This manuscript was accepted by J. Chem. Phys. Click here to see the version of record.

Perspective: Methods for large-scale density functional calculations on metallic systems

Jolyon Aarons,[1] Misbah Sarwar,[2] David Thompsett,[2] and Chris-Kriton Skylaris[1,*]

[1] *School of Chemistry, University of Southampton, Southampton SO17 1BJ, UK*
[2] *Johnson Matthey Technology Centre, Sonning Common, Reading UK*

(Dated: November 25, 2016)

Current research challenges in areas such as energy and bioscience have created a strong need for Density Functional Theory (DFT) calculations on metallic nanostructures of hundreds to thousands of atoms to provide understanding at the atomic level in technologically important processes such as catalysis and magnetic materials. Linear-scaling DFT methods for calculations with thousands of atoms on insulators are now reaching a level of maturity. However such methods are not applicable to metals, where the continuum of states through the chemical potential and their partial occupancies provide significant hurdles which have yet to be fully overcome. Within this perspective we outline the theory of DFT calculations on metallic systems with a focus on methods for large-scale calculations, as required for the study of metallic nanoparticles. We present early approaches for electronic energy minimization in metallic systems as well as approaches which can impose partial state occupancies from a thermal distribution without access to the electronic Hamiltonian eigenvalues, such as the classes of Fermi Operator Expansions and Integral Expansions. We then focus on the significant progress which has been made in the last decade with developments which promise to better tackle the length-scale problem in metals. We discuss the challenges presented by each method, the likely future directions that could be followed and whether an accurate linear-scaling DFT method for metals is in sight.

PACS numbers: Valid PACS appear here

I. INTRODUCTION

Electronic structure theory calculations, using the Density Functional Theory (DFT) approach are widely used to compute and understand the chemical and physical properties of molecules and materials. The study of metallic systems, in particular, is an important area for the employment of DFT simulations as there is a broad range of practical applications. These applications range from the study of bulk metals and surfaces to the study of metallic nanoparticles, which is a rapidly growing area of research due to its technological relevance [1].

For example, metallic nanoparticles have optical properties that are tunable and entirely different from those of the bulk material as their interaction with light is determined by their quantization of energy levels and surface plasmons. Due to their tunable optical properties, metal nanoparticles have found numerous applications in biodiagnostics [2][3] as sensitive markers, such as for the detection of DNA by Au nanoparticle markers. Another very promising area of metallic nanoparticle usage concerns their magnetic properties which intricately depend not only on their size but also on their geometry[4], and can exhibit effects such as giant magnetoresistance[5][6]. However, by far the domain in which metallic nanoparticles have found most application so far is the area of heterogeneous catalysis. Catalytic cracking of hydrocarbons produces the fuels we use, the vehicles we drive contain catalysts to control the emissions released, and the production of certain foodstuffs also rely on catalytic processes, all of which can use metallic nanoparticles. Catalysis also has a significant role to play in Proton Exchange Membrane (PEM) fuel cells, for example, offering a promising source of clean energy, producing electricity by the electrochemical conversion of hydrogen and oxygen to water. Either monometallic or alloyed, metallic nanoparticles are used as catalysts in these processes, and anchored to a support such as an oxide or carbon. The size, shape and composition (e.g. core-shell, bulk alloy, segregated structure)[7] of these nanoparticles influence their chemical properties, and catalytically important sizes of nanoparticles (diameters of 2-10nm) can consist of hundreds to thousands of atoms.

Even though extended infinite surface slabs of one type of crystal plane are often used as models[8], which correspond to the limit of very large nanoparticles ($\gtrsim$ 5nm) we can see from Figure 1 that this limit is not reached before nanoparticles with thousands of atoms are considered. The slab model has been applied successfully in screening metal and metal alloys for a variety of reactions, providing guidance on how to improve current catalysts or identifying novel materials or compositions[9]. However, these types of models, while providing useful insight, do not capture the complexity of the nanoparticle[10], for example, the effect the particle size has on properties or edge effects between different crystal planes. The influence of the support can modify the electronic structure and geometry of the nanoparticle as well as cause "spillover effects" which may, in turn, affect catalytic activity[11]. Metallic nanoparticles are also dynamic,

*Electronic address: c.skylaris@soton.ac.uk

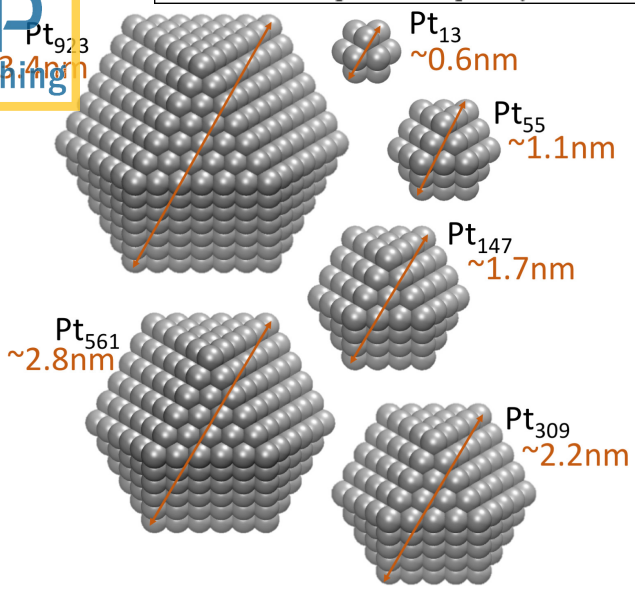

FIG. 1: The first 6 "magic numbers" of platinum cuboctahedral nanoparticles showing how the number of atoms scales cubically with the diameter; 5nm is not reached until the nanoparticle has 2869 atoms.

and interaction with adsorbates can induce a change in their shape and composition, which can modify their behaviour. A fundamental understanding of how catalytic reactions occur on the surfaces of these catalysts is crucial for improving their performance, and DFT simulations play a key role [12][13][14][15][16].

The need to understand and control the rich and unique physical and chemical properties of metallic nanoparticles provides the motivation for the development of suitable DFT methods for their study. Conventional DFT approaches used to model metallic slabs are unsuited to modelling metallic nanoparticles larger than ∼100 atoms as they are computationally very costly with an increasing number of atoms. Therefore, development of DFT methods for metals with reduced (ideally linear) scaling of computational effort with the number of atoms is essential to modelling nanoparticles of appropriate sizes for the applications mentioned above. Such methods also allow for the introduction of increased complexity into the models, such as the effect of the support or the influence of the environment such as the solvent.

In this perspective, we present an introduction to methods for large-scale DFT simulations of metallic systems with metallic nanoparticle applications in mind. We provide the key ideas between the various classes of such methods and discuss their computational demands. We conclude with some thoughts about likely future developments in this area.

II. DENSITY FUNCTIONAL THEORY FOR METALLIC SYSTEMS

Density Functional Theory has become the method of choice for electronic structure calculations of materials and molecules due to its ability to provide sufficient accuracy for practical simulations combined with manageable computational cost. DFT can be performed in its original formulation by Hohenberg and Kohn[17] where observables such as the energy are computed as functionals of the electronic density. This approach is attractive due to its computational simplicity, but its accuracy is limited due to the lack of generally applicable high accuracy approximations for the kinetic energy functional. Thus such "orbital-free" DFT calculations are not usually accurate for chemical properties. However, this is an area of ongoing developments[18] and there have been applications in specific cases such as the physical properties of liquid Li[19].

The majority of DFT calculations are performed with the Kohn-Sham approach which describes the energy of the system as a functional of the electronic density $n(\mathbf{r})$ and molecular orbitals $\{\psi_i\}$

$$E[n] = \sum_i f_i \langle\psi_i|\hat{T}|\psi_i\rangle + \int v_{\text{ext}}(\mathbf{r})n(\mathbf{r})d\mathbf{r} + E_H[n] + E_{xc}[n] \quad (1)$$

where the terms on the right are the kinetic energy, the external potential energy, the Hartree energy and the exchange-correlation energy of the electrons respectively. The molecular orbitals $\{\psi_i\}$ are the solutions of a Schrödinger equation for a fictitious system of non-interacting particles

$$\left[-\frac{1}{2}\nabla^2 + \hat{v}_{\text{KS}}\right]\psi_i(\mathbf{r}) = \epsilon_i\psi_i(\mathbf{r}) \quad (2)$$

where ϵ_i are the energy levels and where the effective potential $\hat{v}_{\text{KS}}$ has been constructed in such a way that the electronic density of the fictitious non-interacting system

$$n(\mathbf{r}) = \sum_i f_i|\psi_i(\mathbf{r})|^2 \quad (3)$$

is the same as the electronic density of the system of interest of interacting particles.

We need to point out that the solution of the Kohn-Sham equations (2) is not a trivial process as the Kohn-Sham potential $\hat{v}_{\text{KS}}[n]$ is a functional of the density n which in turn depends on the occupancies f_i and the one-particle wavefunctions ψ_i via equation 3. Thus the Kohn-Sham equations are non-linear and in practice they need to be solved iteratively until the wavefunctions, occupancies and the density no longer change with respect to each other, which is what is termed a self-consistent solution to these equations. Typically this solution is obtained by a Self-Consistent-Field (SCF) process where

$\hat{v}_{\rm KS}[n]$ is built from the current approximation to the density; then the Kohn-Sham equations are solved to obtain new wavefunctions and occupancies to build a new density; from that new density a new $\hat{v}_{\rm KS}[n]$ is constructed and these iterations continue until convergence.

For a non-spin-polarized system the occupancies f_i are either 2 or 0, depending on whether the orbitals are occupied or not. The extension of the equations to spin polarization, with occupancies 1 or 0, is trivial. This formulation of DFT is suitable for calculations on materials with a band gap (or HOMO-LUMO gap in molecules), and a wide range of algorithms have been developed for the efficient numerical solution of these equations.

The absence of a gap at the Fermi level of metallic systems makes the application of DFT approaches for insulators unsuitable for metallic systems. The extension of DFT to finite electronic temperature by Mermin can overcome this limitation by providing a canonical ensemble statistical mechanics treatment of the electrons. In this approach, the existence of a universal functional $F_T[n]$ of the electronic density for the canonical ensemble electronic system at temperature T is shown, and the Helmholtz free energy of the electronic system is written as:

$$A[n] = F_T[n] + \int v_{\rm ext}(\mathbf{r})\, n(\mathbf{r})\, d\mathbf{r} \tag{4}$$

where $v_{\rm ext}(\mathbf{r})$ is the external potential and $F_T[n]$ contains the kinetic energy, the electron-electron interaction energy and the entropy of the electronic canonical ensemble.

A Kohn-Sham mapping of canonical ensemble DFT to a system of non-interacting electrons can be carried out by analogy with the derivation of standard Kohn-Sham DFT for zero electronic temperature. In this description, the electronic system is represented by the single particle states (molecular orbitals) $\{\psi_i\}$ which are solutions of a Kohn-Sham eigenvalue equation, such as (2). The electronic density is constructed from all the single particle states

$$n(\mathbf{r}) = \sum_i f_i \psi_i(\mathbf{r}) \psi_i^*(\mathbf{r}). \tag{5}$$

where the *fractional* occupancies f_i of the states follow the Fermi-Dirac distribution:

$$f_i^{(FD)}(\epsilon_i) = \left[1 + \exp\left(\frac{\epsilon_i - \mu}{\sigma}\right)\right]^{-1}, \tag{6}$$

where μ is the chemical potential and $\sigma = k_B T$, where T is the electronic temperature and k_B is the Boltzmann constant. For this distribution of occupancies, the electronic entropy is given by

$$S(f_i) = -k_B \sum_i f_i \ln(f_i) + (1 - f_i)\ln(1 - f_i)\ . \tag{7}$$

As in the zero temperature case, the non-interacting system is constructed to have the same density as that of the interacting system. The Helmholtz free energy on the interacting electronic system is expressed as:

$$A[T, \{\varepsilon_i\}, \{\psi_i\}] = \sum_i f_i \langle \psi_i | \hat{T} | \psi_i \rangle + \int v_{\rm ext}(\mathbf{r}) n(\mathbf{r}) d\mathbf{r} + E_H[n] + E_{xc}[n] - TS[\{f_i\}], \tag{8}$$

and consists of the kinetic energy of the non-interacting electrons, and the known expressions for the external potential energy and Hartree energy of the electrons and the unknown exchange-correlation energy expression. Also, the entropic contribution to the electronic free energy $-TS[\{f_i\}]$ is included. In practice, this is a functional not only of the density but also of the molecular orbitals (as they are needed for the calculation of the non-interacting kinetic energy, and the orbital energies, which determine their fractional occupancies.

Another aspect which is particularly relevant for DFT calculations of metallic systems is Brillouin zone sampling. Because of the extremely complicated Fermi surface in some metallic systems, incredibly dense k-point sampling must be done to sample the Brillouin zone adequately, for instance by using a Monkhurst-Pack grid [20], or the VASP tetrahedron method [21]. As the systems become larger, even in metallic systems, the k-point sampling becomes less demanding as the bands flatten.

Solving these canonical ensemble Kohn-Sham equations is not trivial and presents more difficulties than working with the zero temperature Kohn-Sham equations[22]. One has now to determine, in principle, an infinite number of states (instead of just N states in the zero temperature case) - although we can in practice neglect the states whose energy is higher than a threshold beyond which the occupancies are practically zero. Another complication is the fact that most exchange-correlation functionals that are used in practice have been developed for zero temperature, so their behaviour and accuracy in a finite temperature calculation is not well understood.

Due to operations such as diagonalization, the computational effort to perform DFT calculations, whether on insulators or metals, formally increases with the third power in the number of atoms. This scaling constitutes a bottleneck in efforts to apply DFT calculations to more than a few hundred atoms, as is typically the case in problems involving biomolecules and nanostructures. Walter Kohn showed the path for removing this limitation for insulators with his theory of the "nearsightedness of electronic matter"[23] which states that the 1-particle density matrix (or equivalently the Wannier functions) decay exponentially in a system with a band gap:

$$\rho(\mathbf{r}, \mathbf{r}') = \sum_i f_i \psi_i(\mathbf{r}) \psi_i^*(\mathbf{r}') \propto e^{-\gamma|\mathbf{r}-\mathbf{r}'|}\ . \tag{9}$$

Several linear-scaling DFT programs have been developed during the last couple of decades, based on refor-

mulations of DFT in terms of the density matrix or localized (Wannier-like) functions[24][25][26][27][28] and a wide array of techniques have been formulated to allow linear scaling calculations[29][30]. Typically these methods take advantage of the exponential decay to construct highly sparse matrices and use operations such as sparse matrix multiplication and storage where the CPU and memory used scale only linearly with system size.

However, for metallic systems, there is not yet a linear-scaling reformulation. The density matrix for metals is known to have algebraic rather than exponential decay at zero temperature and while exponential decay is recovered at finite temperature, the exponent and the temperature at which it becomes useful for linear-scaling have not been explored adequately for practical use[31].

Due to the importance of calculations on metallic systems, methods have been developed with the aim of performing such calculations in a stable and efficient manner. Here we provide a review of the main classes of methods for such calculations with emphasis on recent developments, that promise to reduce the scaling of the computational effort with the number of atoms, with the aim of simulating larger and more complex metallic systems. We conclude with a discussion of the future directions that could be followed towards the development of improved methods for large-scale (and eventually linear-scaling) calculations on metallic systems.

III. METHODS FOR DFT CALCULATIONS ON METALLIC SYSTEMS

A. Electronic smearing and density mixing

One method for minimising the electronic energy in a DFT calculation is known as direct inversion in the iterative subspace (DIIS). First introduced by Peter Pulay for Hartree-Fock calculations [32] [33], this method was later adapted for DFT calculations by Kresse *et al* [34] and applied successfully to systems of up to 1000 metal atoms using the VASP code[35]. A similar technique has been implemented in the linear scaling DFT code CONQUEST[36], and while not linear scaling for metals, the authors show how such techniques may be applied successfully to density matrix based DFT approaches and perform some operations in a linear scaling way.

The central assumption in this method is that a good approximation to the solution can be constructed as a linear combination of the approximate solutions of the previous m iterations.

$$x^{i+1} = \sum_{j=0}^{m-1} \alpha_{i-j}\, x^{i-j}, \tag{10}$$

where x can represent any of the variables of the solution, such as the Hamiltonian matrix, the density or the one-particle wavefunctions. The DIIS method constructs a set of linear equations to solve which yield the expansion coefficients α_j.

Kresse *et al* discuss two uses for DIIS, RMM-DIIS, for the iterative diagonalization of a Hamiltonian and Pulay mixing of densities. RMM-DIIS, which is a form of iterative diagonalization allows for the first N eigenpairs of a Hamiltonian matrix to be found without performing a full diagonalization of the whole matrix - detailed information can be found in Kresse and Furthmüller[34]. In the following paragraphs, we will discuss the Pulay mixing of densities.

Directly inputting the output density (from equation 3) into the next construction of the Hamiltonian can result in an unstable SCF procedure where large changes in the output density result from small changes in the input, known as charge-sloshing. Density mixing attempts to damp oscillations in the SCF procedure, by mixing densities from previous iterations with the density produced from the wavefunctions and occupancies at the current iteration (the output density).

The simplest approach is linear mixing of densities where the new density, $\mathrm{n}^{i+1}(\mathbf{r})$ is constructed from the density of the previous iteration as

$$\mathrm{n}_{\mathrm{in}}^{i+1}(\mathbf{r}) = \alpha\, \mathrm{n}_{\mathrm{out}}^{i}(\mathbf{r}) + (1-\alpha)\, \mathrm{n}_{\mathrm{in}}^{i}(\mathbf{r}), \tag{11}$$

where $\mathrm{n}_{\mathrm{in}}^{i}(\mathbf{r})$ and $\mathrm{n}_{\mathrm{out}}^{i}(\mathbf{r})$ are the input and output SCF solutions at the ith iteration. A better option regarding stability and efficiency is to use a mixing scheme based upon a history of previous densities, such as Pulay's DIIS procedure. Under the constraint of electron number conservation $\sum_i \alpha^i = 1$, the next input density is given as

$$\mathrm{n}_{in}^{i+1}(\mathbf{r}) = \sum_{j=0}^{m-1} \alpha_{i-j}\, \mathrm{n}_{in}^{i-j}(\mathbf{r}). \tag{12}$$

In DIIS, the mixing coefficients α_i are found by firstly considering the density residual,

$$R[n_{in}^{i}(\mathbf{r})] = n_{out}^{i}(\mathbf{r}) - n_{in}^{i}(\mathbf{r}), \tag{13}$$

which can incidentally be used to reformulate linear mixing as

$$n_{in}^{i+1}(\mathbf{r}) = n_{in}^{i}(\mathbf{r}) + \alpha R[n_{in}^{i}(\mathbf{r})]. \tag{14}$$

If the DIIS assumption is used that the residuals are linear in $n_{in}(\mathbf{r})$, then

$$R[n_{in}^{i+1}(\mathbf{r})] = \sum_{j=0}^{m-1} \alpha_{i-j}\, R[\mathrm{n}_{in}^{i-j}(\mathbf{r})], \tag{15}$$

the Pulay mixing coefficients, α_i which minimize the norm of the residual associated with the current iteration are given as

$$\alpha_i = \frac{\sum_j (\mathbf{A}^{-1})_{ji}}{\sum_{kl} (\mathbf{A}^{-1})_{kl}}, \tag{16}$$

where **A** is the matrix with elements a_{ij} the dot products of the residuals with index i and j.

To apply such a scheme for metals, a preconditioner is additionally required to ensure the convergence has an acceptable rate. The Kerker preconditioner damps long-range components in density changes more than short-range components[37], because, in metals, long-range changes in the density are often the cause of charge-sloshing effects,

$$G(k) = A\frac{k^2}{k^2 + k_0^2} \tag{17}$$

where A is a mixing weight and the parameter k_0 effectively defines a "length-scale" for what is meant by long-range. The k are the wave vectors via which the density is represented. Thus $G(k)$ is used to multiply the n_{in}^{i+1} density in reciprocal space and thus damp the "charge-sloshing" that can occur at long wavelengths. Furthermore, as in all approaches for metals, a smeared occupancy distribution must be used.

For metallic systems, the choice of smearing function is also a major consideration. While a Fermi-Dirac occupation can be used (6), many more options exist which exhibit advantages and disadvantages over Fermi-Dirac. The major benefit of Fermi-Dirac occupation is that the electronic smearing corresponds to a physical thermal distribution at temperature T. If thermally distributed electrons are not of interest and the free energies obtained will be "corrected" to approximate zero Kelvin energies, then any other sigmoidal distribution which converges to a step function in some limit might be used. A significant downside to Fermi-Dirac smearing is that the function tails off very slowly, so a large number of very slightly occupied conduction bands must be used to capture all of the occupied states fully.

Gaussian smearing solves the issue with the long tails of the distribution neatly. The smearing function is given by

$$f_i^{(G)}(\epsilon_i) = \frac{1}{2}\left[1 - \text{erf}\left(\frac{\epsilon_i - \mu}{\sigma}\right)\right], \tag{18}$$

where a shifted and scaled error function is used for state occupancy. Using Gaussian smearing, therefore, means that calculations will need relatively fewer partially occupied conduction states to be effective. In this approach, the "smearing width", σ no longer has a physical interpretation and the free energy functional to be minimized becomes an analogue *generalized* free energy. Despite this, the approach has been used successfully for decades (it is the default smearing scheme in the CASTEP plane-wave DFT code, for instance) and the results can be effectively extrapolated to zero σ [22] by using

$$E_{\sigma=0} = \frac{1}{2}\left(A' + E\right) + O(\sigma^2), \tag{19}$$

where A' is the generalized free energy functional. This is a *post hoc* correction and hence is applied at the end of a calculation, while forces and stresses are not variational with respect to this unsmeared energy.

Another approach to recover non-smeared results for metals is to use a smearing function which knocks out the σ dependence of the generalized entropy. First order Methfessel-Paxton Hermite polynomial smearing [38],

$$f_i^{(MP)}(\epsilon_i) = \frac{1}{\sqrt{\pi}}\left(\frac{3}{2} - \left(\frac{\epsilon_i - \mu}{\sigma}\right)^2\right)e^{-\left(\frac{\epsilon_i-\mu}{\sigma}\right)^2}, \tag{20}$$

has only a quartic dependence on σ, so that results obtained using this approach need not be extrapolated back to zero σ, but may be used directly. A significant disadvantage of this method is that it yields non-physical negative occupancies. This can lead to difficulties in finding the particle number conserving chemical potential, as the thermal distribution has degeneracies, but may also lead to more serious concerns such as areas of negative electron density.

Marzari-Vanderbilt "cold smearing"[39] solves all of these issues by using a form

$$f_i^{(MV)}(x_i) = \frac{1}{\pi}\left(ax^3 - x^2 - \frac{3}{2}ax + \frac{3}{2}\right)e^{-x^2}, \tag{21}$$

where $x_i = (\epsilon_i - \mu)/\sigma$ and a is a free parameter for which the authors suggest a value of -0.5634.

Head-Gordon *et al* have explored expansions of the various smearing functions and present comparisons of convergence with the order of the expansions and the number of operations involved [40].

B. EDFT

Density mixing is non-variational in the sense that it can produce "converged" solutions below the minimum, but more than this it can take a long time to reach convergence, or it can even be unstable without a reliable mixing scheme and preconditioner. This is particularly the case in systems with a large number of degrees of freedom. Ideally, a variational approach, where every step is guaranteed to lower the energy towards the ground state energy would be preferred over a density mixing approach, if it could ensure that the ground state energy would always be reached through a stable progression. Marzari *et al* proposed such a scheme[41] in 1997, which in the literature has become familiar as Ensemble DFT (EDFT). We need to note that EDFT should not be confused with Mermin's original finite temperature DFT formalism that is often referred to with the same name. A variational progression towards the ground state energy is achieved by decoupling the problems of optimising the 1-particle wavefunctions and of optimising the electronic occupancy of these wavefunctions with respect to the energy of the system. This is achieved by performing an occupancy optimization process with fixed wavefunctions at every step in the optimization of the wavefunctions themselves.

The Helmholtz free energy of equation 8 is expressed in an equivalent way in terms of the occupancies of the molecular orbitals $A[T;\{f_i\},\{\psi_i\}]$, due to the existence of a one-to-one mapping between the orbital energies and their occupancies . The method generalizes the occupancies to non-diagonal form $\{f'_{i,j}\}$ by working with molecular orbitals $\{\psi'_i\}$ which can be considered to be a unitary transformation of the orbitals in which the occupancies are diagonal. In practice, the following energy expression is used $A[T;\{f'_{i,j}\},\{\psi'_i\}]$. Working with this expression provides a stable direct energy minimization algorithm because the optimization of occupancies is not slowed down by nearly degenerate unitary rotations of the molecular orbitals. The optimization of the energy is done in two nested loops as follows: within the inner loop, occupancy contribution to the energy is minimized and in the outer loop the orbital contribution is minimized using the projected functional

$$A[T,\{\psi'_i\}] = \min_{\{f'_{ij}\}} A[\mathrm{T};\{\psi'_i\};\{f'_{ij}\}], \tag{22}$$

which allows an unconstrained optimization of both the orbitals and the occupancies.

Despite the obvious advantages of the Marzari method, one weakness is that the mapping from the occupancies to the unbounded range of orbital energies is very ill-conditioned, as typically there is a large number of occupancies close to zero that can map on to very different orbital energies.

Freysoldt, *et al* [42] have attempted to address this issue by developing an equivalent scheme where one works directly with the molecular orbital energies instead of the occupancies. In the spirit of the Marzari approach, they employ a non-diagonal representation of orbital energies, which is the Hamiltonian matrix, and minimize the following functional

$$A[\mathrm{T}] = \min_{\{H'_{ij}\}\{\psi'_i\}} A[\mathrm{T};\{\psi'_i\};\{H'_{ij}\}]. \tag{23}$$

The combined minimization of $\{H'_{ij}\}$ and $\{\psi'_i\}$ is performed using line searches along an augmented search direction in a joint space.

Alternatively, such a scheme can be done in two loops following the Marzari philosophy. In the outer, orbital loop, a preconditioned gradient of the functional with respect to orbitals is used, while in the Hamiltonian space a search direction between the current matrix and that corresponding to the non-self consistent energy minimum is used. The non-self consistent Hamiltonian is constructed by firstly calculating the occupancies $\{f_i\}$ from the Hamiltonian eigenvalues via an occupancy distribution function. A new electronic charge density can then be calculated as

$$n_{new}(\mathbf{r}) = \sum_n f_i \psi_i(\mathbf{r})\psi_i^*(\mathbf{r}), \tag{24}$$

and from this density a new (non self-consistent) Hamiltonian matrix $\tilde{H}$ is constructed which is the end-point to the inner loop line search for updating the Hamiltonian:

$$H_{ij}^{n+1} = (1-\lambda)H_{ij}^n + \lambda\tilde{H}_{ij}^n. \tag{25}$$

Of course, as in the method of Marzari, in order to perform calculations with this approach a diagonalization of the Hamiltonian matrix must occur at each inner loop-step, resulting in a cubically scaling algorithm.

Recently an EDFT method has been implemented within the linear-scaling DFT package ONETEP[43]. ONETEP uses a minimal set of non-orthogonal generalized Wannier functions (NGWFs) which are represented in terms of coefficients of a basis set of periodic sinc (psinc) functions and are strictly localized in space. The psinc basis functions are related to plane waves through a unitary rotation.

ONETEP optimizes the NGWFs variationally with respect to the energy in a manner similar to the outer loop of the EDFT method. The psinc basis set allows a systematic convergence to the complete basis set limit with a single plane-wave cutoff energy parameter much like plane waves do for periodic calculations. Techniques such as PAW, which has gained favour in many codes for the efficient representation of core electrons, may also be applied to the spatially localized NGWFs $\{\phi_\alpha\}$.

ONETEP is usually employed for large systems of hundreds or thousands of atoms within the linear-scaling mode, but a step function occupancy distribution limits its applicability to insulating systems. To exploit nearsightedness of electronic matter, terms in the density matrix,

$$\rho(\mathbf{r},\mathbf{r}') = \sum_n f_n \psi_n(\mathbf{r})\psi_n^*(\mathbf{r}'), \tag{26}$$

which is also equal to the EDFT generalized occupancy matrix (if a Fermi-Dirac occupancy function is used) are truncated based on spatial separation and the rest of the DFT can be written in terms of this quantity as

$$n(\mathbf{r}) = \rho(\mathbf{r},\mathbf{r}). \tag{27}$$

ONETEP constructs the density matrix as an expansion in the NGWFs as:

$$\rho(\mathbf{r},\mathbf{r}') = \sum_{\alpha\beta} \phi_\alpha(\mathbf{r})K^{\alpha\beta}\phi_\beta^*(\mathbf{r}'), \tag{28}$$

where $K^{\alpha\beta}$ is the generalized occupancy of the NGWFs, i.e. its eigenvalues are $\{f_i\}$ and is known as the density kernel. EDFT is implemented [44] by taking the Hamiltonian eigenvalue approach of Freysoldt *et al*, but also taking advantage of the localized nature of the NGWFs which obviously also requires dealing with non-orthogonality.

In ONETEP EDFT, the free energy functional is optimized firstly with respect to the Hamiltonian matrix,

$\{H_{\alpha\beta}\}$ while keeping the NGWFs $\{\phi_\alpha\}$ fixed. The orbital contribution to the free energy can then be minimized by optimising a projected Helmholtz functional,

$$A'[\mathrm{T};\{\phi_\alpha\}] = \min_{\{H_{\alpha\beta}\}} A[\mathrm{T};\{H_{\alpha\beta}\};\{\phi_\alpha\}] \qquad (29)$$

with respect to $\{\phi_\alpha\}$.

In practice, the diagonalization of the Hamiltonian within the inner loop is done by orthogonalizing, and then by solving a standard eigenvalue problem using parallel solvers. It can alternatively be done directly with a generalized parallel eigensolver, but to facilitate the expedient replacement of the eigensolver with an expansion method (discussed in the next section) orthogonalization is performed first. The orthogonalization step can be carried out in many ways; using Gram-Schmidt, Cholesky or Löwdin methods, or simply by taking the inverse of the overlap matrix and left-multiplying the Hamiltonian matrix by this to construct a new orthogonal Hamiltonian.

The Löwdin approach requires the overlap matrix to the power 1/2 and -1/2. Due to work of Jansík *et al* [45], it is possible to construct Löwdin factors in a linear-scaling way using a Newton-Shultz approach, without an expensive eigendecomposition.

The Hamiltonian matrix in the basis of Löwdin orthogonalized orbitals can be written as

$$H'_{ij} = (\mathbf{S}^{-1/2}\mathbf{H}\mathbf{S}^{-1/2})_{ij} \qquad (30)$$

and subsequently, the Kohn-Sham Hamiltonian eigenvalues can be found by diagonalising this orthogonal matrix with a standard parallel eigenvalue solver.

The next step in this EDFT inner-loop method is to construct a non-self consistent density matrix from the Hamiltonian matrix.

This density matrix can be represented as a function of the eigenvalues ε_i of the Hamiltonian as:

$$K^{\alpha\beta} = \sum_i^N M^\alpha{}_i f(\varepsilon_i) M_i^{\dagger\,\beta}, \qquad (31)$$

where the eigenvalues (band occupancies) of the finite-temperature density kernel, $K^{\alpha\beta}$ are given in terms of a smearing function (6): The matrix M contains the eigenvectors of the Hamiltonian eigenproblem:

$$H_{\alpha\beta}M^\beta{}_i = S_{\alpha\beta}M^\beta{}_i\,\varepsilon_i. \qquad (32)$$

Such eigenvalue based approaches will always scale as $O(N^3)$ as they employ matrix diagonalization algorithms. The construction of the Hamiltonian and orthogonalization procedures are however linear-scaling which significantly reduces the prefactor of this approach (See Fig. 2) versus a comparable EDFT implementation in conventional plane-wave codes.

A minimization is performed in the space of Hamiltonians and a search direction is constructed as

$$\Delta^{(n)}_{\alpha\beta} = \tilde{H}^{(n)}_{\alpha\beta} - H^{(n)}_{\alpha\beta}, \qquad (33)$$

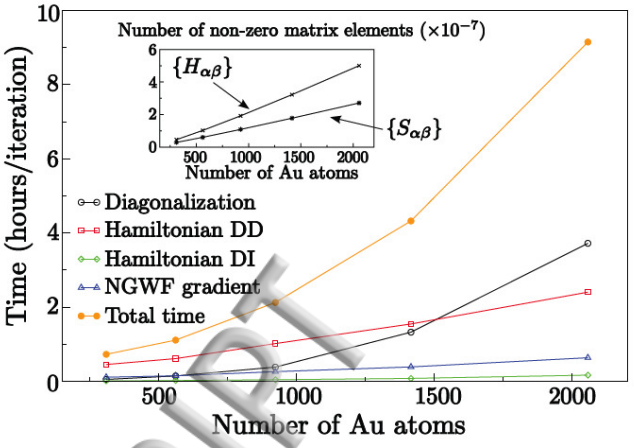

FIG. 2: The scaling of the computational effort with the number of atoms in the EDFT metals method in ONETEP. The computationally demanding steps of the calculation such as the construction of the density dependent (DD) and density independent (DI) parts of the Hamiltonian and the NGWF gradient are linear-scaling and thus allow large numbers of atoms to be treated. However, there is a diagonalization step which is cubic-scaling and eventually dominates the calculation time. Reproduced from [44], with the permission of AIP publishing.

where a new Hamiltonian matrix, $\tilde{H}$ has been constructed from the updated density kernel. The Hamiltonian can then be updated as

$$H^{(n+1)}_{\alpha\beta} = H^{(n)}_{\alpha\beta} - \lambda\Delta^{(n)}_{\alpha\beta}, \qquad (34)$$

where λ is a scalar line search parameter. Once the inner loop has converged, and the contribution to the energy from the Hamiltonian can no longer be reduced, the outer loop is resumed and this process continues to self-consistency. At self-consistency, i.e. when the calculation has converged, the Hamiltonian will commute with the density kernel.

C. Matrix Inversion

In all of the methods which follow in section III D, there is a need for matrix inversion or factorization. In fact, even if one needs an orthogonal representation of the Hamiltonian matrix in general, or for computing derivatives of molecular orbitals for optimizing these orbitals, an inverse or factorization of the overlap matrix is necessary.

The overlap matrix of strictly localized functions is itself a highly sparse matrix, since spatial separation of atoms in systems with many atoms, ensures that most matrix elements will be zero and hence **S** is localized. As for the inverse overlap sparsity of insulators, this can be shown to be exponentially localized by treating **S** as a Hamiltonian matrix and taking its Green's function at a

shift of zero[46]. Since Green's functions are always exponentially localized at shifts outside of the eigenvalue range of the "Hamiltonian" (the overlap matrix is positive definite), the inverse overlap matrix too is exponentially localized. Nunes and Vanderbilt[46] state that the decay length of the exponential localization is dependent upon the ratio of maximum eigenvalue to minimum eigenvalue. So, systems involving overlap matrices with large ℓ^2 condition number will have a long decay length in the inverse overlap matrix.

In many cases it may be desirable to entirely avoid matrix inversion and solve matrix equations directly using a decomposition, but this entirely depends on whether a matrix decomposition can be efficiently computed for a sparse matrix in parallel[47].

If inversion is called for, as in several of the techniques in section III D, because the matrices involved are sparse, conventional inversion techniques cannot be employed efficiently, so more specialized techniques are used. The Newton-Shultz-Hotelling (NSH) inversion is a generalization of the application of the Newton-Raphson method for iterative inversion of scalars to matrices. As in the scalar case, the roots of an equation, in this case $\mathrm{f}(\mathbf{X}) = \mathbf{Q} - \mathbf{X}^{-1}$ are found iteratively with the Newton-Raphson approach. This can be performed simply by recursive application of the iteration:

$$\mathbf{X}_{n+1} = \mathbf{X}_n(2\mathbf{I} - \mathbf{Q}\mathbf{X}_\mathbf{n}), \tag{35}$$

where $\mathbf{Q}$ is the matrix to be inverted and $\mathbf{X}_n$ converges quadratically to the inverse matrix in the limit of $n \to \infty$, provided that $\mathbf{Q}$ is non-singular and $\mathbf{X}_0$ is initialized with a matrix which guarantees convergence of the iteration[48],

$$\mathbf{X}_0 = \alpha\mathbf{Q}^T, \tag{36}$$

where $\alpha = 1/(||\mathbf{Q}||_1\,||\mathbf{Q}||_\infty)$ is a good choice in general, according to Pan and Schreiber[49]. This is discussed in detail in Ozaki[50], along with several other possibilities for matrix inversion in linear-scaling electronic structure theory calculations. This method, however, has been picked up by the community and is used extensively in modern methods[51] and codes[52].

Another approach which has been developed for the inversion of matrices on the poles of a contour integral (see section III D 4) but which may be applicable more widely is Selected Inversion (SI). In effect this amounts to a Cholesky expansion (or LDL or LU decomposition) of the matrix in order to compute selected matrix elements, such as the diagonal, of the inverse matrix exactly [53].

Assuming that the matrix to be inverted is sparse then it can be decomposed as

$$\mathbf{L}\mathbf{D}\mathbf{L}^\dagger = \mathbf{P}\mathbf{Q}\mathbf{P}^T, \tag{37}$$

with sparse $\mathbf{L}$, which are lower triangular factor matrices and where $\mathbf{D}$ is a diagonal matrix and $\mathbf{P}$ is a permutation matrix chosen to reduce the amount of "fill-in" in the sparsity pattern of $\mathbf{L}$ with respect to $\mathbf{Q}$. If we write $\mathbf{P}\mathbf{Q}\mathbf{P}^T \to \mathbf{Q}$ for simplicity, then the inverse matrix,

$$\mathbf{Q}^{-1} = \mathbf{L}^{-\dagger}(\mathbf{L}\mathbf{D})^{-1}, \tag{38}$$

so, provided that the LDL factorization of the sparse $\mathbf{Q}$ matrix can be performed, the potentially sparse (but likely with high fill-in) $\mathbf{Q}^{-1}$ matrix may be calculated trivially by back substitution.

Rather than forming the inverse matrix in this fashion, however, SI may be used to calculated solely the selected elements of the inverse matrix on a pre-defined sparsity pattern. Firstly, the LDL factorization can be computed recursively, using the block factorization:

$$\begin{pmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{pmatrix} = \begin{pmatrix}\mathbf{I} & 0\\ \mathbf{C}\mathbf{A}^{-1} & \mathbf{I}\end{pmatrix}\begin{pmatrix}\mathbf{A} & 0\\ 0 & \mathbf{D}-\mathbf{C}\mathbf{A}^{-1}\mathbf{B}\end{pmatrix}\begin{pmatrix}\mathbf{I} & \mathbf{A}^{-1}\mathbf{B}\\ 0 & \mathbf{I}\end{pmatrix}. \tag{39}$$

In the case that $\mathbf{A}$ is a scalar, a and the matrix to be factorized is block-symmetric, so that $\mathbf{C}$ is a vector $\mathbf{c} = \mathbf{b}^T$, and then

$$\begin{pmatrix}a & \mathbf{b}\\ \mathbf{b}^T & \mathbf{D}\end{pmatrix} = \begin{pmatrix}1 & 0\\ \mathbf{l} & \mathbf{I}\end{pmatrix}\begin{pmatrix}a & 0\\ 0 & \mathbf{D}-\mathbf{b}^T\mathbf{b}/a\end{pmatrix}\begin{pmatrix}1 & \mathbf{l}^T\\ 0 & \mathbf{I}\end{pmatrix} \tag{40}$$

where the vector $\mathbf{l} = \mathbf{b}/a$. The Schur complement, $\mathbf{S} = \mathbf{D} - \mathbf{b}^T\mathbf{b}/a$ can then be factorized recursively and so the inverse can be written using the symmetric block matrix inverse formula, for scalar a, as

$$\mathbf{Q}^{-1} = \begin{pmatrix}a^{-1} + \mathbf{l}^T\mathbf{S}^{-1}\mathbf{l} & -\mathbf{l}^T\mathbf{S}^{-1}\\ -\mathbf{S}^{-1}\mathbf{l} & \mathbf{S}^{-1}\end{pmatrix}. \tag{41}$$

In this way, the inverse can be computed recursively, along with the factorization, by descending through the recursion hierarchy until the Schur complement can be formed and inverted using scalar operations, and then the elements of each successive inverse are calculated from the inverse of the previous level, as the hierarchy is ascended. By computing the inverse recursively in this fashion, the authors[53] show that if the diagonal of the inverse matrix is required, only those elements of the inverse Schur complements which have an index equal to the index of a non-zero element of an $\mathbf{l}$ vector need to be calculated. This corresponds to calculating the inverse matrix for only those elements for which the $\mathbf{L}$ matrix has a non-zero element. The authors report computational scaling of $O(N^{3/2})$ for 1D systems.

Other methods exist for calculating selected elements of the inverse of a sparse matrix with similarly reduced scaling. The FIND algorithm[54], which works by permuting the original matrix to make a desired diagonal element the trivial solution to the equation with one unknown after an LDL or LU decomposition. This is then repeated for every diagonal element, but the computational cost is made manageable by performing partial decompositions and reusing information. Another option is based on Takahashi's equations[55], this has the advantage of not needing to invert the triangular factors.

An important point to note with these techniques for inversion is that a sparse matrix factorization must be used for them to be practical and beneficial. Routines for doing this are available in libraries such as SuperLU [47] and MUMPS [56]. Both of these libraries contain distributed-memory parallel implementations of matrix factorizations for sparse matrices, but it is well known that the parallel scaling of such approaches is somewhat limited [57].

If a general sparse matrix is factorized in such a way, however, then the factors may be dense, so the approach does not exploit the sparsity at all and so is not helpful in making a fast algorithm. In order to circumvent such a situation, the rows and columns of the matrix to be factorized must be reordered prior to factorization with a fill-in reducing reordering. The ideal reordering in terms of number of non-zero elements in the triangular factor is unknown, as calculating it is an NP-complete problem, however good [58] algorithms exist for finding approximations. Unfortunately, the best of these are not well parallelizable. Despite this, libraries such as ParMETIS[59] and PT_SCOTCH[60] exist and do this operation with the best parallel algorithms currently known. It is perhaps worth noting that the best performing reordering methods differ, depending whether serial or parallel computers are used for the calculation.

D. Expansion Approaches

1. *Introduction to expansion approaches*

As in linear-scaling methods for insulators, the idea behind expansion and other reduced scaling approaches is to be able to perform an SCF calculation without using an (inherently) cubic-scaling diagonalization step. Thus these methods attempt to compute a converged density matrix for metallic systems directly from the Hamiltonian, using for example (potentially linear-scaling) matrix multiplications, as the Fermi operator expansion approach which is one of the early approaches of this kind. in this way, the cycle of having to diagonalize the Hamiltonian to obtain wavefunctions and energies from which to build the occupancies and the density is no longer needed.

The idea for Fermi operator expansions (FOE) was first proposed by Goedecker & Colombo in 1994 [61]. The density matrix with finite-temperature occupancies, which can be constructed as a function of the eigenvalues of the one-particle Hamiltonian, as in EDFT, is instead built as a matrix-polynomial expansion of the occupancy function. In most works this occupancy function is the matrix analogue of the Fermi-Dirac occupancy function, but can equally be a generalized occupancy function, such as Gaussian smearing, or either Methfessel-Paxton[62] or Marzari-Vanderbilt[41] cold-smearing. For simplicity, this work will only consider Fermi-Dirac style occupancy smearings, however.

As an alternative to the eigenvalue based methods mentioned in the previous section, FOE methods begin instead by writing the occupancy formula in matrix form:

$$f(\mathbf{X}) = (\mathbf{I} + e^{\mathbf{X}})^{-1}, \tag{42}$$

where I is the identity matrix and

$$\mathbf{X} = (\mathbf{H} - \mu\mathbf{I})\beta. \tag{43}$$

where $\mathbf{H}$ is the Hamiltonian matrix expressed in an orthonormal basis, μ is the chemical potential and $\beta = 1/k_BT$. For instance, we can obtain $\mathbf{H}$ using a Löwdin orthogonalization, see equation 30, using for instance, the previously mentioned iterative refinement method of Jansík *et al*, or the with the combination of a recursive factorization of the overlap matrix and iterative refinement of the approximate result[63], as proposed by Rubensson *et al* [64]. This method for calculating the Löwdin factor (and inverse Löwdin factor) allows for rapid convergence in a Newton-Schultz style refinement, scheme, while providing heuristics for the requisite parameters, without reference to even extremal eigenvalues. As the title of the work suggests, for sparse matrices, this method can be implemented in a linear scaling way.

In practice, the matrix formula of equation 42 cannot be applied directly as the condition number of the matrix to be inverted will be too large in general. Instead, the operation in 42 is performed as an approximate matrix-expansion of this function.

In all of the following methods, matrix inversion plays an important role. In the following sections we will discuss three of the major flavours of expansion approaches; firstly the rational expansions, where the density matrix is made up from a sum involving inverses of functions of the Hamiltonian matrix, secondly the Chebyshev expansions, where the density matrix is formed from as a matrix Chebyshev polynomial approximating the Fermi-Dirac function of the eigenvalues, and thirdly the recursive approaches, where simple polynomial functions are applied recursively to the Hamiltonian matrix to produce the density matrix.

2. *Chebyshev expansion approaches*

A way to perform the operation of applying the Fermi-Dirac occupation function to the eigenvalues of a Hamiltonian matrix was proposed by Goedecker and Teter in 1995 as a Chebyshev expansion [65]. In order to do this, the Fermi function is written as a Chebyshev expansion

$$f(\mathbf{X}) = \sum_{i=0}^{N} a_i \mathbf{T}_i(\mathbf{X}), \tag{44}$$

where $\{\mathbf{T}_i\}$ are the Chebyshev matrices of the first kind, of degree i and $\{a_i\}$ are the expansion coefficients. The

Chebyshev matrices are in the standard form:

$$\begin{aligned} \mathbf{T}_0(\mathbf{X}) &= \mathbf{I} \\ \mathbf{T}_1(\mathbf{X}) &= \mathbf{X} \\ \mathbf{T}_{n+1}(\mathbf{X}) &= 2\mathbf{X}\mathbf{T}_n(\mathbf{X}) - \mathbf{T}_{n-1}(\mathbf{X}). \end{aligned} \tag{45}$$

The coefficients can be developed by simply taking the Chebyshev expansion of the scalar Fermi-Dirac function and applying these to compute a Chebyshev Fermi operator evaluation. As the Chebyshev polynomials can also be defined trigonometrically; $T_n(\cos(\omega)) = \cos(n\omega)$, then the coefficients can be found simply through a Discrete Cosine Transform:

$$\{a_i\} = \mathrm{DCT}\left(\frac{1}{1+e^{\cos(x)}}\right), \tag{46}$$

where $x_i = 2(((e_i - \mu)\beta) - e_0)/(e_N - e_0) - 1$ so that the range of equispaced e_i covers at least the range of the eigenrange of the Hamiltonian matrix (e_0:e_N) and the interval is scaled and shifted to cover the useful interpolative range of Chebyshev polynomials $(-1:1)$. This operation need only be performed once, the coefficients are stored for every subsequent evaluation and is effectively negligible in the complexity and timing of the algorithms based on such an expansion. This way of performing the expansion requires approximately N terms for a given accuracy, where N is a function of the smearing width, β, the required accuracy 10^{-D} and the width of the Hamiltonian eigenvalue spectrum ΔE. This implies that if matrix multiplication operations can be at best $O(N)$ scaling for tight-binding calculations such as those for which this technique was first proposed, then of the order of $M \approx D\beta\Delta E$ matrix multiplications would be required [66].

An improvement to the original Chebyshev series of Goedecker was proposed by Liang and Head-Gordon in 2003; this uses a divide and conquer approach to re-sum the terms of a truncated Chebyshev series[40], which is closely related to the divide and conquer approach for standard polynomials suggested in S. Paterson and L. J. Stockmeyer [67]. This approach factors the polynomial into a number of terms which share common subterms. The authors propose three algorithms, the simplest of which is recursive binary subdivision, where the polynomial terms are grouped into even and odd sub-polynomials, for instance

$$\sum_{i=0}^{N} a_i \mathbf{X}^i = \sum_{j=0}^{N_{\mathrm{even}}} a_{2j}\mathbf{X}^{2j} + \sum_{j=0}^{N_{\mathrm{odd}}} a_{2j+1}\mathbf{X}^{2j+1}, \tag{47}$$

then a factor of $\mathbf{X}$ can be taken out of the odd sum, to give

$$\sum_{i=0}^{N} a_i \mathbf{X}^i = \sum_{j=0}^{N_{\mathrm{even}}} a_{2j}(\mathbf{X}^2)^j + \mathbf{X}\sum_{j=0}^{N_{\mathrm{odd}}} a_{2j+1}(\mathbf{X}^2)^j, \tag{48}$$

If this process is performed recursively, almost a factor of two in matrix multiplications can be saved for each subdivision. If done carefully, the amount of work to produce these terms and combine them to construct the full polynomial is less than the amount of work to construct the polynomial directly, because of the unduplicated effort. To experience the most gain, this approach is repeated recursively, until the sub-division can no longer be performed (efficiently). Using such a divide and conquer scheme reduces the scaling of a Chebyshev decomposition of the Fermi operator to $\mathrm{O}(\sqrt{N})$ number of matrix multiplications.

Rather than decomposing the Fermi operator directly in terms of a truncated matrix-polynomial, Krajewski and Parrinello propose a construction based on an exact decomposition of the grand-canonical potential for a system of non-interacting Fermions[68]:

$$\Omega = -2\mathrm{Tr}\,\ln(\mathbf{I} + e^{-X}). \tag{49}$$

By decomposing the quantity in parentheses as

$$\mathbf{I} + e^{-X} = \prod_{l=1}^{P} \mathbf{M}_l \mathbf{M}_l^*, \tag{50}$$

where

$$\mathbf{M}_l = \mathbf{I} - e^{i(2l-1)\pi/2P} e^{-\mathbf{H}/2P}, \tag{51}$$

where P is an integer which Ceriotti at al. suggest should be in the range 500-1000 for optimal efficiency[69]. Krajewski *et al* show that expectation values of physical observables can be calculated using a Monte-Carlo, stochastic method leading to an $O(N)$ scaling approach, at the expense of noise on values of the calculated properties[70]. In a separate publication, the authors also show that the approach scales linearly without Monte-Carlo sampling in the case of 1D systems such as carbon nanotubes[71].

In later work, Ceriotti, Kühne and Parrinello extend the approach to allow better scaling in general. With this approach the usually expensive and difficult application of the matrix-exponential operator is reduced to a number of more tractable matrix exponential problems, in the sense that approximating them effectively through low-order truncated matrix-polynomial expansions is achievable. Through use of this formalism, the grand-canonical density matrix, otherwise known as the density matrix with Fermi-Dirac occupancy can be expressed exactly in terms of this expansion by taking the appropriate derivative of the grand potential,

$$\frac{\delta\Omega}{\delta\mathbf{H}} = (\mathbf{I} + e^{-X})^{-1} = \frac{2}{P}\sum_{l=1}^{P}[\mathbf{I} - \mathfrak{Re}(\mathbf{M}_l^{-1})]; \tag{52}$$

the authors go on to show that the majority of computational effort in such an approach is in the inversion of the $\mathbf{M}_l$ matrices and that the condition number of those matrices with low values of l is significantly larger than for those with higher l. Furthermore, the condition number drops off rapidly with l, approaching 1 after tens of terms. In practice, the authors have

shown that to efficiently invert all $\mathbf{M}_l$ matrices in an expansion, the majority of low condition-number matrices can be handled through an expansion about an approximate diagonal matrix, with the significant advantage that these terms can all be formed from the same intermediate matrices, reducing the computational effort markedly. The remaining terms are handled via a Newton-Shultz-Hotelling method, which is more expensive, but required only for approximately 10 matrices to achieve maximum efficiency.

The approach of Ceriotti *et al*, makes use of the fast-resumming improvements to the Chebyshev expansion for the decomposition in the high-l regime. It also uses initialization from previous matrices in the low-l regime. This work provides a useful framework for analysis of the problems facing such methods, but alone does not bring the scaling factor down from the previous state of the art. In order to achieve that, Richters and Kühne go on to improve the approach by computing only the real components of high condition number $\mathbf{M}_l^{-1}$ and computing the high-l expansion by approximating the inverse of $\mathbf{M}_l$ directly as a Chebyshev expansion[72]. By taking this approach, the scaling of the method is reduced to $O(\sqrt[3]{N})$ matrix multiplications[73].

3. Recursive methods

Another class of methods which are closely related to purification used in linear-scaling DFT techniques for insulators are the recursive operator expansions. Techniques such as these first appeared in the 1970s with the Haydock approach for tight-binding calculations, which can be performed in a linear scaling manner [74][75].

Niklasson [76] proposed taking a low order Padé approximant of the Fermi operator (42),

$$\mathbf{X}_n(\mathbf{X}_{n-1}) = \mathbf{X}_{n-1}^2\left(\mathbf{X}_{n-1}^2 + (\mathbf{I}-\mathbf{X}_{n-1})^2\right)^{-1} \tag{53}$$

and applying it recursively starting with the initial

$$\mathbf{X}_0 = \frac{1}{2}\mathbf{I} - (\mathbf{H}-\mu_0\mathbf{I})\,\beta/2^{2+N}, \tag{54}$$

where N is the number of iterates in the expansion. With this method the Fermi operator can be approximated as

$$f(\mathbf{H},\mu,\beta) = \mathbf{X}_N(\mathbf{X}_{N-1}(\cdots(\mathbf{X}_0)\cdots)). \tag{55}$$

Niklasson reports that the scheme is quadratically convergent and in practice the number of iterates can often be kept low ($N \approx 10$). Given that one inversion must be performed, or one linear equation solved per iteration which can be seeded from the the previous iterate, if using an iterative method, this technique is expected to be very quick in practice.

4. Rational expansion approaches

Goedecker was the first to introduce methods for the rational series expansion of the finite temperature density matrix [77]. The Fermi operator can be expressed as,

$$f(\mathbf{H},\mu,\beta) = \frac{1}{2k_BT}\int_{-\infty}^{\mu}\left(2\mathbf{I} + \frac{1}{2!}\left(\frac{\mathbf{H}-\mu'\mathbf{I}}{k_BT}\right)^2 + \frac{1}{4!}\left(\frac{\mathbf{H}-\mu'\mathbf{I}}{k_BT}\right)^4 + \cdots\right)^{-1} d\mu', \tag{56}$$

which still contains a very expensive and ill-conditioned inversion operation. Writing the expression in this form does however allow it to be further approximated by expressing the integrand as a partial fraction expansion truncated to order n:

$$f(\mathbf{H},\mu,\beta) = \int_{-\infty}^{\mu}\sum_{\nu=1}^{n}\frac{C_\nu}{(\mathbf{H}-\mu'\mathbf{I}) - k_BT\,(A_\nu + iB_\nu)}d\mu', \tag{57}$$

The coefficients $C_\nu = A_\nu + iB_\nu$, can be calculated, and the partial fraction decomposition is very quickly convergent [78]. The author suggests an n of 16, giving a compact expression:

$$f(\mathbf{H},\mu,\beta) = \sum_{\nu=1}^{n/4}\left[\int_{\Pi_\nu}\frac{2iB_\nu}{\mathbf{H}-z\mathbf{I}}dz + \int_{\Lambda_\nu^+}\frac{A_\nu - iB_\nu}{\mathbf{H}-z\mathbf{I}}dz + \int_{\Lambda_\nu^-}\frac{A_\nu + iB_\nu}{\mathbf{H}-z\mathbf{I}}\,dz\right], \tag{58}$$

where the z values are the complex value points along the path used to evaluate each integral by quadrature. Three paths in the complex plane, Π_ν, Λ_ν^+ and Λ_ν^- are used for quadrature. In essence, the approach works by re-expressing the occupancy formula (42) as in equation 58 which consists of contour integrals. These could be formally evaluated from their residues, but we don't have access to the poles of the function X. Thus, these integrals are computed by numerically integrating around a contour surrounding the eigenspectrum.

At every point z on the quadrature path, a Hermitian matrix, $\mathbf{H} - z\mathbf{I}$ must be inverted. Thus, rational expansion performed as in the contour integral expansion of Goedecker has low prefactor in number of inversions O(ln(M)), but results in a large number of matrix multiplications if performed with an iterative inversion algorithm) [78].

In subsequent work, Lin Lin *et al* propose an alternative scheme based on contour integration of the matrix Fermi-Dirac function, so that:

$$f(\mathbf{X}) = \mathfrak{Im}\sum_{l=1}^{P}\frac{\omega_l}{\mathbf{H} - (z_l+\mu)\mathbf{S}}, \tag{59}$$

where the complex shifts z_l and quadrature weights ω_l can be calculated from solely the chemical potential,

This manuscript was accepted by J. Chem. Phys. Click here to see the version of record.

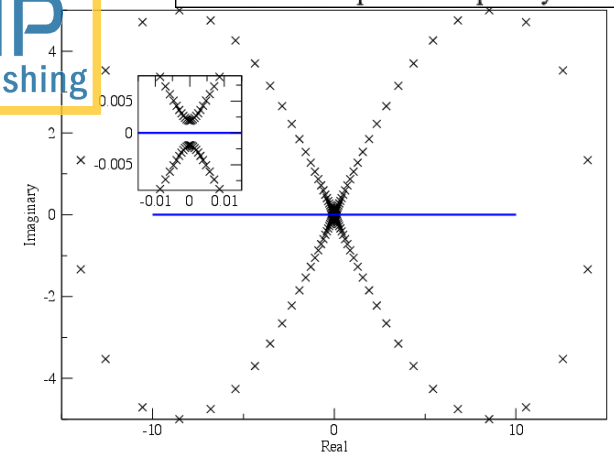

FIG. 3: The typical eigenspectrum of a Hamiltonian matrix for a metallic system is shown on the real axis of this Argand diagram (blue line). The discretized contour integral of a matrix smearing function, as used in the PEXSI method, with poles on the eigenvalues is taken around the black contour, avoiding non-analytic points ($i\pi/\beta$ and $-i\pi/\beta$) on the imaginary axis.

the Hamiltonian eigenspectrum width and the number of points on the contour integral, P. A contour based on the complex shifts suggested by the authors is shown in figure 3. Similar techniques to these are used in KKR and related multi-scattering techniques, see section V.

In performing this integral, the authors are careful to avoid the non-analytic parts of Fermi-Dirac function in the complex plane on the imaginary axis greater than π/β and less than $-\pi/\beta$, by constructing a contour which encompasses the eigenspectrum of the Hamiltonian matrix, but "necks" sufficiently at zero on the real axis to pass though the analytic window while enveloping all of the real eigenvalues. A figure showing such a contour is given in Fig. 3.

Within, the authors show the poor convergence of the Matsubara expansion (expansion into even and odd imaginary frequency components) that Goedecker also reported, though they offer a solution to this in terms of fast multipole (FMP) methods. They do however show that the integral expansion based on a contour integral requires fewer inversions than this Matsubara based approach, even with the FMP method. These methods scale as O(ln(M)) where M is the number of matrix inversions [79].

Lin-Lin & Roberto Car, *et al* have proposed a multipole expansion based on Matsubara theory[80]. Given that

$$(\mathbf{I}+e^{X})^{-1} = (\mathbf{I}-\tanh(X/2))/2, \tag{60}$$

then the Matsubara representation can be written, using the pole-expansion of tanh as

$$(\mathbf{I}+e^{X})^{-1} = \mathbf{I} - 4\mathfrak{Re}\sum_{l=1}^{\infty}(\mathbf{X}-(2l-1)\pi i\mathbf{I})^{-1}, \tag{61}$$

then combining poles into multi-poles, the overall scaling of the method can be reduced to O(ln(N_I), where N_I represents the number of (complex-valued) matrices to be inverted. While this method may be considered almost linear scaling, one could argue that the prefactor is large (when using an inversion method such as NSH).

Such an approach necessarily involves a great deal of complex valued matrix inversions (for each point on the quadrature). This would be a significant problem if an iterative inversion method like NSH was employed, as for instance in a given system, if 50 quadrature points were required for accurate results and perhaps 10 iterations of the NSH algorithm were required to converge each inversion then as many as 1000 matrix multiplications would be required to perform this algorithm. It is necessary, therefore that a more appropriate matrix inversion method be used. This issue led to the development of the selective inversion (SI) method by the same authors (see section III C). Assuming that the Hamiltonian matrix is sparse and so are the matrices $(\mathbf{H}-(z_l+\mu)\mathbf{S})$ on each of the points on the quadrature z_l, then each of these matrices can be decomposed as

$$\mathbf{LDL}^{\dagger} = \mathbf{H}-(z_l+\mu)\mathbf{S} = \mathbf{Q}_l, \tag{62}$$

where $\mathbf{L}$ is a lower triangular matrix and $\mathbf{D}$ is diagonal. and the inverse constructed as in section III C.

Together with the contour integral approach, or the pole-expansion (PEX) for calculating the density matrices, the authors have named this approach PEXSI. SI is useful because up until its development all of the contour integral / rational expansions needed a large number of expensive iterative inversions even though the best rational / contour integral expansions scale as O(ln(M)).

It is also important to note that expansion methods become computationally advantageous when the matrices under consideration are sparse. This is clear in the case of insulators due to the short-sightedness of electronic matter, however the exponential decay of the density matrix is also recovered for metals at finite electronic temperature [31].

IV. CHEMICAL POTENTIAL SEARCH

A significant issue when using an expansion approach in the canonical ensemble is to find the correct chemical potential. With standard cubic-scaling plane-wave DFT, this is not considered an issue, principally because the eigenvalues of the Hamiltonian are readily available. For this reason, it is not computationally expensive to perform a search in the eigenvalue space for the chemical potential which gives the correct electron number. This is not possible without a diagonal representation, as for each trial chemical potential, a new density matrix must be calculated, making the whole process many times more expensive.

Several methods have been proposed to improve the situation, including a finite difference representation of

the change in number of electrons with respect to chemical potential [81].

Niklasson recommends[82] using the analytic derivative of the density matrix with respect to the chemical potential

$$\frac{\partial \mathbf{K}'}{\partial \mu} = \beta \mathbf{K}'(\mathbf{I} - \mathbf{K}'), \tag{63}$$

where $\mathbf{K}'$ is the density matrix represented in an orthogonal basis and $\beta = 1/k_B T$. Then when using a Newton-Raphson optimization process, the electron number can be corrected by altering the chemical potential as:

$$\mu_m = \mu_{m-1} + [N_{\text{occ}} - \text{Tr}(\mathbf{K}')]/\text{Tr}[\beta \mathbf{K}'(\mathbf{I} - \mathbf{K}')] \tag{64}$$

at each step in the optimization.

V. KKR AND RELATED APPROACHES

The Korringa-Kohn-Rostoker method or KKR predates the Hohenberg-Kohn theorem and Kohn-Sham DFT, having been introduced in 1947 by Korringa[83] and 1954 by Kohn and Rostoker[84].

The main principle of the KKR method is that by reproducing the scattering behaviour of electrons and nuclei that the physics of the system will also be reproduced. So, the Lippmann-Schwinger integral, scattering equation is solved rather than the Schrödinger differential equation. The complete and orthogonal set of eigenvectors of the Hamiltonian matrix can be chosen as a basis to expand the Green's function. In extension to this approach, the Green's functions can instead be written in reciprocal space, and then the integrals become contour integrals over the eigenspectrum of the Green's function, where eigenvalues lie on poles[85]. Techniques such as the contour integral approaches to expanding DFT density matrices have very similar analogues in KKR and were applied in such techniques earlier on[86].

KKR approaches remain important for large scale metallic systems, provided that LDA or LSDA quality results are a valid approximation, as they often are in purely metallic systems. KKR-type methods such as the locally self consistent multiple scattering (LSMS) technique have allowed for linear scaling (order N) calculations to be performed on metallic systems within the L(S)DA formalism since the 1990s[87], with the ability to study alloys coming slightly later [88]. Recent work has seen augmented-KKR approaches which combine the benefits of KKR with DFT calculations [89] and methods for linear scaling tight binding KKR on systems of tens of thousands of atoms[90].

It is also worth noting that KKR-type techniques are often applied in a muffin-tin approach and conventionally, the local scattering environments of the individual atoms are joined together through boundary conditions between the atomic environments. These multiple-scattering techniques are particularly suitable for parallelization, as the environments are mostly self-contained and independent, except at the boundary.

VI. CONCLUSIONS AND OUTLOOK

Density Functional Theory calculations on metallic systems (i.e. systems with zero band gap) cannot be done with DFT approaches that have been developed for insulators as such approaches are not designed to cope with a continuum of energy levels at the chemical potential. Mermin's formulation of finite temperature DFT provided the theoretical basis for DFT calculations on metallic systems. A great deal of progress has been made over the last thirty years in studying metallic systems based on Mermin's DFT, starting with approaches such as density mixing and the more stable ensemble DFT. In the last decade rapid progress has been made on expansion methods, which however were introduced much earlier.

Advanced implementations of Mermin's DFT have allowed calculations with over 1000 atoms to be performed. For example Alfè *et al* have performed calculations on molten iron with over 1000 atoms which has provided new unique insights about processes taking place in the Earth's core[35], which can not be obtained by experimental means. Other examples can be found in the work of Nørskov *et al* who have studied small molecule adsorption on platinum nanoparticles of up to 1500 atoms[91] and in Skylaris and Ruiz-Serrano who have reported calculations on gold nanoparticles with up to 2000 atoms[44]. These three examples have used approaches which minimize the number of $O(N^3)$ operations, either the approach of Kresse *et al* as in VASP and GPAW, in the first two examples respectively or the approach of Skylaris *et al* in the latter example. This turns out to be key for calculations on systems up to the low thousands of atoms, at which point the cubic diagonalization step dominates and a different approach with lower computational scaling is required.

We have reviewed the methods for calculations on metallic systems with emphasis on recent developments that promise calculations with larger numbers of atoms. Fractional occupancies of states are needed when dealing with metallic systems and conventionally these are computed after diagonalization of the Hamiltonian which is impractical for large systems. One way to avoid diagonalization is to use expansion methods which construct a finite temperature density matrix via matrix expansion of the Hamiltonian, without needing to access its eigenvalues. Expansion methods have typically been slow and so unsuitable for increasing the size of system one can study or for reducing time to science. Recent developments have improved this situation.

There is considerable option, however in the particular expansion chosen to compute the density matrix. The variants fall roughly into two categories; those based on an expansion via Chebyshev polynomials and those based

on a rational or contour integral expansion. At present, on balance, it seems that the *PEXSI* approach which is based on a contour integral combined with a novel matrix inversion approach is the most computationally efficient option. Time will tell whether the Chebyshev expansion methods can be developed further to have improved scaling with system size and become competitive to the best rational expansion methods, or if stochastic methods become usable.

In the future the ultimate target in methods for metallic systems would be to have computational CPU and memory cost which increase linearly with the number of atoms as is the case for insulators where a number of linear-scaling DFT programs currently exist. This is not at all trivial as the temperature and spatial cutoff at which exponential decay of the density matrix in a metal would be sufficient to have enough matrix sparsity for linear-scaling algorithms and acceptable accuracy is not clear and would need to be explored carefully. Coming from a background of linear-scaling DFT, so far we have a metals method which works in the localized non-orthogonal generalized Wannier function framework of the ONETEP package. This approach is an intermediate step towards the development of a linear-scaling method for metals as it benefits from the framework of a linear-scaling DFT approach (e.g. sparse matrices and algorithms) but it still requires a cubic-scaling diagonalization step as it is based on the EDFT formalism. We expect further work in this direction to involve an expansion method which can exploit the sparsity in the matrices as we do in insulating linear-scaling DFT.

Based on our review of the available literature discussing the application of expansion approaches to DFT calculations, the sizes (i.e. number of atoms) of metallic systems which have been studied with these methods is quite limited, in comparison to, for instance, the sizes of insulating systems studied with the same methods. There are several reasons for this. For example, these algorithms have higher prefactors than do the algorithms involved in calculating idempotent density matrices, as used in linear-scaling DFT. Production calculations of metallic systems with methods which purport to have a reduced or even linear-scaling computational cost with system size have not yet been reported for the large numbers of atoms that one would have expected, because the crossover point at which these methods become advantageous over diagonalization has not yet been reached on currently used High Performance Computing facilities.

Also, algorithms such as density mixing do not work as well for large systems, as has been seen in practical observations of the scaling of SCF iterations with the number of atoms in conventional DFT for metallic systems. If this is indeed a factor, then it is possible that an alternative, such as EDFT will be required.

Even with these caveats, the class of operator expansions applied to density matrix DFT methods appears to be the strongest contender for reducing the scaling of accurate DFT calculations on metallic systems in the future. If such methods are further developed, and actually start to be routinely applied as computational power increases, we expect that they will have a great impact on metallic nanostructure and bulk metal surfaces engineering for industrial applications in optics, magnetics, catalysis and other areas in which the unique properties of metallic systems can be exploited.

Acknowledgments

Jolyon Aarons would like to thank Johnson Matthey and the Engineering and Physical Sciences Research Council (EPSRC) for an industrial CASE PhD studentship.

[1] Riccardo Ferrando, Julius Jellinek, and Roy L Johnston. Nanoalloys: from theory to applications of alloy clusters and nanoparticles. *Chemical reviews*, 108(3):845–910, 2008.

[2] Prashant K. Jain, Wenyu Huang, and Mostafa A. El-Sayed. On the universal scaling behavior of the distance decay of plasmon coupling in metal nanoparticle pairs: A plasmon ruler equation. *Nano Letters*, 7(7):2080–2088, 2007.

[3] Nathaniel L. Rosi and Chad A. Mirkin. Nanostructures in biodiagnostics. *Chemical Reviews*, 105(4):1547–1562, 2005. PMID: 15826019.

[4] Cono Di Paola, Roberto D'Agosta, and Francesca Baletto. Geometrical effects on the magnetic properties of nanoparticles. *Nano letters*, 16(4):2885–2889, 2016.

[5] L. V. Lutsev, A. I. Stognij, and N. N. Novitskii. Giant magnetoresistance in semiconductor/granular film heterostructures with cobalt nanoparticles. *Phys. Rev. B*, 80:184423, Nov 2009.

[6] Wendong Wang, Fengwu Zhu, Jun Weng, Jimei Xiao, and Wuyan Lai. Nanoparticle morphology in a granular cu-co alloy with giant magnetoresistance. *Applied Physics Letters*, 72(9):1118–1120, 1998.

[7] Francesca Baletto and Riccardo Ferrando. Structural properties of nanoclusters: Energetic, thermodynamic, and kinetic effects. *Reviews of modern physics*, 77(1):371, 2005.

[8] Yuguang Ma and Perla B Balbuena. Pt surface segregation in bimetallic pt 3 m alloys: a density functional theory study. *Surface Science*, 602(1):107–113, 2008.

[9] Gustavo E Ramirez-Caballero and Perla B Balbuena. Surface segregation of core atoms in core–shell structures. *Chemical Physics Letters*, 456(1):64–67, 2008.

[10] Ingmar Swart, Frank MF De Groot, Bert M Weckhuysen, Philipp Gruene, Gerard Meijer, and Andre Fielicke. H2 adsorption on 3d transition metal clusters: A combined infrared spectroscopy and density functional study. *The Journal of Physical Chemistry A*, 112(6):1139–1149,

2008.

[11] F. Tournus, K. Sato, T. Epicier, T. J. Konno, and V. Dupuis. Multi-$l1_0$ domain copt and fept nanoparticles revealed by electron microscopy. *Phys. Rev. Lett.*, 110:055501, Jan 2013.

[12] Jens Kehlet Nørskov, Jan Rossmeisl, Ashildur Logadottir, LRKJ Lindqvist, John R Kitchin, Thomas Bligaard, and Hannes Jonsson. Origin of the overpotential for oxygen reduction at a fuel-cell cathode. *The Journal of Physical Chemistry B*, 108(46):17886–17892, 2004.

[13] JR Kitchin, Jens Kehlet Nørskov, MA Barteau, and JG Chen. Modification of the surface electronic and chemical properties of pt (111) by subsurface 3d transition metals. *Journal of Chemical Physics*, 120(21):10240–10246, 2004.

[14] Jeff Greeley and Manos Mavrikakis. Alloy catalysts designed from first principles. *Nature materials*, 3(11):810–815, 2004.

[15] Yuguang Ma and Perla B Balbuena. Surface properties and dissolution trends of pt3m alloys in the presence of adsorbates. *The Journal of Physical Chemistry C*, 112(37):14520–14528, 2008.

[16] Anders Hellman, Andrea Resta, NM Martin, Johan Gustafson, Adriana Trinchero, P-A Carlsson, Olivier Balmes, Roberto Felici, Richard van Rijn, JWM Frenken, et al. The active phase of palladium during methane oxidation. *The journal of physical chemistry letters*, 3(6):678–682, 2012.

[17] Pierre Hohenberg and Walter Kohn. Inhomogeneous electron gas. *Physical review*, 136(3B):B864, 1964.

[18] Tomasz A Wesolowski and Yan Alexander Wang. *Recent Progress in Orbital-free Density Functional Theory*. World Scientific, 2013.

[19] Mohan Chen, Linda Hung, Chen Huang, Junchao Xia, and Emily A. Carter. The melting point of lithium: an orbital-free first-principles molecular dynamics study. *Molecular Physics*, 111(22-23):3448–3456, 2013.

[20] Hendrik J. Monkhorst and James D. Pack. Special points for brillouin-zone integrations. *Phys. Rev. B*, 13:5188–5192, Jun 1976.

[21] Peter E. Blöchl, O. Jepsen, and O. K. Andersen. Improved tetrahedron method for brillouin-zone integrations. *Phys. Rev. B*, 49:16223–16233, Jun 1994.

[22] M J Gillan. Calculation of the vacancy formation energy in aluminium. *Journal of Physics: Condensed Matter*, 1(4):689, 1989.

[23] W Kohn. Density functional and density matrix method scaling linearly with the number of atoms. *PHYSICAL REVIEW LETTERS*, 76(17):3168–3171, APR 22 1996.

[24] D. R. Bowler, R. Choudhury, M. J. Gillan, and T. Miyazaki. Recent progress with large-scale ab initio calculations: the CONQUEST code. *physica status solidi (b)*, 243(5):989–1000, 2006.

[25] Chris-Kriton Skylaris, Arash A. Mostofi, Peter D. Haynes, Oswaldo Diéguez, and Mike C. Payne. Nonorthogonal generalized wannier function pseudopotential plane-wave method. *Phys. Rev. B*, 66:035119, Jul 2002.

[26] Joost VandeVondele, Urban Borštnik, and Jürg Hutter. Linear scaling self-consistent field calculations with millions of atoms in the condensed phase. *Journal of Chemical Theory and Computation*, 8(10):3565–3573, 2012. PMID: 26593003.

[27] E. Artacho, D. Sánchez-Portal, P. Ordejón, A. García, and J. M. Soler. Linear-scaling ab-initio calculations for large and complex systems. *physica status solidi (b)*, 215(1):809–817, 1999.

[28] Luigi Genovese, Alexey Neelov, Stefan Goedecker, Thierry Deutsch, Seyed Alireza Ghasemi, Alexander Willand, Damien Caliste, Oded Zilberberg, Mark Rayson, Anders Bergman, and Reinhold Schneider. Daubechies wavelets as a basis set for density functional pseudopotential calculations. *The Journal of Chemical Physics*, 129(1), 2008.

[29] Stefan Goedecker. Linear scaling electronic structure methods. *Rev. Mod. Phys.*, 71:1085–1123, Jul 1999.

[30] D R Bowler and T Miyazaki. $\mathcal{O}(n)$ methods in electronic structure calculations. *Reports on Progress in Physics*, 75(3):036503, 2012.

[31] S Goedecker. Decay properties of the finite-temperature density matrix in metals. *Physical Review B*, 58(7):3501, 1998.

[32] Peter Pulay. Convergence acceleration of iterative sequences. the case of scf iteration. *Chemical Physics Letters*, 73(2):393–398, 1980.

[33] Peter Pulay. Improved scf convergence acceleration. *Journal of Computational Chemistry*, 3(4):556–560, 1982.

[34] Georg Kresse and Jürgen Furthmüller. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Physical Review B*, 54(16):11169, 1996.

[35] D Alfè, MJ Gillan, and GD Price. Ab initio chemical potentials of solid and liquid solutions and the chemistry of the earth's core. *The Journal of chemical physics*, 116(16):7127–7136, 2002.

[36] Lianheng Tong. Metal CONQUEST, HECToR dCSE report. `http://www.hector.ac.uk/cse/distributedcse/reports/conquest/conquest.pdf`, March 2011. Online; accessed 28-Sep-2016.

[37] GP Kerker. Efficient iteration scheme for self-consistent pseudopotential calculations. *Physical Review B*, 23(6):3082, 1981.

[38] M. Methfessel and A. T. Paxton. High-precision sampling for brillouin-zone integration in metals. *Phys. Rev. B*, 40:3616–3621, Aug 1989.

[39] Nicola Marzari, David Vanderbilt, Alessandro De Vita, and M. C. Payne. Thermal contraction and disordering of the al(110) surface. *Phys. Rev. Lett.*, 82:3296–3299, Apr 1999.

[40] WanZhen Liang, Chandra Saravanan, Yihan Shao, Roi Baer, Alexis T Bell, and Martin Head-Gordon. Improved Fermi operator expansion methods for fast electronic structure calculations. *The Journal of chemical physics*, 119(8):4117–4125, 2003.

[41] Nicola Marzari, David Vanderbilt, and Mike C Payne. Ensemble density-functional theory for ab initio molecular dynamics of metals and finite-temperature insulators. *Physical review letters*, 79(7):1337, 1997.

[42] Christoph Freysoldt, Sixten Boeck, and Jörg Neugebauer. Direct minimization technique for metals in density functional theory. *Phys. Rev. B*, 79:241103, Jun 2009.

[43] Chris-Kriton Skylaris, Peter D. Haynes, Arash A. Mostofi, and Mike C. Payne. Introducing ONETEP: Linear-scaling density functional simulations on parallel computers. *The Journal of Chemical Physics*, 122(8), 2005.

[44] Álvaro Ruiz-Serrano and Chris-Kriton Skylaris. A varia-

tional method for density functional theory calculations on metallic systems with thousands of atoms. *The Journal of chemical physics*, 139(5):054107, 2013.

[45] Branislav Jansík, Stinne Høst, Poul Jørgensen, Jeppe Olsen, and Trygve Helgaker. Linear-scaling symmetric square-root decomposition of the overlap matrix. *The Journal of chemical physics*, 126(12):124104, 2007.

[46] RW Nunes and David Vanderbilt. Generalization of the density-matrix method to a nonorthogonal basis. *Physical Review B*, 50(23):17611, 1994.

[47] Xiaoye S Li and James W Demmel. Superlu_dist: A scalable distributed-memory sparse direct solver for unsymmetric linear systems. *ACM Transactions on Mathematical Software (TOMS)*, 29(2):110–140, 2003.

[48] William H Press, Saul A Teukolsky, William T Vetterling, and Brian P Flannery. *Numerical Recipes: The art of scientific computing (Cambridge*, volume 683. Cambridge Univ. Press, 1992.

[49] Victor Pan and Robert Schreiber. An improved newton iteration for the generalized inverse of a matrix, with applications. *SIAM Journal on Scientific and Statistical Computing*, 12(5):1109–1130, 1991.

[50] T Ozaki. Efficient recursion method for inverting an overlap matrix. *Physical Review B*, 64(19):195110, 2001.

[51] Adam HR Palser and David E Manolopoulos. Canonical purification of the density matrix in electronic-structure theory. *Physical Review B*, 58(19):12704, 1998.

[52] Peter D Haynes, Chris-Kriton Skylaris, Arash A Mostofi, and Mike C Payne. Density kernel optimization in the ONETEP code. *Journal of Physics: Condensed Matter*, 20(29):294207, 2008.

[53] Lin Lin, Chao Yang, Juan C Meza, Jianfeng Lu, Lexing Ying, et al. Selinv—an algorithm for selected inversion of a sparse symmetric matrix. *ACM Transactions on Mathematical Software (TOMS)*, 37(4):40, 2011.

[54] S. Li and E. Darve. Extension and optimization of the FIND algorithm: Computing Green's and less-than Green's functions. *Journal of Computational Physics*, 231(4):1121 – 1139, 2012.

[55] François Henry Rouet. *Thesis: Partial computation of the inverse of a large sparse matrix - application to astrophysics*. Institut national polytechnique de Toulouse, 2009.

[56] Patrick R Amestoy, Iain S Duff, Jean-Yves L'Excellent, and Jacko Koster. A fully asynchronous multifrontal solver using distributed dynamic scheduling. *SIAM Journal on Matrix Analysis and Applications*, 23(1):15–41, 2001.

[57] Nicholas IM Gould, Jennifer A Scott, and Yifan Hu. A numerical evaluation of sparse direct solvers for the solution of large sparse symmetric linear systems of equations. *ACM Transactions on Mathematical Software (TOMS)*, 33(2):10, 2007.

[58] George Karypis and Vipin Kumar. A parallel algorithm for multilevel graph partitioning and sparse matrix ordering. *Journal of Parallel and Distributed Computing*, 48(1):71–95, 1998.

[59] George Karypis, Kirk Schloegel, and Vipin Kumar. Parmetis: Parallel graph partitioning and sparse matrix ordering library. *Dept. of Computer Science, University of Minnesota*, 1997.

[60] Cédric Chevalier and François Pellegrini. Pt-scotch: A tool for efficient parallel graph ordering. *Parallel computing*, 34(6):318–331, 2008.

[61] Stefan Goedecker and L Colombo. Efficient linear scaling algorithm for tight-binding molecular dynamics. *Physical review letters*, 73(1):122, 1994.

[62] MPAT Methfessel and AT Paxton. High-precision sampling for brillouin-zone integration in metals. *Physical Review B*, 40(6):3616, 1989.

[63] Anders M. N. Niklasson. Iterative refinement method for the approximate factorization of a matrix inverse. *Phys. Rev. B*, 70:193102, Nov 2004.

[64] Emanuel H. Rubensson, Nicolas Bock, Erik Holmström, and Anders M. N. Niklasson. Recursive inverse factorization. *The Journal of Chemical Physics*, 128(10), 2008.

[65] Stefan Goedecker and M Teter. Tight-binding electronic-structure calculations and tight-binding molecular dynamics with localized orbitals. *Physical Review B*, 51(15):9455, 1995.

[66] Roi Baer and Martin Head-Gordon. Chebyshev expansion methods for electronic structure calculations on large molecular systems. *The Journal of chemical physics*, 107(23):10003–10013, 1997.

[67] Michael S Paterson and Larry J Stockmeyer. On the number of nonscalar multiplications necessary to evaluate polynomials. *SIAM Journal on Computing*, 2(1):60–66, 1973.

[68] Florian R. Krajewski and Michele Parrinello. Stochastic linear scaling for metals and nonmetals. *Phys. Rev. B*, 71:233105, Jun 2005.

[69] Michele Ceriotti, Thomas D Kühne, and Michele Parrinello. An efficient and accurate decomposition of the Fermi operator. *The Journal of chemical physics*, 129(2):024707, 2008.

[70] Florian R. Krajewski and Michele Parrinello. Linear scaling electronic structure monte carlo method for metals. *Phys. Rev. B*, 75:235108, Jun 2007.

[71] Florian R. Krajewski and Michele Parrinello. Linear scaling for quasi-one-dimensional systems. *Phys. Rev. B*, 74:125107, Sep 2006.

[72] Dorothee Richters and Thomas D Kühne. Self-consistent field theory based molecular dynamics with linear system-size scaling. *The Journal of chemical physics*, 140(13):134109, 2014.

[73] Michele Ceriotti, Thomas D Kühne, and Michele Parrinello. A hybrid approach to Fermi operator expansion. *arXiv preprint arXiv:0809.2232*, 2008.

[74] R Haydock, V Heine, and M J Kelly. Electronic structure based on the local atomic environment for tight-binding bands. ii. *Journal of Physics C: Solid State Physics*, 8(16):2591, 1975.

[75] R Haydock, V Heine, and M J Kelly. Electronic structure based on the local atomic environment for tight-binding bands. *Journal of Physics C: Solid State Physics*, 5(20):2845, 1972.

[76] Anders M. N. Niklasson. Implicit purification for temperature-dependent density matrices. *Phys. Rev. B*, 68:233104, Dec 2003.

[77] S Goedecker. Integral representation of the Fermi distribution and its applications in electronic-structure calculations. *Physical Review B*, 48(23):17573, 1993.

[78] Stefan Goedecker. Low complexity algorithms for electronic structure calculations. *Journal of Computational Physics*, 118(2):261–268, 1995.

[79] Lin Lin, Jianfeng Lu, Lexing Ying, and E Weinan. Pole-based approximation of the Fermi-Dirac function. *Chinese Annals of Mathematics, Series B*, 30(6):729–742,

2009.

[80] Lin Lin, Jianfeng Lu, Roberto Car, and E Weinan. Multipole representation of the Fermi operator with application to the electronic structure analysis of metallic systems. *Physical Review B*, 79(11):115133, 2009.

[81] Lin Lin, Mohan Chen, Chao Yang, and Lixin He. Accelerating atomic orbital-based electronic structure calculation via pole expansion and selected inversion. *Journal of Physics: Condensed Matter*, 25(29):295501, 2013.

[82] Anders M. N. Niklasson, Peter Steneteg, and Nicolas Bock. Extended lagrangian free energy molecular dynamics. *The Journal of Chemical Physics*, 135(16):164111, 2011.

[83] J Korringa. On the calculation of the energy of a Bloch wave in a metal. *Physica*, 13(6-7):392–400, 1947.

[84] W Kohn and N Rostoker. Solution of the schrödinger equation in periodic lattices with an application to metallic lithium. *Physical Review*, 94(5):1111, 1954.

[85] Duane D Johnson, FJ Pinski, and GM Stocks. Fast method for calculating the self-consistent electronic structure of random alloys. *Physical Review B*, 30(10):5508, 1984.

[86] FJ Pinski and GM Stocks. Fast method for calculating the self-consistent electronic structure of random alloys. ii. optimal use of the complex plane. *Physical Review B*, 32(6):4204, 1985.

[87] Yang Wang, G. M. Stocks, W. A. Shelton, D. M. C. Nicholson, Z. Szotek, and W. M. Temmerman. Order-N multiple scattering approach to electronic structure calculations. *Phys. Rev. Lett.*, 75:2867–2870, Oct 1995.

[88] I. A. Abrikosov, A. M. N. Niklasson, S. I. Simak, B. Johansson, A. V. Ruban, and H. L. Skriver. Order- N Green's function technique for local environment effects in alloys. *Phys. Rev. Lett.*, 76:4203–4206, May 1996.

[89] Aftab Alam, Suffian N Khan, Andrei V Smirnov, DM Nicholson, and Duane D Johnson. Green's function multiple-scattering theory with a truncated basis set: An augmented-kkr formalism. *Physical Review B*, 90(20):205102, 2014.

[90] Rudolf Zeller. Towards a linear-scaling algorithm for electronic structure calculations with the tight-binding korringa–kohn–rostoker green function method. *Journal of Physics: Condensed Matter*, 20(29):294215, 2008.

[91] Lin Li, Ask H Larsen, Nichols A Romero, Vitali A Morozov, Christian Glinsvad, Frank Abild-Pedersen, Jeff Greeley, Karsten W Jacobsen, and Jens K Nørskov. Investigation of catalytic finite-size-effects of platinum metal clusters. *The journal of physical chemistry letters*, 4(1):222–226, 2012.