hydrogen Bonds and van der Waals Forces in Ice at Ambient and High Pressures. *Phys. Rev. Lett.* **2011**, *107*, 185701.
40. Kwee, H.; Zhang, S.; Krakauer, H. Finite-Size Correction in Many-Body Electronic Structure Calculations. *Phys. Rev. Lett.* **2008**, *100*, 126404.
41. Drummond, N. D.; Needs, R. J.; Sorouri, A.; Foulkes, W. M. C. Finite-size errors in continuum quantum Monte Carlo calculations. *Phys. Rev. B* **2008**, *78*, 125106.
42. Uejima, Y.; Terashima, T.; Maezono, R. Acceleration of a QM/MM-QMC simulation using GPU. *J. Comput. Chem.* **2011**, *32*, 2264-2272.
43. Uejima, Y.; Maezono, R. GPGPU for orbital function evaluation with a new updating scheme. *J. Comput. Chem.* **2013**, *34*, 83-94.
44. Flyvbjerg, H.; Petersen, H. G. Error estimates on averages of correlated data. *J. Chem. Phys.* **1989**, *91* 461-466.
45. Lee, R. M.; Conduit, G. J.; Nemec, N.; López-Ríos, P.; Drummond, N. D. Strategies for improving the efficiency of quantum Monte Carlo calculations. *Phys. Rev. E* **2011**, *83*, 066706.
46. Sugiyama, M. *Introduction to Statistical Machine Learning*; Morgan Kaufmann Publishers Inc.: San Francisco, 2015.
47. Trail, J. R.; Maezono, R. Optimum and efficient sampling for variational quantum Monte Carlo. *J. Chem. Phys.* **2010**, *133*, 174120.
48. Nakashima, H.; Nakatsuji, H. Solving the electron-nuclear Schrödinger equation of helium atom and its isoelectronic ions with the free iterative-complement-interaction method. *J. Chem. Phys.* **2007**, *127*, 224104.