Computer Physics Communications 128 (2000) 67–92

Computer Physics Communications

www.elsevier.nl/locate/cpc

Linear-scaling parallel algorithms for the first principles treatment of metals ☆

Stuart C. Watson, Emily A. Carter *

Department of Chemistry and Biochemistry, Box 951569, University of California, Los Angeles, CA 90095-1569, USA

Abstract

Orbital-free kinetic energy density functionals provide the means to accurately simulate the behavior of simple *sp*-band metals, on a length scale currently unparalleled by other *ab initio* methods. We present some technical aspects of the efficient parallel implementation of kinetic energy functionals, including a functional with a recently developed density-dependent response kernel that provides a good description of metal surfaces. We further illustrate the ability of this new functional to treat finite metallic systems by examining the metal-insulator transition in a 2-dimensional array of metal quantum dots.

PACS: 71.15.Mb; 73.61.-r; 61.46.+w

1. Introduction

There is much academic and commercial interest in the physical and chemical properties of bulk metals, metal clusters, and metal surfaces. By characterizing the intrinsic attributes of the metallic bond, one gains insight that may be used in the prediction and fabrication of new and improved materials. Metal surfaces and clusters are traditionally of great interest for their magnetic, electronic, and catalytic properties. The recent fabrication of single-electron devices or quantum dots, from metal nano-clusters [1], has produced a great deal of attention. The current paper applies a linear-scaling parallel method to explore such metallic quantum dot arrays.

Many of the phenomena associated with the intrinsic properties of metals are long-range and large-scale. Surface reconstructions of the noble metals are known to extend over hundreds and thousands of unit cells [2]. Metal quantum dots typically have radii of 2 to 7 nm, which can involve hundreds to tens of thousands of atoms. Meaningful investigation of such systems requires both high accuracy (to capture the essential physics), and great efficiency (to be able to investigate the length scales necessary). Traditional *ab initio* methods [3,4] contain the accuracy desired, but are invariably very computationally expensive. Schemes possessing the smallest possible operation count dependence on the system size, preferably linear-scaling, are necessary to study such large systems.

There are difficulties associated with *ab initio* treatment of metals. The itinerant nature of the metallic electrons produces correlation length scales almost impossible to treat within accurate quantum mechanics methods. Density

☆ This paper is published as part of a thematic issue on Parallel Computing in Chemical Physics.

* Corresponding author. E-mail: eac@chem.ucla.edu

0010-4655/00/$ – see front matter
PII: S0010-4655(00)00064-3

functional theory has had some success [4,5]. However, standard orbital-based (i.e. Kohn–Sham DFT) techniques come with added complications over and above those of semiconductors and insulators (such as intensive **k**-point sampling, and charge-"sloshing" [6]). These problems, as well as the large length scales intrinsic to a metal, make the transition to a linear-scaling scheme a difficult undertaking.

The standard path to linear-scaling is to take advantage of some locality within the system, and the sparsity that this produces in derived properties. A recent review of the subject [7] outlined the diversity and complexity associated with many of the methods, which are based upon a relatively simple premise. All properties that may be derived from an electronic structure calculation are expressible in terms of the single-particle density matrix,

$$D(\mathbf{r},\mathbf{r}') = \sum_{\alpha}^{\text{occ.}} \psi_\alpha^*(\mathbf{r})\psi_\alpha(\mathbf{r}'). \tag{1}$$

For an insulator or semiconductor, the elements of this matrix decay exponentially with the separation $|\mathbf{r}-\mathbf{r}'|$ [7,8]. The asymptotic limit of this decay,

$$D(\mathbf{r},\mathbf{r}') \propto \mathrm{e}^{-\gamma|\mathbf{r}-\mathbf{r}'|}, \tag{2}$$

has an exponent that is proportional to the direct band-gap (Δ), the crystal lattice constant (a), and the electron mass (m),

$$\gamma = \frac{a\Delta m}{\hbar^2}, \tag{3}$$

in the limit of the weakly-bound electron. The decay in the tightly-bound limit is also exponential, but with an ill-determined exponent [8].

Linear-scaling is achieved by neglecting terms below a certain magnitude, which implicitly defines a localization region (or alternatively, a localization region is explicitly defined, with an implicit neglection of small terms). It is the choice of methods for obtaining, and then exploiting this localization, which results in the many different $\mathcal{O}(N)$ methods available.

Within a metal, the localization properties are different than in systems with a band-gap. At finite electronic temperatures, the metallic density matrix has an asymptotic decay which is also exponential [7–9], but with a different exponent,

$$D(\mathbf{r},\mathbf{r}') \propto k_\mathrm{F}\frac{\cos(k_\mathrm{F}|\mathbf{r}-\mathbf{r}'|)}{|\mathbf{r}-\mathbf{r}'|^2}\exp\left(-c\frac{k_\mathrm{B}T}{k_\mathrm{F}}|\mathbf{r}-\mathbf{r}'|\right), \tag{4}$$

where c, a constant, is of order 1. Most metals have a Fermi vector (k_F) of the order 0.1–1.0 a.u. Since the Boltzmann constant (k_B) is 3.1667×10^{-6} in atomic units, the exponential decay is unexploitable for all but the highest temperatures, significantly higher than 10^4 K. One example of a method which does exploit this high temperature locality is the energy renormalization group method [10,11]. An expansion of the density operator is made about a high temperature (significantly higher than 10^4 K). Quasi-linear-scaling is achieved by treating corrections from decreasing temperature, with increasing coarseness. While the scaling is beneficial ($N(\ln N)^2$), results presented so far have been for limited dimensionality tight-binding models.

It is clear that metals require a non-traditional approach in order to achieve the linear-scaling desired. Examining the historical evolution of *ab initio* methods reveals a solution from the origins of density functional theory.

2. Linear-scaling density functional theory for metals

The orbital-free kinetic energy density functional (OF-KEDF) method arises from the original form of the Hohenberg–Kohn theory, which is at the heart of density functional theory.

2.1. General density functional theory

Hohenberg–Kohn theory [12] states that the physical properties of a system can be derived purely in terms of the ground-state electronic density, $\rho(\mathbf{r})$, a 3-dimensional, real object. For a given number of electrons, and a fixed external potential (arising from nuclei, or in the case of pseudopotentials and effective-core-potentials, an effective combined nuclear/core-electron potential), there is a direct mapping between the potential and the ground-state wavefunction, and also between the ground-state wavefunction and the density. The ground-state energy is a variational minimum with respect to the electronic density [4,5,12].

By considering the nature of the electrons, one can subdivide the electronic energy into kinetic and potential terms,

$$E^{\text{elec.}} = T^{\text{elec.}} + U^{\text{elec.}}. \tag{5}$$

The potential energy ($U^{\text{elec.}}$) can be further divided into the effect of the external potential, a classical Coulomb (or Hartree) interaction, a correction to this term which takes into account the quantum mechanical Pauli exchange, and the "correlation" energy,

$$U^{\text{elec.}} = E^{\text{Ext.}} + E^{\text{Har.}} + E^{\text{X}} + E^{\text{C}}. \tag{6}$$

This final term is defined as the difference between the Hartree–Fock (HF) energy, and the "exact" energy, and corrects for the mean-field nature of the Hartree expression. It also includes terms accounting for the kinetic energy of the interacting electrons [4,5].

The energy arising from the external potential is a simple expression,

$$E^{\text{Ext.}} = \int \mathrm{d}\mathbf{r}\, \rho(\mathbf{r}) V^{\text{Ext.}}(\mathbf{r}), \tag{7}$$

with an equally trivial potential acting upon the density (the potential being the functional derivative of the energy with respect to the density [4]),

$$\frac{\delta E^{\text{Ext.}}}{\delta \rho(\mathbf{r})} = V^{\text{Ext.}}(\mathbf{r}). \tag{8}$$

The construction of this potential from atom-centered functions depends upon the exact form of the potential, and the formalism used. For example, local atomic potentials (and pseudopotentials) involve a simple sum over atomic sites, I,

$$V^{\text{Ext.}}(\mathbf{r}) = \sum_I V_I\big(|\mathbf{r} - \mathbf{R}_I|\big). \tag{9}$$

The Hartree energy is the classical expression for the Coulomb interaction of two electron densities,

$$E^{\text{Har.}} = \frac{1}{2} \int\int \mathrm{d}\mathbf{r}\, \mathrm{d}\mathbf{r}' \, \frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}. \tag{10}$$

The evaluation of this term can be performed more efficiently by expressing the real-space density as a plane-wave expansion,

$$\rho(\mathbf{r}) = \frac{1}{\Omega} \sum_{\mathbf{g}} \mathrm{e}^{i\mathbf{g}\cdot\mathbf{r}} \rho(\mathbf{g}), \tag{11}$$

where Ω is the system volume, and the reciprocal-space density ($\rho(\mathbf{g})$) are the expansion coefficients, and also correspond to the inverse Fourier transform coefficients of the real-space density,

$$\rho(\mathbf{g}) = \int \mathrm{d}\mathbf{r}\, \mathrm{e}^{-i\mathbf{g}\cdot\mathbf{r}} \rho(\mathbf{r}). \tag{12}$$

Performing this expansion allows the real-space double integral for the Hartree energy to be re-expressed as a single reciprocal-space summation,

$$E^{\text{Har.}} = \frac{1}{2\Omega} \sum_{\mathbf{g}\neq 0} \frac{4\pi}{|\mathbf{g}|^2} \rho^*(\mathbf{g})\rho(\mathbf{g}), \tag{13}$$

effectively reducing the order of the evaluation from quadratic to linear in the system size. This transformation will also be exploited several times in the OF-KEDF method.

The Hartree potential can be expressed on a real- or reciprocal-space grid,

$$\frac{\delta E^{\text{Har.}}}{\delta \rho(\mathbf{r})} = \int \mathrm{d}\mathbf{r}' \frac{\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}, \tag{14}$$

$$\frac{\delta E^{\text{Har.}}}{\delta \rho^*(\mathbf{g})} = \int \mathrm{d}\mathbf{r}\, \mathrm{e}^{-i\mathbf{g}\cdot\mathbf{r}} \frac{\delta E^{\text{Har.}}}{\delta \rho(\mathbf{r})} \tag{15}$$

$$= \frac{4\pi}{|\mathbf{g}|^2} \rho(\mathbf{g}). \tag{16}$$

The exchange and correlation energies are historically grouped together [13]. Many forms of these functionals exist [14,15], and the development of a general transferable functional is an area of active research [16–18]. For the "simple" metals under study here, the Local Density Approximation (LDA), wherein the exchange-correlation properties at a point in space are equated to those of the uniform electron gas with the same density, has been seen to be appropriate [4]. The energy and potential are expressed in terms of functionals of the local density,

$$E^{\text{XC}} = \int \mathrm{d}\mathbf{r}\, \rho(\mathbf{r}) \epsilon^{\text{XC}}\big[\rho(\mathbf{r})\big], \tag{17}$$

$$\frac{\delta E^{\text{XC}}}{\delta \rho(\mathbf{r})} = \epsilon^{\text{XC}}\big[\rho(\mathbf{r})\big] + \rho(\mathbf{r}) \mu^{\text{XC}}\big[\rho(\mathbf{r})\big], \tag{18}$$

$$\frac{\delta E^{\text{XC}}}{\delta \rho^*(\mathbf{g})} = \int \mathrm{d}\mathbf{r}\, \mathrm{e}^{-i\mathbf{g}\cdot\mathbf{r}} \frac{\delta E^{\text{XC}}}{\delta \rho(\mathbf{r})}, \tag{19}$$

where the exchange-correlation potential functional, μ^{XC}, is the derivative of the energy functional ($\mu^{\text{XC}} = \partial\epsilon^{\text{XC}}/\partial\rho$). The local nature of most exchange-correlation energy and potential functionals, allows a linear-scaling evaluation. The Ceperley–Alder [19] exchange-correlation functionals, as parameterized by Perdew–Zunger [20] are used in the present work.

We have shown that each term of the electronic potential energy, and their potentials, can be formulated in a manner that allows evaluation with a computational effort that scales linearly with system size. We shall now turn our attention to the electronic kinetic energy, and the consequences of its form on computational scaling within a metallic calculation.

2.2. Kohn–Sham kinetic energy

It was Kohn and Sham [21] who first observed that the potential obtained from the Density Functional Theory energy expression was the same as that applied to a system of non-interacting electrons with the Schrödinger equation,

$$\left(-\frac{1}{2}\nabla^2 + \frac{\delta U^{\text{elec.}}}{\delta \rho(\mathbf{r})} \right) \psi_\alpha(\mathbf{r}) = \epsilon_\alpha \psi_\alpha(\mathbf{r}). \tag{20}$$

By solving the set of single-particle Schrödinger (or Kohn–Sham) equations, one obtains the electronic energy, density, and importantly for this discussion, the electronic kinetic energy of a system of non-interacting electrons

with the same density. The solution must be obtained self-consistently, as the potential depends upon the density and the density, in turn, depends on the single-particle wavefunctions,

$$\rho(\mathbf{r}) = \sum_{\alpha}^{\text{occ.}} \psi_{\alpha}^{*}(\mathbf{r})\psi_{\alpha}(\mathbf{r}). \tag{21}$$

The orbitals are occupied, with increasing energy ϵ_{α}, until the correct total number of electrons is obtained. This reformulation of the DFT problem in terms of the Kohn–Sham orbitals makes the problem tractable. It also has some very important consequences.

The non-interacting Kohn–Sham kinetic energy is given by,

$$T^{\text{KS}} = \sum_{\alpha}^{\text{occ.}} \langle \psi_{\alpha} | -\tfrac{1}{2}\nabla^2 | \psi_{\alpha} \rangle, \tag{22}$$

provided that $\{\psi\}$ form an orthonormal set,

$$\langle \psi_{\alpha} | \psi_{\beta} \rangle = \delta_{\alpha,\beta}. \tag{23}$$

The imposition of orthogonality on a set of orbitals generally makes the method scale as the third power of the system size (although, as discussed previously, local techniques can reduce this scaling [7,22]).

Another important point is that the Kohn–Sham orbitals are not directly observable. It is the sum of the square of their magnitudes which equates to the density (Eq. (21)). Wavefunctions differing by some complex phase will produce the same density,

$$\left(\mathrm{e}^{i\mathbf{k}}\psi_{\alpha}(\mathbf{r})\right)^{*}\left(\mathrm{e}^{i\mathbf{k}}\psi_{\alpha}(\mathbf{r})\right) = \mathrm{e}^{-i\mathbf{k}}\mathrm{e}^{i\mathbf{k}}\psi_{\alpha}^{*}(\mathbf{r})\psi_{\alpha}(\mathbf{r}) = \left|\psi_{\alpha}(\mathbf{r})\right|^2. \tag{24}$$

Were this phase a function of the position (i.e. $\mathbf{k}\cdot\mathbf{r}$ not simply $\mathbf{k}$), it would have a contribution to the kinetic energy. For a periodic system (for example, a simulation cell with Born–von Karman boundary conditions) this is a restatement of Bloch's theorem [23] which tells us that each wavefunction can be written as the product of a complex phase (which does not necessarily have the periodicity of the system) and a cell-periodic function,

$$\psi_{\alpha,\mathbf{k}}(\mathbf{r}) = \mathrm{e}^{i\mathbf{k}\cdot\mathbf{r}}\psi_{\alpha}(\mathbf{r}). \tag{25}$$

The allowed values of this phase are determined by the boundary conditions that apply to the "bulk" (not the simulation cell). For infinite metals, the type modeled using periodic boundary conditions, an infinitely dense set is allowed, with terms up to the Fermi energy affecting the electronic density and energy. Approximating this set (**k**-point sampling) is difficult for metals. The Fermi surface needs to be represented accurately to ensure correct orbital occupation, often requiring very dense **k**-point grids, which comes at a computational cost that scales linearly with the number of additional points. With larger systems, more of the complex-phase is included explicitly in the calculation, requiring less-dense **k**-point grids for equivalent accuracy. However, this increase in size comes at the usual cubic computational cost. What this invariably means for metals is that there is a need for very intensive **k**-point sampling, which can increase the computational cost by orders of magnitude over a corresponding calculation for an insulator.

Summarizing, the introduction of Kohn–Sham orbitals allows the calculation of the non-interacting electronic kinetic energy within Density Functional Theory. The computational cost generally scales as the third power of the system size, and the necessity for **k**-point sampling has a dramatic effect on the prefactor of this scaling. DFT does not require such orbitals and was originally cast in a form without them. A density-only formulation, as we shall see, does not suffer from these computational expenses.

2.3. Orbital-free kinetic energy density functional

Reverting back to a density-derived kinetic energy functional, as suggested by Hohenberg and Kohn [12], and in the spirit of the original density functional theory, is a very attractive proposition. Without the introduction of the

single-particle orbitals there is no need for the orbital orthogonalization that dominates the scaling of the Kohn–Sham scheme. Also, as we are dealing with the density, an observable quantity, there is no phase information to consider and no need for expensive Bloch summations or **k**-point sampling. For a kinetic energy which is a functional solely of the local density (Thomas–Fermi [12,24,25], or von Weizsäcker [26] are two historical examples) one would expect scaling similar to that already seen in the potential energy functionals, which scale linearly with the system size. However, there is a major obstacle to overcome to make this a useful approach.

As yet, no known functionals exist that can (to within reasonable accuracy and for general applications) reproduce the results of the orbital-based Kohn–Sham method. There are several functionals which are known to give the correct non-interacting kinetic energy at certain limits (the Thomas–Fermi KEDF for a uniform electron gas and the von Weizsäcker KEDF for a single spatial orbital [4,5]). Recently a class of functionals have arisen that are based upon the response behavior of the uniform Hartree gas [27–33]. They have been shown to give results comparable in accuracy to Kohn–Sham, for the properties of relatively simple bulk metals. The latest functionals have even reproduced surface energetics for aluminum [33], well beyond where one would expect the linear response expansion to hold.

2.3.1. Linear response kinetic energy density functionals

All functionals of this class based upon the response behavior of the Hartree gas have a general form. They include the Thomas–Fermi kinetic energy, the von Weizsäcker kinetic energy, and a "correction" that explicitly gives the "correct" (Hartree) response behavior of the electrons,

$$T_{\mathrm{s}} = T_{\mathrm{TF}} + T_{\mathrm{vW}} + T_{\mathrm{Res.}}. \tag{26}$$

The Thomas–Fermi kinetic energy yields the correct result for the non-interacting uniform electron gas. It also includes the Fermi statistics necessary for the electrons (statistics that are not included in the other terms). Its energy and potential are simple functionals of the density,

$$T_{\mathrm{TF}} = C_{\mathrm{TF}} \int \mathrm{d}\mathbf{r}\, \rho^{5/3}(\mathbf{r}), \qquad C_{\mathrm{TF}} = \tfrac{3}{10}\left(3\pi^2\right)^{2/3}, \tag{27}$$

$$\frac{\delta T_{\mathrm{TF}}}{\delta \rho(\mathbf{r})} = \tfrac{5}{3} C_{\mathrm{TF}} \rho^{2/3}(\mathbf{r}). \tag{28}$$

The von Weizsäcker kinetic energy can be formulated as that of the Laplacian applied to the square-root of the density,

$$T_{\mathrm{vW}} = \int \mathrm{d}\mathbf{r} \sqrt{\rho(\mathbf{r})}\left(-\tfrac{1}{2}\nabla^2\right)\sqrt{\rho(\mathbf{r})}. \tag{29}$$

However, it can also be expressed in terms of whole functions of the density, which aids in the formulation of the potential,

$$T_{\mathrm{vW}} = \int \mathrm{d}\mathbf{r} \left\{ -\frac{1}{4}\nabla^2\rho(\mathbf{r}) + \frac{1}{8}\frac{(\nabla\rho(\mathbf{r}))^2}{\rho(\mathbf{r})} \right\} = \frac{1}{8}\int \mathrm{d}\mathbf{r} \frac{(\nabla\rho(\mathbf{r}))^2}{\rho(\mathbf{r})}, \tag{30}$$

$$\frac{\delta T_{\mathrm{vW}}}{\delta \rho(\mathbf{r})} = -\frac{1}{4}\frac{\nabla^2\rho(\mathbf{r})}{\rho(\mathbf{r})} + \frac{1}{8}\left(\frac{\nabla\rho(\mathbf{r})}{\rho(\mathbf{r})}\right)^2 \tag{31}$$

(the second equality in Eq. (30) holds for any physical density in a closed system, or Hilbert space).

An elegant form for both of these terms, is that of double integration [27,33]:

$$T_{\mathrm{TF}} = C_{\mathrm{TF}} \left\langle \rho^{5/6}(\mathbf{r}) \middle| \delta(\mathbf{r}-\mathbf{r}') \middle| \rho^{5/6}(\mathbf{r}') \right\rangle, \tag{32}$$

$$T_{\mathrm{vW}} = -\tfrac{1}{4}\left\langle \rho^{1/2}(\mathbf{r}) \middle| \delta(\mathbf{r}-\mathbf{r}')\nabla^2 + \nabla^2\delta(\mathbf{r}-\mathbf{r}') \middle| \rho^{1/2}(\mathbf{r}') \right\rangle. \tag{33}$$

The response functional developed by Wang and Teter [27], amongst others [28–30,32], and extended by Wang et al. [33], can be expressed as a natural generalization of this double integration form,

$$T_{\text{Res.}} = C_{\text{TF}}\langle\rho^{\alpha}(\mathbf{r})|\omega_{\alpha,\beta}(\mathbf{r}-\mathbf{r}')|\rho^{\beta}(\mathbf{r}')\rangle, \tag{34}$$

where α and β are parameters. Requiring the functional to explicitly give the correct linear response immediately gives an expression for the response kernel ($\omega_{\alpha,\beta}$), in a reciprocal space representation, as

$$\widetilde{\omega}_{\alpha,\beta}(|\mathbf{g}|) = -\frac{\chi_{\text{Lind.}}^{-1} - \chi_{\text{TF}}^{-1} - \chi_{\text{vW}}^{-1}}{2\alpha\beta C_{\text{TF}}\rho_0^{\alpha+\beta-2}} \tag{35}$$

$$= \frac{5G(q)}{9\alpha\beta\rho_0^{\alpha+\beta-5/3}}, \tag{36}$$

where ρ_0 is the average electronic density, $\widetilde{\omega}_{\alpha,\beta}$ is the Fourier transform of the real-space kernel, $q = |\mathbf{g}|/2k_{\text{F}}$ is the magnitude of the reciprocal lattice vector, normalized by the magnitude of the Fermi vector k_{F}, which is in turn, related to ρ_0,

$$k_{\text{F}} = \sqrt[3]{3\pi^2\rho_0}. \tag{37}$$

χ_{TF}, χ_{vW}, and $\chi_{\text{Lind.}}$ are the susceptibilities of the electron gas within the Thomas–Fermi, von Weizsäcker, and Lindhard models, respectively (the Lindhard model being the uniform Hartree gas) [23]. The response function G, is given by,

$$G(q) = \left(\frac{1}{2} + \frac{1-q^2}{4q}\ln\left|\frac{1+q}{1-q}\right|\right)^{-1} - 3q^2 - 1. \tag{38}$$

Consideration of the limiting forms of Eq. (34) by Wang et al., produced values of α and β,

$$\alpha, \beta = \frac{5}{6} \pm \frac{\sqrt{5}}{6}. \tag{39}$$

With these values, spurious contributions to first order in the density deviations are removed at the large $|\mathbf{g}|$ limit [33]. Wang and Teter [27], by a different analysis, produced values of $\alpha, \beta = \frac{5}{6} \pm (\frac{5}{6} - \frac{\sqrt{8}}{3})$.

Calculating the response kinetic energy and potential within linear-scaling time is a matter of using the appropriate representation (as it was for the Hartree energy/potential, Eq. (13)),

$$T_{\text{Res.}} = \int\int \mathrm{d}\mathbf{r}\,\mathrm{d}\mathbf{r}'\rho^{\alpha}(\mathbf{r})\,\rho^{\beta}(\mathbf{r}')\omega_{\alpha,\beta}(\mathbf{r}-\mathbf{r}')$$

$$= \frac{1}{\Omega}\sum_{\mathbf{g}}\rho_{\alpha}^{*}(\mathbf{g})\rho_{\beta}(\mathbf{g})\omega_{\alpha,\beta}(\mathbf{g}), \tag{40}$$

$$\frac{\delta T_{\text{Res.}}}{\delta\rho(\mathbf{r})} = \alpha\rho^{\alpha-1}(\mathbf{r})\int\mathrm{d}\mathbf{r}'\rho^{\beta}(\mathbf{r}')\omega_{\alpha,\beta}(\mathbf{r}-\mathbf{r}') + \beta\rho^{\beta-1}(\mathbf{r})\int\mathrm{d}\mathbf{r}'\rho^{\alpha}(\mathbf{r}')\omega_{\alpha,\beta}(\mathbf{r}-\mathbf{r}') \tag{41}$$

$$= \frac{\alpha\rho^{\alpha-1}(\mathbf{r})}{\Omega}\sum_{\mathbf{g}}\mathrm{e}^{i\mathbf{g}\cdot\mathbf{r}}\rho_{\beta}(\mathbf{g})\omega_{\alpha,\beta}(\mathbf{g}) + \frac{\beta\rho^{\beta-1}(\mathbf{r})}{\Omega}\sum_{\mathbf{g}}\mathrm{e}^{i\mathbf{g}\cdot\mathbf{r}}\rho_{\alpha}(\mathbf{g})\omega_{\alpha,\beta}(\mathbf{g}), \tag{42}$$

where $\rho_{\alpha}(\mathbf{g})$ are the Fourier coefficients of $\rho^{\alpha}(\mathbf{r})$. The two summations in the final term are Fourier transforms and can be precomputed. This allows the evaluation of both the energy and potential in a computation time which scales linearly with system size.

2.3.2. Linear response functional with a density-dependent kernel

The Lindhard response function introduced in Eq. (35) gives the correct linear response behavior about a uniform electron density. Within the double density integration (Eq. (34)), the single reference density (ρ_0, used in Eq. (36) and in normalizing the reciprocal lattice vector) is not always appropriate, especially for systems with greatly varying densities (surfaces and alloys, for example). Wang et al. [33] have produced a response functional with a density-dependent kernel, in the spirit of those developed by Chacón et al. [34], while maintaining linear-scaling.

The response correction with the density-dependent kernel has the same form as in Eq. (34),

$$T_{\rho\text{-Res.}} = C_{\mathrm{TF}}\big\langle \rho^{\alpha}(\mathbf{r})\big|\omega_{\alpha,\beta}(\xi_{\gamma}(\mathbf{r},\mathbf{r}'),\mathbf{r}-\mathbf{r}')\big|\rho^{\beta}(\mathbf{r}')\big\rangle, \tag{43}$$

where the exact form of the local two-body Fermi vector $\xi_{\gamma}(\mathbf{r},\mathbf{r}')$ is related to the local one-body Fermi vector ($k_{\mathrm{F}}(\mathbf{r})$) by Natural Variable arguments [35],

$$\xi_{\gamma}(\mathbf{r},\mathbf{r}') = \sqrt[\gamma]{\frac{k_{\mathrm{F}}(\mathbf{r})^{\gamma}+k_{\mathrm{F}}(\mathbf{r}')^{\gamma}}{2}}, \tag{44}$$

$$k_{\mathrm{F}}(\mathbf{r}) = \sqrt[3]{3\pi^2\rho(\mathbf{r})}. \tag{45}$$

Enforcing the linear response of the Hartree gas (as in Eq. (35)), now produces a universal second-order differential equation for each fixed magnitude of the reciprocal lattice vector $|\mathbf{g}|$,

$$q^2\widetilde{\omega}''_{\alpha,\beta}(q,\rho_0)+\big[\gamma+1-6(\alpha+\beta)\big]q\widetilde{\omega}'_{\alpha,\beta}(q,\rho_0)+36\alpha\beta\widetilde{\omega}_{\alpha,\beta}(q,\rho_0)=20G(q)\rho_0^{5/3-(\alpha+\beta)}, \tag{46}$$

where derivatives are with respect to the *scaled* reciprocal lattice vector $q(=|\mathbf{g}|/2k_{\mathrm{F},0})$.

Examining the limiting behavior of this object, Wang et al. [33] suggested values of α and β as before,

$$\alpha,\beta = \tfrac{5}{6}\pm\tfrac{\sqrt{5}}{6}, \tag{47}$$

and a value for γ of 2.7, which was seen to reduce spurious effects linear in the density deviation, for the $\mathbf{g}\to 0$ limit.

Evaluation of the density-dependent kernel, $\omega_{\alpha,\beta}(\xi_{\gamma}(\mathbf{r},\mathbf{r}'),\mathbf{r}-\mathbf{r}')$, in linear-scaling computation time, relies on a Taylor expansion about a uniform density ρ_*, which is given by,

$$\begin{aligned}
\omega_{\alpha,\beta}\big(\xi_{\gamma}(\mathbf{r},\mathbf{r}'),\mathbf{r}-\mathbf{r}'\big) &= \omega_{\alpha,\beta}(k_{\mathrm{F}}^{*},\mathbf{r}-\mathbf{r}')\\
&\quad+\frac{\partial\omega_{\alpha,\beta}(\xi_{\gamma}(\mathbf{r},\mathbf{r}'),\mathbf{r}-\mathbf{r}')}{\partial\rho(\mathbf{r})}\bigg|_{\rho_*}\Delta\rho(\mathbf{r})\\
&\quad+\frac{\partial\omega_{\alpha,\beta}(\xi_{\gamma}(\mathbf{r},\mathbf{r}'),\mathbf{r}-\mathbf{r}')}{\partial\rho(\mathbf{r}')}\bigg|_{\rho_*}\Delta\rho(\mathbf{r}')\\
&\quad+\frac{\partial^2\omega_{\alpha,\beta}(\xi_{\gamma}(\mathbf{r},\mathbf{r}'),\mathbf{r}-\mathbf{r}')}{\partial\rho^2(\mathbf{r})}\bigg|_{\rho_*}\frac{(\Delta\rho(\mathbf{r}))^2}{2}\\
&\quad+\frac{\partial^2\omega_{\alpha,\beta}(\xi_{\gamma}(\mathbf{r},\mathbf{r}'),\mathbf{r}-\mathbf{r}')}{\partial\rho^2(\mathbf{r}')}\bigg|_{\rho_*}\frac{(\Delta\rho(\mathbf{r}'))^2}{2}\\
&\quad+\frac{\partial^2\omega_{\alpha,\beta}(\xi_{\gamma}(\mathbf{r},\mathbf{r}'),\mathbf{r}-\mathbf{r}')}{\partial\rho(\mathbf{r})\partial\rho(\mathbf{r}')}\bigg|_{\rho_*}\Delta\rho(\mathbf{r})\Delta\rho(\mathbf{r}')\\
&\quad+\mathcal{O}\big(\big(\Delta\rho(\mathbf{r})\big)^3\big),
\end{aligned} \tag{48}$$

where $\Delta\rho(\mathbf{r})=\rho(\mathbf{r})-\rho_*$. The derivatives within this Taylor-expansion can be related (up to second-order) to the objects arising in the second-order differential equation already seen (Eq. (46)), by means of their Fourier transforms,

$$\widehat{F}\left(\left.\frac{\partial w_{\alpha,\beta}(\xi_\gamma(\mathbf{r},\mathbf{r}'),\mathbf{r}-\mathbf{r}')}{\partial\rho(\mathbf{r})}\right|_{\rho_*}\right)=-\frac{q_*\widetilde{\omega}'_{\alpha,\beta}(q_*,\rho_*)}{6\rho_*}, \tag{49}$$

$$\widehat{F}\left(\left.\frac{\partial^2 w_{\alpha,\beta}(\xi_\gamma(\mathbf{r},\mathbf{r}'),\mathbf{r}-\mathbf{r}')}{\partial\rho^2(\mathbf{r})}\right|_{\rho_*}\right)=\frac{q_*^2\widetilde{\omega}''_{\alpha,\beta}(q_*,\rho_*)+(7-\gamma)q_*\widetilde{\omega}'_{\alpha,\beta}(q_*,\rho_*)}{36\rho_*^2}, \tag{50}$$

$$\widehat{F}\left(\left.\frac{\partial^2 w_{\alpha,\beta}(\xi_\gamma(\mathbf{r},\mathbf{r}'),\mathbf{r}-\mathbf{r}')}{\partial\rho(\mathbf{r})\partial\rho(\mathbf{r}')}\right|_{\rho_*}\right)=\frac{q_*^2\widetilde{\omega}''_{\alpha,\beta}(q_*,\rho_*)+(1+\gamma)q_*\widetilde{\omega}'_{\alpha,\beta}(q_*,\rho_*)}{36\rho_*^2}. \tag{51}$$

Each of these terms can be evaluated in a computation time scaling linearly with system size. The double integral form of the response kinetic energy (Eq. (43)) can also be evaluated (in reciprocal-space) in linear-scaling time,

$$T_{\rho\text{-Res.}}=C_{\mathrm{TF}}\iint \mathrm{d}\mathbf{r}\,\mathrm{d}\mathbf{r}'\rho^\alpha(\mathbf{r})\rho^\beta(\mathbf{r})\omega_{\alpha\beta}\big(\xi_\gamma(\mathbf{r},\mathbf{r}'),\mathbf{r}-\mathbf{r}'\big), \tag{52}$$

$$=\frac{C_{\mathrm{TF}}}{\Omega}\sum_{\mathbf{g}}\rho_\alpha^*(\mathbf{g})\rho_\beta(\mathbf{g})\omega_{\alpha\beta}(\mathbf{g}). \tag{53}$$

The potential arising from the density-dependent kernel is very similar to that of the density-independent kernel (Eqs. (41) and (42)), with an additional term which includes the potential due to the kernel,

$$\frac{\delta T_{\rho\text{-Res.}}}{\delta\rho(\mathbf{r})}=\frac{\delta T_{\mathrm{Res.}}}{\delta\rho(\mathbf{r})}+\rho^\alpha(\mathbf{r})\int \mathrm{d}\mathbf{r}'\frac{\delta\omega_{\alpha\beta}(\xi_\gamma(\mathbf{r},\mathbf{r}'),\mathbf{r}-\mathbf{r}')}{\delta\rho(\mathbf{r})}\rho^\beta(\mathbf{r}')$$
$$+\rho^\beta(\mathbf{r})\int \mathrm{d}\mathbf{r}'\frac{\delta\omega_{\alpha\beta}(\xi_\gamma(\mathbf{r},\mathbf{r}'),\mathbf{r}-\mathbf{r}')}{\delta\rho(\mathbf{r})}\rho^\alpha(\mathbf{r}'). \tag{54}$$

The density-dependence of the kernel is easily obtained from the Taylor expansion (Eq. (48)). Each term is computable in linear-scaling time, as is the final potential (again, the terms are precomputable as Fourier expansions).

2.3.3. Quadratic response kernels

A kinetic energy functional which explicitly includes (approximate) quadratic response behavior has been formulated by Wang and Teter [27] (with further analysis by Foley and Madden [31,36]). The separable approximation of the quadratic response function allows its computation in linear-scaling time. While the explicit inclusion of the correct higher order terms produces improved bulk properties [31], the lack of a density-dependence in the quadratic response kernel makes the treatment of surfaces difficult, and can produce results for surfaces that are not size-consistent [33,37].

3. Scaling of OF-KEDF versus conventional Kohn–Sham

The analysis of the computational effort presented above, and in previous works, demonstrates a quasi-linear scaling for all elements of the calculation of the electronic energy. It is the transformation to and from a Fourier representation for the electronic density which should dominate the computational effort. However, prefactors in the computational cost for each element of the calculation can vary greatly, and linear in "principle" is not always linear in "practice."

Fig. 1 shows two sets of comparable simulations, for FCC aluminum at the experimental density. The KS method is implemented by the commercial code CASTEP (CASTEP 3.9 from Molecular Simulations Inc., San Diego). It uses ultra-soft pseudopotentials, allowing a planewave expansion cutoff energy of only 140 eV. While this does not affect the scaling of the method, it is considerably more efficient than a norm-conserving scheme which would require a higher cutoff energy. The code takes full advantage of the decreased necessity for **k**-point sampling

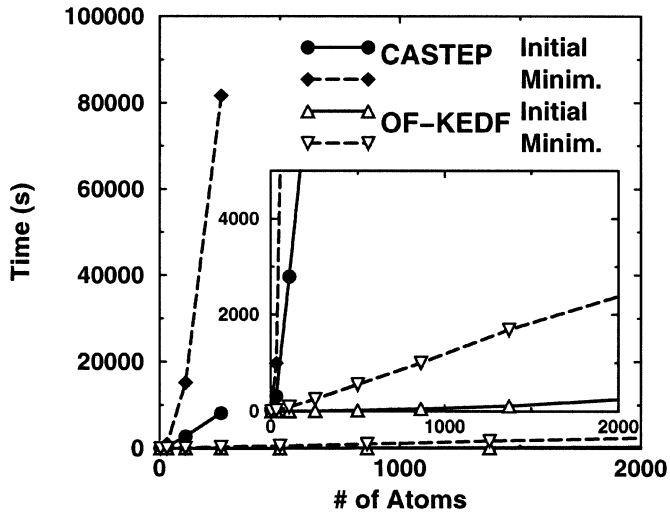

Fig. 1. The elapsed CPU time for the total energy minimization for FCC aluminum at the experimental density. Calculations were performed on a single node of a Compaq ES40. The KS-DFT code, CASTEP, used an ultra-soft pseudopotential, with a plane-wave cutoff of 140 eV. The OF-KEDF DFT code used a local norm-conserving potential and a cutoff of 680 eV. The initial (**Initial**) step includes initialization, and calculation of the ion–ion interactions. The minimization (**Minim.**) process had an energy tolerance of ∼1 meV. The insert has an expanded scale, to show detail of the OF-KEDF results.

with increasing simulation size (for the largest simulations, only a single **k**-point was needed). The OF-KEDF is implemented with the density-dependent kernel of Wang et al. [33], as outlined above. The norm-conserving local pseudopotential used, requires a higher cutoff than the ultra-soft pseudopotential (680 eV).

The observed scaling of the KS method very quickly makes the computational effort unmanageable for large systems. The scaling of the minimization procedure is clearly non-linear, and appears to be the expected cubic scaling. Note that the initialization is dominated by the initialization of the wavefunctions, and appears to scale linearly with system size.

The scaling of the minimization within the OF-KEDF scheme is seen to be linear, allowing the simulation of much larger systems than available to the KS method. The initialization step of the OF-KEDF includes the calculation of the ion–ion interactions (a term not calculated during the electronic minimization). The quadratic-scaling of this term can be seen in the timings of the initial step, and for very large systems (>5000 atoms), the ion–ion interaction was seen to dominate the computation for *ab initio* molecular dynamics simulations [39].

The results shown in Fig. 1 demonstrate conclusively that the computational effort within the OF-KEDF method not only scales linearly, but that the prefactors are small. Metallic systems with thousands of atoms are open for study, without the need for expensive supercomputers.

4. Parallelism and load-balancing

The OF-KEDFs attain quasi-linear scaling by taking advantage of the convolution operation, as seen in the expression for the Hartree energy (Eqs. (10) and (13)). Two representations are needed for the density, and several intermediate objects, one in real-space and one in reciprocal-space, which are related by,

$$\rho(\mathbf{r}) = \frac{1}{\Omega}\sum_{\mathbf{g}} \mathrm{e}^{i\mathbf{g}\cdot\mathbf{r}}\rho(\mathbf{g}), \tag{55}$$

$$\rho(\mathbf{g}) = \int \mathrm{d}\mathbf{r}\, \mathrm{e}^{-i\mathbf{g}\cdot\mathbf{r}}\rho(\mathbf{r}). \tag{56}$$

In order to achieve effective parallelism, both representations must be considered.

4.1. Real-space representation

The real-space representation of an object are the values of that object on a uniformly spaced grid (the use of the highly efficient Fast Fourier Transform (FFT) restricts the grid to be of this form). Computationally, it is stored and manipulated as a dense vector.

There are several terms, derived from the density, which are best formulated and evaluated in real-space (for example, T^{TF}, T^{vW}, and, for LDA, E^{XC}). All of these terms involve local functionals of the density,

$$E = \int \mathrm{d}\mathbf{r}\, \rho(\mathbf{r}) F\big[\rho(\mathbf{r})\big]. \tag{57}$$

Gradient terms (such as those within the von Weizsäcker and, for GGA, E^{XC}) can be evaluated locally by means of a pair of (fast) Fourier transforms,

$$\nabla_x \rho(\mathbf{r}) = \frac{1}{\Omega} \sum_{\mathbf{g}} \left(\mathrm{e}^{i\mathbf{g}\cdot\mathbf{r}} i\mathbf{g}_x \left(\int \mathrm{d}\mathbf{r}\, \mathrm{e}^{-i\mathbf{g}\cdot\mathbf{r}}\rho(\mathbf{r}) \right) \right), \tag{58}$$

where only the x component of the (vector) gradient (∇_x) is shown as illustration, using the x component of the reciprocal-grid vector ($\mathbf{g}_x$). This local evaluation makes the parallelism of the real-space objects, and the evaluation of properties from them (invariably simple summations) almost trivial. Each grid point requires the same computation. Load-balancing is a simple matter of ensuring that all nodes have an equal number of points. Interprocessor communication is limited to collection of local data sums, an insignificant task compared to their computation.

4.2. Reciprocal-space representation

The reciprocal-space representation are the coefficients of a Fourier expansion of the real-space object. These coefficients are evaluated by using the Fast Fourier Transform. It is this transform (and its inverse) which has the worst asymptotic scaling of the entire electronic structure calculation within the OF-KEDF method, scaling as $N_{\mathrm{grid}} \log_2 N_{\mathrm{grid}}$. The use of the FFT imposes a uniform reciprocal-space grid, whose spacing is determined by the size of our orthorhombic simulation cell,

$$\mathbf{g} = \left(n_x \frac{2\pi}{L_x}, n_y \frac{2\pi}{L_y}, n_z \frac{2\pi}{L_z} \right) \tag{59}$$

(L_x, L_y, L_z are the simulation cell lengths).

The Fourier expansion is truncated at the plane-wave of energy, E_{cut},

$$|\mathbf{g}| = \sqrt{2E_{\mathrm{cut}}}, \tag{60}$$

producing a sphere of reciprocal-space grid points. By using only the reciprocal-space sphere, and not the full cube, we treat each reciprocal lattice direction equivalently, but with the possible introduction of truncation errors which need to be considered when choosing E_{cut}. While the real-space objects (the density and objects derived thereof) are real, their Fourier coefficients are complex. However, as they are the Fourier coefficients of real objects there is a conjugated inversion symmetry,

$$\rho(-\mathbf{g}) = \rho^*(\mathbf{g}). \tag{61}$$

This symmetry can be exploited so that only the positive half-sphere of reciprocal-space grid points needs to be calculated and stored, as illustrated in Fig. 2.

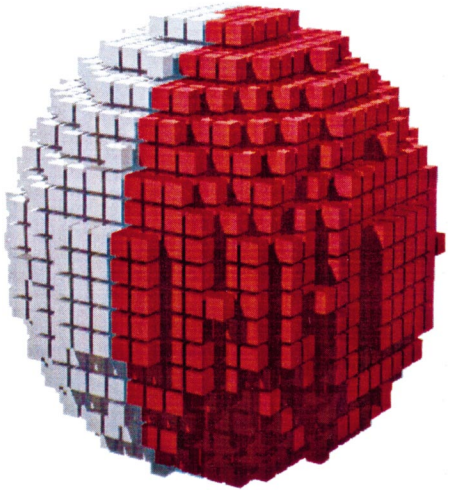

Fig. 2. The reciprocal-grid positive half-sphere. The grid origin is at the center. Only the red half-sphere need be calculated and stored.

Requiring only the positive half-sphere makes integration and evaluation within the reciprocal-space representation significantly more efficient than within the real-space representation. The number of active grid points in reciprocal-space is approximately given by

$$N_{\text{grid}} \approx \frac{1}{2}\left(\frac{4\pi}{3} N_{\mathbf{g}_x} N_{\mathbf{g}_y} N_{\mathbf{g}_z}\right), \tag{62}$$

where

$$N_{\mathbf{g}_x} = \text{int}\left(\frac{L_x}{2\pi}\sqrt{2E_{\text{cut}}}\right) + 1, \tag{63}$$

includes all grid-points within the cutoff energy. This should be compared with the number of real-space grid points, ($\sim 8 N_{\mathbf{g}_x} N_{\mathbf{g}_y} N_{\mathbf{g}_z}$ for the linear response KE functional, $\sim 27 N_{\mathbf{g}_x} N_{\mathbf{g}_y} N_{\mathbf{g}_z}$ for the quadratic response [31, 36]). Remembering that the reciprocal-space representation is complex, this still makes it approximately twice as efficient to calculate a property on the reciprocal-space grid ($\sim$7 times for the quadratic response).

Like the real-space grid, properties calculated on the reciprocal-space grid are local to each grid-point, and again each grid-point requires the same amount of computation. However, unlike real-space, the reciprocal-space grid is sparse, with much of the standard FFT grid not requiring computation (only the positive half-sphere). Standard vector manipulation would not result in the improved efficiency possible. Therefore, the sparse vector is mapped onto a dense vector, by simply ignoring non-required grid points (points beyond the sphere radius, or with a negative **g**), and keeping track of the sparse-dense map. Evaluations can now be performed using dense vector techniques. The only interprocessor communication necessary, as with the real-space grid, is the collection and distribution of local data summations.

This mapping of sparse – dense reciprocal-space representation vectors in a parallel environment can involve interprocessor communication (sparse data on one node would become dense on another, or vice-versa). Therefore, it is intimately linked with the load-balancing and data-distribution of the grid, and with the most computationally demanding part of the OF-KEDF method, the $N_{\text{grid}} \log_2 N_{\text{grid}}$ scaling Fast Fourier Transform.

4.3. Mapping and the FFT

As shown previously, all terms within the OF-KEDF method can be calculated with a computational effort which scales linearly with system size, provided that the correct real or reciprocal representation is used. It is

the transformation between the representations which requires the most effort. The transformation (Eq. (56), and its inverse Eq. (55)) generally scales as the square of the system size (N^2). However, the use of the Fast Fourier Transform (FFT) [40] allows the transform to be performed in time $N_{\text{grid}} \log_2 N_{\text{grid}}$, with the restriction that uniform grids in both representations be used.

The general procedure for performing a 3-dimensional FFT, is by a series of 1-dimensional transforms, followed by an axis rotation. For example, a 1D FFT is performed along the initial primary index direction (x). For a system of size $a \times b \times c$, this would involve $b \times c$ 1D FFTs of size a (for a cost of $b \times c \times a \log_2 a$). The result is then rotated so that a different coordinate (y) lies along the primary index. This then undergoes a 1D FFT ($a \times c \times b \log_2 b$), the result is again rotated so that the remaining coordinate (z) lies along the index direction. The final 1D FFT ($a \times b \times c \log_2 c$) can be followed by a final rotation, although it is not really necessary. Not performing the final rotation does save one set of communications, at the inconvenience of working in a set of rotated coordinates. The total computational cost scales as $a \times b \times c \times \log_2(a \times b \times c)$, while the communication cost depends upon the exact data distribution (but at worst it is $3 \times a \times b \times c$, including a final rotation). The balance of computational loading (the 1D FFT), and communication (the rotation) depends upon the data distribution, and the relative speeds of each operation. Three possible schemes for distributing the data amongst a series of processor nodes are shown in Fig. 3.

When choosing the optimal data distribution, another consideration, beyond the speed of the inter-processor communications, is the load-balancing within the reciprocal-space calculation. The "sparsity" of the reciprocal-space representation means that care must be taken to ensure an even spread of the grid points amongst the nodes, so as to load-balance the calculation (the majority of the OF-KEDF computation, excluding the transform, is performed in reciprocal-space). It is possible during the sparse – dense mapping to map the grid between nodes, but this involves communication. With current massively parallel computers, it is communication which has greater latency, taking considerably more cycles than computation. Efficient computation invariably means keeping communication to a minimum.

Ideal load-balancing is obtained when distributing the grid in a point-wise fashion (Fig. 3(a)). In this manner we can ensure that each node gets the optimal number of points. However, efficient computation of the 1D FFT would require significant data redistribution, and large numbers of communications. Maintaining a full data-row on a single node (as in Fig. 3(b)) allows a single 1D FFT to be performed without communication. When rotating a data-row ($x \to y$, for example), a complete reorganization of the data amongst the nodes is necessary, a process which should require N communications (where N is the total number of data points). However, a proportion of pre-rotated data already resides on the node to which it would be rotated. For a row-wise distribution, this proportion has a large dataset limit of $1 - \frac{1}{M^2}$, for an M-node computer. This gives a total communication cost (neglecting the final rotation) of,

$$2N\left(1 - \frac{1}{M^2}\right). \tag{64}$$

Looking at Fig. 3(b), one can see that the number of relevant grid-points in each row varies significantly, depending upon exactly where in the half-sphere they are taken from. Load-balancing can be maintained, provided that there are enough data-rows to average out this variation between the nodes.

Interprocessor communication can be reduced even further by distributing the data by planes (Fig. 3(c)). By holding a full data-plane on a node, the communication associated with the first rotation is removed ($x \to y$ if the the xy-plane is held). For the remaining rotation, the proportion of the data already residing on a given node is greater ($1 - \frac{1}{M}$), leading to a significantly lower total communication cost,

$$N\left(1 - \frac{1}{M}\right). \tag{65}$$

This decrease in communication comes with a possible load-balancing penalty. There are now significantly fewer planes (than lines or points), and the imbalance of computational cost per plane is larger (there are many more

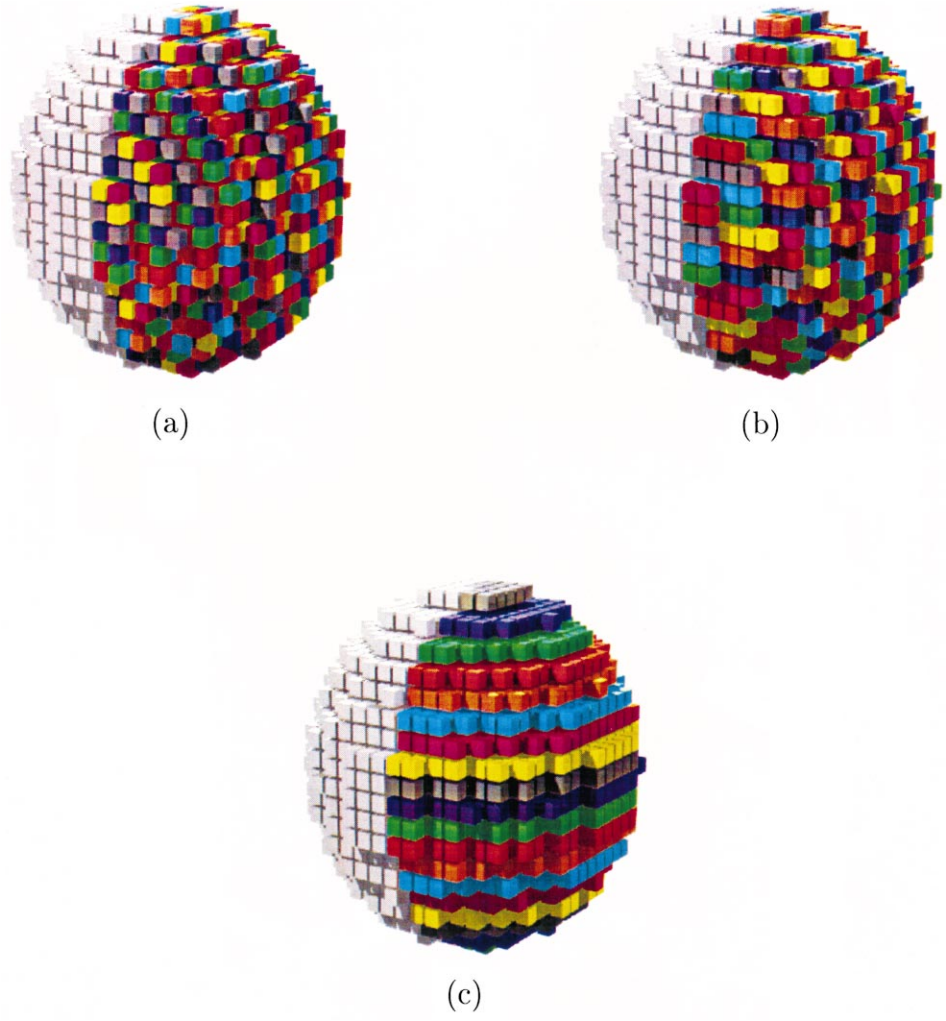

Fig. 3. Possible partitioning schemes for the reciprocal-space grid. In this simplified example, the different colors indicate which of eight possible nodes a grid point may reside on, being partitioned by (a) points, (b) rows, or (c) planes. The amount of communication and the degree of load-balancing is different for each case.

grid points in a plane at the center of the half-sphere than one at the edge). This makes it less likely that the computational cost can be averaged out over nodes.

Fig. 4 shows the scaling of the OF-KEDF calculation (electronic terms only), using the plane-wise distribution. The calculation is almost perfectly parallel ($\sim$99%). However, for the "small" system shown, and a large number of nodes, there are not enough data for a proper balance. For very large massively parallel computers, this could produce enough of a degradation that redistribution of the data amongst the nodes during the sparse-dense mapping would become more efficient.

The development of highly optimized "commercial" FFT libraries (Cray SciLib, SGI Complib, Compaq DXMLP) makes much of this discussion academic. Using these routines, it was seen that both computation and communication out-performed any "home-grown" routines, even those taking advantage of the reciprocal-space

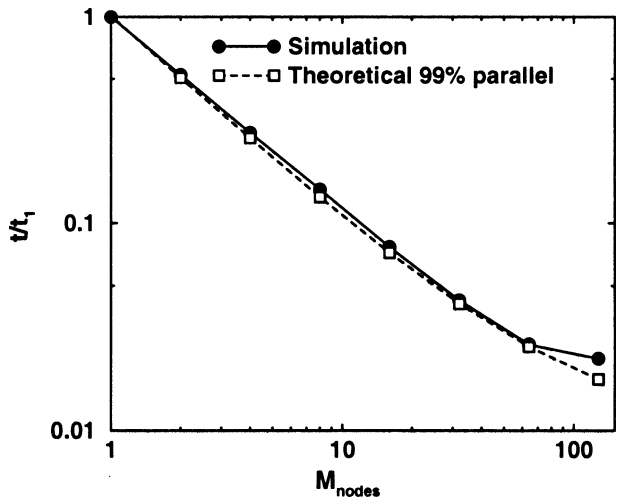

Fig. 4. The elapsed time (solid line) versus number of nodes, relative to a single node, for the electronic part of a "small" 128 atom calculation of bulk BCC sodium, performed on a Cray T3D. The scaling is almost perfect, comparable to 99% (dashed line). Performance degrades when the number of data-planes is less than the number of nodes ($\geqslant$ 128 nodes).

sparsity. These libraries have their own requirements for distribution of the data amongst the nodes. We have found that, for the OF-KEDF method, it is invariably more efficient to take advantage of these commercial developments.

5. Technical details

5.1. The variational parameter

Historically [28,29,31,41] it is the density, in either its real or reciprocal-space representation, which is taken to be the variational parameter (the intrinsic variable of an iteration scheme, that which is converged to produce the ground-state or energy minimum) in a Hohenberg–Kohn orbital-free calculation. Using the reciprocal-space representation has the advantage of being more efficient (as seen earlier, there are significantly less points in reciprocal-space than in real-space) as well as being suitable for successful preconditioning [28,42].

One problem with the reciprocal-space representation is that there is no assurance of a positive definite real-space density, a property which is not only physical, but essential for a feasible iteration scheme. In simulations where the ground-state density approaches zero (the "vacuum" of a surface calculation, for example) a very accurate iteration scheme is required, with very short propagation time-steps. This results in very lengthy (and computationally expensive) convergence. The real-space representation also experiences very similar problems.

An alternative scheme [33,43,44] is to use an object that is related to the density by some transform and which guarantees a positive definite density at all stages. One such object is the "square root" of the density, or to be technically correct, the object ($\chi(\mathbf{r})$) that, when squared, gives the real-space density,

$$\rho(\mathbf{r}) = \big(\chi(\mathbf{r})\big)^2. \tag{66}$$

Restricting $\chi(\mathbf{r})$ to be real and choosing it as the variational parameter (VP), ensures that we have a positive definite density at all stages of the iteration process.

One might consider that, like the density, a reciprocal-space implementation of the VP would be more efficient, as used in Refs. [43,45]. However, the transform between the VP and the density (Eq. (66)), means that the VP no longer has the same periodicity as the density. This can result in a "poorly" converged uniform Fourier expansion

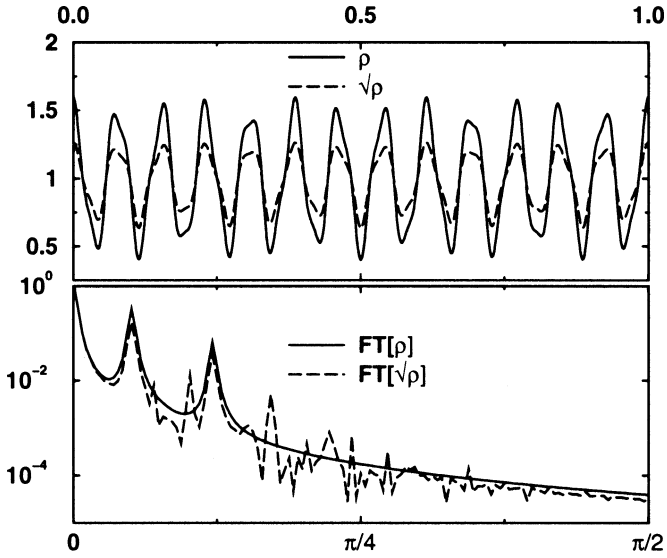

Fig. 5. The upper figure shows a simple periodic function, $y = 1 + \frac{1}{2}\cos(13 \cdot 2\pi x) + \frac{1}{10}\cos(31 \cdot 2\pi x)$, and its square-root. The lower figure is the simple Fourier transform (using a Blackman window [40]) of both functions. The transformed square-root shows considerable more structure, extending to much higher frequencies than the simple transform.

(as illustrated in Fig. 5). It is exactly this type of expansion which would be used to give the reciprocal-space representation. Therefore $\chi(\mathbf{g})$ should not, and will not, be used.

The simple *local* relationship between the new VP and the density allows an analytic expression for the new potential (the functional derivative of the energy *with respect to the variational parameter*),

$$\frac{\delta E}{\delta \chi(\mathbf{r})} = \int \mathrm{d}\mathbf{r}' \frac{\delta \rho(\mathbf{r}')}{\delta \chi(\mathbf{r})} \frac{\delta E}{\delta \rho(\mathbf{r}')}, \tag{67}$$

where

$$\frac{\delta \rho(\mathbf{r}')}{\delta \chi(\mathbf{r})} = \delta(\mathbf{r} - \mathbf{r}') 2\chi(\mathbf{r}), \tag{68}$$

leads immediately to

$$\frac{\delta E}{\delta \chi(\mathbf{r})} = 2\chi(\mathbf{r}) \frac{\delta E}{\delta \rho(\mathbf{r})}. \tag{69}$$

This apparently trivial relationship between the potential for the density and the potential for the VP has some very important consequences on the iteration procedure.

The form of the potential on the VP, as a product of the VP and the potential on the density, means that should the VP go to zero, then it will remain there, regardless of the potential affecting the density. For the ground-state of a finite system experiencing a finite external potential, this will not happen. In this situation, the density can approach, but not become, zero. However, truncation errors (arising from the plane-wave expansion, Eq. (11), for example) and overstepping [41] make such an event possible during the convergence process. Therefore, care must be taken to ensure that the VP is always nonzero. This also implies that there are no nodes in the variational parameter, and it is all positive (or all negative), similar to the density that it represents. However, one advantage that the new VP does possess over the density, is that a change of sign due to truncation errors (however unphysical) does not lead to catastrophic breakdown of the convergence procedure.

Propagating the variational parameter in real-space means that conservation of charge is no longer a trivial matter (ensuring charge conservation when propagating a reciprocal-space density, $\rho(\mathbf{g})$, is simply a matter of setting the $\mathbf{g}=0$ term of the potential ($\frac{\delta E}{\delta\rho(\mathbf{g}=0)}$) to zero). Some form of constrained propagation is necessary. Traditionally, this is included by means of a Lagrange multiplier μ [46],

$$E^{\text{const.}} = E - \mu\left(\int \mathrm{d}\mathbf{r}\,\rho(\mathbf{r}) - N_e\right). \tag{70}$$

In the case of the VP, this produces an additional term in the potential,

$$\frac{\delta E^{\text{const.}}}{\delta\chi(\mathbf{r})} = 2\chi(\mathbf{r})\left(\frac{\delta E}{\delta\rho(\mathbf{r})} - \mu\right). \tag{71}$$

Within a steepest descent minimization (or during the line-minimization of a conjugate gradient scheme), the VP is propagated as,

$$\chi(\mathbf{r};t+\Delta t) = \chi(\mathbf{r};t) - \Delta t\cdot 2\chi(\mathbf{r};t)\left(\frac{\delta E}{\delta\rho(\mathbf{r};t)} - \mu\right). \tag{72}$$

The Lagrange multiplier can be calculated [33] by performing an initial step (without the multiplier) followed by multiplication of both sides of Eq. (72) by $\chi(\mathbf{r};t)$ and integrating over all space,

$$\int \mathrm{d}\mathbf{r}\,\chi(\mathbf{r};t)\chi(\mathbf{r};t+\Delta t) = \int \mathrm{d}\mathbf{r}\,\chi(\mathbf{r};t)\left(\chi(\mathbf{r};t) - \Delta t\cdot 2\chi(\mathbf{r};t)\left(\frac{\delta E}{\delta\rho(\mathbf{r};t)} - \mu\right)\right). \tag{73}$$

This leads to

$$\mu = \frac{N_e - I_1 - I_2}{2\Delta t N_e}, \tag{74}$$

where

$$N_e = \int \mathrm{d}\mathbf{r}\left(\chi(\mathbf{r};t)\right)^2, \tag{75}$$

$$I_1 = \int \mathrm{d}\mathbf{r}\,\chi(\mathbf{r};t)\chi'(\mathbf{r};t+\Delta t), \tag{76}$$

$$I_2 = \int \mathrm{d}\mathbf{r}\,\Delta t\cdot 2\left(\chi(\mathbf{r};t)\right)^2\frac{\delta E}{\delta\rho(\mathbf{r};t)}, \tag{77}$$

and $\chi'(\mathbf{r};t+\Delta t)$ is the propagated VP *without* the Lagrange multiplier. This approximate form, which has linearized the μ dependence, does *not* guarantee charge conservation from one step to the next.

We can ensure charge conservation by multiplying Eq. (72) by $\chi(\mathbf{r};t+\Delta t)$, and integrating over all space. This leads to a quadratic form for μ, with solutions,

$$\mu = \frac{I_1}{N_e} - \frac{1 \pm \sqrt{1 + 4\Delta t^2\left(\frac{I_1^2}{N_e^2} - \frac{I_2}{N_e}\right)}}{2\Delta t}, \tag{78}$$

where

$$N_e = \int \mathrm{d}\mathbf{r}\left(\chi(\mathbf{r};t)\right)^2 = \int \mathrm{d}\mathbf{r}\left(\chi(\mathbf{r};t+\Delta t)\right)^2, \tag{79}$$

$$I_1 = \int \mathrm{d}\mathbf{r}\left(\chi(\mathbf{r};t)\right)^2\frac{\delta E}{\delta\rho(\mathbf{r};t)}, \tag{80}$$

and

$$I_2 = \int \mathrm{d}\mathbf{r}\left(\chi(\mathbf{r};t)\right)^2\left(\frac{\delta E}{\delta\rho(\mathbf{r};t)}\right)^2. \tag{81}$$

Not only does this non-linear form ensure that charge is conserved, but it naturally provides an upper limit on the step that can be taken while maintaining charge conservation (that which gives real solutions to the Lagrange multiplier),

$$\Delta t_{\max} = \frac{1}{2}\sqrt{\frac{N_e^2}{N_e I_2 - I_1^2}}. \tag{82}$$

This non-linear form allows arbitrary size steps to be attempted (of the type found within a line minimization), provides recognition of inappropriate size steps, and fully conserves charge. Similar non-linear expressions can be formulated for dynamical propagation (for example, using the Verlet or velocity Verlet algorithms [46]), with the exact form depending on which propagation algorithm is used.

5.2. Local pseudopotentials

The Kohn–Sham orbitals have an added benefit beyond allowing the calculation of the kinetic energy. They also allow a non-local representation of the effective ionic potential (pseudopotential). Pseudopotentials from all-electron calculations on the atom produce different potentials for each angular momentum component. The appropriate mixing of the potential in a solid phase calculation is obtained by projecting the atomic orbitals (s, p, d, etc.) onto the Kohn–Sham orbitals, and applying the relevant pseudopotential [47,48]. The information necessary to calculate this projection is not present when a density-only representation is used.

A scheme for producing the necessary local pseudopotentials, based upon *ab initio* calculations was recently proposed by Watson et al. [38]. The density of a fully converged orbital-based calculation performed using a non-local pseudopotential (ρ_{KS}), is mapped, using the OF-KEDF, to produce a 3-dimensional external potential,

$$V^{\mathrm{Ext.}}(\mathbf{r}) = -\frac{\delta T_{\mathrm{s}}}{\delta\rho(\mathbf{r})}\bigg|_{\rho_{\mathrm{KS}}} - \frac{\delta E^{\mathrm{Har.}}}{\delta\rho(\mathbf{r})}\bigg|_{\rho_{\mathrm{KS}}} - \frac{\delta E^{\mathrm{XC}}}{\delta\rho(\mathbf{r})}\bigg|_{\rho_{\mathrm{KS}}}. \tag{83}$$

This potential is then projected onto the atoms, and spherically averaged, a process best performed on the reciprocal lattice,

$$V_I(g) = \frac{1}{N_{|\mathbf{g}|=g}} \sum_{|\mathbf{g}|=g} \frac{V^{\mathrm{Ext.}}(\mathbf{g})}{S_I(\mathbf{g})}, \tag{84}$$

where $N_{|\mathbf{g}|=g}$ is the number of number of reciprocal lattice vectors ($\mathbf{g}$) with magnitude g, and the partial ionic structure factor $S_I(\mathbf{g})$ is for species I,

$$S_I(\mathbf{g}) = \sum_{J}^{\text{Ions of type } I} \exp(i\mathbf{g}\cdot\mathbf{R}_J). \tag{85}$$

The result is an *ab initio* local pseudopotential for the orbital-free scheme, derived from a *bulk crystal* reference (not from the traditional atom), which produces densities and energies comparable to the non-local potentials of the orbital-based scheme. Fig. 6 shows the energy for the FCC phase of aluminum, calculated by an orbital-based non-local potential, and with the orbital-free local potential. The agreement of properties, to within a few percent, is excellent. Note that the pseudopotentials generated and used here are consistent with the density-dependent kernel, and are subtly different from those calculated using the density-independent and quadratic response kinetic energy density functionals [38,41].

6. Large-scale simulation results

The power of this linear-scaling, easily parallelizable method for the DFT treatment of metals has been demonstrated previously for bulk systems. In particular, the OF-KEDF technique was used in perhaps the largest,

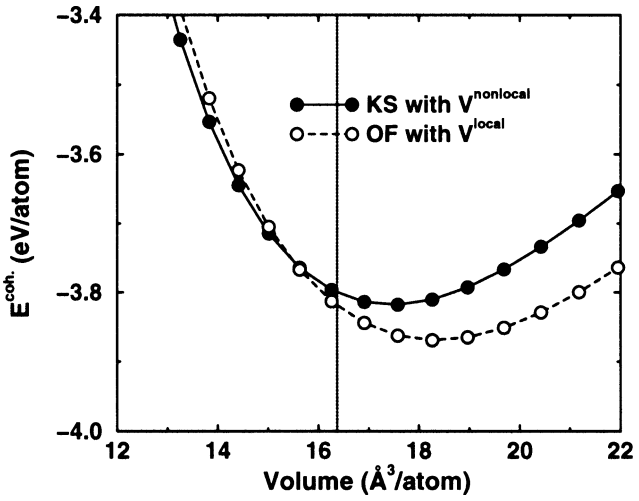

Fig. 6. The cohesive energy/volume curve for FCC aluminum. The experimental volume is the vertical line ($\Omega = 16.4$ Å^3/atom, $a_0 = 4.03$ Å). Comparison with the experimental cohesive energy of -3.4 eV, is not appropriate due to LDA errors in the atom [41].

fully self-consistent DFT, *ab initio* molecular dynamics simulation to date [39,41] to examine grain boundary annealing in sodium, modeled by 6714 atoms over a period of 1.5 ps. This study of nano-scale, bulk-like structures (micro-crystallites) used the density-independent kernel of Pearson et al. [28]. Wang et al. [33] have recently shown that an OF-KEDF with a density-dependent kernel can be used to study highly inhomogeneous systems. They were able to reproduce, with reasonable accuracy, Kohn–Sham predictions for the absolute and relative energies for several different surfaces of aluminum. This advance extends the realm of problems to which the OF-KEDF method may be applied.

An interesting problem, which would appear to be ideally suited to the density-dependent OF-KEDF, is the electronic properties of metal quantum dot (QD) arrays. The length scales involved in even the most minimalistic study of such an array, are beyond conventional Kohn–Sham methods, while the possible itinerant nature of the states make traditional localized $\mathcal{O}(N)$ methods inappropriate.

Thin-film QD arrays comprised of small (2–7 nm) diameter clusters of Ag and Au, passivated by alkylthiols, have been observed to self-assemble, forming 2-dimensional superlattices [49]. The electronic properties of these arrays depend upon the inter-cluster separation which, in turn, depends upon the length of the passivating alkyl tails, and the degree of compression of the film. Such films can be made to undergo a continuous metal-insulator transition [1,49–51].

Fig. 7 shows a 2-dimensionally hexagonal array of spherical aluminum clusters, each with a diameter ($2R$) of 2 nm. The passivating groups, present in the experiment only to act as an insulating medium and to keep the array mechanically stable, are here modeled electronically by vacuum (the “perfect” insulator). Equivalent arrays of silver clusters with similar radii, were recently seen to undergo the insulator to metal phase transition under compression [50], at a cluster surface separation of about 40% of the cluster radius. The transition was attributed to the narrowing of the Coulombic band-gap due to quantum-exchange interactions (a Mott–Hubbard transition [52]). In this work we will study a range of densities, from close-packed, to dilute, with Fig. 7 showing two of the extremal situations for the arrangement of our simulated hexagonal structures, with surface separations of 0.2 nm and 1 nm (20 and 100% of the radius, respectively).

Simulations were performed using the density-dependent OF-KEDF kernel. The pseudopotential for Al was derived from the Troullier–Martins non-local pseudopotential [53], as outlined in Section 5.2. A plane-wave cutoff energy of 680 eV was seen to be well converged with respect to the pseudopotential. The periodically repeated simulation cell used had a vacuum gap of 10 Å in the z-direction, so a general simulation cell has the dimensions

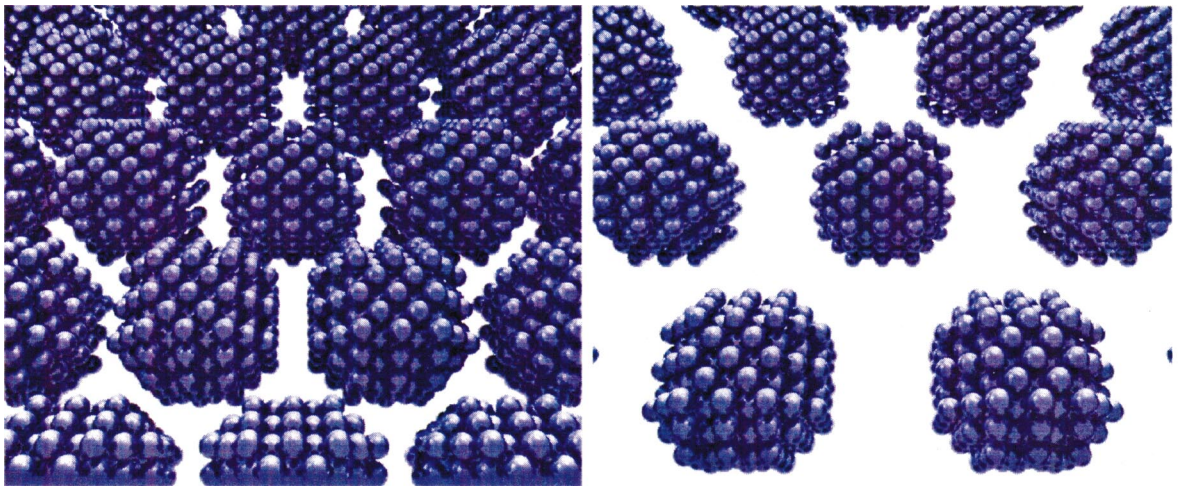

Fig. 7. The 2-dimensional, close-packed array of Al_{249} spherical clusters with diameter 2 nm. The cluster center separation distance is 2.2 nm (left) and 3 nm (right). The relaxed ion positions (shown), are almost identical to their starting positions (<0.1 Å displacements).

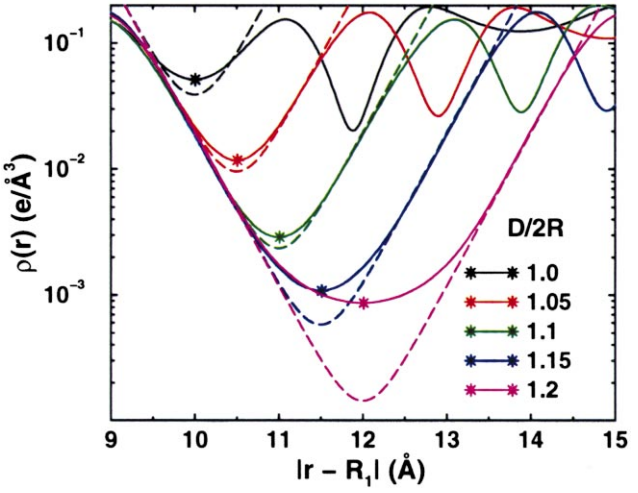

Fig. 8. The density along the [100] neighbor separation vector, for various separations. A $D/2R$ value of 1 (black line) indicates "touching" clusters. The dashed lines show an exponential density decay, given by Eq. (86). $|\mathbf{r} - \mathbf{R}_1|$ indicates the distance from the cluster centered at the origin ($\mathbf{R}_1$). The mid-point between clusters is indicated by $*$.

$D \times \sqrt{3}D \times 30$ Å (D being the cluster center-center separation). This elongated shape allows the simulation of the 2-dimensional hexagonal symmetry in an orthorhombic cell. Each cell contains two Al_{249} clusters, which are generated by extracting a sphere of radius 10 Å, centered on an atomic site, from the FCC crystal with the experimental lattice parameter (4.03 Å). Close examination of the cluster reveals its surface to be a collection of low-index (100), (110), and (111) surfaces. Initial calculations showed that the relaxation in the cluster, as for extended low index aluminum surfaces [33], was small (< 0.1 Å). Therefore unrelaxed ionic configurations are used in the results presented here.

Fig. 8 shows the self-consistent electronic density of two neighboring clusters (one centered at $\mathbf{R}_1$, the other at a distance D), along one of the cluster-to-cluster nearest neighbor directions, for several separations D. The

mid-point between cluster centers (and surfaces) is indicated by the $*$'s. The initial configuration of $D/2R = 1$ corresponds to the close-packed structure, where the "spherical" clusters are touching. One can see that the inter-cluster density is very similar to that seen in the bulk of the cluster (the right-hand side of the curve carries on into the inner regions of the neighboring cluster, illustrating the cluster "bulk" density). As the cluster separation increases, the density between the clusters drops rapidly.

At intermediate cluster separations ($D/2R \geqslant 1.05$), the density decay away from the cluster surface can be matched (over several orders of magnitude) to a single exponential,

$$\rho_{\text{surf}}(r) = \rho_0 \exp\bigl(-k_c(r - R_c)\bigr), \tag{86}$$

where ρ_0 is the average electronic density of the metal cluster ($= \rho_{\text{bulk}} \sim 0.18$ e/Å^3), R_c is approximately the cluster radius (9.2 Å), and k_c indicates the rapidity of the density decay (2.8 Å^{-1}). The latter two terms have values which were chosen to best describe the self-consistent densities. The density derived from two exponential surface densities (one centered on each cluster) is also shown in Fig. 8. One can see that this exponential form is appropriate for the inter-cluster density, for all but the smallest separations ($D/2R = 1$), out to $\sim$2 Å from the surface, which covers almost three orders of magnitude in the density.

As the inter-cluster separation increases, the simulated regions of very small density are not so well described by the exponential (for $D/2R > 1.25$, which are not shown, the low density is seen to "plateau" at approximately the minimum density seen for $D/2R = 1.25$). These regions of low density are far from the linear response regime for which the kinetic energy potential is explicitly correct, and the Taylor expansion in the density-dependent kernel, Eq. (48), appropriate. However, they do not contribute significantly to the energy since the density is so small. Not only is it possible that the response-based OF-KEDF is not capturing these low density regions accurately, but the insensitivity of the energy to such small densities, makes absolute convergence of these regions very difficult. This combination, while making total energy calculations for this region possible, makes the resulting density unreliable. Therefore, we shall be making use of our exponential fit in what follows, which gives excellent agreement for almost three orders of magnitude of the density.

The insulating behavior of the QD arrays at large separation, attributed to the Coulomb blockade [54], is not directly accessible from the static ground-state energy calculations performed. However, properties of our simulated array can be used to parameterize simple models which can capture the processes involved. Of main interest is the behavior of the density external to the QD (the surface decay). Attributing the exponential decay of this density (seen in Fig. 8 and quantified by Eq. (86)) to a single, "*s*"-like surface state, one can assign useful properties to the QD (the surface-state wavefunction would be given by the square-root of the density, $\psi_{\text{surf}}(\mathbf{r}) = \sqrt{\rho_{\text{surf}}(\mathbf{r})}$). The energy associated with such a state is given by [23],

$$\epsilon_{\text{surf}} = -\frac{\hbar^2}{2m}\left(\frac{k_c}{2}\right)^2, \tag{87}$$

where k_c is the same as the exponent in Eq. (86). For the Al clusters under study, this corresponds to an energy of ~ -6.1 eV. This energy is not too far removed from the bulk work-function (4.2 eV), and does not take into account any structure of the states associated with the geometry of the cluster surface, or similarly, any higher angular momentum components of the cluster surface wavefunction which would almost certainly be present in our *approximately* spherical cluster. This energy should be considered a lower-bound of the surface-state energy.

The Coulomb blockade can be modeled as a pair of electronic bands. The fully occupied valence band has a lower energy (ϵ_l) than the (totally or partially) empty conduction band (ϵ_u). These bands are formed from the surface-states of the QDs and are separated by a Coulombic blockade energy (ϵ_{C}). This energy arises due to the finite energy cost associated with transference of single electrons between QDs. For isolated QDs, ϵ_{C} is related to capacitance of the QD (C) by [54],

$$\epsilon_{\text{C}} = \frac{e^2}{C}. \tag{88}$$

For well separated clusters (large D), the valence and conduction bands are narrow, with energies,

$$\epsilon_l = \epsilon_{\text{surf}}, \tag{89}$$

$$\epsilon_u = \epsilon_l + \epsilon_{\text{C}} = \epsilon_{\text{surf}} + \epsilon_{\text{C}}. \tag{90}$$

As the QDs come together in the 2-dimensional hexagonal array, interactions between the QDs gives each band width (structure). The hexagonal symmetry of the problem, and the effect on this structure, is similar to the π_z bands in graphite [55]. These bands can be effectively modeled using a tight-binding Hamiltonian [23]. The effect of interactions between neighboring clusters moves the maximum and minimum in each band by up to three times the neighbor interaction strength β,

$$\epsilon_{l,u} = \epsilon_{l,u} \pm 3\beta_{l,u} \tag{91}$$

(β also corresponds to the off-diagonal tight-binding Hamiltonian matrix elements). At the most simple level, β can be approximated using the Wolfsberg–Helmholz formula [56],

$$\beta_{l,u} = \epsilon_{l,u} K S, \tag{92}$$

where the Wolfsberg–Helmholz constant K is symmetry-dependent, and in this instance $K = 3.2$ [56]. Overlap between the clusters can be included using the Wheland correction, which in this instance, replaces our band-splitting 3β, with an effective term [57],

$$\epsilon_{l,u} \pm 3\beta_{l,u} \to \epsilon_{l,u} \pm \frac{3}{1 \pm 3S}(\beta - \epsilon_{l,u} S). \tag{93}$$

The states that are of interest in this study are those that are first to overlap with decreasing separation. These are the highest energy states of the lower (valence) band, and the lowest energy states of the upper (conduction) band. Considering the effects of neighboring interactions and overlap, the energies of these states are modeled by,

$$\epsilon_l = \epsilon_{\text{surf}} - \frac{3\epsilon_{\text{surf}} S(K-1)}{1-3S}, \tag{94}$$

$$\epsilon_u = \epsilon_{\text{surf}} + \epsilon_{\text{C}} + \frac{3(\epsilon_{\text{surf}} + \epsilon_{\text{C}}) S(K-1)}{1+3S}. \tag{95}$$

The overlap (S) is between surface states on neighboring clusters. We can extract this object from our simulation results, as we did for the surface energy, using the exponential fit for our surface density, and the resultant surface wavefunction, obtaining the overlap as,

$$S\big(|\mathbf{R}_1 - \mathbf{R}_2|\big) = \int \mathrm{d}\mathbf{r}\, \psi^*_{\text{surf}}\big(|\mathbf{r} - \mathbf{R}_1|\big)\psi_{\text{surf}}\big(|\mathbf{r} - \mathbf{R}_2|\big). \tag{96}$$

The overlap for the simulated Al clusters is shown in Fig. 9. Like the surface density and wavefunction, it is seen to decay exponentially.

One object required within our tight-binding model, which is unavailable from our current simulations, is the Coulomb blockade energy (ϵ_{C}). For 2.6 nm diameter Ag clusters, this energy was measured to be 0.34 eV [51]. One might expect that the Coulomb blockade for similar size clusters of Al would be of the same magnitude as that seen for Ag, given the similarities in the work functions ($< 5\%$ difference). Fig. 9 shows the tight-binding band energies for our 2 nm Al clusters, using this Coulomb blockade value. Shown, are the highest energy valence state, and the lowest energy conduction state. One can see that, as the simulated array is compressed (decreasing D), the bands get closer together, crossing well before the clusters would be considered to be touching ($D/2R = 1$). Indeed, the bands cross when the overlap is extremely small (~ 0.01).

Of prime importance when performing a tight-binding analysis of electronic conductivity, is the band-gap,

$$\Delta\epsilon = \epsilon_u - \epsilon_l. \tag{97}$$

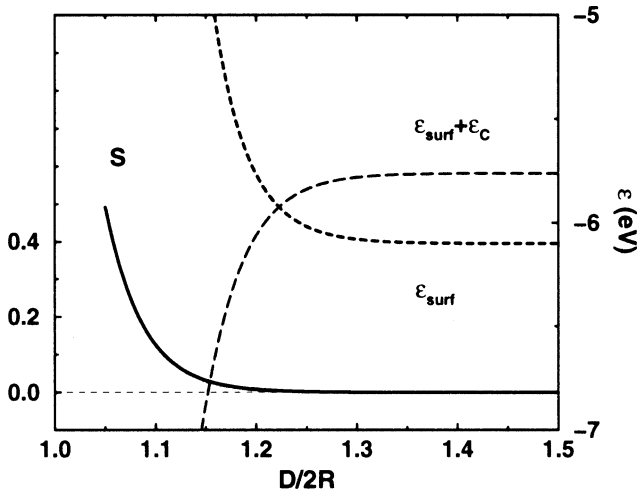

Fig. 9. The overlap of neighboring cluster surface wavefunctions, normalized to obtain the correct density (the left-hand scale). Also shown (right-hand scale) are the highest and lowest energies of the lower and upper bands, respectively, as given by Eqs. (94) and (95), for $\epsilon_{\text{surf}} = -6.1$ eV, and $\epsilon_{\text{C}} = +0.34$ eV.

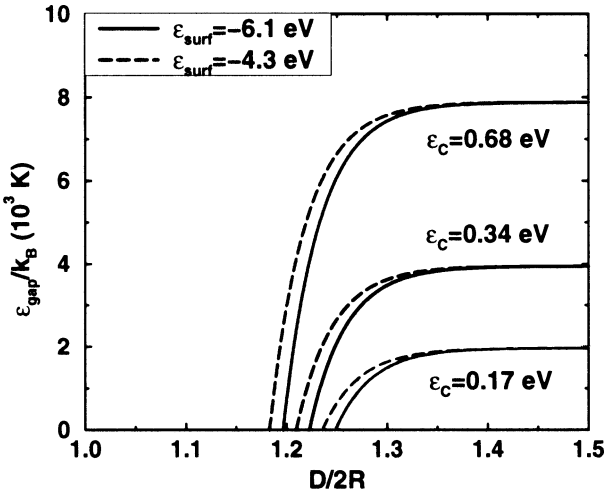

Fig. 10. The band-gap, for the two band model, using surface energies from the OF-KEDF simulation (−6.1 eV), or the negative of the experimental bulk work function (−4.3 eV), and a Coulomb blockade energy taken as that seen for similar sized Ag QDs (0.34 eV), as well as twice and half this value.

For a conduction process to occur, the conduction band must be accessible to the electrons in the valence band, either by the bands overlapping (which would be indicated by $\epsilon_u < \epsilon_l$, i.e. $\Delta\epsilon < 0$) or by thermal excitations. Such excitations, from the valence band to the conduction band, would have a probability which is related to the expression for the Fermi–Dirac occupation probability (f) [23],

$$f(\Delta\epsilon) = \frac{1}{\exp(\Delta\epsilon/k_{\text{B}}T) + 1}. \tag{98}$$

For this to have a significant value, the band-gap ($\Delta\epsilon$) must be of the order of $k_{\text{B}}T$ or less (where T is the sample temperature).

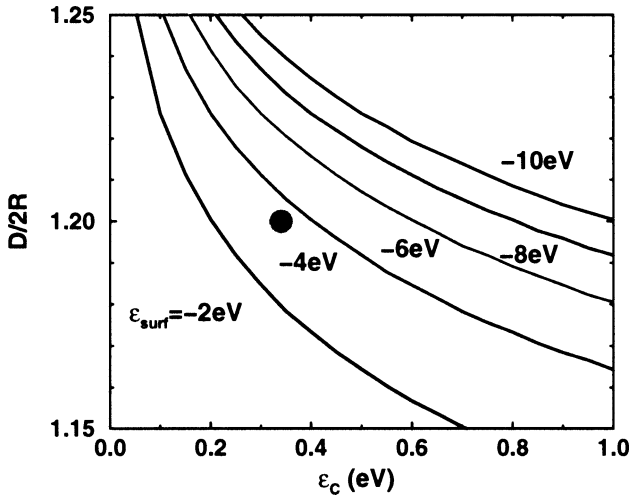

Fig. 11. The position of the band-gap closure, versus Coulomb blockade energy, for several surface energies. The simulated value of the surface energy is −6.1 eV, while negative of the bulk work-function for aluminum is −4.3 eV. Experimental results for 2.6 nm diameter silver clusters, are shown as the circle.

Fig. 10 shows the band gap, as a function of cluster separation, calculated using the tight-binding model (through Eqs. (94), (95), and (97)). The surface energy used is that calculated from the simulated density. Also shown are results which utilize a surface-state energy corresponding to the negative of the work function for bulk aluminum. This change in surface energy (by ∼2 eV) is not seen to have a substantial effect on where the band-gap goes to zero. The Coulomb blockade energy used is the experimental value for the 2.6 nm silver cluster. Also shown are results using twice and half this value, indicating that the band-gap is quite sensitive to the Coulomb blockade energy. While there is a significant change in the band-gap at large separation, all three curves have a gap which closes at very similar values of the separation, $D/2R \sim 1.2$.

The ability to investigate the importance of the parameters within the tight-binding model, allows for a deeper insight into the metal-insulator transition. Fig. 11 shows the position of the band-gap closure, as a function of the (in this case uncertain) Coulomb blockade energy. One can see that, as the blockade energy decreases, the separation at which the metal-insulator transition would occur increases. Also shown is the effect of different surface energies. One can see that the separation (and QD density) at which the metal-insulator transition occurs is affected by both the Coulomb blockade energy, and the energy of the surface state, but that the transition separation is an order of magnitude more sensitive to the Coulomb blockade energy.

The experimentally observed metal-insulator transition in 2.6 nm diameter Ag QDs, occurs at a cluster separation of $D/2R \sim 1.2$. This is approximately the same as that indicated by the closure of the valence-conduction band-gap, calculated for Al, using the same Coulomb blockade energy. However, there are several clear differences (beyond the elemental) between the Ag experiment and the current Al simulation studies. The experimental system contains passivating species (alkylthiols) between the clusters, while the simulation has vacuum. The effect of the passivating groups on the transition can be estimated by considering the passivating groups as an effective dielectric. The alkyl tails, being relatively unpolarizable, would have a low dielectric constant, while the thiol heads, in contact with the metal clusters, are more polarizable, and thus would correspond to a higher dielectric. One would expect this to have the effect of extending the metal-cluster surface state wavefunctions into the region of the thiol heads, while the alkyl tails would act in a similar manner to the vacuum. This would increase the inter-cluster overlap, broadening the bands. Another aspect of the experiment not present in the simulation, is the uncertainty in the cluster sizes (size distributions), and motion of the clusters (thermal fluctuations). Both effects

applied to the simulation would also be expected to broaden the bands. This band broadening would reduce the band-gap, resulting in an earlier onset of the transition.

An important aspect of the simulation results is the sensitivity of the metal-insulator transition to the Coulomb blockade energy, and the relative insensitivity to the surface-state energy. Were one wanting to "tune" the transition, it would be more effective to "turn" the blockade energy "dial", than altering the surface energy. Both properties are strongly (though not equally) dependent upon the electronic properties of the materials used to form the clusters, and the size of clusters prepared. One would expect to be able to create experimental conditions (by doping or enlarging/reducing the metal cluster, for example) which would produce the required electronic properties.

7. Conclusion

Once again, the orbital-free kinetic energy density functional method has been shown to be applicable in the study of complex metallic problems. The new density-dependent-kernel functional allows the meaningful study of systems, well beyond its explicitly exact regime. While the research of metal quantum dots is relatively new, this is to our knowledge, the first self-consistent *ab initio* study of such a system. The problem would appear to be too large for a standard Kohn–Sham treatment, and too "metallic" for the locality reliant $\mathcal{O}(N)$ methods. The OF-KEDF method provided self-consistent densities that, when input into a tight binding model, yielded new in sights into the QD metal-insulator transition. For example, extremely small (0.01) overlap between QD surface wavefunctions is all that is required for the band gap to close.

The linear-scaling method is not only parallel, but efficient enough to allow the simulation of nano-scale processes to be performed on relatively inexpensive workstations (the current work was performed on a 4-way SMP Compaq ES40). The scaling of the electronic calculation is such that, beyond a thousand atoms (a simulation adequately performed on a medium-sized workstation), the electronic calculation is overwhelmed by the evaluation of ionic properties, such as the structure factor, ion–ion and ion–electron forces (terms which currently scale as $N_{\text{grid}}N_{\text{Ion}}$, although efforts are currently underway to reduce this scaling [58]).

Recent reviews of linear-scaling methods appeared to have overlooked the orbital-free schemes. While there is currently a limit to the applicability of the functionals, with their continued improvements, it is likely only a matter of time before accurate universal functionals of the density are developed. "... it is not unduly optimistic to say that $\rho(\mathbf{r})$ may be the unifying link between the microscopic world and our perception of it" [59].

Acknowledgements

We thank Dr. Niranjan Govind and Dr. Yan Alexander Wang for helpful discussions. Financial support for this work was provided by the National Science Foundation, The Army Research Office, and the Air Force Office of Scientific Research.

References

[1] C.P. Collier et al., Science 277 (1997) 1978.
[2] P.J. Estrup, Chemistry and Physics of Solid Surfaces V, R. Vanselow, R. Howe (Eds.) (Springer, Berlin, 1984) Section 9.
[3] A. Szabo, N.S. Ostlund, Modern Quantum Chemistry (McGraw-Hill, New York, 1989).
[4] R.G. Parr, W. Yang, Density-Functional Theory of Atoms and Molecules (Oxford University Press, New York, 1989).
[5] R.M. Dreizler, E.K.U. Gross, Density Functional Theory: An Approach to the Quantum Many-Body Problem (Springer, Berlin, 1990).
[6] M.C. Payne et al., Rev. Mod. Phys. 64 (1992) 1045.
[7] S. Goedecker, Rev. Mod. Phys. 71 (1999) 1085.
[8] S. Ismail-Beigi, T.A. Arias, Phys. Rev. Lett. 82 (1999) 2127.
[9] S. Goedecker, Phys. Rev. B 58 (1998) 3501.

[10] R. Baer, M. Head-Gordon, Phys. Rev. B 58 (1998) 15 296.
[11] R. Baer, M. Head-Gordon, J. Chem. Phys. 109 (1998) 10 159.
[12] P. Hohenberg, W. Kohn, Phys. Rev. B 136 (1964) 864.
[13] M. Levy, J.P. Perdew, Phys. Rev. A 32 (1985) 2010.
[14] A.D. Becke, Phys. Rev. A 38 (1988) 3098.
[15] J.P. Perdew, Y. Wang, Phys. Rev. B 45 (1992) 13 244.
[16] F.A. Hamprecht, A.J. Cohen, D.J. Tozer, N.C. Handy, J. Chem. Phys. 109 (1998) 6264.
[17] W. Yang, J. Chem. Phys. 109 (1998) 10 107.
[18] D. Joubert, J. Chem. Phys. 110 (1999) 1873.
[19] D.M. Ceperley, B.J. Alder, Phys. Rev. Lett. 45 (1980) 566.
[20] J.P. Perdew, A. Zunger, Phys. Rev. B 23 (1981) 5048.
[21] W. Kohn, L.J. Sham, Phys. Rev. 140 (1965) 1133.
[22] G. Galli, M. Parrinello, Phys. Rev. Lett. 69 (1992) 3547.
[23] N.W. Ashcroft, N.D. Mermin, Solid State Physics (Holt-Saunders, Philadelphia, 1976).
[24] L.H. Thomas, Proc. Cambridge Philos. Soc. 23 (1927) 542.
[25] E. Fermi, Z. Phys. 48 (1928) 73.
[26] C.F. von Weizsäcker, Z. Phys. 96 (1935) 431.
[27] L.-W. Wang, M.P. Teter, Phys. Rev. B 45 (1992) 13 196.
[28] M. Pearson, E. Smargiassi, P.A. Madden, J. Phys. Cond. Matter 5 (1993) 3221.
[29] E. Smargiassi, P.A. Madden, Phys. Rev. B 49 (1994) 5220.
[30] F. Perrot, J. Phys. Cond. Matt. 6 (1994) 431.
[31] M. Foley, P.A. Madden, Phys. Rev. B 53 (1996) 10 589.
[32] Y.A. Wang, N. Govind, E.A. Carter, Phys. Rev. B 58 (1998) 13 465.
[33] Y.A. Wang, N. Govind, E.A. Carter, Phys. Rev. B 60 (1999) 16 350.
[34] E. Chacón, J.E. Alverllos, P. Tarzona, Phys. Rev. B 32 (1985) 7868.
[35] Y.A. Wang, Phys. Rev. A 55 (1997) 4589.
[36] M. Foley, D.Phil. Thesis, Oxford University, UK (1995).
[37] S.C. Watson, E.A. Carter (unpublished).
[38] S. Watson, B.J. Jesson, E.A. Carter, P.A. Madden, Europhys. Lett. 41 (1998) 37.
[39] S.C. Watson, P.A. Madden, Phys. Chem. Commun. (1998) 1. (http://www.rsc.org/is/journals/current/PhysChemComm/pccpub.htm).
[40] W.H. Press, B.P. Flannery, S.A. Teukolsky, W.T. Vetterling, Numerical Recipes (Cambridge University Press, Cambridge, 1986).
[41] S.C. Watson, D.Phil. Thesis, Oxford University, UK (1996).
[42] M.N. Pearson, Masters Thesis, Oxford University (1992).
[43] V. Shah, D. Nehete, D.G. Kanhere, J. Phys. Cond. Matt. 6 (1994) 10 773.
[44] N. Govind, J. Wang, H. Guo, Phys. Rev. B 50 (1994) 11 175.
[45] A. Vichare, D.G. Kanhere, J. Phys. Cond. Matt. 10 (1998) 3309.
[46] M.P. Allen, D.J. Tildesley, Computer Simulation of Liquids (Clarendon Press, Oxford, 1987).
[47] L. Kleinman, D.M. Bylander, Phys. Rev. Lett. 48 (1982) 1425.
[48] P.E. Blöchl, Phys. Rev. B 41 (1990) 5414.
[49] J.R. Heath, C.M. Knobler, D.V. Leff, J. Phys. Chem. B 101 (1997) 189.
[50] G. Markovich, C.P. Collier, J.R. Heath, Phys. Rev. Lett. 80 (1998) 3807.
[51] G. Medeiros-Ribeiro, D.A.A. Ohlberg, E.S. Williams, J.R. Heath, Phys. Rev. B 59 (1999) 1633.
[52] N.F. Mott, Metal-Insulator Transitions (Taylor & Francis, London, 1990).
[53] N. Troullier, J.L. Martins, Phys. Rev. B 43 (1991) 1993.
[54] C.W.J. Beenakker, Phys. Rev. B 44 (1991) 1646.
[55] J.K. Burdett, Chemical Bonding in Solids (Oxford University Press, New York, 1995).
[56] M. Wolfsberg, L. Helmholz, J. Chem. Phys. 20 (1952) 837.
[57] G.W. Wheland, J. Amer. Chem. Soc. 63 (1941) 2025.
[58] S.C. Watson, E.A. Carter (unpublished).
[59] A.S. Bamzai, B.M. Deb, Rev. Mod. Phys. 53 (1981) 95.