

Efficient calculation of electronic structure using $O(N)$ density functional theory

Ayako Nakata,^{*,†} Yasunori Futamura,^{*,‡} Tetsuya Sakurai,^{‡,¶} David R Bowler,^{§,||,⊥}
and Tsuyoshi Miyazaki^{†,§}

[†]*First-principles Simulation Group, Nano-Theory Field, International Center for Materials Nanoarchitectonics (MANA), National Institute for Materials Science (NIMS), 1-1 Namiki, Tsukuba, Ibaraki 305-0044, JAPAN*

[‡]*Department of Computer Science, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, Japan*

[¶]*CREST, Japan Science and Technology Agency, 4-1-8 Hon-cho, Kawaguchi, Saitama 332-0012, Japan*

[§]*Department of Physics & Astronomy, University College London, Gower St, London WC1E 6BT, UK*

^{||}*WPI-MANA, National Institute for Materials Science (NIMS), 1-1 Namiki, Tsukuba, Ibaraki 305-0044, Japan*

[⊥]*London Centre for Nanotechnology, University College London, 17-19 Gordon Street, London WC1H 0AH, UK*

E-mail: NAKATA.Ayako@nims.go.jp; futamura@cs.tsukuba.ac.jp

Abstract

We propose an efficient way to calculate the electronic structure of large systems by combining a large-scale first-principles density functional theory code, CONQUEST, and an efficient interior eigenproblem solver, the Sakurai–Sugiura method. The electronic

Hamiltonian and charge density of large systems are obtained by CONQUEST and the eigenstates of the Hamiltonians are then obtained by the Sakurai–Sugiura method. Applications to a hydrated DNA system, and adsorbed P₂ molecules and Ge hut clusters on large Si substrates demonstrate the applicability of this combination on systems with 10,000+ atoms with high accuracy and efficiency.

1 Introduction

First-principles density functional theory (DFT) is a powerful tool to investigate the atomic and electronic structures of molecules and condensed matter. However, most DFT calculations to date are limited in the system size which they can treat to about a thousand atoms, because the computational costs of conventional DFT calculations scale cubically to the numbers of atoms in target systems, N . This size limitation makes it difficult to treat complex systems, such as biomolecules, surfaces with defects or dopants with realistic concentrations, using conventional DFT methods.

In order to overcome this size limitation, we have developed a DFT code for large-scale systems, CONQUEST,^{1–3} which can routinely treat systems with more than 10,000 atoms (and has been shown to scale to million atom systems^{3,4}). There are two important techniques to achieve large-scale calculations in CONQUEST. One is the use of “support functions” to express Kohn-Sham density matrices:¹ these are real-space local orbitals having values only in finite spatial regions. The electronic Hamiltonian and overlap matrices in this support-function basis are thus sparse, and calculations with sparse matrices are highly efficiently parallelized in CONQUEST.⁵ The other technique is the linear-scaling, or $O(N)$, method based on the density matrix minimization (DMM) method.^{6–9} In this method, the energy is minimized with respect to an auxiliary density matrix,⁶ without performing diagonalization. The range of the auxiliary density matrix is restricted to a finite spatial region, based on the near-sightedness principle;¹⁰ the auxiliary density matrix is thus also sparse and the DMM calculations scale linearly with N .

$O(N)$ calculations which work with the density matrix implicitly integrate over energy, and produce only the sum of the occupied eigenvalues, and not any of the eigenstates, i.e. the one-electron wavefunctions (molecular orbitals and bands) of the system. We often want to know the individual eigenstates to analyse the electronic structure of the system, though generally within a relatively small energy range. These can be found efficiently from the converged ground-state Hamiltonian.

There are several methods to obtain eigenstates in the interior of the spectrum of a system, such as the shift-and-invert Lanczos method,¹¹ the Sakura–Sugiura method,¹² and the FEAST algorithm.¹³ These methods enable us to avoid full diagonalization of a matrix. The Sakurai–Sugiura method (SSM) was proposed by Sakurai and Sugiura¹² for solving interior generalized eigenproblems of sparse matrices. Several applications of SSM on electronic structure calculations have been reported: band structure calculations of a large system containing 10,000 silicon atoms, in conjunction with real-space finite-difference DFT;¹⁴ molecular orbital (MO) calculations of biomolecules with the fragment molecular orbital (FMO) method;¹⁵ and core-excited-state calculations using time-dependent DFT.¹⁶

The purpose of this study is to present arguments for combining CONQUEST and SSM. CONQUEST and SSM are well matched: the sparse electronic Hamiltonian can be efficiently constructed by CONQUEST and interior eigenproblem of this sparse matrix can be solved efficiently by SSM. In the next section, the theoretical background of CONQUEST and SSM are presented. In the third section, we demonstrate that the combination enables us to investigate the electronic structure of large systems containing several thousand or several hundred thousand atoms with DFT (hydrated DNA; P_2 molecules adsorbed on the Si(100) surface; and Ge hut clusters formed on the Si(001) surface). The final section gives our conclusions.

2 Theoretical background

2.1 CONQUEST

We briefly explain two key techniques in CONQUEST, the support functions¹ and the $O(N)$ method,⁶⁻⁹ to give context to the combination of CONQUEST with SSM.

2.1.1 Support functions

In DFT, the Kohn-Sham (KS) density matrix ρ is expressed as

$$\rho(\mathbf{r}, \mathbf{r}') = \sum_n f_n \psi_n(\mathbf{r}) \psi_n(\mathbf{r}')^*, \quad (1)$$

where ψ_n and f_n are n th KS orbital and its occupation number, respectively. This would necessitate cubic scaling, so instead, in CONQUEST, ρ is expanded using the support functions ϕ :

$$\rho(\mathbf{r}, \mathbf{r}') = \sum_{i\alpha, j\beta} \phi_{i\alpha}(\mathbf{r}) K_{i\alpha, j\beta} \phi_{j\beta}(\mathbf{r}')^*. \quad (2)$$

Here, \mathbf{K} is the density matrix in the support function basis, and i, j and α, β are the indices of atoms and support functions, respectively.

Support functions are constructed as linear combinations of given basis functions χ . CONQUEST supports two kinds of basis functions, B-spline (blip) finite element functions¹⁷ and pseudo atomic orbitals (PAO).¹⁸ The use of blip functions enables us to improve the accuracy systematically by decreasing the width and the grid spacing of the functions, with the grid spacing being directly equivalent to a plane-wave cutoff. The PAOs are atomic-orbital basis functions found from the pseudo-potentials and consist of radial functions (ξ) multiplied by spherical harmonic functions.^{19,20} Although the systematic improvement of the basis set completeness for PAOs is not straightforward, it can be improved by increasing the number of radial functions for each spherical harmonic (i.e., multiple- ξ functions) as well as adding extra angular momentum shells. The computational cost of minimisation with PAOs

is at present much lower than that with blip basis functions.

Since PAOs are intrinsically localized around atoms, we can use them as support functions without modification, which we call “primitive” support functions. On the other hand, we often contract the PAOs by taking linear combinations in order to reduce the number of support functions, because the computational cost increases cubically with the number of support functions *per atom* in CONQUEST, even with the $O(N)$ method. Conventionally, we take a linear combination of PAOs on each atom,

$$\phi_{i\alpha}(\mathbf{r}) = \sum_{\mu} c_{i\alpha,i\mu} \chi_{i\mu}(\mathbf{r}), \quad (3)$$

where \mathbf{c} are the linear-combination coefficients and $\chi_{i\mu}$ is the μ th PAO on atom i . \mathbf{c} is then optimized for all atoms in the target system. A similar contraction method was reported by Ozaki and Kino.^{21,22} Since the linear combinations are taken only over the PAOs which are centred at the target atom, the contracted “single-site” functions have to keep the point-group symmetry of the target atom.¹⁸

We have recently proposed “multi-site” support functions (MSSFs),^{23,24} which are linear combinations of PAOs on the target atom and its neighbouring atoms in a finite region (r_{MS}):

$$\phi_{i\alpha}(\mathbf{r}) = \sum_k^{\text{neighbors}} \sum_{\mu \in k} C_{i\alpha,k\mu} \chi_{k\mu}(\mathbf{r}), \quad (4)$$

where k runs over the neighbouring atoms of atom i within the region r_{MS} . The coefficients \mathbf{C} are optimized numerically.²⁴ The initial values of \mathbf{C} are found as localized occupied MOs using the local filter diagonalization (LFD) method.^{23,25,26} Because MSSFs are localised MO-like functions, we can use a number of MSSFs equivalent to a minimal basis. Therefore, by determining \mathbf{C} with the LFD method and subsequently by the numerical optimization, the MSSFs can reproduce the results of the primitive functions with high accuracy while significantly reducing the computational cost.

Because the PAOs are localized functions which are restricted to finite spatial regions,¹⁹

the support functions constructed from PAOS are also localized. Although more delocalized than single-site support functions, MSSFs are still localized within the region $(r_{\text{PAO}} + r_{\text{MS}})$. Therefore, the Hamiltonian and overlap matrices in the support-function basis are always sparse.

2.1.2 $O(N)$ calculations with DMM

CONQUEST supports two methods to optimize the electronic states: conventional diagonalization method, scaling as $O(N^3)$ and the DMM method, scaling as $O(N)$. In the diagonalization method, the density matrix \mathbf{K} in (2) is obtained as the sum of the outer products of the eigenvectors of the electronic Hamiltonian \mathbf{H} . On the other hand, in the DMM calculations, \mathbf{K} is found by optimizing the energy with respect to an auxiliary density matrix \mathbf{L} . \mathbf{L} is introduced to guarantee the weak idempotency of \mathbf{K} :^{1,8,9}

$$K = 3LSL - 2LSLSL, \quad (5)$$

where \mathbf{S} is the overlap matrix between support functions. The partial derivative of the total energy with respect to \mathbf{L}^1 is:

$$\frac{\partial E_{\text{tot}}}{\partial L_{\alpha\beta}} = \begin{aligned} & [6(SLH + HLS)_{\beta\alpha} \\ & -4(SLSLH + SLHLS + HLSLS)_{\beta\alpha}]. \end{aligned} \quad (6)$$

The energy minimization is performed with the following restriction:

$$L_{i\alpha,j\beta} = 0, \quad |\mathbf{R}_i - \mathbf{R}_j| > r_L, \quad (7)$$

where r_L corresponds to a chosen cutoff radius. This restriction is supported by Kohn’s “near-sightedness” principle.¹⁰ The effect of equation (7) is to make \mathbf{L} a localized, and thus sparse, matrix. Thus, all of the matrices in eq. (6) are sparse so that the matrix multiplications

in eq. (6) can be performed with $O(N)$ scaling. The total energy obtained by the $O(N)$ method with $r_L = \infty$ should be equal to the one obtained by diagonalization with infinitely fine Brillouin zone sampling.

Since diagonalization is not used to solve eq. (6), no eigenstates of the electronic Hamiltonian are calculated during the $O(N)$ calculation. It is usually difficult to apply conventional eigenproblem solvers to large-scale systems; however, in most practical applications, once the Hamiltonian is optimized, only a small number of eigenstates in a given energy region of interest are required. Any eigenstate can be found in linear scaling time (as the number of basis set coefficients per eigenstate scales linearly with system size). Therefore, an efficient approach to eigenstate solution such as SSM, is an ideal choice to pair with $O(N)$ calculations in CONQUEST.

2.2 Sakurai–Sugiura method

We now turn to a brief presentation of the Sakurai-Sugiura method. SSM is an eigenvalue solver which computes eigenvalues located in a specified contour path Γ on the complex plane and their associated eigenvectors. Variants of the method using the Rayleigh–Ritz procedure²⁷ and block versions^{28,29} of the method have been proposed. In this paper, we refer to these variants as the Sakurai–Sugiura method.

We denote the dimension of \mathbf{H} and \mathbf{S} as n . The key part of SSM is a set of M complex moments found as contour integrals over the contour Γ :

$$W_k := \frac{1}{2\pi i} \oint_{\Gamma} z^k (zS - H)^{-1} S V dz, \quad (8)$$

where V is an $n \times L_s$ random real matrix whose column vectors are linearly independent, L_s is greater than or equal to the maximum multiplicity of the eigenvalues inside Γ and k varies between 0 and $M - 1$ so that $W := [W_0, W_1, \dots, W_{M-1}]$. By Cauchy’s integral theorem, the column vectors of W become linear combinations of eigenvectors corresponding

to the eigenvalues located within Γ ; we write m_Γ as the number of eigenvalues (including multiplicity) inside Γ . Then W spans the same space as the eigen-subspace associated with the eigenvalues inside Γ , provided that $L_s M \geq m_\Gamma$. Either using the Rayleigh–Ritz procedure for $\text{Range}(W)$, or solving a small generalized eigenvalue problem of block Hankel matrices consist of $(SV)^T W_k$ ($k = 0, 1, \dots, 2M - 1$), we can compute all the eigenvalues within Γ and their corresponding eigenvectors.

In order to approximate the contour integral eq. (8), we apply numerical quadrature:

$$\hat{W}_k \approx W_k = \sum_{j=1}^{N_q} w_j \zeta_j^k (z_j S - H)^{-1} SV, \quad (9)$$

where N_q is the number of quadrature points, z_j is a quadrature point, w_j is a quadrature weight. The normalized quadrature point ζ_j is used instead of z_j in the k -th order calculation of eq. (9) for numerical stability. We define $\zeta_j = (z_j - \gamma_s)/\rho_s$ with real scalars γ_s and $\rho_s > 0$. For instance, when Γ is a circle, γ_s and ρ_s are the center and radius of the circle, respectively. In SSM, we need to solve linear systems (with multiple right-hand sides)

$$(z_j S - H)Y_j = SV \quad (j = 1, 2, \dots, N_q) \quad (10)$$

for computing \hat{W}_k .

There are three levels of parallelism in the SSM algorithm. First, eq. (10) can be solved independently for each value of j . Second, each of these linear systems can be solved in a distributed manner. Finally, we can divide the energy range of interest into multiple regions, and the eigenvalues in each region can be computed in parallel. More details for the implementation and an efficient way of setting the parameters of SSM are described in Ref. 30.

3 Practical applications

In order to demonstrate the usefulness of the present combination of CONQUEST and SSM, we results from calculations on three systems: a hydrated DNA system; P_2 molecules adsorbed on the Si(100) surface; and Ge hut clusters on the Si(001) surface. Since the hydrated DNA system and P_2 on Si(100), which consist of several thousand atoms, can be treated with conventional diagonalization when using MSSFs, the SCF calculations in CONQUEST were performed with diagonalization in these cases, allowing us to check the accuracy of the results from the SSM. For Ge clusters on Si surfaces, which consist of 10,000+ atoms, structural optimisation was performed with non-self-consistent $O(N)$ calculations, leading to the Hamiltonian for the relaxed system. In the CONQUEST calculations, PBE96 generalized-gradient-approximation (GGA) exchange-correlation functional³¹ and PAOs with norm-conserving pseudo potentials, which were generated with the SIESTA code,²⁰ were used.

3.1 Hydrated DNA

First, we investigate the accuracy of SSM by comparing with results from exact diagonalisation, using the ScaLAPACK routine PDSYGV on the CONQUEST matrices. We calculated the eigenstates of a hydrated DNA system, the B-DNA decamer 5'-d(CCATTAATGG)₂-3' which contains 634 atoms in the DNA with 932 hydrating water molecules and 9 Mg counter ions, totalling 3,439 atoms. More detailed information for the geometry used is given in Ref. 32. In CONQUEST, the electronic structure of the system was optimized self-consistently with the MSSFs ($r_{MS} = 8.0$ bohr). For C, N, O and P atoms, DZP (2s2p1d) PAOs³³ were contracted into four support functions using the most delocalized s and p PAOs as trial vectors. For H and Mg, DZP (2s1p) PAOs³³ were contracted to one MSSF using the most delocalized s PAOs as trial vectors. The dimension of the Hamiltonian is reduced from 27,883 with DZP PAOs to 4,774 with MSSFs. The eigenvalues and eigenvectors of the optimized Hamiltonian were obtained by both SSM and the ScaLAPACK routine to investigate both

the efficiency and accuracy of SSM. In the SSM calculation, we calculated the eigenstates in the energy range $[-0.4:0.4]$ Hartree with an interval of 0.01 Hartree, giving 6093 states in total. The parameters N_q (number of quadrature points), M (number of complex moments), and L_s (number of columns of the source matrix) are set to 16, 8 and 64, respectively.

Table 1 and Figure 1 shows the calculated eigenvalues and eigenvectors around the Fermi level (-0.215 eV), which correspond to the energies and charge distributions of eight MOs around the HOMO-LUMO gap. Table 1 shows that the difference in eigenvalues calculated by SSM and those by Scalapack are negligibly small, less than 10^{-9} eV. The calculated HOMO-LUMO gap is 2.16 eV, which is smaller than typically-reported values for DNA systems (about 3 - 4 eV³⁴). This underestimation is likely due to the use of a GGA functional.

Table 1: Calculated eigenvalues of hydrated DNA systems [eV].

	SSM	Scalapack
HOMO-3	-1.7330847154	-1.7330847153
HOMO-2	-1.5343553564	-1.5343553563
HOMO-1	-1.4458947736	-1.4458947736
HOMO	-1.2563299881	-1.2563299879
LUMO	0.9071560149	0.9071560148
LUMO+1	0.9438977187	0.9438977188
LUMO+2	0.9673219957	0.9673219956
LUMO+3	0.9725452608	0.9725452608
HOMO-LUMO gap	2.1634860030	2.1634860027

We can also see that the molecular orbitals (MOs) calculated by SSM and Scalapack are very similar to each other in Figure 1. The figure clearly shows that the MOs of the hydrated DNA system around the HOMO-LUMO gap are quite localized. We suggest that the present method (i.e. the combination of CONQUEST and SSM) can be a powerful tool to investigate charge-transfer properties of DNA systems.

We also investigate the ability of the present method to improve the accuracy of the unoccupied states calculated with MSSFs. Figure 2 shows the density of states (DOS) of the hydrated DNA system calculated with the MSSFs (blue line in Figure 2a). The results with the primitive DZP PAOs (Figure 2b) are also shown for comparison. The difference between

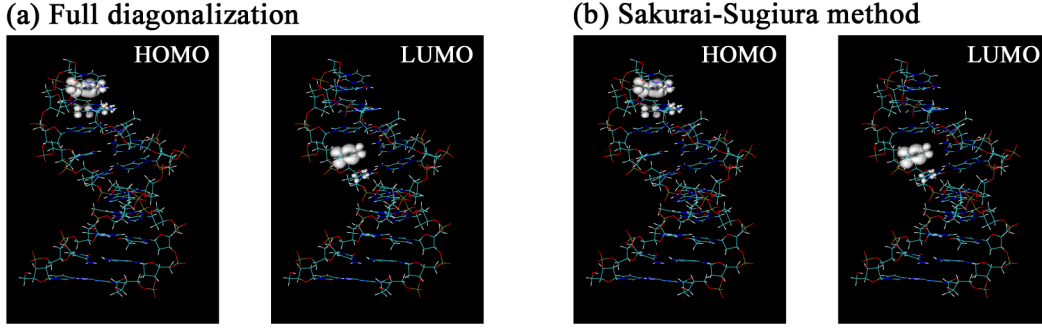


Figure 1: Molecular orbital pictures of the hydrated DNA calculated by (a) full diagonalization by Scalapack and (b) Sakurai-Sugiura method.

these results is shown in Figure 2c with the blue line, where it can clearly be seen that the difference in the occupied states is very small (the fractional difference is around 0.001) but the difference becomes larger in higher-lying unoccupied states, as already reported in Ref. 23,24. This low accuracy in unoccupied states comes from the minimal size of the MSSF basis, and the optimisation process, which only accounts for the occupied states. This error will be critical when we consider properties related to the unoccupied states such as excitation energies, although for most calculations it is only the accuracy of the occupied states that is important.

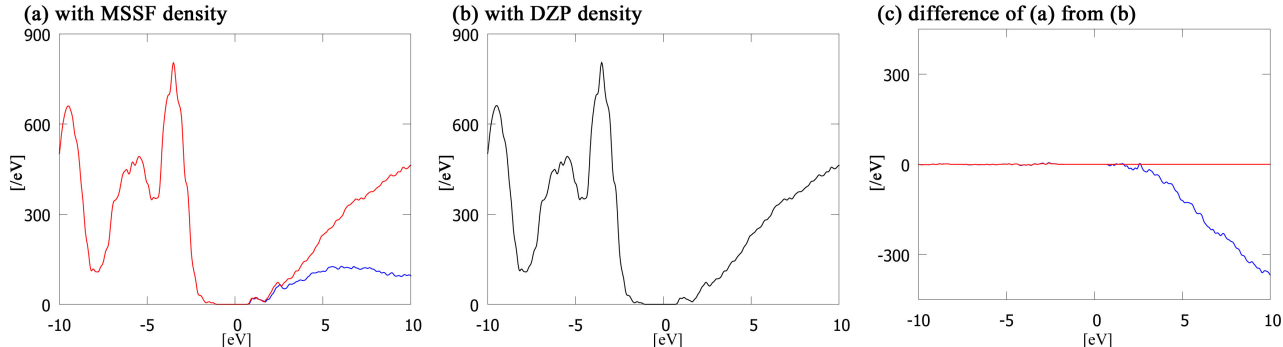


Figure 2: Density of states of hydrated DNA obtained with (a) multi-site support functions (MSSF) (blue) and DZP PAOs using MSSF charge density (red) and (b) DZP PAOs using DZP SCF charge density (black). r_{MS} is set to be 8.0 bohr. The difference of (a) MSSF (blue) and DZP with the MSSF density (red) from (b) DZP is also shown in (c).

The accuracy of the unoccupied states can be improved by the following procedure: (i) optimize the charge density by SCF calculations with MSSFs; (ii) build the Hamiltonian

using the primitive PAO basis, but with the optimized MSSF charge densities; and (iii) perform one-shot SS calculations to obtain the eigenvalues and eigenstates of the Hamiltonians. Because the dimension of the reconstructed Hamiltonian in the primitive PAO basis is much larger (27,883 for the DNA system) than the Hamiltonian in MSSF basis (7,447), diagonalization in the primitive PAO basis is often difficult even if is possible in the MSSF basis. The use of SSM (procedure (iii)) can then solve this problem. In Figure 2 we also show the reconstructed results: the red line in Figure 2a) is very similar to that of the primitive PAOs (Figure 2b). Figure 2c shows the difference from the primitive PAO results, which is almost zero throughout the eigenspectrum.

3.2 P_2 molecules on Si(100) surfaces

In this section, we demonstrate STM simulations of large Si(100) surface samples with adsorbed P_2 molecules. Simulations of STM images and STS spectra based on DFT calculations have an important role to play in understanding the atomic and electronic structure of surfaces and adsorbates, particularly in collaboration with experiments. In these simulations, we need only the eigenstates in limited energy regions around the Fermi levels because STM and STS measurements are only consider biases up to a few eV. However, due to the limitation of the system size in DFT simulations, it has until now been difficult to treat large and complex surface structures.

P_2 molecules adsorbed on Si(001) tend to form two key structures, either bridging two Si dimers along the Si dimer row (P_2 (I)), or just above a Si dimer (P_2 (II)).³⁵ We modelled the P_2 (I) and P_2 (II) structures on large areas substrate, with ten layers of silicon, each layer consisting of 8 dimer rows each containing 16 Si dimers. The lowest layer is terminated with H atoms. Thus, each system consists in total of 3,074 atoms (2,560 Si, 512 H and 2 P atoms). The geometry was optimized by fixing the bottom two Si layers and H atoms. The MSSFs ($r_{MS} = 17.0$ bohr) were used for the optimization. For Si and P atoms, four MSSFs are constructed from TZDP (3s3p2d) PAOs³⁶ using the most delocalized s and p PAOs as trial

vectors. The MSSFs for H atoms were constructed from DZP PAOs³³ as in Section 3.1. The optimized geometries are shown in Figure 3(a). As reported in Ref. 35, P₂(I) is calculated to be more stable than P₂(II) (by 0.92 eV in the present calculations).

In order to obtain the theoretical STM images corresponding to the experimental results,³⁵ the eigenvectors only in the regions of the experimental sample biases, -1.5 eV and +1.4 eV, are required. Therefore, we calculated the eigenstates from -1.5 eV to +1.4 eV around the Fermi energy (e_F) with SSM. STM images were generated using the Tersoff-Hamann approach.³⁷ In order to improve the accuracy of the unoccupied states, as in Section 3.1, we reconstructed the Hamiltonians in primitive DZP PAO basis with the MSSF charge densities, and obtained the eigenstates by SSM subsequently. Figures 3(b) and 3(c) present the theoretical STM images corresponding to the sample biases -1.5 eV and +1.4 eV, which reproduce the experimental images³⁵ very well. We would like to emphasise that only about 800 eigenstates in the interval [$e_F - 1.5$ eV: $e_F + 1.4$ eV] were needed (precisely, 776 states for P₂(I) and 778 states for P₂(II)), instead of the whole 58,924 eigenstates of the Hamiltonian which would have been required by conventional full diagonalization. These results suggest that our method will enable us to simulate the electronic structure of complex surfaces containing dopants and defects with realistic concentrations.

Table 2: Comparison of the computation times (in second) of ScaLAPACK’s PDSYGVX and SSM for the P₂ on Si(001) system. n_p indicates the number of computational nodes.

n_p	4	8	16	32	64
PDSYGVX	5,064	3,032	1,591	1,043	629
SSM	924	469	239	134	85

It is also instructive to compare the computation time for SSM and ScaLAPACK’s PDSYGVX subroutine for the system P₂(I). The computational environment used for this test was the supercomputer COMA (PACS-IX) at University of Tsukuba. Each node has Two Intel Xeon E5-2670v2 CPUs, each with 10 cores, and a clock frequency of 2.5 GHz. While COMA has Intel Xeon Phi accelerators, we used only the CPUs for the performance evaluation. For the parameters of SSM, we set $N_q = 16$, $M = 8$, and $L_s = 100$. We di-

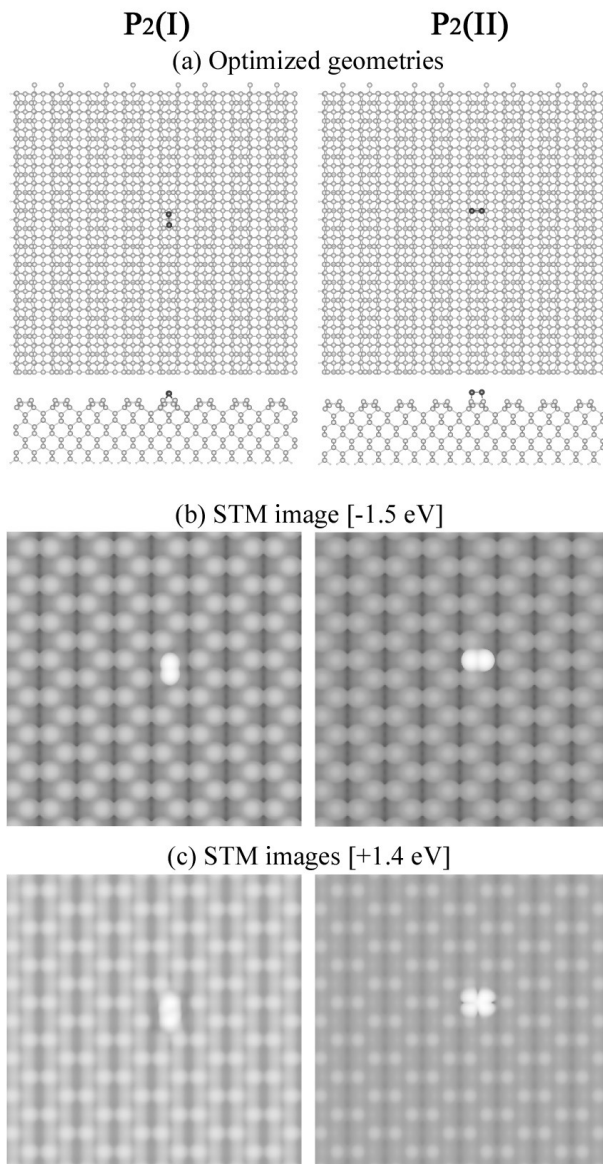


Figure 3: P₂ molecules on Si(100) surfaces: (a) Top and side views of the optimized geometries (Si, P and H atoms are colored in black, grey and light grey), (b) STM images at biases -1.5 eV and (c) STM images at biases +1.4 eV.

vided the interval into four equal parts. For the numerical quadrature, we used Chebyshev points in the interval $[-1, 1]$ for ζ_j and the associated barycentric weights for w_j in the way presented in Ref. 38. We set γ_s and ρ_s as the center and half width of the target energy range. As the setting of multi-processing for SSM, we run four MPI processes per node and use five OpenMP threads per MPI process. For PDSYGVX, we run one MPI processes per node and use 20 OpenMP threads per MPI process. Table 2 shows the comparison of the computational times with SSM and PDSYGVX subroutine. The table indicates that for all cases, SSM is at least five times faster than PDSYGVX, and more than seven times faster for 32 and 64 nodes. This significant speedup comes from the high parallel efficiency of SSM, which takes advantage of the sparsity of the matrices.

3.3 Ge hut clusters on Si(001) surfaces

Finally in this section, we consider systems that are out of reach of conventional DFT methods: Ge hut clusters on the Si(001) surfaces (Ge/Si(001)) consisting of (a) 23,737 and (b) 194,573 atoms. The optimized geometry of system (a) is shown in Figure 4. 8×13 and 20×30 Ge hut-shaped clusters were placed on 14×19 and 29×39 SiGe substrates in systems (a) and (b), respectively. The dimensions of \mathbf{H} and \mathbf{S} matrices for system (a) and (b) are 213,633 and 1,751,157 with SZP PAOs. These dimensions are beyond what could be addressed with conventional exact diagonalization. The geometry optimization of systems (a) and (b) was performed with our $O(N)$ method; the detail of the geometry optimizations will be given in a forthcoming paper.³⁹ We used the PZ81 LDA exchange-correlation functional⁴⁰ and SZP PAOs.⁴¹

By combining CONQUEST and SSM, we can now explore not just the atomic structure but also the electronic structure of these systems. Figure 4(b) shows the sum of eigenstates in the energy window $[-0.01: +0.02]$ eV around Fermi level (summation of 43 bands). From the figure we can see that the electronic charge around the Fermi level is localized at the surface of the Ge hut cluster. Although the detailed discussion of the atomic and electronic

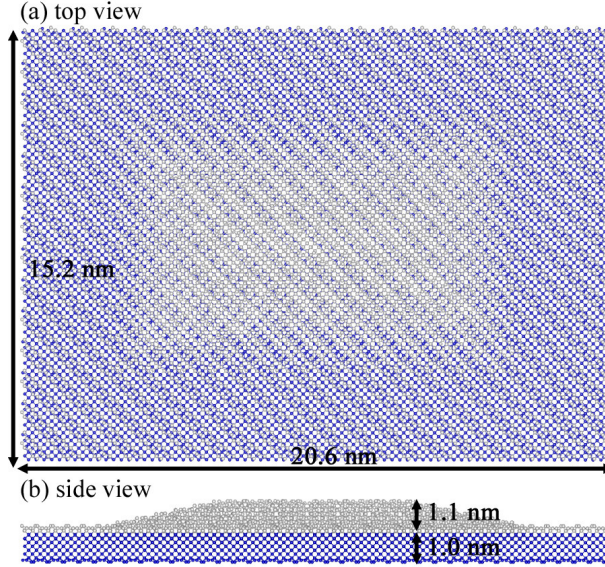


Figure 4: Top and side views of the optimized geometry of the Ge hut clusters on the Si(001) (23,737 atoms).

structures of the systems, which will be investigated in a future publication,³⁹ goes beyond the purpose of this paper, the result is promising as an example of the electronic structure of one of the largest systems treated by first-principle DFT calculations.

To indicate the computation time required, we show the time for SSM calculations on both systems (a) and (b) in Table 3. For the parameters of SSM, we set $N_q = 16$, $M = 8$, and $L_s = 64$. We used the K computer at RIKEN Advanced Institute for Computational Science. The K computer has SPARC64TM VIIIfx CPUs, with eight cores per compute node. We used 64 nodes and 6400 nodes of the K computer for systems (a) and (b), respectively. We run four MPI processes per node and use 8 OpenMP threads per MPI process.

Table 3: The computation times of SSM for Ge hut cluster systems (a) and (b). n_p indicates the number of computational nodes.

	# of atoms	n_p	Time [sec]
System (a)	23,737	64	146
System (b)	194,573	6,400	2,399

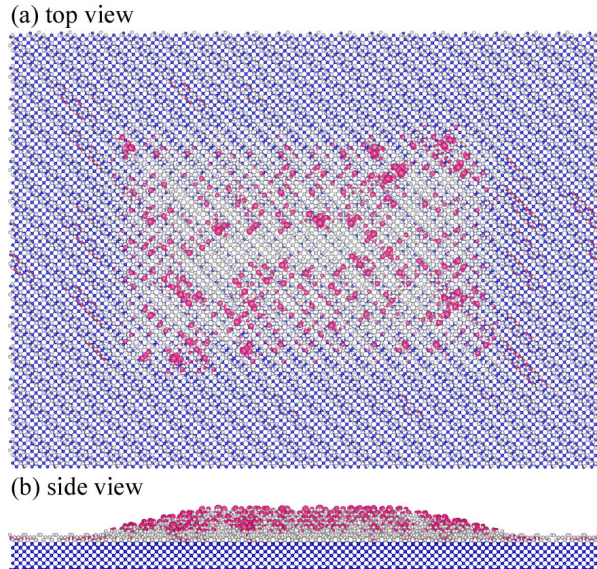


Figure 5: Top and side views of the electronic density distributions of the Ge hut clusters on the Si(001) (23,737 atoms) around Fermi level.

4 Conclusions

We have shown that the combination of CONQUEST and the Sakurai-Sugiura method (SSM) is a powerful tool to analyze the electronic structure of very large systems. CONQUEST can perform self-consistent energy minimization and geometry optimization of very large systems based on first-principles DFT by using local orbital (support) functions and a linear-scaling method. The use of the support functions makes the overlap and electronic Hamiltonian matrices sparse. The recently-developed multi-site support function method enables us to reduce the number of support functions to a minimal basis set. However, there are two key problems in analyzing the electronic structure of large systems with CONQUEST: the multi-site support function method gives high accuracy for occupied states but not for unoccupied states; and the linear-scaling method in CONQUEST does not provide the eigenvalues or eigenstates of the Hamiltonian. SSM is a method to solve interior eigenproblems of large sparse matrices. SSM can provide the eigenvalues and eigenvectors in a finite eigenvalue range using contour integrals in complex planes, without performing diagonalization. The combination of these methods is a powerful new tool.

We have shown that: i) the combined method is perfectly accurate in the case of a hydrated DNA system; ii) the accuracy of the unoccupied states found with multi-site support functions can be greatly improved; and iii) the method can be used to perform STM and STS simulations of complex surface structures such as P_2 molecules on Si(001). We have also shown that we can analyse the electronic structure of systems over 10,000 atoms, considering two Ge hut clusters on Si(001) surfaces with 23, 737 and 194, 573 atoms.

Our new method enables us to analyze the atomic and electronic structure of very large systems accurately and efficiently.

Acknowledgement

We gratefully acknowledge the discussion about the geometries of the DNA system, Si surfaces with P_2 and the hut cluster systems with Dr. Takao Otsuka, Dr. Keisuke Sagisaka and Dr. Sergiu Arapan, respectively. This work was supported by JSPS Grant-in-Aid for Scientific Research: Grant Number 15H01052, 25286097 and 26246021, Japan. This work was partly supported by "World Premier International Research Center Initiative (WPI Initiative) on Materials Nanoarchitectonics" and "Exploratory Challenge on Post-K computer" (Frontiers of Basic Science: Challenging the Limits) by MEXT, Japan. Calculations were performed in the Numerical Materials Simulator at NIMS. This research used computational resources of the K computer provided by the RIKEN Advanced Institute for Computational Science through the HPCI System Research project (Project ID:hp120170, hp140069, hp150248, hp160138). This research in part used computational resources of COMA (PACS-IX) provided by Interdisciplinary Computational Science Program in Center for Computational Sciences, University of Tsukuba.

References

- (1) Hernández, E.; Gillan, M. J.; Goringe, C. M. *Phys. Rev. B* **1996**, *53*, 7147–7157.

- (2) Bowler, D. R.; Choudhury, R.; Gillan, M. J.; Miyazaki, T. Recent progress with large-scale ab initio calculations: The CONQUEST code. *Phys. Status Solidi Basic Res.* 2006; pp 989–1000.
- (3) Bowler, D. R.; Miyazaki, T. *J. Phys. Condens. Matter* **2010**, *22*, 074207.
- (4) Arita, M.; Arapan, S.; Bowler, D. R.; Miyazaki, T. *J. Adv. Simul. Sci. Eng.* **2014**, *1*, 87–97.
- (5) Bowler, D.; Miyazaki, T.; Gillan, M. *Comput. Phys. Commun.* **2001**, *137*, 255–273.
- (6) Hernández, E.; Gillan, M. J. *Phys. Rev. B* **1995**, *51*, 10157–10160.
- (7) Bowler, D. R.; Gillan, M. J. *Comput. Phys. Commun.* **1999**, *120*, 95–108.
- (8) McWeeny, R. *Rev. Mod. Phys.* **1960**, *32*, 335–369.
- (9) Li, X.-P.; Nunes, R. W.; Vanderbilt, D. *Phys. Rev. B* **1993**, *47*, 10891–10894.
- (10) Kohn, W. *Phys. Rev. Lett.* **1996**, *76*, 3168–3171.
- (11) Thomas Ericsson, A. R. *Mathematics of Computation* **1980**, *35*, 1251–1268.
- (12) Sakurai, T.; Sugiura, H. *J. Comput. Appl. Math.* **2003**, *159*, 119–128.
- (13) Polizzi, E. *Phys. Rev. B* **2009**, *79*, 115112.
- (14) Futamura, Y.; Sakurai, T.; Furuya, S.; Iwata, J. I. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* **2013**, *7851 LNCS*, 226–235.
- (15) Umeda, H.; Inadomi, Y.; Watanabe, T.; Yagi, T.; Ishimoto, T.; Ikegami, T.; Tadano, H.; Sakurai, T.; Nagashima, U. *J. Comput. Chem.* **2010**, *31*, 2381–2388.
- (16) Tsuchimochi, T.; Kobayashi, M.; Nakata, A.; Imamura, Y.; Nakai, H. *J. Comput. Chem.* **2008**, *29*, 2311–2316.

- (17) Hernández, E.; Gillan, M. J.; Goringe, C. M. *Phys. Rev. B* **1997**, *55*, 13485–13493.
- (18) Torralba, A. S.; Todorović, M.; Brázdová, V.; Choudhury, R.; Miyazaki, T.; Gillan, M. J.; Bowler, D. R. *J. Phys. Condens. Matter* **2008**, *20*, 294206.
- (19) Sankey, O. F.; Niklewski, D. J. *Phys. Rev. B* **1989**, *40*, 3979–3995.
- (20) Soler, J. M.; Artacho, E.; Gale, J. D.; García, A.; Junquera, J.; Ordejón, P.; Sánchez-Portal, D. *J. Phys. Condens. Matter* **2002**, *14*, 2745–2779.
- (21) Ozaki, T. *Phys. Rev. B* **2003**, *67*, 155108.
- (22) Ozaki, T.; Kino, H. *Phys. Rev. B* **2004**, *69*, 195113.
- (23) Nakata, A.; Bowler, D. R.; Miyazaki, T. *J. Chem. Theory Comput.* **2014**, *10*, 4813–4822.
- (24) Nakata, A.; Bowler, D.; Miyazaki, T. *Phys. Chem. Chem. Phys.* **2015**, *17*, 31427–31433.
- (25) Rayson, M. J.; Briddon, P. R. *Phys. Rev. B - Condens. Matter Mater. Phys.* **2009**, *80*, 205104.
- (26) Rayson, M. J. *Comput. Phys. Commun.* **2010**, *181*, 1051–1056.
- (27) Sakurai, T.; Tadano, H. *Hokkaido Math. J.* **2007**, *36*, 745–757.
- (28) Ikegami, T.; Sakurai, T.; Nagashima, U. *J. Comput. Appl. Math.* **2010**, *233*, 1927–1936.
- (29) Ikegami, T.; Sakurai, T. *Taiwanese J. Math.* **2010**, *14*, 825–837.
- (30) Sakurai, T.; Tadano, H.; Futamura, Y. *J. Algo. Comput. Tech.* **2013**, *7*, 249–269.
- (31) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.
- (32) Otsuka, T.; Miyazaki, T.; Ohno, T.; Bowler, D. R.; Gillan, M. *J. Physics-Condensed Matter* **2008**, *20*, 294201.

- (33) The range of two s, two p and a d PAOs are (4.6, 3.4), (5.7, 3.7) and (5.7) bohr for C, (4.1, 2.9), (5.0, 3.1) and (5.0) bohr for N, (3.7, 2.5), (4.6, 2.6) and (4.6) bohr for O and (5.2, 4.0), (6.5, 4.6) and (6.5) bohr for P. Those of two s and a p PAOs are (5.5, 4.0) and (5.5) bohr for H and (7.7, 6.5) and (7.7) bohr for Mg.
- (34) Apalkov, V.; Wang, X.-F.; Chakraborty, T. *Physics Aspects of Charge Migration Through DNA*; Springer Berlin Heidelberg, 2007; pp 77–119.
- (35) Sagisaka, K.; Marz, M.; Fujita, D.; Bowler, D. *Phys. Rev. B - Condens. Matter Mater. Phys.* **2013**, *87*, 155316.
- (36) The range of three s, three p and two d PAOs are (8.5, 6.6, 4.0), (8.5, 6.6, 4.0) and (8.5, 6.6) bohr for Si and (7.0, 5.6, 4.0), (7.0, 5.6, 4.0) and (7.0, 5.0) bohr for P.
- (37) Tersoff, J.; Hamann, D. R. *Phys. Rev. B* **1985**, *31*, 805–813.
- (38) Austin, A. P.; Trefethen, L. N. *SIAM Journal on Scientific Computing* **2015**, *37*, A1365–A1387.
- (39) Arapan S.; Bowler D. R.; Miyazaki T. A linear scaling DFT study of the growth of a new {105} facet layer on a Ge hut cluster. arXiv:1510.00526.
- (40) Perdew, J. P.; Zunger, A. *Phys. Rev. B* **1981**, *23*, 5048–5079.
- (41) The range of s, p and d PAOs are 4.4, 5.3 and 5.3 bohr for Si and 4.3, 5.4 and 5.4 bohr for Ge.