Linear-scaling first-principles molecular dynamics of complex biological systems with the CONQUEST code

Takao Otsuka[1], Makoto Taiji[1], David R. Bowler[2,3,4] and Tsuyoshi Miyazaki[2,3,5]

[1]*Quantitative Biology Center (QBiC), RIKEN, 6-2-4 Furuedai, Suita, Osaka 565-0874, Japan*

[2]*International Center for Materials Nanoarchitechtonics (MANA), National Institute for Materials Science (NIMS), 1-1 Namiki, Tsukuba, Ibaraki 305-0044, Japan*

[3]*Department of Physics and Astronomy, University College London (UCL), Gower Street, London WC1E 6BT, U.K.*

[4]*London Centre for Nanotechnology, UCL, 17-19 Gordon Street, London WC1H 0AH, U.K.*

[5]*Division of Bio-organometallics, Research Institute for Science and Technology, Tokyo University of Science, 2641 Yamasaki, Noda, Chiba 278-8510, Japan*

The recent progress of linear-scaling or $\mathcal{O}(N)$ methods in the density functional theory (DFT) is remarkable. In this paper, we show that all-atom molecular dynamics simulations on complex biological systems based on DFT are now possible using our linear-scaling DFT code CONQUEST. We first overview the calculation methods used in CONQUEST and explain the method introduced recently to realise efficient and robust first-principles molecular dynamics (FPMD) with $\mathcal{O}(N)$ DFT. Then we show that we can perform reliable all-atom FPMD simulations on a hydrated DNA model containing about 3400 atoms. We also report that the velocity scaling method is both reliable and useful to control the temperature of the FPMD simulation of this system. Based on these results, we conclude that CONQUEST is ready to do reliable FPMD simulations on complex biological systems.

1. Introduction

Molecular simulation technology is now commonly used to explore biological phenomena of biomolecular systems. It helps us understand the mechanism of various biological phenomena, including enzyme reactions, photoexcitations, molecular interactions and so on.[1)] Although most molecular simulations of biological systems use parametrised interatomic potentials, the reliability of such empirical potentials in various environments or for some phenomena is sometimes doubtful. So it is important that we should be able to perform molecular simulations based on quantum mechanics. However, the cost of quantum simulations, such as first-principles (FP) simulations based on the density functional theory (DFT), is usually very expensive, especially for large systems. As is well known, the CPU time of normal DFT calculations is proportional to the cube

of the number of atoms N in the simulation cell. It is very difficult and expensive to treat systems containing more than 1000 atoms within DFT. To reduce this demanding cost, quantum mechanics and molecular mechanics (QM/MM) hybrid method or its molecular dynamics version (QM/MM-MD) is often used for molecular simulations on biological systems. However, it is usually impossible to remove the effect of the artificial boundary between the two regions introduced in the hybrid calculations. There are increasing demands for all-atom DFT simulations on complex biological systems.

In this respect, the recent advances in computational techniques for large-scale DFT calculations called linear-scaling or $\mathcal{O}(N)$ method, whose calculation cost is only proportional to N, is encouraging.[2)] We have been developing our own linear-scaling DFT code called CONQUEST[3)] and have recently demonstrated that we can treat million-atom systems with DFT using the code.[4,5)] We also recently introduced a method to realise efficient and reliable first-principles molecular dynamics (FPMD) on large systems, by combining the $\mathcal{O}(N)$ DFT and extended Lagrangian Born-Oppenheimer molecular dynamics (XL-BOMD) methods.[6)] We investigated the requirements for calculations with accurate $\mathcal{O}(N)$ FPMD simulations and actually performed FPMD on a very large crystalline silicon system, containing 32,768 atoms.[7)] We expect that we can also employ this technique on large and complex biological systems.

In this paper, we overview the calculation methods used in CONQUEST, and explain the combined method for efficient and accurate FPMD simulations. Then, we show that we can do reliable all-atom FPMD simulations on a test DNA model, a DNA decamer hydrated with a large number of water molecules, consisting of about 3400 atoms. We demonstrate that the FPMD simulations on the hydrated DNA system are robust and accurate.

2. Linear-scaling first-principles molecular dynamics method

2.1 Linear-scaling DFT code CONQUEST

In this subsection, we first overview the computational methods and recent progress of the CONQUEST code.

In CONQUEST, we use the Kohn-Sham density matrix defined as

$$\rho(\boldsymbol{r}, \boldsymbol{r}') = \sum_n f_n \Psi_n(\boldsymbol{r}) \Psi_n^*(\boldsymbol{r}'), \tag{1}$$

where $\Psi_n(\boldsymbol{r})$ is the eigenfunction (Kohn-Sham orbitals) of the Kohn-Sham Hamiltonian for the band index n, and f_n is its occupation number.[8–10)] The total energy based on

DFT can be calculated from the density matrix, with the use of the pseudopotential method and standard exchange-correlation functionals such as the local density approximation (LDA) or a generalised gradient approximation (GGA). It should be noted that an efficient technique to calculate the exact exchange term has been recently introduced[11] to the code, and thus hybrid functionals are also now available.

In CONQUEST, we represent the density matrix by localised orbitals called "support functions", $\phi_{i\alpha}(\boldsymbol{r})$, with the matrix elements $K_{i\alpha,j\beta}$ which are the coefficients of the density matrix expressed in this non-orthogonal basis of support functions.

$$\rho(\boldsymbol{r},\boldsymbol{r}') = \sum_{i\alpha,j\beta} \phi_{i\alpha}(\boldsymbol{r}) K_{i\alpha,j\beta} \phi_{j\beta}(\boldsymbol{r}'), \tag{2}$$

The support function $\phi_{i\alpha}(r)$ for the orbital α is centred on the atom i and is non-zero only inside the "support region". The support functions themselves are represented in terms of basis functions, and two types of basis sets are available in CONQUEST: B-splines on regular grids;[12] and numerical pseudo-atomic orbitals (PAOs).[13–15] When B-splines are used, we can systematically improve the accuracy of the basis set by reducing the grid spacing and can reach the planewave accuracy. On the other hand, we can employ efficient calculations with a reasonable accuracy by using PAOs as basis sets. Even with PAO basis sets, we can improve the accuracy by increasing the number of basis functions, but the computational cost usually increases very rapidly. We have recently introduced a method to treat such accurate but large PAO basis sets efficiently.[16–18] With this method, called the multisite support function (MSSF) method, we can perform accurate calculations without increasing the CPU time significantly.

The matrix $K_{i\alpha,j\beta}$ is obtained either by the conventional diagonalization method, or by a linear-scaling (or $\mathcal{O}(N)$) method. In the case of $\mathcal{O}(N)$ calculations, CONQUEST uses the desnity matrix minimisation (DMM) method proposed by Li *et al.*[19] In this method, we express the matrix K following McWeeny's purification transformation,[20]

$$K = 3LSL - 2LSLSL, \tag{3}$$

to impose weak idempotency on the density matrix. Here, the matrix L is called the auxiliary density matrix and S ($S_{i\alpha,j\beta} = \langle \phi_{i\alpha} | \phi_{j\beta} \rangle$) is the overlap matrix between the support functions. To achieve the $\mathcal{O}(N)$ behaviour using the locality of density matrix, we introduce a spatial cut-off R_L on the L-matrix: $L_{i\alpha,j\beta} = 0$ for $|\mathbf{R}_i - \mathbf{R}_j| > R_L$, where $\mathbf{R}_i$ are the atomic positions. Then we calculate the matrix elements $L_{i\alpha,j\beta}$ which minimise the DFT total energy using numerical optimisation methods, such as the

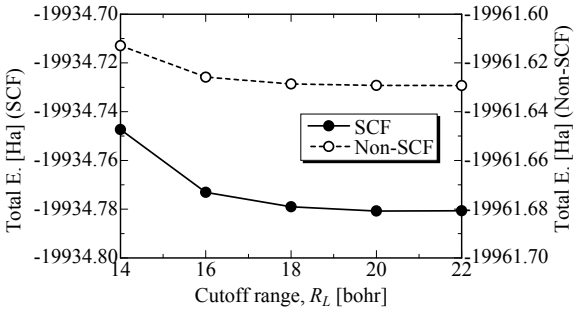

Fig. 1. Total energy of a hydrated DNA system, whose structure is shown in Fig. 2, as a function of R_L, the cutoff range of the auxiliary density matrix L. Total energy obtained by non-self-consistent technique (NonSCF) is also prensented.

residual minimisation method[21,22] One of the big advantages of DMM method is that it satisfies the variational principle and we can monitor the accuracy of the $\mathcal{O}(N)$ calculations by checking the R_L dependence of the total energy. Figure 1 shows the R_L dependence of the total energy for a DNA system, which is the target of the MD study shown in the next section. Here, the total energy using Harris-Foulkes functional obtained by non-self-consisten (NSC) technique[23,24] is also presented together with the DFT total energy using the self-consistent-field (SCF) charge density. This graph shows that the total energy converges as the cutoff applied to the L matrix is increased, regardless of the self-consistency. We can also see that the result with the NSC technique converges faster than the SCF result. It is probably due to the fact that the electronic structure by NSC usually has a larger energy gap than SCF and is more localised. We also note that tests on smaller systems (dry DNA with NSC) show that it converges to the exact diagonalisation result.[25]

Another strong point of CONQUEST is its excellent efficiency on massively parallel computers. Since CONQUEST uses the locality of the electronic structure, it also has an advantage in parallelisation. We recently reported its parallel efficiency on K computer and showed that it has almost ideal parallel efficiency even when we use more than 200,000 cores.[5,26] It was also demonstrated that we can now treat million-atom systems using the CONQUEST code on such large-scale parallel computers. Using this ability of the code, the code has been used for structure relaxations on the nano-scale systems of semiconductor surfaces.[27,28]

2.2 Molecular dynamics with the CONQUEST code

Even though we can now calculate the total energy and atomic forces[29,30] of very large systems using the $\mathcal{O}(N)$ DFT method, this does not guarantee that stable, efficient and accurate FPMD simulations are also possible in practice. There are two types of methods widely used in the conventional FPMD simulations: Car-Parrinello MD (CPMD) and Born-Oppenheimer MD (BOMD). To realise $\mathcal{O}(N)$ FPMD simulations, we adopt the BOMD method since we do not want to have the ambiguity of fictitious mass, which is used in CPMD simulations. Another advantage of the BOMD method is we can use larger time step than CPMD. However, it should be noted that the stability or accuracy of the BOMD simulations strongly depends on the accuracy of the calcualted forces. We usually use an iterative method to calculate the ground state of the electronic structure even in the conventional methods. It is well known that we can have an unphysical energy drift, if the electronic structure is not well converged and the calculated forces are not accurate enough. This problem is closely related to the time reversibility of the optimised electronic structure. In order to solve this problem, Niklasson *et al.* recently proposed a new method called the extended Lagrangian Borm-Oppenheimer MD (XL-BOMD) method. With this method, the time reversibility of the electronic structure is maintained and the stability of the BOMD simulations are greatly improved.[6,31,32]

Recently, we combined this XL-BOMD scheme with the DMM method and demonstrated that the combined method enables us to do efficient and reliable FPMD with the $\mathcal{O}(N)$ method.[7] The Lagrangian in the XL-BOMD scheme $\mathcal{L}^{\mathrm{XBO}}$ is defined in the following way, using the Lagrangian in the usual BOMD method $\mathcal{L}^{\mathrm{BO}}$,

$$\mathcal{L}^{\mathrm{XBO}}\left(X,\dot{X},\boldsymbol{R},\dot{\boldsymbol{R}}\right)=\mathcal{L}^{\mathrm{BO}}\left(\boldsymbol{R},\dot{\boldsymbol{R}}\right)+\frac{\mu}{2}\mathrm{Tr}\left[\dot{X}^2\right]-\frac{\mu\omega^2}{2}\mathrm{Tr}\left[(LS-X)^2\right] \tag{4}$$

where the matrix X is a sparse matrix introduced to prepare the initial guess of L matrix at each MD step. μ is the fictitious electronic mass, and ω is the curvature of the electronic harmonic potential. As in the original XL-BOMD method, if we take the limit $\mu \to 0$, $\mathcal{L}^{\mathrm{XBO}}$ becomes $\mathcal{L}^{\mathrm{BO}}$ and we have two equations of motion for nuclear positions and X, respectively. If we apply the time-reversible Verlet scheme to calculate X using the equation of motion, we have

$$X(t+\delta t)=2X(t)-X(t-\delta t)+\delta t^2\omega^2(L(t)S(t)-X(t)) \tag{5}$$

which shows that $X(t)$ is time reversible and evolves in a harmonic potential centred

around the ground-state $L(t)S(t)$. We can expect that a good initial guess for the L-matrix, which will obey time reversal symmetry, can be calculated by multiplying X and S^{-1} (in CONQUEST, the sparse approximate inverse S is computed using Hotelling's method[33]). As a result, the optimised L matrix also satisfies time reversibility and the trajectories of FPMD simulations become stable and accurate. In practice, for the numerical propagation of the matrix X, we use an equation of motion with a dissipative term to maintain numerical stability of the matrix X.[34]

In our previous study, we clarified the effects of control parameters used in the DMM method in the FPMD simulations. We have found that even when the total energy is not fully converged, MD trajectories are almost the same as those in more accurate MD simulations. We also demonstrated that reliable MD simulations can be actually performed on 32,768-atom crystalline silicon system using 1024 CPUs (8192 cores) of the K computer. Since we have already shown that parallel efficiency of CONQUEST is ideal even when using more than 200,000 cores, we can conclude that FPMD simulations on million-atom systems are now available using a big supercomputer like the K computer.

3. All-atom FPMD simulations on a hydrated DNA system with the CONQUEST code

Although we already demonstrated the practical ability of the combined (DMM+XL-BOMD) method, the examples of FPMD simulations using CONQUEST have been limited to simple systems so far, such as crystalline silicon or bulk water. In this section, we present another example of MD simulations on a more complex system, a hydrated DNA system, whose structure is shown in Fig. 2. The system was studied in our previous work,[25] and consists of DNA 10 base pairs (d(CCATTAATGG)2 in PDB ID: 1WQZ) of 634 atoms, 9 Mg atoms as counter ions, and 932 H_2O molecules, being 3,439 atoms in total. The initial structure is prepared by classical MD simulation using the AMBER9 with the force fields of PARM99[35] for the DNA atoms and TIP3P for water molecules. In classical MD simulations, the system is equilibrated with constant pressure and the structure at the last step is adopted for the initial structure of the FPMD simulation. Figure 3 shows the energy profile of the FPMD simulation in the micro-canonical case (NVE simulation). In this O(N) FPMD simulation, periodic boundary condition, single-zeta with polarization (SZP) basis set, Perdew-Burke-Ernzerhof (PBE)[36] exchange-correlation functional, non-self-consistent (NSC) technique with the Harris-Foulkes energy functional, the cutoff range of 16 bohr for L matrix, and the nu-

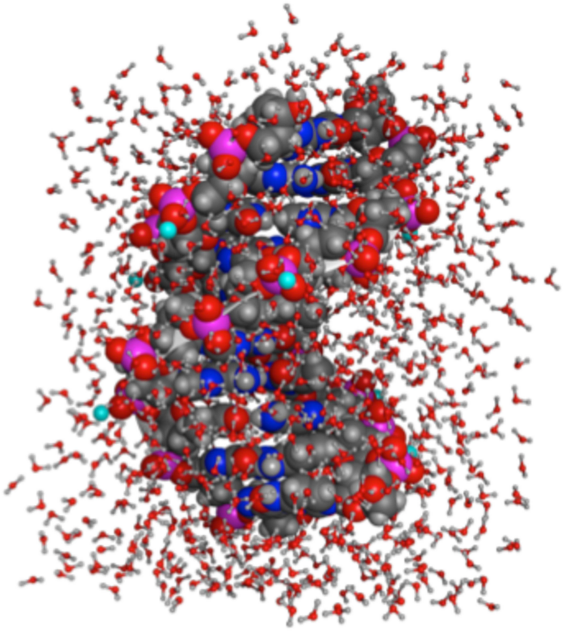

Fig. 2. Structure of a DNA decamer (PDB ID:1WQZ) hydrated with 934 water molecules, consisting of about 3400 atoms.

merical integration grid cutoff of 75 Ha are used. It should be noted that the SZP basis set used in the present MD simulations tends to show slightly larger error compared with DZP basis set. For example, the mean absolute error of the bond lengths in the adenine molecule is 0.05 Å using the SZP basis set, while it is 0.02 Å with DZP. However, we believe that the qualitative aspects reported in this paper are not affected by the choice of the basis set. As a dissipation term in the equation of motion for the X matrix, we consider the terms up to the order of 5. The time step is 0.5 fs and the initial temperature for the atomic velocity is 300K.

The most important point we can see from Fig. 3 is that the total energy, which is the sum of the potential energy (DFT total energy) and kinetic energy of nuclei, is constant during the simulation. This means that the present method is reliable also for this complex system. We can also see that the potential and the kinetic energies of the system both fluctuate relatively large in the early stage (up to about 50 MD steps). This probably shows a rapid response to the the differences of chemical bonds between classical force fields and the CONQUEST calculation with the present conditions. However, after about one hundred MD steps, these two energies show much slower changes. We do not clearly understand why we have such behaviour, but we expect that using the initial structure given by classical MD simulations would help to reduce

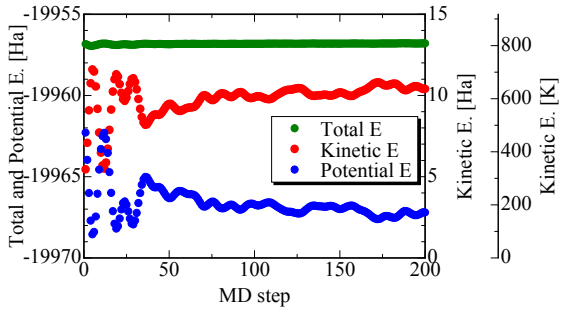

Fig. 3. Energy profile in the FPMD simulation of the hydrated DNA system (NVE simulation). Total energy (green), potential energy (blue), and kinetic energy (red) are shown.

the simulation time for the equilibration of FPMD simulations.

Next we investigate the temperature-controlled FPMD simulations by the velocity scaling method. Figure 4 shows the energy profiles of the FPMD simulation of the same hydrated DNA system, at the temperature of 300K (a) and 600K (b), respectively. In this method, we simply rescale the velocity at each MD step to make the average of the kinetic energy same as a given temperature, and use the corrected velocity in the equation of motion for nuclei. In Fig. 4, the profiles of the kinetic energy calculated with the velocities before the correction are plotted, together with the potential energy and total energy.

As can be seen in Fig. 4, although the fluctuation of the kinetic energy defined by the velocities before the correction is large in the early stages (0-50 MD steps), the change of the kinetic energy becomes very small after 50 MD steps. In both cases, i.e. at 300K and 600K, the kinetic energy is close to the correct temperature after around 120 steps, and the profile of the total energy is becoming flat. This rapid convergence should be useful to control the temperature in FPMD simulations. We expect that it is also possible to do stable micro-canonical (NVE) MD simulations around the given temperature after we employ the velocity scaling method for a short time. This may be a useful technique for FPMD on similar hydrated biological systems. We have also recently implemented another method to control the temperature, the Nose-Hoover-chain method. The stability of this method with the DMM+XL-BOMD scheme will be reported elsewhere. We believe these techniques will contribute to realise many efficient, reliable and accurate FPMD simulations on complex biological systems in the near future.

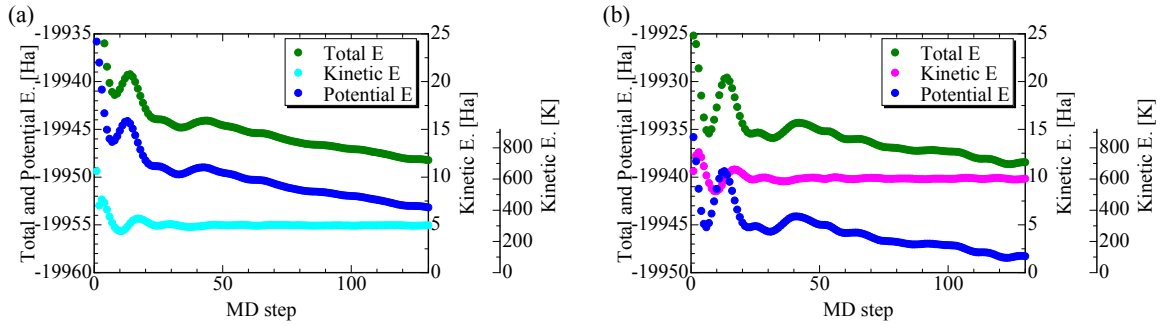

Fig. 4. Energy profile of the FPMD simulation on the hydrated DNA system (NVT simulation) with the velocity scaling method at 300K (a) and 600K (b); total energy (green), potential energy (blue), and kinetic energy (red) calcualted from the atomic velocities before the correction.

4. Summary

The linear-scaling or $\mathcal{O}(N)$ code CONQUEST has the ability to treat million-atom systems based on DFT. In this paper, we first gave an overview of the methods used in the code and introduced recent progress, especially the newly introduced method which combines the DMM and XL-BOMD methods to realise accurate and efficient FPMD simulations with the $\mathcal{O}(N)$ method.

Then, we demonstrated that the method can be also applied to a complex biological system, a hydrated DNA system, containing about 3400 atoms. We have shown that the total energy is conserved accurately in the micro-canonical simulations when the combined method (DMM+XL-BOMD) is applied. Furthermore, we found the velocity scaling method is useful to control the temperature of the FPMD simulation of this system. Based on these results, we can conclude that CONQUEST is ready to do reliable FPMD simulations on complex biological systems.

Acknowledgment

This work is partly supported by JSPS KAKENHI project (Grant Number 22790122, 26610120 and 26246021). We also acknowledge the Strategic Programs for Innovative Research (SPIRE) of MEXT and the Computational Materials Science Initiative (CMSI), Japan. The calculations were performed in part on the NIMS material numerical simulator and the COMA (PACS-IX) system of University of Tsukuba. The authors also acknowledge Prof. M. J. Gillan (University College London) for useful discussion and Dr. M. Arita (Tokyo University of Science) for his contribution to the implementaion of DMM+XL-BOMD method into CONQUEST.

References

1) For example, Novel prize of chemistry 2013, Martin Karplus, Michael Levitt, Arieh Warshel, "for the development of multiscale models for complex chemical systems".
2) Bowler D R and Miyazaki T 2012 *Reports on Progress in Physics* **75** 036503
3) http://www.linear-scaling.org/
4) Bowler D R and Miyazaki T 2010 *Journal of Physics: Condensed Matter* **22** 074207
5) Arita M, Arapan S, Bowler D R, and Miyazaki T 2014 *Journal of Advanced Simulation in Science and Engineering* **1** 87
6) Niklasson A M N 2008 *Phys. Rev. Lett.* **100** 123004
7) Arita M, Bowler D R, and Miyazaki T 2014 *Journal of Chemical Theory and Computation* **10** 5419
8) Hernández E and Gillan M J 1995 *Phys. Rev. B* **51** 10157
9) Bowler D R, Miyazaki T, and Gillan M J 2002 *J. Phys.:Condens. Matter* **14** 2781
10) Bowler D R, Choudhury R, Gillan M J, and Miyazaki T 2006 *Phys. Stat. Sol. (b)* **243** 989
11) Truflandier L A, Miyazaki T, and Bowler D R 2011 Eprint arXiv:1112.5989v2
12) Hernández E, Gillan M J, and Goringe C M 1997 *Phys. Rev. B* **55** 13485
13) Sankey O F and Niklewski D J 1989 *Phys. Rev. B* **40** 3979
14) Soler J M, Artacho E, Gale J D, García A, Junquera J, Ordejón P, and Sánchez-Portal D 2002 *J. Phys.:Condens. Matter* **14** 2475
15) Torralba A S, Todorović, M. Brázdová V, Choudhury R, Miyazaki T, Gillan M J, and Bowler D R 2008 *J. Phys.:Condens. Matter* **20** 294206
16) Rayson M J and Briddon P R 2009 *Phys. Rev. B* **80** 205104
17) Nakata A, Bowler D R, and Miyazaki T 2015 *Physical Chemistry Chemical Physics* **17** 31427
18) Nakata A, Bowler D R, and Miyazaki T 2014 *Journal of Chemical Theory and Computation* **10** 4813
19) Li X P, Nunes R W, and Vanderbilt D 1993 *Phys. Rev. B* **47** 10891
20) McWeeny R 1960 *Rev. Mod. Phys.* **32** 335
21) Bowler D R and Gillan M J 1999 *Comp. Phys. Commun.* **120** 95
22) Bowler D R and Gillan M J 2000 *Chemical Physics Letters* **325** 473

23) Harris J 1985 *Phys. Rev. B* **31** 1770

24) Foulkes W M C and Haydock R 1989 *Phys. Rev. B* **39** 12520

25) Otsuka T, Miyazaki T, Ohno T, Bowler D R, and Gillan M J 2008 *J. Phys.:Condens. Matter* **20** 294201

26) Miyazaki T 2013 Ultra-large-scale first-principles calculations by the k computer NIMS NOW Vol. 11(9), 6

27) Miyazaki T, Bowler D R, Choudhury R, and Gillan M J 2007 *Phys. Rev. B* **76** 115327

28) Miyazaki T, Bowler D R, Gillan M J, and Ohno T 2008 *J. Phys. Soc. Jpn.* **77** 123706

29) Miyazaki T, Bowler D R, Choudhury R, and Gillan M J 2004 *J. Chem. Phys.* **121** 6186

30) Torralba A S, Bowler D R, Miyazaki T, and Gillan M J 2009 *J. Chem. Theory Comput.* **5** 1499

31) Niklasson A M N, Tymczak C J, and Challacombe M 2006 *Phys. Rev. Lett.* **97** 123001

32) Souvatzis P and Niklasson A M N 2014 *The Journal of Chemical Physics* **140** 044117

33) Press W H, Flannery B P, Teukolsky S A, and Vetterling W T 1992 *Numerical Recipes in FORTRAN: The Art of Scientific Computing* Cambridge University Press, Cambridge, UK 2nd edition

34) Niklasson A M N, Steneteg P, Odell A, Bock N, Challacombe M, Tymczak C J, Holmström E, Zheng G, and Weber V 2009 *J. Chem. Phys.* **130** 214109

35) Wang J, Cieplak O, and Kollman P A 2000 *J. Comp. Chem.* **21** 1049

36) Perdew J P, Burke K, and Ernzerhof M 1996 *Phys. Rev. Lett.* **77** 3865