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

Electronic annealing Fermi Operator Expansion for DFT calculations on metallic systems

Jolyon Aarons and Chris-Kriton Skylaris*
Department of Chemistry, University of Southampton, Highfield, Southampton SO17 1BJ, UK
(Dated: January 25, 2018)

Density Functional Theory (DFT) calculations with computational effort which increases linearly with the number of atoms (linear-scaling DFT) have been successfully developed for insulators, taking advantage of the exponential decay of the one-particle density matrix. For metallic systems, the density matrix is also expected to decay exponentially at finite electronic temperature and linear-scaling DFT methods should be possible by taking advantage of this decay. Here we present a method for DFT calculations at finite electronic temperature for metallic systems which is effectively linear-scaling ($O(N)$). Our method generates the elements of the one-particle density matrix and also finds the required chemical potential and electronic entropy using polynomial expansions. A fixed expansion length is always employed to generate the density matrix, without any loss in accuracy by the application of a high electronic temperature followed by successive steps of temperature reduction until the desired (low) temperature density matrix is obtained. We have implemented this method in the ONETEP linear-scaling (for insulators) DFT code which employs local orbitals that are optimised in situ. By making use of the sparse matrix machinery of ONETEP, our method exploits the sparsity of Hamiltonian and density matrices to perform calculations on metallic systems with computational cost that increases asymptotically linearly with the number of atoms. We demonstrate the linear-scaling computational cost of our method with calculation times on Palladium nanoparticles with up to ∼13,000 atoms.

PACS numbers: Valid PACS appear here

I. INTRODUCTION

Density Functional Theory (DFT) calculations are now routinely applied to all sorts of materials as they offer a good combination of computational efficiency and accuracy. Most commonly the Kohn-Sham variant of DFT is used for applications and a lot of effort has been expended in developing increasingly accurate exchange-correlation functionals for such calculations. These calculations are performed for insulators (i.e. materials with a finite band gap) but DFT is also used for calculations on metallic systems (zero band gap) following the extension to metals by Mermin [1] , which calculates the electronic free energy rather than the energy and utilizes fractional occupancies for the molecular orbitals. DFT calculations with computational effort which increases linearly with the number of atoms (linear-scaling DFT) have been successfully developed for insulators[2–6], taking advantage of the exponential decay of the one-particle density matrix - as formulated in the nearsightedness of electronic matter principle of Walter Kohn [7]. For metallic systems, the density matrix is also expected to decay exponentially at finite electronic temperature[8] and linear-scaling DFT methods should be possible by taking advantage of this decay[9].

To compute a density matrix in DFT calculations, an occupancy function is applied to the Hamiltonian matrix. In DFT calculations for metallic systems, the Mermin finite electronic temperature formulation of Kohn-Sham DFT leads to the density matrix being obtained by diagonalizing the Hamiltonian matrix, applying a sigmoidal function, such as the Fermi-Dirac function to the energy eigenvalues and then using the Hamiltonian eigenvectors to transform these occupancy eigenvalues back into the original space of the Hamiltonian. This direct approach is, however, dependent on a cubically scaling diagonalization, so it becomes rapidly intractable as the number of basis states increases.

In the case where the Hamiltonian matrix is very large and sparse, many operations, including matrix products and even inversions may be performed in a much reduced computational complexity [3]. In such cases, it could be beneficial to avoid the eigendecomposition completely by forming a matrix function analogue of the scalar occupation function which constructs a density matrix from a Hamiltonian matrix. The sigmoidal function is likely to be non-linear, so an approximation to it can be made based on linear operations (matrix products, for instance) which is accurate within a pre-defined domain of Hamiltonian eigenvalues. Such methods are often known as Fermi operator expansions (FOE) [10], when the Fermi-Dirac function is used as the occupancy function, though they are generalizable to any occupancy smearing function[11]. The functional form of the expansion is also flexible, with options including Taylor expansions, among various more efficient alternatives. These improved alternatives [10–16], have been developed with the aim that the pre-factors involved in computing the resultant matrices are lower. The pre-factors of these

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

methods are, however, still larger than for eigendecomposition, meaning that these methods are unattractive for systems with fewer than thousands of electronic states (bands or molecular orbitals). It is also worth mentioning that FOE based techniques have also been applied successfully to insulators, such as in the recursive TC2 method of Niklasson[17]. The combined recursive and divide and conquer method of Ozaki has also been applied to insulators and tested on metallic systems [18].

We must point out, however, that for performing large-scale calculations of metallic systems, there are many options. These include FOE approaches as well as KKR-style approaches and orbital-free DFT. We have presented a review of such methods as well as current techniques for energy minimization of metals and FOE techniques in our perspective article [19] . Orbital-free DFT has been applied in specific cases such as to calculate the physical properties of liquid lithium [20]. KKR / Multi-scattering approaches are also competitive with DFT in specific cases, such as bulk metals with defects and slabs. These methods used to require comparatively homogenous systems, however, KKR has since been applied to layered slabs [21] and multicomponent systems, such as in the combined real/reciprocal space approach of Johnson *et al* [22]. Research is ongoing in both OF-DFT[23] and KKR.

In this work, we present a new method for DFT calculations on metallic systems where a fixed expansion length of the FOE is always employed, without any loss in accuracy. This is achieved by the application of a high electronic temperature followed by successive steps of temperature halving until the desired (low) temperature density matrix is obtained via a quenching approach. We call this method the Annealing and QUenching Algorithm FOE or AQuA-FOE and we have implemented it in the ONETEP [3] linear-scaling (for insulators) DFT code which employs local orbitals that are optimised in situ (Non-orthogonal Generalised Wannier Functions, NGWFs) so that it provides large basis set accuracy in the calculations. Making use of the sparse matrix machinery of ONETEP [24–26], our method exploits the sparsity of Hamiltonian and density matrices (which is expected to decay exponentially for metallic systems at high temperatures [8, 27]) to perform calculations on metallic systems with reduced computational cost. Our method is expected to be linear-scaling with the number of atoms, for large enough systems, and we attempt to reach this limit in this work, with timings for self-consistent iterations on metallic nanoparticles with up to ˜13,000 atoms. An integral part of our method is a sparse matrix algorithm for finding the chemical potential and for calculating the electronic entropy, as these are also essential to finite temperature DFT. ONETEP is a parallel code and continuous effort has been put into improving its parallel performance, as described in the paper by Wilkinson *et al* [26], where the MPI/OpenMP parallelism of the code is presented. The method presented in this work is fully MPI/OpenMP parallel as it utilizes the parallel libraries developed in ONETEP over many years.

In sections II-V we describe the general theory behind our new method without yet introducing the complexities associated with non-orthogonal bases and sparse matrices. These are introduced in sections VI and VII. In section VII we provide numerical validation by comparing calculations against the standard diagonalisation-based EDFT approach of ONETEP on metallic nanoparticles and then in section VIII explore the timings of AQUA-FOE on Pd nanoparticles ranging from ˜2,400 to ˜13,000 atoms. We finish with some conclusions and thoughts for future applications of this method.

II. KOHN-SHAM DFT IN THE MERMIN FINITE TEMPERATURE FORMULATION

For calculations of metallic systems, we must work with electrons with fractional occupancy distribution. This can be achieved by using the finite temperature Kohn-Sham equations [28] inspired by the Mermin formulation of DFT[1]. In this prescription, we minimize the Helmholtz free energy of the interacting electronic system, which is expressed as:

$$A[T,\{\varepsilon_i\},\{\psi_i\}] = \sum_i f_i \langle\psi_i|\hat{T}|\psi_i\rangle + \int v_{\text{ext}}(\mathbf{r})n(\mathbf{r})d\mathbf{r} + E_H[n] + E_{xc}[n] - TS[\{f_i\}], \tag{1}$$

this consists of the kinetic energy of the non-interacting electrons, expressions for the external potential energy and Hartree energy of the electrons and the unknown exchange-correlation energy expression. The entropic contribution to the electronic free energy $-TS[\{f_i\}]$ is included.

To calculate the occupancies of the electronic states with finite temperature, the Fermi-Dirac (F-D) distribution can be used

$$f(\epsilon,\mu,\beta) = \frac{1}{1+e^{(\epsilon-\mu)\beta}} \ . \tag{2}$$

Comparing this with the logistic function;

$$l(x) = \frac{1}{1+e^{x}}, \tag{3}$$

it can be seen that the F-D function can be written in terms of the logistic function where $x \to (\epsilon-\mu)\beta$. This is useful because the logistic function can, in turn, be written in terms of the hyperbolic tangent. This representation allows the use of trigonometric identities to simplify computations, as we will see later.

$$l(x) = \frac{1}{2}\left(1+\tanh\left(\frac{x}{2}\right)\right). \tag{4}$$

The finite temperature formulation of Kohn-Sham density functional theory results in occupancies of states

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

which follow the F-D distribution. The energies of the orbitals are the eigenvalues of the Hamiltonian matrix which is often non-diagonal. Their occupancies, which follow the F-D distribution are the eigenvalues of the density matrix, which is also non-diagonal and it is essentially the matrix F-D function of the Hamiltonian.

If we have a Hamiltonian $\mathbf{H}$, then its eigenvalue expansion is:

$$\mathbf{H} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{\dagger}, \tag{5}$$

where $\mathbf{Q}$ is a unitary matrix of eigenvectors, and $\mathbf{\Lambda}$ is a diagonal matrix of eigenvalues. If we know the largest absolute eigenvalue ($|\epsilon|_{\max} = \alpha$), we can say that the eigenspectrum of $\mathbf{H}$ lies in the interval $[-\alpha, \alpha]$. As the occupancy function will be applied to all eigenvalues, any approximation to it must be accurate to some tolerance within this interval.

In this work, we show that in order to compute efficiently an FOE-type expansion, the range of energy eigenvalues can be scaled to reduce its spectral radius and then the resulting density matrix can be quenched to have the occupancy eigenvalues corresponding to the original energy eigenvalues. This is effectively an annealing and quenching procedure on the electronic temperature of the system and is done using solely matrix multiplication and without a need for any matrix diagonalisation or inversion. So this approach could result in linear-scaling computational effort if the matrices involved have sufficient sparsity and sparse matrix storage and multiplication algorithms are employed.

III. ELECTRONIC QUENCHING

Whichever FOE method is used to generate density matrices from Hamiltonian matrices, the expansion is valid only within a given energy eigenvalue interval. For accurate results, a large number of terms in the expansion must be used, such that the expansion is valid over the full range of energy eigenvalues of the system. It has been shown recently that far fewer terms are necessary, however, at high temperature [29].

We propose that rather than increasing the number of terms for systems of increasing eigenvalue spectrum width (which is exacerbated by low electronic temperature), a fixed-length FOE is used which is valid on a predefined interval, but the electronic temperature is increased by a sufficient multiples of the target temperature so that the full spectrum of the hot-electron Hamiltonian lies within the interval of validity of the FOE. Then, the low (target) temperature density matrix can be recovered by quenching steps, as we will show in this section.

The Hamiltonian matrix is firstly shifted so that the trial chemical potential is at zero and scaled into units of smearing widths,

$$\mathbf{H}' = (\mathbf{H} - \mu\mathbf{I})\beta \ . \tag{6}$$

In multiplying the Hamiltonian matrix by $\beta = 1/k_B T$, this matrix is unit-less, but we refer to it as the "scaled and shifted Hamiltonian matrix" in this section. The scaled and shifted Hamiltonian matrix is then annealed by dividing by a sufficiently large power of 2:

$$\mathbf{H}_{\text{hot}} = \mathbf{H}'/2^n, \tag{7}$$

so that its spectrum lies on the desired interval as the implicit temperature of this "hot" Hamiltonian goes from $\beta = 1/k_B T$ to $\beta_{\text{hot}} = 1/(2^n k_B T)$. The exponent, n can be determined as

$$n = \text{ceiling}\left(\log_2(\rho_{\mathbf{H}'}/c)\right), \tag{8}$$

where $\rho_{\mathbf{H}'}$ is the spectral radius (largest absolute eigenvalue) of $\mathbf{H}'$ and c is the domain of energy eigenvalues within which the FOE is valid. $\rho_{\mathbf{H}'}$ can be determined either by using the Gershgorin circle theorem or using a power series iteration, as we do in this work. The chosen FOE is then applied to $\mathbf{H}_{\text{hot}}$ to obtain $\mathbf{K}_{\text{hot}}$. Then $\mathbf{K}_{\text{hot}}$ is annealed n times to obtain the density matrix corresponding to the low (target) temperature Hamiltonian matrix at β.

Following equation 4, the density matrix is firstly transformed in such a way that it has eigenvalues of the hyperbolic tangent function:

$$\mathbf{R}_{\text{hot}} = 2\mathbf{K}_{\text{hot}} - \mathbf{I} \ , \tag{9}$$

or in other words, the range of the eigenvalues of the density matrix is scaled from [0,1] to [-1,1]. To this transformed density matrix, the matrix analogue of the hyperbolic double angle formula can be applied, to obtain a matrix with half the electronic temperature of the $\mathbf{R_{old}}$:

$$\mathbf{R}_{\text{new}} = \frac{2\mathbf{R}_{\text{old}}}{(\mathbf{I} + \mathbf{R}^2_{\text{old}})}. \tag{10}$$

This formula is first applied to $\mathbf{R}_{\text{old}} = \mathbf{R}_{\text{hot}}$ and repeated n times to give the desired density matrix. With each iteration of the above formula, the temperature of the density matrix is halved. The way this works is by noting that in the argument to the exponential in the Fermi-Dirac function: $(\varepsilon - \mu)\beta$, the energies are effectively divided by the temperature in energy units. By implicitly multiplying $\mathbf{H}'$ by 2 as happens with the application of the hyperbolic double angle formula, the temperature is effectively halved with each application. We should note that the high temperature is simply a device for reducing the number of terms in the FOE expansion but the actual temperature which is used in the DFT is the "low" target temperature. Equation 10 is exactly the same as the scalar double angle formula given by

$$\tanh(2x) = \frac{2\tanh(x)}{1 + \tanh^2(x)}, \tag{11}$$

where we have already calculated the value at higher temperature, i.e. $\tanh(x)$.

The target temperature density matrix can eventually be recovered as

$$\mathbf{K} = (\mathbf{R} + \mathbf{I})/2. \tag{12}$$

It is worth noting that to apply equation 10 requires an inversion or to compute the solution to a system of linear equations. To avoid this, we take a Chebyshev expansion of equation 10 which requires 37 terms to reach machine precision, although an accuracy of 10^{-9} can be obtained with only 25 terms and this can be evaluated with as few as 12 matrix multiplications using the divide and conquer approach of Head-Gordon [11]. We should note here that the Chebyshev expansion of the hyperbolic double angle formula that we discuss here is different from the Chebyshev expansion of the Fermi-Dirac distribution which can be used to compute $\mathbf{K}_{\text{hot}}$. It is also interesting to note that here we use a recursion in the temperature scaling of the FOE, which is itself computed using a Chebyshev expansion, while Niklasson *et al* proposed a recursion method directly for the FOE[16, 30] , which however requires many explicit matrix inversions, which are absent from our approach.

For a given chemical potential, the scaling in terms of number of matrix products to apply this algorithm is given by:

$$N_{\text{MP}} = N_{\text{FOE}}(c) + nN_{\text{HI}}, \tag{13}$$

where N_{FOE} is the number of matrix products required to compute the FOE, N_{HI} is the number of matrix products required to compute an application of the hyperbolic double angle formula as a Chebyshev expansion. The number of matrix products required to apply the FOE depends on c, but in practice this is constant and we always use a value of 15 smearing widths (β_{hot}) as there is no need to change this aside from optimization of the total number of matrix multiplications. This is a compromise between the FOE length and number of hyperbolic double angle formula evaluations. The hyperbolic double angle formula is always evaluated with a fixed number of terms in the Chebyshev expansion and hence, a fixed number of matrix multiplications. So, the only variable in the total number of matrix products is the spectral width of the scaled and shifted Hamiltonian matrix, $\rho_{\mathbf{H}'}$.

As the number of atoms increases, if the material is homogeneous, we do not expect the spectral width of the Hamiltonian matrix to increase as it asymptotically reaches the bulk value. However, as an extreme upper bound, if the spectral width were to increase linearly with system size, then the number of matrix multiplications required to apply this algorithm would only increase logarithmically, according to equation 8. In practice, this increase will be somewhere between zero and logarithmic. So, with a non-increasing, or logarithmically increasing number of matrix multiplications for an increase in system size (number of atoms), if the matrix multiplication can be made to be linear-scaling with dimension, as is the case for sparse matrices in the ONETEP linear-scaling DFT program for sufficient sparsity, then the annealing algorithm should have linear-scaling computational cost with system size.

IV. COMBINING HAMILTONIAN ANNEALING AND QUENCHING WITH CHEBYSHEV FERMI OPERATOR EXPANSION

The Chebyshev FOE, as described originally by Goedecker and Teter [31] is achieved by taking a Chebyshev expansion of the 1D Fermi-Dirac function:

$$\{a_i\} = \text{DCT}\left(\frac{1}{1 + e^{\cos(x_i)}}\right), \tag{14}$$

where $x_i = 2(((\epsilon_i - \mu)\beta) - \epsilon_{\text{min}})/(\epsilon_{\text{max}} - \epsilon_{\text{min}}) - 1$ so that the range of equispaced grid points e_i covers at least the interval of the negative to positive spectral radius of the Hamiltonian matrix (ϵ_{min},ϵ_{max}). The interval is scaled and shifted to cover the useful interpolative range of Chebyshev polynomials $[-1, 1]$. DCT() refers to the Discrete Cosine Transform operation. $\{a_i\}$ are the Chebyshev expansion coefficients. The density matrix is formed by taking Chebyshev polynomials of the Hamiltonian matrix and summing using the weights, a_i:

$$\mathbf{K}(\mathbf{H}') = \sum_{i=0}^{P} a_i \mathbf{T}_i(\mathbf{H}'), \tag{15}$$

where $\{\mathbf{T}_i\}$ are the Chebyshev matrices of the first kind, of degree i. The Chebyshev matrices are in the standard form:

$$\begin{aligned} \mathbf{T}_0(\mathbf{H}') &= \mathbf{I} \\ \mathbf{T}_1(\mathbf{H}') &= \mathbf{H}' \\ \mathbf{T}_{n+1}(\mathbf{H}') &= 2\mathbf{H}'\mathbf{T}_n(\mathbf{H}') - \mathbf{T}_{n-1}(\mathbf{H}'). \end{aligned} \tag{16}$$

This assumes that either $\mathbf{H}'$ is in an orthogonal basis or it has been orthogonalized prior to application. If non-orthogonal basis functions are used, then $\mathbf{H}'_{\alpha\beta}$ is a covariant quantity and $\mathbf{K}^{\alpha\beta}$ should be contravariant. This can be achieved by raising one index of the Hamiltonian matrix, $\mathbf{H}'^{\alpha}{}_{\beta}$ either by multiplying by the inverse overlap matrix or solving $\mathbf{S}_{\gamma\alpha}\mathbf{H}'^{\alpha}{}_{\beta} = \mathbf{H}'_{\gamma\beta}$, to make it contra-covariant. In so doing, the Chebyshev products are all well defined and the resultant $\mathbf{K}^{\alpha}{}_{\beta}$ matrix is also contra-covariant. To make the density matrix fully contravariant in the form necessary for computing the electronic density, the column index can be raised by, for instance multiplying by the inverse overlap matrix on the right, or solving another linear matrix equation – this is explored in more detail in section VI.

This way of performing the expansion requires approximately P terms for a given accuracy, where P is a function of the smearing width, β, the required accuracy 10^{-D} and the Hamiltonian spectral width, ρ. An alternative form to determine the number of terms required

in the expansion depends on the chemical potential and is given in Suryanarayana [15] .

As we did for equation 10, here we once again use the improvement to the original Chebyshev series of Goedecker, proposed by Liang and Head-Gordon in 2003 [11] which reduces the cost of a Chebyshev representation of the Fermi operator to $O(\sqrt{N})$ number of matrix multiplications. In the best case scenario of a local orbital method where sufficient matrix sparsity leads to a cost of matrix multiplication being $O(N)$ this would lead to a cost per SCF iteration proportional to $O(N^{3/2})$.

In order to reduce this cost from $O(N^{3/2})$ to $O(N)$ we limit the width of the FOE to the constant interval $[-c, c]$ as we mention in section III. We apply this FOE to a high temperature Hamiltonian, with a spectral radius less than c. In effect this means that we keep on raising the temperature with Hamiltonian spectral radius while keeping the FOE expansion length small and independent of $\rho_{\mathbf{H}'}$. Then the quenching formula of equation equation 10 is applied n times which is proportional to the system by equation 8 so the overall scaling per SCF iteration is $O(N \log_2(\rho_{\mathbf{H}'}/c))$. So both the quenching procedure of section III and the FOE procedure of this section have $O(N)$ computational cost, and as a result the overall cost of the method scales as $O(N)$.

V. ELECTRONIC ENTROPY

Since we have a finite electronic temperature, the energy which is minimised to find the ground state of a system is the Helmholtz free energy. Therefore, an electronic entropy term must be calculated.

The entropic contribution of the i*th* state is well known, as we are dealing with non-interacting, finite temperature Fermions:

$$s_i = f_i \ln(f_i) + [1 - f_i] \ln(1 - f_i), \tag{17}$$

on a state-by-state basis where f_i is the occupancy of the ith state. So, if the eigenvalues of the Hamiltonian matrix and hence the occupancies were easily accessible, then calculating the entropy would be simply a matter of using (17) for every occupancy, f_i and then summing them to get the contribution to the entropy:

$$\mathcal{S} = \sum_i s_i \ . \tag{18}$$

In the case where we are performing a Fermi-operator-expansion and eigenvalues are not available, another approach must be taken. In this case we aim to calculate an entropy matrix, in the same basis as the density matrix, with eigenvalues s_i.

This can be achieved in a number of ways but conceptually, using a matrix logarithm is the simplest method. Analogously to (17), the entropy matrix can be calculated as:

$$\mathcal{S} = \mathrm{tr}[\mathbf{K} \ \ln(\mathbf{K}) + [\mathbf{I} - \mathbf{K}] \ \ln(\mathbf{I} - \mathbf{K})] \ , \tag{19}$$

which concentrates all of the computation to the matrix logarithm function and the entropic contribution to the free energy can be written as

$$-T \sum_i s_i = -T\mathcal{S}. \tag{20}$$

The difficulty arises when attempting to compute the matrix logarithm, two of which must be performed to evaluate (19). Firstly, in order to calculate a matrix logarithm, the matrix argument must be positive definite, but $\mathbf{K}$ necessarily has eigenvalues close to zero by construction, which may be indistinguishable from zero on a finite precision computer.

In order to overcome this limitation, those eigenvalues closer to zero than some pre-defined threshold can be projected out, with the justification that the corresponding eigenvalue from the pre-multiplicative matrix, $\mathbf{K}$ or $\mathbf{I} - \mathbf{K}$, would have zeroed out this very large negative number.

Performing this operation adds a good deal of computation onto an already expensive calculation, as two extra step function projection operations must be performed for every electronic entropy calculation.

Computing the matrix logarithm is a very computationally demanding operation in itself. One way to calculate a matrix logarithm is by using the Mercator series

$$\ln(1 + x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots , \tag{21}$$

which can be applied to a general matrix $\mathbf{A}$ by linearity:

$$\ln(\mathbf{A}) = \ln(\mathbf{I} + \mathbf{X}) = \mathbf{X} - \frac{\mathbf{X}^2}{2} \cdots \tag{22}$$

Doing this in practice is ill-advised due to the poor convergence properties of the series. An alternative proposed by Kenney and Laub[32], which has been rigorously explored and revised in [33] is to use an inverse scaling and squaring approach. This is roughly the scaling and squaring algorithm for calculating matrix exponentials with the operations in reverse order.

The steps in the inverse scaling and squaring algorithm are firstly to calculate $\mathbf{B} = \mathbf{A}^{1/(2^m)}$, where m is sufficiently large that the result $\mathbf{B}$ is arbitrarily close to an identity matrix. $\ln(\mathbf{B})$ can then be approximated by an n^{th} order Padé approximant to $\ln(1 + x)$ as

$$\ln(\mathbf{B}) \approx \mathrm{pa}^{[m/m]}(\mathbf{B} - \mathbf{I}), \tag{23}$$

so that the final result may be calculated as:

$$\ln(\mathbf{A}) = 2^m \ln(\mathbf{B}). \tag{24}$$

Applying this algorithm, presented by Higham[34] costs $(12 + 2m - 2/3)$ matrix multiplications, which may be a huge expense, depending on the cost of matrix products.

An alternative to the direct approach of calculating (19) with matrix logarithms is to find a function which approximates (19) with reduced computational overhead.

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

One possible form, which we propose here, is:

$$s(x) \simeq ax^2 + \frac{b}{c + dx - dx^2} - e - ax = y(x), \qquad (25)$$

where $a = 1.96056$, $b = 0.0286723$, $c = 0.114753$, $d = 1.98880$ and $e = 0.249860$.

This is an inverse quadratic with a quadratic fit to the error subtracted from it. This form approximates the entropy expression, and it has a mean squared error of 1.0×10^{-6} for each s_i approximated, as demonstrated in figure 1. A further refinement can be made to this by fitting to the error of this expression with another Padé approximant, such as

$$s(x) \simeq y + \frac{a'y + y^4 + b'y^3 + c'y^2 - d'}{e' - h'y - b'y^2 - k'y^3} = z(y), \qquad (26)$$

where $a' = 0.01548792$, $b' = 1.1542349$, $c' = 0.3418894$, $d' = 4.964447 \times 10^{-6}$, $e' = 0.04446540$, $h' = 1.806607$, $k' = 5.0611405$ and where y are values calculated with (25). This form approximates the entropy expression, and it has a mean squared error of 1.0×10^{-8}, for each s_i approximated, as shown in figure 1. We use equation 26 in all of our calculations.

In order to avoid the inversions which can be seen in equations 25 and 26 we take these expressions and compute Chebyshev expansions of them. We apply such Chebyshev expansions to the density kernel to get the entropy matrix, from which we can compute the entropy as the trace. In this way we avoid the need for matrix inverses in applying equations 25 and 26.

VI. IMPLEMENTATION IN TERMS OF NON-ORTHOGONAL LOCAL ORBITALS IN THE ONETEP CODE

A major problem in applying the techniques described in the previous sections is that the orthogonalization of the Hamiltonian matrix prior to application of the expansion method is costly and severely limits the sparsity of the orthogonalized Hamiltonian matrix. This has the knock-on effect that each of the matrix products in the expansion method is more expensive than using non-orthogonal matrices and the feasibility of the method is limited to systems where the full or almost full N^2 Hamiltonian matrix may be stored in memory.

We aim to implement our approach within the framework of the ONETEP [3] linear-scaling (up to now for insulators) DFT code which is based on a representation of the molecular orbitals in terms of a set of non-orthogonal strictly localised orbitals $\{\phi_\alpha\}$ which are optimised in situ [35], and remain atom centered throughout the calculation. These localised orbitals are called Non-orthogonal Generalised Wannier Functions (NGWFs) and are expressed in a basis set of psinc functions [36] which are equivalent to a plane wave basis set. In the conventional diagonalisation-based EDFT method implemented in ONETEP [37] the covariant Hamiltonian matrix $H_{\alpha\beta}$ in the representation of the NGWFs is optimised in the inner loop (the EDFT loop) of a two nested loop procedure, while the NGWFs are optimised in the outer loop using a conjugate gradients approach. We have used this energy minimizer unmodified in this work, because it has been pre-validated in [37] , where it has been shown that with increasing NGWF radii the calculations converge to plane-wave basis set results.

The implementation of EDFT in ONETEP prior to this work can be used to calculate any property that the linear-scaling implementation of LNV can be used to calculate. This includes forces, including Pulay terms [38] for MD and structure relaxations, as well as properties calculations including local and projected density of states calculations. In previous work geometry optimisations were shown on large metallic clusters [39] .

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

$$\rho(\mathbf{r}, \mathbf{r}') = \sum_{\alpha\beta} \phi_\alpha(\mathbf{r}) K^{\alpha\beta} \phi^*_\beta(\mathbf{r}'), \qquad (27)$$

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. 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,

$$A'[\mathrm{T}; \{\phi_\alpha\}] = \min_{\{H_{\alpha\beta}\}} A[\mathrm{T}; \{H_{\alpha\beta}\}; \{\phi_\alpha\}] \qquad (28)$$

with respect to $\{\phi_\alpha\}$. In conventional EDFT the density kernel can be generated explicitly as a function of the eigenvalues ε_i (obtained by diagonalisation) of the Hamiltonian as:

$$K^{\alpha\beta} = \sum_i^N M^\alpha{}_i f(\varepsilon_i) M_i^{\dagger\,\beta}, \qquad (29)$$

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

$$H_{\alpha\beta} M^\beta{}_i = S_{\alpha\beta} M^\beta{}_i \, \varepsilon_i. \qquad (30)$$

Such eigenvalue based approaches will always scale as $O(N^3)$ as they employ matrix diagonalization algorithms.

The aim of this work is to replace the diagonalisation, the building of the density kernel, the chemical potential search and the computation of the entropy in EDFT with our new AQuA-FOE method. Up to now ONETEP has used internally fully covariant and fully contravariant tensors but for our work here we will follow the contra-covariant approach suggested by Gibson, Haydock and

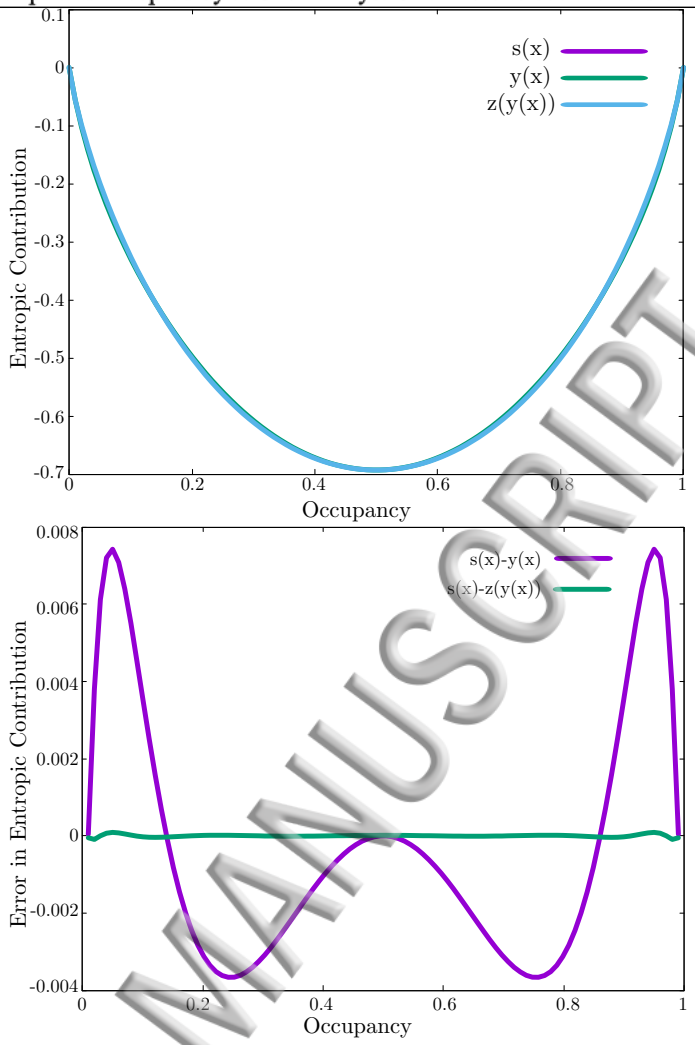


FIG. 1: The two entropy approximations suggested in this work are overlayed in (a) and the relative errors shown in (b)

LaFemina [40]. This involves working directly with the Hamiltonian in the natural representation, which is the contra-covariant Hamiltonian matrix, or $H^{\alpha}_{\ \beta}$.

$H^{\alpha}_{\ \beta}$ can be used in place of the H in any of the matrix products described and the result is tensorially correct, and as both matrices have the same eigenvalues, the resulting contra-covariant density matrix has the correct eigenvalues. The standard contravariant density matrix can then be formed by either multiplying on the right by the inverse overlap matrix to raise the index, or by solving

$$K^{\alpha}_{\ \beta} = K^{\alpha\gamma} S_{\gamma\beta}, \tag{31}$$

for $K^{\alpha\gamma}$.

The method of Gibson, *et al* works by assuming that the Hamiltonian is a local operator, and that when applied to a state, the value of the result at position $\mathbf{r}$ depends only upon the value in a small localization region surrounding the state at position $\mathbf{r}$. This must be the case because when applying the Hamiltonian operator, the contra-covariant form is the correct matrix representation of the operator and this preserves the locality of the states. Hence, $H^{\alpha}_{\ \beta}$ is a sparse matrix, if $H_{\alpha\beta}$ is sparse even if $S^{\alpha\beta}$ is not. Using this logic, the authors suggest solving the matrix equation

$$S_{\gamma\alpha} H^{\alpha}_{\ \beta} = H_{\gamma\beta}, \tag{32}$$

for $H^{\alpha}_{\ \beta}$, without constructing the inverse overlap matrix, $S^{\gamma\alpha}$. This can be done in practice approximately using the same locality arguments as above by taking the sparsity pattern of the resultant contra-covariant Hamiltonian matrix as the subspace of local interactions: that is to construct small matrices for each column of the resulting matrix from the non-zero elements of $H_{\gamma\beta}$ on the mask:

$$M_i = \mathbf{v}_i \mathbf{v}_i^T, \tag{33}$$

where $\mathbf{v}_i$ is the column vector of nonzero elements from the ith column of $H_{\alpha\beta}$. This "nonzero elements of the ith

column squared matrix problem" can be solved to give the ith column of the resultant matrix:

$$ {}^{(i)}S_{\gamma\alpha}\,{}^{(i)}H^{\alpha}_{\ \beta} = {}^{(i)}H_{\gamma\beta}, \tag{34} $$

and then the elements are put back into the sparsity pattern to produce the approximate $H^{(A)\alpha}_{\ \ \beta}$.

The accuracy of the approximation can be checked trivially by using

$$ \Delta = S_{\gamma\alpha}H^{(A)\alpha}_{\ \ \beta} - H_{\gamma\beta}. \tag{35} $$

and the norm of Δ gives an estimate for the error. This can be minimized by increasing the size of the localization regions of the local functions.

In practice, we use the same sparsity pattern for $H^{\alpha}_{\ \beta}$ as we do for $H_{\alpha\beta}$, as was suggested in Gibson *et al.* In the calculations we have performed so far, this has proved adequate based on the error estimate given by taking $||\Delta||_F$, but if this error ever proved too large, we have the option to reduce the sparsity of $H^{\alpha}_{\ \beta}$ by multiplying the radii of the spherical localization regions by a value > 1.0 and constructing a more accurate sparsity pattern from these more delocalized regions.

The sparsity pattern of the contra-covariant Hamiltonian matrix is only one of the considerations for sparsity, however. We must also pay close attention to the sparsity pattern of the density matrices. In most linear-scaling DFT packages, the single particle density matrices are kept sparse through either a geometric cutoff between atomic centres, or a truncation by estimating the absolute magnitude of matrix elements and setting to zero, those elements which are smaller than the threshold.

In this work, we have opted to use an alternative method based on the nature of the FOE-type approaches. Since we perform a power series expansion, we opt to limit the sparsity of the contra-covariant density matrix to be no more dense than the $H^{\alpha}_{\ \beta}H^{\beta}_{\ \gamma}$ term. In so doing, we limit the accuracy of small systems, but calculations on these systems are likely to be possible with conventional methods, or with dense matrices so the compromise is worth it. In large systems, where resulting sparsity may be adequate for linear or near-linear scaling, care must be taken to ensure that an appropriate electronic smearing is used, as the error induced by the sparsity is reduced when raising the electronic temperature.

This sparsity pattern is set at the beginning of a calculation (or at the beginning of each geometry iteration, if performing a structural relaxation) based on the overlap of the spheres containing the NGWFs. In ONETEP, the sparsity patterns are based on such geometric considerations, so the sparsity pattern of the Hamiltonian matrix will not change through the calculations and hence, neither will the density kernel matrix. We find that with our Hamiltonian squared sparsity pattern that the non-zero elements correlate particularly well with the largest elements of the density kernel when compared with a density kernel calculated with no sparsity enforced (see figure 2).

The sparsity pattern chosen for a given system type should be tested on that system as it may not be transferrable in general. The sparsity pattern in this work gets less sparse with increasing NGWF radius, which is the main parameter which controls the accuracy of this approximation in our metals method calculations.

We find that for a truncated octahedral Al_{2406} nanoparticle that electronic temperature has a large effect on the achievable accuracy (fig. 2). At 2.5 eV smearing, we reach a maximum error of about 10^{-4} in the elements of the density kernel , whereas if we increase to 25 eV or 250 eV, we get maximum errors of 10^{-6} and 10^{-8}, respectively.

VII. CHEMICAL POTENTIAL SEARCH

When applying the Fermi-Dirac distribution at a particular electronic temperature to the eigenvalues of a Hamiltonian matrix, the chemical potential μ is found trivially by applying the Fermi function to all of the eigenvalues with a trial value of chemical potential μ_0, summing all of the results of these applications. The resulting scalar (number of electrons) is compared with the desired number of electrons. This process is repeated with a modification to the chemical potential until the difference between calculated and desired number of electrons is below some arbitrary threshold.

This approach is equivalent to using a root finding method on the following equation, with the chemical potential μ as the independent variable:

$$ \Delta N_e = N_e - \sum_{i=1}^{N} \frac{1}{1 + e^{(\epsilon_i - \mu)\beta}}, \tag{36} $$

where N_e is the target number of electrons, N is the size of the eigenspace, ϵ_i are the eigenvalues of the Hamiltonian matrix, μ is the chemical potential and β is thermodynamic temperature $\beta = 1/kT$.

This direct approach is efficient if the eigenvalues are known. However, when performing a Fermi-operator-expansion (FOE) in the non-diagonal space, the eigenvalues are not known, but the number of particles can still be calculated as the trace of the density matrix. A similar approach could then be used:

$$ \Delta N_e = N_e - \mathrm{trace}\left(\frac{\mathbf{I}}{\mathbf{I} + e^{(\mathbf{H} - \mu\mathbf{I})\beta}}\right), \tag{37} $$

where $\mathbf{H}$ is the Hamiltonian matrix (which is assumed to be orthogonal or be contra-covariant as described in section VI) and $\mathbf{I}$ is the identity matrix of dimension N. The problem with this approach is that each evaluation of the density matrix is a relatively expensive operation, which one would not want to perform many times per SCF iteration.

A solution can be found by referring again to equation 4 and 9 and noticing that a density matrix calculated at

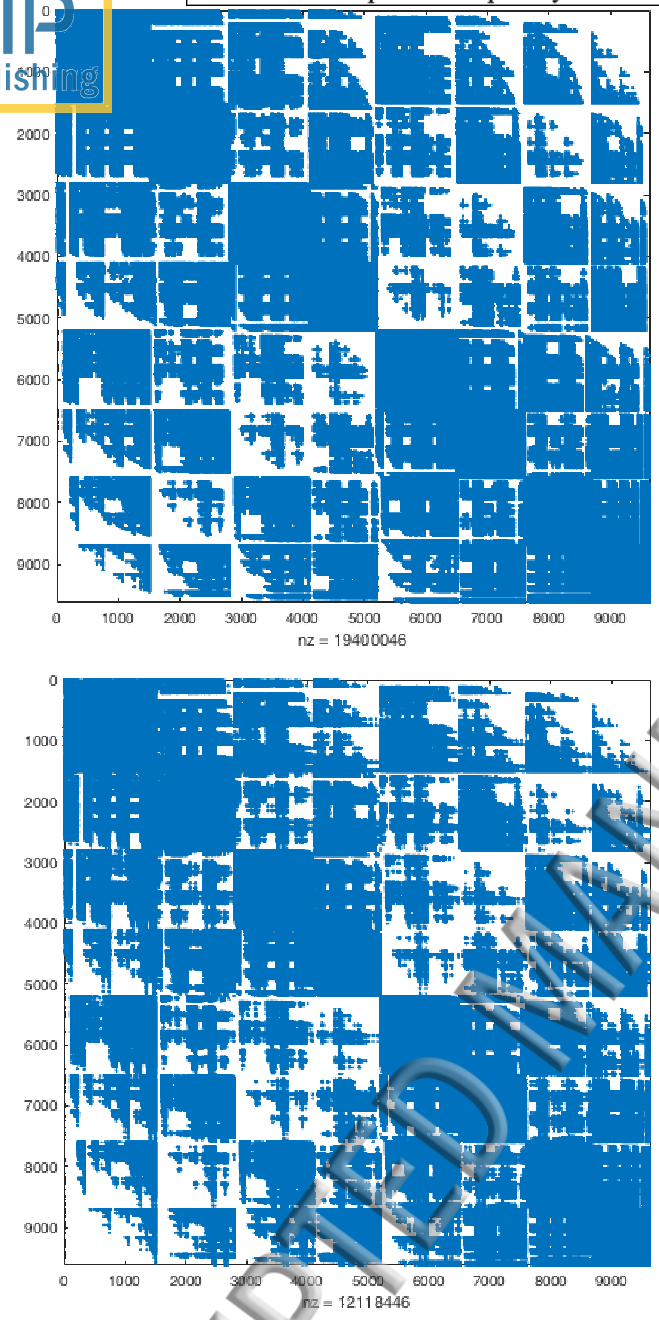


FIG. 2: A comparison showing the sparsity pattern of the ${H^{\alpha}}_{\beta}{H^{\beta}}_{\gamma}$ matrix (above) and the sparsity pattern resulting from a 2.5 eV smeared density kernel calculated without truncation and then having a truncation applied to its elements which are smaller than a threshold of 10^{-4} (below). These are calculations on a truncated octahedral Al nanoparticle with 2406 atoms (4 NGWFs per atom). The axes labels in this figure represent matrix row and column indices and the number beneath each of them shows the number of non-zero elements in each.

a particular chemical potential can be modified to be at a different chemical potential at a lower cost than recalculation.

With this in mind, hyperbolic, trigonometric identities can be used to provide increments to the chemical potential, using

$$\tanh(x \pm y) = \frac{\tanh(x) \pm \tanh(y)}{1 \pm \tanh(x)\tanh(y)}, \tag{38}$$

or for matrices, where $\rho_{\mathbf{H},\mu,\beta}$ is the scaled and shifted density matrix of equation, as in equation 9 calculated for a Hamiltonian, $\mathbf{H}$, at a chemical potential μ, we can use

$$\tanh\left(\frac{((\mathbf{H}-\mu\mathbf{I})\beta)}{2} \pm \frac{\beta\Delta\mu\mathbf{I}}{2}\right) = \frac{\rho_{\mathbf{H},\mu,\beta} \pm \tanh(\frac{\beta\Delta\mu}{2})\mathbf{I}}{\mathbf{I} \pm \tanh(\frac{\beta\Delta\mu}{2})\rho_{\mathbf{H},\mu,\beta}}. \tag{39}$$

Provided that $\mathbf{I} \pm \tanh(\frac{\beta\Delta\mu}{2})\rho_{\mathbf{H},\mu,\beta}$ can be inverted efficiently, this method should involve far less computation than recalculation of the density matrix.

If the matrix is inverted with a Newton-Shulz-Hotelling algorithm, then each time the root-finding algorithm calls for a new trial chemical potential, then the inverse can be initialised with the inverse from the previous trial point saving significant calculation effort.

What we do in practice is to actually to use a further Chebyshev expansion of equation 39. The coefficients for this expansion have to be computed every time a new $\Delta\mu$ is used, but this is a relatively inexpensive operation. When computing the expansion we take the scalar form of equation 39 as:

$$q(x) = \frac{x+c}{1+xc}, \tag{40}$$

on the domain [-1,1], where c is $\beta\Delta\mu/2$. We then evaluate the Chebyshev expansion of $\rho_{\mathbf{H},\mu,\beta}$ using these coefficients, again using the divide and conquer algorithm of Liang and Head-Gordon, *et al.* to provide us with the density matrix at the updated chemical potential, $\rho_{\mathbf{H},\mu+\Delta\mu,\beta}$.

We can go further than this, however, and use derivatives with respect to the chemical potential to speed up the root search. The first derivative of the density matrix with respect to chemical potential, is:

$$\frac{\partial\rho}{\partial\mu} = -\frac{\beta}{4}\left(\mathbf{I}-\rho^2\right), \tag{41}$$

because

$$\frac{d(\tanh(x))}{dx} = \mathrm{sech}^2(x) = 1 - \tanh^2(x). \tag{42}$$

Furthermore, because the trace operator commutes with the differential operator, we can write:

$$\frac{\partial N_e}{\partial\mu} = -\frac{\beta}{4}\left(1 - \mathrm{trace}(\rho^2)\right). \tag{43}$$

As $N_e(\mu)$ is a monotonic function (when using the Fermi-Dirac distribution), the root finding ought to be simple with Newton's method. It is complicated slightly, however, because $N_e(\mu)$ has multiple stationary points which increase in breadth with Hamiltonian eigenvalue-spacing and decreasing temperature.

Using Newton's method at one of these points would send the next trial point off to ±infinity, so a safe-guarded version is used in practice, where a trust-region is defined and if Newton's method is going to send the next point outside of these bounds, then a bisection step is performed.

VIII. VALIDATION TESTS

In ONETEP, we have implemented the Annealing and QUenching Algorithm - FOE (AQuA-FOE) approach that we present in this paper for calculating the density matrix from a given Hamiltonian matrix and also to calculate an electronic entropy matrix. This type of FOE is applied at every step of the ensemble-DFT (EDFT) approach, which is used as the electronic energy minimization technique.

To test the AQuA-FOE method we have carried out numerical comparisons against the conventional diagonalisation-based EDFT method which is available in ONETEP [37]. For this we use small cuboctahedral platinum nanoparticles. We performed all of these calculations in ONETEP with EDFT, at 500 eV kinetic energy cut-off, 9.0 Bohr radius localization spheres for the NGWFs, and with PAW[41], using the data from the GBRV pseudopotential dataset [42]. The platinum-platinum distance was set to the bulk value of 2.8 Å. We ran the AQuA-FOE scheme where the chemical potential search was configured to halt once the chemical potential had been found to within 10^{-6}E$_H$ - 10^{-8}E$_H$. We also used the refined approximation to the entropy which was described in section V and an electronic smearing width of 0.1 eV. We want to make clear that these systems are too small for demonstrating linear-scaling behaviour, because we do not have any sparsity in the density kernel, but are effectively tests of the methodology against exactly the same EDFT scheme, but with diagonalisation. These calculations have been done to demonstrate convergence of the energy with respect to our chemical potential search stopping criterion.

With this prescription for a cuboctahedral platinum nanoparticle with 55 atoms, the convergence in energy of the AQuA-FOE scheme with chemical potential search stopping criterion can be seen in Table I. We observe rapid convergence (shown in table I) of the energy to a value of -4914.74445 E_H while the the same converged total energy with diagonalization is -4914.74442 E_H. With a 147-atom Pt nanoparticle we calculated an energy of -13137.33174 E_H with a diagonalization based technique and the converged AQuA-FOE value is -13137.33169 E_H. There was also no difference in the number of iterations

TABLE I: The convergence in total energy in E$_H$ of AQuA-FOE for cuboctahedral platinum nanoparticles with respect to chemical potential search stopping criterion (tolerance)

μ tolerance	Pt_{55}	Pt_{147}
10^{-6}	-4914.74211	-13137.32070
10^{-7}	-4914.74475	-13137.33179
10^{-8}	-4914.74445	-13137.33169

(either optimization of the Hamiltonian in the inner loop or optimization of the NGWFs in the outer loop) between all the AQuA-FOE and the diagonalisation calculations. The tightest chemical potential convergence criterion value of 10^{-8}E$_H$ has been used for all of the rest of the calculations in this paper.

Next we examined the scaling of the AQuA-FOE method with system size using truncated octahedral Pd nanoparticles with 2406, 4033, 6266, 9201, 12934 atoms. We chose these as truly 3-dimensional examples as they are a much more stringent test of density matrix decay as compared to 1- and 2-dimensional systems which can be more straightforward to demonstrate reduced or linear-scaling performance with the number of atoms due to a lower crossover point with cubic-scaling methods. We chose to run with the Hamiltonian matrix squared sparsity for the 1-particle density matrix, as described in section VI and an electronic smearing of 0.5 eV. We also used the refined approximation to the entropy which was described in section V. A psinc basis set kinetic energy cut-off of 500 eV was used and 9 NGWFs per Pd atom were employed, with NGWF radii of 6.0 a$_0$.

We ran all of these calculations on Archer, the UK's national supercomputer using 2400 MPI processes with 2 OpenMP threads per process. The calculations were compared against the timings of ONETEPs diagonalization based EDFT method for each system with the same number of processes and threads. The eigendecomposition was performed by the Scalapack implementation in the Intel MKL library and all of the sparse matrix operations in the AQuA-FOE calculations were performed using the SPAM3 sparse matrix algebra library which is an integral part of the ONETEP code. We did not have sufficient computing resources to perform all of the calculations to convergence, running instead a fixed number of four inner and four outer loop iterations to assess average time per outer loop iteration. From previous experience with the EDFT approach, we expect a small increase in the number of outer loop iterations with system size [37].

With ONETEP, the current implementation of EDFT with AQuA-FOE and Hamiltonian matrix squared sparsity has crossover point with ONETEPs highly optimized dense-matrix, diagonalization based EDFT scheme at ∼2000 atoms (see figure 3). The observed linear-scaling ($O(N)$, to be precise) of the AQuA-FOE method shows that is it 5 times quicker when we reach Pd_{9201} (4 hours per EDFT outer loop iteration (1 hour per inner loop iteration) than the cubically scaling diagonalization technique (20 hours per EDFT outer loop iteration). Figure

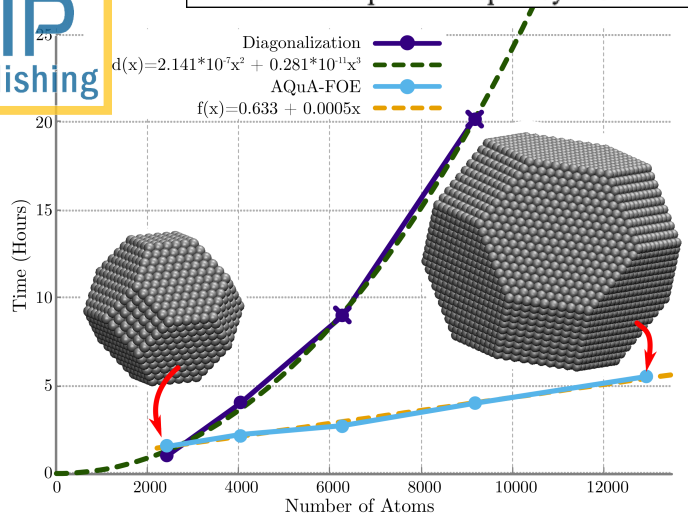


FIG. 3: A comparison of EDFT in ONETEP using both Diagonalization with Scalapack (deep purple) and AQuA-FOE (light blue). We have performed calculations with 4 inner and 4 outer loop (EDFT) iterations for regular truncated octahedral nanoparticles of Palladium with 2406, 4033, 6266, 9201 and 12934 atoms. The average run-times (per outer loop iteration) are presented on the y-axis, a cubic fit is shown through the diagonalization timings (deep green) and a linear fit is shown though the AQuA-FOE timings (orange). The geometries of the Pd_{2406} and Pd_{12934} nanoparticles are inset.

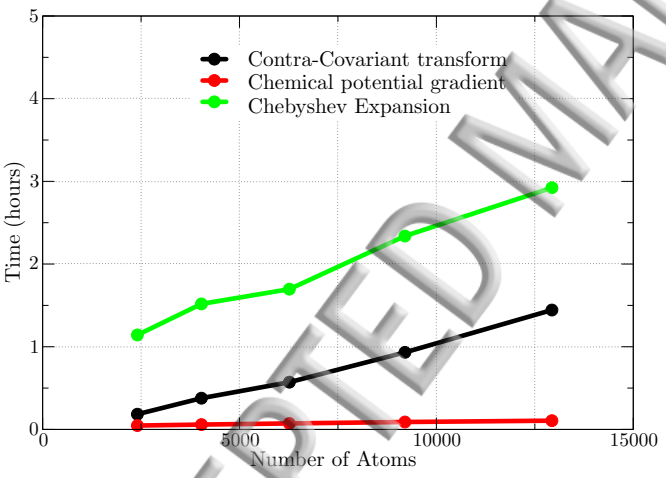


FIG. 4: A breakdown of the average run-times (per outer loop iteration) into the times taken to compute the contra-covariant solve as described by Haydock *et al*, the time taken to compute the energy derivatives with respect to the chemical potential and the time taken to compute all of the Chebyshev expansions used in this method.

4 presents a breakdown of the timings into component parts for the FOE calculations. This linear scaling results from exploitation of the exponential decay properties of the metal density matrix at finite electronic temperature [27], by imposing a sparsity pattern on the density matrix throughout our calculations. The 12934 atom system with diagonalization was excessively demanding in terms of computational expense and so we were not able to run it for comparison.

We expect that with the ever increasing performance of computational resources, and after further optimization of the AQuA-FOE method and code (perhaps through the use of the SelInv[43] algorithm rather than Chebyshev expansion to perform the quenches), that this method will become increasingly useful for calculations of industrially important large, metallic systems in technological applications in fields such as heterogeneous catalysis and biosensing.

We expect the cross-over point to the linear-scaling regime to be different for different materials and methods. For palladium we found the cross-over point to be around 2000 atoms, but this is not representative of crossover against methods that contain cubic scaling components, because the EDFT of ONETEP already has several components (e.g. Hamiltonian matrix construction) which are sparse due to NGWF localisation even though the method involves a cubic scaling diagonalization as the bottleneck. So a cross-over with a fully cubic-scaling method at around 1000 atoms is more plausible as mentioned by Suryanarayana *et al*[9].

IX. CONCLUSIONS

We have presented the AQuA-FOE method for running DFT calculations on large metallic systems with a computational cost which increases effectively linearly with the number of atoms. The ONETEP code, which has already supported linear-scaling DFT calculations on insulators for many years provided the framework for developing our new method. Most of the machinery of ONETEP such as non-orthogonal local orbitals optimised *in situ* (strictly localized orbitals, which we call NGWFs), and sparse, CPU-distributed matrices was used within our method. As in linear-scaling calculations on insulators, the exponential decay in the one-particle density matrix, which is also exhibited in finite temperature metallic systems was exploited to achieve reduced scaling.

The AQuA-FOE method works by running a Fermi Operator Expansion (FOE) on a Hamiltonian at several times the desired temperature. The resultant "hot" density matrix is then quenched repeatedly until the desired (lower) temperature is reached. This means that only a fixed, constant number of matrix multiplications is required to perform the FOE. Since each temperature-halving quenching operation also has a fixed cost, increase in the number of matrix products for different systems can only come from increase in the required number of quenches. However, for a given material, the Hamiltonian spectral width is asymptotically constant with increasing number of atoms. Hence, the number of quenching operations will also remain constant with the number of atoms and the full method is effectively linear-scaling with system size as the cost of each matrix product is

linear-scaling with system size in the limit of sufficient matrix sparsity for large systems.

Our method finds the elements of the one-particle density matrix along with the electron number conserving chemical potential, and also electronic entropy using polynomial expansions which also have a constant number of terms with system size. We have shown validation calculations of AQuA-FOE inside the EDFT procedure by comparing numerically with the diagonalisation based EDFT that is already in ONETEP showing agreement in the energies to better than 10^{-5} E_H per atom. We have also demonstrated the linear-scaling computational cost of our method with calculation times on Palladium nanoparticles ranging from 2,406 to 12,934 atoms.

We expect that this method will become increasingly useful as supercomputing power becomes greater and more available. Complex metallic materials are possible to study with this approach, including large metallic nanoparticles which have a growing number of applications in important technological areas such as catalysis and biomolecular markers. We should note that we have not yet run calculations with the AQuA-FOE method on bulk metallic systems. Other methods based on the FOE concept have been tested successfully on bulk metallic systems[29, 44], so we intend to explore how AQuA-FOE performs for bulk metallic systems in future work, including further developments as necessary.

Acknowledgments

Jolyon Aarons would like to thank Johnson Matthey and the Engineering and Physical Sciences Research Council (EPSRC) for an industrial CASE PhD studentship. The calculations for this work were carried out on the Iridis4 Supercomputer of the University of Southampton and the ARCHER supercomputer via the UKCP consortium (EPSRC grants EP/K013556/1 and EP/P022030/1).

[1] N David Mermin. Thermal properties of the inhomogeneous electron gas. *Physical Review*, 137(5A):A1441, 1965.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] Emil Prodan and Walter Kohn. Nearsightedness of electronic matter. *Proceedings of the National Academy of Sciences of the United States of America*, 102(33):11635–11638, 2005.

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

[9] Phanish Suryanarayana. On nearsightedness in metallic systems for o (n) density functional theory calculations: A case study on aluminum. *Chemical Physics Letters*, 679:146–151, 2017.

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

[11] 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.

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

[13] 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.

[14] 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.

[15] Phanish Suryanarayana. On spectral quadrature for linear-scaling density functional theory. *Chemical Physics Letters*, 584:182 – 187, 2013.

[16] Anders MN Niklasson, Marc J Cawkwell, Emanuel H Rubensson, and Elias Rudberg. Canonical density matrix perturbation theory. *Physical Review E*, 92(6):063301, 2015.

[17] Anders MN Niklasson. Implicit purification for temperature-dependent density matrices. *Physical Review B*, 68(23):233104, 2003.

[18] Taisuke Ozaki. O (n) krylov-subspace method for large-scale ab initio electronic structure calculations. *Physical Review B*, 74(24):245101, 2006.

[19] Jolyon Aarons, Lewys Jones, Aakash Varambhia, Katherine E MacArthur, Dogan Ozkaya, Misbah Sarwar, Chris-Kriton Skylaris, and Peter D Nellist. Predicting the oxygen-binding properties of platinum nanoparticle ensembles by combining high-precision electron microscopy

and density functional theory. *Nano Letters*, 17(7):4003–4012, 2017.

[20] 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.

[21] K Wildberger, R Zeller, and PH Dederichs. Screened kkr-green's-function method for layered systems. *Physical Review B*, 55(15):10074, 1997.

[22] A. V. Smirnov and D. D. Johnson. Accuracy and limitations of localized Green's function methods for materials science applications. *Phys. Rev. B* , 64(23):235129, December 2001.

[23] Johannes M Dieterich and Emily A Carter. Opinion: Quantum solutions for a sustainable energy future. *Nature Reviews Chemistry*, 1:0032, 2017.

[24] Nick DM Hine, Peter D Haynes, Arash A Mostofi, C-K Skylaris, and Mike C Payne. Linear-scaling density-functional theory with tens of thousands of atoms: Expanding the scope and scale of calculations with onetep. *Computer Physics Communications*, 180(7):1041–1053, 2009.

[25] NDM Hine, PD Haynes, AA Mostofi, and MC Payne. Linear-scaling density-functional simulations of charged point defects in alo using hierarchical sparse matrix algebra. *The Journal of chemical physics*, 133:114111, 2010.

[26] Karl A Wilkinson, Nicholas DM Hine, and Chris-Kriton Skylaris. Hybrid mpi-openmp parallelism in the onetep linear-scaling electronic structure code: Application to the delamination of cellulose nanofibrils. *Journal of chemical theory and computation*, 10(11):4782–4794, 2014.

[27] Michele Benzi, Paola Boito, and Nader Razouk. Decay properties of spectral projectors with applications to electronic structure. *SIAM Review*, 55(1):3–64, 2013.

[28] Walter Kohn and Lu Jeu Sham. Self-consistent equations including exchange and correlation effects. *Physical review*, 140(4A):A1133, 1965.

[29] Phanish Suryanarayana, Phanisri P. Pratapa, Abhiraj Sharma, and John E. Pask. SQDFT: Spectral quadrature method for large-scale parallel $\mathcal{O}(n)$ Kohn-Sham calculations at high temperature. *Computer Physics Communications*, 2017.

[30] Petros Souvatzis and Anders M. N. Niklasson. Extended lagrangian born-oppenheimer molecular dynamics in the limit of vanishing self-consistent field optimization. *The Journal of Chemical Physics*, 139(21):214102, 2013.

[31] 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.

[32] Charles Kenney and Alan J. Laub. Condition estimates for matrix functions. *SIAM Journal on Matrix Analysis and Applications*, 10(2):191–209, 1989.

[33] Awad H. Al-Mohy and Nicholas J. Higham. Improved inverse scaling and squaring algorithms for the matrix logarithm. *SIAM Journal on Scientific Computing*, 34(4):C153–C169, 2012.

[34] Sheung Hun Cheng, Nicholas J Higham, Charles S Kenney, and Alan J Laub. Approximating the logarithm of a matrix to specified accuracy. *SIAM Journal on Matrix Analysis and Applications*, 22(4):1112–1125, 2001.

[35] 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.

[36] Arash A Mostofi, Peter D Haynes, Chris-Kriton Skylaris, and Mike C Payne. Preconditioned iterative minimization for linear-scaling electronic structure calculations. *The Journal of chemical physics*, 119(17):8842–8848, 2003.

[37] Álvaro Ruiz-Serrano and Chris-Kriton Skylaris. A variational method for density functional theory calculations on metallic systems with thousands of atoms. *The Journal of chemical physics*, 139(5):054107, 2013.

[38] NDM Hine, M Robinson, PD Haynes, C-K Skylaris, MC Payne, and AA Mostofi. Accurate ionic forces and geometry optimization in linear-scaling density-functional theory with local orbitals. *PHYSICAL REVIEW B*, 83, 2011.

[39] Jolyon Aarons, Lewys Jones, Aakash Varambhia, Katherine E MacArthur, Dogan Ozkaya, Misbah Sarwar, Chris-Kriton Skylaris, and Peter D Nellist. Predicting the oxygen-binding properties of platinum nanoparticle ensembles by combining high-precision electron microscopy and density functional theory. *Nano Letters*, 17(7):4003–4012, 2017.

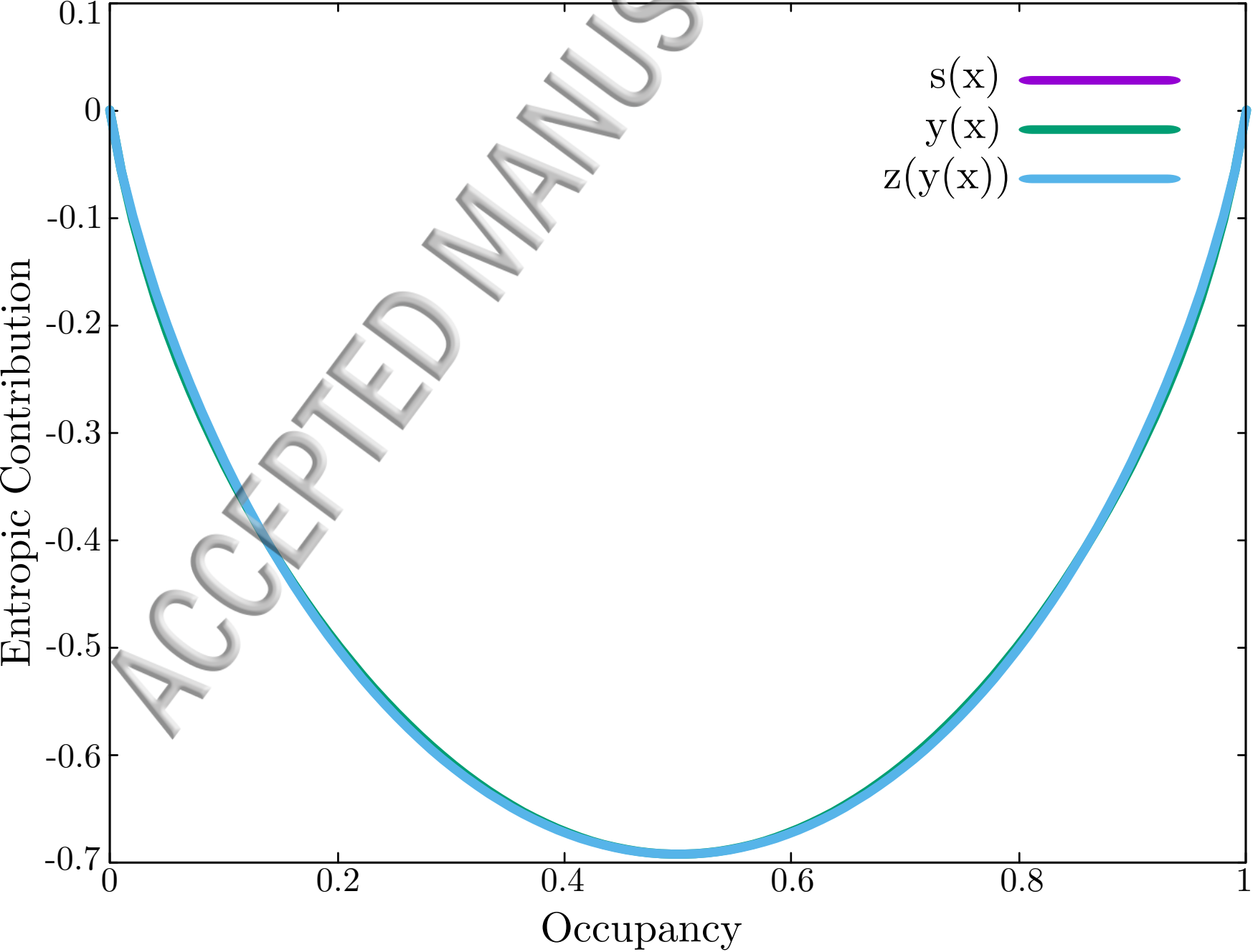
[40] Andrew Gibson, Roger Haydock, and John P LaFemina. Ab initio electronic-structure computations with the recursion method. *Physical Review B*, 47(15):9229, 1993.

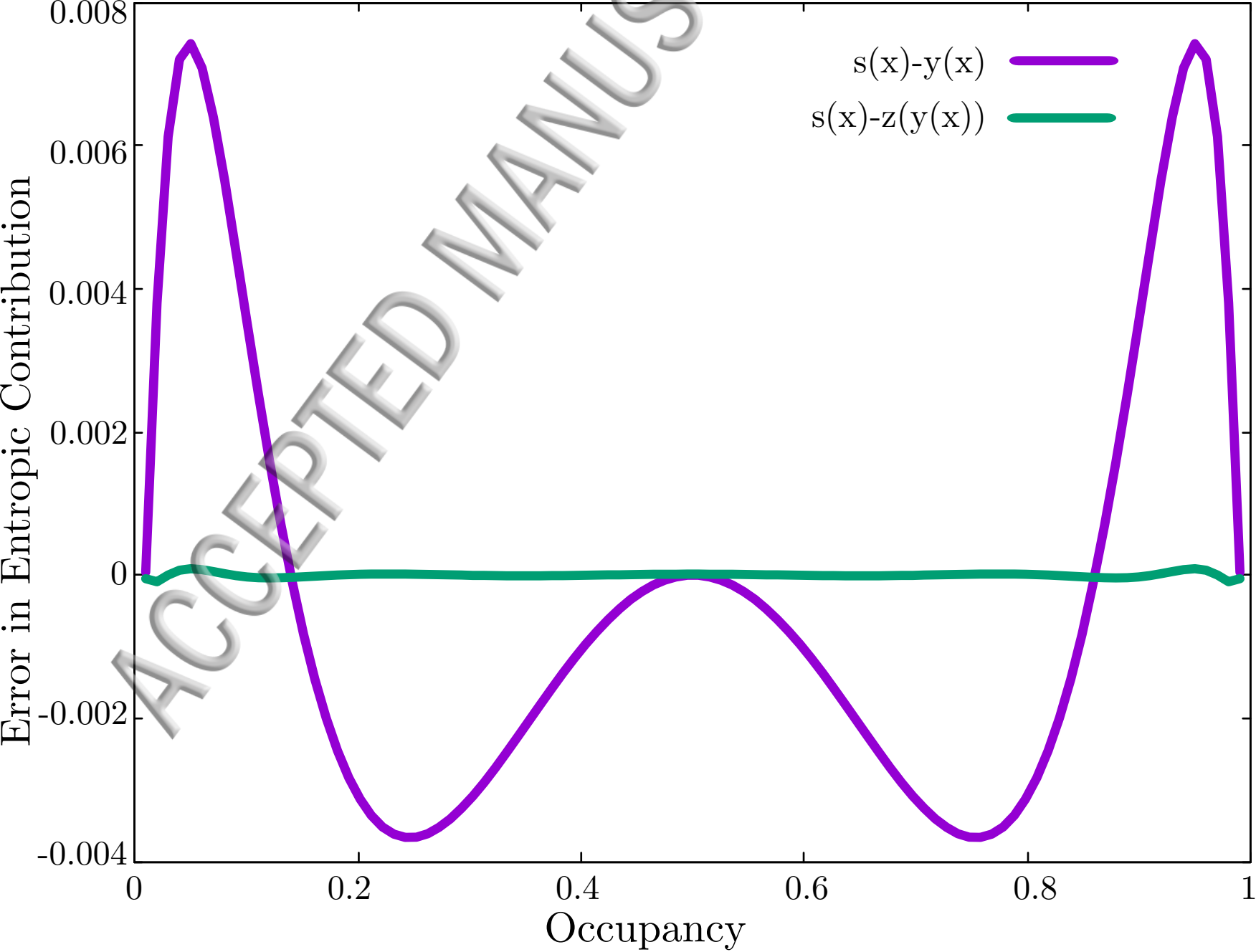
[41] Nicholas DM Hine. Linear-scaling density functional theory using the projector augmented wave method. *Journal of Physics: Condensed Matter*, 29(2):024001, 2016.

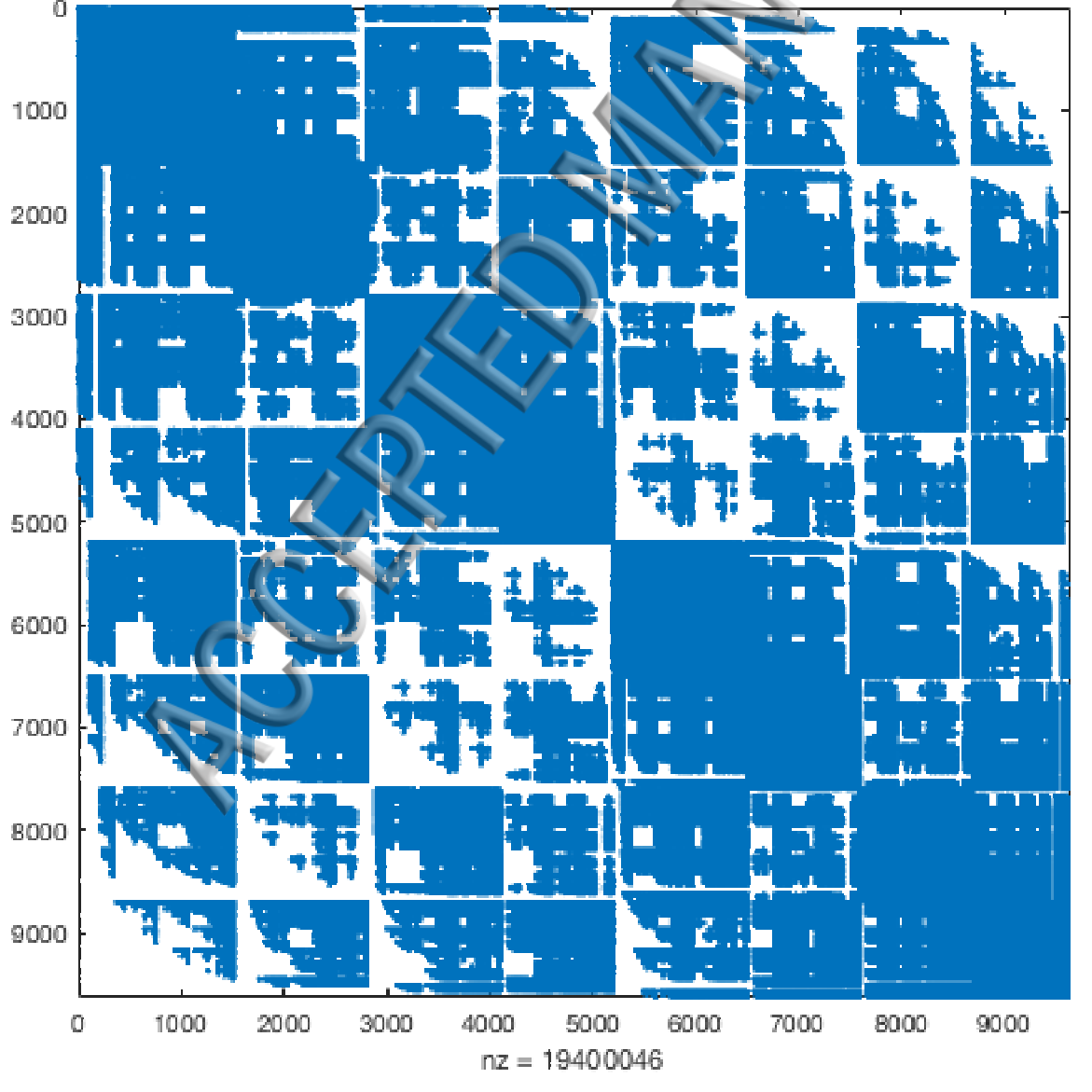
[42] Kevin F. Garrity, Joseph W. Bennett, Karin M. Rabe, and David Vanderbilt. Pseudopotentials for high-throughput dft calculations. *Computational Materials Science*, 81:446 – 452, 2014.

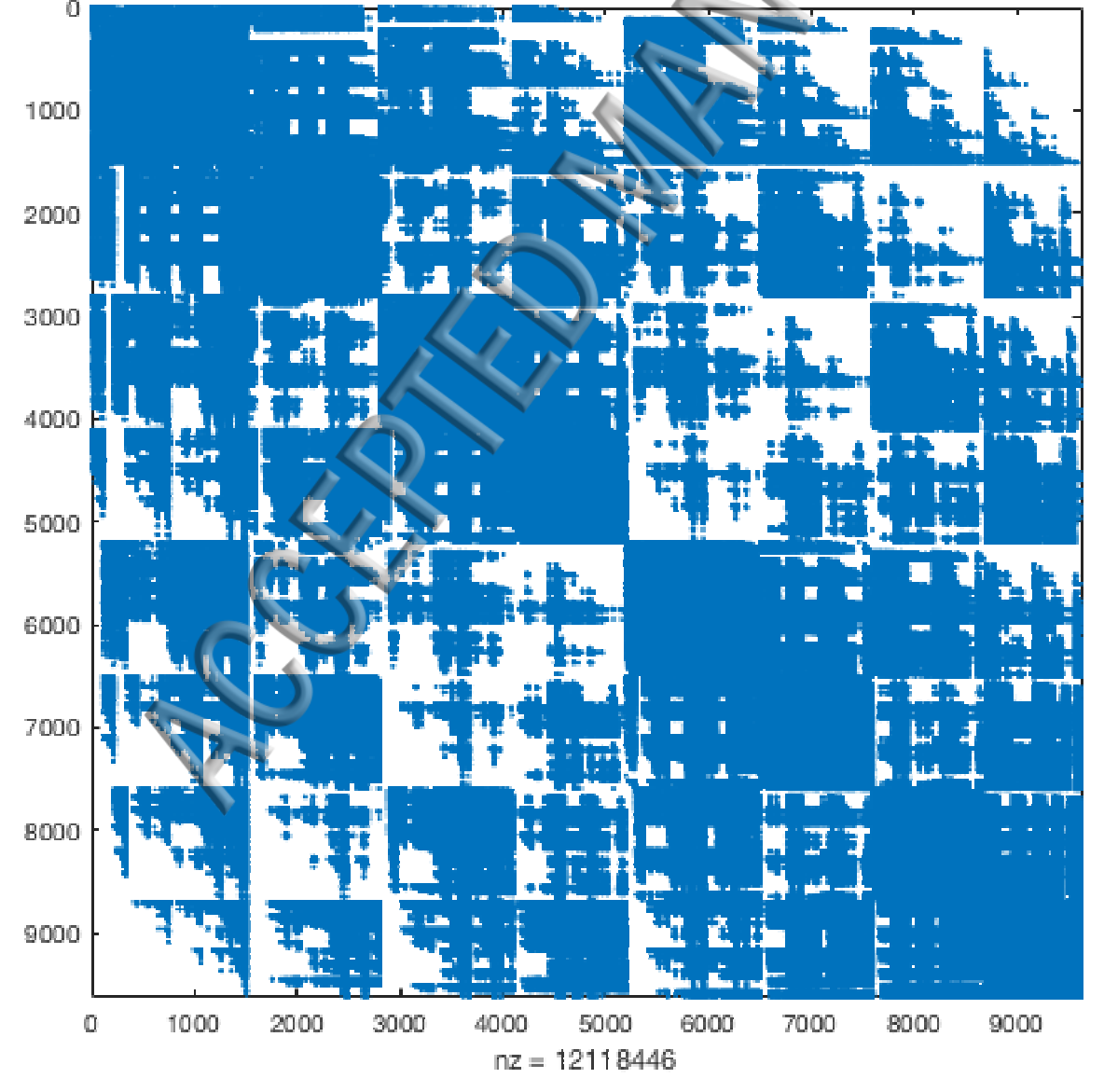
[43] Lin Lin, Chao Yang, Juan C Meza, Jianfeng Lu, Lexing Ying, et al. Selinv—an algorithm for selected inversion of a sparse symmetric matrix. *ACM Transactions on Mathematical Software (TOMS)*, 37(4):40, 2011.

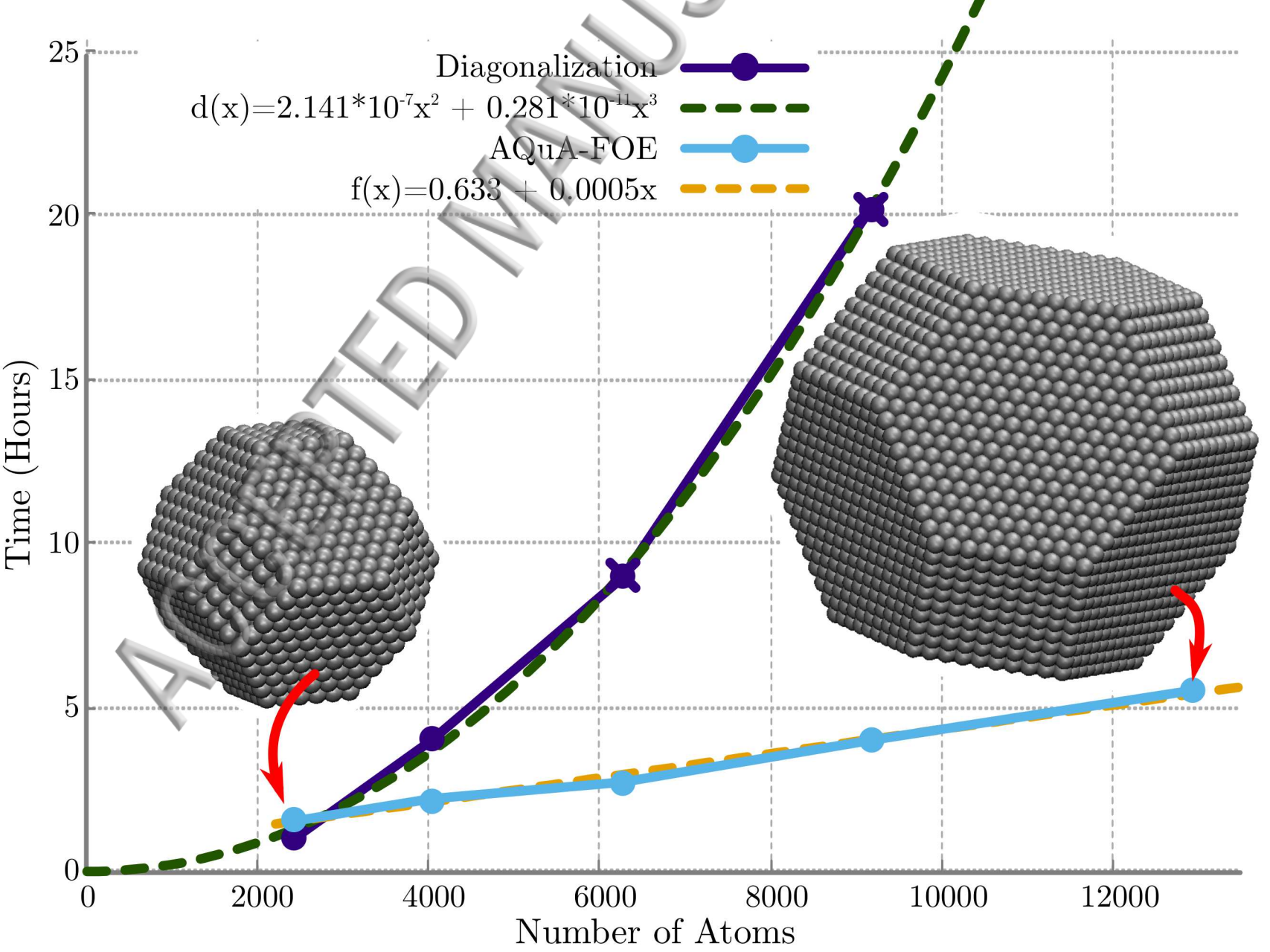
[44] Stephan Mohr, Marc Eixarch, Maximilian Amsler, Mervi J Mantsinen, and Luigi Genovese. Linear scaling dft calculations for large tungsten systems using an optimized local basis. *arXiv preprint arXiv:1711.10993*, 2017.










Diagonalization
d(x)=2.141*10^-7x^2 + 0.281*10^-11x^3
AQuA-FOE
f(x)=0.633 + 0.0005x
Time (Hours)
Number of Atoms
0
5
10
15
20
25
0
2000
4000
6000
8000
10000
12000