The **Siesta** method for ab initio order-\(N\) materials simulation

José M. Soler,\(^1\) Emilio Artacho,\(^2\) Julian D. Gale,\(^3\) Alberto García,\(^4\) Javier Junquera,\(^1,\(^5\) Pablo Ordejón,\(^6\) and Daniel Sánchez-Portal\(^7\)

\(^1\) Dep. de Física de la Materia Condensada, C-III, Universidad Autónoma de Madrid, E-28049 Madrid, Spain
\(^2\) Department of Earth Sciences, University of Cambridge, Downing St., Cambridge CB2 3EQ, United Kingdom
\(^3\) Department of Chemistry, Imperial College of Science, Technology and Medicine, South Kensington SW7 2AY, United Kingdom
\(^4\) Departamento de Física de la Materia Condensada, Universidad del País Vasco, Apt. 644, 48080 Bilbao, Spain
\(^5\) Institut de Physique, Bâtiment B5, Université de Liège, B-4000 Sart-Tilman, Belgium
\(^6\) Instituto de Ciencia de Materiales de Barcelona, CSIC, Campus de la UAB, Bellaterra, 08193 Barcelona, Spain
\(^7\) Dep. de Física de Materiales and DIPC, Facultad de Química, UPV/EHU, Apt. 1072, 20080 Donostia, Spain

(Dated: February 1, 2008)

We have developed and implemented a self-consistent density functional method using standard norm-conserving pseudopotentials and a flexible, numerical LCAO basis set, which includes multiple-zeta and polarization orbitals. Exchange and correlation are treated with the local spin density or generalized gradient approximations. The basis functions and the electron density are projected on a real-space grid, in order to calculate the Hartree and exchange-correlation potentials and matrix elements, with a number of operations that scales linearly with the size of the system. We use a modified energy functional, whose minimization produces orthogonal wavefunctions and the same energy and density as the Kohn-Sham energy functional, without the need of an explicit orthogonalization. Additionally, using localized Wannier-like electron wavefunctions allows the computation time and memory, required to minimize the energy, to also scale linearly with the size of the system. Forces and stresses are also calculated efficiently and accurately, thus allowing structural relaxation and molecular dynamics simulations.

I. INTRODUCTION

As the improvements in computer hardware and software allow the simulation of molecules and materials with an increasing number of atoms \(N\), the use of so-called order-\(N\) algorithms, in which the computer time and memory scales linearly with the simulated system size, becomes increasingly important. These \(O(N)\) methods were developed during the 1970’s and 80’s for long range forces and empirical interatomic potentials, but only in the last 5-10 years for the much more complex quantum-mechanical methods, in which atomic forces are obtained by solving the interaction of ions and electrons together. Even among quantum mechanical methods, there are very different levels of approximation: empirical or semiempirical orthogonal tight-binding methods are the simplest ones; ‘ab-initio’ nonorthogonal-tight-binding and non-self-consistent Harris-functional methods are next; and fully self-consistent density functional theory (DFT) methods are the most complex and reliable. The implementation of \(O(N)\) methods in quantum-mechanical simulations has also followed these steps, with several methods already well established within the tight-binding formalism, but much less so in self-consistent DFT. The latter also require, in addition to solving Schrödinger’s equation, the determination of the self-consistent Hamiltonian in \(O(N)\) iterations. While this is difficult using plane waves, a localized basis set appears to be the natural choice. One proposed approach are the ‘blips’ of Hernandez and Gillan, regularly-spaced Gaussian-like splines that can be systematically increased, in the spirit of finite-element methods, although at a considerable computational cost.

We have developed a fully self-consistent DFT, based on a flexible linear combination of atomic orbitals (LCAO) basis set, with essentially perfect \(O(N)\) scaling. It allows extremely fast simulations using minimal basis sets and very accurate calculations with complete multiple-zeta and polarized bases, depending on the required accuracy and available computational power. In previous papers, we have described preliminary versions of this method, that we call **Siesta** (Spanish Initiative for Electronic Simulations with Thousands of Atoms). There is also a review of the tens of studies performed with it, in a wide variety of systems, like metallic surfaces, nanotubes, and biomolecules. In this work we present a more complete description of the method, as well as some important improvements.

Apart from that of Born and Oppenheimer, the most basic approximations concern the treatment of exchange and correlation, and the use of pseudopotentials. Exchange and correlation (XC) are treated within Kohn-Sham DFT. We allow for both the local (spin) density approximation (LDA/LSD) or the generalized gradient approximation (GGA). We use standard norm-conserving pseudopotentials in their fully non-local form. We also include scalar-relativistic effects and the
nonlinear partial-core-correction to treat exchange and correlation in the core region.24

The SIESTA code has been already tested and applied to dozens of systems and a variety of properties. Therefore, we will just illustrate here the convergence of a few characteristic magnitudes of silicon, the archetypical system of the field, with respect to the main precision parameters that characterize our method: basis size (number of atomic basis orbitals); basis range (radius of the basis orbitals); fineness of the real-space integration grid; and confinement radius of the Wannier-like electron states. Other parameters, like the $k$-sampling integration grid, are common to all similar methods and we will not discuss their convergence here.

II. PSEUDOPOTENTIAL

Although the use of pseudopotentials is not strictly necessary with atomic basis sets, we find them convenient to get rid of the core electrons and, more importantly, to allow for the expansion of a smooth (pseudo)charge density on a uniform spatial grid. The theory and usage of first principles norm-conserving pseudopotentials is already well established. SIESTA reads them in semilocal form (a different radial potential for each angular momentum $l$, optionally generated scalar-relativistically) from a data file that users can fill with their preferred choice. We generally use the Troullier-Martins parameterization.21,22 We transform this semilocal form into the fully non-local form proposed by Kleinman and Bylander (KB)23:

$$\hat{V}^{PS} = V_{\text{local}}(r) + \hat{V}^{KB}$$

$$\hat{V}^{KB} = \sum_{l=0}^{l_{\text{max}}} \sum_{m=-l}^{m_{l}} \sum_{n=1}^{N_{l}^{KB}} |\chi_{l}^{KB}(r)| \langle \chi_{l}^{KB} | V_{\text{local}} | \chi_{l}^{KB} \rangle$$

where $\delta V_{l}(r) = V_{l}(r) - V_{\text{local}}(r)$. $\chi_{l}^{KB}(r)$ are spherical harmonics) are the KB projection functions

$$\chi_{l}^{KB}(r) = \delta V_{l}(r) \phi_{l}(r).$$

The functions $\phi_{l}$ are obtained from the eigenstates $\psi_{l}$ of the semilocal pseudopotential (screened by the pseudo-valence charge density) at energy $\epsilon_{l}$ using the orthogonalization scheme proposed by Blöch:24

$$\phi_{l}(r) = \psi_{l}(r) - \sum_{n'=1}^{n-1} \phi_{l_{n}'} \delta V_{l_{n}'}(r) \psi_{l_{n}}(r)$$

where

$$\delta V_{l}(r) = V_{l}(r) - V_{\text{local}}(r).$$

$V_{H}$ and $V_{xc}$ are the Hartree and XC potentials for the pseudo-valence charge density, and we are using atomic units ($\hbar = m_e = 1$) throughout all of this paper.

The local part of the pseudopotential $V_{\text{local}}(r)$ is in principle arbitrary, but it must join the semilocal potentials $V_{l}(r)$ which, by construction, all become equal to the (unscreened) all-electron potential beyond the pseudopotential core radius $r_{\text{core}}$. Thus, $\delta V_{l}(r) = 0$ for $r > r_{\text{core}}$. Ramer and Rappe have proposed that $V_{\text{local}}(r)$ be optimized for transferability,25 but most plane wave schemes make it equal to one of the $V_{l}(r)$’s for reasons of efficiency. Our case is different because $V_{\text{local}}(r)$ is the only pseudopotential part that needs to be represented in the real space grid, while the matrix elements of the non-local part $V_{KB}$ are cheaply and accurately calculated by two-center integrals. Therefore, we optimize $V_{\text{local}}(r)$ for smoothness, making it equal to the potential created by a positive charge distribution of the form

$$\rho_{\text{local}}(r) \propto \exp[-(\sinh(abr)/\sinh(b))^{2}],$$

where $a$ and $b$ are chosen to provide simultaneously optimal real-space localization and reciprocal-space convergence.26,27 After some numerical tests we have taken $b = 1$ and $a = 1.82/r_{\text{core}}$. Figure 1 shows $V_{\text{local}}(r)$ for silicon.

Since $V_{l}(r) = V_{\text{local}}(r)$ outside $r_{\text{core}}$, $\chi_{l}^{KB}(r)$ is strictly zero beyond that radius, irrespective of the value of $\epsilon_{l}$. Generally it is sufficient to have a single projector $\chi_{l}^{KB}$ for each angular momentum (i.e. a single term in the sum on $n$). In this case we follow the normal practice of making $\epsilon_{l}$ equal to the valence atomic eigenvalue $\epsilon_{l}$, and the function $\phi_{l}(r)$ in Eq. 3 is identical to the corresponding eigenstate $\psi_{l}(r)$. In some cases, particularly for alkaline metals, alkaline earths, and transition metals of the first few columns, we have sometimes found it necessary to include the sicoare states together with the valence states. In these cases, we also include two independent KB projectors, one for the sicoare and one for the valence states. However, our pseudopotentials are still norm-conserving rather than “ultrasoft.”28 This is because, in our case, it is only the electron density that needs to be accurately represented in a real-space grid, rather than each wavefunction. Therefore, the ultrasoft pseudopotential formalism does not apply in SIESTA the same savings as it does in PW schemes. Also, since the non-local part of the pseudopotential is a relatively cheap operator within SIESTA, we generally (but not necessarily) use a larger than usual value of $r_{\text{max}}^{KB}$ in Eq. (3), making it one unit larger than the $l_{\text{max}}$ of the basis functions.

III. BASIS SET

Order-N methods rely heavily on the sparsity of the Hamiltonian and overlap matrices. This sparsity requires either the neglect of matrix elements that are small enough or the use of strictly confined basis orbitals, i.e., orbitals that are zero beyond a certain radius.29 We have
adopted this latter approach because it keeps the energy strictly variational, thus facilitating the test of the convergence with respect to the radius of confinement. Within this radius, our atomic basis orbitals are products of a numerical radial function times a spherical harmonic. For atom $I$, located at $\mathbf{R}_I$, 

$$\phi_{I\text{mn}}(\mathbf{r}) = \phi_{I\text{mn}}(r_I) Y_{lm}(\hat{r}_I)$$  

(8)

where $\mathbf{r}_I = \mathbf{r} - \mathbf{R}_I$, $r = |\mathbf{r}|$ and $\hat{r} = \mathbf{r}/r$. The angular momentum (labelled by $l, m$) may be arbitrarily large and, in general, there will be several orbitals (labelled by index $n$) with the same angular dependence, but different radial dependence, which is conventionally called a ‘multiple-\(\zeta\)’ basis. The radial functions are defined by a cubic spline interpolation from the values given on a fine radial mesh. Each radial function may have a different cutoff radius and, up to that radius, its shape is completely free and can be introduced by the user in an input file. In practice, it is also convenient to have an automatic procedure to generate sufficiently good basis sets. We have developed several such automatic procedures, and we will describe here one of them for completeness, even though we stress that the generation of the basis set, like that of the pseudopotential is to a large extent up to the user and independent of the SIESTA method itself.

In the case of a minimal (single-\(\zeta\)) basis set, we have found convenient and efficient the method of Sankey and Niklewski. Their basis orbitals are the eigenfunctions of the (pseudo)atom within a spherical box (although the radius of the box may be different for each orbital, see below). In other words, they are the (angular-momentum-dependent) numerical eigenfunctions $\phi_I(r)$ of the atomic pseudopotential $V_I(r)$, for an energy $\epsilon_I + \delta \epsilon$ chosen so that the first node occurs at the desired cutoff radius $r^*_I$:

$$\left(-\frac{1}{2r}\frac{d^2}{dr^2} + \frac{l(l+1)}{2r^2} + V_I(r)\right) \phi_I(r) = (\epsilon_I + \delta \epsilon)\phi_I(r)$$  

(9)

with $\phi_I(r^*_I) = 0$ (we omit indices $I$ and $n$ here for simplicity). In order to obtain a well balanced basis, in which the effect of the confinement is similar for all the orbitals, it is usually better to fix a common ‘energy shift’ $\delta \epsilon$, rather than a common radius $r^*$, for all the atoms and angular momenta. This means that the orbital radii depend on the atomic species and angular momentum.

One obvious possibility for multiple-\(\zeta\) bases is to use pseudopotential eigenfunctions with an increasing number of nodes. They have the virtue of being orthogonal and asymptotically complete. However, the efficiency of this kind of basis depends on the radii of confinement of the different orbitals, since the excited states of the pseudopotential are usually unbound. Thus, in practice we have found this procedure rather inefficient. Another possibility is to use the atomic eigenstates for different ionization states.\[2\] We have implemented a different scheme\[3\], based on the ‘split-valence’ method which is standard in quantum chemistry\[4\]. In that method, the first-\(\zeta\) basis orbitals are ‘contracted’ (i.e. fixed) linear combinations of Gaussians, determined either variationally or by fitting numerical atomic eigenfunctions. The second-\(\zeta\) orbital is then one of the Gaussians (generally the slowest-decaying one) which is ‘released’ or ‘split’ from the contracted combination. Higher-\(\zeta\) orbitals are generated in a similar way by releasing more Gaussians. Our scheme adapts this split-valence method to our numerical orbitals. Following the same spirit, our second-\(\zeta\) functions $\phi_{2\zeta}(r)$ have the same tail as the first-\(\zeta\) orbitals $\phi_{1\zeta}(r)$ but change to a simple polynomial behaviour inside a ‘split radius’ $r^*_1$:

$$\phi_{2\zeta}(r) = \begin{cases} r^l(a_l - b_l r^2) & \text{if } r < r^*_1 \\ \phi_{1\zeta}(r) & \text{if } r \geq r^*_1 \end{cases}$$  

(10)

where $a_l$ and $b_l$ are determined by imposing the continuity of value and slope at $r^*_1$. These orbitals therefore combine the decay of the atomic eigenfunctions with a smooth and featureless behaviour inside $r^*_1$. We have found it convenient to set the radius $r^*_1$ by fixing the norm of $\phi_{1\zeta}$ in $r > r^*_1$. We have found empirically that a reasonable value for this ‘split-norm’ is $\sim 0.15$. Actually,
instead of $\phi_l^2 \zeta$ thus defined, we use $\phi_l^1 \zeta - \phi_l^2 \zeta$, which is zero beyond $r_l^*$, to reduce the number of nonzero matrix elements, without any loss of variational freedom.

To achieve well converged results, in addition to the atomic valence orbitals, it is generally necessary to also include polarization orbitals, to account for the deformation induced by bond formation. Again, using pseudoatomic orbitals of higher angular momentum is frequently unsatisfactory, because they tend to be too extended, or even unbound. Instead, consider a valence pseudoatomic orbital $\phi_l(r) = \phi_l(r)Y_l^m(\hat{r})$, such that there are no valence orbitals with angular momentum $l + 1$. To polarize it, we apply a small electric field $E$ in the $z$-direction. Using first-order perturbation theory

\[(H - E)\delta \phi = - (\delta H - \delta E)\phi, \tag{11}\]

where $\delta H = E_z$ and $\delta E = \langle \phi | \delta H | \phi \rangle = 0$ because $\delta H$ is odd. Selection rules imply that the resulting perturbed orbital will only have components with $l' = l \pm 1, m' = m$:

\[\delta H \phi_l^m(r) = \langle \phi(r) | E_z | Y_l^m(\hat{r}) \rangle = E r \phi_l(r)(c_{l-1} Y_{l-1,m} + c_{l+1} Y_{l+1,m}) \tag{12}\]

and

\[\delta \phi_l^m(r) = \varphi_{l-1}(r) Y_{l-1,m}(\hat{r}) + \varphi_{l+1}(r) Y_{l+1,m}(\hat{r}). \tag{13}\]

Since in general there will already be orbitals with angular momentum $l - 1$ in the basis set, we select the $l + 1$ component by substituting (12) and (13) in (11), multiplying by $Y_{l+1,m}^*(\hat{r})$ and integrating over angular variables. Thus we obtain the equation

\[-\frac{1}{2r} \frac{d^2}{dr^2} r + \frac{(l+1)(l+2)}{2r^2} + V_l(r) - E_l \varphi_{l+1}(r) = -r \phi_l(r) \tag{14}\]

where we have also eliminated the factors $E$ and $c_{l+1}$, which affect only the normalization of $\varphi_{l+1}$. The polarization orbitals are then added to the basis set: $\phi_{l+1,m}^*(\hat{r}) = N \varphi_{l+1}(r) Y_{l+1,m}(\hat{r})$, where $N$ is a normalization constant.

We have found that the previously described procedures generate reasonable minimal single-ζ (SZ) basis sets, appropriate for semiquantitative simulations, and double-ζ plus polarization (DZP) basis sets that yield high quality results for most of the systems studied. We thus refer to DZP as the ‘standard’ basis, because it usually represents a good balance between well converged results and a reasonable computational cost. In some cases (typically alkali and some transition metals), semiconductor states also need to be included for good quality results. More recently, we have obtained extremely efficient basis sets optimized variationally in molecules or solids. Figure 2 shows the performance of these atomic basis sets compared to plane waves, using the same pseudopotentials and geometries. It may be seen that the SZ bases are comparable to planewave cutoffs typically used in Car-Parrinello molecular dynamics simulations, while DZP sets are comparable to the cutoffs used in geometry relaxations and energy comparisons. As expected, the LCAO is far more efficient, typically by a factor of 10 to 20, in terms of number of basis orbitals. This efficiency must be balanced against the faster algorithms available for plane waves, and our main motivation for using an LCAO basis is its suitability for O(N) methods. Still, we have generally found that, even without using the O(N) functional, SIESTA is considerably faster than a plane wave calculation of similar quality.

Figure 3 shows the convergence of the total energy curve of silicon, as a function of lattice parameter, for different basis sizes, and table 3 summarizes the same information numerically. It can be seen that the ‘standard’ DZP basis offers already quite well converged results, comparable to those used in practice in most plane wave calculations.

Figure 4 shows the dependence of the lattice constant, bulk modulus, and cohesive energy of bulk silicon with the range of the basis orbitals. It shows that a cutoff radius of 3 Å for both s and p orbitals yields already very well converged results, specially when using a ‘standard’ DZP basis.
FIG. 3: Total energy per atom versus lattice constant for bulk silicon, using different basis sets, noted as in Fig. 2. PW refers to a very well converged (50 Ry cutoff) plane wave calculation. The dotted line joins the minima of the different curves.

TABLE I: Comparisons of the lattice constant \( a \), bulk modulus \( B \), and cohesive energy \( E_c \) for bulk Si, obtained with different basis sets. The basis notation is as in Fig. 2. PW refers to a 50 Ry-cutoff plane wave calculation. The LAPW results were taken from ref. 37, and the experimental values from ref. 38.

<table>
<thead>
<tr>
<th>Basis</th>
<th>( a ) (Å)</th>
<th>( B ) (GPa)</th>
<th>( E_c ) (eV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SZ</td>
<td>5.521</td>
<td>88.7</td>
<td>4.722</td>
</tr>
<tr>
<td>DZ</td>
<td>5.465</td>
<td>96.0</td>
<td>4.841</td>
</tr>
<tr>
<td>TZ</td>
<td>5.453</td>
<td>98.4</td>
<td>4.908</td>
</tr>
<tr>
<td>SZP</td>
<td>5.424</td>
<td>97.8</td>
<td>5.227</td>
</tr>
<tr>
<td>DZP</td>
<td>5.389</td>
<td>96.6</td>
<td>5.329</td>
</tr>
<tr>
<td>TZP</td>
<td>5.387</td>
<td>97.5</td>
<td>5.355</td>
</tr>
<tr>
<td>TZDP</td>
<td>5.389</td>
<td>96.0</td>
<td>5.340</td>
</tr>
<tr>
<td>TZTP</td>
<td>5.387</td>
<td>96.0</td>
<td>5.342</td>
</tr>
<tr>
<td>TZTPF</td>
<td>5.385</td>
<td>95.4</td>
<td>5.359</td>
</tr>
<tr>
<td>PW</td>
<td>5.384</td>
<td>95.9</td>
<td>5.369</td>
</tr>
<tr>
<td>LAPW</td>
<td>5.41</td>
<td>96</td>
<td>5.28</td>
</tr>
<tr>
<td>Exp.</td>
<td>5.43</td>
<td>98.8</td>
<td>4.63</td>
</tr>
</tbody>
</table>

IV. ELECTRON HAMILTONIAN

Within the non-local-pseudopotential approximation, the standard Kohn-Sham one-electron Hamiltonian may be written as

\[
\hat{H} = \hat{T} + \sum_I V_I^{local}(\mathbf{r}) + \sum_I \hat{V}_I^{KB} + V^H(\mathbf{r}) + V^{xc}(\mathbf{r})
\]

where \( \hat{T} = -\frac{1}{2} \nabla^2 \) is the kinetic energy operator, \( I \) is an atom index, \( V^H(\mathbf{r}) \) and \( V^{xc}(\mathbf{r}) \) are the total Hartree and XC potentials, and \( V_I^{local}(\mathbf{r}) \) and \( \hat{V}_I^{KB} \) are the local and non-local (Kleinman-Bylander) parts of the pseudopotential of atom \( I \).

In order to eliminate the long range of \( V_I^{local} \), we screen it with the potential \( V_I^{atom} \), created by an atomic electron density \( \rho_I^{atom} \), constructed by populating the basis functions with appropriate valence atomic charges. Notice that, since the atomic basis orbitals are zero beyond the cutoff radius \( r_c^I = \text{max}_{l} (r_{c_I}^l) \), the screened ‘neutral-atom’ (NA) potential \( V_I^{NA} \equiv V_I^{local} + V_I^{atom} \) is also zero beyond this radius (see Fig. 1). Now let \( \delta \rho(\mathbf{r}) \) be the difference between the self-consistent electron density \( \rho(\mathbf{r}) \) and the sum of atomic densities \( \rho_I^{atom} = \sum_I \rho_I^{atom} \), and let \( \delta V^H(\mathbf{r}) \) be the electrostatic potential generated by \( \delta \rho(\mathbf{r}) \),
which integrates to zero and is usually much smaller than \( \rho(r) \). Then the total Hamiltonian may be rewritten as

\[
\hat{H} = \hat{T} + \sum_i \hat{V}^{KB}_i + \sum_i V^{NA}_i(r) + \delta V^H(r) + V^{xc}(r)
\]

The matrix elements of the first two terms involve only two-center integrals which are calculated in reciprocal space and tabulated as a function of interatomic distance. The remaining terms involve potentials which are calculated on a three-dimensional real-space grid. We consider these two approaches in detail in the following sections.

**V. TWO-CENTER INTEGRALS**

The overlap matrix and the largest part of the Hamiltonian matrix elements are given by two-center integrals. We calculate these integrals in Fourier space, as proposed by Sankey and Niklewski, but we use some implementation details explained in this section. Let us consider first overlap integrals of the form

\[
S(R) \equiv \langle \psi_1 | \psi_2 \rangle = \int \psi_1^* (r) \psi_2 (r - R) dr,
\]

where the integral is over all space and \( \psi_1, \psi_2 \) may be basis functions \( \phi_{lm} \), KB pseudopotential projectors \( \chi_{lmn} \), or more complicated functions centered on the atoms. The function \( S(R) \) can be seen as a convolution: we take the Fourier transform

\[
\psi(k) = \frac{1}{(2\pi)^{3/2}} \int \psi(r) e^{-i kr} dr
\]

where we use the same symbol \( \psi \) for \( \psi(r) \) and \( \psi(k) \), as its meaning is clear from the different arguments. We also use the plane wave expression of Dirac’s delta function, \( \int e^{i(k' - k) r} dr = (2\pi)^3 \delta(k' - k) \), to find the usual result that the Fourier transform of a convolution in real space is a simple product in reciprocal space:

\[
S(R) = \int \psi_1^* (k) \psi_2 (k) e^{-i kR} dk
\]

Let us assume now that the functions \( \psi(r) \) can be expanded exactly with a finite number of spherical harmonics:

\[
\psi(r) = \sum_{l = 0}^{l_{max}} \sum_{m = -l}^{l} \psi_{lm}(r) Y_{lm}(\hat{r})
\]

\[
\psi_{lm}(r) = \int_0^\pi \sin \theta d\theta \int_0^{2\pi} d\varphi Y_{lm}^* (\theta, \varphi) \psi(r, \theta, \varphi)
\]

This is clearly true for basis functions and KB projectors, which contain a single spherical harmonic, and also for functions like \( x\psi(r) \), which appear in dipole matrix elements. We now substitute in the expansion of a plane wave in spherical harmonics

\[
e^{i k r} = \sum_{l = 0}^{\infty} \sum_{m = -l}^{l} 4\pi i^{l} j_l(kr) Y_{lm}^*(\hat{k}) Y_{lm}(\hat{r}),
\]

to obtain

\[
\psi(k) = \sum_{l = 0}^{l_{max}} \sum_{m = -l}^{l} \psi_{lm}(k) Y_{lm}(\hat{k})
\]

\[
\psi_{lm}(k) = \frac{2}{\pi} (-i)^l \int_0^\infty r^2 dr j_l(kr) \psi_{lm}(r).
\]

Substituting now (23) and (22) into (19) we obtain

\[
S(R) = \sum_{l = 0}^{2l_{max}} \sum_{m = -l}^{l} S_{lm}(R) Y_{lm}(\hat{R})
\]

where

\[
S_{lm}(R) = \sum_{l_1, l_2, m_1, m_2} G_{l_1, l_2, m_1, m_2} S_{l_1, l_2, m_1, m_2}(R)
\]

\[
G_{l_1, l_2, m_1, m_2} = \int_0^\pi \sin \theta d\theta \int_0^{2\pi} d\varphi Y_{l_1 m_1}^* (\theta, \varphi) Y_{l_2 m_2} (\theta, \varphi) Y_{lm} (\theta, \varphi)
\]

\[
S_{l_1, l_2, m_1, m_2}(R) = 4\pi i^{l_1 - l_2 - l} \int_0^\infty k^2 dk j_l(kR)
\]

\[
	imes \int \psi_{l_1 m_1}^* (k) i^m j_l(k) Y_{l_2 m_2} (k).
\]

Notice that \( i^{-l_1} \psi_{l_1}(k), i^m j_l(k), \) and \( i^{l_1 - l_2 - l} \) are all real, since \( l_1 - l_2 - l \) is even for all \( l \)'s for which \( G_{l_1, l_2, m_1, m_2} \neq 0 \). The Gaunt coefficients \( G_{l_1, l_2, m_1, m_2} \) can be obtained by recursion from Clebsch-Gordan coefficients. However, we use real spherical harmonics for computational efficiency:

\[
Y_{lm}(\theta, \varphi) = C_{lm} P_m^n (\cos \theta) \times \left\{ \begin{array}{ll}
\sin (m \varphi) & \text{if } m < 0 \\
\cos (m \varphi) & \text{if } m \geq 0
\end{array} \right.
\]

where \( P_m^n(z) \) are the associated Legendre polynomials and \( C_{lm} \) normalization constants. This does not affect the validity of any of previous equations, but it modifies the value of the Gaunt coefficients. Therefore, we find it is simpler and more general to calculate \( G_{l_1, l_2, m_1, m_2} \) directly from Eq. (27). To do this, we use a Gaussian quadrature,

\[
\int_0^\pi \sin \theta d\theta \int_0^{2\pi} d\varphi \rightarrow 4\pi \frac{1}{N_\theta} \sum_{i = 1}^{N_\theta} \sin \theta_i \frac{1}{N_\varphi} \sum_{j = 1}^{N_\varphi} \psi_{l_i m_i}^* (k) i^m j_l(k) Y_{l_2 m_2} (k)
\]
with \( N_z = 1 + 3l_{\text{max}}, N_\theta = 1 + \text{int}(3l_{\text{max}}/2) \), and the points \( \cos \theta_i \) and weights \( w_i \) are calculated as described in ref. [21]. This quadrature is exact in equation (27) for spherical harmonics \( Y_{lm} \) (real or complex) of \( l \leq l_{\text{max}} \), and it can be used also to find the expansion of \( \psi(r) \) in spherical harmonics (eq. (21)).

The coefficients \( G_{l_1,m_1,l_2,m_2,lm} \) are universal and they can be calculated and stored once and for all. The functions \( S_{l_1,m_1,l_2,m_2,lm}(R) \) depend, of course, on the functions \( \psi_{1,2}(r) \) being integrated. For each pair of functions, they can be calculated and stored in a fine radial grid \( R_i \), up to the maximum distance \( R_{\text{max}} = r_1^c + r_2^c \) at which \( \psi_1 \) and \( \psi_2 \) overlap. Their value at an arbitrary distance \( R \) can then be obtained very accurately using a spline interpolation.

Kinetic matrix elements \( T(R) \equiv \langle \psi^*_1 \mid -\frac{1}{2} \nabla^2 \mid \psi_2 \rangle \) can be obtained in exactly the same way, except for an extra factor \( k^2 \) in Eq. (28):

\[
T_{l_1,m_1,l_2,m_2,lm}(R) = 4\pi i^{l_1-l_2} \int_0^\infty \frac{1}{2} k^4 dk j_l(kR) \times i^{-l_1} \psi^{*}_{1,l_1,m_1}(k) j_{l_2} \psi_{2,l_2,m_2}(k). \tag{31}
\]

Since we frequently use basis orbitals with a kind \[ we need rather fine radial grids to obtain accurate kinetic matrix elements, and we typically use grid cutoffs of more than 2000 Ry for this purpose. Once obtained, the fine grid does not penalize the execution time, because the interpolation effort is independent of the number of grid points. It also affects very marginally the storage requirements, because of the one-dimensional character of the tables. However, even though it needs to be done only once, the calculation of the radial integrals (24), (28), and (31) is not negligible if performed unwisely. We have developed a special fast radial Fourier transform for this purpose, as explained in appendix [3].

Dipole matrix elements, such as \( \langle \psi_1 \mid x \mid \psi_2 \rangle \), can also be obtained easily by defining a new function \( \chi_1(r) \equiv x\psi_1(r) \), expanding it using (21), and computing \( \langle \chi_1 \mid \psi_2 \rangle \) as explained above (with the precaution of using \( l_{\text{max}} + 1 \) instead of \( l_{\text{max}} \)).

VI. GRID INTEGRALS

The matrix elements of the last three terms of Eq. (16) involve potentials which are calculated on a real-space grid. The fineness of this grid is controlled by a 'grid cut-off' \( E_{\text{cut}} \); the maximum kinetic energy of the planewaves that can be represented in the grid without aliasing [3].

The short-range screened pseudopotentials \( V_{\text{NA}}(r) \) in (3) are tabulated as a function of the distance to atoms \( I \) and easily interpolated at any desired grid point. The last two terms require the calculation of the electron density on the grid. Let \( \psi_i(r) \) be the Hamiltonian eigenstates, expanded in the atomic basis set

\[
\psi_i(r) = \sum_\mu \phi_\mu(r) c_{i\mu}, \tag{32}
\]

where \( c_{i\mu} = \langle \psi_i \mid \phi_\mu \rangle \) and \( \phi_\mu^* \) is the dual orbital of \( \phi_\mu \): \( \langle \phi_\mu \mid \phi_\nu \rangle = \delta_{\mu\nu} \). We use the compact index notation \( \mu = \{ I \{ n,m \} \} \) for the basis orbitals, Eq. (5). The electron density is then

\[
\rho(r) = \sum_i n_i |\psi_i(r)|^2 \tag{33}
\]

where \( n_i \) is the occupation of state \( \psi_i \). If we substitute (32) into (33) and define a density matrix

\[
\rho_{\mu\nu} = \sum_i c_{i\mu} n_i c_{i\nu} \tag{34}
\]

where \( c_{i\mu} \equiv c_{vi}^* \), the electron density can be rewritten as

\[
\rho(r) = \sum_{\mu} \rho_{\mu\nu} \phi_\mu^*(r) \phi_\mu(r) \tag{35}
\]

We use the notation \( \phi_\mu^* \) for generality, despite our use of real basis orbitals in practice. Then, to calculate the density at a given grid point, we first find all the atomic basis orbitals, Eq. (6), at that point, interpolating the radial part from numerical tables, and then we use (35) to calculate the density. Notice that only a small number of basis orbitals are non-zero at a given grid point, so that the calculation of the density can be performed in \( O(N) \) operations, once \( \rho_{\mu\nu} \) is known. The storage of the orbital values at the grid points can be one of the most expensive parts of the program in terms of memory usage. Hence, an option is included to calculate and use these terms on the fly, in the the spirit of a direct-SCF calculation. The calculation of \( \rho_{\mu\nu} \) itself with Eq. (34) does not scale linearly with the system size, requiring instead the use of special \( O(N) \) techniques to be described below. However, notice that in order to calculate the density, only the matrix elements \( \rho_{\mu\nu} \) for which \( \phi_\mu \) and \( \phi_\nu \) overlap are required, and they can therefore be stored as a sparse matrix of \( O(N) \) size. Once the valence density is available in the grid, we add to it, if necessary, the non-local core correction [4], a spherical charge density intended to simulate the atomic cores, which is also interpolated from a radial grid. With it, we find the exchange and correlation potential \( V_{xc}(r) \), trivially in the LDA and using the method described in ref. [4] for the GGA. To calculate \( \delta V(H)(r) \), we first find \( \rho_{\mu\nu}(r) \) at the grid points, as a sum of spherical atomic densities (also interpolated from a radial grid) and subtract it from \( \rho(r) \) to find \( \delta \rho(r) \). We then solve Poisson’s equation to obtain \( \delta V(H)(r) \) and find the total grid potential \( V(r) = V_{\text{NA}}(r) + \delta V(H)(r) + V_{xc}(r) \). Finally, at every grid point, we calculate \( V(r)\phi_\mu^*(r) \phi_\nu(r) \Delta r^3 \) for all pairs \( \phi_\mu, \phi_\nu \) which are not zero at that point (\( \Delta r^3 \) is the volume per grid point) and add it to the Hamiltonian matrix element \( H_{\mu\nu} \).

To solve Poisson’s equation and find \( \delta V(H)(r) \) we normally use fast Fourier transforms in a unit cell that is either naturally periodic or made artificially periodic by a supercell construction. For neutral isolated molecules,
our use of strictly confined basis orbitals makes it trivial to avoid any direct overlap between the repeated molecules, and the electric multipole interactions decrease rapidly with cell size. For charged molecules we supress the $G = 0$ Fourier component (an infinite constant) of the potential created by the excess of charge. This amounts to compensating this excess with a uniform charge background. We then use the method of Makov and Payne to correct the total energy for the interaction between the repeated cells. Alternatively, we can solve Poisson’s equation by the multigrid method, using finite differences and fixed boundary conditions, obtained from the multipole expansion of the molecular charge density. This can be done in strictly $O(N)$ operations, unlike the FFT’s, which scale as $N \log N$. However, the cost of this operation is typically negligible and therefore has no influence on the overall scaling properties of the calculation.

Figures 3 and 4 show the convergence of different magnitudes with respect to the energy cutoff of the integration grid. For orthogonal unit cell vectors this is simply, in atomic units, $E_{\text{cut}} = (\pi/\Delta x)^2/2$ with $\Delta x$ the grid interval.

![Figure 5](image)

**FIG. 5:** (a) Convergence of the total energy and pressure in bulk silicon as a function of the energy cutoff $E_{\text{cut}}$ of the real space integration mesh. Circles and continuous line: using a grid-cell-sampling of eight refinement points per original grid point. The refinement points are used only in the final calculation, not during the self-consistency iteration (see text). Triangles: two refinement points per original grid point. White circles: no grid-cell-sampling. (b) Bond length and angle of the water molecule as a function of $E_{\text{cut}}$.

**FIG. 6:** Same as Fig. 3 for the total energy and pressure of bulk iron. This is presented as a particularly difficult case because of the very hard partial core correction ($r_m = 0.7$ au) required for a correct description of exchange and correlation.

repeated twice in an almost independent way: only to calculate $V_x^\alpha(r)$ need they be combined. However, in the non-collinear spin case [44, 45, 46, 47], the density at every point is not represented by the up and down values, but also by a vector giving the spin direction. Equivalently, it may be represented by a local spin density matrix

$$
\rho^{\alpha\beta}(r) = \sum_i n_i \psi_i^{\beta*}(r) \psi_i^{\alpha}(r) = \sum_{\mu\nu} \rho_{\mu\nu} \phi_{\mu}(r) \phi_{\nu}(r) \quad (36)
$$

$$
\psi_i^{\alpha}(r) = \sum_{\mu} \phi_{\mu}(r) c_{\mu i}^{\alpha} \quad (37)
$$

$$
\rho_{\mu\nu}^{\alpha\beta} = \sum_i c_{\mu i}^{\alpha} n_i c_{\nu i}^{\beta} \quad (38)
$$

where $\alpha, \beta$ are spin indices, with up or down values. The coefficients $c_{\mu i}^{\alpha}$ are obtained by solving the generalized eigenvalue problem

$$
\sum_{\nu,\beta} (H_{\mu\nu}^{\alpha\beta} - E_i S_{\mu\nu} \delta^{\alpha\beta}) c_{\nu i}^{\beta} = 0 \quad (39)
$$

where $H_{\mu\nu}^{\alpha\beta}$, like $\rho_{\mu\nu}^{\alpha\beta}$, is a $(2N \times 2N)$ matrix, with $N$ the number of basis functions:

$$
H_{\mu\nu}^{\alpha\beta} = \langle \phi_{\mu}| \hat{T} + \hat{V}^{KB} + V^{NA}(r) + \delta V^H(r) + V_{XC}^{\alpha\beta}(r)| \phi_{\nu} \rangle. \quad (40)
$$

This is in contrast to the collinear spin case, in which the Hamiltonian and density matrices can be factorized into two $N \times N$ matrices, one for each spin direction. To calculate $V_{XC}^{\alpha\beta}(r)$ we first diagonalize the $2 \times 2$ matrix $\rho^{\alpha\beta}(r)$ at every point, in order to find the up and down spin densities $\rho^\uparrow(r), \rho^\downarrow(r)$ in the direction of the local spin.
vector. We then find \( V_X(r), V_Y(r) \) in that direction, with the usual local spin density functional** and we rotate back \( V_X(r) \) to the original direction. Thus, the grid operations are still basically the same, except that they need now be repeated three times, for the \( \uparrow \uparrow, \downarrow \downarrow \) and \( \uparrow \downarrow \) components. Notice that \( \rho_{\alpha\beta}(r) \) and \( V_{\alpha\beta}(r) \) are locally Hermitian, while \( H_{\mu\nu} \) and \( \rho_{\mu\nu} \) are globally Hermitian \( (H_{\beta\alpha} = H^{\alpha\beta}) \), so that their \( \downarrow \uparrow \) components can be obtained from the \( \uparrow \downarrow \) ones.

VIII. BRILLOUIN ZONE SAMPLING

Integration of all magnitudes over the Brillouin zone (BZ) is essential for small and moderately large unit cells, especially of metals. Although Siesta is designed for large unit cells, in practice it is very useful, especially for comparisons and checks, to be able to also perform calculations efficiently on smaller systems without using expensive superlattices. On the other hand, an efficient \( k \)-sampling implementation should not penalize, because of the required complex arithmetic, the \( \Gamma \)-point calculations used in large cells. A solution used in some programs is to have two different versions of all or part of the code, but this poses extra maintenance requirements.

We have dealt with this problem in the following way: around the unit cell (and comprising itself) we define an auxiliary supercell large enough to contain all the atoms whose basis orbitals are non-zero at any of the grid points of the unit cell, or which overlap with any of the basis orbitals in it. We calculate all the non-zero two-center integrals between the unit cell basis orbitals and the supercell orbitals, without any complex phase factors. We also calculate the grid integrals between all the supercell basis orbitals \( \phi_{\mu'} \) and \( \phi_{\nu'} \) (primed indices run over all the supercell), but within the unit cell only. We accumulate these integrals in the corresponding matrix elements, thus making use of the relation

\[
< \phi_{\mu'}|V(r)|\phi_{\nu'}> = \sum_{(\mu'\nu')=(\mu\nu')} < \phi_{\mu'}|V(r)f(r)|\phi_{\nu'}> .
\]

(41)

\( f(r) = 1 \) for \( r \) within the unit cell and is zero otherwise. \( \phi_{\mu} \) is within the unit cell. The notation \( \mu' \equiv \mu \) indicates that \( \phi_{\mu'} \) and \( \phi_{\mu} \) are equivalent orbitals, related by a lattice vector translation. \( (\mu'\nu') \equiv (\mu\nu') \) means that the sum extends over all pairs of supercell orbitals \( \phi_{\mu'} \) and \( \phi_{\nu'} \) such that \( \mu' \equiv \mu, \nu' \equiv \nu \), and \( R_{\mu'} - R_{\nu'} = R_{\mu} - R_{\nu} \). Once all the real overlap and Hamiltonian matrix elements are calculated, we multiply them, at every \( k \)-point by the corresponding phase factors and accumulate them by folding the supercell orbital to its unit-cell counterpart. Thus

\[
H_{\mu\nu}(k) = \sum_{\mu'\nu'} H_{\mu'\nu'} e^{ik(R_{\nu'}-R_{\mu})}
\]

(42)

where \( \phi_{\mu} \) and \( \phi_{\nu} \) are within the unit cell. The resulting \( N \times N \) complex eigenvalue problem, with \( N \) the number of orbitals in the unit cell, is then solved at every sampled \( k \) point, finding the Bloch-state expansion coefficients \( c_{\mu i}(k) \):

\[
\psi_i(k, r) = \sum_{\mu'} c_{\mu'i}(k) e^{ikR_{\nu'}}(r)
\]

(43)

where the sum extends over all basis orbitals in space, \( i \) labels the different bands, \( c_{\mu'i} = c_{\mu i} \) if \( \mu' \equiv \mu \), and \( \psi_i(k, r) \) is normalized in the unit cell.

The electron density is then

\[
\rho(r) = \sum_i \int_{BZ} n_i(k)|\psi_i(k, r)|^2 dk
\]

(44)

where the sum is again over all basis orbitals in space, and the density matrix

\[
\rho_{\mu\nu} = \sum_i \int_{BZ} c_{\mu i}(k)n_i(k)c_{\nu i}(k)e^{ik(R_{\nu'}-R_{\mu})}dk
\]

(45)

is real (for real \( \phi_{\mu} \)'s) and periodic, i.e. \( \rho_{\mu\nu} = \rho_{\mu'\nu'} \) if \( (\mu, \nu) \equiv (\nu', \mu') \) (with \( \equiv \) meaning again ‘equivalent by translation’). Thus, to calculate the density at a grid point of the unit cell, we simply find the sum \( (44) \) over all the pairs of orbitals \( \phi_{\mu}, \phi_{\nu} \) in the supercell that are non-zero at that point.

In practice, the integral in \( (45) \) is performed in a finite, uniform grid of the Brillouin zone. The fineness of this grid is controlled by a \( k \)-grid cutoff \( l_{cut} \), a real-space radius which plays a role equivalent to the plane-wave cutoff of the real-space grid as**. The origin of the \( k \)-grid may be displaced from \( k = 0 \) in order to decrease the number of inequivalent \( k \)-points.

If the unit cell is large enough to allow a \( \Gamma \)-point-only calculation, the multiplication by phase factors is skipped and a single real-matrix eigenvalue problem is solved (in this case, the real matrix elements are accumulated directly in the first stage, if multiple overlaps occur). In this way, no complex arithmetic penalty occurs, and the differences between \( \Gamma \)-point and \( k \)-sampling are limited to a very small section of the code, while all the two-center and grid integrals use always the same real-arithmetic code.

IX. TOTAL ENERGY

The Kohn-Sham** total energy can be written as a sum of a band-structure (BS) energy plus some correction terms, sometimes called ‘double count’ corrections. The BS term is the sum of the energies of the occupied states \( \psi_i \):

\[
E^{BS} = \sum_i n_i(\psi_i|H|\psi_i) = \sum_{\mu\nu} H_{\mu\nu}\rho_{\mu\nu} = \text{Tr}(H \rho)
\]

(46)

where \( \rho_{\mu\nu} \) is again the density matrix obtained from the \( \phi_{\mu}, \phi_{\nu} \) orbitals.

[**] Local spin density functional method

[***] Plane-wave basis

[****] Real-matrix approach
where spin and \( k \)-sampling notations are omitted here for simplicity. At convergence, the \( \psi_i \)'s are simply the eigenvectors of the Hamiltonian, but it is important to realize that the Kohn-Sham functional is also perfectly well defined outside this so-called 'Born-Oppenheimer surface', i.e. it is defined for any set of orthonormal \( \psi_i \)'s.

The correction terms are simple functionals of the electron density, which can be obtained from equation (35), and the atomic positions. The Kohn-Sham total energy can then be written as

\[
E_{KS} = \sum_{\mu \nu} H_{\mu \nu} \rho_{\mu \nu} - \frac{1}{2} \int V^H(r) \rho(r) d^3r + \frac{1}{2} \int \left( \epsilon^{xc}(r) - V^{xc}(r) \right) \rho(r) d^3r + \sum_{i < j} Z_i Z_j \frac{\rho_{ij}(r)}{|R_{ij}|}.
\]

(47)

where \( I, J \) are atomic indices, \( R_{IJ} \equiv |R_J - R_I| \), \( Z_i, Z_J \) are the valence ion pseudoatom charges, and \( \epsilon^{xc}(r) \rho(r) \) is the exchange-correlation energy density. In order to avoid the long range interactions of the last term, we construct from the local-pseudopotential \( V_{iJ}^{local} \) which has an asymptotic behavior of \(-Z_I/r \) a diffuse ion charge, \( \rho_i^{local}(r) \), whose electrostatic potential is equal to \( V_{iJ}^{local}(r) \):

\[
\rho_i^{local}(r) = -\frac{1}{4\pi} \nabla^2 V_{iJ}^{local}(r).
\]

(48)

Notice that we define the electron density as positive, and therefore \( \rho_i^{local} \leq 0 \). Then, we write the last term in (47) as

\[
\sum_{i < J} \frac{Z_i Z_J}{R_{IJ}} = \frac{1}{2} \sum_{I,J} U_{IJ}^{local}(R_{IJ}) + \sum_{i < J} \delta U_{IJ}^{local}(R_{IJ}) - \sum_i U_i^{local}
\]

(49)

where \( U_{IJ}^{local} \) is the electrostatic interaction between the diffuse ion charges in atoms \( I \) and \( J \):

\[
U_{ij}^{local}(|R|) = \int V_i^{local}(r) \rho_j^{local}(r - R) d^3r,
\]

(50)

\( \delta U_{IJ}^{local} \) is a small short-range interaction term to correct for a possible overlap between the soft ion charges, which appears when the core densities are very extended:

\[
\delta U_{IJ}^{local}(R) = \frac{Z_i Z_J}{R} - U_{IJ}^{local}(R),
\]

(51)

and \( U_i^{local} \) is the fictitious self interaction of an ion charge (notice that the first right-hand sum in (49) includes the \( I = J \) terms):

\[
U_i^{local} = \frac{1}{2} U_i^{local}(0) = \frac{1}{2} \int V_i^{local}(r) \rho_i^{local}(r) 4\pi r^2 dr.
\]

(52)

Defining \( \rho_i^{NA} \) from \( V_i^{NA} \), analogously to \( \rho_i^{local} \), we have that \( \rho_i^{NA} = \rho_i^{local} + \rho_i^{atom} \), and equation (47) can be transformed, after some rearrangement of terms, into

\[
E_{KS} = \sum_{\mu \nu} (T_{\mu \nu} + V_{\mu \nu}^{KB}) \rho_{\mu \nu} + \frac{1}{2} \sum_{i,j} U_{ij}^{NA}(R_{ij}) + \sum_{i < J} \delta U_{ij}^{local}(R_{ij}) + \sum_i U_i^{local}
\]

\[
+ \int V^{NA}(r) \delta \rho(r) d^3r + \frac{1}{2} \int \left( \delta V^H(r) \delta \rho(r) d^3r + \int \epsilon^{xc}(r) \rho(r) d^3r
\]

(53)

where \( V^{NA} = \sum_i V_i^{NA} \) and \( \delta \rho = \rho - \sum_i \rho_i^{NA} \).

\[
U_{ij}^{NA}(R) = \int V_{ij}^{NA}(r) \rho_j^{NA}(r - R) d^3r
\]

(54)

is a radial pairwise potential that can be obtained from \( V_{ij}^{NA}(r) \) as a two-center integral, by the same method described previously for the kinetic matrix elements:

\[
T_{\mu \nu} = \langle \phi_\mu | -\frac{1}{2} \nabla^2 | \phi_\nu \rangle
\]

(55)

\[
V_{\mu \nu}^{KB} \text{ is also obtained by two-center integrals:}
\]

\[
V_{\mu \nu}^{KB} = \sum_\alpha \langle \phi_\mu | \chi_\alpha \rangle V_{\alpha \beta}^{KB} \langle \chi_\alpha | \phi_\nu \rangle
\]

(56)

where the sum is over all the KB projectors \( \chi_\alpha \) that overlap simultaneously with \( \phi_\mu \) and \( \phi_\nu \).

Although (53) is the total energy equation actually used by SIESTA, its meaning may be further clarified if the \( I = J \) terms of \( \frac{1}{2} \sum_{i,j} U_{ij}^{NA}(R_{ij}) \) are combined with \( \sum_i U_i^{local} \) to yield

\[
E_{KS} = \sum_{\mu \nu} (T_{\mu \nu} + V_{\mu \nu}^{KB}) \rho_{\mu \nu} + \sum_{i,j} U_{ij}^{NA}(R_{ij}) + \sum_{i < j} \delta U_{ij}^{local}(R_{ij}) + \sum_i U_i^{local}
\]

\[
+ \int V^{NA}(r) \delta \rho(r) d^3r
\]

(57)

where

\[
U_i^{atom} = \int_0^\infty \left( V_i^{local}(r) + \frac{1}{2} U_i^{atom}(r) \right) \rho_i^{atom}(r) 4\pi r^2 dr
\]

(58)

is the electrostatic energy of an isolated atom.
The last three terms in Eq. (53) are calculated using the real space grid. In addition to getting rid of all long-range potentials (except that implicit in $\delta V(H(r))$), the advantage of (53) is that, apart from the relatively slowly-varying exchange-correlation energy density, the grid integrals involve $\rho(r)$, which is generally much smaller than $\rho(\mathbf{r})$. Thus, the errors associated with the finite grid spacing are drastically reduced. Critically, the kinetic energy matrix elements can be calculated almost exactly, without any grid integrations.

It is frequently desirable to introduce a finite electronic temperature $T$ and/or a fixed chemical potential $\mu$, either because of true physical conditions or to accelerate the self-consistency iteration. Then the functional that must be minimized is the free energy

$$F(R, \psi, n) = E^{KS}(R, \psi, n) - \mu \sum_i n_i - k_B T \sum_i (n_i \log n_i + (1 - n_i) \log(1 - n_i)). \quad (59)$$

Minimization with respect to $n_i$ yields the usual Fermi-Dirac distribution $n_i = 1/(1 + e^{(\varepsilon_i - \mu)/k_B T})$.

X. HARRIS FUNCTIONAL

We will mention here a special use of the Harris energy functional, that is generally defined as

$$E_{Harris}[\rho^{in}] = \sum_i n_i^{-1} \langle \psi_i^{out} | \hat{H}^{in} | \psi_i^{out} \rangle - \frac{1}{2} \int \int \frac{\rho^{in}(\mathbf{r}) \rho^{in}(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d^3\mathbf{r} d^3\mathbf{r}' + \int (\varepsilon_{xc}(\mathbf{r}) - \varepsilon_{xc}^{in}(\mathbf{r})) \rho^{in}(\mathbf{r}) d^3\mathbf{r} + \sum_{I,J} Z_I Z_J \frac{Z_I Z_J}{R_{IJ}} \quad (60)$$

where $\hat{H}^{in}$ is the KS Hamiltonian produced by a trial density $\rho^{in}$ and $\psi_i^{out}$ are its eigenvectors (which in general are different from those whose density is $\rho^{in}$). As in Eq. (49), the first term in (60) can be written as $Tr(H^{in} \rho^{out})$, and the rest are the so-called ‘double count corrections’. An important advantage of eq. (60) is that it does not require $\rho^{in}$ to be obtained from a set of orthogonal electron states $\psi_i^{in}$, and in fact $\rho^{in}$ is frequently taken as a simple superposition of atomic densities. However, we will assume here that the states $\psi_i^{in}$ are indeed known. In this case, the Kohn-Sham energy $E^{KS}[\rho^{in}]$, Eq. (47), obeys exactly the same expression (60), except that $\psi_i^{out}$ and $n_i^{out}$ must be replaced by $\psi_i^{in}$ and $n_i^{in}$. Thus, a simple subtraction gives

$$E_{Harris}[\rho^{in}] = E^{KS}[\rho^{in}] + \sum_{\mu} H_{\mu}^{in} \left( \rho^{out}_{\mu} - \rho^{in}_{\mu} \right). \quad (61)$$

Generally the Harris functional is used nonself-consistently, with a trial density given by the sum of atomic densities. But here we want to comment on its usefulness to improve dramatically the estimate of the converged total energy, by taking $\rho^{in}_{\mu}$ as the density matrix of the $(n-1)$th self-consistency iteration and $\rho^{out}_{\mu}$ of the $n$th iteration. In fact, $E^{Harris}$ frequently gives, after just two or three iterations, a better estimate than $E^{KS}$ after tens of iterations. Unfortunately, we have found that there is hardly any improvement in the convergence of the atomic forces thus estimated, and therefore the self-consistent Harris functional is less useful for geometry relaxations or molecular dynamics.

XI. ATOMIC FORCES

Atomic forces and stresses are obtained by direct differentiation of (53) with respect to atomic positions. They are obtained simultaneously with the total energy, mostly in the same places of the code, under the general paradigm “a piece of energy ⇒ a piece of force/stress” (except that some pieces are calculated only in the last self-consistency step). This ensures that all force contributions, including Pulay corrections, are automatically included. The force contribution from the first term in (53) is

$$\frac{\partial}{\partial \mathbf{R}_I} \left( \sum_{\mu} (T_{\mu \mu} + V_{KB}^{\mu \mu}) \rho_{\mu} \right) =$$

$$\sum_{\mu} (T_{\mu \mu} + V_{KB}^{\mu \mu}) \frac{\partial \rho_{\mu}}{\partial \mathbf{R}_I} + 2 \sum_{\mu \nu} \sum_{I \in \ell} \frac{dT_{\mu \nu}}{\partial \mathbf{R}_I} \rho_{\mu \nu} +$$

$$2 \sum_{\mu \nu} \sum_{\alpha I} S_{\mu \alpha} v_{KB \alpha}^{\mu} \frac{dS_{\alpha \nu}}{\partial \mathbf{R}_I} \rho_{\mu \nu} -$$

$$2 \sum_{\mu \nu} \sum_{\alpha I} S_{\mu \alpha} v_{KB \alpha}^{\mu} \frac{dS_{\alpha \nu}}{\partial \mathbf{R}_I} \rho_{\mu \nu} \quad (62)$$

where $\alpha$ are KB projector indices, $I$ indicates orbitals or KB projectors belonging to atom $I$, and we have considered that

$$\frac{\partial S_{\mu \nu}}{\partial \mathbf{R}_{I_\mu}} = - \frac{\partial S_{\mu \nu}}{\partial \mathbf{R}_{I_\mu}} = \frac{dS_{\mu \nu}}{d \mathbf{R}_{I_{\mu}}} \quad (63)$$

where $\mathbf{R}_{I_{\mu}}$ is the position of atom $I_{\mu}$, to which orbital $\phi_{\mu}$ belongs and $\mathbf{R}_{\mu} = \mathbf{R}_{I_{\mu}} - \mathbf{R}_{I_{\mu}}$.

Leaving aside for appendix A the terms containing $\partial \rho_{\mu \nu}/\partial \mathbf{R}_I$, the other derivatives can be obtained by straightforward differentiation of their expansion in spherical harmonics (Eq. (24)). However, instead of using the spherical harmonics $Y_{\ell \mu}(\hat{r})$ themselves, it is convenient to multiply them by $r^\ell$, in order to make them
analytic at the origin. Thus
\[
\frac{dS_{\mu\nu}(R)}{dR} = \sum_{lm} \nabla \left( \frac{S_{\mu\nu}^{lm}(R)}{R^l} R^l Y_{lm}(\hat{R}) \right)
\]
\[
= \sum_{lm} \frac{d}{dR} \left( \frac{S_{\mu\nu}^{lm}(R)}{R^l} \right) R^l Y_{lm}(\hat{R}) \hat{R}
\]
\[
+ \sum_{lm} \frac{S_{\mu\nu}^{lm}(R)}{R^l} \nabla (R^l Y_{lm}(\hat{R}))
\]

(64)

In fact, it is \( S_{\mu\nu}^{lm}(R)/R^l \), rather than \( S_{\mu\nu}^{lm}(R) \), that is stored as a function of \( R \) on a radial grid. Its derivative, \( d(S_{\mu\nu}^{lm}(R)/R^l)/dR \), is then obtained from the same cubic spline interpolation used for the value itself. The value and gradient of \( R^l Y_{lm}(\hat{R}) \) are calculated analytically from explicit formulae (up to \( l = 2 \)) or recurrence relations. Entirely analogous equations apply to \( dT_{\mu\nu}/dR_{\mu\nu} \).

The second and third terms in Eq. (53) are simple interatomic pair potentials whose force contributions are calculated trivially from their radial spline interpolations. The fourth term is a constant which does not depend on the atomic positions. Taking into account that \( V_{\mu\nu}^{NA}(r) = \sum_{i \in I} V_{\mu\nu}^{NA}(r - \mathbf{R}_i) \), and therefore \( \partial V_{\mu\nu}^{NA}(r)/\partial \mathbf{R}_i = -\nabla V_{\mu\nu}^{NA}(r - \mathbf{R}_i) \), the force contribution from the fifth term is
\[
\frac{\partial}{\partial \mathbf{R}_i} \int V_{\mu\nu}^{NA}(r) \delta \rho(r) d^3 r = \int \nabla V_{\mu\nu}^{NA}(r) \delta \rho(r) d^3 r + \int V_{\mu\nu}^{NA}(r) \frac{\partial \delta \rho(r)}{\partial \mathbf{R}_i} d^3 r
\]

(65)

The sixth term is the electrostatic self-energy of the charge distribution \( \delta \rho(r) \):
\[
\frac{\partial}{\partial \mathbf{R}_i} \frac{1}{2} \int \delta V^H(r) \delta \rho(r) d^3 r = \int \delta V^H(r) \frac{\partial \delta \rho(r)}{\partial \mathbf{R}_i} d^3 r
\]

(66)

In the last term, we take into account that \( d(\rho e^{\rho(r)})/d\rho = e^{\rho(r)} \) to obtain
\[
\frac{\partial}{\partial \mathbf{R}_i} \int e^{\rho(r)} \delta \rho(r) d^3 r = \int V_{\mu\nu}^{c}(r) \frac{\partial \rho(r)}{\partial \mathbf{R}_i} d^3 r
\]

(67)

Now, using Eq. (53) and that, for \( \nu \in I \), \( \partial \phi_{\nu}(r)/\partial \mathbf{R}_i = -\nabla \phi_{\nu} \), the change of the self-consistent and atomic densities are
\[
\frac{\partial \rho(r)}{\partial \mathbf{R}_i} = \text{Re} \sum_{\mu} \frac{\partial \rho_{\mu\nu}}{\partial \mathbf{R}_i} \phi_{\mu}^*(r) \phi_{\nu}(r)
\]
\[
- 2 \text{Re} \sum_{\mu} \sum_{\nu \in I} \rho_{\mu\nu} \phi_{\mu}^*(r) \nabla \phi_{\nu}(r)
\]

(68)

\[
\frac{\partial \rho_{\text{atom}}(r)}{\partial \mathbf{R}_i} = -2 \text{Re} \sum_{\mu} \rho_{\mu\nu} \phi_{\mu}^*(r) \nabla \phi_{\nu}(r)
\]

(69)

where we have taken into account that the density matrix of the separated atoms is diagonal. Thus, leaving still aside the terms with \( \partial \rho_{\mu\nu}/\partial \mathbf{R}_i \), the last term in Eq. (63), as well as those in (60) and (17), have the general form
\[
\text{Re} \sum_{\mu} \sum_{\nu \in I} \rho_{\mu\nu} \int V(r) \phi_{\mu}^*(r) \nabla \phi_{\nu}(r) d^3 r
\]
\[
= \text{Re} \sum_{\mu} \sum_{\nu \in I} \rho_{\mu\nu} \langle \phi_{\mu} | V(r) | \nabla \phi_{\nu} \rangle.
\]

(70)

These integrals are calculated on the grid, in the same way as those for the total energy (i.e. \( \langle \phi_{\mu} | V(r) | \phi_{\nu} \rangle \)). The gradients \( \nabla \phi_{\nu}(r) \) at the grid points are obtained analytically, like those of \( \phi_{\nu}(r) \) from their radial grid interpolations of \( \phi(r)/r^l \):
\[
\nabla \phi_{I\mu}(r) = \frac{d}{dr} \left( \frac{\phi_{I\mu}(r)}{r^l} \right) r^l Y_{lm}(\hat{r}) \hat{r}
\]
\[
+ \frac{\phi_{I\mu}(r)}{r^l} \nabla (r^l Y_{lm}(\hat{r})).
\]

(71)

In some special cases, with elements that require hard partial core corrections or explicit inclusion of the semi-core, the grid integrals may pose a problem for geometry relaxations, because they make the energy dependent on the position of the atoms relative to the grid. This 'egg-box effect' is small for the energy itself, and it decreases fast with the grid spacing. But the effect is larger and the convergence slower for the forces, as they are proportional to the amplitude of the energy oscillation, but inversely proportional to its period. These force oscillations complicate the force landscape, especially when the true atomic forces become small, making the convergence of the geometry optimization more difficult. Of course, the problem can be avoided by decreasing the grid spacing but this has an additional cost in computer time and memory. Therefore, we have found it useful to minimize this problem by recalculating the forces, at a set of positions, determined by translating the whole system by a set of points in a finer mesh. This procedure, which we call 'grid-cell sampling', has no extra cost in memory. And since it is done only at the end of the self-consistency iteration, for fixed \( \rho_{\mu\nu} \), it has only a moderate cost in CPU time.

At finite temperature, the forces are really the derivatives of the free energy with respect to atomic displacements since
\[
\frac{dF(R_i, \psi_i(r), n_i)}{dR_i} = \frac{\partial F}{\partial R_i} + \sum_i \frac{\partial F}{\partial n_i} \frac{\partial n_i}{\partial R_i} + \int \frac{\partial F}{\partial \psi_i^*(r)} \frac{\partial \psi_i(r)}{\partial R_i} d^3 r
\]

(72)

In this particular equation we have used the notation \( dR_i \), as opposed to \( \partial \mathbf{R}_i \), to indicate the inclusion of the change in \( \psi_i(r) \) and \( n_i \) when we move the atom, in calculating the derivative. But we have used also that \( \partial F/\partial n_i = \partial F/\partial \psi_i(r) = 0 \) and that the last two terms in (72) do not depend on \( R_i \), so that \( \partial F/\partial R_i = \partial E/\partial R_i \).
The latter are the atomic forces actually calculated. Notice, however, that \( \frac{dE}{dR_I} \neq \frac{dE}{dR} \), so that the calculated forces are indeed the total derivatives of the free, not the internal energy.

We would like to also mention the calculation of forces using the non-self-consistent Harris functional, in which the ‘in’ density is a superposition of atomic densities. We have implemented this as an option for ‘quick and dirty’ calculations because, used with a minimal basis set, it makes SIESTA competitive with tight binding methods, which are much faster than density functional calculations. The problem that we address here is that, although \( E^{Harriss} \) is stationary with respect to \( \rho \), it is not so with respect to \( \rho^{in} \). In particular, there appears a force term

\[
\int \frac{\partial V^{in}(r)}{\partial R_I} \rho^{out}(r) d^3r.
\]  

(73)

A similar term appears for the electrostatic interaction between the input and output density, but it presents no special problems because of the linear character of the Hartree potential. However, evaluation of (73) requires the change of the exchange-correlation potential with density, a quantity also required to evaluate the linear response of the electron gas, but not in normal energy and force calculations. Finally, notice that, apart from this minor difficulty, the Harris-functional forces are perfectly well defined at the first iteration only. For later iterations (but still not converged) there is no practical way to calculate \( \partial \rho^{in}/\partial R_I \) and, without the help of Hellman-Feynman’s theorem (which applies only at convergence), the forces are not well defined. Of course, the omission of the terms depending on this quantity produces an estimate of the forces, but we have found that their convergence is not appreciably faster than those estimated from the Kohn-Sham functional.

**XI. STRESS TENSOR**

We define the stress tensor as the positive derivative of the total energy with respect to the strain tensor

\[
\sigma_{\alpha\beta} = \frac{\partial E^{KS}}{\partial \varepsilon_{\alpha\beta}}
\]

(74)

where \( \alpha, \beta \) are Cartesian coordinate indices. To translate to standard units of pressure, we must simply divide by the unit-cell volume and change sign. During the deformation, all vector positions, including those of atoms and grid points (and of course lattice vectors), change according to

\[
r'_{\alpha} = \sum_{\beta=1}^{3} (\delta_{\alpha\beta} + \varepsilon_{\alpha\beta}) r_{\beta}
\]

(75)

The shape of the basis functions, KB projectors, and atomic densities and potentials do not change, but their origin gets displaced according to (75). From this equation, we find that

\[
\frac{\partial r_{\gamma}}{\partial \varepsilon_{\alpha\beta}} = \delta_{\gamma\alpha} r_{\beta}
\]

(76)

The change in \( E^{KS} \) is essentially due to these position displacements, and therefore the calculation of the stress is almost perfectly parallel to that of the atomic forces, thus being performed in the same sections of the code. For example:

\[
\frac{\partial T_{\mu\nu}}{\partial \varepsilon_{\alpha\beta}} = \sum_{\gamma=1}^{3} \frac{\partial T_{\mu\nu}}{\partial r_{\alpha\gamma}} \frac{\partial r_{\gamma}}{\partial \varepsilon_{\alpha\beta}} - \frac{\partial T_{\mu\nu}}{\partial r_{\alpha\beta}} r_{\beta}
\]

(77)

Since \( \frac{\partial T_{\mu\nu}}{\partial r_{\alpha\beta}} \) is evaluated to calculate the forces, it takes very little extra effort to multiply it also by \( r_{\beta} \) for the stress. Equally, force contributions like (70) have their obvious stress counterpart

\[
\sum_{\mu\nu} \rho_{\mu\nu} (\phi_{\mu}(V(r))(\nabla_{\alpha} \phi_{\nu}) r_{\beta})
\]

(78)

However, there are three exceptions to this parallelism. The first concerns the change of the volume per grid point or, in other words, the Jacobian of the transformation (72) in the integrals over the unit cell. This Jacobian is simply \( \delta_{\alpha\beta} \), and it leads to a stress contribution

\[
\left[ \int \left( V^{NA}(r) + \frac{1}{2} \delta V^{H}(r) \right) \delta \rho(r) d^3r + E^{xc} \right] \delta_{\alpha\beta}
\]

(79)

Notice that the renormalization of the density, required to conserve the change when the volume changes, enters through the orthonormality constraints, to be discussed in appendix A. The second special contribution to the stress lies in the fact that, as we deform the lattice, there is a change in the factor \( 1/|r - r'| \) of the electrostatic energy integrals. We deal with this contribution in reciprocal space, when we calculate the Hartree potential by FFTs, by evaluating the derivative of the reciprocal-space vectors with respect to \( \epsilon_{\alpha\beta} \). Since \( G' = \sum_{\beta} G(\delta_{\beta\alpha} - \epsilon_{\beta\alpha}) \):

\[
\frac{\partial}{\partial \varepsilon_{\alpha\beta}} \frac{1}{G^2} = 2G_{\alpha}\delta_{\beta\alpha}
\]

(80)

Finally, the third special stress contribution arises in GGA exchange and correlation, from the change of the gradient of the deformed density \( \rho(r) \rightarrow \rho(r') \). The treatment of this contribution is explained in detail in reference 12.

**Electric Polarization**

The calculation of the electric polarization, as an integral in the grid across the unit cell, is standard and
almost free for molecules, chains and slabs (in the directions perpendicular to the chain axis, or to the surface). For bulk systems, the electric polarization cannot be found from the charge distribution in the unit cell alone. In this case, we need the so-called Berry-phase theory of polarization, which allows to compute quantities like the dynamical charges and piezoelectric constants.

Here we comment some details of our implementation. If $R_\alpha$ are the lattice vectors and $P^c = \sum_{\alpha=1}^3 P_{\alpha}^c R_\alpha$ is the electronic contribution to the macroscopic polarization, then we have

$$2\pi P^c_\alpha = G_\alpha \cdot P^c = -\frac{2e}{(2\pi)^3} \int d\mathbf{k} \ G_\alpha \cdot \frac{\partial}{\partial \mathbf{k}'} \Phi(\mathbf{k}, \mathbf{k}') \bigg|_{\mathbf{k}'=\mathbf{k}} \tag{81}$$

where $G_\alpha$ is the corresponding reciprocal lattice vector, $e$ is the electron charge, $u_i(\mathbf{k}, \mathbf{r}) = e^{-i\mathbf{k} \cdot \mathbf{r}} \psi_i(\mathbf{k}, \mathbf{r})$ is the periodic part of the Bloch function, and the factor of two comes from the spin degeneracy. The quantum phase $\Phi(\mathbf{k}, \mathbf{k}')$ is defined as

$$\Phi(\mathbf{k}, \mathbf{k}') = \text{Im} \left[ \ln (\det \langle u_i(\mathbf{k}, \mathbf{r}) | u_j(\mathbf{k}', \mathbf{r}) \rangle) \right] \tag{82}$$

The derivative in (82) depends on a gauge that must be chosen such that $u(\mathbf{k} + G, \mathbf{r}) = e^{-iG \cdot \mathbf{r}} u(\mathbf{k}, \mathbf{r})$. In practice, the integral is replaced by a discrete summation, and a finite-difference approximation is made for the derivative

$$G_1 \cdot P^c \approx -\frac{2e}{\Omega N_2^2 N_3^3} \sum_{i_2=0,i_3=0}^{N_2-1,i_1=0} \sum_{i_1=0}^{N_1-1} \Phi(\mathbf{k}, \mathbf{k} + \Delta \mathbf{k}_i) \tag{83}$$

where we have split the sum to stress the fact that we have a two-dimensional integral in the plane defined by $G_2$ and $G_3$, and a linear integral along $G_1$. Due to the approximation in the derivative, the linear integral usually requires a finer mesh than the surface integral. To evaluate $\Phi(\mathbf{k}, \mathbf{k} + \Delta \mathbf{k})$ we use our LCAO basis:

$$\langle u_i(\mathbf{k}) | u_j(\mathbf{k} + \Delta \mathbf{k}) \rangle = \langle \psi_i(\mathbf{k}) | e^{-i\Delta \mathbf{k} \cdot \mathbf{r}} | \psi_j(\mathbf{k} + \Delta \mathbf{k}) \rangle = \sum_{\mu'} \sum_{\mu} c_{i\mu}(\mathbf{k}) c_{j\mu'}^*(\mathbf{k} + \Delta \mathbf{k}) e^{-i\mathbf{k} \cdot (\mathbf{R}_\mu - \mathbf{R}_{\mu'})} \langle \phi_{\mu'} | e^{-i\Delta \mathbf{k} \cdot (\mathbf{R} - \mathbf{R}_{\mu'})} | \phi_{\mu} \rangle \tag{84}$$

Formulas similar to (84) have been implemented by several authors, mainly in the context of Hartree-Fock calculations, in which the basis orbitals are expanded in gaussians whose matrix elements can be found analytically. Our numerical, localized pseudo-atomic basis orbitals are not well suited for a gaussian expansion. Instead, we expand the plane-waves appearing in equation (84) to first order in $\Delta \mathbf{k}$, $e^{-i\Delta \mathbf{k} \cdot (\mathbf{R} - \mathbf{R}_{\mu})} \approx 1 - i\Delta \mathbf{k} \cdot (\mathbf{R} - \mathbf{R}_{\mu}) + \mathcal{O}(\Delta \mathbf{k}^2)$, and then we calculate the matrix elements of the position operator as explained in section \[\text{V}\]. It is interesting to note that, since the discretized formula (83) only holds to $\mathcal{O}(\Delta \mathbf{k}^2)$, the approximation of the matrix elements in (83) does not introduce any further errors in the calculation of the polarization. In a symmetrized version, we approximate equation (84) as

$$\sum_{\nu} \sum_{\mu'} c_{i\mu}(\mathbf{k}) c_{\mu' j}(\mathbf{k} + \Delta \mathbf{k}) e^{-i(\mathbf{k} + \Delta \mathbf{k} \cdot \mathbf{R}_{\nu})} [ \langle \phi_{\nu} | \phi_{\mu'} \rangle - i\Delta \mathbf{k} \cdot (\mathbf{R} - \mathbf{R}_{\mu'}) \langle \phi_{\nu} | (\mathbf{r} - \mathbf{R}_{\mu'}) | \phi_{\mu'} \rangle ] \tag{85}$$

XIII. ORDER-N FUNCTIONAL

The basic problem for solving the Kohn-Sham equations in $\mathcal{O}(N)$ operations is that the solutions (the Hamiltonian eigenvectors) are extended over the whole system and overlap with each other. Just to check the orthogonality of $N$ trial solutions, by performing integrals over the whole system, involves $\sim N^3$ operations. Among the different methods proposed to solve this problem, we have chosen the localized-orbital approach because of its superior efficiency for non-orthogonal basis sets. The initially proposed functional used a fixed number of occupied states, equal to the number of electron pairs, and it was found to have numerous local minima in which the electron configuration was easily trapped. A revised functional which uses a larger number of states than electron pairs, with variable occupations, has been found empirically to avoid the local minima problem. This is the functional that we use and recommend.

Each of the localized, Wannier-like states, is constrained to its own localization region. Each atom $I$ is assigned a number of states equal to $\text{int}(Z_f^{\text{val}}/2 + 1)$ so that, if doubly occupied, they can contain at least one excess electron (they can also become empty during the minimization of the energy functional). These states are confined to a sphere of radius $R_c$ (common to all states) centered at $\mathbf{R}_I$. More precisely, the expansion (Eq. (82)) of a state $\psi_i$ centered at $\mathbf{R}_I$ may contain only basis orbitals $\phi_{\mu}$ centered on atoms $J$ such that $|\mathbf{R}_{IJ}| < R_c$. This implies that $\psi_i(\mathbf{r})$ may extend to a maximum range $R_{e} + r_{\text{max}}^{\text{max}}$, where $r_{\text{max}}^{\text{max}}$ is the maximum range of the basis orbitals. For covalent systems, a localization region centered on bonds rather than atoms is more efficient (it leads to a lower energy for the same $R_c$), but it is less suitable to a general algorithm, especially in case of ambiguous bonds. Therefore, we generally use the atom-centered localization regions.

In the method of Kim, Mauri, and Galli (KMG), the
The minimization proceeds without need to orthonormalize the electron states $\psi_i$. Instead, the orthogonality, as well as the correct normalization (one below $\eta$ and zero above it) result as a consequence of the minimization of $E^{KMG}$. This is because, in contrast to the KS functional, $E^{KMG}$ is designed to penalize any nonorthogonality. The KS ground state, with all the occupied $\psi_i$'s orthonormal, is also the minimum of $E^{KS}$, at which $E^{KMG} = E^{KS}$. If the variational freedom is constrained by the localization of the $\psi_i$'s, the orthogonality cannot be exact, and the resulting energy is slightly larger than for unconstrained wavefunctions. In insulators and semiconductors, the Wannier functions are exponentially localized, and the energy excess due to their strict localization decreases rapidly as a function of the localization radius $R_c$, as can be seen in Fig. 7.

![FIG. 7: Convergence of the lattice constant, bulk modulus, and cohesive energy as a function of the localization radius $R_c$ of the Wannier-like electron states in silicon. We used a supercell of 512 atoms and a minimal basis set with a cutoff radius $r_c = 5$ a.u. for both $s$ and $p$ orbitals.](image)

If the system is metallic, or if the chemical potential is not within the band gap (for example because of the presence of defects), the KMG functional cannot be used in practice. In fact, although some $O(N)$ methods can handle metallic systems in principle, we are not aware of any practical calculations at a DFT level. In such cases we copy the Hamiltonian and overlap matrices to standard expanded arrays and solve the generalized eigenvalue problem by conventional order-$N^3$ diagonalization techniques. However, even in this case, most of the operations, and particularly those to find the density and potential, and to set up the Hamiltonian, are still performed in $O(N)$ operations.

Irrespective of whether the $O(N)$ functional or the standard diagonalization is used, an outer self-consistency iteration is required, in which the density matrix is updated using Pulay’s Residual Metric Minimization by Direct Inversion of the Iterative Subspace (RMM-DIIS) method. Even when the code is strictly $O(N)$, the CPU time may increase faster if the number of iterations required to achieve the solution increases with $N$. In fact, it is a common experience that the required number of selfconsistency iterations increases with the size of the system. This is mainly because of the ‘charge sloshing’ effect, in which small displacements of charge from one side of the system to another give rise to larger changes of the potential, as the size increases. Fortunately, the localized character of the Wannier-like wavefunctions used in the $O(N)$ method help to solve also this problem, by limiting the charge sloshing. Ta-
Table II: Average number of selfconsistency (SCF) iterations (per molecular dynamics step) and average number of conjugate-gradient (CG) iterations (per SCF iteration) required to minimize the $O(N)$ functional, during a simulation of bulk silicon at $\sim 300$ K. We used the Verlet method at constant energy, with a time step of 1.5 fs, and a minimal basis set with a cutoff radius $r_c = 5$ a.u. $R_c$ is the localization radius of the Wannier-like wavefunctions used in the $O(N)$ functional (see text). $N$ is the number of atoms in the system.

<table>
<thead>
<tr>
<th>$N$</th>
<th>$R_c = 4\AA$</th>
<th>$R_c = 5\AA$</th>
</tr>
</thead>
<tbody>
<tr>
<td>64</td>
<td>5.8</td>
<td>9.3</td>
</tr>
<tr>
<td>512</td>
<td>4.9</td>
<td>11.4</td>
</tr>
<tr>
<td>1000</td>
<td>4.3</td>
<td>11.5</td>
</tr>
</tbody>
</table>

Figure 8 shows the essentially perfect $O(N)$ behaviour of the overall CPU time and memory. This is not surprising in view of the completely strict enforcement of $O(N)$ algorithms everywhere in the code (except the marginal $N \log N$ factor in the FFT used to solve Poisson’s equation, which represents a very small fraction of CPU time even for 4000 atoms).

XIV. OTHER FEATURES

Here we will simply mention some of the possibilities and features of the SIESTA implementation of DFT:

- A general-purpose package, the flexible data format (fdf), initially developed for the SIESTA project, allows the introduction of all the data and precision parameters in a simple tag-oriented, order-independent format which accepts different physical units. The data can then be accessed from anywhere in the program, using simple subroutine calls in which a default value is specified for the case in which the data are not present. A simple call also allows the read pointer to be positioned in order to read complex data ‘blocks’ also marked with tags.

- The systematic calculation of atomic forces and stress tensor allows the simultaneous relaxation of atomic coordinates and cell shape and size, using a conjugate gradients minimization or several other minimization/annealing algorithms.

- It is possible to perform a variety of molecular dynamics simulations, at constant energy or temperature, and at constant volume or pressure, also including Parrinello-Rahman dynamics with variable cell shape. The geometry relaxation may be restricted, to impose certain positions or coordinates, or more complex constraints.

- The auxiliary program VIBRA processes systematically the atomic forces for sets of displaced atomic positions, and from them computes the Hessian matrix and the phonon spectrum. An interface to the PHONON program is also provided within SIESTA.

- A linear response program (LINRES) to calculate phonon frequencies has also been developed. The code reads the SCF solution obtained by Siesta, and calculates the linear response to the atomic displacements, using first order perturbation theory. It then calculates the dynamical matrix, from which the phonon frequencies are obtained.

- A number of auxiliary programs allows various representations of the total density, the total and local density of states, and the electrostatic or total potentials. The representations include both two-dimensional cuts and three-dimensional views, which may be colored to simultaneously represent the density and potential.
• Thanks to an interface with the Transiesta program, it is possible to calculate transport properties across a nanocontact, finding self-consistently the effective potential across a finite voltage drop, at a DFT level, using the Keldish Green’s function formalism.

• The optical response can be studied with SIESTA using different approaches. An approximate dielectric function can be calculated from the dipolar transition matrix elements between occupied and unoccupied single-electron eigenstates using first order time-dependent perturbation theory. For finite systems, these are easily calculated from the matrix elements of the position operator between the basis orbitals. For infinite periodic systems, we use the matrix elements of the momentum operator. It is important to notice, however, that the use of non-local pseudopotentials requires some correction terms.

We have also implemented a more sophisticated approach to compute the optical response of finite systems, using the adiabatic approximation to time-dependent DFT. The idea is to integrate the time-dependent Schrödinger equation when a time dependent perturbation is applied to the system. From the time evolution, it is then possible to extract the optical adsorption and dipole strength functions, including some genuinely many-body effects, like plasmons. Using this approach we have successfully calculated the electronic response of systems such as fullerenes and small metallic clusters.

ACKNOWLEDGMENTS

We are deeply indebted to Otto Sankey and David Drabold for allowing us to use their code as an initial seed for this project, and to Richard Martin for continuous ideas and support. We thank Jose Luis Martins for numerous discussions and ideas, and Jürgen Kübler for helping us implement the noncollinear spin. The exchange-correlation methods and routines were developed in collaboration with Carlos Balbas and Jose L. Martins. We also thank In-Ho Lee, Maider Machado, Juana Moreno, and Art R. Williams for some routines, and Eduardo Anglada and Oscar Paz for their computational help. This work was supported by the Fundación Ramón Areces and by Spain’s MCYT grant BFM2000-1312. JDG would like to thank the Royal Society for a Ramón Areces and by Spain’s MCYT grant BFM2000-77.

APPENDIX A: ORTHOGONALITY FORCE AND STRESS

We have yet to comment on the force and stress terms containing ∂ρν/∂RI. Substituting the first term of Eq. (68) into Eqs. (67) and adding the first term of Eq. (62) we obtain a simple expression: ∑μμνHρν/∂RI. Now ρν is a function of the Hamiltonian eigenvector coefficients and occupations only (Eq. (14)). On the Born-Oppenheimer surface (BOS), EKS is stationary with respect to those coefficients and occupations, and the Hellman-Feynmann theorem guarantees that any change of them will not modify the total energy to first order, and therefore will not affect the forces. In other words, the atomic forces are the partial derivatives ∂EKS/∂RI at constant ρμ and nμ. Even in the Car-Parrinello scheme, in which the system moves out of the BOS, making the Hellman-Feynmann theorem invalid, the atomic forces are nevertheless defined as derivatives at constant ρμ and nμ.

Thus, it may seem that the terms ∂ρμ/∂RI are irrelevant for the calculation of the forces. However, in the previous discussion we have omitted to say that the KS energy must be minimized under the constraint of orthonormality of the occupied states and that, therefore, at the BOS the energy is stationary only with respect to changes of ψi which do not violate the orthonormality. With an atomic basis set, the displacement of atoms (and the deformation of the unit cell) modifies the basis, and therefore the occupied states ψi = ∑μ μμψμ, even at constant ρμ’s. And the change of the states affects their orthonormality. Thus, in order to calculate the new total energy, we need to re-orthonormalize the occupied states, by changing their coefficients cμi. Schematically, we must solve

\[ \langle \psi_i | δS | \psi_j \rangle + \langle δ\psi_i | S | \psi_j \rangle + \langle \psi_i | S | δ\psi_j \rangle = 0 \]  

where δS represents the change of Sμν due to the atomic displacements, and δψi the modification of ψi due to the change of cμi. Without lack of generality, we can expand δψi in the basis of the eigenvectors ψi as δψi = j ψj λij. Substituting this expansion into (A1) and using that \[ \langle \psi_i | S | \psi_j \rangle = \delta_{ij} \]  

we obtain \[ λ_{ij} = -\frac{1}{2} \langle \psi_j | δS | ψ_i \rangle. \]  

In terms of the coefficients cμi, we have \[ \langle \psi_j | δS | ψ_i \rangle = \sum_{μν} c_{μj} S_{μν} c_{νi} \] and

\[ \delta c_{μi} = \frac{1}{2} \sum_{ν} \sum_{ν′} c_{μj} c_{νj} δS_{νν′} c_{νi} \]

where we have used that \[ c_{μi} = \langle \bar{ψ}_μ | ψ_i \rangle, \] and

\[ \sum_{i} c_{μi} c_{νν} = \langle \bar{ψ}_μ | \bar{ψ}_ν \rangle = S_{μν}^{-1} \]
Differentiating now Eq. (B4) and using (A3) we obtain
\[ \delta \rho_{\mu \nu} = -\frac{1}{2} \sum_{\zeta \nu} \left( \rho_{\mu \eta} \delta S_{\eta \zeta} S_{\zeta \nu}^{-1} + S_{\mu \eta}^{-1} \delta S_{\eta \zeta} \rho_{\zeta \nu} \right) \] (A5)
And
\[ \sum_{\mu \nu} H_{\mu \nu} \delta \rho_{\mu \nu} = -\sum_{\mu \nu} E_{\mu \nu} \delta S_{\mu \nu} \] (A6)
where \( E_{\mu \nu} \) is the so-called energy-density matrix:
\[ E_{\mu \nu} = \frac{1}{2} \sum_{\zeta \nu} \left( S_{\mu \eta}^{-1} H_{\eta \zeta} \rho_{\zeta \nu} + \rho_{\mu \eta} H_{\eta \zeta} S_{\zeta \nu}^{-1} \right) \]
\[ = \sum_{\zeta \nu} \rho_{\mu \eta} H_{\eta \zeta} S_{\zeta \nu}^{-1} \] (A7)
where \( \epsilon_i \) are the eigenstate energies. To calculate the orthogonalization force or stress, \( \delta S_{\mu \nu} \) must be substituted by the appropriate derivative:
\[ F_{\mu \nu}^{\text{orthog}} = 2 \sum_{\nu \in l} \sum_{\nu \in l} E_{\mu \nu} \frac{\partial S_{\mu \nu}}{\partial R_{\mu \nu}} \] (A8)
This equation has been derived before in different ways, and Ordejón et al. found it also for the \( O(N) \) functional, even though it does not require the occupied states to be orthogonal. In this case, Eq. (A8) must be substituted by a more complicated expression.

Similarly, the stress contribution is
\[ \sigma_{\alpha \beta}^{\text{orthog}} = -\sum_{\mu \nu} E_{\mu \nu} \frac{\partial S_{\mu \nu}}{\partial R_{\mu \nu}} R_{\mu \nu} \] (A9)

APPENDIX B: RADIAL FAST FOURIER TRANSFORM

We consider here how to perform fast integrals of the form
\[ \psi_l(k) = \int_0^{\infty} r^2 dr j_l(kr) \psi_l(r) \] (B1)
where \( j_l(kr) \) is a spherical Bessel function and \( \psi_l(r) \) is a radial function which behaves as \( \psi_l(r) \sim r^l \) for \( r \to 0 \). Although methods to perform fast Bessel and Hankel transforms have been described previously in different fields, we have developed a simple method adapted to our needs. It is based on the fact that \( j_l(x) \) has the general form \( P_l^1(x) \sin(x) + P_l^0(x) \cos(x) \), where \( P_l^1(x), P_l^0(x) \) are simple polynomials \( P_l^\nu(x) = \sum_{n=0}^l c_n^\nu x^n \). Thus, the method involves computing \( l+1 \) fast sine and cosine transforms and add the different terms:
\[ \psi_l(k) = \sum_{n=0}^l \frac{c_n^1}{2k^{l+1-n}} \int_{-\infty}^{\infty} \psi_l(r) \sin(kr) dr + \sum_{n=0}^l \frac{c_n^0}{2k^{l+1-n}} \int_{-\infty}^{\infty} \psi_l(r) \cos(kr) dr \] (B2)

Notice that we have extended the integral to the whole real axis, defining \( \psi_l(-r) = (-1)^l \psi_l(r) \), in accordance to the behaviour \( \psi_l(r) \sim r^l, r \to 0 \). The coefficients \( c_n^\nu \) can be obtained by defining a complex polynomial \( P_l(x) = P_l^1(x) + ip_l^0(x) \), which obeys the recurrence relations:
\[ P_0(x) = i \equiv \sqrt{-1} \]
\[ P_1(x) = i - x \]
\[ P_{l+1}(x) = (2l + 1)P_l(x) - x^2P_{l-1}(x) \] (B3)

In order to perform the integrals in (B2) using discrete FFT's, we need to calculate \( \psi_l(r) \) on a regular radial grid, up to a maximum radius \( r_{\text{max}} \), beyond which \( \psi_l(r) \) is assumed to be strictly zero. The separation \( \Delta r \) between grid points determines a cutoff \( k_{\text{max}} = \pi/\Delta r \) in reciprocal space, and vice versa, \( \Delta k = \pi/r_{\text{max}} \). For convolutions, such as those involved in Eq. (28), we need \( r_{\text{max}} = r_{l}^1 + r_{l}^2 \) and \( k_{\text{max}} = \max(k_{l}^1, k_{l}^2) \), where \( r_{l}^1, r_{l}^2, k_{l}^1, k_{l}^2 \) are the cutoff radii and maximum wavevectors of \( \psi_{l,1}, \psi_{l,2} \), respectively. We must then pad with zeros the intervals \( [r_{l,1}, r_{\text{max}}], [r_{l,2}, r_{\text{max}}] \) for the forward transforms \( \psi_{l,1}(r) \to \psi_{l,2}(k) \). In practice, we set \( r_{\text{max}} = 2\max\mu(r_{l}^0), k_{\text{max}} = \max\mu(k_{l}^0) \), where \( \mu \) labels all the basis orbitals and KB projectors, and we use the same real and reciprocal grids for all orbital pairs. In this way, we need to perform the forward transform only once for each radial function \( \psi_{\mu}(r) \). Finally, notice that in Eq. (28), \( \psi_{l,1} r_{l,1}, m_{l,1}(k) r_{l,1} 2, m_{l,2}(k) \sim k^{l+1} + l^2 \) for \( k \to 0 \), while \( l+1-l \) is even and nonnegative, so that the integrands of Eq. (B2) for the backward transform are all even and well behaved at the origin.

APPENDIX C: EXTENDED-MESH ALGORITHM

We describe here a simple and efficient algorithm to handle mesh indices in three-dimensional periodic systems. Its versatility makes it suitable for several different tasks in SIESTA like neighbor-list constructions, basis orbital evaluation in the real-space integration grid, density-gradient calculations in the GGA, etc. It would be also very appropriate for other problems, like the solution of partial differential equations by real-space discretization or the calculation of the interaction energy in lattice models. For clarity of the exposition, we will describe the algorithm for a specially simple application, namely the evaluation of the Laplacian of a function \( f(r) \) using finite differences, even though the algorithm is not used in SIESTA for this purpose. In three dimensions, one generally discretizes space in all three periodic directions, using an index for each direction. For simplicity, let us consider an orthorhombic unit cell, with mesh steps \( \Delta x, \Delta y, \Delta z \). Then the simplest formula for the Laplacian is
\[ \nabla^2 f(x, y, z) = \frac{f(x+1, y, z) - 2f(x, y, z) + f(x-1, y, z)}{\Delta x^2} \]
\[ + \frac{f(x, y+1, z) - 2f(x, y, z) + f(x, y-1, z)}{\Delta y^2} \]
\[ + \frac{f(x, y, z+1) - 2f(x, y, z) + f(x, y, z-1)}{\Delta z^2} \]
A direct translation of this expression into Fortran90 code might read

\[
L_f(i_x, i_y, i_z) = \begin{cases} 
& ( f(\text{modulo}(i_x+1,n_x), i_y, i_z) + f(\text{modulo}(i_x-1,n_x), i_y, i_z) )/dx^2 \\
& + ( f(i_x,\text{modulo}(i_y+1,n_y), i_z) + f(i_x,\text{modulo}(i_y-1,n_y), i_z) )/dy^2 \\
& + ( f(i_x, i_y, \text{modulo}(i_z+1,n_z)) + f(i_x, i_y, \text{modulo}(i_z-1,n_z)) )/dz^2 \\
& - f(i_x, i_y, i_z) * (2/dx^2+2/dy^2+2/dz^2) 
\end{cases}
\]

where the indices \( i_x = (x, y, z) \) of the arrays \( f \) and \( Lf \) run from 0 to \( n_a - 1 \), as in C. There are two problems with this construction. First, the \textbf{modulo} operations are required to bring the indices back to the allowed range \([0, n_a - 1]\). And second, the use of three indices to refer to a mesh point implies implicit arithmetic operations, generated by the compiler, to translate them into a single index giving its position in memory.

A straightforward solution to these inefficiencies would be to create a neighbor-point list \( j_{\text{neighb}}(i, \text{neighb}) \), of the size of the number of mesh points times the number of neighbor points. However, although the latter are only six in our simple example, they may frequently be as many as several hundred, which generally makes this approach unfeasible. A partial solution, addressing only the first problem, is to create six (or more for longer ranges) one-dimensional tables \( j_{\text{neighb}}(i) = \text{mod}(i_a \pm 1, n_a) \) to avoid the modulo computations. Here, we describe a multidimensional generalization of this method, which solves both problems at the expense of a very reasonable amount of extra storage.

The method is based on an \textit{extended mesh}, which extends beyond the periodic unit cell, by as much as required to cover all the space that can be reached from the unit cell by the range of the interactions or the finite-difference operator. The extended mesh range is \( i^{\text{min}}_\alpha = -\Delta n_\alpha \) and \( i^{\text{max}}_\alpha = n_\alpha - 1 + \Delta n_\alpha \), where \( \Delta n_\alpha = 1 \) in our particular example, in which the Laplacian formula extends just to first-neighbor mesh points. In principle, in cases with a small unit cell and a long range, the mesh extension may be larger than the unit cell itself, extending over several neighbor cells. However, in the more relevant case of a large system, we will expect the extension region to be small compared to the unit cell.

We then consider two combined indices, one associated to the normal unit-cell mesh, and another one associated to the extended mesh

\[
i = i_x + n_x i_y + n_x n_y i_z,
\]

\[
i^{\text{ext}} = (i_x - i^{\text{min}}_x) + n_x^{\text{ext}} (i_y - i^{\text{min}}_y) + n_x^{\text{ext}} n_y^{\text{ext}} (i_z - i^{\text{min}}_z),
\]

where \( n_x^{\text{ext}} = i^{\text{max}}_\alpha - i^{\text{min}}_\alpha + 1 = n_\alpha + 2\Delta n_\alpha \). The key observation is that, if \( i^{\text{ext}} \) is a mesh point \textit{within} the unit cell \((0 \leq i_\alpha \leq n_\alpha - 1)\), and if \( j^{\text{ext}} \) is a neighbor mesh point \textit{within} its interaction range, \( i_\alpha - i_\alpha \leq \Delta n_\alpha \), then the arithmetic difference \( j^{\text{ext}} - i^{\text{ext}} \) depends only on the relative positions of \( i^{\text{ext}} \) and \( j^{\text{ext}} \) (i.e., on \( j_\alpha - i_\alpha \)), but not on the position of \( i^{\text{ext}} \) within the unit cell. We can then create a list of neighbor strides \( \Delta j^{\text{ext}} \), and two arrays to translate back and forth between \( i \) and \( i^{\text{ext}} \). One of the arrays maps the unit cell points to the central region of the extended mesh, while the other one folds back the extended mesh points to their periodically equivalent points within the unit cell. Then, to access the neighbors of a point \( i \), we \( a) \) translate \( i \rightarrow i^{\text{ext}} \); \( b) \) find \( j^{\text{ext}} = i^{\text{ext}} + \Delta j^{\text{ext}} \); and \( c) \) translate \( j^{\text{ext}} \rightarrow j \). Notice that several points of the extended mesh will map to the same point within the unit cell and that, in principle, a unit cell point \( j \) may be neighbor of \( i \) through different values of \( j^{\text{ext}} \). In our example, the innermost loop would then read

\[
L_f(i) = 0 \\
do \text{neighb} = 1, n_{\text{neighb}} \\
j^{\text{ext}} = i_{\text{extended}}(i) + \text{ij_delta}(\text{neighb}) \\
j = i_{\text{cell}}(j^{\text{ext}}) \\
L_f(i) = L_f(i) + L(\text{neighb}) * f(j) 
end do
\]

where the number of neighbor points would be \( n_{\text{neighb}} = 7 \), including the central point itself, and

\[
\begin{align*}
\text{ij_delta}(1) &= 1 & L(1) &= 1/dx^2 \\
\text{ij_delta}(2) &= -1 & L(1) &= 1/dx^2 \\
\text{ij_delta}(3) &= n_x^{\text{ext}} & L(3) &= 1/dy^2 \\
\text{ij_delta}(4) &= -n_x^{\text{ext}} & L(4) &= 1/dy^2 \\
\text{ij_delta}(5) &= n_x^{\text{ext}} n_y^{\text{ext}} & L(5) &= 1/dz^2 \\
\text{ij_delta}(6) &= -n_x^{\text{ext}} n_y^{\text{ext}} & L(6) &= 1/dz^2 \\
\text{ij_delta}(7) &= 0 & L(7) &= -2/dx^2-2/dy^2-2/dz^2 
\end{align*}
\]

Notice that the above loop is completely general, for any linear operator, using an arbitrary number of neighbor points for its finite difference representation. In fact, it is even independent of the space dimensionality. Furthermore, the index operations required are just one addition and three memory calls to arrays of range one. This inner loop simplicity comes at the expense of the two extra arrays \( i_{\text{extended}} \) and \( i_{\text{cell}} \) (of the size of the normal and extended meshes, respectively) which are generally an acceptable memory overhead. Notice, however, that the the neighbor-point list \( \text{ij_delta} \) is independent of the mesh index \( i \), what makes this array quite small in most problems of interest.

**APPENDIX D: SPARSE MATRIX TECHNIQUES**

We will describe here some of the sparse-matrix multiplication techniques used in evaluating Eqs. (87), and (88). There is a large variety of sparse-matrix representations and algorithms, each one optimized for a different kind of sparsity. The main constraint for choosing our representation and algorithms is that they must be \( O(N) \) in both memory and CPU time. We enforce this condition strictly by requiring, for example, that a vector of size \( N \) will not be reset to zero a number...
\( \sim N \) of times. In our sparse matrices, like \( S_{\mu\nu} \), \( H_{\mu\nu} \), \( c_\mu \), \( \rho_{\mu\nu} \), and \( \phi_\mu(r) \), the number \( p \) of non-zero elements in a row is typically much larger than one (but still of order \( \sim N^0 \)) and much smaller than the row size \( m \sim N^1 \). Such matrix rows are efficiently stored as a real vector of size \( p \), containing the non-zero elements, and an integer vector of the same size containing the column index of each non-zero element. The whole matrix \( A \) of \( n \) rows is then represented by two arrays \( A \) and \( j\text{col} \), of size \( n \times p \), such that \( A_{ij} = A(i,k) \), where \( j = j\text{col}(i,k) \). The problem with this representation is that, given a value \( j \) of the column index, there is no simple way to access the element \( A_{ij} \) without scanning the whole row, what is frequently too costly. One solution is to unpack a row \( i \), that will be repeatedly used, into ‘expanded form’, i.e. to transfer it to a vector \( \text{Arow} \) of the full row size \( m \) (containing also all the zeros), so that \( A_{ij} = \text{Arow}(j) \). Since \( p >> 1 \), the size of \( \text{Arow} \) is negligible compared to that of \( A \) and \( j\text{col} \).

To find the matrix product \( C \) of two sparse matrices \( A \) and \( B \)

\[
C_{ik} = \sum_j A_{ij}B_{jk}
\]

we proceed iteratively for each row \( i \) of \( A \) (which will generate the same row of \( C \)): each non-zero element \( j \) of the row is multiplied by every non-zero element of the \( j \)th row of \( B \) (whose column index is, say \( k \)) and the result is accumulated in the \( k \)th position of an auxiliary ‘expanded’ vector. After finishing with that row of \( A \) we pack the vector in sparse format into the \( i \)th row of \( C \) and restore the auxiliary vector to zero. In fact, the packing can be performed simultaneously with the product, using an auxiliary index vector instead:

\[
\text{pos}(k) = 0
\]

\[
\text{Arow}(j) = 0.
\]

\[
\text{Arow}(j) = A(i,jA)
\]

\[
\text{pos}(k) = j\text{C}
\]

\[
C(i,jC) = C(i,jC) + A(i,jA)*B(j,jB)
\]

We then unpack a row \( i \) of \( A \) and multiply it by a column \( j \) of \( B \) (a row of its transpose) for each required matrix element \( C_{ij} \) of their product:

\[
C(0).
\]

\[
\text{Arow}(0). \quad \text{Auxiliary vector}
\]

\[
\text{Arow}(0) = 0.
\]

We then unpack a row \( i \) of \( A \) and multiply it by a column \( j \) of \( B \) (a row of its transpose) for each required matrix element \( C_{ij} \) of their product:
The combination of these two matrix multiplication algorithms allows an efficient evaluation of Eqs. (84-88). Since these equations involve a trace or a relatively small subset of a final matrix, it is important to control the order and sparsity of the intermediate products, in order to keep them as sparse as possible. Notice that, once a row of $A \times B$ has been evaluated, it may be multiplied by a third matrix, to obtain a row of the final product, without need of storing the whole intermediate matrix.

To calculate the density at a grid point using Eq. (35) we need to access the matrix elements $\rho_{\mu \nu}$, and this is inefficient if they are stored in sparse format. Thus, we first copy the matrix elements, between the $n_r$ basis orbitals which are non-zero at the grid point $r$, into an auxiliary matrix array, of size $n_{aux} \times n_{aux}$, with $n_{aux} \geq n_r$. We also create a lookup table $pos$, of size equal to the total number of basis orbitals, such that $pos(\mu)$ is the position, in the auxiliary matrix, of the matrix elements of orbital $\mu$ (or zero if they have not been copied to it). If there are new non-zero orbitals at the next grid points, we keep copying them into the auxiliary matrix, until all its $n_{aux}$ slots are full, at which point we erase it and restart the process. Since successive grid points tend to contain the same non-zero basis orbitals, these copies and erasures are not frequent.

25. Some integrals, like $\langle \phi_{l_{mn}} | V_{r} | \phi_{l'_{m'n'}} \rangle$ could also be calculated in this way, but this is not the case of $\langle \phi_{l_{nn}} | V_{r}^{NA} | \phi_{l'_{n'n'}} \rangle$, which involves rather cumbersome three-center integrals of arbitrary numerical functions. Therefore, it is simpler to find the total neutral-atom potential and to calculate a single integral $\langle \phi_{l_{nn}} | V_{r}^{NA} | \phi_{l'_{n'n'}} \rangle$ in the uniform spatial grid.
Notice that our grid cutoff to represent the density is not directly comparable to the energy cutoff in the context of plane wave codes, which usually refers to the wavefunctions. Strictly speaking, the density requires a value four times larger.


