Title	Automated error control in divide-and-conquer self-consistent field calculations
Author(s)	Kobayashi, Masato; Fujimori, Toshikazu; Taketsugu, Tetsuya
Citation	Journal of computational chemistry, 39(15), 909-916 https://doi.org/10.1002/jcc.25174
Issue Date	2018-06-05
Doc URL	http://hdl.handle.net/2115/74535
Rights	This is the peer reviewed version of the following article: Journal of computational chemistry. Volume39, Issue15. June 5, 2018. Pages 909-916 which has been published in final form at https://dx.doi.org/10.1002/jcc.25174. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.
Type	article (author version)
File Information	JCC_AdaptiveDC_revised_ver2.pdf

Automated error control in divide-and-conquer self-consistent field calculations

Masato Kobayashi, [a,b,c] *Toshikazu Fujimori,*[d] *and Tetsuya Taketsugu*[a,b]

---

[a]   *Masato Kobayashi, Tetsuya Taketsugu*

Department of Chemistry, Faculty of Science, Hokkaido University, Sapporo 060-0810,

Japan

E-mail: k-masato@sci.hokudai.ac.jp

[b]   *Masato Kobayashi, Tetsuya Taketsugu*

ESICB, Kyoto University, Kyoto 615-8520, Japan

[c]   *Masato Kobayashi*

PRESTO, Japan Science and Technology Agency, Kawaguchi 332-0012, Japan

[d]   *Toshikazu Fujimori*

Graduate School of Chemical Sciences and Engineering, Hokkaido University, Sapporo

060-0810, Japan

ABSTRACT:

In linear-scaling divide-and-conquer (DC) electronic structure calculations, a buffer region is used to control the error introduced by the DC approximation. In this study, an energy-based error estimation scheme is proposed for the DC self-consistent field method with a two-layer buffer region scheme. Based on this scheme, a procedure to automatically determine the appropriate buffer region in the DC method is proposed. It was confirmed that the present method works satisfactorily in calculations of water clusters and proteins, although its performance was insufficient for the calculation of a delocalized graphene system.

Introduction

Since the advent of computational quantum chemistry, the rapid increase in computational power has allowed the electronic structure calculation of ever-larger systems. In variational quantum chemical methods, the major computational task is the diagonalization of the Hamiltonian matrix that scales cubically with respect to the number of basis functions. To enable the electronic structure calculation of very large systems, many types of approximate electronic structure methods have been proposed in the last two decades that show linear-scaling computational time with respect to the system size. Almost all linear-scaling methods are approximations of existing matured electronic structure methods, such as Hartree–Fock (HF),[1] Kohn–Sham density functional theory (DFT),[2] and post-HF correlation calculations.[3] The results of linear-scaling methods bear two types of errors, i.e., those derived from the methodology and those from the linear-scaling approximation itself, the latter of which is desired to be controlled by the linear-scaling method itself.

In many linear-scaling methods, procedures for distance-based control and/or accuracy evaluation have been introduced. For example, in the density matrix minimization method,[4] a cutoff distance was introduced for the construction of an auxiliary density matrix of the support function.[5,6] In the molecular tailoring approach,[7] Gadre and coworkers defined the R-goodness parameter,[8,9] which indicates the quality of a fragmentation scheme based on the distance. The generalized energy-based fragmentation (GEBF) approach[10] can also employ a distance-based accuracy control scheme. [11] In the divide-and-conquer (DC) method,[12–14] the size of the buffer region controls primarily the accuracy of the approximation. Although methods with distance-based control parameters can systematically improve the accuracy of the approximations, it is still difficult to estimate the error in energy, which is the most important property in electronic structure

calculations. For some linear-scaling methods, density-based or energy-based error estimation schemes have also been developed. For example, for the density matrix purification method,[15,16] Rubensson and coworkers proposed a scheme to control the density-matrix error derived from iterative purification.[17,18] Niklasson *et al.* proposed a graph-based Fermi-operator expansion scheme, in which the accuracy was controlled by thresholded sparse matrix algebra.[19] However, in fragmentation-based linear-scaling approaches, such as those with DC and molecular tailoring methods, it can be difficult to control the accuracy without careful prior testing.[19] Another example of accuracy control can be found in the fragment molecular orbital method,[20] in which the results can be improved by increasing the order of the many-body expansion.[21,22]

In this study, a scheme to estimate the energy error introduced in DC-HF or DC-DFT calculations[23] is proposed. Nakai and coworkers extended the DC method to open-shell systems[13,24] and proposed an energy gradient.[25] Recently, they have also applied this method to the density-functional tight-binding (DFTB) theory,[26,27] which has enabled us to perform quantum mechanical calculation of one million atom systems within one minute with the Japanese K supercomputer.[28,29] In the present error estimation method, the two-layer buffer region scheme originally introduced by Dixon and Merz[30] was adopted. Guided by this error estimation scheme, an algorithm to automatically determine the appropriate buffer size was established.

Method

DC self-consistent field (SCF) method with a two-layer buffer region

Before summarizing the DC-SCF method with a two-layer buffer region, we note that, in the DC method, each basis function should be connected to an atom. Therefore, it is simply called an atomic orbital (AO) and denoted with a Greek letter index, $\{\mu, \nu, \dots\}$. In the DC method, the entire

system is first divided into N_{sub} disjoint subsystems, each of which is referred to as the central region. A set of basis functions connected to the central region of subsystem α is denoted by $S(\alpha)$. For each subsystem, the buffer region is added to the central region to construct a localization region, where the subsystem molecular orbitals (MOs) are constructed. In the two-layer buffer scheme introduced by Dixon and Merz,[30] the buffer region is hierarchically divided into two sub-regions, denoted as the inner and outer buffer regions (Figure 1). The inner buffer region, in which the set of AOs is denoted by $B_{\text{i}}(\alpha)$, is used to construct the subsystem MOs as well as to contribute to the density matrix; while the outer buffer region, in which the set of AOs is denoted by $B_{\text{o}}(\alpha)$, is only used to construct the subsystem MOs.

According to the DC-SCF scheme, the one-body density matrix of the entire system is approximated by the sum of subsystem contributions:

$$D_{\mu\nu} \approx D_{\mu\nu}^{\text{DC}} = \sum_{\alpha}^{\text{subsystem}} P_{\mu\nu}^{\alpha} D_{\mu\nu}^{\alpha} , \tag{1}$$

where $\mathbf{D}^{\alpha}$ expresses the density matrix of subsystem α, which is given in closed-shell case by

$$D_{\mu\nu}^{\alpha} = \sum_{p} f_{\beta}(\varepsilon_{\text{F}} - \varepsilon_{p}^{\alpha}) C_{\mu p}^{\alpha} C_{\nu p}^{\alpha*} . \tag{2}$$

The subsystem MOs, $\{\psi_{p}^{\alpha}\}$, are expanded in the two-layer buffer scheme with the AOs, $\{\phi_{\mu}\}$, in the outer localization region, $L_{\text{o}}(\alpha) \equiv S(\alpha) \cup B_{\text{i}}(\alpha) \cup B_{\text{o}}(\alpha)$:

$$\psi_{p}^{\alpha}(\mathbf{r}) = \sum_{\mu \in L_{\text{o}}(\alpha)} C_{\mu p}^{\alpha} \phi_{\mu}(\mathbf{r}) . \tag{3}$$

The MO coefficients, $\{\mathbf{C}_{p}^{\alpha}\}$, and MO energies, $\{\varepsilon_{p}^{\alpha}\}$, are obtained by solving the following subsystem Roothaan equation:

$$\mathbf{F}^{\alpha}[\mathbf{D}^{\text{DC}}]\mathbf{C}_{p}^{\alpha} = \varepsilon_{p}^{\alpha}\mathbf{S}^{\alpha}\mathbf{C}_{p}^{\alpha} . \tag{4}$$

$\mathbf{F}^{\alpha}[\mathbf{D}^{\mathrm{DC}}]$ and $\mathbf{S}^{\alpha}$ are the subsystem effective Hamiltonian and overlap matrices, respectively, which are the submatrices of the entire effective Hamiltonian and overlap matrices,

$$F_{\mu\nu}[\mathbf{D}^{\mathrm{DC}}] = H_{\mu\nu}^{\mathrm{core}} + \sum_{\lambda\sigma} D_{\lambda\sigma}^{\mathrm{DC}} \left[2\langle\mu\sigma|\nu\lambda\rangle - \langle\mu\sigma|\lambda\nu\rangle \right], \tag{5}$$

$$S_{\mu\nu} = \langle\phi_{\mu}|\phi_{\nu}\rangle, \tag{6}$$

for $\boldsymbol{L}_{\mathrm{o}}(\alpha)$ with two-electron integral notation of $\langle\mu\sigma|\nu\lambda\rangle = \iint d\mathbf{r}_1 d\mathbf{r}_2 \phi_{\mu}^{*}(\mathbf{r}_1)\phi_{\sigma}^{*}(\mathbf{r}_2)r_{12}^{-1}\phi_{\nu}(\mathbf{r}_1)\phi_{\lambda}(\mathbf{r}_2)$.

Although the Fock matrix is shown in eq. (5) as a typical example, the effective Hamiltonian generally depends on the density matrix. $\mathbf{P}^{\alpha}$ in eq. (1) is the partition matrix, which is defined in the two-layer buffer scheme by

$$P_{\mu\nu}^{\alpha} = \begin{cases} 1 & \text{for } \mu \in S(\alpha) \wedge \nu \in S(\alpha) \\ 1/2 & \text{for } (\mu \in S(\alpha) \wedge \nu \in B_{\mathrm{i}}(\alpha)) \text{ or } vice\ versa, \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

and $f_{\beta}(x) = [1 + \exp(-\beta x)]^{-1}$ is the Fermi distribution function with the inverse temperature parameter β. ε_{F} represents the universal Fermi level, which is determined by solving the following non-linear equation to conserve the total number of electrons, n_{e}, in the entire system:

$$n_{\mathrm{e}} = 2\mathrm{Tr}(\mathbf{D}^{\mathrm{DC}}\mathbf{S}). \tag{8}$$

The density matrix of eq. (1) and the effective Hamiltonian matrix of eq. (5) are determined self-consistently. The electronic energy can be obtained as the functional of the density matrix:

$$E[\mathbf{D}^{\mathrm{DC}}] = \mathrm{Tr}\left[\mathbf{D}^{\mathrm{DC}}(\mathbf{H}^{\mathrm{core}} + \mathbf{F}[\mathbf{D}^{\mathrm{DC}}])\right], \tag{9}$$

when the effective Hamiltonian is linear with respect to the density matrix, which is satisfied in HF and semiempirical MO calculations, but is not in typical DFT calculations.

Estimation of the DC energy error with a two-layer buffer region

If the outer buffer region is transferred into the inner buffer region, the density matrix changes by

$$\Delta D_{\mu\nu} = \sum_{\alpha} \left[P'^{\alpha}_{\mu\nu} \sum_{p} f_{\beta}(\varepsilon'_F - \varepsilon^{\alpha}_p) C^{\alpha}_{\mu p} C^{\alpha*}_{\nu p} - P^{\alpha}_{\mu\nu} \sum_{p} f_{\beta}(\varepsilon_F - \varepsilon^{\alpha}_p) C^{\alpha}_{\mu p} C^{\alpha*}_{\nu p} \right], \tag{10}$$

where relaxation of the subsystem MOs is neglected. $\mathbf{P}'^{\alpha}$ is the auxiliary partition matrix

$$P'^{\alpha}_{\mu\nu} = \begin{cases} 1 & \text{for } \mu \in S(\alpha) \wedge \nu \in S(\alpha) \\ 1/2 & \text{for } [\mu \in S(\alpha) \wedge \nu \in (B_i(\alpha) \cup B_o(\alpha))] \text{ or } \textit{vice versa}, \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

and ε'_F is the auxiliary Fermi level. The first-order energy variation can be estimated with the density matrix correction, $\Delta\mathbf{D}$, as

$$\Delta E = 2\text{Tr}\left[\Delta\mathbf{D}\mathbf{F}[\mathbf{D}^{DC}] \right], \tag{12}$$

where the effective Hamiltonian is assumed to be linear with respect to the density matrix.

There are two ways of obtaining the auxiliary Fermi level, ε'_F. The first one is to consider $\varepsilon'_F = \varepsilon_F$, which simplifies eq. (10) to

$$\Delta D_{\mu\nu} = \sum_{\alpha} (P'^{\alpha}_{\mu\nu} - P^{\alpha}_{\mu\nu}) \sum_{p} f_{\beta}(\varepsilon_F - \varepsilon^{\alpha}_p) C^{\alpha}_{\mu p} C^{\alpha*}_{\nu p} = \sum_{\alpha} \Delta P^{\alpha}_{\mu\nu} D^{\alpha}_{\mu\nu}, \tag{13}$$

where

$$\Delta P^{\alpha}_{\mu\nu} = \begin{cases} 1/2 & \text{for } \left(\mu \in S(\alpha) \wedge \nu \in B_o(\alpha) \right) \text{ or } \textit{vice versa} \\ 0 & \text{otherwise} \end{cases}. \tag{14}$$

Substituting eq. (13) into eq. (12) gives

$$\Delta E = 2\sum_{\mu\nu}\sum_{\alpha} \Delta P^{\alpha}_{\mu\nu} D^{\alpha}_{\mu\nu} F^{\alpha}_{\nu\mu}[\mathbf{D}^{DC}] = \sum_{\alpha} \sum_{\mu \in S(\alpha)} \sum_{\nu \in B_o(\alpha)} 2 D^{\alpha}_{\mu\nu} F^{\alpha}_{\nu\mu}[\mathbf{D}^{DC}]. \tag{15}$$

According to the energy density analysis (EDA),[31] which is analogous to the Mulliken population analysis, the variation in energy can be separated into the contributions from the atoms in the outer buffer regions:

$$\Delta E = \sum_{\alpha} \sum_{\mu \in S(\alpha)} \sum_{A \in \boldsymbol{B}_o(\alpha)} \sum_{\nu \in A} 2D_{\mu\nu}^{\alpha} F_{\nu\mu}^{\alpha}[\mathbf{D}^{DC}] = \sum_{\alpha} \sum_{A \in \boldsymbol{B}_o(\alpha)} \Delta E_A^{\alpha} , \qquad (16)$$

where

$$\Delta E_A^{\alpha} = \sum_{\mu \in S(\alpha)} \sum_{\nu \in A} 2D_{\mu\nu}^{\alpha} F_{\nu\mu}^{\alpha}[\mathbf{D}^{DC}] , \qquad (17)$$

and index A designates an atom.

The other way to obtain the auxiliary Fermi level relies on the electron number constraint, i.e., ε_F' is found by solving the following equation:

$$n_e = 2\mathrm{Tr}\left[\left(\mathbf{D}^{DC} + \Delta\mathbf{D}\right)\mathbf{S}\right] . \qquad (18)$$

Note that, in semiempirical MO calculations with a zero differential overlap (ZDO) approximation, the solution of eq. (18) is $\varepsilon_F' = \varepsilon_F$ (as in the first case), since $\mathbf{S} = \mathbf{I}$ and the diagonal elements of $\Delta\mathbf{D}$ with $\varepsilon_F' = \varepsilon_F$ [i.e., eq. (13)] are zero.

Automatic determination of the buffer region based on the estimated energy error

If one chooses $\varepsilon_F' = \varepsilon_F$ for DC calculations with a two-layer buffer region, the energy error introduced by the DC method can be estimated as the sum of contributions from the outer buffer atoms in each subsystem according to eq. (16). Also, it is known that the density matrix $\rho(\mathbf{r}_1, \mathbf{r}_2) = D_{\mu\nu}\phi_{\mu}(\mathbf{r}_1)\phi_{\nu}(\mathbf{r}_2)$ decays exponentially with the distance $|\mathbf{r}_1 - \mathbf{r}_2|$ in the case of an insulator.[32] Based on these facts, the following automatic extension scheme for the buffer region was developed:

i)   Evaluation of ΔE_A^{α} according to eq. (16) after constructing $\mathbf{F}[\mathbf{D}^{DC}]$ for each SCF cycle.

ii)   Transferring all atoms in the outer buffer region of subsystem α to its inner buffer region.

iii)   Inclusion of the atoms in the sphere with radius r_{ext} centered on atom A with $\Delta E_A^\alpha \geq e_{\text{thresh}}$

into the new outer buffer region of subsystem α.

iv)   Calculation of the subsystem MOs with eq. (4), construction of the density matrices with

eqs. (1) and (13), and back to step i).

After several cycles, the outer buffer region automatically vanishes when all ΔE_A^α become less

than the threshold. Following this scheme, it may become possible to choose the appropriate buffer

region for each subsystem while preserving the energy error per atom. In the actual implementation,

the subsystem density matrix element required in eq. (16) is approximated as $D_{\mu\nu}^\alpha \sim \Delta D_{\mu\nu}$ to avoid

the need for storing the density matrices of all subsystems. This approximation can be validated

because $\Delta D_{\mu\nu} = \Delta P_{\mu\nu}^{\alpha_1} D_{\mu\nu}^{\alpha_1} + \Delta P_{\mu\nu}^{\alpha_2} D_{\mu\nu}^{\alpha_2} = 1/2 \left(D_{\mu\nu}^{\alpha_1} + D_{\mu\nu}^{\alpha_2} \right) \sim D_{\mu\nu}^{\alpha_1}$ for $\mu \in S(\alpha_1) \wedge \mu \in \boldsymbol{B}_{\text{o}}(\alpha_2)$ and

$\nu \in S(\alpha_2) \wedge \nu \in \boldsymbol{B}_{\text{o}}(\alpha_1)$, where $D_{\mu\nu}^{\alpha_2}$ is considered to be similar to $D_{\mu\nu}^{\alpha_1}$. Here, $\mu \in \boldsymbol{B}_{\text{o}}(\alpha_2)$ is not

always the case. Therefore, $D_{\mu\nu}^\alpha \sim 2\Delta D_{\mu\nu}$ is the other choice of the approximation, while it is

equivalent to halve e_{thresh}.

Numerical Assessment

Computational details

We implemented the automatically controlled DC method to the GAMESS package[33,34] and

assessed its accuracy and efficiency for different types of systems. In the DC method, the inverse

temperature parameter, β, in Eq. (2) was set to 200 a.u. The parameters for the automated DC

method were set to $e_{\text{thresh}} = 0.1 \ \mu E_{\text{h}}$ and $r_{\text{ext}} = 3.0$ Å unless otherwise noted.

To discuss quantitatively the size of the localization region determined in the present scheme, we defined the major axis radius of localization region α, $l_{local}(\alpha)$, as half of the maximum atom pair distance in localization region α. The major axis radius at the initial SCF step, l_{local}^{ini}, where the outer buffer region is excluded from the localization region, should strongly correlate with the initial buffer size, while that at the final SCF step, l_{local}^{fin}, is expected to be barely dependent on the initial buffer size.

Estimated DC-HF energy error

We first compared two estimation schemes of the DC-HF energy error with eq. (12): (a) where $\varepsilon_F' = \varepsilon_F$ and (b) where ε_F' was determined for $(\mathbf{D}^{DC} + \Delta\mathbf{D})$. The estimated and actual energy errors were obtained for calculations of the crambin protein, as summarized in Table 1. Here, the 6-31G basis set[35] was adopted. The geometry of crambin was obtained via the protein data bank (PDB, identification number 1CRN) and the hydrogen atoms were then added with the FU program.[36] The estimated energy errors obtained from the second and final SCF steps are given for both estimation schemes. The initial guess density, which affects the estimation at the second SCF step, was obtained by the DC extended Hückel method implemented in GAMESS. In the DC calculations, the entire protein was cut between the carbonyl C and the α-C, and each fragment was treated as a central region. The buffer size was defined by r_b^{in} and r_b^{out}, where the unions of the spherical regions with radius r_b^{in} and r_b^{out} centered on each atom in the central region were considered as the inner and outer localization regions, respectively. As expected, the actual energy error decreased with the increasing buffer size, except for the smallest buffer size where an error cancellation seems to have occurred. The two estimation schemes did not display significant

differences. At both the second and final SCF steps, the difference in the errors estimated by the two schemes was less than 10% for $r_b^{in} \geq 4.0$ Å. The order of the estimated energy error at the final SCF step was consistent with that of the actual error. This estimation scheme worked reasonably even at the early SCF step, although the estimated error at the second SCF step was two or more times larger than that at the final step for $r_b^{in} \geq 4.0$ Å. The method was also tested in calculations of delocalized polyene system and the similar results were obtained (see Table S1 in the Supporting Information).

From the following section on, we will mainly focus on the semiempirical PM3 method,[37,38] which adopts the ZDO approximation.

Accuracy and computational time of automated DC calculations

The accuracy of the present method and its computational time requirements were examined in the calculation of a cubic system containing N_{water} randomly oriented water molecules with a weight density of 1.0 g cm^{-3}. In the DC calculations, each water molecule was treated as the central region. The initial buffer size was determined by r_b^{in} and r_b^{out}, the definitions of which are the same as in the previous section. Table 2 summarizes the initial buffer-size dependence of the automated DC-PM3 energy, the wall-clock computational time, and the number of SCF cycles for $N_{water} = 1000$. The computational time for the SCF calculations was measured using a computer node equipped with an Intel Xeon E5-1650 CPU (6 cores, 3.50 GHz), and the average of three measurements was calculated. The energy difference from the standard PM3 results divided by the number of atoms (3000) is also shown in parentheses. For $r_b^{in} \leq 6.0$ Å, the energy difference values are comparable: ~0.5 μE_h atom^{-1}. For $r_b^{in} \geq 6.5$ Å, the energy difference gradually decreases to zero because the

estimation for the initial buffer size is smaller than the threshold for most of the subsystems. In fact, the energy error at a single fixed buffer size of $r_b = 7.5$ Å is 0.50 μE_h atom^{-1}, in good agreement with the result for $r_b^{out} = 7.5$ Å. Although the number of SCF cycles is slightly larger than that for standard PM3 calculations, the computational time is ~10 times shorter for $r_b^{in} \leq 6.5$ Å. It is also suggested that a smaller initial buffer size results in the deterioration of the SCF convergence, which in turn leads to longer computational times.

Table 3 summarizes the average ($<l_{local}>$) and standard deviation ($\sigma[l_{local}]$) of the major axis radii among all localization regions in the automated DC-PM3 calculations of the water system (N_{water} = 1000). As expected, $<l_{local}^{ini}>$ increased linearly with the initial buffer size, and $\sigma[l_{local}^{ini}]$ was found to be relatively small. Interestingly, $<l_{local}^{fin}>$ was found to be larger for small initial buffer sizes up to $r_b^{in} = 6.0$ Å, although the difference was fairly small. Accordingly, $\sigma[l_{local}^{fin}]$ displayed smaller values for larger initial buffer sizes. It was thus suggested that large initial buffer sizes efficiently aid the selection of the appropriate buffer region and hence may reduce the computational time, although this effect does not largely affect the energy error. For readers with particular interest, the behavior of $<l_{local}>$ during the SCF iteration is given in Figure S1 in the Supporting Information.

Next, the dependence of the computational time on the system-size was determined, as shown in Figure 2. The initial buffer size was set to $r_b^{in} = 5.0$ Å and $r_b^{out} = 6.0$ Å. Even for $N_{water} = 400$, the time for the automated DC-PM3 calculation (54 s) was around four times shorter than that for the standard PM3 calculation (204 s). Furthermore, the time required for the standard PM3 calculations increased steeply with the system size. The scaling analysis with the double logarithmic plot indicated that the times for the standard and automated DC-PM3 calculations scaled as $O(n^{2.7})$ and $O(n^{1.6})$, respectively. For all systems, the DC energy error per atom was

within a narrow range: 0.44–0.57 μE_h. It was thus confirmed that the present method is able to control the accuracy of the DC method while maintaining an almost linear-scaling computational cost.

Table 4 summarizes the dependence of the energy error, computational time, and average of the major axis radii at the final step ($<l_\mathrm{local}^\mathrm{fin}>$) on the energy-based threshold, e_thresh, in the automated DC-PM3 calculations of the water system ($N_\mathrm{water} = 1000$). The initial buffer size was set to $r_\mathrm{b}^\mathrm{in} = 3.5$ and $r_\mathrm{b}^\mathrm{out} = 4.5$ Å. The result confirmed that the energy error is almost proportional to the energy threshold, e_thresh. As expected, $<l_\mathrm{local}^\mathrm{fin}>$ decreases gradually as the energy threshold increases. Accordingly, the computational time decreases as the energy threshold increases, while it shows more significant dependence than $<l_\mathrm{local}^\mathrm{fin}>$ does. Therefore, it is important to set e_thresh appropriately to enjoy both of good accuracy and less computational time.

Finally, the parallel efficiency of the program was examined, although the present source code is not optimized for the parallelization. Table 5 summarizes the dependence of the wall-clock times (t) for the automated DC-PM3 SCF calculation of the system with $N_\mathrm{water} = 1000$ on the number of CPU cores (N_core). The times were measured using a computer node equipped with two Intel Xeon E5-2667 CPU (8 cores, 3.20 GHz) and the average of three measurements was calculated. The initial buffer size was set to $r_\mathrm{b}^\mathrm{in} = 5.0$ Å and $r_\mathrm{b}^\mathrm{out} = 6.0$ Å. The parallel scalability S, given at the last column of the Table, is defined as the wall-clock time ratio $S = t(N_\mathrm{core} = 1) / [N_\mathrm{core} \times t(N_\mathrm{core})]$. Up to $N_\mathrm{core} = 4$, the scalability is higher than 0.7, while it rapidly decreases for $N_\mathrm{core} > 4$. There are two main reasons for the deterioration: (i) the reordering of the processing subsystem, which is effective for minimizing load imbalance, is not optimized for the varying subsystem size in the present automated DC method, and (ii) the semiempirical Hamiltonian matrix construction is not

13

efficiently parallelized in GAMESS. Although there is room for improvement, the present automated DC implementation is moderately parallelized, which especially works better for larger systems.

Dependence of the accuracy on the system

We then applied the method to covalently bound systems. Table 6 shows the initial buffer-size dependence of the automated DC-PM3 energy for the crambin system treated in the previous section. The energy difference from the standard PM3 results is also shown in parentheses. Again, it was confirmed that the energy difference was suppressed to small enough values: <1.4 μE_h atom^{-1}. The results for the crambin system did not show a systematic decrease of the energy difference up to $r_b^{in} = 5.5$ Å, as the initial buffer size was sufficiently smaller than the major axis radius of the final localization region, as summarized in Table 7. From these data, it was again confirmed that $<l_{local}^{fin}>$ and $\sigma[l_{local}^{fin}]$ tend to be smaller for larger initial buffer sizes. In comparison with Table 3, the $<l_{local}^{fin}>$ value for the crambin system is ~1 Å longer than that of the water system as the decay rate of the density matrix elements through covalent bonds is slower than that through hydrogen bonds.

Next, the present method was examined in calculations of the conjugated graphene system depicted in Figure 3 ($C_{180}H_{48}$). All atoms were placed on a plane and the C–C and C–H bond lengths were fixed to 1.42 and 1.09 Å, respectively. Table 8 shows the initial buffer-size dependence of the DC-PM3 energy for $C_{180}H_{48}$. In the DC calculation, the entire system was divided by a lattice spacing of 3.5 Å and each fragment was treated as a central region. The definitions of the initial buffer sizes, r_b^{in} and r_b^{out}, were the same as those in the previous sections. The energies obtained with a fixed buffer size are given in Table 8, together with the estimated

energy errors at the final SCF step and $<l_{\text{local}}^{\text{fin}}>$. Unlike the results for the water and crambin systems, the present automated DC method afforded in some cases a large energy deviation of $>10 ~\mu E_{\text{h}}$ atom^{-1}. The estimated energy error with the fixed buffer size was found to be about one order of magnitude smaller than the actual error. Due to the significantly slow decay of the density matrix for conjugated systems, the energy error estimated in the outer buffer region may be insufficient to reproduce the actual energy error. In addition, the energy error does not converge to the standard PM3 result due to the finite temperature approximation in the DC method. Actually, the finite-temperature PM3 energy with $\beta = 200$ a.u. is $-810.643352 ~E_{\text{h}}$, which is much closer to the converged DC-PM3 energy.

Finally, the dependence of the energy error on the energy-based threshold, e_{thresh}, was assessed. Figure 4 shows the dependence of the final energy error on e_{thresh} for the automated DC-PM3 calculation of 1000 water molecules and the crambin and graphene systems. The initial buffer size was set to $r_{\text{b}}^{\text{in}} = 3.5$ and $r_{\text{b}}^{\text{out}} = 4.5$ Å (or $r_{\text{b}}^{\text{out}} = 5.0$ Å for the graphene system). For the water and crambin systems, which were adequately treated by the automated DC method, the energy error increased proportionally to e_{thresh}, as expected. For the graphene system, however, the energy error did not show a systematic trend but oscillated throughout the e_{thresh} value range, even at low e_{thresh} values. Although there is still some room for improvement in the present automated DC scheme, it has been demonstrated that the energy error can be suppressed with the present method even for conjugated systems.

Performance in the HF and DFT calculations

We applied the present scheme to the HF method and DFT with the pure BLYP[39,40] and hybrid B3LYP[41,42] functionals. Here, the DC energy error is estimated with eq. (9) even for DFT calculations, where the Hamiltonian matrix is not linear with the density matrix. The option to use the HF Hamiltonian (Fock) matrix at the early SCF stage of the DFT calculation, which is adopted in the default setting of the GAMESS program, was switched off. Table 9 shows the initial buffer-size dependence of the DC-HF, DC-B3LYP, and DC-BLYP energies for a n-alkane ($C_{150}H_{302}$) with the 6-31G* basis set.[43] In the DC calculations, a C_2H_4 (or C_2H_5 for the edges) group is adopted as a central region. For the DC-HF and DC-B3LYP calculations, the energy error could be controlled within 0.7 μE_h atom^{-1}, while that for the pure DFT (DC-BLYP) calculation is one order larger in magnitude. The final localization region for central subsystems contains $(C_2H_4)_{10-12}$ in the DC-HF and DC-B3LYP calculations, while that does $(C_2H_4)_{7-8}$ in the DC-BLYP calculations. This result suggests that the magnitude of ΔE with eq. (9) is underestimated without the HF exchange term. Actually, the energy error of the DC-BLYP calculation with the DC-HF final localization region was 0.09 μE_h atom^{-1} for $r_b^{in} = 3.0$ and $r_b^{out} = 4.5$ Å. Therefore, in practical DC-DFT calculation, it is recommended to determine the appropriate buffer size with the early DC-HF SCF cycles, which can be performed in GAMESS by switching back on the default option to use the Fock matrix at the early SCF stage

Concluding Remarks

In this study, we have proposed an energy-based error estimation scheme for the linear-scaling DC quantum chemical method with the help of two-layer buffer regions. Exploiting the fact that the estimated energy error can be divided into contributions from the atoms in the outer buffer region

16

of each subsystem, our error estimation scheme was utilized for the automatic determination of the appropriate buffer region for the DC method. The present automated DC method worked satisfactorily in calculations of water, protein, and alkane systems, although its performance was insufficient in the calculation of a delocalized graphene system. Improvement of the present scheme for delocalized systems will be the scope of future studies. Furthermore, in the present scheme, the buffer region was gradually extended during the SCF cycles. There is an alternative approach to reduce the buffer region from a large initial buffer size, which may be preferably used when the method is applied to a series of quantum chemical calculations, such as geometry optimizations, where the appropriate buffer region of the previous step is available.

An energy-based error control scheme such as the present method will be indispensable for quantum chemical molecular dynamics simulations, especially for microcanonical ensembles, where the total energy conservation is rigorously examined. Recently, Nakai and coworkers have published a series of studies performing quantum chemical molecular dynamics simulations with the DC-DFTB method.[28,29] The present automated DC method can straightforwardly be extended to the so-called DFTB2 Hamiltonian, which is linear with respect to the density matrix. Furthermore, the present error estimation scheme is expected to work even for non-linear Hamiltonians such as DFTB3. The development of an automated DC-DFTB molecular dynamics program is desirable not only to reduce the effort of preliminary assessments before the production runs but also to guarantee the accuracy of the results. The application of the present scheme to the DC Hartree–Fock–Bogoliubov method,[44] which can effectively treat the static electron correlation of large systems,[45] is also straightforward, as well as that to the open-shell DC unrestricted HF method.[13,24] However, the present method cannot be combined with DC post-HF correlation methods such as the second-order Møller–Plesset perturbation (MP2)[46–48] and coupled

cluster[49–51] theories. As pointed out by Kobayashi and Nakai,[52] the appropriate buffer size for DC post-HF correlation calculations is generally smaller than that for DC-HF calculations. Furthermore, especially in DC-MP2 calculations, the appropriate buffer size should be determined before carrying out the MP2 calculations as the procedure is not iterative. The solution to this issue will pave the way toward the development of an automated DC-MP2 scheme in the near future.

Acknowledgment

REFERENCES

[1] A. Szabo, N. S. Ostlund, Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory; Dover: New York, **1996**.

[2] T. Tsuneda, Density Functional Theory in Quantum Chemistry; Springer: Tokyo, **2014**.

[3] T. Helgaker, P. Jørgensen, J. Olsen, Molecular Electronic-Structure Theory; John Wiley & Sons: New York, **2000**.

[4] X.-P. Li, R. W. Nunes, D. Vanderbilt, *Phys. Rev. B* **1993**, *47*, 10891.

[5] R. W. Nunes, D. Vanderbilt, *Phys. Rev. B* **1994**, *50*, 17611.

[6] D. R. Bowler, R. Choudhury, M. J. Gillan, T. Miyazaki, *Phys. Stat. Sol. B* **2006**, *243*, 989.

[7] N. Sahu, S. R. Gadre, *Acc. Chem. Res.* **2014**, *47*, 2739.

[8] N. Sahu, S. D. Yeole, S. R. Gadre, *J. Chem. Phys.* **2013**, *138*, 104101.

[9] V. Ganesh, R. K. Dongare, P. Balanarayan, S. R. Gadre, *J. Chem. Phys.* **2006**, *125*, 104109.

[10] S. Li, W. Li, J. Ma, *Acc. Chem. Res.* **2014**, 47, 2712.

[11] W. Li, S. Li, Y. Jiang, *J. Phys. Chem. A* **2007**, 111, 2193.

[12] W. Yang, T.-S. Lee, *J. Chem. Phys.* **1995**, *103*, 5674.

[13] M. Kobayashi, H. Nakai, *Phys. Chem. Chem. Phys.* **2012**, *14*, 7629.

[14] M. Kobayashi, H. Nakai, In *Linear-Scaling Techniques in Computational Chemistry and Physics,* Chapter 5; R. Zaleśny, M. G. Papadopoulos, P. G. Mezey, J. Leszczynski, Eds.; Springer: Dordrecht, **2011**; pp. 97–127.

[15] A. M. N. Niklasson, *Phys. Rev. B* **2002**, *66*, 155115.

[16] J. Kim, Y. Jung, *Int. J. Quantum Chem.* **2016**, *116*, 563.

[17] E. H. Rubensson, *SIAM J. Sci. Comput.* **2012**, *34*, B1.

[18] E. H. Rubensson, E. Rudberg, P. Salek, *J. Chem. Phys.* **2008**, *128*, 074106.

[19] A. M. N. Niklasson, S. M. Mniszewski, C. F. A. Negre, M. J. Cawkwell, P. J. Swart, J. Mohd-Yusof, T. C. Germann, M. E. Wall, N. Bock, E. H. Rubensson, H. Djidjev, *J. Chem. Phys.* **2016**, *144*, 234101.

[20] D. G. Fedorov, K. Kitaura, The Fragment Molecular Orbital Method: Practical Applications to Large Molecular Systems; CRC: Boca Raton, **2009**.

[21] D. G. Fedorov, K. Kitaura, *J. Chem. Phys.* **2004**, *120*, 6832.

[22] T. Nakano, Y. Mochizuki, K. Yamashita, C. Watanabe, K. Fukuzawa, K. Segawa, Y. Okiyama, T. Tsukamoto, S. Tanaka, *Chem. Phys. Lett.* **2012**, *523*, 128.

[23] T. Akama, M. Kobayashi, H. Nakai, *J. Comput. Chem.* **2007**, *28*, 2003.

[24] M. Kobayashi, T. Yoshikawa, H. Nakai, *Chem. Phys. Lett.* **2010**, *500*, 172.

[25] M. Kobayashi, T. Kunisada, T. Akama, D. Sakura, H. Nakai, *J. Chem. Phys.* **2011**, *134*, 034105.

[26] M. Elstner, D. Porezag, G. Jungnickel, J. Elsner, M. Haugk, Th. Frauenheim, S. Suhai, G. Seifert, *Phys. Rev. B* **1998**, *58*, 7260.

[27] M. Gaus, Q. Cui, M. Elstner, *J. Chem. Theory Comput.* **2011**, *7*, 931.

[28] H. Nishizawa, Y. Nishimura, M. Kobayashi, S. Irle, H. Nakai, *J. Comput. Chem.* **2016**, *37*, 1983.

[29] H. Nakai, A. W. Sakti, Y. Nishimura, *J. Phys. Chem. B* **2016**, *120*, 217.

[30] S. L. Dixon, K. M. Merz Jr., *J. Chem. Phys.* **1997**, *107*, 879.

[31] H. Nakai, *Chem. Phys. Lett.* **2002**, *363*, 73.

[32] S. Goedecker, *Rev. Mod. Phys.* **1999**, *71*, 1085.

[33] M. S. Gordon, M. W. Schmidt, In *Theory and Applications of Computational Chemistry: The First Forty Years*; C. Dykstra, G. Frenking, K. Kim, G. Scuseria, Eds.; Elsevier: Amsterdam, **2005**; pp 1167–1189.

[34] M. W. Schmidt, K. K. Baldridge, J. A. Boatz, S. T. Elbert, M. S. Gordon, J. H. Jensen, S. Koseki, N. Matsunaga, K. A. Nguyen, S. Su, T. L. Windus, M. Dupuis, J. A. Montgomery Jr., *J. Comput. Chem.* **1993**, *14*, 1347.

[35] W. J. Hehre, R. Ditchfield, J.A. Pople, *J. Chem. Phys.* **1972**, *56*, 2257.

[36] K. Kitaura, FU. https://sourceforge.net/projects/fusuite/ (accessed Feb 6, 2017).

[37] J. J. P. Stewart, *J. Comput. Chem.* **1989**, *10*, 209.

[38] J. J. P. Stewart, *J. Comput. Chem.* **1989**, *10*, 221.

[39] A. D. Becke, *Phys. Rev. A* **1988**, *38*, 3098.

[40] C. Lee, W. Yang, R. G. Parr, *Phys. Rev. B* **1988**, *37*, 785.

[41] A. D. Becke, *J. Chem. Phys.* **1993**, *98*, 5648.

[42] P. J. Stephens, F. J. Devlin, C. F. Chabalowski, M. J. Frisch, *J. Phys. Chem.* **1994**, *98*, 11623.

[43] P. C. Hariharan, J. A. Pople, *Theor. Chim. Acta* **1973**, *28*, 213.

[44] V. N. Staroverov, G. E. Scuseria, *J. Chem. Phys.* **2002**, *117*, 11107.

[45] M. Kobayashi, T. Taketsugu, *Chem. Lett.* **2016**, *45*, 1268.

[46] M. Kobayashi, Y. Imamura, H. Nakai, *J. Chem. Phys.* **2007**, *127*, 074103.

[47] M. Kobayashi, T. Taketsugu, *Theor. Chem. Acc.* **2015**, *134*, 107.

[48] M. Kobayashi, H. Nakai, *J. Chem. Phys.* **2013**, *138*, 044102.

[49] M. Kobayashi, H. Nakai, *J. Chem. Phys.* **2008**, *129*, 044103.

[50] M. Kobayashi, H. Nakai, *J. Chem. Phys.* **2009**, *131*, 114108.

[51] T. Yoshikawa, M. Kobayashi, H. Nakai, *Int. J. Quantum Chem.* **2013**, *113*, 218.

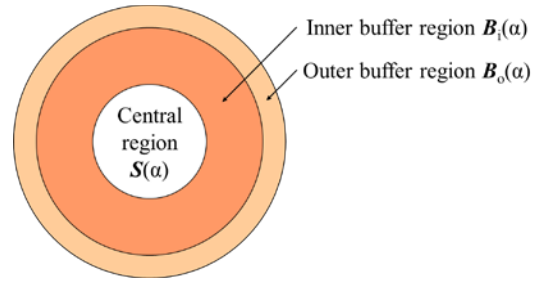[52] M. Kobayashi, H. Nakai, *Int. J. Quantum Chem.* **2009**, *109*, 2227.

Figure 1. Structure of central, inner buffer, and outer buffer regions in the two-layer DC method.
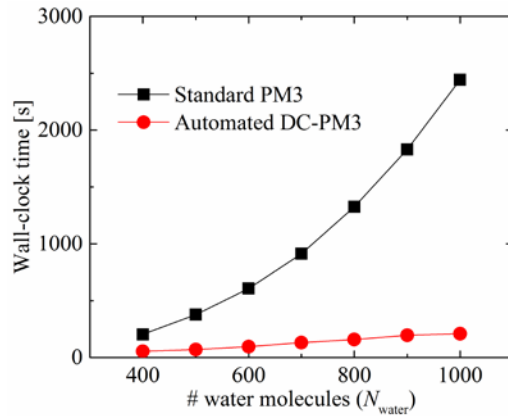
Figure 2. System-size dependence of the wall-clock computational time of standard PM3 and automated DC-PM3 calculations for the model system containing N_{water} water molecules. The initial buffer size for the DC calculations was fixed to $r_{\text{b}}^{\text{in}} = 5.0$ and $r_{\text{b}}^{\text{out}} = 6.0$ Å.

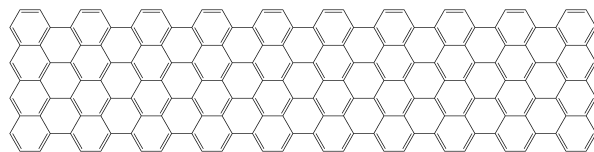**Figure 3.** Structure of the graphene system, $C_{180}H_{48}$.
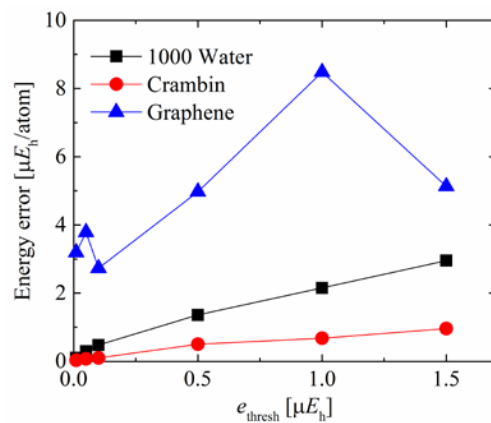
Figure 4. Dependence of the energy error on the energy-based threshold, e_{thresh}, in the automated DC-PM3 calculations of 1000 water molecules and the crambin and graphene systems.

Table 1. Buffer size dependence of the actual and estimated DC-HF energy errors for the crambin protein. Standard HF energy is $-17996.926754\ E_h$

r_b^{in} /Å	r_b^{out} /Å	Actual error /E_h	$-\Delta E$ by scheme (a) /E_h		$-\Delta E$ by scheme (b) /E_h	
			2nd step	Final step	2nd step	Final step
3.5	4.5	−0.144241	−0.768550	−0.886510	−0.682481	−0.890704
4.0	5.0	−0.348118	−1.005190	−0.532129	−1.071929	−0.526987
4.5	5.5	−0.067862	−0.504153	−0.126115	−0.507108	−0.125820
5.0	6.0	−0.017408	−0.123892	−0.038836	−0.124559	−0.038845
5.5	6.5	−0.005229	−0.084293	−0.016880	−0.084118	−0.016881

Table 2. Initial buffer-size dependence of the total energy, the wall-clock computational time, and the number of SCF cycles for the automated DC-PM3 calculation of the model system containing 1000 water molecules

r_b^{in} /Å	r_b^{out} /Å	Energy /E_h	(Diff. /$\mu E_h \bullet atom^{-1}$)	Time /s	# cycles
3.5	4.5	−11945.190938	(+0.48)	250	14
4.0	5.0	−11945.190942	(+0.48)	246	14
4.5	5.5	−11945.190837	(+0.51)	233	13
5.0	6.0	−11945.190719	(+0.55)	209	12
5.5	6.5	−11945.190414	(+0.65)	209	13
6.0	7.0	−11945.190229	(+0.72)	202	12
6.5	7.5	−11945.191077	(+0.43)	246	12
7.0	8.0	−11945.191791	(+0.19)	325	12
Standard PM3		−11945.192376		2443	11

Table 3. Average and standard deviation of the major axis radii of all localization regions at the initial and final SCF steps in the automated DC-PM3 calculation of the model system containing 1000 water molecules

r_b^{in} /Å	r_b^{out} /Å	$<l_{local}^{ini}>$ /Å	$\sigma[l_{local}^{ini}]$ /Å	$<l_{local}^{fin}>$ /Å	$\sigma[l_{local}^{fin}]$ /Å
3.5	4.5	3.577	0.306	8.105	0.626
4.0	5.0	4.105	0.314	8.030	0.604
4.5	5.5	4.659	0.319	8.031	0.581
5.0	6.0	5.194	0.305	8.176	0.635
5.5	6.5	5.703	0.313	7.981	0.766
6.0	7.0	6.213	0.318	7.640	0.647
6.5	7.5	6.708	0.342	7.832	0.507
7.0	8.0	7.218	0.375	8.228	0.430

Table 4. Energy threshold (e_{thresh}) dependence of the energy error, computational time, and average of the major axis radii of all localization regions at the final SCF steps in the automated DC-PM3 calculation of the model system containing 1000 water molecules. The initial buffer size was set to $r_{\text{b}}^{\text{in}} = 3.5$ and $r_{\text{b}}^{\text{out}} = 4.5$ Å.

e_{thresh} /μE_{h}	Energy error /μE_{h} atom^{-1}	Time /s	$<l_{\text{local}}^{\text{fin}}>$/Å
0.01	0.103	451	8.994
0.05	0.295	319	8.396
0.10	0.479	243	8.105
0.50	1.360	179	7.421
1.00	2.153	173	7.156
1.50	2.953	144	7.024

Table 5. Parallelization efficiency of the automated DC-PM3 calculations of the model system containing 1000 water molecules

N_{core}	Time (t) /s	Scalability (S)
1	710	1.000
2	381	0.933
4	252	0.706
8	229	0.387
16	272	0.163

Table 6. Initial buffer-size dependence of the automated DC-PM3 energy for the crambin system

r_b^{in} /Å	r_b^{out} /Å	Energy /E_h	(Diff. /$\mu E_h \bullet atom^{-1}$)
3.5	4.5	−2117.084675	(+0.10)
4.0	5.0	−2117.084601	(+0.21)
4.5	5.5	−2117.084647	(+0.14)
5.0	6.0	−2117.083858	(+1.37)
5.5	6.5	−2117.084462	(+0.43)
Standard PM3		−2117.084739	

Table 7. Average and standard deviation of the major axis radii of all localization regions at the initial and final SCF steps in the automated DC-PM3 calculation of the crambin system

r_b^{in} /Å	r_b^{out} /Å	$<l_{local}^{ini}>$ /Å	$\sigma[l_{local}^{ini}]$ /Å	$<l_{local}^{fin}>$ /Å	$\sigma[l_{local}^{fin}]$ /Å
3.5	4.5	5.348	0.890	9.082	1.204
4.0	5.0	5.835	0.981	9.158	1.208
4.5	5.5	6.338	0.948	9.043	1.178
5.0	6.0	6.781	0.969	9.072	1.198
5.5	6.5	7.204	1.065	8.899	1.191

Table 8. Initial buffer-size dependence of the buffer-size fixed and automated DC-PM3 energies for the graphene system, $C_{180}H_{48}$. The estimated energy error at the final SCF step in the buffer-size fixed calculation is also given. Standard PM3 energy is $-810.650309\ E_h$

r_b^{in} /Å	r_b^{out} /Å	Buffer size fixed		Automated		$<l_{local}^{fin}>$ /Å
		Actual error /E_h	$-\Delta E$ /E_h	Energy /E_h	(Diff. /$\mu E_h \bullet atom^{-1}$)	
3.5	5.0	+0.466022	+0.141160	−810.650931	(−2.73)	10.672
4.0	5.5	+0.163445	+0.023277	−810.651058	(−3.29)	10.820
4.5	6.0	+0.107509	+0.024784	−810.644480	(+25.56)	12.251
5.0	6.5	+0.137531	+0.013554	−810.650325	(−0.07)	11.458
5.5	7.0	+0.065487	+0.008674	−810.648093	(+9.72)	11.427
6.0	7.5	+0.041446	+0.008377	−810.651029	(−3.16)	10.926
6.5	8.0	+0.044827	+0.004561	−810.644494	(+25.50)	12.033
7.0	8.5	+0.039109	+0.003055	−810.643122	(+31.52)	11.796
7.5	9.0	+0.016598	+0.001494	−810.655281	(−21.81)	10.846
8.0	9.5	+0.011964	+0.001004	−810.644095	(+27.25)	12.199
8.5	10.0	+0.014299	+0.001242	−810.642821	(+32.84)	11.992
9.0	10.5	+0.017775	+0.000638	−810.643328	(+30.61)	12.310
9.5	11.0	+0.010530	+0.000513	−810.644435	(+25.76)	12.543
10.0	11.5	+0.008625	+0.000339	−810.644620	(+24.95)	12.211

Table 9. Initial buffer-size dependence of the automated DC-HF and DC-DFT energies for the n-alkane system, $C_{150}H_{302}$.

r_b^{in} /Å	r_b^{out} /Å	HF		B3LYP		BLYP	
		Energy /E_h	(Diff.) /$\mu E_h \cdot atom^{-1}$	Energy /E_h	(Diff.) /$\mu E_h \cdot atom^{-1}$	Energy /E_h	(Diff.) /$\mu E_h \cdot atom^{-1}$
3.0	4.5	−5855.829391	(−0.20)	−5897.795934	(+0.64)	−5893.578891	(+4.98)
4.0	5.5	−5855.829339	(−0.08)	−5897.796218	(+0.02)	−5893.578904	(+4.95)
5.0	6.5	−5855.829321	(−0.04)	−5897.796217	(+0.02)	−5893.578904	(+4.95)
6.0	7.5	−5855.829329	(−0.06)	−5897.795958	(+0.59)	−5893.578954	(+4.84)
7.0	8.5	−5855.829347	(−0.10)	−5897.796192	(+0.07)	−5893.579524	(+3.58)
Standard		−5855.829302		−5897.796225		−5893.581143	