Analysis and Optimisation of the Basis Set Filtration Algorithm

Andrew Lawson. Submitted for the qualification of Doctor of Philosophy, from the faculty of Science, Agriculture and Engineering in Newcastle University in May 2014.

Abstract

The filtration algorithm has recently been introduced as a way of increasing the speed of *ab initio* modelling calculations using Cartesian Gaussian basis functions. It works by developing a novel set of basis functions which are constructed specifically for the system being modelled. It has been implemented in the *ab initio* density functional theory based modelling package AIMPRO. The standard filtration process is found to be accurate when the filtration radius is increased to at least 10 Bohr radii in silicon.

The standard filtration process uses all the basis functions centred on points inside a sphere centred on each atom in turn. By rejecting some of these functions (a *trimming* process), the filtration process can be speeded up, however there will be a resulting loss of accuracy. Three approaches to developing a filtered basis for an atom are considered, and compared. The most successful critertion for function trimming is found to be where functions are kept which exceed a threshold value on the surface of a sphere.

Structural optimisation using filtration produce accurate final structures, even when using parameters that give rise to poorly converged absolute energies. For the most time consuming elements of a calculation, a rapid filtration process is possible. However, very poor filtration thresholds introduce small inconsistencies between energies and forces, which can make optimisation difficult if algorithms are chosen that use both the energy and force. Algorithms that only use forces are implemented, and shown to be stable and produce accurate structures. This is further demonstrated using a new implementation of the Lanczos method for determining transition states. This is compared against the current AIMPRO method, the nudged elastic band. The new method is far superior in terms of speed, and offers greater stability towards the end of calcuations.

Acknowledgements

I would like to thank my supervisors Prof. Patrick Briddon and Dr. Jonathan Goss for their continuous and most excellent support throughout the entirety of my PhD, especially in the last few months. I would like to thank my wife Sarah, for putting up with me during my more stressed moments. Finally I would like to thank my sister and parents, without whom this would not have been possible. Special thanks goes to my mother who has helped me more than she will know.

Contents

List of Tables

List of Figures

Chapter 1

CHAPTER 1: INTRODUCTION

Ab-initio modelling permits the calculation of the properties of molecules and solids. It does not reply on experimental data, and so there is a huge variety of potential applications. The results are far more accurate than semi-empirical [18] or force field [1] methods. However the time taken to perform these calculations can be significant depending on the size of the system being modelled. The time required scales with the cube of the system size [34], so doubling the size of the system increases the time of a calculation by at least a factor of eight. The in-house modelling package AIMPRO [12, 38] is used for all developments and calculations, and exhibits this behaviour along with most other electronic structure codes. This effectively imposes a limit on the sizes of system that can be modelled, routinely 200-500 atoms at present.

The ability to model large systems is however particularly important when complex problems are being considered. For example, a point defect can usually (but not always) be modelled using a unit cell of 1000 atoms. This cannot be done so easily if the problem involves an imperfect interface between two materials (possibly with misfit dislocations), and the interest is on the behaviour of point defects interacting with this complex environment. When modelling imperfect interfaces, to incorporate the defects, strain and dislocations, large numbers of atoms are required before the structure shows a repeating unit. Larger systems also allow for more types of interface phenomena to be modelled, yielding results more appropriate for comparison to real life situations and more accurate calculations of properties.

The filtration algorithm, introduced by Rayson and Briddon in 2009 [56], produces filtered basis functions by analysing the basis functions for the structure that is being modelled. The resulting filtered functions reduce the time required for the Hamiltonian diagonalisation process, by a factor of at least 100, to a factor of 1000

and beyond. This process is the time dominant step for calculations for all but the smallest of systems. This thesis investigates the effect of this algorithm on the speed and accuracy of energy and structural optimisation calculations, for systems consisting of defects in unit cells of silicon. As filtration offers greater and greater efficiency savings for larger and larger systems, the system sizes investigated are at the small to medium scale, mainly between 64 and 216 atoms. Some results for larger systems, of up to 1000 atoms, are presented.

Alongside this, the implementation of a technique to identify transition states is detailed, the Lanczos method [43]. Results for the new method are compared against the existing Nudged Elastic Band method [16].

Chapter 2

GENERAL THEORY

In this chapter the scientific theories and methods that form the foundation of modern computational modelling will be outlined. The first topic outlines the system of units adopted for convenience when dealing with energies and lengths at atomic scales. Following this the Schrödinger equation, Born-Oppenheimer approximation, Density Functional Theory, basis sets, pseudopotentials, supercells, k-points and finally self-consistency are examined.

2.1 Atomic units

For computational modelling, the familiar SI units such as energies in Joules and lengths in metres are not appropriate, as most quantities would involve large negative exponents. The Hartree atomic units system (as opposed to the Rydberg atomic units system used in spectroscopy) define the electron mass m_e, the charge on a proton e, the Dirac constant $\frac{h}{2\pi}$ and Coulomb's constant $k_e = \frac{1}{4\pi\epsilon_0}$ as having a value of 1.

This results in changes to other familiar units. Two such units frequently encountered throughout this thesis are lengths, expressed in Bohrs/atomic units (a.u.) and energies, expressed in Hartrees (Ha). 1 a.u.=0.529 Å. 1 Ha=27.2114 eV.

The main advantage is that the majority of quantities of interest in an atomic or molecular environment are of order unity. Another advantage is the simplification of equations by removing the need for many constants, such as seen in equation (3).

2.2 The Schrödinger Equation and the many body problem

A molecule, or a basis (i.e. the repeating sub-unit) of a periodic structure such as a crystal, can be specified through the spatial coordinates of the nuclei and electrons.

The energy E_i of the i_{th} state of this system can then be found by solving the many-body Schrödinger equation [65], using an expression for the Hamiltonian operator $\hat{H}$, and solving for the wavefunction of the i_{th} state of the system, Ψ_i:

$$\hat{H}\Psi_i = E_i\Psi_i \tag{1}$$

The Hamiltonian represents the sum of the kinetic and potential energies of the system. It can be split into 5 terms for a molecular system:

$$H = T_e + T_n + V_{ee} + V_{ne} + V_{nn} \tag{2}$$

where T_e and T_n are the kinetic energy operators for electrons and nuclei respectively, and V_{ee}, V_{ne} and V_{nn} describe the interactions between electrons, electrons and nuclei, and nuclei respectively. If we represent the positions of the M nuclei by $R_1, R_2.....R_M$ and the N electrons by $r_1, r_2.....r_n$, (2) can be expanded to:

$$\begin{aligned}\hat{H} = &-\frac{1}{2}\sum_{i=1}^{N}\nabla_i^2 - \frac{1}{2}\sum_{k=1}^{M}\frac{1}{M_k}\nabla_k^2 + \sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{|r_i - r_j|} \\ &-\sum_{k=1}^{M}\sum_{i=1}^{N}\frac{Z_k}{|R_k - r_i|} + \sum_{k=1}^{M}\sum_{l>k}^{M}\frac{Z_k Z_l}{|R_k - R_l|}\end{aligned} \tag{3}$$

where M_k and Z_k are the mass and charge of nucleus k respectively. It is possible to solve this equation mathematically for systems containing one electron and one nucleus, such as a hydrogen atom or a He^+ ion [36]. For more complex systems, computational methods must be employed. However due to the large dimensionality of equation (3) it quickly becomes a practical impossibility for anything but the simplest, smallest system. To be useful, controlled approximations must be made. These approximations must also simplify (3), and reduce its dimensionality. The following sections in this chapter present such approximations utilised by AIMPRO, most of which are employed by commonly used computational software.

2.3 Born-Oppenheimer Approximation

As discussed above, the equations (1)-(3) present a considerable challenge. For a molecule as simple as ethane there are 24 nuclear $(R_1, R_2, R_3...)$ and 54 electronic $(r_1, r_2, r_3...)$ spatial coordinates, totalling 78 for the time independent Schrödinger equation. Born-Oppenheimer realised the equation could be approximated by separating the nuclear and electronic terms [9]. This is possible because of the high mass ratio between nuclei and electrons, and hence the kinetic energy of the electrons will be of roughly an equal ratio larger than the kinetic energy of the nuclei. The nuclei can hence be thought of as moving sufficiently slowly that the electrons are moving in the Coulombic potential instantaneously created by them. The minimum mass ratio is in hydrogen, and even here it is over 1836. If the wavefunction of the complete system is written as Ψ_T, we can approximately separate the electronic and nuclear wavefunctions as in 4:

$$\Psi_T(r, R) = \Psi_e(r; R)\Psi_n(R) \tag{4}$$

in which the total wavefunction is a product of a nuclear wavefunction $\Psi_n(R)$ and an electron wavefunction $\Psi_e(r)$. This allows us to develop an electronic only version of the Schrödinger equation, and to ignore the motion of the nuclei, and the associated kinetic energy term, $\hat{T}_n$. This gives us an equation as a function of r, with R as parameters:

$$\hat{H}_e\Psi_e(r_1, r_2...) = E_e\Psi_e(r_1, r_2...) \tag{5}$$

$$\hat{H}_e = -\frac{1}{2}\sum_{i=1}^{N}\nabla_i^2 + \sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{|r_i - r_j|} - \sum_{k=1}^{M}\sum_{i=1}^{N}\frac{Z_k}{|R_k - r_i|} \tag{6}$$

This is solved for a fixed set of nuclear coordinates, R_k, and then the nuclear energy term $\hat{T}_n$ is reintroduced, and the nuclear Schrödinger equation solved to give the total energy of the system:

$$\hat{H}_t\Psi_t = E_t\Psi_t \tag{7}$$

$$[T_n + E_e(R)]\Psi_t = E_t\Psi_t \tag{8}$$

2.4 Density Functional Theory

2.4.1 Overview

Density functional theory (DFT) is an *ab-initio* quantum mechanical method for calculating an approximation to the solution of the Schrödinger equation for a system of one or more atoms. Atoms, molecules, liquids and crystals can be studied using DFT.

The fundamental quantity in this theory is the electron density, a function that depends only on the three spatial coordinates. It is demonstrated that all ground state properties of the system can be calculated as functionals of this quantity, for example to get the total ground energy of a system. This function of a function is referred to as a functional, so in DFT a functional of the electron density, hence the name. Its usefulness as a computational modelling method can be quickly seen by comparing the number of variables it uses to that in a wavefunction based approach such as Hartree-Fock [23, 28, 63].

In Hartree-Fock, the energy levels for a system of N electrons are calculated through the use of a Slater determinant [61, 62] of spin-orbitals (one for each electron). As this $N \times N$ determinant can be generated from just the main diagonal (or any row/column) there are N terms, each with 3 spatial and 1 spin variable, or $4N$ variables in total.

In DFT the energy is a functional of the electron density, which depends on the familiar 3 Euclidean spatial coordinates. This does not scale with the size of the system one is attempting to model - for 10 or 1000 electrons the problem has been reduced to the evaluation of a functional of a quantity, whose input argument has only 3 degrees of freedom.

There is a problem however. The exact functional that gives the ground state energy from the electron density is currently unknown. However, despite this drawback, the use of DFT with approximate functionals has proved to be a successful and efficient method for calculating the properties of atomic systems.

2.4.2 Derivation

The emergence of DFT as we know it today started with the publication of the 1964 paper, *Inhomogenous Electron Gas by Hohenberg and Kohn*. One of the central concepts of DFT was introduced early on in this paper, that a particular electron density $n(r)$ corresponds to a unique external potential $V(r)$. There is thus a one to one relationship in between the electron density function $n(r)$, and the external potential $V(r)$. Here a proof is presented using the method of *reductio ad absurdum*, in which we start by proposing which two systems have the same electron density $n(r)$ and hence potential arising from electron-electron interactions, V_{ee}, but different external potentials V_1 and V_2. It follows closely the original proof from the 1964 Hohenberg/Kohn paper. Assume the two systems have external potentials V_1 and V_2, where V_1 cannot be expressed as $V_2 +$constant. In general, the electronic Hamiltonian may be written:

$$H = T + V + V_{\text{ee}} \tag{9}$$

For the two systems, only V is different, so we have:

$$\begin{aligned} H_1 =& T + V_1 + V_{\text{ee}} \\ H_2 =& T + V_2 + V_{\text{ee}} \end{aligned} \tag{10}$$

Hence

$$H_1 = H_2 - V_2 + V_1 \tag{11}$$

As we have different Hamiltonians, and $V_1 \neq V_2 + c$, we must have different Schrödinger equations, and hence two different ground state wavefunctions, Ψ_1 and

Ψ_2. The ground state energies for the two systems are:

$$\begin{aligned} E_1 &= \langle \Psi_1 \,|\, H_1 \,|\, \Psi_1 \rangle \\ E_2 &= \langle \Psi_2 \,|\, H_2 \,|\, \Psi_2 \rangle \end{aligned} \tag{12}$$

Due to the variational principle of quantum mechanics, the ground state wavefunction produces the lowest expectation value of energy for a particular Hamiltonian. This means we can write:

$$\begin{aligned} E_1 &= \langle \Psi_1 \,|\, H_1 \,|\, \Psi_1 \rangle < \langle \Psi_2 \,|\, H_1 \,|\, \Psi_2 \rangle \\ E_2 &= \langle \Psi_2 \,|\, H_2 \,|\, \Psi_2 \rangle < \langle \Psi_1 \,|\, H_2 \,|\, \Psi_1 \rangle \end{aligned} \tag{13}$$

Using:

$$\begin{aligned} H_1 &= H_2 - V_2 + V_1 \\ H_2 &= H_1 - V_1 + V_2 \end{aligned} \tag{14}$$

we get:

$$\begin{aligned} E_1 < \langle \Psi_2 \,|\, H_2 - V_2 + V_1 \,|\, \Psi_2 \rangle &= E_2 + \int (V_1(r) - V_2(r)\,)\, n(r)\, dr = E_2 + x \\ E_2 < \langle \Psi_1 \,|\, H_1 - V_1 + V_2 \,|\, \Psi_1 \rangle &= E_1 + \int (V_2(r) - V_1(r)\,)\, n(r)\, dr = E_1 - x \end{aligned} \tag{15}$$

where:

$$x = \int (V_1(r) - V_2(r)\,)\, n(r)\, dr \tag{16}$$

Adding these two together gives:

$$E_1 + E_2 < E_2 + E_1 \tag{17}$$

This is clearly a contradiction if the ground state is non-degenerate, so the original assumptions were incorrect, i.e. $V_1 \neq V_2 + c$.

Thus for a given density $n(r)$ there exists a specific potential $V(r)$ and vice versa. This is known as the Hohenberg-Kohn theorem. If we take the operator $\hat{F}$ defined by:

$$\hat{F} = \hat{T} + \hat{V}_{\text{ee}} \tag{18}$$

we can, for a system of N electrons with a density function $n(r)$, write the Hamiltonian as:

$$\hat{H} = \hat{F} + \hat{V}_{\text{ext}} \tag{19}$$

with both of the right hand operators being specified fully by the system size N and external potential (or equivalently of course, by the electron density function). This in turn shows that the associated system wavefunction Ψ must also be determined uniquely by the density. This allows use to write the energy of the system as a functional of the density. Take first a system in the ground state, with a ground state density, wavefunction and energy:

$$E_{\text{gs}} = E[n(r)_{\text{gs}}] = \langle \Psi_{\text{gs}} \,|\, H_{\text{gs}} \,|\, \Psi_{\text{gs}} \rangle \tag{20}$$

If we have a different wave function, Ψ_x, corresponding to a different density $n_x(r)$using the variational principle of quantum mechanics, we have:

$$E_{\text{gs}} = E[n(r)_{\text{gs}}] = \langle \Psi_{\text{gs}} \,|\, H_{\text{gs}} \,|\, \Psi_{\text{gs}} \rangle < \langle \Psi_x \,|\, H_{\text{gs}} \,|\, \Psi_x \rangle = E[n_x(r)] \tag{21}$$

Thus the density that minimises the energy is the ground state density. This is the second theorem of Hohenberg and Kohn.

The problem now is to find $E[n(r)]$. We can write

$$E[n(r)] = \int n(r) V_{\text{ext}} dr + F[n(r)] \tag{22}$$

The classical electron-electron interaction term can be moved out of the functional F[n(r)] to create a new functional G[n(r)]:

$$E[n(r)] = \int n(r) V_{\text{ext}} dr + \frac{1}{2} \int \frac{n(r) n(r')}{|r - r'|} dr dr' + G[n(r)] \tag{23}$$

The problem is now to find an expression for $G[n(r)]$. A much earlier attempt by Thomas-Fermi [21, 66] produced reasonable energies of molecules in isolation, but when these energies were subtracted to provide a description of bonding in molecules the method proved inadequate [64]. A big step was taken towards this by Kohn and

Sham in the 1965 paper *Self-Consistent Equations Including Exchange and Correlation Effects* [40]. Kohn-Sham proposed an approach that allowed an expression for the energy functional to be written down that yielded energies and differences of energies close to experiment. The idea was to introduce a fictional system of a non-interacting gas of N electrons, with an electron density $n(r)$ the same as the real system. The expectation of the kinetic energy of the true system can be written in terms of the kinetic energy of this fictitious non-interacting system $T_{\mathrm{s}}[n(r)]$ as:

$$\left\langle \Psi \left| \hat{T} \right| \Psi \right\rangle = T_{\mathrm{s}}[n(r)] + \Delta T \tag{24}$$

with the difference from the interacting system represented by ΔT. The hope is that this 2nd term is small. This allows us to rewrite $G[n(r)]$ as

$$G[n(r)] = T_{\mathrm{s}}[n(r)] + E_{xc}[n(r)] \tag{25}$$

This would give the electron density of the system, $n(r)$ as:

$$n(r) = \sum_{\lambda=1}^{N} |\Psi_\lambda(r)|^2 \tag{26}$$

where the Ψ_k's are the states of this non-interacting system. Then we may write:

$$T_{\mathrm{s}}[n(r)] = \sum_{\lambda=1}^{N} \int \Psi_\lambda^*(-\frac{1}{2}\nabla^2)\Psi_\lambda dr \tag{27}$$

$$\text{and } E[n(r)] = T_{\mathrm{s}}[n(r)] + \int n(r)V_{\mathrm{ext}}(r)dr + \frac{1}{2}\int \frac{n(r)n(r')}{|r-r'|}drdr' + E_{\mathrm{xc}}[n(r)] \tag{28}$$

$E_{\mathrm{xc}}[n(r)]$ contains the difference in the kinetic energy between the real and non-interacting system, and all the contributions to the electron-electron term and the Hartree energy. The real electron-electron energy will differ from the Hartree energy because electrons do interact, and will tend to stay away from each other, reducing the energy of a system. This is the correlation part of E_{xc}. The Pauli exclusion principle has not been factored in, and this will also affect the energy of the system by keeping electrons of parallel spin further away from each other. This is the exchange part of E_{xc}. A significant effort has been done to obtain an expression for $E_{\mathrm{xc}}[n(r)]$, which is not a trivial problem, and involves further approximations.

2.4.3 Estimating $E_{\mathbf{xc}}$: The Local-Density Approximation (LDA)

The LDA [49, 67] provides an approximation to E_{xc}, using the homogeneous electron gas (HEG) model as a starting point. If $n(r)$ does not vary too rapidly, one may write:

$$E_{\mathrm{xc}}[n(r)] = \int n(r) E_{\mathrm{xc}}(n) dr \tag{29}$$

where $E_{\mathrm{xc}}(n)$ is the exchange correlation energy per electron of a HEG of density n. This is the LDA. E_{xc} is split into separate terms for the exchange and correlation terms:

$$E_{\mathrm{xc}} = E_{\mathrm{x}}^{\mathrm{LDA}} + E_{\mathrm{c}}^{\mathrm{LDA}} \tag{30}$$

The energy density is calculated locally at points on a grid, and if $n(r)$ varies slowly, it can be shown that for a HEG:

$$E_{\mathrm{x}}^{\mathrm{LDA}} \propto n(r)^{\frac{1}{3}} \tag{31}$$

$$V_{\mathrm{x}}^{\mathrm{LDA}} \propto \int n(r)^{\frac{4}{3}} dr \tag{32}$$

This leaves $E_{\mathrm{c}}^{\mathrm{LDA}}$. Analytic expressions are available for high and low density limits, which correspond to infinitely weak or strong correlations respectively. Quantum Monte-Carlo simulations have been performed [15]. By interpolating between these accurate results, using the high/low density information, and using theorems about the limits of the functional forms of $E_{\mathrm{c}}^{\mathrm{LDA}}$, approximations for $E_{\mathrm{c}}^{\mathrm{LDA}}$ can be made, such as described by Perdew-Zunger in 1981 [50], and Perdew-Wang in 1992 (PW92) [49].

There are also other functionals, such as the GGA [6, 27, 37, 48, 49] where the first derivative of the electron density $n(r)$ is used. The calculations in this thesis use the LDA PW92 functional.

2.5 Basis sets

We now turn to the solutions of the Kohn-Sham equation.

$$-\frac{1}{2}\nabla^2\psi(r) + V(r)\psi(r) = \epsilon\psi(r) \tag{33}$$

The potential V(r) includes the external potential due to atomic nuclei, together with the interactions between electrons. It is a somewhat surprising but important result that this complex interaction can be represented in a simple potential, V(r). For computational calculations these solutions need to be discretised to be represented on digital computers. This is done by expressing the solution as a summation of functions which can be described and stored as coefficients, c_i.

$$\psi(r) = \sum_i c_i\phi_i(r) \tag{34}$$

In AIMPRO these functions are Gaussians. Gaussians are chosen as the calculation of matrix elements produces integrals that are much easier to compute than is the case with other functions, such as Slater-type orbitals [25]. Slater-type orbitals follow the true wavefunction more closely than Gaussians, having the same rate of decay (e^{-ar} compared to e^{-ar^2}). This means more Gaussians are required to model the true wavefunction. Although other types of functions may require fewer functions to be used, using more computationally efficient Gaussians is many times faster than using fewer less efficient functions.

The Gaussians themselves are positioned on each atom:

$$\phi_1(r) = e^{-\alpha_i(r-R_i)^2} \tag{35}$$

The exponents α_i determine the width of the Gaussian, a larger value meaning a narrower Gaussian. The centre of the Gaussian is located at R_i. The exponent parameters are difficult to determine, and are pre-calculated for every type of atom in the solid being modelled-. The parameters are determined by varying them and seeing which produces the lowest energy for each type of atom/solid. They are then

fixed and transferred to other related systems (for example parameters determined for bulk silicon are used for defects in silicon). The coefficients c_i are determined using the variational principle of quantum mechanics, i.e. they are varied until the energy is minimised. This is done as part of each calculation. In practice, for an atom such as silicon, four different exponents are used. For each exponent, there are almost always multiple basis functions, created by pre-multiplying the basic Gaussians to form Cartesian Gaussian Orbitals (CGO). They are akin to the s, p and d orbitals from atomic orbital theory. The s-type function is the basic Gaussian (35). There are 3 p-type CGOs shown in (36) created by multiplying by 3 pre-factors, one for each of the 3 Cartesian axes:

$$\phi_2(r) = (x - R_{ix})e^{-[\alpha_i(r-R_i)^2]} \tag{36a}$$

$$\phi_3(r) = (y - R_{iy})e^{-[\alpha_i(r-R_i)^2]} \tag{36b}$$

$$\phi_4(r) = (z - R_{iz})e^{-[\alpha_i(r-R_i)^2]} \tag{36c}$$

When the s and p CGOs are used they provide 4 independent functions sharing the exponent. This approach is extended further with the addition of 6 more functions, created by multiplying by combinations of 2 of the 3 pre-factors. This yields 5 d-type CGOs and an additional s-type CGO as shown in (37).

$$\phi_5(r) = (x - R_{ix})^2e^{-[\alpha_i(r-R_i)^2]} \tag{37a}$$

$$\phi_6(r) = (y - R_{iy})^2e^{-[\alpha_i(r-R_i)^2]} \tag{37b}$$

$$\phi_7(r) = (z - R_{iz})^2e^{-[\alpha_i(r-R_i)^2]} \tag{37c}$$

$$\phi_8(r) = (x - R_{ix})(y - R_{iy})e^{-[\alpha_i(r-R_i)^2]} \tag{37d}$$

$$\phi_9(r) = (x - R_{ix})(z - R_{iz})e^{-[\alpha_i(r-R_i)^2]} \tag{37e}$$

$$\phi_10(r) = (y - R_{iy})(z - R_{iz})e^{-[\alpha_i(r-R_i)^2]} \tag{37f}$$

When the s, p and d CGOs are used they provide 10 basis functions for that value of α. A common set of these basis functions, or basis set, involves ten functions for each of the two smallest exponents and four functions for each of the two largest

exponents. This means 28 basis functions per atom, and is referred to in this thesis as a ddpp basis set. Each letter corresponds to one exponent, with the smallest exponent listed first, the largest last. The majority of time spent in a large calculation is in a routine whose time requirement is proportional to the cube of the total number of basis functions. Filtration contracts these 28 functions per atom to a much smaller number, typically 4 for an atom like silicon; most calculations in this thesis are done using 4 contracted functions per atom. Another popular basis set for silicon uses 10 functions for each of the 4 exponents, i.e. a dddd basis set, with 40 functions per atom.

As stated above, CGOs have the advantage of fast computation over Slater-type orbitals. They are also flexible. Difficult elements, such as those with populated valence f-orbitals can have extra f-functions placed on them, without having to change the basis set for other atoms. Their rapid decay also aids in reducing the number of elements of the Hamiltonian matrix formed as part of the AIMPRO calculation.

2.5.1 Alternative basis sets - Plane waves

Plane waves are another type of basis set, and they highly effective for systems with periodic boundary conditions, such as the silicon crystals modelled in this thesis. They are effectively a Fourier series, with each basis function a term in the series. As these functions are periodic, they are a natural choice for modelling periodic systems. They have many useful properties.

1. They are systematically convergent. In the case of Gaussians, adding extra functions can lead to instability in the calculations due to small errors in the storage of double precision numbers causing a singular overlap matrix (zero valued eigenvalues). Plane waves do not suffer from this problem, are always orthogonal to each other, hence adding terms converges the result correctly.

2. To increase the accuracy of a calculation there is a single parameter that can be

changed, referred to as E_{cut}. This refers to the maximum energy of a plane wave, and quickly tells you the quality of a calculation. It is also easy to change. This compares to Gaussians where the process and quality involves many parameters, some of which can effect each other.

3. Gaussians are placed on atoms. Placing them elsewhere such as on bonds leads to problems, such as when bonds break. This means there is a nucleocentric bias in the basis set. Plane waves cover the entire unit cell removing this restriction.

They also however present some limitations.

1. Typically a very large number of functions are needed, leading to large memory requirements. AIMPRO using CGOs can model up to 4000 atoms on a desktop PC with 16GB of memory, a task currently unachievable using plane waves.

2. Although convergence is theoretically possible, in practice due to the large number of functions required it is rarely achieved. As with CGO based calculations, energy differences are obtained and converged, utilising the systematic nature of the errors.

3. Molecules cannot be modelled, unlike cluster calculations in AIMPRO.

4. In AIMPRO if one atom is added, such as an 1 atom of oxygen into 1000 atoms of silicon, the time required for the calculation will barely be affected as only of the order of 28-40 extra basis functions are required. In plane waves the addition of a single atom from the 2p (C, N, O, F) or 3d (Fe, Co, Ni, Cu, Zn...) series will require the addition of a large number of basis functions.

For both CGO and especially plane wave based calculations, the use of pseudopotentials is extremely useful in reducing the number of basis functions required to achieve an accurate calculation. We consider this next.

2.6 Pseudopotentials

After DFT, the concept and implementation of pseudopotentials [2,7,26,39] is the second most important approximation used in a modern modelling calculation. The idea is based on the notion that only valence electrons are involved in bonding/chemical reactions. The core electrons are so tightly bound that they take little part in the chemistry taking place [59].

Core states are more difficult than valence states because they vary rapidly, and because of the large Coulombic potential have a cusp at the nucleus. They also contain many nodes, the position of which are important for the accuracy of the calculation. They make the energy very large, so that even a small percentage error leads to a large overall error, even when calculating energy differences such as for formation energies. The potential they feel and the energies themselves are so large that a relativistic treatment is required. Also they have a knock-on effect on the valence states, causing them to oscillate. Rapid oscillations in the core, and consequently valence states require large amounts of Gaussian basis functions, dramatically increasing the length of a calculation. For plane wave based methods, this presents an insurmountable problem.

By replacing the $\frac{-Z}{r}$ potential with a pseudopotential $V^{\mathrm{ps}}(r)$ that is identical to the original potential beyond a certain distance from the nucleus, a cut-off radius, but simplified inside it, these problems can be removed. Most importantly, the number of basis functions required for an accurate calculation is reduced. Figure 1 shows a real potential and pseudopotential.

The development of pseudopotentials included two concepts that led to transferable pseudopotentials, i.e. potentials that can be developed for an atom such as carbon, and then used to give accurate calculations on systems such as diamond, graphite or hydrocarbons. Splitting the charge density and norm conservation [26].

Splitting the charge density ($n(r)$) into a valence ($n^{\mathrm{v}}(r)$) and core ($n^{\mathrm{c}}(r)$) charge

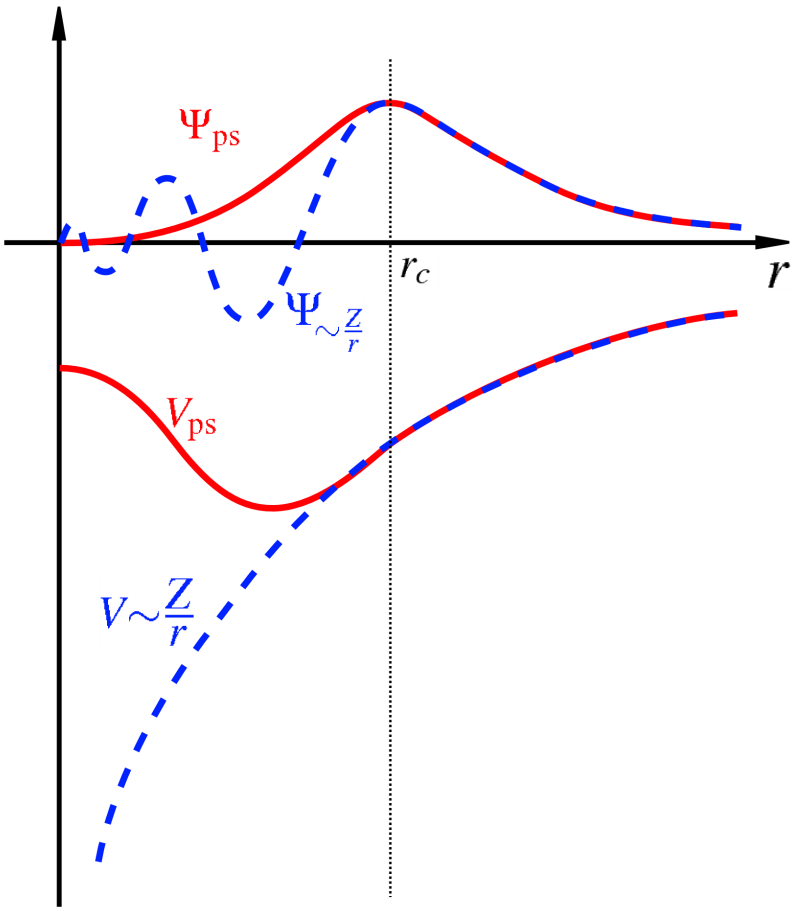


Figure 1: All electron wavefunction and potential in blue, and the pseudopotential and resulting wavefunction in red. Note that the true wavefunction has rapid oscillations near the core, and the pseudowavefunction is nodeless.

density (38) enables the removal of the valence potential from the pseudopotential (39), creating an ionic pseudopotential $V_{\text{ion}}^{\text{ps}}(r)$. Without this potentials would vary from system to system, and this method dramatically improves transferability. The density is split:

$$n(r) = n^{\text{v}}(r) + n^{\text{c}}(r) \tag{38}$$

so that the ionic-pseudopotential can be calculated as:

$$V_{\text{ion}}^{\text{ps}}(r) = V^{\text{ps}}(r) - \int \frac{n^{\text{v}}(r)dr}{|r - r'|} - V_{\text{xc}}[n^{\text{v}}(r)] \tag{39}$$

Before norm conservation the pseudopotential $V^{\text{ps}}(r)$ had to equal the true poten-

tial $V(r)$ beyond a certain distance from the nucleus of the atom. Norm conservation adds an additional requirement, that beyond this cut-off point the wavefunction solution under the pseudopotential matches the wavefunction solution under the real potential.

Determining which electrons to treat as valence can sometimes be tricky. For elements like carbon and silicon the choice is straightforward.

C Core: $1s^2$ Valence: $2s^2$ $2p^2$

Si Core: $1s^2$ $2s^2$ $2p^6$ Valence: $3s^2$ $3p^2$

However for some elements this is not as straightforward. Transition elements can have the outermost s and d shells lying close in energy to each other, and different reactions will involve different orbitals. For example in ZnSe, if the $3d$ electrons are treated as core electrons, the lattice constant is short by up to 10% when compared to experiment. When they are treated as valence electrons the difference to experiment is only 1% [41].

The pseudopotentials used in AIMPRO and in this thesis were developed by Hartwigsen, Goedecker and Hutter (1998) [29]. These authors tabulated all the parameters necessary for all the elements in the periodic table, so the user must simply state which pseudopotential is being used.

2.7 Supercells and clusters

Once a system (molecule(s), crystal...) has been decided upon, there are two main methods of modelling it. In a cluster calculation only one isolated copy of the system exists. In a supercell calculation the system becomes the unit cell, which is repeated in space, with periodic boundary conditions introduced to simulate an infinite crystal.

Creating a supercell from a number of primitive cells of the crystal, then creating a defect in this supercell, is a quick and useful way to model crystal defects. When using this approach to extract useable real world data, it must be ensured the supercell is large enough to prevent defects interacting with the equivalent defects in the neigh-

bouring unit cell. For the purposes of testing filtration this is largely irrelevant, as the focus is on filtration producing the same results as unfiltered calculations. However both small and large systems have been investigated for completeness.

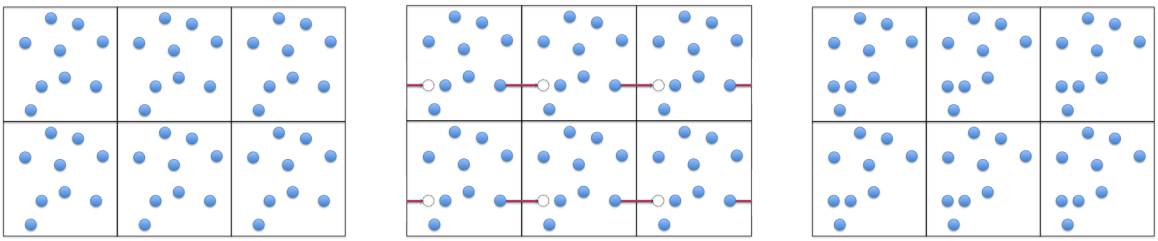


Figure 2: 6 unit cells in a supercell calculation, showing the effect of one atom moving into an adjacent cell.

In calculations involving atoms moving, such as structural optimisations and nudged elastic band (NEB) runs, when an atom moves out of one cell it appears on the opposite side of the cell. This is represented in figure 2.

In AIMPRO when performing a supercell calculation the lattice type and associated lattice parameters must be specified. Although it is possible to specify the lattice through 3 lattice basis vectors, almost without exception the lattice is specified by type. The lattice types are the 14 conventional 3-dimensional Bravais lattices. So for a cubic lattice (simple, body-centred or face-centred) only one parameter representing the length of a side of the lattice is required. For a simple tetragonal system it is two, one for the length of the depth/width, and another for the height.

When building a supercell for systems containing a defect a typical strategy is to take the 8 atom primitive cell and copy it n times in each direction. This leads to supercells containing 8n×n×n atoms. For n=2 the supercell contains 2×2×2×8=64 atoms. For n=3, 216 atoms. For n=4 512 atoms, and for n=5 1000 atoms. Defects can then be introduced into these large systems.

Supercell calculations also require the specification of k-points, which is described in the next section.

2.8 k-points

The first Brillouin zone is formed by taking the Wigner-Seitz cell in the reciprocal lattice. When performing calculations on periodic systems with the supercell method, most properties of the system such as the charge density, total energy or density of states require a Brillouin zone integration. In practice this can only be done by sampling at various points inside the Brillouin zone [4, 8]. There are a variety of ways to choose these points, but two methods have been employed for the calculations in this thesis.

The most basic method, Γ point sampling, involves one sampling point at the centre of the Brillouin zone. For increased accuracy more points are required. Monkhorst-Pack [45] defined an unbiased method where a grid of evenly spaced points, sharing the lattice symmetry, are placed inside the Brillouin zone. As more and more points, referred to as k-points, are used the answer will converge. For bulk silicon of 216 atoms or more a 2×2×2 grid provides reasonably converged results. This would be referred to as MP 2 2 2 sampling in the text. The number of k-points can be reduced using the symmetry of the reciprocal lattice. In the 2-atom FCC unit cell of bulk silicon using MP 2 2 2 sampling the number of k-points is reduced from 8 to 2.

Although there are an infinite number of Brillouin zones, the first Brillouin zone contains all the information necessary to fully describe the wavefunctions that are solutions to the Kohn-Sham equations. When reference is made to the Brillouin zone, it means specifically the first Brillouin zone.

The silicon calculations in this thesis use a simple cubic Bravais lattice, with all three sides of length a. The corresponding reciprocal lattice and Brillouin zone is also simple cubic, with all sides of length $\frac{2\pi}{a}$. In this case the reciprocal lattice will share the same symmetry as the lattice in real space. As the number of atoms used in the calculations increases, the Brillouin zone becomes smaller, and consequently fewer k-points are required to maintain a specific level of accuracy.

When energy differences are being calculated, such as for formation energies, the use of the same (or equivalent) k-points in all the systems typically leads to lower errors than expected, due to cancellation.

2.9 Self-consistency

Self-consistency ensures the resulting electron density minimises the total energy of the system. It consists of a repeated series of steps, in which the electron density $n(r)$ is changed using an iterative process. Each cycle of this process is referred to as a self-consistent field (SCF) iteration/cycle/step. Starting with iteration k=0, and an input density $n_k^{\text{in}}(r)$, you use the input density $n_k^{\text{in}}(r)$ to generate a potential, and then solve the Kohn-Sham equations. This will yield a new density, $n_k^{\text{out}}(r)$. The current and previous densities are then used to form a new input density.

$$n_{k+1}^{\text{in}}(r) = \alpha n_k^{\text{out}}(r) + (1-\alpha) n_k^{\text{in}}(r) \tag{40}$$

The value of α is small, usually 0.1 or above, and under 0.4. Larger values would lead to energy densities changing too much from iteration to iteration, and smaller values would take too many iterations to reach the state that minimised the energy. When the difference between $n_k^{\text{in}}(r)$ and $n_k^{\text{out}}(r)$ is smaller than a pre-defined tolerance value, $n_k^{\text{out}}(r)$ is accepted. Otherwise the iteration counter is increased, $k \to k+1$, and the process is repeated.

Details of the steps that move from the input potential to the energy density within an iteration of the SCF process can be found in section 3.2.

2.10 From Standard Theory to Filtration

This chapter has outlined the various bodies of theory that have come together to produce a standard AIMPRO calculation. This thesis concerns itself with both comparing the accuracy of, and improving the speed of, calculations using filtration alongside the

techniques presented in this chapter. The next chapter introduces the theory behind the filtration process, and explains both how the filtered functions are created and used.

Chapter 3

FILTRATION THEORY

3.1 Basis set filtration - the concept

Filtration is a method of reducing the number of basis functions in, and hence increasing the speed of, a calculation without having a significant effect on the accuracy. A typical silicon calculation, using Gaussian orbitals in AIMPRO, has 28 basis functions per atom, and standard calculations produce $28N_{\mathrm{atom}}$ (N_{atom} is the number of atoms in the system) states of increasing energy. Of these, only the first $2N_{\mathrm{atom}}$ states are occupied by the non-core electrons (all silicon calculations presented in this thesis use a pseudopotential with 4 non-core electrons, and 2 of these non-core electrons are able to occupy each state). The other states are not of direct interest, but are necessary outcomes for an accurate calculation. The extra freedom offered to the calculation through the large underlying basis set results in a higher accuracy.

The filtration method creates a fixed number of custom basis functions for each atom in the system, which we will term filtered functions. For each atom, these filtered functions are formed by analysing the basis functions on and close to the atom in question. Filtered functions are created that span the occupied states, and far fewer of the unoccupied states. This allows a much smaller basis set to be created that still produces accurate calculations. The Hamiltonian diagonalisation step in filtration calculations then uses a subspace Hamiltonian matrix, computed using filtered functions instead of the full Gaussian set. If N basis functions are present in the system in the unfiltered calculation, and the filtration step produces n filtered functions for the system, this diagonalisation step will be $(N/n)^3$ times faster. Typical values for unfiltered basis sets for silicon include 28 functions in a 'ddpp' basis set or 40 in a 'dddd' basis set (see section 2.5). A filtered basis set for silicon is usually

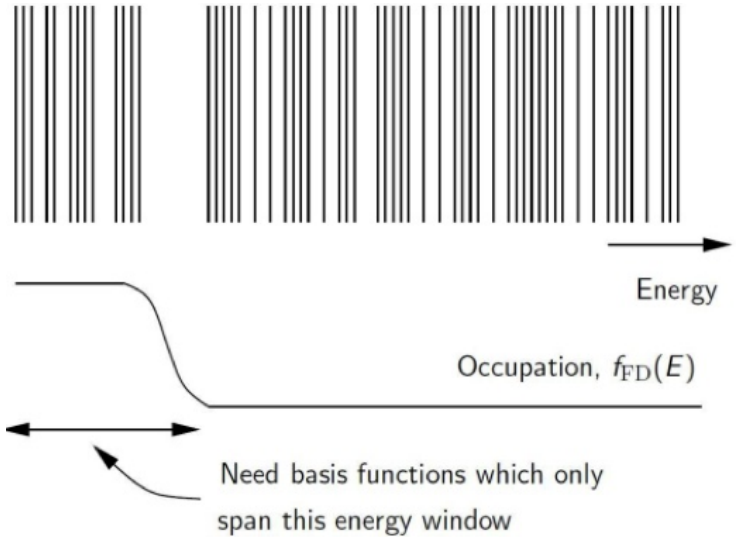


Figure 3: Energy levels produced by standard computational methods, and the much smaller energy window which is of interest.

chosen to include 4 filtered functions. This means the Hamiltonian diagonalisation step for a silicon calculation would be 343 (ddpp) or 1000 (dddd) times faster. This cubic relationship has been demonstrated already in previous work [55].

For clarity, the original basis functions and the space they span are referred to as the primitive basis set or primitive (basis) functions, and the primitive space. The filtered equivalents are filtered functions, and the subspace. Terms such as subspace Hamiltonian refer to the Hamiltonian after transformation from the primitive space to the subspace.

3.2 A conventional calculation

In a conventional calculation the Hamiltonian matrix H is created using the full set of basis functions $\phi_i(r)$ (41). H will be of size $N \times N$, where N is the total number

of basis functions in the system. We have

$$H_{ij} = \int \phi_i(r)(-\frac{1}{2}\nabla^2 + V(r))\phi_j(r)dr + \iint \phi_i(r)V^{nl}(r,r')\phi_j(r')drdr' \tag{41}$$

$V(r)$ is the local potential, with the second term using $V^{nl}(r,r')$ necessary when using non-local pseudopotentials, such as those developed by Hartwigsen, Goedecker and Hutter (1998) [29], used extensively in the calculations in this thesis. The overlap matrix is then calculated, again using the full set of basis functions.

$$S_{ij} = \int \phi_i(r)\phi_j(r)dr \tag{42}$$

H and S are used to form the generalised eigenvector problem (43).

$$Hc_\alpha = \lambda_\alpha S c_\alpha \tag{43}$$

If the vectors c_α are combined to form the columns of a matrix c, and the eigenvalues λ_α to form the non-zero elements of a diagonal matrix Λ, (43) can be rewritten as

$$Hc = Sc\Lambda \tag{44}$$

When solved, this yields a matrix c, which is used to produce the density matrix b

$$b = 2cf_{\mathrm{FD}}(\Lambda)c^T \tag{45}$$

where f_{FD} represents the Fermi-Dirac function, where the factor of 2 accounts for spin in a restricted calculation.

The matrix b in turn gives us the electron density $n(r)$.

$$n(r) = \sum_{i,j=1}^{N} b_{ij}\phi_i(r)\phi_j(r) \tag{46}$$

$n(r)$ is then used to create a new input $n(r)$ (as outlined in section 2.9), and hence a new $V(r)$, and (41)-(46) cycled through until the self-consistency criteria is/are achieved, and hence producing a self consistent $n(r)$.

3.3 Creating the filtered functions

The Kohn-Sham levels, $\psi_\lambda(r)$, are expressed as a sum of every primitive basis function, $\phi_i(r)$, of the whole system [56].

$$\psi_\lambda(r) = \sum_{i=1}^{N} c_{i\lambda}\phi_i(r) \tag{47}$$

By inverting this expression, the basis functions can be expressed as a sum of the Kohn-Sham levels.

$$\phi_i(r) = \sum_{\lambda} d_{i\lambda}\psi_\lambda(r) \tag{48}$$

λ takes values from 1 up to the number of electrons in the calculation (or half the number of electrons in spin-restricted calculations such as in this thesis), and $c_{i\lambda}$ and $d_{i\lambda}$ are coefficients, where

$$d_{i\lambda} = \int \psi_\lambda(r)\phi_i(r)\,dr = \sum_{j=1}^{N} S_{ij}c_{j\lambda} \tag{49}$$

with S_{ij} being the overlap matrix, defined in (42).

By inserting a Fermi-Dirac function into (48), a filtered function $\Phi_i(r)$ is created that only spans the energy window indicated by the occupation function $f_{FD}(E)$ in figure 3 [56]. $f(E_\lambda)$ is thus the occupancy at the KS state of energy E_λ.

$$\Phi_i(r) = \sum_{\lambda} f(E_\lambda)d_{i\lambda}\psi_\lambda(r) \tag{50}$$

By putting the expressions for $\psi_\lambda(r)$ and $d_{i\lambda}$ above into this we arrive at

$$\Phi_i(r) = \sum_{\lambda} f(E_\lambda) \sum_{j} S_{ij}c_{j\lambda} \sum_{i} c_{k\lambda}\phi_k(r) \tag{51}$$

$$\text{or } \Phi_i(r) = \sum_{j,k} b_{jk}S_{ij}\phi_k(r) \tag{52}$$

$$\text{where } b_{jk} = \sum_{\lambda} f(E_\lambda)c_{j\lambda}c_{k\lambda} \tag{53}$$

These equations can be combined to give

$$\Phi_i(r) = \sum_k K_{ki}\phi_k(r) \tag{54}$$

where K is termed the filtration matrix.

The trick now is to realise that $\Phi_i(r)$ is localised when a high temperature Fermi-Dirac function is used [56]. The argument for this is in section 3.6. This allows each $\Phi_i(r)$ to be constructed from the basis functions on atoms contained in a small sphere around the atom (see figure 4). The radius of this sphere is the filtration radius, R_{cut}, and can be altered but tends to produce converged results when it is chosen to contain around 30 atoms. The calculation of $\Phi_i(r)$ has to be done for each atom in the system, but as the sphere does not increase in size with the overall system, this leads to $O(N)$ scaling [56]. The time dominant step in a filtered basis calculation still has $O(N^3)$ scaling, so as the system sizes increases, the filtration step itself will become less and less significant in terms of its contribution to the overall length of a calculation.

3.4 Creating the K matrix

This following process takes place for each atom in the system, one at a time. Firstly as described above and illustrated in figure 4 the basis functions within the radius R_{cut} around the atom are identified. The intersection of the rows and columns of the Hamiltonian H and overlap matrix S that correspond to these identified functions are kept, creating a smaller versions H' and S'. This is represented pictorially in figure 5, where 5 basis functions are kept from the possible 36 in a fictitious system.

H' and S' then form the generalised eigenvector problem

$$H'c' = S'c'\Lambda' \tag{55}$$

where Λ' is a diagonal matrix of the eigenvalues. The matrix c' is then used to

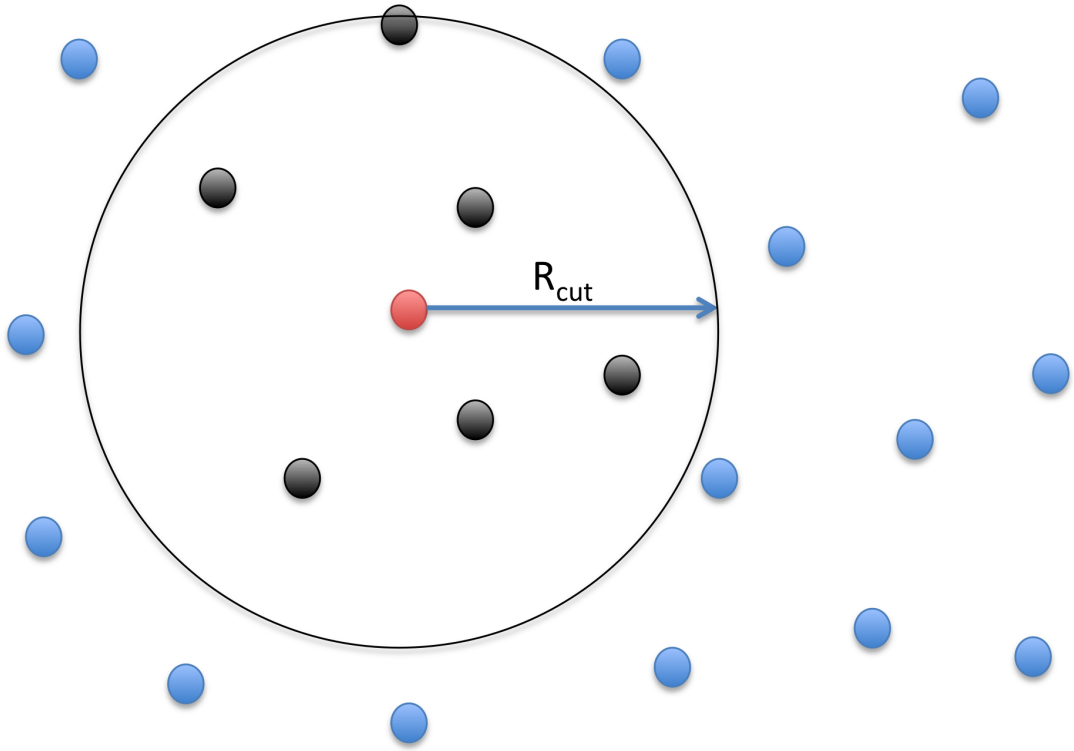


Figure 4: 2D representation of the process where Gaussians on atoms within a sphere of radius R_{cut} centred around the red atom, here coloured black, are used to create the filtered functions for the red atom. Gaussians on blue atoms are ignored. This process is repeated for each atom in the system, so each atom's sphere may contain different numbers of atoms.

calculate the density matrix b'_{pq}

$$b'_{pq} = \sum_{\lambda} f(E_\lambda) c'_{p\lambda} c'_{q\lambda} \tag{56}$$

where p and q label the (small, system size independent) number of functions retained during the filtration step [56].

The filtration matrix K is formed by creating rows from the related columns of b'_{pq} using (52) and (54).

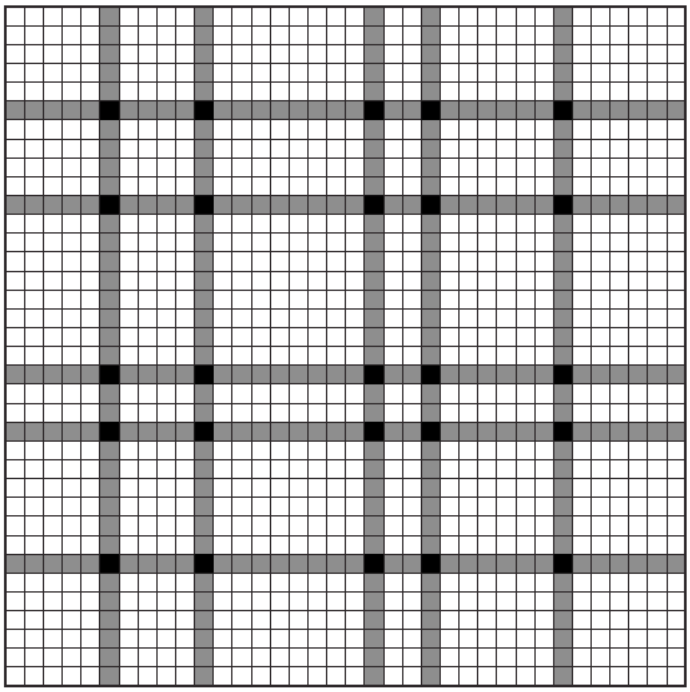


Figure 5: Representation of H or S for a system with a total of 36 basis functions. For a particular atom 5 basis functions are within the sphere of radius R_{cut}. The related rows and columns are shaded, and the black elements where they intersect are used to form H' and S'. The reduction in size of the two matrices, and hence of the generalised eigenvector problem they form is clear. [56]

3.5 Using the filtered functions

The procedure for a conventional calculation outlined in section 3.2 using the full set of basis functions is altered when using the filtered functions. After the Hamiltonian, overlap matrix and filtered functions are formed, a transformation to the subspace is performed. The density matrix is calculated in the subspace, after which a transformation back to the primitive space gives the full density matrix and hence electron density in the primitive space.

3.5.1 Primitive to subspace transformation

With the matrix K having been created, the subspace Hamiltonian H^{sub} and overlap matrix S^{sub} are required for the subspace generalised eigenproblem. We start with an expression for H^{sub}.

$$H_{IJ}^{\text{sub}} = \int \Phi_I(r)\hat{H}\Phi_J(r)dr = \sum_{i=1}^{N}\sum_{j=1}^{N} K_{iI}K_{jJ}\int \phi_i(r)\hat{H}\phi_j(r)dr \tag{57}$$

The terms integrated over r in this expression represent the element H_{ij} of the primitive Hamiltonian.

$$H_{IJ}^{\text{sub}} = \sum_{i=1}^{N}\sum_{j=1}^{N} K_{iI}K_{jJ}H_{ij} \tag{58}$$

$$H^{\text{sub}} = K^T H K \tag{59}$$

Following similar steps we find the result

$$S^{\text{sub}} = K^T S K \tag{60}$$

forming the subspace generalised eigenproblem

$$H^{\text{sub}}c^{\text{sub}} = S^{\text{sub}}c^{\text{sub}}\Lambda^{\text{sub}} \tag{61}$$

Once solved and the matrix c^{sub} is formed, the subspace density matrix b^{sub} is found using

$$b_{IJ}^{\text{sub}} = 2\sum_{\lambda} f(\Lambda_{\lambda\lambda}^{\text{sub}})c_{I\lambda}^{\text{sub}}c_{J\lambda}^{\text{sub}} \tag{62}$$

with the factor of two again accounting for spin.

With b_{IJ}^{sub} calculated, the next step is to transform back to the primitive space to obtain the primitive density matrix b and consequently the electron density $n(r)$.

3.5.2 Subspace to primitive transformation

Writing the total number of filtered functions in the system as n^{sub}, we can calculate the electron density $n(r)$ using

$$n(r) = \sum_{I,J=1}^{n^{\text{sub}}} b_{IJ}^{\text{sub}} \Phi_I(r)\Phi_J(r) = \sum_{I,J=1}^{n^{\text{sub}}} \sum_{i,j=1}^{N} K_{iI}K_{jJ} b_{IJ}^{\text{sub}} \phi_i(r)\phi_j(r) = \sum_{i,j=1}^{N} b_{ij}\phi_i(r)\phi_j(r) \tag{63}$$

as

$$b_{ij} = \sum_{I,J=1}^{n^{\text{sub}}} b_{IJ}^{\text{sub}} K_{iI}K_{jJ} \tag{64}$$

which is a matrix product of the form

$$b = K b^{\text{sub}} K^T \tag{65}$$

(46) can then be used as before to give the electron density $n(r)$, and hence the total energy. The strength of this approach is that once b is obtained in the primitive basis, the code required to compute the energy density does not need to be modified in any way.

3.6 Reasons why Filtered Functions are Localised

We start by defining a filtration operator $\hat{F}$ that transforms our N basis functions into n filtered functions ($N \geq n$).

$$\bar{\phi}_j(r) = \hat{F}\phi_i (i = 1...N, j = 1...n) \tag{66}$$

Assume we know the Kohn-Sham solutions, the ψ_ns. We can then write, following equations 51 - 54

$$\hat{F}\phi_i(r) = \int \sum_n f_n \psi_n(r)[\psi_n^*(r')\phi_i(r')dr'] \tag{67}$$

$$= \int \rho(r, r')\phi_i(r')dr' \tag{68}$$

where $f_n = f_{\mathrm{FD}}(E_n)$, and $\rho(r, r')$ is the charge density matrix:

$$\rho(r, r') = \sum_n f_n \psi_n(r) \psi_n^*(r') \tag{69}$$

To proceed, we look at some of the properties of the charge density matrix. Firstly as $|r - r'| \to \infty$, $\rho(r, r') \to 0$. In other words as the two points go further apart the density matrix tends to zero. Secondly for an insulator at any temperature (or any material with a band gap) or at the other end of the band gap scale, a metal at high temperature, $\rho(r, r') \to e^{-\alpha|r-r'|}$. For a metal however, as $T \to 0, \rho(r, r') \to \frac{1}{|r-r'|^\alpha}$ [44].

Referring back to (68) we note that $\phi_i(r')$ is a Gaussian and hence a localised function. Because of these properties of $\rho(r, r')$, if we choose a high temperature Fermi-Dirac function we ensure that $\rho(r, r')$ not only decreases faster than $\phi_i(r')$, but that it also is a localised function. This means we are guaranteed the integral in (68) will produce a localised function of r, and hence $\hat{F}\phi_i(r)$, the filtered functions are localised. The degree of localisation is dependent amongst other things, on the value of the exponent of the basis functions α, and on the temperature of the Fermi-Dirac function. The higher the temperature is set the more localised the filtered functions are, but conversely the more of the energy window that they span. It should be noted here that the temperature used in the filtration process is unrelated to the temperature used in the main calculation (equation 45). The filtration temperature is chosen to be sufficiently high to guarantee a well localised ρ.

The discussion in this section is not sufficient to allow filtered functions to be generated as they rely on knowledge of the Kohn-Sham solutions, the ψ_i's which are required to calculate $\rho(r, r')$. However only a knowledge of $\rho(r, r')$ for a small region of the system is required. Thus $\rho(r, r')$ is calculated for a subsystem - this is the process involving R_{cut}. A higher temperature Fermi-Dirac function reduces the size of the subsystem required for a set level of accuracy, but may require more filtered functions to be calculated and used in the calculation in the filtered subspace. A lower temperature Fermi-Dirac function will span a smaller energy window and thus lead to a smaller H' in the filtered subspace, and thus a much faster calculation. However it

will also lead to less localised filtered functions, increasing the value of the parameter R_{cut} required for an accurate calculation.

3.7 The Fermi-Dirac function in Filtration

The Fermi-Dirac function $f(E_\lambda)$ as illustrated in figure 3 in section 3.1 has two main parameters - the temperature of the function kT and the chemical potential E_f. The chemical potential determines at what energy the Fermi-Dirac function reaches half its maximum value. The temperature controls how fast the function drops off, with a higher temperature giving a slower drop off. Hence a higher temperature means more energy levels are spanned by the filtered functions, but a high temperature function is required to ensure the filtered functions that are created are localised. If the temperature of the function is not high enough, a large value of R_{cut} is required for an accurate calculation [56].

It is possible to specify these parameters in a calculation, or to allow AIMPRO to optimise these values automatically, based on the value of R_{cut} and the number of filtered functions which are created and used. Both methods are used in the calculations presented in this thesis, with details of which method was used is stated for each set of results. When specifying a filtration temperature for silicon, a value of between 2-3 eV is typically used.

The temperature of the Fermi-Dirac function used in filtration is not to be confused with the temperature of the system used to populate the Kohn-Sham levels, which will usually be much lower, typically 0.01 eV for a metallic system.

Having outlined the theoretical background behind the filtration process, the next chapter investigates the performance of filtration, and the effect of changing the filtration parameters such as R_{cut}, when it is applied to the calculation of energies of defects in semiconductors.

Chapter 4

ENERGY CALCULATIONS USING FILTRATION

The previous two chapters introduced the most significant theories and approximations used in a CGO *ab-initio* calculation, the theory behind the filtration process, and an outline of how it is implemented within the AIMPRO code. In this chapter filtration is applied to calculate formation energies (FEs) for various defects in silicon. The accuracy of these AIMPRO calculations is compared to the results for corresponding AIMPRO calculations performed without filtration. There are two types of filtration calculation, the previously published (standard) process [56] introduced in the previous chapter, and another in which an additional step is added to reduce the time taken to produce the filtered functions. Firstly a recap of the standard method is outlined, then the new method involving the additional step is described. The results for both methods are then presented.

4.1 Filtration Method 1 - Standard Filtration

This section commences with a summary of the SF process. This is followed by looking at the effect of changing the parameter R_{cut} on the speed of a SCF iteration, and on the overall calculation time for unit cells of bulk silicon of increasing size. After this, the details of a proposed new step, and the anticipated differences to the speed and accuracy of the calculations are outlined. This extra step also has a parameter τ, which like R_{cut}, can be adjusted, and which should give more accurate but slower calculations as it increases.

4.1.1 Recap of Standard Filtration Process

The filtration process introduced in the previous chapter can be summarised as follows:

1. Start with a system of atoms, with each atom having a number of primitive basis functions centred on it.

2. For each atom in turn, capture the basis functions centred on the atom itself, and on all atoms contained in a sphere of radius R_{cut} around it.

3. Use these functions to create filtered functions for the atom in question.

4. Repeat steps 2 and 3 for every atom in the system, creating custom filtered functions for each atom.

5. Using these custom filtered functions, calculate the properties of the system required. This chapter will investigate energies, the next chapter energies and forces.

Throughout this thesis calculations performed without filtration will be referred to as NF (no filtration) calculations. Calculations performed with the filtration process outlined above will be referred to as SF (standard filtration) calculations. Filtered calculations performed with the extra step (mentioned above, and described in detail in section 4.2) will be referred to as AF (advanced filtration).

4.1.2 Effect of R_{cut} on calculation times

When analysing calculation times, there are two obvious choices for what is measured from a timing point of view. These are the total time taken for the calculation, and the time taken per self-consistent field (SCF) iteration. Each has its merits. A user is simply interested in the total time taken, but this doesn't allow for scientific analysis, due to the large number of steps in which filtration and Hamiltonian matrix calculations play no part.

The time taken per SCF iteration is more useful. It contains the part of AIMPRO that filtration speeds up, the diagonalisation step. This is true for all calculations. It does include some other tasks, but has the advantage of being easily extracted from the AIMPRO output files, and as system size gets bigger the diagonalisation step takes up a larger and larger percentage of the total SCF time. The fixed portion of the SCF time can also easily be differentiated from the variable part when a list of results is available, such as seen in table 21. Another drawback is the danger of comparisons between two runs, where the computational power differs, be it through a system wide slowdown, or differing CPU power or number of nodes being used. These can largely be eliminated by ensuring all runs for comparison take place in similar environments, and by looking at the timings for parts of the run filtration doesn't affect. These timings should theoretically be the same, as they are doing the same calculation. By looking at the variation here, one can get a sense of the inherent randomness of timings, whether there was a problem with one of the runs, or a hardware disparity between them. This has been checked for all SCF time data in this thesis and will not be further referenced in the text.

In a NF calculation, the SCF process is dominated by the Hamiltonian diagonalisation step [30], especially for large systems, and systems with a fine k-point mesh. In a SF calculation there are two main processes, the filtration step (where the filtered functions are created) and the subspace Hamiltonian diagonalisation. The cubic relationship to the number of atoms for the diagonalisation step, contrasts with the time required for the filtration step being linearly proportional to the number of atoms in a system. So as the system size increases, the saving in the Hamiltonian diagonalisation step increasingly outweighs the increase in the time required for the filtration step. Conversely for systems of a certain size and below, using filtration will increase the total time, as the filtration step will outweigh the savings in the diagonalisation step. By reducing the time taken for the filtration step, the system size for which this happens can be reduced. Also for small to medium sized systems (hundreds of

atoms) significant time savings can be made. How to do this without significantly affecting the accuracy of the calculation forms the main part of this thesis. For SF calculations, the only way to do this is by reducing the parameter R_{cut}.

The next two chapters focus entirely on defects in unit cells of silicon, where most, if not all, atoms are silicon, and the total number of atoms in the system deviates from the corresponding bulk system by only 1-3 atoms. This means theoretically, by changing R_{cut} for bulk silicon, the effects on the numbers of functions kept, and hence time taken by SCF steps will be broadly the same as seen for the defect systems. Also a determination of the crossover point where filtration makes a calculation faster as a function of R_{cut} can be done for bulk silicon, and the results expected to apply to the defect systems used in the next two chapters.

It is useful to see the link between the number of functions presented to the filtration process, and the parameter R_{cut}, and this is presented in table 1. The number of atoms and hence functions, have an approximately cubic relationship with the value of R_{cut}, due to the increasing volume of the sphere. Values of R_{cut} of 10-12 a.u. are typical values for a silicon based calculation.

Table 2 shows the average SCF times for an energy calculation of bulk silicon systems of increasing size, for various values of R_{cut}. It is unusual to run a filtration calculation with an R_{cut} of greater than 12 a.u. so results for greater values are not presented. As the system size increases the average SCF time for a NF calculation increases at a much faster rate than for the SF calculations. A 1000 atom calculation's SCF time is 6.6 times longer than for a 512 atom calculation, not far off the value of $(1000/512)^3 = 7.5$ for a N^3 relationship. For system sizes of 512 atoms or more calculations are faster when using filtration. As the system size becomes lower than this, increasingly smaller values of R_{cut} are required to observe faster SCF iterations than a NF calculation. For 1000 atom calculations, SCF iterations are roughly 7 times faster than when using SF, even when R_{cut}=12 a.u. If an R_{cut} of 10 a.u. is used, this rises to a factor of 19.

Table 1: Number of atoms (N_{atom}), and functions (N_{keep}), inside a sphere of radius R_{cut} centred on an atom, for bulk silicon with lattice parameter 10.24 a.u. using a ddpp basis set.

R_{cut} (a.u.)	N_{atom}	N_{keep}
1-4	1	28
5-7	5	140
8	17	476
9-10	29	812
11	35	980
12	47	1316
13	71	1988
14	87	2436
15	99	2772
16	123	3444

Table 2: Average SCF times for bulk silicon energy calculations. For each system size, results for SF calculations for varying values of R_{cut} and the corresponding NF calculation are shown.

R_{cut} (a.u.)	Average SCF time (s)			
	64 atoms	216 atoms	512 atoms	1000 atoms
5	8	28	82	215
8	13	50	136	345
10	31	115	282	638
12	98	341	841	1735
(NF)	11	157	1826	12091

In generating this table the same MP sampling has been used for all cell sizes. In fact it is often the case that MP grids with more points are needed for smaller systems than larger ones, as the Brillouin zone being sampled is correspondingly larger. The filtration step is however performed in real space, and therefore is done only once, independent of the sampling grid. The speed up generated by filtration is therefore multiplied by the number of k-points in these runs and hence will give a much greater performance boost. Especially when looking at metallic systems, this can provide an additional order of magnitude improvement for systems of less than 100 atoms.

One other point to observe in table 2 is the effect of increasing R_{cut} has on the SCF time. As more functions are presented to the filtration step, it takes longer and longer to produce the filtered functions. If it was possible to reduce the number of functions presented without affecting the accuracy of the resulting filtered functions it would be possible to decrease the time required for this step, and hence to lower the size of system for when filtration becomes effective. A method of achieving this is proposed here. The details of the idea, and its implementation into the AIMPRO code are detailed in the next section.

4.2 Filtration Method 2 - Advanced Filtration

As the system size increases, filtration becomes more and more effective at speeding up a calculation. The system size below which it is faster to use a NF calculation is dependent on the time taken by the step where the filtered functions are created from the primitive basis set. This step is dependent on the number of atoms in the system, and the number of primitive basis functions captured in the sphere of radius R_{cut}. To make the filtration process faster, a version of this step known as advanced filtration (AF) was developed, tested and implemented.

AF works by rejecting some functions captured in the sphere of radius R_{cut}. The original process selects functions based only on the position of the centre of the Gaussians. The first AF implementation, and the one used in this and the following chapter, uses the integral of a product of each Gaussian in turn with a trial Gaussian centred on the atom for which the filtered functions are being created. If this integral is greater than a parameter, the function is kept and presented to the filtration process. This has the effect of keeping Gaussian functions of all exponents if they are close to the central atom, but to keep only the more delocalised functions if they are further away. As the number of functions trimmed in the AF process increases, the calculation will deviate more and more from the corresponding SF calculation, and the filtration step will take less and less time.

Another way of viewing this would be to focus on accuracy rather than speed, and to use this overlap rather than R_{cut} as a way of selecting the functions. This may be simulated by using a very large R_{cut}, such that it will be the overlap factor rather than R_{cut} that determines the inclusion or otherwise of each function. It may be (and in fact it will be shown to be true) that 1000 functions chosen in this way will produce better filtered functions and hence a more accurate energy than 1000 functions produced using SF.

4.2.1 Theory Behind Advanced Filtration

In AF, each Gaussian basis function φ_j (centred at $\vec{R}_j$ with an exponent α_j) inside the sphere of radius R_{cut} is examined in turn. The expression (70) is calculated for each one. φ_i is a Gaussian centred on the central atom at $\vec{R}_i$, with an exponent α_i equal to the minimum exponent of the basis set. φ_i is referred to as the trial Gaussian. In the ddpp basis set used for silicon in this thesis, the Gaussians have four exponents of values 0.16145, 0.46343, 1.31473 and 3.75324, so the trial Gaussian φ_i will have an exponent $\alpha_i = 0.16145$.

$$\int \varphi_i \varphi_j d\vec{r} = \left(\frac{\pi}{\alpha_i + \alpha_j}\right)^{\frac{3}{2}} \exp\left[-\frac{\alpha_i \alpha_j}{\alpha_i + \alpha_j}(\vec{R}_i - \vec{R}_j)^2\right] \tag{70}$$

As $|\vec{R}_i - \vec{R}_j| \to \infty$ the value of this expression is controlled mainly by the exponent term. The prefactor only varies weakly based on the values of the exponents. For $\alpha_j = 0.16145$ the value of the prefactor is 30.9, for $\alpha_j = 3.75324$ the value is 3.9. This is the full range of the prefactor, roughly one order of magnitude, as it does not depend on the distance between the two Gaussians. In comparison, the exponent term will vary by many orders of magnitude. This means the prefactor can be largely ignored, and we use the exponent term to define a test for each Gaussian surrounding the central atom. Specifically we reject the surrounding Gaussian if

$$\exp\left(-\frac{\alpha_i \alpha_j}{\alpha_i + \alpha_j}(\vec{R}_i - \vec{R}_j)^2\right) < \exp(-\tau) \tag{71}$$

which can be simplified to:

$$\frac{\alpha_i \alpha_j}{\alpha_i + \alpha_j}(\vec{R}_i - \vec{R}_j)^2 > \tau \tag{72}$$

If we increase the distance between the two Gaussians, the value of the left hand side of (72) increases due to $(\vec{R}_i - \vec{R}_j)^2$. If we narrow either Gaussian by increasing its exponent, the value also increases due to $\frac{\alpha_i \alpha_j}{\alpha_i + \alpha_j}$. In both of these cases this corresponds to the integral of the product of the two Gaussians decreasing, the left hand side of (72) increasing, and the function only being kept for larger and larger values of τ. τ is

Table 3: The maximum useful setting for the AF parameter τ for silicon, at various values of R_{cut}.

R_{cut}	τ_{max}
6	5.6
7	7.6
8	9.9
9	12.5
10	15.5
11	18.7
12	22.3
13	26.2
14	30.3

therefore a parameter which can be increased to keep more and more of the functions captured within the sphere of radius R_{cut}. For a fixed value of R_{cut}, the maximum value of the left hand side of (72), and hence the maximum useful value of τ, can be calculated. We take a Gaussian on the edge of the sphere of radius R_{cut}, with the narrowest spread and hence largest exponent α_{max} overlapping with the trial Gaussian of exponent α_i. The value of τ_{max} above which no further functions will be included is then

$$\tau_{\text{max}} = \frac{\alpha_i \alpha_{\text{max}}}{\alpha_i + \alpha_{\text{max}}} (R_{\text{cut}})^2 \tag{73}$$

For silicon using the ddpp basis set in this thesis, the various values of τ_{max} for values of R_{cut} can be seen in table 3.

One final point to note about this technique is that is does not take into account the angular momentum of the basis function. A p or d function is treated as if it were an s function, that is to say the expressions seen in (36-37) have their prefactors removed, to look like the expression in (35). This is necessary, as p and d functions are highly

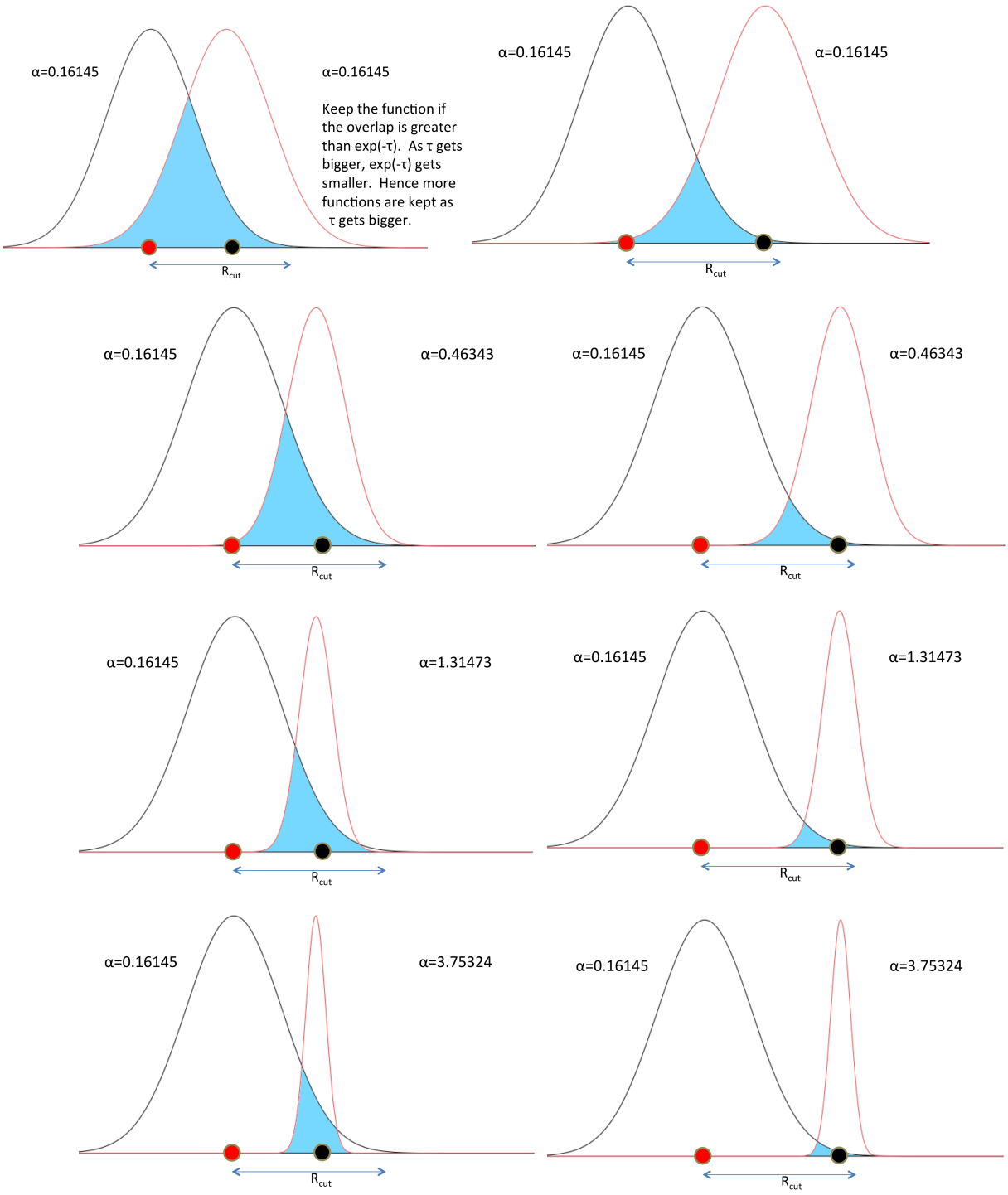


Figure 6: A schematic diagram to illustrate the effect of increasing both the distance between Gaussians and of varying exponent values, on the value of the overlap integral used in AF. The images on the right have the atoms further apart than those on the left, to show schematically how quickly the overlap drops with distance and exponent.

directional. If a p-type orbital points in a direction orthogonal to its displacement from the central atom, the overlap would always be zero, and these functions would then be ignored. This is illustrated in figure 7.

4.2.2 Implementation and Testing of AF

When the SF R_{cut} based process is finished for an atom, a reduced size Hamiltonian, overlap matrix and density matrix (H', S' and b'_{ij}) are produced. They are reduced in size further by AF, in a process illustrated in figure 8. These twice-reduced matrices are then used to populate the relevant rows and columns of the K matrix.

After this has been done for all the atoms, the full K matrix has been created. It will be the same size as it would have been using just SF. However the rows and columns relating to the basis functions that have been rejected will be populated with zeros, i.e. K will be sparser. It is obviously possible at this point to reduce the size of the K matrix by removing these rows and columns, or to simply create a smaller matrix in the first place. To do this would require some coding changes to create the smaller matrix in the first place, but more importantly require extensive coding changes and testing to the filtration section of AIMPRO. By keeping the K matrix the same size, these changes are avoided. The changes and testing would have required weeks if not months of effort to ensure no errors were introduced.

There is a small downside to this approach, in the transformations between the primitive space and subspace. As the K matrix is larger, this will take longer than necessary. It should be noted however that the interactions between atoms beyond a certain distance are very small. For this reason, sparse matrix algorithms are pre-built into the filtration process. The increased sparsity of the K matrix because of the AF procedure will increase the efficiency of these algorithms. Although it will never be as efficient as reducing the size of the K matrix, this novel approach allows most of the benefit to still be present without the huge effort of recoding the majority of the filtration code. The full impact of the reduced size of the N_{atom} eigenproblems is of

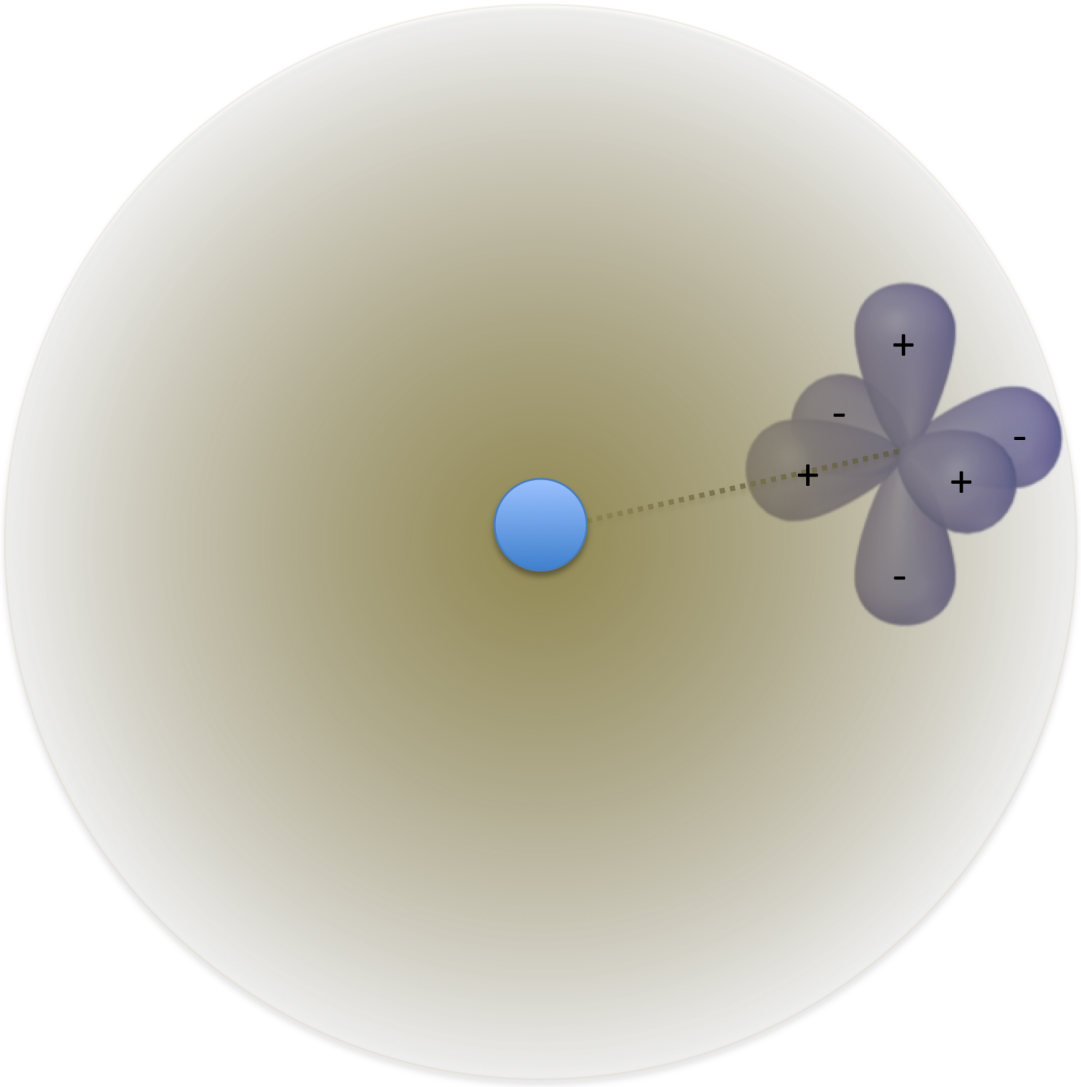


Figure 7: Illustration of the problems encountered if overlap was calculated using CGOs other than s-type. The central blue atom has a trial Gaussian (always s-type) represented by the brown sphere. If the angular momentum of the surrounding CGOs were taken into account, such as in the 3 orthogonal purple p-type CGOs shown here, the use of the full overlap (including angular variation) in AF would only include the p-type orbital pointing towards the central atom. Hence all surrounding CGOs are treated as s-type when calculating whether or not they are rejected in the AF process.

course unaffected, which is the main time saving feature of AF.

This way only two procedures were required to be developed that sit around the existing code. They fit around the existing code as follows:

1. Procedure 1 identifies the primitive functions that have failed the AF test, and creates the reduced size matrices H'', S'' and b''_{ij} (as in steps C-E in figure 8).

2. H'', S'' and b''_{ij} are passed to the filtration algorithm.

3. The filtration process produces the columns of the filtration matrix K related to the central atom.

4. H' and S' are not required any more, but b'_{ij} is. It is thus recreated in procedure 2 from b''_{ij} by expanding into a matrix the size of b'_{ij}, filling the gaps with zeros.

5. This process is repeated for every atom in the system, after which the full K matrix is available.

4.2.3 Effect of τ on Calculation Times

As with R_{cut} in table 2, it is useful to see the effect of the AF parameter τ on the time taken by a SCF step, and how many functions are kept. This was carried out on a system of 216 atoms of bulk silicon, using R_{cut}=10 and 12 a.u. and Γ point sampling. AF was applied for values of τ of 6, 8 and 10. The results are in tables 4 and 5. It can be seen τ has a dramatic effect on the time taken for a SCF iteration. For example cutting the number of functions by just 20% halves the average SCF time for an R_{cut} of 10 a.u.. Cutting the number of functions by 25% obtains a similar speed increase for AF calculations with an R_{cut} of 12 a.u. What must now be established is the accuracy of calculations using various values of R_{cut} and τ. This is the focus of the next section.

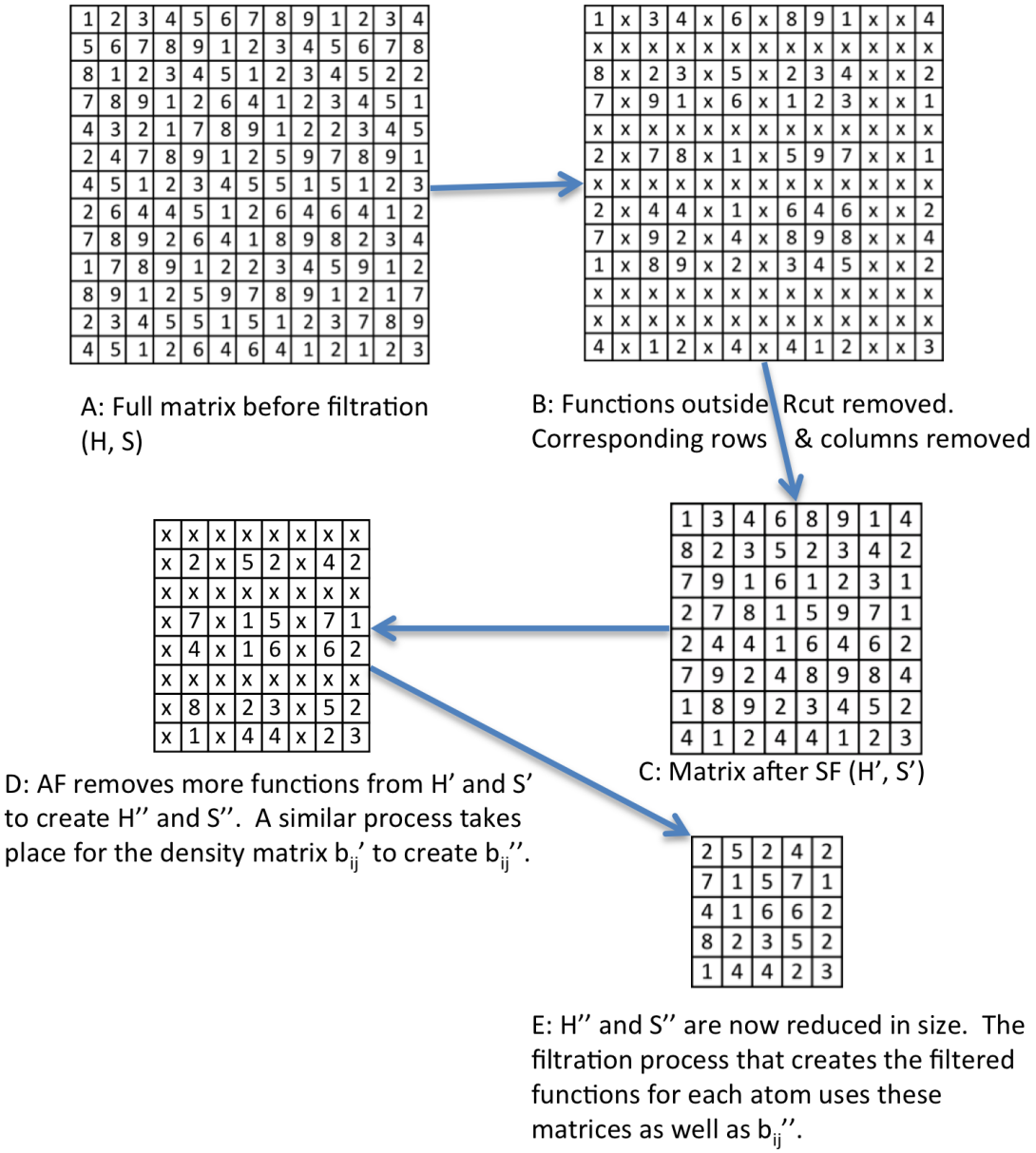


Figure 8: Illustration of SF and AF reducing the size of the Hamiltonian or overlap matrix (H' and S'). A shows the full size matrix, which is reduced to matrix C in the SF process. AF then further reduces this, creating matrix E. A similar process operates on the density matrix b'_{ij}, the process being slightly different as b'_{ij} and b''_{ij} are not square matrices.

Table 4: The effect of changing τ on average SCF time and number of functions presented to the filtration process. The system is 216 atoms of bulk silicon, with R_{cut}=10 a.u..

Filtration type	N_{keep}	Average SCF time (s)
AF, $\tau = 6$	380	41
AF, $\tau = 8$	584	58
AF, $\tau = 10$	716	88
SF	812	115
NF	-	157

Table 5: The effect of changing τ on average SCF time and number of functions presented to the filtration process. The system is 216 atoms of bulk silicon, with R_{cut}=12 a.u..

Filtration type	N_{keep}	Average SCF time (s)
AF, $\tau = 8$	548	58
AF, $\tau = 10$	776	102
AF, $\tau = 12$	992	172
SF	1316	341
NF	-	157

4.3 Filtration Results - Comparing Accuracies of Calculations

When determining the accuracy of calculations using SF, the goal or 'ideal answer' is the answer given by AIMPRO under the same conditions but without using filtration, a NF calculation. When the AF algorithm is being tested, the goal is then the answer using SF, i.e. no trimming of the basis functions within the sphere of radius R_{cut} around the central atom. Note the same value of R_{cut} must be used in both the SF and AF calculation. To state this another way, the process is perfect from an accuracy viewpoint if AF calculations produce no changes to the energy obtained without function trimming taking place. This concept is central to the analysis presented in this thesis.

These points can be reinforced with two examples. If a formation energy (FE) is quoted as 14.5 meV, and AIMPRO without filtration comes up with -3.5 meV, we want AIMPRO with SF to return an answer as close to -3.5 meV as possible. The task of getting AIMPRO's answer closer to the accepted one is a separate topic entirely.

Similarly, if the SF calculation ran on this system returned say -3.0 meV, and the calculation ran for a third time, now using AF (with the same value of R_{cut} as the SF calculation), we want the answer to be as close as possible to -3.0 meV, not -3.5 meV. When discussing the AF process, we are focussed only on that and not the accuracy of the untrimmed but filtered (SF) calculations.

The idea behind AF is to reduce the time taken for the filtration step, while sacrificing only a small amount of accuracy, typically less than 10 meV. As the number of functions trimmed in the AF process increases, the calculation will deviate more and more from the corresponding SF calculation, and the filtration step will take less and less time. The main goal is to see how far the filtration parameters can be pushed before the accuracy of the calculation is compromised too much, i.e. beyond the accepted 10 meV threshold.

When data is presented, typically greater accuracy is recorded than presented. For example when single energies are calculated, 10 decimal places are recorded and used to calculate FEs, as well as absolute and percentage differences. As data is rounded for display in each table sometimes two FEs that appear to be the same are only the same to the nearest meV, and can still show a percentage and/or absolute difference.

4.4 Filtration Results - Ideal Vacancy In Silicon

This section presents results of evaluations of total energies and FEs for an ideal vacancy in silicon, for various sized systems. An ideal vacancy consists of taking the regular bulk silicon structure and removing one atom, without then relaxing the structure. The FE is defined as the difference in energy of N silicon atoms in a regular structure of N sites, to that of N atoms in a regular structure of $N+1$ sites. Here we calculate the energy of the regular structure, $E(N)$, and the same structure with one atom removed, $E(N-1)$. Hence the FE can be calculated as.

$$\text{Formation Energy} = E(N-1) - \frac{N-1}{N}E(N) \tag{74}$$

4.4.1 Details of Systems Modelled

- Three system sizes were investigated, N=64, 216 and 512 atoms.
- Calculations were performed using a simple cubic lattice, with a lattice constant of multiples of 5.419Å, and Γ point sampling of the Brillouin zone.
- The DFT calculation used an LDA functional [49].
- The pseudopotential used were as presented in Hartwigsen, Goedecker and Hutter (1998) [29].
- SF calculations used values of R_{cut} ranging from 5 to 12 a.u., AF calculations used values for the AF parameter τ of 5-8.

- The basis set for silicon was ddpp, and 4 filtered functions per silicon atom were used.
- The Fermi-Dirac function used in the filtration process had a Fermi energy of 0.2 Ha, with kT set to 0.1 Ha.

4.4.2 Link Between $R_{\mathbf{cut}}$ and the Time Required for a SCF step

Before the accuracy of the calculations is examined, it is interesting to see the effect on the length of time for a SCF iteration the R_{cut} parameter has. Table 6 shows the average SCF time for the 63 and 64 silicon atom systems for both the NF run, and the SF runs at the relevant R_{cut} radii. The 63 atom calculations are a little quicker than the 64 atom ones. As well as speed increases due to the reduced system size as the values of N_{keep} are lower, the filtration process will be quicker for each atom as well as having to run for one fewer atom. As predicted in section 4.1.2, the SCF times for the 63 atom system calculations are very close to the 64 atom system calculations, previously seen in table 2. This allows us to look to the value of R_{cut} when analysing the effect on the time required for a SCF iteration, instead of having to calculate average SCF times. Looking at R_{cut} and N_{keep} instead of actual SCF times allows effects such as computational power to be ignored.

4.4.3 Overview of Results

Three sections of results are presented for the bulk and ideal vacancy silicon calculations.

1. Section 4.4.4 looks at the differences between total energy calculations performed using NF and SF. SF calculations were performed using values of R_{cut} ranging from 5 to 12 a.u., for system sizes of 63/64, 215/216 and 511/512 atoms of silicon (for the ideal vacancy and bulk structures respectively).

Table 6: The effect of filtration radius R_{cut} on the number of functions presented to the filtration algorithm (N_{keep}), and average SCF times, for unit cells containing 64 atoms of bulk silicon, and 63 atom of bulk silicon with an ideal vacancy. For SF calculations, as N_{keep} increases, so does the time required for an average SCF iteration. At this small system size, only the lowest value of R_{cut} leads to a faster SCF iteration than when using NF.

R_{cut} (a.u.)	N_{keep} 63 atoms	Average SCF time(s) 63 atoms	N_{keep} 64 atoms	Average SCF time(s) 64 atoms
5	138.2	7	140	8
8	468.9	12	476	13
10	799.5	29	812	31
12	1295.6	92	1316	98
NF	-	10	-	11

2. Section 4.4.5 investigates the calculation of the FE of an ideal vacancy in silicon, using the same system sizes and filtration methods/parameters seen in section 4.4.4.

3. Section 4.4.6 analyses the performance of AF, by comparing AF calculations of FEs of an ideal vacancy in silicon to corresponding SF calculations. R_{cut} was set to 10 a.u. for all calculations, with the AF parameter τ ranging from 5 to 8. Calculations were performed on system sizes of 215/216, 511/512 and 999/1000 atoms.

4.4.4 Total Energy Calculations - NF vs SF

This section examines AIMPRO calculations of the total energy of bulk silicon and the corresponding ideal vacancy systems. Calculations without filtration (NF) are compared to ones using filtration (SF) for varying values of R_{cut}.

Table 7: The effect of filtration radius R_{cut} on the energies of unit cells containing 64 atoms of bulk silicon, and 63 atoms of bulk silicon with an ideal vacancy. The SF calculations becomes closer to the corresponding NF calculations as R_{cut} is increased. The same effect can be seen in tables 8 and 9.

R_{cut} (a.u.)	$E(R_{\text{cut}})$ (Ha) 64 atoms	$E(R_{\text{cut}})$ (Ha) 63 atoms	$E(R_{\text{cut}}) - E(NF)$ (meV per atom) 64 atoms	$E(R_{\text{cut}}) - E(NF)$ (meV per atom) 63 atoms
5	-253.12570	-249.05343	136.5	134.5
8	-253.41976	-249.33542	11.5	12.7
10	-253.43218	-249.34873	6.2	7.0
12	-253.44338	-249.36117	1.4	1.6
NF	-253.44674	-249.36488	-	-

Table 8: The effect of filtration radius R_{cut} on the energies of unit cells containing 216 atoms of bulk silicon, and 215 atoms of bulk silicon with an ideal vacancy. The SF calculations becomes closer to the corresponding NF calculations as R_{cut} is increased. The same effect can be seen in tables 7 and 9.

R_{cut} (a.u.)	$E(R_{\text{cut}})$ (Ha) 216 atoms	$E(R_{\text{cut}})$ (Ha) 215 atoms	$E(R_{\text{cut}}) - E(NF)$ (meV per atom) 216 atoms	$E(R_{\text{cut}}) - E(NF)$ (meV per atom) 215 atoms
5	-855.10331	-850.99417	132.7	132.9
8	-856.05438	-851.94019	12.9	13.2
10	-856.09488	-851.98162	7.8	7.9
12	-856.14012	-852.02744	2.1	2.1
NF	-856.15658	-852.04432	-	-

Table 9: The effect of filtration radius R_{cut} on the energies of unit cells containing 512 atoms of bulk silicon, and 511 atoms of bulk silicon with an ideal vacancy. The SF calculations becomes closer to the corresponding NF calculations as R_{cut} is increased. The same effect can be seen in tables 7 and 8.

R_{cut} (a.u.)	$E(R_{\text{cut}})$ (Ha) 512 atoms	$E(R_{\text{cut}})$ (Ha) 511 atoms	$E(R_{\text{cut}}) - E(NF)$ (meV per atom) 512 atoms	$E(R_{\text{cut}}) - E(NF)$ (meV per atom) 511 atoms
5	-2027.18610	-2023.06812	131.2	131.3
8	-2029.40988	-2025.28801	13.0	13.1
10	-2029.50611	-2025.38519	7.9	7.8
12	-2029.61556	-2025.49514	2.1	2.1
NF	-2029.65436	-2025.53435	-	-

In tables 7, 8 and 9 the calculated energies of each structure are presented for SF calculations of varying R_{cut}, and the corresponding NF calculation. The differences of each SF calculation to the corresponding NF calculation in meV per atom is also presented. When comparing single energy calculations, the energy difference per subunit (in this case a silicon atom) of the structure should be examined, not the total difference. If the total energy difference was used, it would increase linearly with the size of the system, which doesn't lend itself to useful analysis. Note this is different to when examining FE differences. FEs can be directly compared between different system sizes. As only one vacancy is present in each of the pairs of systems, and it is only the energy of the vacancy that is calculated, the FE should converge to a fixed answer as the system size is increased, with vacancy-vacancy interactions between adjacent supercells tending to zero with increasing number of atoms per supercell.

Four very basic conclusions can be drawn immediately.

For each of the six systems investigated, it can be seen that as the cut-off radius

increases, the results converge towards the unfiltered value. This is to be expected, as large values of R_{cut} reduce the truncation in filtration.

There is a clear trend of a less negative total energy with decreasing filtration radius. This is due to the reduced freedom when constructing the filtered basis set, therefore the energy is higher due to the variational principle.

The differences per atom for the three pairs of systems are very similar, with the results for 215/216 and 511/512 atoms very close to each other for all values of R_{cut}, especially $R_{\text{cut}} > 5$ a.u.. This is not a particularly surprising find, but is reassuring in terms of the observed consistency of the filtration method. As gamma point sampling of the Brillouin zone was used for these calculations, the large reciprocal lattice size for 63/64 atom systems will mean the FE is not close to the convergence value.

The energy differences per atom at $R_{\text{cut}} = 5$ a.u. of over 130 meV in all 3 cases is very high. This is much too large a difference in the absolute energies for filtration to be an acceptable replacement method for unfiltered calculations. Typically differences of the order of 1-5 meV per atom would be regarded as an acceptable difference. $R_{\text{cut}} = 10$ a.u. has differences of 6-8 meV/atom, which is still a bit too high. Only a value of $R_{\text{cut}} = 12$ a.u. yields a reasonable difference of 1.4-2.1 meV/atom. However such calculations using filtration are commonplace, and still yield the time-saving properties of filtration described earlier [56].

This being said, in computational physics or chemistry only differences of energy are important and the error in the absolute energies are expected to be largely systematic, and indeed this is necessary for filtration to be a useful technique. This is examined in the next section.

4.4.5 Formation Energy Calculations - NF vs SF

The previous section examined a single NF calculation against a single SF calculation. In this section we look at formation energies, calculated using the formula outlined in the introduction to section 4.4. This formula is based around the difference of two

total energies, both of which use the same filtration method (i.e. E[NF_1] - E[NF_2] or E[SF_1] - E[SF_2], never E[NF_1] - E[SF_2]). It is hoped this will lead to cancellation of systematic errors.

Looking at the results in table 10 we can see the error in an energy difference such as a FE is much smaller than the error in a total energy calculation. The differences in the vacancy FE are far smaller than the individual differences making up the calculation. They also don't show a great deal of difference for different system sizes. This means our analysis here, and for similar ones in the future, should focus on quantities based on energy differences, such as a FE in this particular case. The idea that basis set errors are largely systematic and cancel when looking at differences, is well established. It is used routinely in quantum chemistry [30].

Table 10: Changes in total energies (ΔE) and FEs (ΔFE) for ideal vacancy formation in bulk silicon systems, modelled using SF with R_{cut}=12 a.u. compared to NF calculations. The much larger differences in the total energies that scale with system size contrast to the much smaller differences seen in the formation energies. This demonstrates the cancellation of systematic errors necessary for filtration to provide accurate results.

System size N-1/N atoms	ΔE=E[NF]-E[SF] N atoms (mHa)	ΔE=E[NF]-E[SF] N-1 atoms (mHa)	ΔFE=FE[NF]-FE[SF] (mHa)
63/64 atoms	-3.36	-3.71	0.40
215/216 atoms	-16.46	-16.88	0.50
511/512 atoms	-38.80	-39.21	0.49

Although for the purposes of this and following analyses, single energy AIMPRO results are not required, they are detailed for almost all results in this and the following chapter. This is because they serve as useful data, and allow the possibility of further analysis of this data, in ways not carried out in this thesis. With the volume of data presented, and the overlap of some parameters in many of the calculations, it

is possible some correlations will go unseen. There are a handful of cases where the source single energy results are not present, in cases where a main point is being supported by further results, and these further results are not central to the current theme. An example of this is the result of moving a silicon atom by 0.1 pm increments and finding the FE at each value, as this data is presented to support the observed shifts in atoms by giving a feel for the energy changes at this level.

Table 11: Calculations of the FE of an ideal vacancy in a 64 atom unit cell of silicon, using NF, and SF for a range of values of R_{cut}. The FEs are calculated using the formula 74 in the introduction to section 4.4. As the filtration radius R_{cut} is increased, the SF calculations converge to the NF result.

R_{cut} (a.u.)	E[64] (Ha)	E[63] (Ha)	FE (eV)	FE[R_{cut}]-FE[NF] (meV)	|% Change|
5	-253.12570	-249.05343	3.189	-124	3.76
8	-253.41976	-249.33542	3.392	79	2.38
10	-253.43218	-249.34873	3.363	49	1.49
12	-253.44338	-249.36117	3.324	11	0.33
NF	-253.44674	-249.36488	3.313	-	-

A more detailed analysis is now provided. For each of the 3 system sizes, FEs were calculated using different values of R_{cut}, and each compared to the corresponding NF result. The results are shown in tables 11-13. Looking at table 11 reveals that as R_{cut} is increased, the magnitude of the FE difference to the NF calculation decreases. For an R_{cut} of 10 a.u., the difference is around 50 meV, and for 12 a.u. around 10 meV. 10 meV is about the accuracy of any unfiltered calculation. Certainly the calculation done with an R_{cut} of 12 a.u. gives an acceptable answer. Any filtration radius smaller than 10 a.u. would be unacceptable for calculating FEs. With most electronic structure calculations the energies lie above the true answer, true in the sense of an ideal measurement at 0K. However for a filtration radius of 5 a.u. we get an answer lower

than our target FE, the unfiltered calculation that yields a FE of 3.313 eV. This is because the FE is a function of two calculated energies. If say the 63 atom calculation is further away from its unfiltered 63 atom counterpart than the 64 atom calculation is from its 64 atom unfiltered counterpart, the FE will be lower than the unfiltered calculation.

Table 12: Calculations of the FE of an ideal vacancy in a 216 atom unit cell of silicon, using NF, and SF for a range of values of R_{cut}. The FEs are calculated using the formula 74 in section 4.4. As the filtration radius R_{cut} is increased, the SF calculations converge to the NF result.

R_{cut} (a.u.)	E[216] (Ha)	E[215] (Ha)	FE (eV)	E[R_{cut}]-E[NF] (meV)	|% Change|
5	-855.10331	-850.99417	4.091	48	1.18
8	-856.05438	-851.94019	4.108	65	1.62
10	-856.09488	-851.98162	4.078	35	0.87
12	-856.14012	-852.02744	4.056	14	0.33
NF	-856.15658	-852.04432	4.043	-	-

An examination of tables 12 and 13 shows the same trends as table 11. An R_{cut} of 5 or 8 a.u. is poor with differences to the unfiltered calculations of around 50-75 meV. R_{cut}=12 a.u. FEs are 13-14 meV, just larger than the 10 meV threshold previously discussed. These would be acceptable results. If a value of R_{cut} of 12 a.u. is required for an accurate calculation, referring to table 2 shows for calculations of this type, filtration becomes effective (i.e. reduces the time required for a SCF iteration) for systems whose size is somewhere between 216 and 512 atoms. It is worth nothing however that all these calculations were performed using Γ point sampling of the Brillouin zone. For systems of this size the sampling would always take place on a grid, typically with 4 or more symmetry-distinct k-points. The filtration process only

Table 13: Calculations of the FE of an ideal vacancy in a 512 atom unit cell of silicon, using NF, and SF for a range of values of R_{cut}. The FEs are calculated using the formula 74 in section 4.4. As the filtration radius R_{cut} is increased, the SF calculations converge to the NF result.

R_{cut} (a.u.)	E[512] (Ha)	E[511] (Ha)	FE (eV)	E[R_{cut}]-E[NF] (meV)	|% Change|
5	-2027.18610	-2023.06812	4.317	76	1.79
8	-2029.40988	-2025.28801	4.304	63	1.5
10	-2029.50611	-2025.38519	4.273	32	0.77
12	-2029.61556	-2025.49514	4.254	13	0.31
NF	-2029.65436	-2025.53435	4.241	-	-

needs to run once and can then be applied to all k-points. SF calculations using 4 k-points will suffer the penalty of one filtration step, but offer the advantage of a reduced Hamiltonian diagonalisation four times in each SCF step. This will favour SF calculations. Table 14 shows the effect of increasing the number of k-points on the average time required for a SCF step. Filtration is effective for 216 atom systems when using a MP grid of 4 4 4 or finer. In fact it reduces the time required for a SCF iteration by over a factor of 3. For 64 atom systems filtration is still not effective for a MP grid of 4 4 4, but only takes twice as long per SCF iteration, instead of the factor of 9 seen for Γ-point sampling.

It should be noted that these results are for bulk silicon, which displays a large amount of symmetry, helping to reduce the number of independent k-points. Almost all systems to be modelled will not show this degree of symmetry, and consequently the number of k-points will be larger when using the same 222 or 444 MP grid. Repeating the 64 atom calculations, but moving a few atoms slightly to break the symmetry, allows the effect of the higher number of k-points on the average SCF time to be seen. These results are in table 15. Now the calculation using a MP 4 4 4 sampling grid is

Table 14: Average times for an SCF iteration for bulk silicon energy calculations with increasing number of k-points (Γ-point, MP 2 2 2, MP 4 4 4). Two system sizes, 64 and 216 atoms, were modelled. SF calculations used an R_{cut} of 12 a.u.. The effect of the filtration step required to be performed only once per SCF iteration can be seen through the rapid increase of the NF SCF times, compared to the gradual increase seen for the SF calculations.

Filtration	Average SCF time (s)					
Method	64 atoms			216 atoms		
	Γ	2 2 2	4 4 4	Γ	2 2 2	4 4 4
SF	98	101	118	341	370	505
NF	11	19	57	157	457	1762

faster under SF than a NF calculation, even for 64 atoms.

The value of R_{cut}=12 a.u. has been shown to produce accurate calculations for the ideal vacancy formation energy in silicon, for system sizes dependent on the k-point sampling options. To achieve converged results, as the system size is reduced a finer grid is usually necessary, which would mean filtration can be effective for system sizes of 216 atoms, and possibly smaller. Once the system size approaches 64 atoms, the filtration step itself takes longer than the savings it produces, unless a very fine k-point sampling grid is employed in the calculation. However, the time required for the filtration step can be reduced using AF. Hence the above calculations were then repeated using AF, albeit for a different range of system sizes, the results of which are detailed in the next section.

4.4.6 Formation Energy Calculations, SF vs AF

The calculations in this section were performed using AF, with values of the tolerance parameter, τ, of 5-8. This time 215/216, 511/512 and 999/1000 atom systems were used. R_{cut} was chosen to be 10 a.u., all other parameters the same as in section

Table 15: Average times for an SCF iteration for total energy calculations for a unit cell of 64 atoms of silicon, with 5 atoms slightly displaced to break the symmetry of the unit cell. SF calculations used an R_{cut} of 12 a.u.. The effect of the filtration step being performed only once per SCF iteration is even more pronounced than was witnessed in table 14.

Filtration	Average SCF time (s)	
Method	2 2 2	4 4 4
SF	117	277
NF	56	408

4.4. Tables 16, 17 and 18 show the results for 215/216, 511/512 and 999/1000 atom systems respectively.

Note the values at $\tau = \infty$ still represent a filtered calculation, with the filtration radius set to 10 a.u.. The lower the value of τ, the more functions that are dropped, and the lower the accuracy. When $\tau = \infty$ no functions are dropped, producing a SF calculation.

Before examining the FE results, one observation is worth noting from the data presented in tables 16, 17 and 18. Previously when comparing single energy calculations using SF and NF, large energy differences were seen. However the systematic nature of these allowed them to be used when comparing energy differences with much more accuracy, as the differences tended to 'follow' each other. This is also seen here when comparing SF calculations against AF ones.

Looking at the results for the three system sizes at once, the results are divisible into two sets. For $\tau = 5$, the difference between the SF calculation and the trimmed one is about 28 meV in all 3 cases. This is an appreciable difference. For values of $\tau = 6$ or above, very small differences are seen, between 1 and 5 meV, less significant than other sources of error inherent in computational calculations. A value of $\tau = 6$ or more would be an acceptable filtration parameter for this and similar systems.

Looking back to table 4 in section 4.2.3, we see a value of 6 for τ reduces the average SCF time for this system from 115 seconds for a SF calculation, to 41 seconds, and to 58 seconds for τ=8. A NF calculation had an average SCF time of 157 seconds. This is an excellent result both from a timing and accuracy perspective.

Table 16: The effect of the AF parameter τ on calculations of FEs of an ideal vacancy in a 216 atom unit cell of silicon. R_{cut} was set to 10 a.u.. Even very small values of τ, down to as low as 6, lead to negligible differences in the resulting FE compared to that achieved using SF.

τ	E[216] (Ha)	E[215] (Ha)	FE (eV)	FE[τ]-FE[SF] (meV)	|% Change|
5	-855.94782	-851.83409	4.109	28	0.68
6	-855.97395	-851.86131	4.076	-5	-0.13
7	-856.06890	-851.95583	4.076	-5	-0.13
8	-856.07790	-851.96454	4.083	1	0.03
∞ (SF)	-856.09501	-851.98162	4.081	-	-

4.4.7 Conclusions from Results for Ideal Vacancy in Silicon

A summary of the main findings from this section is as follows:

1. Total energy calculations performed using SF or AF produce results that can be different from the corresponding NF results, but that can be controlled with R_{cut} and τ. The differences depend on system size, as well as the chosen values of R_{cut} and τ.

2. Differences of total energy calculations using SF or AF, such as FEs, produce small deviations from the equivalent NF calculations.

3. For the FE of an ideal vacancy in silicon, a SF calculation using a filtration radius R_{cut} of 12 a.u. gives an acceptable result.

Table 17: The effect of the AF parameter τ on calculations of FEs of an ideal vacancy in a 512 atom unit cell of silicon. R_{cut} was set to 10 a.u.. Even very small values of τ, down to as low as 6, lead to negligible differences in the resulting FE compared to that achieved using SF.

τ	E[512] (Ha)	E[511] (Ha)	FE (eV)	FE[τ]-FE[SF] (meV)	|% Change|
5	-2029.15237	-2025.03687	4.145	28	0.69
6	-2029.21574	-2025.10123	4.114	-2	0.05
7	-2029.44644	-2025.33143	4.116	1	0.02
8	-2029.46723	-2025.35198	4.121	5	0.12
∞ (SF)	-2029.50613	-2025.39098	4.116	-	-

Table 18: The effect of the AF parameter τ on calculations of FEs of an ideal vacancy in a 1000 atom unit cell of silicon. R_{cut} was set to 10 a.u.. Even very small values of τ, down to as low as 6, lead to negligible differences in the resulting FE compared to that achieved using SF.

τ	E[1000] (Ha)	E[999] (Ha)	FE (eV)	FE[τ]-FE[SF] (meV)	|% Change|
5	-3963.26535	-3959.14745	4.208	27	0.63
6	-3963.39010	-3959.27316	4.178	-3	0.07
7	-3963.84342	-3959.72591	4.182	0	0.00
8	-3963.88352	-3959.76580	4.186	5	0.11
∞ (SF)	-3963.95965	-3959.84203	4.181	-	-

4. When using AF to calculate the FE of an ideal vacancy in silicon, values of the AF parameter τ as low as 6 produced acceptable results. This corresponds to using only 40% of the functions present in the original sphere of radius R_{cut}, and reduces the average SCF time to a similar percentage.

The next section will the performance of SF and AF when used to calculate the energy of a reaction, which involves calculating total energies for four silicon based structures.

4.5 Oxygen defect in silicon

As a second example to test filtration, defects containing oxygen were considered in silicon, specifically the energy of a reaction of two structures forming two others.

4.5.1 Details of Systems Modelled

The calculations in this section are of the energy of the reaction in which an oxygen-vacancy centre captures an interstitial oxygen atom to create the VO_2 defect [17].

$$\mathrm{E[Reaction]} = \mathrm{E[216\ Si]} + \mathrm{E[VO_2]} - \mathrm{E[VO]} - \mathrm{E[O_i]} \tag{75}$$

The four structures that appear in this reaction are as follows:

1. A unit cell of silicon containing 216 atoms of silicon in the relaxed configuration, referred to as 216 Si.

2. A unit cell of silicon containing an extra interstitial oxygen atom, referred to as O_i. To create this structure, an oxygen atom is introduced to an otherwise perfect silicon unit cell of 216 atoms. This forms a defect in which a Si-Si bond breaks and the oxygen atom inserts itself into this broken bond, forming a bridging configuration [17].

3. A unit cell of silicon in which a silicon atom is replaced with an off-centre oxygen atom. This structure is referred to as VO [17].

4. A unit cell of silicon in which a silicon atom is replaced with two oxygen atoms. This structure is referred to as VO_2 [17]. As the VO_2 defect can be considered to be formed when a VO centre captures a O_i defect, the energy of this transformation E[Reaction] is equivalent to equation 75.

The total energy calculations for each of these four structures used the following parameters:

- The supercells were made using a 3x3x3 grid of primitive cubic cells of 8 silicon atoms, with a lattice parameter of 10.19 a.u..
- For the O_i, VO and VO_2 structures, initial best guesses as to the relaxed structure were made. This was done by initially removing a silicon atom (for the VO and VO_2), then placing the oxygen atom(s) using a best guess. The structure was then optimised in AIMPRO using the conjugate gradient algorithm. These optimised structures were used for all the total energy calculations.
- The k-point sampling of the Brillouin zone used an MP 2 2 2 sampling grid [45].
- The Fermi-Dirac function used in the filtration process had a Fermi energy of 0.2 Ha, with kT set to 0.1 Ha.
- The filling of the energy levels took place at a temperature with kT set to 0.04 eV.
- The pseudopotentials for oxygen and silicon were as presented in Hartwigsen, Goedecker and Hutter (1998) [29].

4.5.2 Overview of Results

Section 4.5.3 looks at the difference in the energy of the reaction using NF, and SF using a filtration radius R_{cut} of 10 a.u..

Section 4.5.4 compares the calculated reaction energy between the SF result from section 4.5.3 to AF calculations using values of the AF parameter τ of 5, 6, 8 and 10.

Finally in section 4.5.5 an analysis of the effect of the different filtration methods, and parameters used in each, on the time required for SCF iterations is provided.

Table 19: Comparison of NF and SF calculations of energy of reactants, energy of products and overall reaction energy for the reaction (75). SF calculations used a value of 10 a.u. for R_{cut}.

	E_{NF} (Ha)	E_{SF}, R_{cut}=10 a.u. (Ha)	E_{NF}-E_{SF} (meV)
E[O_i] (Ha)	-872.20824	-872.14660	-1677
E[216] (Ha)	-856.16318	-856.10183	-1669
E[VO_2] (Ha)	-884.28124	-884.21883	-1698
E[VO] (Ha)	-868.18638	-868.12429	-1690
E[Reaction] (eV)	-1.355	-1.354	-1

4.5.3 Energy of Reaction - NF vs SF

Looking first at the SF results in table 19, we see the unfiltered reaction energy is -1.355 eV, and the SF calculation gives an answer of 1.354 eV, a difference of only 1 meV. Again the differences in the individual energies are far larger than the differences in the reaction energy, here by a factor of over 1600. This is an outstanding result, and suggests we may have room to make efficiency savings via AF.

4.5.4 Energy of Reaction - SF vs AF

In table 20 the comparison is between the SF energies, and the function trimmed energies (the SF result is the target result for AF). For values of $\tau = 5$ or $\tau = 6$ large differences of 86 meV and 116 meV respectively are obtained. This is an order of magnitude larger than the acceptance threshold of 10 meV. When the quality of the calculation is increased with the AF parameters of $\tau = 8$ and $\tau = 10$ work much better giving differences of 19 meV and 7 meV respectively. Only $\tau = 10$ produces a calculation of acceptable quality. It should be noted that the individual energy differences are still much larger in absolute terms than the differences in the reaction energy. That is to say the differences are still cancelling out, but not to the extent

Table 20: The effect of parameter τ on AF calculations, for energy of reactants, energy of products and reaction energy for the reaction (75). All calculations used a value of 10 a.u. for R_{cut}.

τ	5	6	8	10
E[O_i] (Ha)	-871.97734	-872.01506	-872.12799	-872.14492
E[216] (Ha)	-855.94265	-855.96951	-856.08427	-856.10065
E[VO_2] (Ha)	-884.04694	-884.08478	-884.20092	-884.21704
E[VO] (Ha)	-867.95932	-867.99371	-868.10675	-868.12273
E[Reaction] (eV)	-1.440	-1.239	-1.373	-1.362
Difference to SF (meV)	-86	116	-19	-7
Difference to SF mod (%)	6.3	8.5	1.4	0.5

seen when just applying SF. This is to be expected to a certain degree, as lower values of τ are reducing the quality of the basis set. However the differences are larger than those seen in the ideal vacancy FE results.

4.5.5 Link Between R_{cut}, τ and the average time required for an SCF iteration.

It would be useful to see what these values of τ mean, in terms of how much faster the part of the calculation they affect runs. The average numbers of functions available from which to construct the filtered functions in each of the calculations from table 20 are presented in table 21, along with the average time for a SCF step. The reduction in this time reflects the faster speed of the generation of the filtered functions, which depends cubically on N_{keep}. The effect of this is also proportional to the number of atoms, so that the speed up is greater as the system size reduces,.

The dramatic change in SCF time at the small end of the system spectrum due to the speeding up of the filtration process is clear, with dropping the number of functions by a factor of roughly 3.5 leading to a reduction in the SCF time by approximately a

Table 21: Number of functions presented to the filtration algorithm (N_{keep}) and average SCF times in seconds for SF and AF calculations for reactants and products of reaction (75).

τ	O_i		216		VO_2		VO	
	N_{keep}	SCF (s)	N_{keep}	SCF (s)	N_{keep}	SCF (s)	N_{keep}	SCF (s)
5	261	28	260	15	261	27	260	26
6	384	65	380	41	380	66	379	65
8	555	86	548	58	565	87	562	85
10	723	115	716	88	725	113	720	113
SF	823	136	812	108	827	139	820	138
NF	-	172	-	146	-	171	-	170

factor of 5. This is at the expense of accuracy however, with only a value of 10 for τ giving an answer within 10 meV of the SF result. However this does reduce the time required for an SCF iteration to roughly two thirds of that seen for a NF calculation, without an appreciable loss of accuracy. It must be noted that this is at the smallest end of the spectrum of system sizes for which filtration operates. For larger systems the time savings obtained with filtration will be increased dramatically, tending much more rapidly to two orders of magnitude, which comes the $(N/n)^3$ theoretical limit outlined in section 3.1. The ability to produce faster calculations for systems of this small size is a good result.

4.5.6 Summary of Findings

The use of SF with an R_{cut} of 10 a.u. produces results only 1 meV different to the NF result. When AF is introduced into this SF calculation, a value of τ=10 produced good results. Use of these filtration parameters for these systems will produce improvements for both small and large systems, an important result. This has been presented as a trimming down of the functions captured using R_{cut}=10 a.u.. However, the possibility

clearly exists to start with R_{cut}=15 a.u. or a suitably large value, and then use τ as the fundamental parameter, rather than R_{cut}. This is examined in chapter 6.

4.6 Conclusions

Total energy calculations performed using filtration can produce results significantly different to their NF counterparts. This difference, being an extensive quantity, increases with system size. The energy difference per atom however does not, and can be made sufficiently small with a large enough value of R_{cut}. When the systematic nature of these differences are removed, by subtracting total energies, the resulting SF and AF calculations lie much closer to the corresponding NF calculations.

For the systems studied here, SF calculations using a filtration radius R_{cut} of 12 a.u. produce results of acceptable accuracy, that is to say within 10 meV of the NF result. For some systems R_{cut} can be lowered to 10 a.u.. For systems of 216 atoms and above this leads to time savings, through the reduction in the time required for SCF iterations, even when using Γ-point sampling of the Brillouin zone. When using a finer sampling grid, filtration becomes effective for lower and lower sized systems. MP 4 4 4 grids and finer are faster using filtration than without for bulk silicon calculations of only 64 atoms.

AF is a method of reducing the number of functions presented to the filtration algorithm, and consequently speeding up the filtration process, which in turn reduces the time required for an SCF iteration. Its use produces some savings, but care must be used to ensure too many functions are not removed before the filtered functions are created, as this can lead to divergence in calculated energies from the SF/NF result. However in the next section, this is shown to not necessarily be a limitation, as filtration is applied to structural optimisation, where filtration is used to calculate forces rather than energies.

The use of τ rather than R_{cut} as a parameter to control the accuracy of a filtered calculation has also been suggested. This is investigated in chapter 6, where various

different methods of trimming the functions are applied to calculate four further formation energies, to see which is the most effective. In these calculations, the value of R_{cut} is made large enough so that only the AF parameters affect the functions included in the creation of the filtered functions.

Chapter 5

OPTIMISATION OF STRUCTURES USING FILTRATION

Possibly the most common type of calculation performed using AIMPRO is the determination of the equilibrium structure of a molecule or solid. This process is referred to as structural optimisation. It relies on AIMPRO providing the forces on each atom for a particular set of coordinates for each atom. In this chapter, the use of filtration in such calculations will be examined. The use of filtration in this area introduces some problems, which require the development of new optimisation routines. After some basic theory, these developments will be presented, then finally these new optimisation routines will be used in both SF and AF calculations, and the results presented and analysed from perspectives of both accuracy and speed.

To reduce the time required for a structural optimisation, there are two main parts of the calculation where this can take place. Firstly the number of times AIMPRO is required to provide the forces on each atom (known as force calls) can be reduced. This is mainly achieved by using more efficient optimisation algorithms. Filtration requires the development of a new line minimiser, which forms an integral part of the optimisation routine used by AIMPRO. This has been designed to use as few force calls as possible, while still being accurate and stable. This development takes up the first half of this chapter.

Secondly, each force call can be made faster. This is achieved through the use of filtration. The more aggressively parameters can be set, namely R_{cut} and τ (introduced in the previous chapter), without affecting the resulting structure, the quicker each force call will be, and consequently the overall calculation. The application of this methodology to defects structures in silicon forms the second half of this chapter.

Before this, some basic concepts relevant to the discussions in this chapter are outlined.

5.1 Structural Determination Calculations in AIMPRO

As explained in section 2.3, the Born-Oppenheimer approximation allows us to define a specific energy for every set of atomic coordinates. Expressed in Cartesian coordinates, this gives the energy as a scalar function of 3N dimensions for the x, y and z values of the positions of the N nuclei in the structure under examination. This surface is referred to as a potential energy surface (PES), and the concept forms the backbone of the structural determination methods used in this thesis.

5.1.1 Potential Energy Surfaces

Two of the most commonly seen features on PESs are minima and saddle points. An order n saddle point on a PES is a maximum in n dimensions and a minimum in the other $3N - n$. Unless specifically stated, this work deals with first order saddle points, so it a point on the potential energy surface which is a maximum in one direction, and a minimum in all the others. A minimum on the PES is referred to as the equilibrium position, where the forces on the atoms of the structure represented by this point are zero. If the temperature was 0K and zero point energy ignored, the atoms would remain at this position indefinitely according to classical physics. Around the equilibrium position, the shape of the potential energy curve with respect to the distance between two atoms is roughly quadratic, that is it can be approximated to:

$$\mathrm{PE} = ax^2 + bx + c \tag{76}$$

for this one-dimensional system.

The closer the positions of the atoms are to equilibrium, the more accurate this approximation is. Most of the optimisation techniques employed assume a quadratic PES. Although this clearly won't be appropriate when far enough away from a local

minimum or saddle point on the PES, the techniques used will still generally move towards their goal of a local minimum or saddle point. As they do the methods become more and more accurate. Examples of methods like this include conjugate gradient (CG) [33, 52], BFGS [13, 14, 22, 24, 52, 58] and direction inversion in the iterative subspace method (DIIS) [10, 19, 20, 53, 54].

For structures which are so far away from the quadratic region of their PES that methods of this nature could be unstable, methods that do not rely on this quadratic approximation exist. The simplest, safest and often slowest to converge method in this category is known as steepest descent. A safe way to move a structure towards the desired quadratic region of a PES can be to start out with the steepest descent method, then when closer to the minimum/saddle point, employing one of the faster quadratically based methods.

A potential energy surface is multi-dimensional, a hypersurface. Hence they can contain many minima. Even a simple PES with 2 spatial axes can be extremely complicated and contain many minima and saddle points. The minima closest to the starting point may not be the lowest energy minimum of all minima contained within the region spanned by the PES. The lowest energy minimum is referred to as the global minimum. It is not guaranteed the closest minimum to a starting structure is a global minimum, and could instead be a local minimum. Unless you are dealing with a very simple structure, usually or no more than 4 atoms, it is excessively time consuming if the minimum you have located is the global minimum just by using optimisation techniques. A knowledge of the chemistry of the system under investigation through experience of similar systems is typically required to put the results into a useful context.

This chapter deals with the effect of filtration on structural optimisation calculations. Chapter 7 is unrelated to filtration, and concerns itself with a faster method of identifying saddle point structures without the identification of a minimum energy path (MEP). The previous technique known as the nudged elastic band (NEB)

method, identifies both the saddle point and the MEP, but is a computationally intensive procedure.

5.1.2 Why Determine Minimum Energy Structures

The determination of minimum energy structures is a fundamental process of computational chemistry. These structures are seen experimentally, and the results of calculations can be compared to the results of x-ray diffraction experiments, which allow the structure to be determined directly. Through computational methods, the stable or meta-stable structures of molecules which are difficult to produce, short lived or not yet realised experimentally, can also be predicted. The properties of such structures can then be determined through further calculations, such as hyperfine couplings, IR absoprtion properties, and using, when combined with knowledge of transition state structures, activation energies for reactions. Calculations such as the NEB, which reveal MEPs, require minimum energy structures for both the start and end images. The algorithm used by AIMPRO to achieve this is now outlined.

5.1.3 Minima Finding Techniques - The Conjugate Gradient Algorithm

Of the variety of methods that can be employed to identify minima on a PES, AIMPRO uses the CG algorithm. It can be broken down into a series of repeated steps.

1. Set the initial search direction $\vec{d_0}(=\vec{d_i})$, to be equal to the initial force $\vec{F_0}(=\vec{F_i})$. Here vectors such as $\vec{F}$ are of length $3N$, where N is the number of atoms.

2. Move along this direction $\vec{d_i}$ to a position $\vec{R}_{i+1} = \vec{R}_i + \alpha\vec{d_i}$, until the force $\vec{F}_{i+1}$ is orthogonal to $\vec{d_i}$, i.e. $\vec{F}_{i+1}.\vec{d_i} = 0$. This process is known as line minimisation. In practice a tolerance value τ_{LM} is set, and the point $\vec{R}_{i+1}$ is accepted when $|\vec{F}_{i+1}.\vec{d_i}| < \tau_{LM}$, where $\vec{F}_{i+1}$ is the force at point R_{i+1}.

3. If at this point $\vec{R}_{i+1}$, the maximum component of the force vector $\vec{F}_{i+1}$, is less than another pre-defined tolerance τ_{CG}, the minimum has been reached.

4. If not, calculate a new search direction using (77).

$$\vec{d}_{i+1} = \vec{F}_{i+1} + \alpha \vec{d}_i \tag{77}$$

where α is given by (78) [52].

$$\alpha = \vec{F}_{i+1}.(\vec{F}_{i+1} - \vec{F}_i)/|\vec{F}_i|^2 \tag{78}$$

5. Repeat the procedure from step 2, using the new force and direction, until the check in step 3 is satisfied.

6. The CG direction is updated based on the previous directions and the current force. After a fixed number of iterations, this direction is reset to the current force, which effectively restarts the algorithm. Due to the non-quadratic nature of PESs, and the fact that the line minimisations that take place are not exact, this can prove restrictive. This is necessary for systems with small numbers of atoms, or with few degrees of freedom. For example if only one atom is free to move, only 3 conjugate directions are possible, so the CG algorithm needs to be reset every 3 iterations. The actual formula used is the smaller of 10, and the number of degrees of freedom in the system.

In its most basic terms, the CG method chooses a direction, and a line minimiser determines how far to move along this direction. For filtration, the direction choosing algorithm remains unchanged. However the line minimiser requires changes. AIMPRO has traditionally used a line minimiser that uses both energies and forces, using cubic interpolation. However, there is a slight discrepancy between the energy and forces produced by a filtration calculation [56]. This means that a line minimiser should not use energies and forces to determine how far to move in the search direction.

Two different solutions were implemented. The first, used for all the results in this chapter, is an optimisation method that does not require a line minimisation

step, based on the DIIS algorithm. This was implemented, then tested to ensure it produced the correct minimum energy structures, using the same or less force calls than the previous method.

The second method, used for all the results in chapter 7, involved a complete rewrite of the line minimiser. The development and testing of this was a significant investment, and details are provided in chapter 7.

The next section uses the DIIS algorithm to calculate minimum energy structures for unit cells of silicon, containing a variety of defects. Calculations are performed using NF, SF and AF, and the effect of R_{cut} and τ on the accuracy of the resulting structures investigated.

5.2 Comparing Structures

This chapter deals with relaxed structures produced by structural optimisation calculations. Calculations using NF, SF and AF, with varying filtration parameters are compared against each other. In the previous chapter, the comparisons were energy-based, which lends itself to a straight forward comparison. When comparing structures, this is not as straightforward. Methods which quantify the amount by which two structures are different are required. Two are used in this chapter.

An idea of how close two structures are to each other can be obtained by calculating the total energy of the resulting structures. As in the previous section, if different filtration methods are used to calculate the final total energies, they should not be compared directly. Instead a measure such as a FE should be calculated for the NF calculation, and for the SF calculation, and the differences in the two resulting formation energies is taken to be a measure of the difference in the structures obtained.

However, the filtration method used for the structural determination does not need to be the same as the filtration method used for the final total energy calculation. For example, AF could be used to calculate a structure, then NF used to provide the total energy. In the examples in this chapter, the formation energies involve a total

energy of a defect structure, and the total energy of bulk silicon. As the bulk silicon structure is already a minimum energy structure, this does not undergo a structural optimisation. Hence the energy of the bulk silicon structure only depends on the filtration method and parameters used in the final total energy calculation. So if the final energy calculations use the same filtration method and parameters, either the total energies or formation energies can be directly compared, as the differences in each will be exactly the same.

There is the possibility of two different structures having very similar energies however, so another method is required to supplement this one. To really ensure two structures are the same, it is necessary to compare the positions of the atoms in each structure. This can be done either using the atomic coordinates themselves, or using bond lengths and bond angles. Using atomic coordinates introduces some complications. If a structure was to expand slightly, whilst maintaining the same geometric centre, the change in position of outer atoms would be greater than that observed for atoms closer to the centre, and for a large cluster, the maximum difference of coordination would increase with cluster size. This would not be seen when using bond angles and lengths. Rotations of one of the two structures relative to the other, whilst not changing the actual structure or its total energy, can also produce this effect when using absolute positions as opposed to the relative nature of bond angles and lengths. However, both of these effects can only occur for clusters, not unit cell based calculations performed at constant volume. Alternatively, if two atoms were to switch positions in one structural optimisation, but not the other, bond length and angle information would not instantly reveal this. Atomic position comparison would. All of the results presented in this chapter are unit cell based, so atomic position shifts are analysed. This is done in the following way:

1. Rotations in unit cells produce energy changes, translations do not. To ensure translations did not take place in one of the structures being compared, when comparing a set of structures, the geometric centre of the first structure is

calculated. Then the geometric centre of all of the other structures are forced to be the same as the first. In practice it was found the shift in geometric centres between structures was insignificant.

2. The final position of each atom is given in atomic units, with each atom being numbered. This allows direct comparison of the position of atom 1 in structure 1, to that of atom 1 in structure 2 for example. The change in the x, y and z coordinate for each atom 'pair' is then calculated, creating a list of Δx, Δy and Δz values.

3. The maximum absolute value and standard deviation of each of these three sets of numbers is calculated. This way, an indication of the largest individual shift, and the average shift, is provided.

Having provided the analysis framework, the first set of results is now presented. The first structures to be looked at involve unit cells of silicon with a single interstitial silicon atom.

5.3 Single Silicon Interstitial in Bulk Silicon

This section presents formation energies of a single silicon atom interstitial in 216 atom silicon unit cells, for calculations performed using NF, SF and AF. Three different positions of the interstitial atom are modelled, creating three defect structures, detailed in the next section. Once a total energy of the optimised defect structure X is calculated, along with the total energy of the bulk silicon unit cell (using the same filtration method and parameters), the FE is given by:

$$\text{Formation Energy of X} = \text{E[X]} - \frac{217}{216}\text{E(216 bulk)} \tag{79}$$

5.3.1 Details of Systems Modelled

Three defect structures are examined in this section.

1. The interstitial is placed in the centre of a tetrahedral cage in the silicon structure. This structure is referred to as the tetrahedral defect, or T_d for short. [42]

2. A split interstitial is created, replacing an atom with a pair of silicon atoms along the [110] direction. This structure is referred to as the [110] defect, or [110] for short. [42]

3. The interstitial is placed at a hexagonal site, the point at the centre of one of the hexagonal chair structures in the silicon lattice. This structure is referred to as the H defect, or H for short. [42]

The calculation details are as follows:

- The supercells were made using a $3 \times 3 \times 3$ grid of primitive cubic cells of 8 silicon atoms, with a lattice parameter of 10.195 a.u..
- The k-point sampling of the Brillouin zone used an MP 2 2 2 sampling grid [45].
- The Fermi-Dirac function used in the filtration process had a Fermi energy of 0.2 Ha (roughly in mid-gap), with kT set to 0.1 Ha.
- The temperature used to fill the Kohn-Sham levels was 0.04 eV.
- The pseudopotentials for silicon were as presented in Hartwigsen, Goedecker and Hutter (1998) [29].
- The optimisation routine used for the structural optimisations was the DIIS algorithm, outlined in section 5.1.3. The structure was considered to be optimised when the maximum component of the force was less than 10^{-4} Ha/a.u..

5.3.2 Comparison of Formation Energies — Use of Same Filtration Method Throughout Calculation

The results in this section all use the same method of filtration for both the structural optimisation and final total energy calculation. SF and AF calculations had a filtration

Table 22: FEs of [110], T_d and H interstitials in unit cells of 216 atoms of silicon, calculated using three filtration methods. Both the structure and final energy within an individual calculation used the same filtration method. Significant differences when changing filtration method are seen. SF and AF calculations had a filtration radius R_{cut} of 10 a.u. and AF calculations used a value of τ of 6.

Filtration Method	FE[110] eV	FE[T_d] eV	FE[H] eV
NF	3.541	3.691	3.648
SF	3.565	3.722	3.668
AF	3.626	3.788	3.751

radius R_{cut} of 10 a.u. and AF calculations used a value of τ of 6. The formation energies of each of the three defects, for NF, SF and AF calculations, are shown in table 22.

The differences between NF and SF formation energies are 24, 31 and 20 meV for the [110], T_d and H defects respectively. These are over twice the 10 meV threshold, below which is considered an acceptable calculation. When comparing AF to NF the differences are much larger - 85, 97 and 103 meV. These differences are far too large, about an order of magnitude too large. These differences can be thought of as arising from two distinct parts of the calculation. Firstly the equilibrium positions of the atoms in the defect structures due to the change in the PES. Secondly, the final energy calculation itself will contribute to the difference. It is possible using further calculations to separate these two factors. This is examined in the next section, where the filtration method for the structural optimisation is varied separately to the filtration method for the final total energy calculation.

5.3.3 Comparison of Formation Energies - Use of Different Filtration Methods for Calculating Structure and Total Energy

The discussions that follow are going to use NF, SF and AF for finding the optimised positions of the atoms. Then the final defect structure's energy will be calculated using either NF, SF or AF. To simplify matters first the method for optimisation will be listed, then the method for the final energy. For example AF/SF means an optimisation done using advanced filtration, and the final energy calculated using standard filtration. Under this nomenclature, in the previous section we discussed NF/NF, SF/SF and AF/AF calculations. In this fashion, a table for each defect structure can be presented where the data along a row shows the same structure, but varying the filtration method for only the final energy. Similarly the data down a column shows three different structures obtained through three filtration methods, but using the same filtration method for the energy. By looking at the spread of the results in a column compared to that of a row the relative effects on the final energy of the filtration methods from the positional and final energy portions of the calculation can be assessed.

This information is presented in tables 23, 24 and 25. Looking initially at the [110] results in table 23, it can be seen that changing the filtration method for the structural optimisation part of a calculation between NF, SF and AF results in a change in the final energy of 5 meV or less, regardless of which filtration method is used to calculate the final energy. However when the same filtration method is used for the structural optimisation, and the filtration method for just the final energy calculation is varied instead, much larger differences of 79-89 meV are observed, about 40 times the size seen when varying the method used for the structural optimisation. This is an important result, as it indicates that forces are much less sensitive to filtration than energies. In a structural optimisation, most (typically 90% or more - for large systems far from an equilibrium structure this can rise to over 99%) of the effort is spent calculating the structure. This means that the highly efficient AF

method, using extremely aggressive values of the AF parameter τ, can be used to derive accurate structures, and then either a NF calculation, or a SF calculation with a suitable value of R_{cut} performed at the end to produce an accurate total energy.

Looking at the results for T_d in table 24, we see a 95-97 meV spread for the final energy calculation dependence on the filtration method, and only a 1-2 meV variance when varying the filtration method for the structural optimisation. In table 25 for the H defect, a 101-117 meV energy spread compares to 1 meV for structural calculations using the 3 filtration methods when the final energy is calculated using NF or SF, and 17 meV when using AF for the final energy.

These results are impressive, especially when it is noted that for these calculations R_{cut} is set to 10 a.u. a setting shown in the previous chapter to produce significant efficiency gains, and that τ is 6, a low value, which again has been shown to greatly reduce the time required to produce the filtered functions.

The formation energies have been shown to be very close using this technique of AF then NF/SF. There is a chance however that the structures, although close in energy, are different in terms of the positions of the atoms. This is examined next.

5.3.4 Comparison of Atomic Positions

In the output from AIMPRO each atom is numbered. This allows us to compare the positions of each atom resulting from a structural optimisation using NF against one performed using SF or AF. Each x, y and z component is analysed individually. The maximum absolute change observed in any pair of atoms is recorded in table 27, as well as the standard deviation of the absolute change.

The maximum shifts are reasonably low. In the SF calculation, a maximum shift of 0.288 pm, 0.097 pm and 0.406 pm were seen for the [110], T_d and H defects respectively. For AF, apart from a value of 1.677 pm for the z shift of one atom in [110] defect, all the shifts are under 1 pm for the [110] defect, and under 0.5 pm for the T_d and H defects. This is compared to a bulk silicon-silicon bond length in this system of

Table 23: FE of [110] interstitial in silicon, performed using three filtration methods to calculate the structure (for example pos NF indicates no filtration was used to calculate the minimum energy structure), then three filtration methods to calculate the final energy (for example energy SF indicates standard filtration was used to calculate the final total energy). The variation in the formation energies across rows, and then down columns, can be seen by examining the max-min data. This shows varying the method used for the structure results in far less variation in FE than observed when varying the method used for the final total energy calculation. SF and AF calculations had a filtration radius R_{cut} of 10 a.u. and AF calculations used a value of τ of 6.

[110] eV	Energy NF eV	Energy SF eV	Energy AF	max-min
pos NF	3.541	3.564	3.630	0.089
pos SF	3.543	3.565	3.630	0.087
pos AF	3.547	3.570	3.626	0.079
max-min	0.005	0.005	0.004	

Table 24: FE of T_d interstitial in silicon, performed using three filtration methods to calculate the structure (for example pos NF indicates no filtration was used to calculate the minimum energy structure), then three filtration methods to calculate the final energy (for example energy SF indicates standard filtration was used to calculate the final total energy). The variation in the formation energies across rows, and then down columns, can be seen by examining the max-min data. This shows varying the method used for the structure results in far less variation in FE than observed when varying the method used for the final total energy calculation. SF and AF calculations had a filtration radius R_{cut} of 10 a.u. and AF calculations used a value of τ of 6.

$T_(d)$	Energy NF	Energy SF	Energy AF	max-min
pos NF	3.691	3.722	3.788	0.097
pos SF	3.691	3.722	3.788	0.097
pos AF	3.693	3.723	3.788	0.095
max-min	0.002	0.001	0.000	

Table 25: FE of H interstitial in silicon, performed using three filtration methods to calculate the structure (for example pos NF indicates no filtration was used to calculate the minimum energy structure), then three filtration methods to calculate the final energy (for example energy SF indicates standard filtration was used to calculate the final total energy). The variation in the formation energies across rows, and then down columns, can be seen by examining the max-min data. This shows varying the method used for the structure results in far less variation in FE than observed when varying the method used for the final total energy calculation. Only the AF/AF to AF/NF formation energies differs by more than the 10 meV accuracy threshold, and only by a small margin. SF and AF calculations had a filtration radius R_{cut} of 10 a.u. and AF calculations used a value of τ of 6.

H	Energy NF	Energy SF	Energy AF	max-min
pos NF	3.648	3.667	3.766	0.117
pos SF	3.649	3.668	3.749	0.100
pos AF	3.650	3.668	3.751	0.101
max-min	0.001	0.001	0.017	

Table 26: For reference purposes, the individual total energy data behind the formation energies seen in tables 22 to 25.

	Energy NF	Energy SF	Energy AF
Si	-856.29060	-856.22769	-856.09459
110 pos[NF]	-860.12477	-860.06072	-859.92460
110 pos[SF]	-860.12470	-860.06070	-859.92460
110 pos[AF]	-860.12457	-860.06052	-859.92474
Td pos[NF]	-860.11926	-860.05493	-859.91879
Td pos[SF]	-860.11925	-860.05493	-859.91878
Td pos[AF]	-860.11920	-860.05488	-859.91879
h pos[NF]	-860.12083	-860.05695	-859.91960
h pos[SF]	-860.12081	-860.05691	-859.92021
h pos[AF]	-860.12078	-860.05690	-859.92014

Table 27: Maximum observed differences in atomic positions for structures optimised using two filtration methods. For each of the three structures ([110], T_d and H interstitials in unit cells of 216 atoms of silicon) NF vs SF, and NF vs AF results are presented. All values are in picometers. SD(x) refers to the standard deviation of the change in the x-coordinate of each atom. SF and AF calculations had a filtration radius R_{cut} of 10 a.u. and AF calculations used a value of τ of 6.

	Max. Diff.(x)	Max. Diff.(y)	Max. Diff.(z)	SD(x)	SD(y)	SD(z)
110 NF vs SF	0.288	0.288	0.253	0.052	0.052	0.104
110 NF vs AF	0.872	0.872	1.677	0.217	0.217	0.394
T_d NF vs SF	0.097	0.097	0.097	0.035	0.035	0.035
T_d NF vs AF	0.444	0.444	0.444	0.107	0.107	0.107
H NF vs SF	0.406	0.406	0.406	0.112	0.112	0.112
H NF vs AF	0.380	0.380	0.380	0.112	0.112	0.112

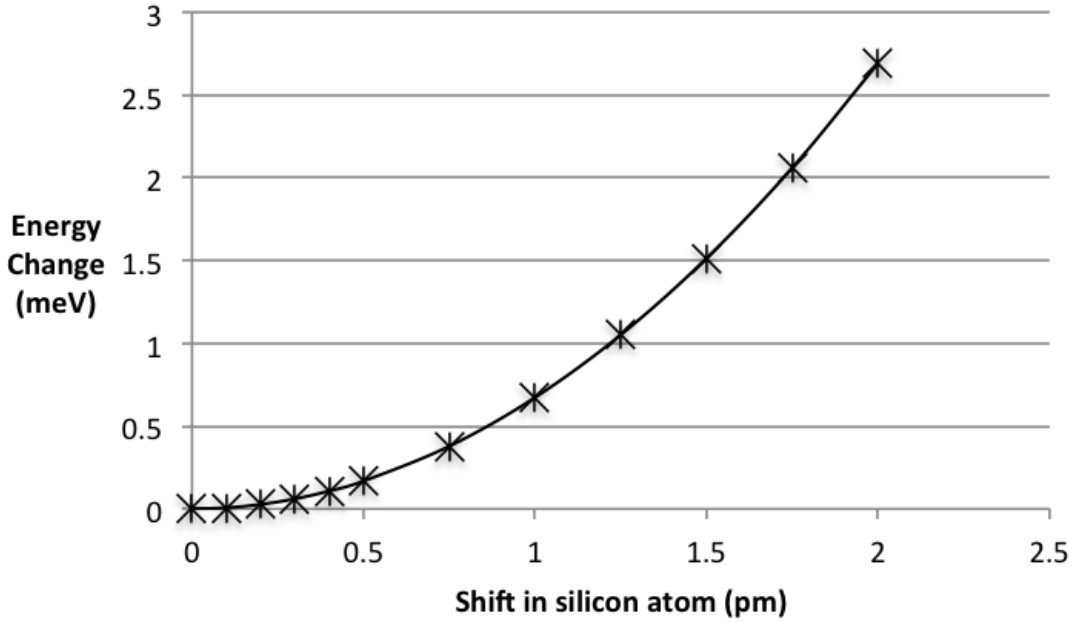


Figure 9: Graph to show change in energy in a unit cell of bulk silicon as one silicon atom is moved towards its second nearest neighbour. The quadratic nature of displacements of atoms at this level can be clearly seen. The changes in position recorded in table 27 produce negligible changes in the total energy of the system.

234 pm. This is the maximum shift; the standard deviation which gives an indication of the shift per atom pair is even lower in each case. For the SF calculations the values are around 0.1 pm, an insignificant amount. For AF calculations the standard deviations are also extremely low, 0.217 pm, 0.107 pm and 0.112 pm for the [110], T_d and H defects respectively.

The change in energy for shifts of 0.1 pm to 2 pm for one atom in a unit cell of silicon are shown in figure 9, to give a sense of the energy changes that variations in this range could generate. A shift of 0.1 pm leads to an energy change of less than a hundreth of 1 meV, 0.4 pm about a tenth of 1 meV. Even though these changes have to be summed over all the atoms, the differences are so small that they can be neglected around the 0.1 pm range, and cause only small differences in the 1-2 pm range.

5.3.5 Conclusions for Single Silicon Interstitial in Bulk Silicon

It is clear that the structures produced when using either SF or AF during the structural optimisations performed in this section are very close to those produced when performing an NF calculation. The differences in formation energies, maximum observed differences of atomic positions, and standard deviations of differences of atomic positions are all extremely low. In the next section, we revisit the reaction analysed previously in section 4.5.

5.4 Oxygen defects in bulk silicon

This section will compare results using NF, SF and AF for formation energies for the reaction (75), and the positional changes in the optimised structures O_i, VO_2 and VO. Initially NF is compared against SF for formation energies, and then changes in the three structures. Then SF is compared against AF using values of τ of 5, 6, 8 and 10, again for formation energies and changes in the three structures. Finally the structures resulting from AF/SF and SF/SF calculations are compared.

5.4.1 Details of Systems Modelled

The three systems were described in section 4.5. The calculation details were as seen in section 4.5, with the following additions:

- The tolerance for the SCF iterative process was set to 1.0×10^{-7} Ha.
- The optimisation routine used for the structural optimisations was the DIIS algorithm, introduced in section 5.1.3. The structure was considered to be optimised when the maximum component of the force was less than 10^{-4} Ha/a.u..

5.4.2 Comparison of Reaction Energy - Standard Filtration

Table 28: Total energy and FE from reaction (75) results for SF/SF and SF/NF calculations, with comparison to NF/NF results. R_{cut} was set to 10 a.u..

	SF/SF	SF/NF	NF/NF
O_i	-872.26800	-872.33111	-872.33111
Si 216	-856.22763	-856.29051	-856.29051
VO_2	-884.33844	-884.40231	-884.40232
VO	-868.24039	-868.30386	-868.30387
Energy (Ha)	-0.05768	-0.05785	-0.05785
Energy (eV)	-1.570	-1.574	-1.574
Diff. to NF/NF (meV)	5	0	-

Table 28 shows the formation energies using SF/SF, SF/NF and NF/NF. Using SF for both positions and final energy gives a 5 meV difference to the NF result. This is a very good result because it is less than the 10 meV target. If an SF/NF calculation is performed, the difference drops to less than 1 meV, which is an excellent result. To see how the structures differ in terms of atomic position, table 29 shows the standard deviation in the x, y and z coordinates varies from 0.024-0.028 pm for

Oi, to 0.051-0.084 pm for VO and 0.056-0.093 pm for VO2. The maximum shift of any atom in the Oi case was 0.21 pm, for VO2 0.435 pm and 0.525 pm for VO. These are all incredibly small differences and would be unnoticable in a graphic representation of the unit cell, and of the order of 0.23% of a silicon silicon bond length for the maximum shift, and 0.04% or less for the standard deviation. This shows that SF can be applied to the whole calculation in this case without any significant change in the formation energy being produced, and negligible change if the final energy is calculated using no filtration.

Table 29: Maximum and standard deviation (SD) of the differences of position of atoms produced using NF and SF optimisation of the three defect structures. All values in picometers.

	Max. Diff.(x)	Max. Diff.(y)	Max. Diff.(z)	SD(x)	SD(y)	SD(z)
O_i	0.210	0.210	0.105	0.028	0.028	0.024
VO_2	0.435	0.234	0.234	0.093	0.056	0.056
VO	0.525	0.256	0.256	0.084	0.051	0.051

5.4.3 Comparison of Reaction Energy - Advanced Filtration

Comparing SF/SF against AF/AF, the results in table 30 show the formation energy resulting when both the position of the atoms through structural relaxation, and the resulting energy of this structure are calculated using AF with varying values of τ, i.e. AF/AF. This is compared to the results using SF/SF. Large differences in formation energy are obtained for $\tau = 5$ and $\tau = 6$. $\tau = 8$ is a borderline acceptable result, whilst $\tau = 10$ is a good result. Now the structures determined using AF are put into a SF energy calculation, i.e. an AF/SF calculation, the results being shown in table 31.

The results for $\tau = 5, 6$ are now very good at 10 meV and 3 meV respectively, and

Table 30: Total energy and FE from reaction (75) results for four sets of AF/AF calculations, with comparison to SF/SF results. R_{cut} was set to 10 a.u., and the AF parameter τ between 5 and 10.

	$\tau = 5$	$\tau = 6$	$\tau = 8$	$\tau = 10$	SF
O_i	-872.09591	-872.13112	-872.25021	-872.26652	-872.26800
Si 216	-856.06551	-856.09330	-856.21076	-856.22667	-856.22763
VO_2	-884.16367	-884.20185	-884.32141	-884.33675	-884.33844
VO	-868.07224	-868.10764	-868.22357	-868.23897	-868.24039
Energy (Ha)	-0.06103	-0.05639	-0.05839	-0.05793	-0.05768
Energy (eV)	-1.661	-1.534	-1.589	-1.576	-1.570
Diff. to SF (meV)	-91	35	-19	-7	-

Table 31: Total energy and FE results for reaction reaction (75). Structures were optimised using SF, and then AF with 4 values of the AF parameter τ. The final structures obtained were in all cases calculated using SF. A comparison of AF/SF results to SF/SF results is also presented. Using SF for the final energy greatly reduces the differences in FE between SF and AF. R_{cut} was set to 10 a.u., and the AF parameter τ between 5 and 10.

	$\tau = 5$	$\tau = 6$	$\tau = 8$	$\tau = 10$	SF
O_i	-872.26776	-872.26788	-872.26799	-872.26801	-872.26800
Si 216	-856.22763	-856.22763	-856.22763	-856.22763	-856.22763
VO_2	-884.33826	-884.33835	-884.33844	-884.33844	-884.33844
VO	-868.24010	-868.24032	-868.24039	-868.24039	-868.24039
FE (Ha)	-0.05803	-0.05778	-0.05769	-0.05767	-0.05768
FE (eV)	-1.579	-1.572	-1.570	-1.569	-1.570
Diff. to SF (meV)	-10	-3	0	0	-

for $\tau = 8, 10$ the results are less than 1 meV, an excellent result, an order of magnitude less than the 10 meV acceptance threshold. The structures are now analysed from an atomic position perspective.

Tables 32, 33 and 34 show the positional shifts in the AF structures compared to the SF ones. The differences are larger than seen when comparing SF to NF in table 29. The results for O_i and VO_2 are still acceptable for all values of τ, although τ=5 or 6 produce much larger differences than 8 or 10. All measures are under 1.5 pm however, and this is a small distance, approximately 0.6% of a silicon silicon bond length. The VO results for τ=5 show differences that could be classed as significant. One atom (the oxygen atom) is translated nearly 6 pm in the x-direction, and the standard deviations are approximately 20 times larger than seen if τ is increased to 8 or 10. Overall, a value of 6 or above for τ would be an appropriate choice, which is an excellent result. At this level, results from table 4 indicate a reduction in average SCF times from a SF calculation of a factor of 3. If greater precision in the final structure was required, one the maximum component of the force vector had dropped beneath a certain value, for example 10 times the exit force tolerance, the value of τ could be raised to 10. Using this method, the same minimum on the PES would be located using aggressive and hence efficient filtration parameters, then the final few iterations performed using more accurate filtration parameters.

5.4.4 Conclusions for Oxygen Defects in Silicon

The FEs and positional analyses have shown using an AF calculation with a low value of τ (as low as 6) and reasonable value of R_{cut} (10 a.u.) to produce a minimum energy structure, followed by a total energy calculation on the resulting structure using SF, produces a result equivalent to that of using NF throughout the calculation. The NF/NF result for this FE was 1.574 eV, an AF/AF calculation (τ was 6, R_{cut} was 10 a.u.) produces a result of 1.572 meV. The two sets of structures obtained using NF and AF differed from each other in positional terms by under 2 pm, with the average

Table 32: Maximum and standard deviation (SD) of the differences of position of atoms produced using AF and SF optimisation of the O_i defect structure. All values in picometers. The changes in position are extremely small (less than 1 pm).

τ	Max. Diff.(x)	Max. Diff.(y)	Max. Diff.(z)	SD(x)	SD(y)	SD(z)
5	0.636	0.636	0.628	0.143	0.143	0.141
6	0.874	0.874	0.338	0.115	0.115	0.104
8	0.594	0.594	0.111	0.047	0.047	0.023
10	0.097	0.097	0.038	0.013	0.013	0.010

Table 33: Maximum and standard deviation (SD) of the differences of position of atoms produced using AF and SF optimisation of the VO_2 defect structure. All values in picometers. The changes in position are extremely small (around 1 pm for $tau = 5$, less than 1 pm for greater values of τ).

τ	Max. Diff.(x)	Max. Diff.(y)	Max. Diff.(z)	SD(x)	SD(y)	SD(z)
5	1.402	0.888	0.888	0.293	0.202	0.202
6	1.004	0.290	0.290	0.165	0.064	0.064
8	0.127	0.085	0.085	0.033	0.021	0.021
10	0.042	0.041	0.041	0.016	0.012	0.012

Table 34: Maximum and standard deviation (SD) of the differences of position of atoms produced using AF and SF optimisation of the VO defect structure. All values in picometers. τ=5 produces changes that are becoming significant. τ=6 or above produces changes in position that are extremely small (around 1 pm or less).

τ	Max. Diff.(x)	Max. Diff.(y)	Max. Diff.(z)	SD(x)	SD(y)	SD(z)
5	5.892	2.685	2.685	0.517	0.403	0.403
6	1.062	0.427	0.427	0.160	0.093	0.093
8	0.122	0.091	0.091	0.023	0.018	0.018
10	0.063	0.202	0.202	0.013	0.024	0.024

shift much lower than this. This demonstrates the power of filtration in terms of efficient calculations. If further precision is required, the final structure from an AF structural optimisation can be tidied up using a more precise one. The possibility also exists to link the value of τ during a calculation to the maximum component of force, which would achieve this automatically.

5.5 Conclusions

The concept of using AF with low, extremely aggressive parameters, followed by an NF or SF total energy calculation has been proved to be successful for modelling defects in silicon. Extremely low values of the AF parameter τ, that proved unusable for energy calculations, can be used for structural optimisations. This is particularly significant, as the vast majority of CPU time required in modelling is used in this activity. The calculation of a final energy with improved R_{cut} adds an insignificant overhead. It should also be noted that $\tau = 10$ also produces excellent structures, with differences of only 10^{-4} Å— insignificant in real calculations. Even more approximate values such as $\tau = 6$ produced structures good enough to give final energy differences differing by only 1 meV.

The next chapter looks at different methods of removing functions in the AF process, to see if a more efficient method than the overlap-based one employed in this and the previous chapter is possible.

Chapter 6

INVESTIGATION OF ADVANCED FILTRATION METHODS

In this chapter, three different methods of AF will be compared against one another. The specific measure of each method will be its ability to produce FEs for four different systems, each consisting of a defect in a unit cell of 64 silicon atoms. The goal is to produce AF FEs as close to the NF FE, whilst using filtered functions created from the minimum number of primitive basis functions.

One method of AF has already been introduced and explained, in section 4.2. The work presented in this chapter alters this method, by introducing more parameters, making the method more flexible. Specifically, a parameter is introduced for s-type functions, another for p-type, and another for d-type. This introduces the concept of differentiating between functions based on their angular momentum, without the complications of building this into the actual calculation of the overlap integral, outlined in section 4.2.1. This method is referred to as toltrim.

This concept of separate parameters will also be included as an extension to the SF method considered previously, by allowing different values of R_{cut} for s-type, p-type and d-type functions. This will be the second method, referred to as autofilt.

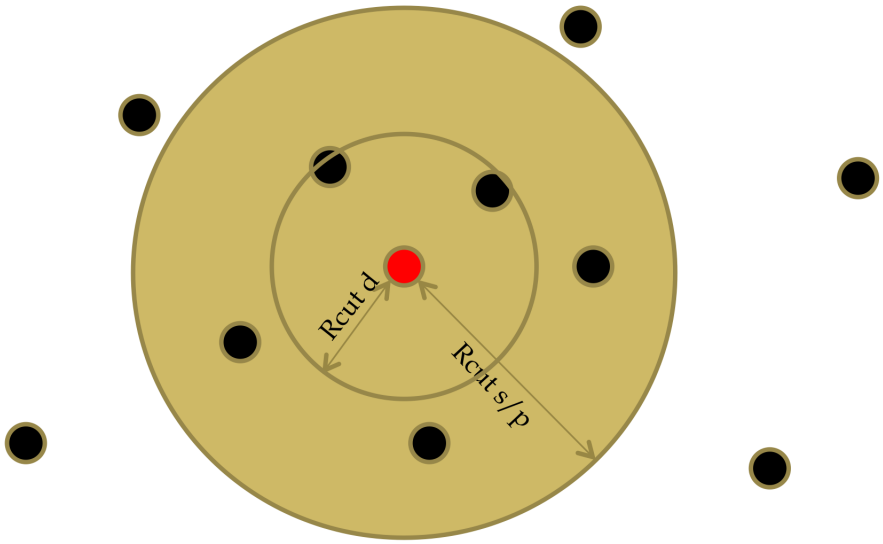
A third method, referred to as radtrim, is also detailed. This also will have different parameters for s-type, p-type and d-type functions.

The next section will explain these three methods in detail.

6.1 Explanation of Advanced Filtration Methods

The three methods are autofilt, toltrim and radtrim. Toltrim and autofilt are variations of the AF and SF processes used in previous chapters, hence the explanations

Figure 10: The autofilt method employing differing radii for s, p and d functions. The two atoms inside the inner sphere of radius $R_{\text{cut-d}}$ have s, p and d functions kept. The three atoms within the outer sphere of radius $R_{\text{cut}} = R_{\text{s}} = R_{\text{p}}$ only, have just s and p functions retained for the creation of the filtered functions.



will only show the differences arising from differentiating primitive basis functions based on their angular momenta. Radtrim however is a new technique, and will be explained in greater detail.

6.1.1 Autofilt

S, p and d functions can each be assigned a different value of R_{cut}, giving R_{s}, R_{p} and R_{d}. Internally, within the code, R_{cut} is still used to generate the initial list of functions, and is consequently set to be the largest value from R_{s}, R_{p} and R_{d}. An illustration is provided by figure 10.

6.1.2 Toltrim

The required overlap τ between the two Gaussians can be set for s functions, another value for p and for d. Along with the value for the exponent of the Gaussian trial function α_{Fix}, we have τ_{s}, τ_{p} and τ_{d}, a total of four adjustable parameters. R_{cut} is set

to be just large enough to include all atoms whose basis functions have a chance of being included.

6.1.3 Radtrim

The third method analysed is a form of compromise between these two methods. A fixed radius R_{sphere} is drawn around the central atom. All functions inside this sphere are kept. Functions outside this sphere, but inside the sphere of radius R_{cut} are kept if the value of the Gaussian when it touches the sphere of radius R_{sphere} is greater than a tolerance $e^{-\tau}$. Figure 11 shows a 2d representation of this, with the one atom within the inner sphere having all its functions kept. The four atoms between the two spheres are subjected to a further test, shown in figures 12 and 13. In figure 12 the red atom is the atom for which filtered functions are being constructed. The black atom lies between the two spheres. Gaussians A and B are kept as their values at the edge of the inner sphere are greater than the tolerance parameter τ. In figure 13 the atoms are further apart, and only Gaussian A is kept. Note that in practice, $e^{-\tau}$ is very much smaller than shown in these schematic diagrams, as τ typically takes values of 5 and above.

The cut-off value for the basis function where it touches the sphere of radius R_{sphere} can be set for s, p and d functions giving 4 parameters. R_{sphere}, τ_{s}, τ_{p} and τ_{d}. Again R_{cut} is set to be the smallest value than encompasses all possible functions that could be kept.

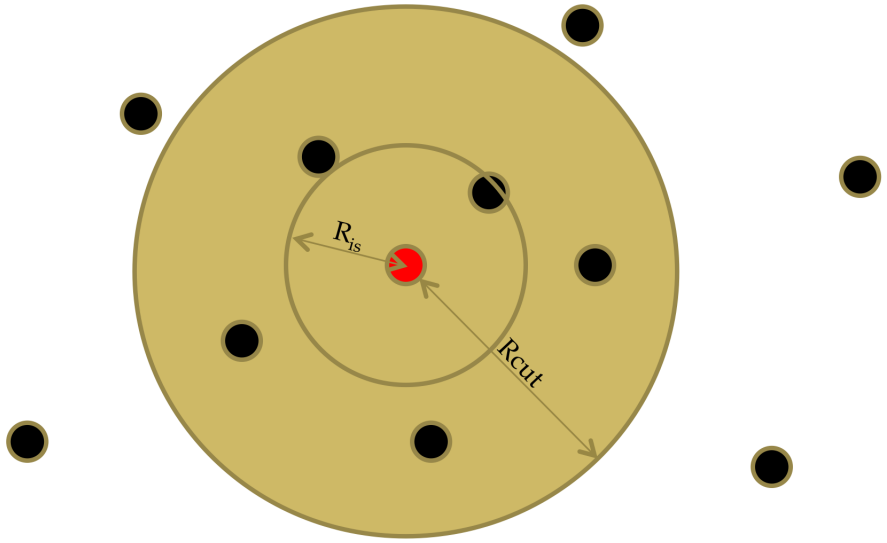
Now the methods used in this chapter have been explained, the structures to which they will be applied are detailed.

6.2 Systems under investigation

The systems chosen for investigation were 4 defects in 64 atoms of bulk silicon. The defects structures chosen were [51, 57]:

- A [110] split interstitial. The extra silicon atom was placed into a unit cell of

Figure 11: The radtrim AF method. All functions on atoms inside the inner sphere of radius R_{is} are kept. Functions outside this sphere, but inside the sphere of radius R_{cut} are kept if their value at the edge of the inner sphere is above a pre-defined tolerance, $e^{-\tau}$.



bulk silicon and the resulting structure optimised without filtration. Referred to as [110] .

- A vacancy structure, formed by removing one silicon atom from a unit cell of bulk silicon, followed by a structural optimisation without filtration. Referred to as V.

- An I_3 tri-interstitial. The three extra silicon atoms were placed into a unit cell of bulk silicon, and then the resulting structure optimised without filtration. Referred to as I_3.

- Amorphous silicon, a non-crystalline form of silicon, containing some atoms without the four-fold coordination of bulk silicon. Referred to as amorphous, or Am.

Figure 12: The central atom for which filtered functions are being created is red. The black atom is one of the surrounding atoms, whose primitive basis functions are subjected to the radtrim test, to decide whether or not they are kept, i.e. included in the process which creates the filtered functions. Primitive basis functions are kept if their value at the edge of the inner sphere (radius R_{is}) is greater than $e^{-\tau}$ (A and B).

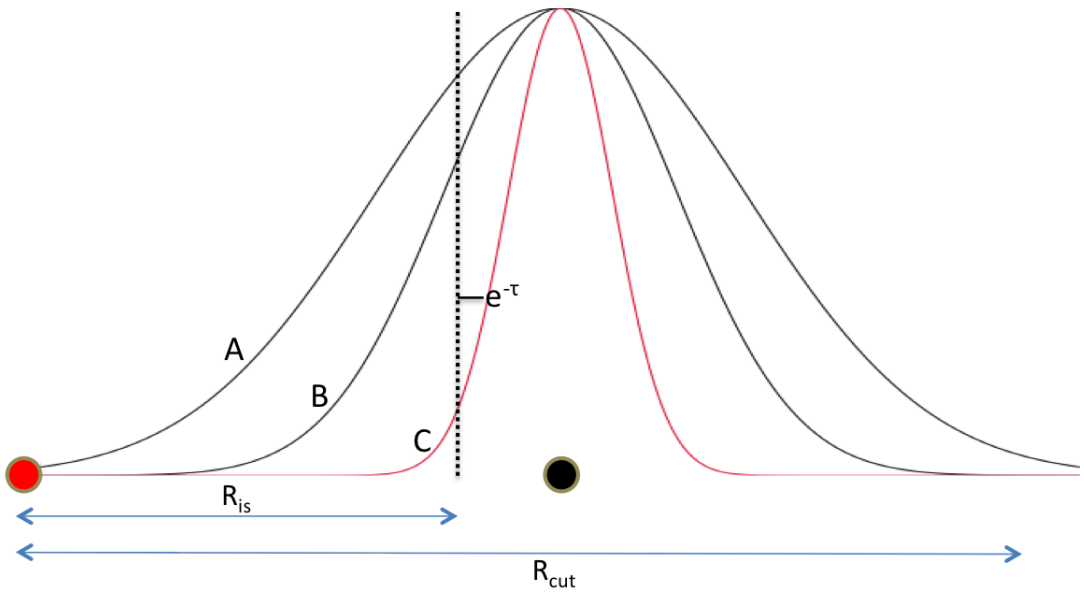
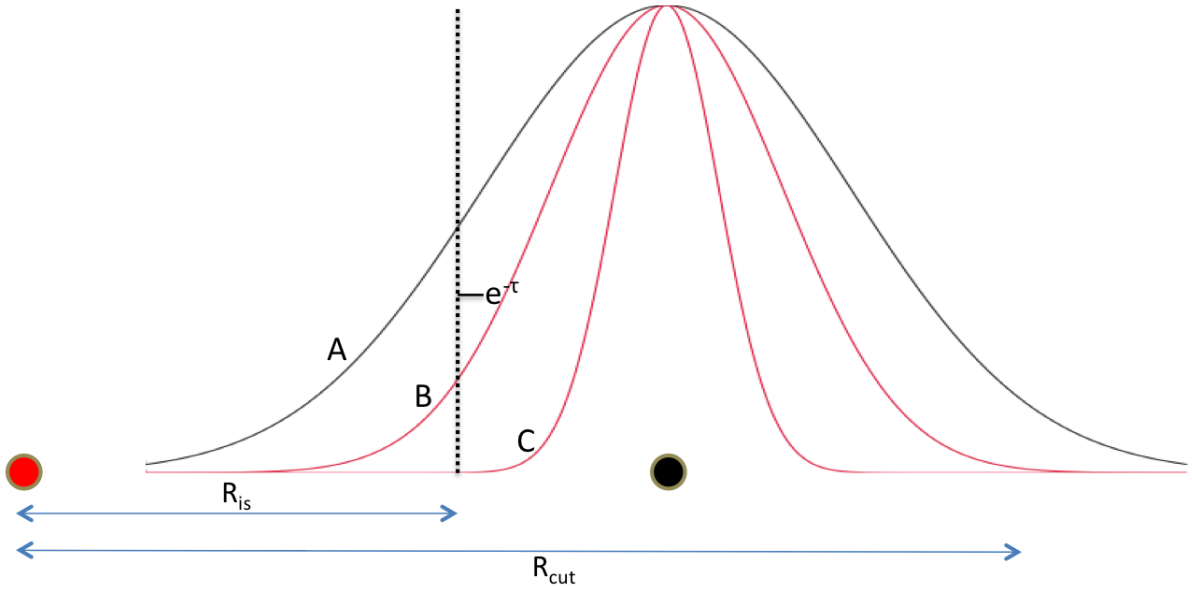


Figure 13: Following on from figure 12. Another surrounding atom is located further away from the red central atom. Now only function A is kept, as both functions B and C have decayed below the tolerance $e^{-\tau}$ at the surface of the inner sphere of radius R_{is}.



6.2.1 Details of Systems Modelled

Of the five structures used in the calculations in this chapter, four required structural optimisations to obtain the minimum energy structures required for the FE calculations. The bulk silicon structure is already relaxed. For the amorphous structure, a data file of a 64 atom unit cell containing a model of amorphous silicon was provided [60]. This required both a structural relaxation, and a lattice optimisation. Details are provided in section 6.2.3. For the vacancy structure, one atom was removed, and then this resulting structure optimised. For the [110] and I_3 interstitial systems, one and three atoms respectively were placed into a unit cell of silicon into roughly the correct place, and the resulting structure optimised. Optimisations were performed using the following parameters:

- The supercells were made using a $2\times2\times2$ grid of primitive cubic cells of 8 silicon atoms, with a lattice parameter of 10.193231 a.u., calculated through a lattice optimisation of the 8 atom unit cell using the same computational parameters as the structural optimisations.
- K-point sampling of the Brillouin zone used an MP 2 2 2 sampling grid [45].
- No filtration was used for the structural optimisations.
- The filling of the energy levels took place at a temperature with kT set to 0.04 eV.
- The optimisation routine used for the structural optimisations was the CG algorithm, using the newly developed force only based line minimiser (section 5.1.3). The structure was considered to be optimised when the maximum component of the force was less than 10^{-5} Ha/a.u., i.e. less than 1meV Å.
- The tolerance for the SCF iterative process was set to 1.0×10^{-9} Ha. This smaller SCF tolerance was required because of the lower maximum force component tolerance in the previous item.

- A 'ddpp' basis set was employed, with exponents values of 0.16145, 0.46343, 1.31473 and 3.75324.
- The pseudopotentials for silicon were as presented in Hartwigsen, Goedecker and Hutter (1998) [29].

The total energy calculations used the same relevant parameters as used for the structural optimisations. For the total energy calculations using filtration, the Fermi-Dirac function used in the filtration process had the Fermi energy and kT optimised by AIMPRO during each calculation.

6.2.2 Calculation of FEs

The formation energies were calculated using:

$$\mathrm{FE} = \mathrm{E}(110) - \frac{65}{64}\mathrm{E}(64\,\mathrm{bulk}) \tag{80}$$

$$\mathrm{FE} = \mathrm{E}(\mathrm{I_3}) - \frac{67}{64}\mathrm{E}(64\,\mathrm{bulk}) \tag{81}$$

$$\mathrm{FE} = \mathrm{E}(\mathrm{V}) - \frac{63}{64}\mathrm{E}(64\,\mathrm{bulk}) \tag{82}$$

$$\mathrm{FE} = \mathrm{E}(\mathrm{amorphous}) - \mathrm{E}(64\,\mathrm{bulk}) \tag{83}$$

6.2.3 Note on Amorphous Silicon

For amorphous silicon the lattice parameter was determined through lattice optimisation to be 10.25898 au. Calculations using this material therefore used different lattice parameters for the amorphous and bulk calculations, which as they are both the relaxed values is the usual method. This does lead to another question. Should the filtration parameters relating to distances from the central atom, such as R_{cut},

Table 35: Effect of slight increase in R_{cut} for amorphous silicon on FE per atom.

	$R_{\text{cut}} = 8$	$R_{\text{cut}} = 8.051602$	NF
Bulk (Ha)	-253.68663	-253.68663	-253.71240
Amorphous (Ha)	-253.23580	-253.23686	-253.27594
FE (Ha)	0.45085	0.44977	0.43646
$\lvert\Delta$FE$\rvert$ per atom (meV)	6.118	5.659	-

be accordingly increased for the amorphous runs. Theoretically there will always be a different number of basis functions passed to the filtration algorithm for each of the two calculations making up a FE. Without forcing filtration parameters to ensure matching number of presented functions, for say a bulk and an ideal vacancy defect system, this has to be accepted. The scenario when moving to having two different lattice parameters in the two systems making up a FE is no different. However it is useful to see what difference it does make.

Table 35 shows the FE per atom for 64 atoms of amorphous silicon firstly without filtration, secondly with filtration using an R_{cut} of 8 a.u. in both systems, and last using an R_{cut} of 8 a.u. for bulk and 8*10.25898/10.193231=8.051602 for amorphous. A difference of 0.46 meV/atom, although a shift of nearly 10%, is quite small for such a small value of R_{cut}. When this is repeated for a value of $R_{\text{cut}} = 12$ a.u., this difference drops to 0.15 meV/atom, showing the results are converging as R_{cut} is increased.

6.3 Choice of parameter sets and how to interpret the resulting data

It is straightforward enough to increase the accuracy of a filtered calculation by increasing the number of functions presented to the filtration algorithm, either through increasing R_{cut} in a SF or AF calculation, or reducing the amount of trimming in an AF calculation. To quantify which of the three AF methods is the most efficient,

calculations using similar number of functions must be compared against each other. To ensure comparable data is gathered, the number of functions kept, i.e. the N_{keep} values, must be known in advance. As this is not a trivial calculation, pre-work using bulk silicon was carried out, and the resulting parameter sets obtained applied to the bulk and 4 defect structures.

For bulk silicon only, a small set of parameters for each method was created, and AIMPRO altered so that the program stopped after the first N_{keep} calculation. These results gave a rough idea of the link between a choice of parameters, and the resulting value of N_{keep}. This allowed the creation of a large set of parameters, sets of 300 for each method. These were again ran through the altered AIMPRO to obtain a large set of parameters and their resulting N_{keep} values. For the autofilt method a parameter set was chosen that gave a spread of N_{keep} values, and the number of results in ranges of N_{keep} 100 wide (i.e. 400-500, 500-600...) counted. Then it was ensured the parameter sets for the radtrim and toltrim method produced the same count in each band. This would ensure comparable results.

The value of N_{keep} produced by a particular parameter set, would obviously change when the parameter set was transferred to one of the defect structures. However, N_{keep} bands which were under-represented could be filled in by the addition of a few more runs. A total of 85 autofilt, 94 radtrim and 85 toltrim results were obtained for each of the 5 structures (bulk silicon and the 4 defect structures). The range of N_{keep} includes some very low values, and consequently large $|\Delta\text{FE}|$ values. This is deliberate, as it is hoped when they are applied to calculations such as structural optimisations, this would not significantly degrade the calculation, especially near the start when exact forces are not as critical.

The FEs calculated (using equations 80 to 83) using AF were not compared to the corresponding SF result. This was because the calculations were using different values of R_{cut}, which would have produced different SF results for each structure. An unfiltered result does not suffer from this problem, and a fixed baseline simplifies

what is already quite a complex comparison due to the number of variables present in the method. The results are shown in figures 14, 15, 16 and 17.

As well as the obvious goal of as low a difference as possible in AF FEs to the corresponding NF FE, another desirable property of the results is a small spread of $|\Delta\text{FE}|$ for a range of N_{keep}. In practice when choosing the parameters for a calculation we will not have the luxury of a chart showing the most efficient parameters for that particular structure, so an AF with a range and hence uncertainty of $\pm 20\,\text{meV}$ is preferable to a one with a range of $\pm 50\,\text{meV}$.

The results are presented for each of the four defect structures in turn, starting with the [110] interstitial.

6.4 Results - [110] Interstitial

The graph in figure 14 shows the absolute differences in calculated FEs of a [110] interstitial in silicon for the three AF methods, when compared to the NF FE. For each method, a varied parameter set was used to produce calculations with a varied range for N_{keep}.

As N_{keep} increases, it is expected that the FE will approach the corresponding NF calculation. This trend is clearly shown by all three methods. More significant is the change in the variation of the $|\Delta\text{FE}|$s for a small range of N_{keep}. For low values of N_{keep} there is an extremely large range of $|\Delta\text{FE}|$, about 140 meV for autofilt, 100 meV for toltrim and 60 meV for radtrim. The radtrim and toltrim methods show a rapid decrease in this variation as N_{keep} increases. However the autofilt method even for large N_{keep} has a much larger range of values of $|\Delta\text{FE}|$, over 30 meV. Radtrim offers the smallest variation in $|\Delta\text{FE}|$ for all values of N_{keep}.

Having established the the radtrim method produces results in the narrowest range, qualitative analysis of the average $|\Delta\text{FE}|$ is required. This is provided in table 36, which shows the average value of $|\Delta\text{FE}|$ for bands of N_{keep} 100 wide. In the data, the upper range for N_{keep} was ignored due to the sparsity of data for some

methods/structures, with the analysis focusing on the range of N_{keep}=500-1400. In the results alongside the average |ΔFE|, N_{keep} averages are given to show where the average point lies. Due to the different methods having different function selection methods, some averages may lie to the left or right of the band. When for example comparing an average N_{keep} of 1020 for one method, against an average N_{keep} of 1090 for a different method, the difference in the number of functions should be taken into consideration.

Table 36 shows these results for the [110] interstitial. The average |ΔFE| tends to decrease with increasing N_{keep}, though each method does display the odd jump. The average N_{keep} values are quite close to each other, with the maximum variation lying in the 700-800 band, where the average value ranges from 736.7 for autofilt to 767.5 for the radtrim method. This is still not a terribly significant difference, and allows us to directly compare the average |ΔFE| results between the three methods for each band of N_{keep}.

At lower values of N_{keep} of 500-800 the autofilt and radtrim methods perform better than the toltrim method. By N_{keep}=800, the radtrim method is the best performer, coming in at under half the average |ΔFE| of the other two methods. In every band from 800 upwards, radtrim has the lowest average |ΔFE|. The toltrim method is particularly poor at very low values of N_{keep} (500-800), giving average |ΔFE|s of 56.9-79.3 meV.

For the calculation of the formation energy of the [110] interstitial, radtrim is the best method with regards to the two measures of average |ΔFE|, and the range of |ΔFE| results.

6.5 Results - Amorphous Silicon

For the amorphous structure figure 15 shows the distribution of |ΔFE| against N_{keep} for the 3 methods. From a range perspective, autofilt performs significantly worse than the other two methods. Toltrim and radtrim are fairly similar, with radtrim

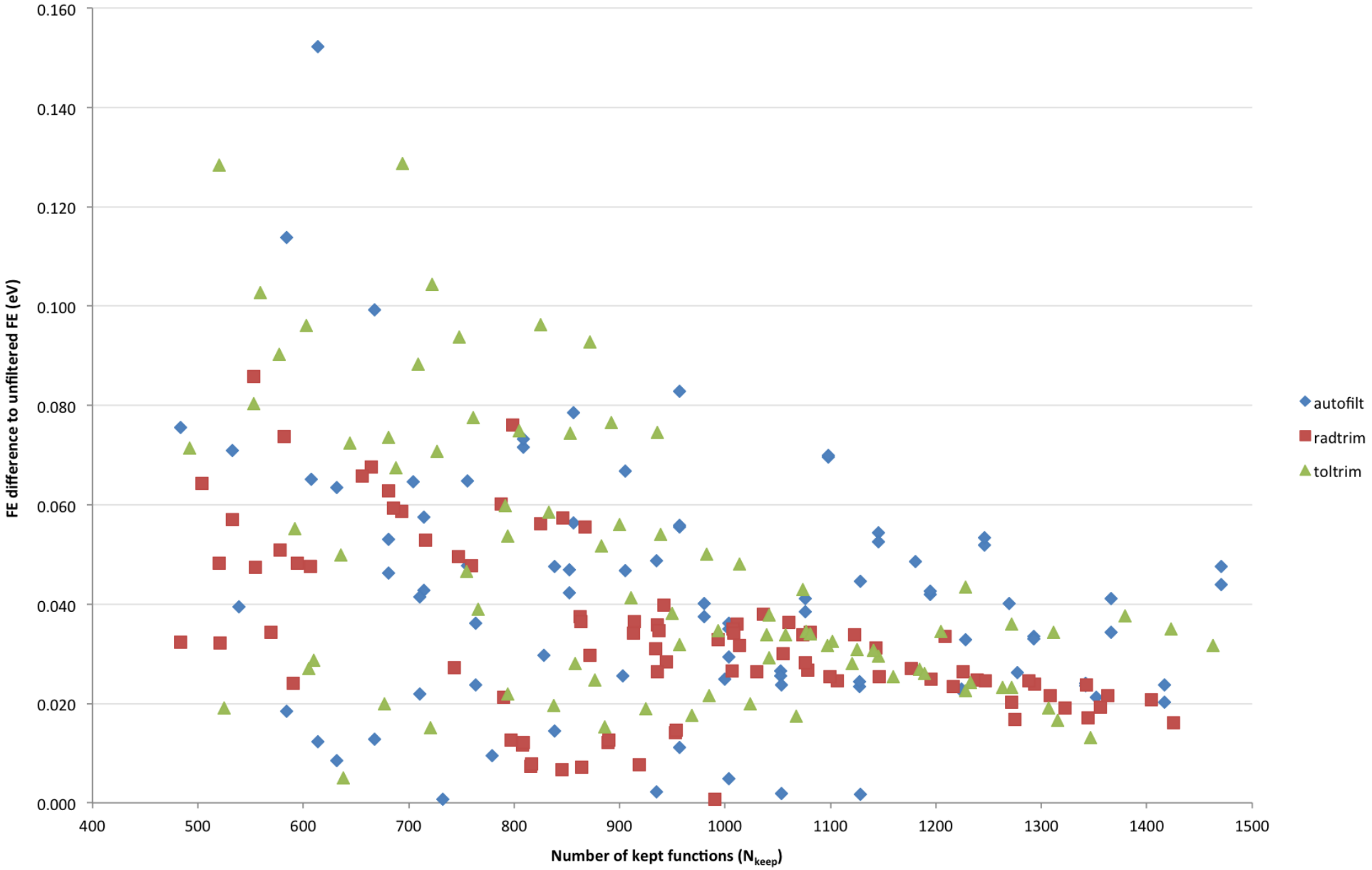


Figure 14: Differences in AF FEs to the NF FE of a [110] interstitial in a unit cell of 64 silicon atoms. The radtrim method outperforms the toltrim and autofilt methods, both in terms of variation of, and average value of $|\Delta FE|$.

Table 36: Average |ΔFE| and N_{keep} values for each AF method, for [110] interstitial in a unit cell of 64 silicon atoms, analysed for 100 wide N_{keep} bands.

	|ΔFE| (meV)			N_{keep}	
N_{keep} range	Autofilt	Radtrim	Toltrim	Autofilt	Radtrim	Toltrim
500-600	60.6	51.5	79.3	560.2	554.5	554.5
600-700	57.0	58.5	56.9	643.9	656.4	647.7
700-800	37.3	43.4	61.0	736.7	767.5	753.5
800-900	51.2	25.0	55.7	837.7	848.2	856.6
900-1000	41.5	25.9	39.9	947.5	943.4	949.8
1000-1100	33.5	31.6	33.0	1048.1	1045.7	1056.0
1100-1200	37.1	27.8	28.8	1152.5	1148.6	1145.8
1200-1300	35.2	24.2	29.6	1255.7	1251.7	1243.1
1300-1400	28.9	20.4	24.1	1353.8	1339.5	1332.1

offering slightly more consistent results. It is also clear looking at the graph, the radtrim results lie closer to the unfiltered result. A qualitative analysis is available in table 37. The maximum range of N_{keep} averages between the three methods is under 20 for every N_{keep} 100-wide band, allowing direct method-method comparison using just |ΔFE| results. All the |ΔFE| values tend to drop with increasing N_{keep} as before. Again radtrim is easily seen as the most efficient method throughout the whole range of N_{keep} values. Toltrim is the next most efficient method, with autofilt the least.

For calculations of the amorphous silicon FE, as for the [110] interstitial FE, radtrim produces the best results in terms of both low, and relatively consistent |ΔFE|.

6.5.1 Results - I_3 Tri-Interstitial

Figure 16 shows the |ΔFE| results against N_{keep} for the I_3 tri-interstitial. Autofilt results again show a large variation in |ΔFE| for the whole range of N_{keep}. Only at

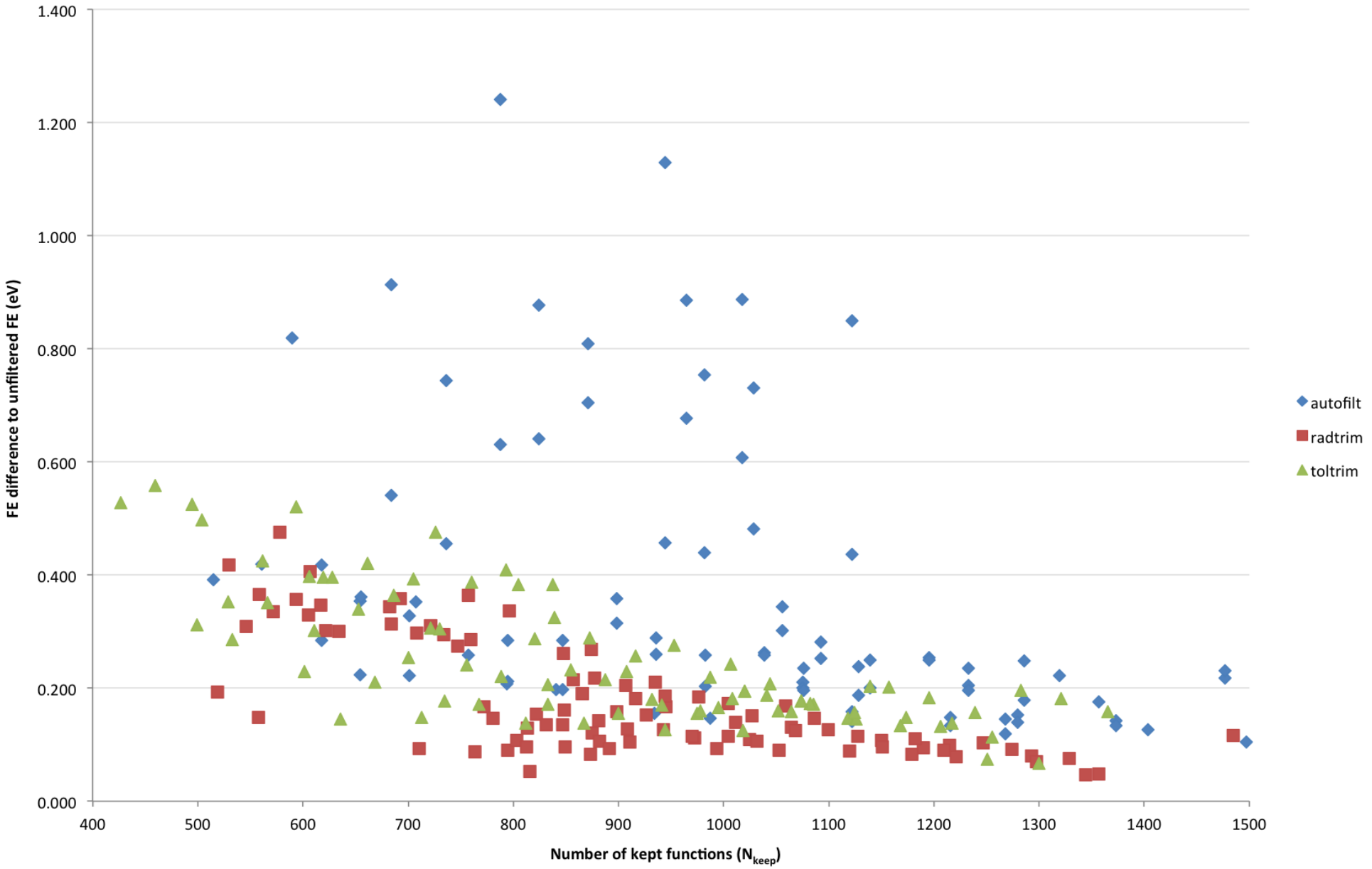


Figure 15: Differences in AF FEs to the NF FE of amorphous silicon from bulk.

Table 37: Average |ΔFE| and N_{keep} values for each AF method, for amorphous silicon, analysed for 100 wide N_{keep} bands.

	|ΔFE| (meV)			N_{keep}	
N_{keep} range	Autofilt	Radtrim	Toltrim	Autofilt	Radtrim	Toltrim
500-600	543.0	325.3	405.4	555.2	556.8	547.9
600-700	442.1	333.6	320.1	652.3	642.0	636.9
700-800	448.7	229.1	290.7	754.3	753.7	741.4
800-900	487.2	146.4	251.7	858.2	853.4	842.0
900-1000	471.3	151.2	190.5	961.7	942.2	948.1
1000-1100	374.9	131.5	180.0	1054.9	1044.6	1045.2
1100-1200	296.6	99.4	164.7	1141.5	1157.1	1150.2
1200-1300	173.0	87.5	125.4	1254.2	1251.0	1250.1
1300-1400	168.4	57.2	169.9	1355.6	1343.3	1343.1

N_{keep}=1200+ does the variation start to reduce, and then only by a small amount from about 230 meV below 1200 to about 160 meV above 1200. Radtrim and toltrim both have variations of up to 150 meV from N_{keep}=500-850. By N_{keep}=850 onwards radtrim becomes more consistent with a variation of about 75 meV to N_{keep}=1030, then reducing to about 30 meV from N_{keep}=1030+. Toltrim starts to become more consistent after N_{keep}=1000, keeping the same level of variation to the highest values of N_{keep}, about 75 meV. Radtrim |ΔFE| values from N_{keep}=850+ are closer to zero than toltrim, but some autofilt results lie closer to zero. From N_{keep}=1250+ radtrim results are closer to zero than most autofilt and toltrim results.

For a numerical analysis of the |ΔFE| values, the I_3 results in table 38 show average N_{keep} values are close to each other for the three methods, apart from in the N_{keep} band 1200-1300 where the autofilt method lies at the upper end at 1275.6, with radtrim in the centre at 1250.7 and toltrim slightly to the lower end at 1246.5. This still does not alter the results, as in this range autofilt performs the worst of the three

Table 38: Average |ΔFE| and N_{keep} values for each AF method, for I_3 tri-interstitial, analysed for 100 wide N_{keep} bands.

	|ΔFE| (meV)			N_{keep}	
N_{keep} range	Autofilt	Radtrim	Toltrim	Autofilt	Radtrim	Toltrim
500-600	163.0	184.3	221.3	546.4	552.1	552.6
600-700	172.8	163.8	175.2	644.3	637.7	653.9
700-800	112.5	152.6	152.6	748.0	739.6	753.1
800-900	105.8	88.0	148.3	860.2	847.9	854.0
900-1000	120.9	90.0	138.8	959.5	948.1	945.2
1000-1100	97.3	97.4	104.3	1050.0	1054.7	1043.7
1100-1200	122.3	89.1	95.2	1156.6	1139.2	1139.1
1200-1300	121.7	76.1	88.5	1275.6	1250.7	1246.5
1300-1400	96.0	61.6	72.5	1360.1	1354.2	1337.5

methods despite this slight advantage of more functions. In the lower N_{keep} range of 500-800 the autofilt method is the most efficient, but by a small margin. There is not much to separate the three methods. Beyond this however, radtrim once again is the most efficient method with the lowest |ΔFE| values. The other two methods in this structure are much closer to the radtrim method in this mid-upper range of N_{keep}, and by N_{keep}=1100 or more, toltrim is a close second to radtrim.

Radtrim is the obvious choice again, it performing better than autofilt and toltrim in both categories of variation, and having results closer to the unfiltered answer (|ΔFE|=0).

6.6 Results - Vacancy

Figure 17 clearly shows a large variation in toltrim |ΔFE| values in N_{keep} values from 400 to 950 of about 130 meV, falling dramatically at N_{keep}=950 onwards to

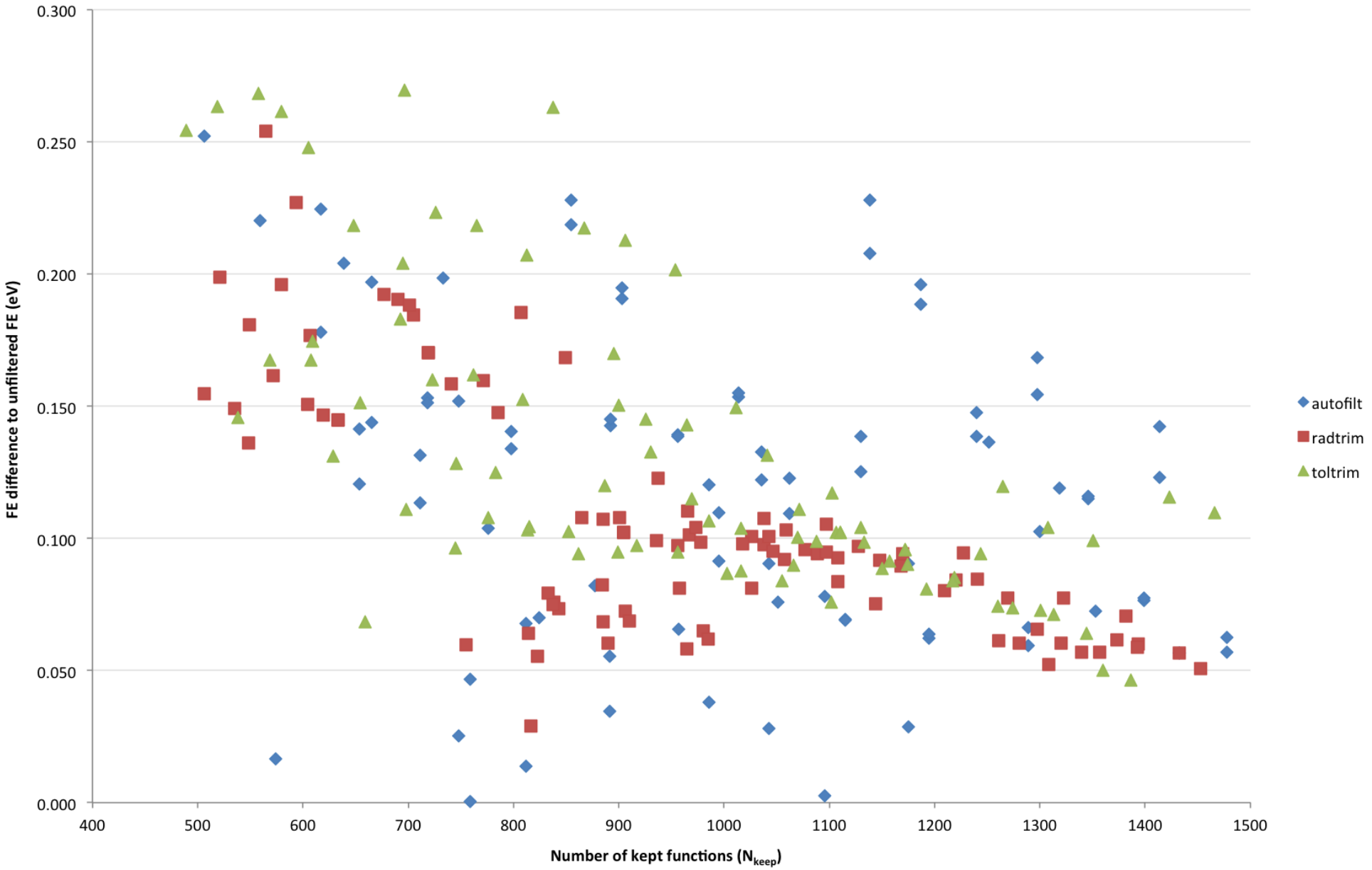


Figure 16: Differences in AF FEs to the NF FE of a I_3 tri-interstitial in a unit cell of 64 silicon atoms.

Table 39: Average |ΔFE| and N_{keep} values for each AF method, for vacancy in silicon, analysed for 100 wide N_{keep} bands.

	|ΔFE| (meV)			N_{keep}	
N_{keep} range	Autofilt	Radtrim	Toltrim	Autofilt	Radtrim	Toltrim
500-600	56.8	39.1	77.0	568.6	548.9	552.5
600-700	54.0	40.6	74.4	663.0	659.6	656.2
700-800	44.8	22.4	59.5	749.0	759.2	750.0
800-900	42.9	16.6	33.4	847.5	852.2	852.0
900-1000	34.6	19.3	31.3	951.3	946.5	946.5
1000-1100	39.6	22.6	23.1	1055.0	1045.7	1043.7
1100-1200	34.1	18.3	20.0	1153.7	1155.2	1153.4
1200-1300	26.3	12.9	16.2	1244.0	1252.3	1255.1
1300-1400	25.5	13.2	30.3	1339.2	1335.0	1359.8

about 25 meV. Autofilt results have a variation for all values of N_{keep} that is fairly consistent, about 75 meV. Radtrim results have a variation of about 55 meV up to N_{keep}=800 where the variation in |ΔFE| falls to about 30 meV, then falls again at N_{keep}=1000 to about 10 meV. The radtrim method again displays the lowest degree of variation.

Qualitatively it is clear the radtrim results lie closer to the |ΔFE|=0 goal, and this is analysed quantitatively in table 39. Table 39 shows similar trends as for the other three defect structures. The N_{keep} averages are close to each other, again allowing direct method-method comparisons. Radtrim is the method giving the lowest |ΔFE| values, and does so in every 100-wide N_{keep} band. The second best method is autofilt for lower values of N_{keep}, and toltrim for higher values of N_{keep}.

Again radtrim seems to be the best choice for the AF algorithm, both in terms of giving results close to the unfiltered result, and the lowest variation.

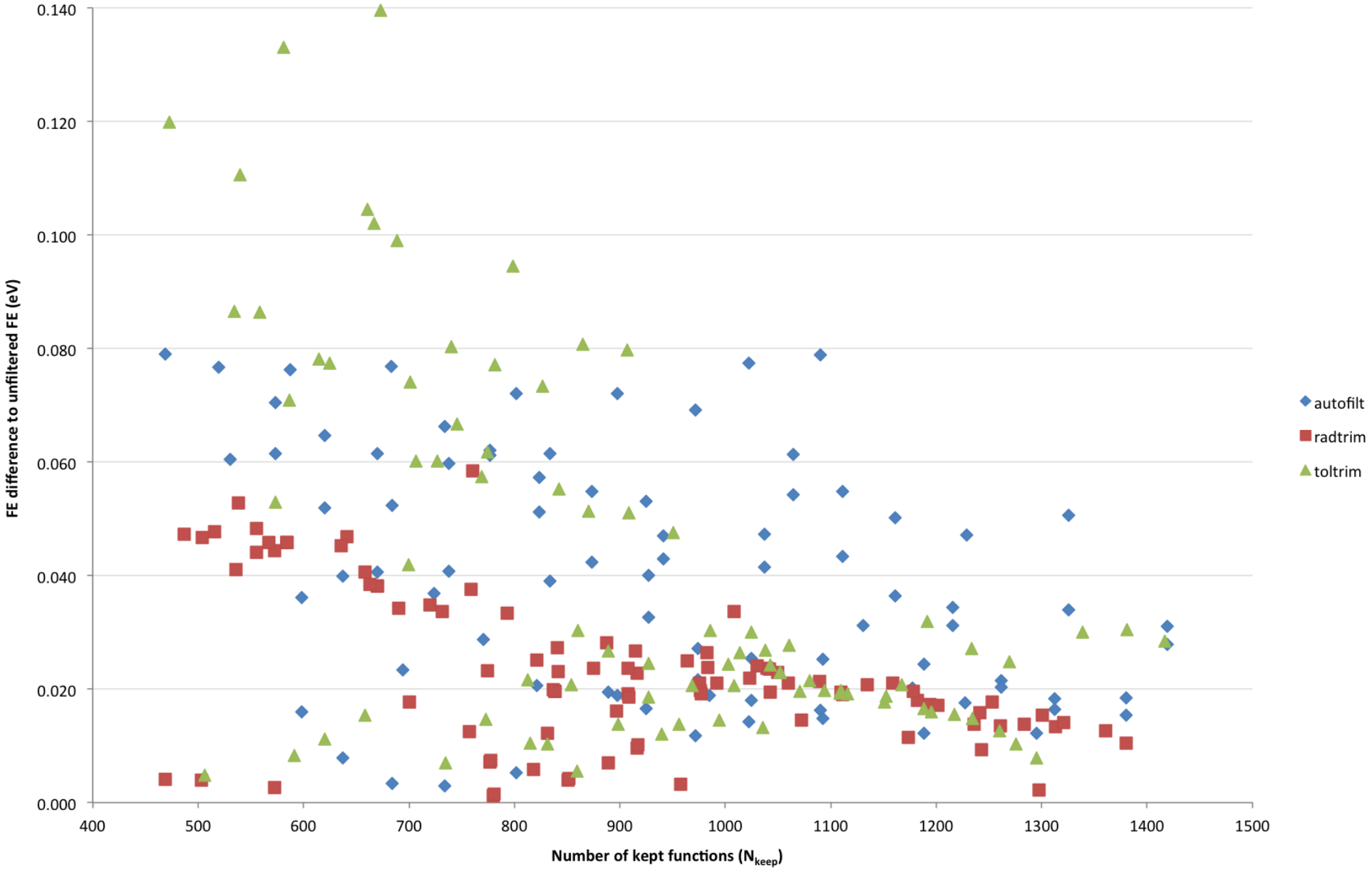


Figure 17: Differences in AF FEs to the NF FE of a vacancy in a unit cell of 64 silicon atoms.

6.7 AF Method Comparison - Conclusions

Taking all four structures into account, it would seem the radtrim method offers not only the lowest $|\Delta FE|$ for most ends of the N_{keep} spectrum, but also the lowest variation in $|\Delta FE|$. The results clearly point to radtrim being the most effective trimming algorithm for both low, medium and high values of N_{keep}. For low values of N_{keep}, toltrim and autofilt perform badly compared to radtrim. For medium-high and greater values of N_{keep} autofilt continues to be poor, but toltrim is a reasonable, but still second, choice to radtrim.

6.8 Radtrim R_{sphere} Parameter Investigation

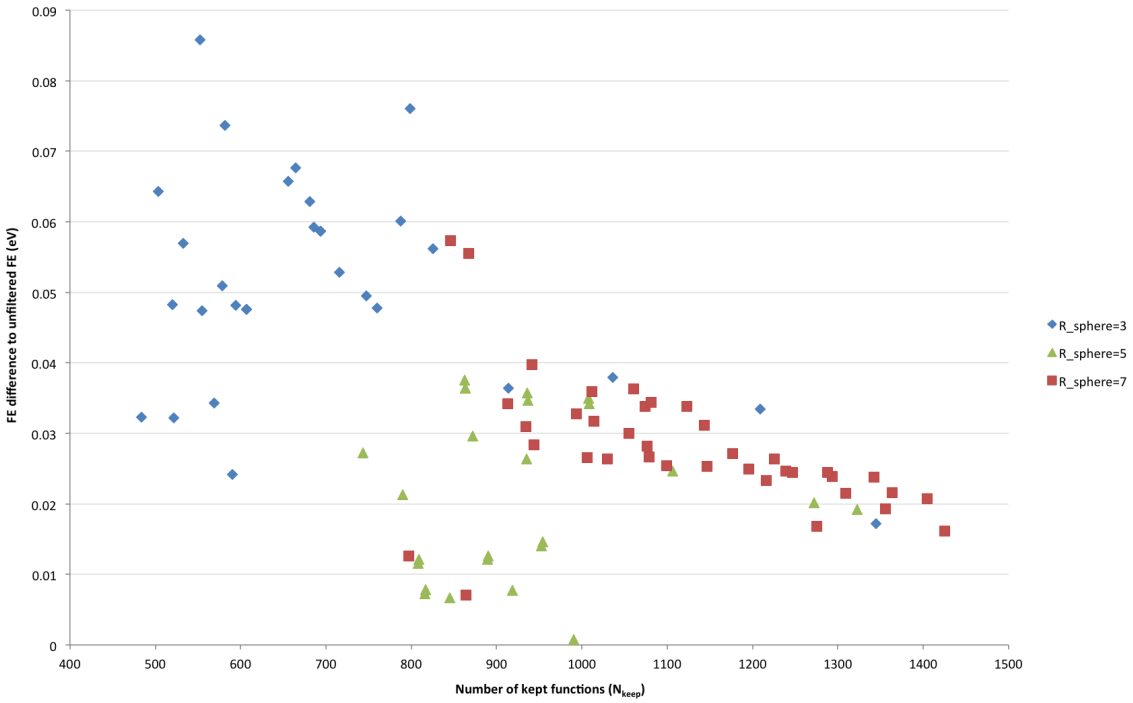
The previous results and findings were based on using varying parameters for each method. Radtrim proved to be the most effective AF method. The radius of the inner sphere is clearly an important parameter for this method, and from the data available it is possible to perform an analysis, to determine if a particular value performs better than others. The 3 different inner sphere radii used as part of the parameter set for radtrim were 3, 5 and 7 a.u. Scatter graphs of $|\Delta FE|$ against N_{keep} were produced for each of the four structures, with the three different values of R_{sphere} represented by three different colours.

Figures 18a to 19b show the radtrim results separated by the value of R_{sphere} - 3, 5 or 7 a.u.. The four charts are very similar, and will be discussed simultaneously. The first thing to notice is the narrow region where the results from the 3 values of R_{sphere} overlap. Due to the way radtrim works, the larger the sphere the more functions that are going to be included, regardless of the values of the s, p and d parameters. This means each value of R_{sphere} will occupy a distinct region of the chart. There is enough data in the overlap region to compare the effectiveness of the 3 values, for values of N_{keep} between 750 and 900. Results where R_{sphere} is 3 a.u. have the highest values of $|\Delta FE|$. The increase in functions presented to the filtration algorithm doesn't

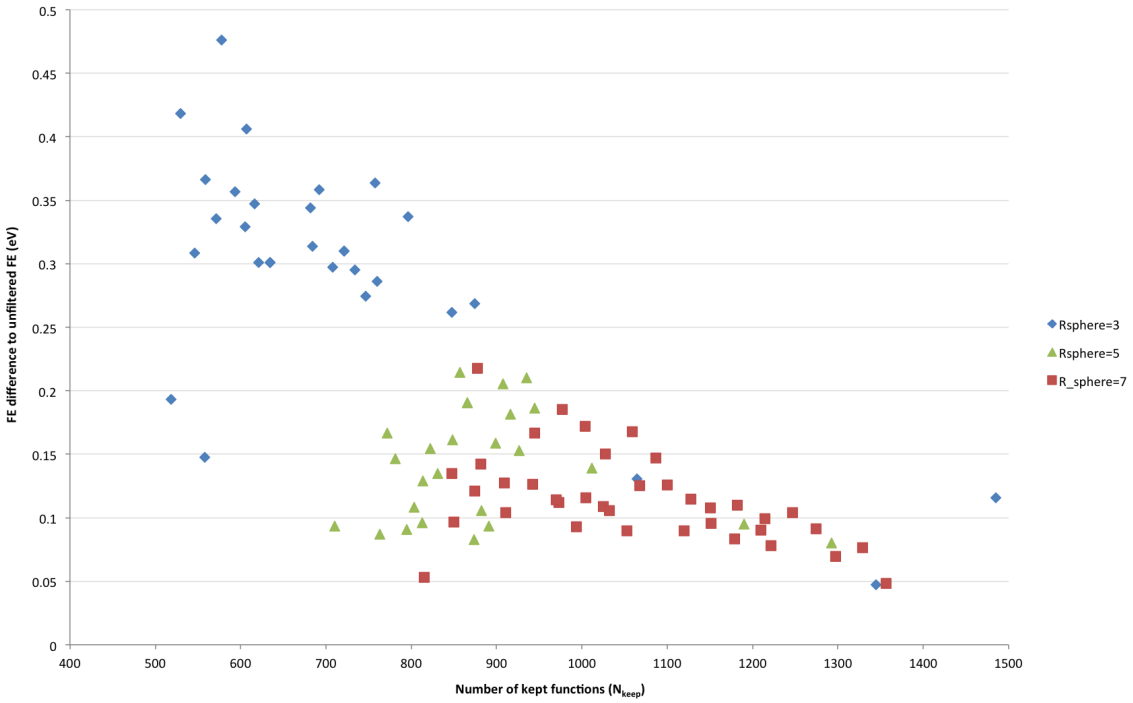
help move the results in the regions occupied by results where R_{sphere} is 5 or 7 a.u., they lie above them or at the upper limits. Looking at just values of 5 or 7 a.u. for R_{sphere}, 5 a.u. is clearly a much more efficient choice for this parameter. These results offer the lowest $|\Delta\text{FE}|$ values at their ranges of N_{keep}. In fact for all structures but amorphous, they offer the lowest $|\Delta\text{FE}|$ values for any calculations using much higher values of N_{keep}, over 1400 functions. Even for the amorphous structure FE, R_{sphere}=5 a.u. results perform as well as R_{sphere}=7 a.u. results where N_{keep}=1100+, despite only using between 700 and 950 functions.

R_{sphere}=5 a.u. performing so well, seems to indicate a good calculation requires two types of functions to be presented to the filtration algorithm. Firstly it is important to include a certain number of functions close to the central atom. For bulk silicon a value of R_{cut} of 5-7 a.u. includes the first shell of nearest neighbours, the closest 4 tetrahedrally arranged atoms. A value of 3 a.u. doesn't include this shell. Although we are not dealing with bulk silicon here, it is a good approximation for where the majority of the atoms will lie. This value leads to poor $|\Delta\text{FE}|$ results, even when other parameters are increased to compensate for the reduced number of kept functions. Secondly it is important to use the lowest value of R_{sphere} that includes this shell, so any functions on the next shell of atoms are only included if they are long ranged enough to have some penetration towards the central atom. A value of R_{sphere}=7 a.u. puts the edge of the inner sphere right next to the second shell of atoms, meaning they will always be included unless extremely low parameters are used for τ_{s}, τ_{p} and τ_{d}.

In summary, for AF FE calculations using the radtrim method, for the systems investigated here, a choice of the internal sphere of 5 a.u. produces results closer to the corresponding NF results than 3 a.u., or results using 7 a.u. where the total number of kept functions is less than 1400. Using the parameters R_{sphere}=5, and values of τ_{s}, τ_{p} and τ_{d} that produce values of N_{keep} of about 700-950 functions when a reasonably good calculation is required of greater speed. To achieve better accuracy, over 1400

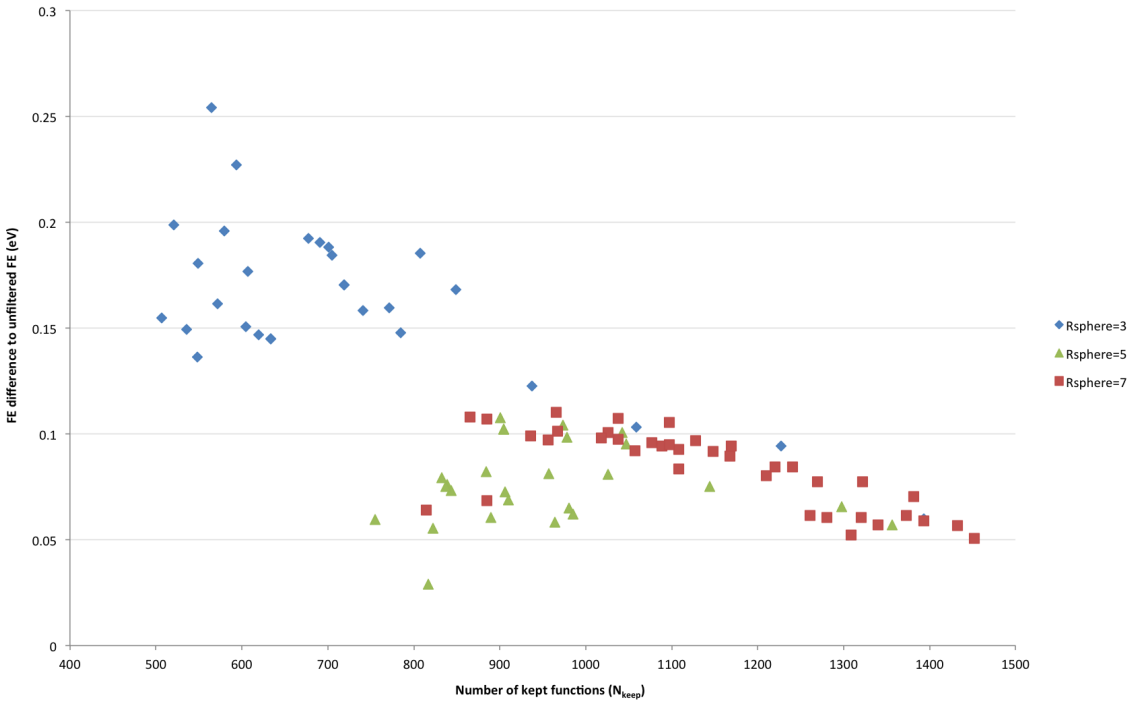


(a) Radtrim 110 results split by R_{sphere} parameter

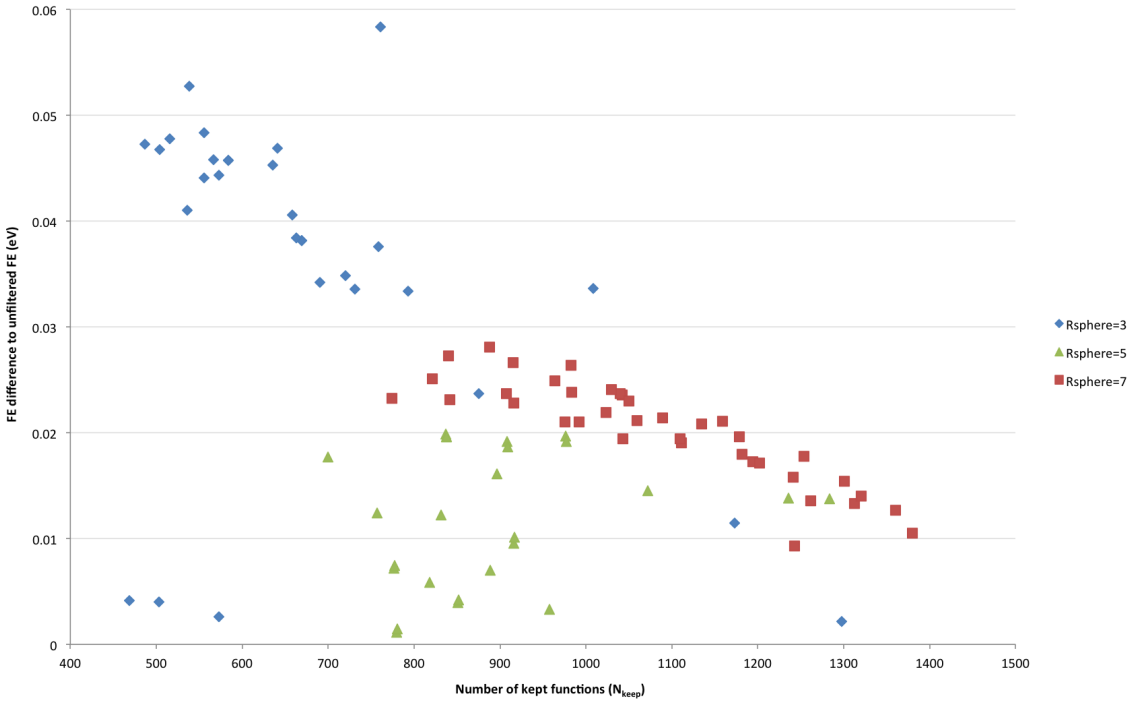


(b) Radtrim amorphous results split by R_{sphere} parameter

Figure 18: Radtrim results - effect of R_{sphere} parameter



(a) Radtrim I_3 results split by R_{sphere} parameter



(b) Radtrim vacancy results split by R_{sphere} parameter

Figure 19: Radtrim results - effect of R_{sphere} parameter

functions are required before any improvement is seen.

6.9 Conclusions

When calculating FEs for silicon systems, the radtrim method has been shown to be the most efficient AF algorithm of the three investigated. It performs better than the currently implemented method, toltrim, and should clearly replace it in future releases of the code. With the findings from chapter 5, structural optimisations for silicon systems performed using radtrim, setting the parameter R_{sphere} to 5 a.u., could use low values of the AF parameter τ significantly speeding up the filtration process. This would enable the efficiency advantages of the filtration method to be applicable for all but the smallest of systems. It also means the results in chapters 4 and 5 can be improved, whilst simultaneously maintaining or reducing the filtration time.

Chapter 7

TRANSITION STATE IDENTIFICATION - THE LANCZOS METHOD

This chapter takes a departure from filtration. The current method of choice for the identification of the structure of a transition state is the Nudged Elastic Band (NEB) method. It is a powerful tool, but has some drawbacks, primarily the number of force calls required to complete a calculation. A new method is implemented that could reduce this, and compared to the NEB method.

As a reaction or structural change proceeds, the energy of the system will typically rise and fall between two states that are minima on the PES. Along this path, the state with the highest energy is referred to as a transition state. A transition state is always a saddle point on a PES, almost always (but not necessarily) a first-order saddle point. The identification of transition states is an important aspect of computational modelling. By determining the difference between the energy of the initial state of the system, and the energy of the system in its transition state configuration, the activation energy of the reaction has been determined. The activation energy is an important measure of how quickly a reaction will proceed, and determines the reaction rate's dependence on temperature.

The current method of choice in AIMPRO for calculating a transition state is the nudged elastic band method (NEB) [16]. It suffers from some drawbacks, the primary one being that the whole reaction path must be modelled, using a series of images to represent the system at different points on the reaction path. All these images must be optimised simultaneously, leading to an optimisation of high dimensionality, which typically means a long calculation, with a significant number of force calls. This reaction path is referred to as the minimum energy path (MEP).

Other methods are available, and one class involves the identification of an 'uphill direction'. This is the direction on a PES in which movement uphill (i.e. in a direction of increasing energy) is necessary to arrive at a transition state. This approach is not without its difficulties, but has the potential to greatly reduce the number of force calls required to identify the transition state. The Dimer method [31] is an example of an uphill direction identification algorithm. The method used to identify the uphill direction in this thesis, is an implementation of the Lanczos algorithm. The Lanczos algorithm [43] forms the basis of transition state location methods such as the Activation Relaxation Technique (ART) [5, 43, 46].

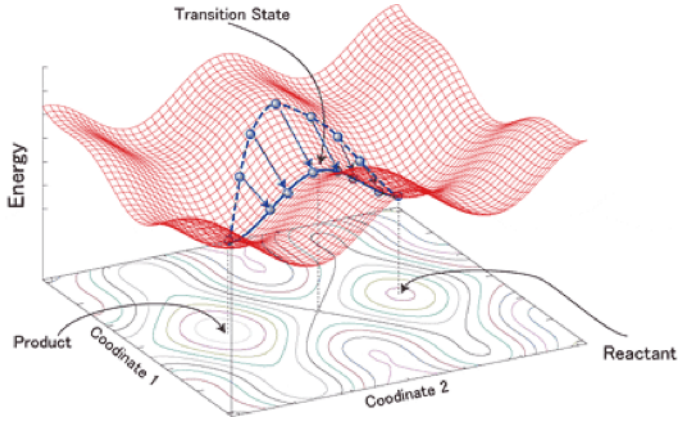
Once the uphill direction has been identified, the component of the force is 'flipped' in this direction, and effectively a minimisation is performed on the altered PES. This can be achieved using the methods employed for structural optimisation outlined in chapter 5. The uphill direction will of course change as the minimisation moves the structure on the PES, so periodically the uphill direction is recalculated.

Before the Lanczos method and its implementation in AIMPRO is explained, the current NEB method is explained, as well as the difficulties encountered in using the method, which prompted this work.

7.1 The Nudged Elastic Band Method

The NEB method requires the knowledge of the structures of the products and reactants. The transition state can be guessed by simply interpolating the positions of the atoms between the initial and final structures. This of course will not work if the reaction is a structural change where the product and reactant are the same. Even if this is not the case, this method often produces structures too far away from the actual transition state. Some knowledge of the the likely structure or properties of the transition state lead to much more efficient and successful NEB calculations. These three structures, or images, are points on the PES. More images are created between these three images using linear interpolation. These images are then connected by

Figure 20: Illustration of the NEB method. The blue spheres represent images on a PES, with the arrows showing what happens when the energy of each image is minimised. The sequence of images provides an approximation to the minimum energy path. As more images are used, the accuracy of both this path and the energy of the transition state become more accurate. Usually between 9 and 21 images are used.



virtual elastic bands, so movement away from the starting position causes a restoring force on the PES. This ensures the images are equally spaced on the PES, and do not slip into the minima on the PES representing the structures of the products and reactants. This is represented in figure 20 below.

Getting the maximum component of force below 10^{-2} Ha/a.u. is much harder than reducing it to this value. Small changes in one image can be transferred via the forces in the elastic band to other images. Also the minimisation routines used often are based on assuming quadratic behaviour of the energy with respect to changes in position. This assumption works well for structural optimisations. In NEB calculations, only the forces from the elastic bands parallel to the tangent of the MEP are kept. Also only the forces from the gradient of the PES perpendicular to the tangent of the MEP are retained [32]. This projection can disrupt the effectiveness of quadratic-based minimisers [32].

Typical behaviour is shown in the NEB calculations towards the end of this chapter, in figures 23 and 24.

As each image needs to be minimised at each iteration, large number of force calls are required. This combined with the oscillatory behaviour make an alternative to the NEB method an attractive prospect.

7.2 The Lanczos Method

The Lanczos method requires only a single structure as a starting point. The starting structure will preferably be located near a saddle point, but it can be located near or even at a minimum on the PES. The procedure requires an initial direction, which can be specified, or more commonly a random direction is generated. An iterative procedure then refines this uphill direction, which is an eigenvector of the Hessian matrix, calculating the corresponding eigenvalue. This is the lowest eigenvalue. The procedure starts with a structure represented by a position $\vec{x}_0$ on the PES, a random direction vector $\vec{r}_0$, $\beta = ||\vec{r}_0||$, and $\vec{q}_0 = 0$. The following iterative procedure (with k as the iteration counter) then takes place [47]:

$$\vec{q}_k = \frac{\vec{r}_{k-1}}{\beta_{k-1}} \tag{84}$$

$$\vec{u}_k = \frac{\nabla V(\vec{x}_k + 10^{-3}\vec{q}_k) - \nabla V(\vec{x}_k)}{10^{-3}} \tag{85}$$

where $\nabla V(\vec{x}_k)$ is the gradient of the energy at the point $\vec{x}_k$.

$$\vec{r}_k = \vec{u}_k - \beta_{k-1}\vec{q}_{k-1} \tag{86}$$

$$\alpha_k = \vec{q}_k^t \vec{r}_k \tag{87}$$

$$\vec{r}_k = \vec{r}_k - \alpha_k \vec{q}_k \tag{88}$$

$$\beta_k = ||\vec{r}_k|| \tag{89}$$

α_k and β_k are used to form a tridiagonal matrix T:

$$T = \begin{pmatrix} \alpha_1 & \beta_1 & 0 & 0 & 0 \\ \beta_1 & \alpha_2 & \beta_2 & 0 & 0 \\ 0 & \beta_2 & \alpha_3 & ... & 0 \\ 0 & 0 & ... & ... & \beta_{i-1} \\ 0 & 0 & 0 & \beta_{i-1} & \alpha_i \end{pmatrix} \tag{90}$$

The size of this matrix increases after each iteration, starting out as a 1x1 matrix. The lowest eigenvalue λ^{T_k} of this matrix is calculated using standard Fortran routines. When equation (91) is satisfied, the uphill direction $\vec{u}_k$ is considered to be the true one.

$$|\frac{\lambda^{T_k} - \lambda^{T_{k-1}}}{\lambda^{T_{k-1}}}| < 10^{-3} \tag{91}$$

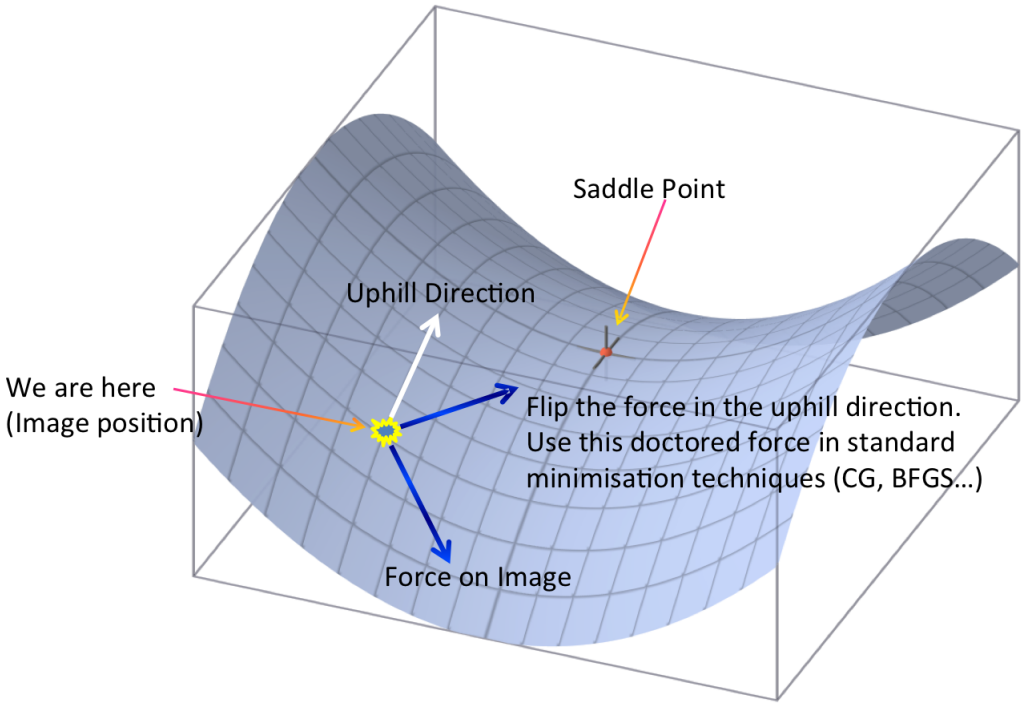
With the uphill direction identified, the force on the structure at this point on the PES is 'flipped' in this direction, effectively converting the problem into one of minimisation. This is illustrated in figure 21.

With this uphill direction, a pre-defined number of minimisation iterations (typically 6) take place before the uphill direction is recalculated. This procedure of calculating the uphill direction, then performing 6 minimisation steps, and repeating, is continued until, as in structural optimisations, the maximum component of the force falls below a pre-defined tolerance.

7.3 Choice of Minimisation Algorithm

When coding and testing the Lanczos method, a set of small organic molecules, known as the Baker set [3], close to a transition state for various reactions and structural rearrangements, were chosen as the test structures. This was because they offered steep and shallow areas on their respective PESs. The hydrogen atoms in particular could move around without significantly affecting the energy and forces on the structure. The internal parameters required for the method to work were optimised so that the

Figure 21: Illustration of how the uphill direction is used to convert the force vector to point towards the saddle point/transition state in the Lanczos method.



Lanczos method would perform well, even for these difficult structures. While this testing was taking place, it was realised the DIIS minimisation technique, although offering excellent efficiencies (DIIS requires only one force call per iteration, as it does not use a line minimiser), was not sufficiently stable. The conjugate gradient algorithm was used, and proved to be much more stable. However, this required the development of a line minimiser which used only forces, if this was ever to be used in conjunction with the filtration algorithm. The development of this is detailed in the next section.

7.4 Force-only Based Line Minimiser

The function of the line minimiser is to move along a direction $\vec{d}$ on a PES until the force vector $\vec{f}$ is orthogonal to $\vec{d}$. This exit criteria will be said to be achieved when

$$|\vec{f}.\vec{d}| < \tau_{LM}. \tag{92}$$

in which $\vec{d}$ is a unit vector, and τ_{LM} is set to be a tenth of the initial value of $\vec{f}.\vec{d}$ when the line minimiser was called.

7.4.1 Outline of Main Steps

The basic method of the line minimiser will be as follows:

1. The line minimiser will start at the initial point, $\vec{R}_1$, and move a trial amount to the trial point $\vec{R}_2$, in the direction $\vec{d}_1$ provided by the conjugate gradient algorithm. The initial direction is given by the direction on the PES equivalent to the initial force $\vec{f}_1$ on the structure. How the trial point is determined is outlined in section 7.4.2.

2. The force vector is calculated at this trial point, giving $\vec{f}_2$. There is a possibility that by chance this point will satisfy the exit criteria for the line minimiser, so $\vec{f}_2$ is checked against this (92).

3. If this check is unsuccessful, using the information from the initial point and the trial point, a third calculated point $\vec{R}_3$ is calculated, where the force $\vec{f}_3$ is hoped to satisfy the exit criteria. How this third point is chosen is detailed in section 7.4.3.

4. $\vec{f}_3$ is calculated and checked against the exit criteria.

5. If the check is unsuccessful, two of the points $\vec{R}_1$, $\vec{R}_2$ and $\vec{R}_3$ are kept as the initial point and the trial point, and the algorithm returns to step 3. The process of determining which point becomes which is detailed in section 7.4.4.

7.4.2 Choice of Trial Point

In the following sections, the trial point will be referred to as point 2, the initial point point 1, and point 3 is the calculated point using the information from points 1 and 2. $\vec{f}_1^{\max}$ refers to the component of $\vec{f}_1$ with the maximum absolute value. The trial point is always along $\vec{d}$, so $\vec{R}_2$ can be specified by α_{trial}, where:

$$\vec{R}_2 = \vec{R}_1 + \alpha_{\text{trial}}\vec{d} \tag{93}$$

and the calculated point $\vec{R}_3$ can be similarly specified by α_{calc}, where:

$$\vec{R}_3 = \vec{R}_1 + \alpha_{\text{calc}}\vec{d} \tag{94}$$

For the first trial point of a structural optimisation calculation, α is set to be a multiple of the lower of unity or $\vec{f}_1^{\max}$. Every subsequent line minimisation, say the $(i+1)_{\text{th}}$, sets the trial step to be of a size such that:

$$\alpha_{\text{trial}(i+1)}\vec{f}_{i+1}^{\max} = 2\alpha_{\text{trial}calc(i)}\vec{f}_i^{\max} \tag{95}$$

This ensures the maximum distance any component of any atom moves from $\vec{R}_1$ to $\vec{R}_2$ at the start of a line minimisation, is exactly double the maximum distance moved in the previous successful line minimisation. This is useful, as this will usually lead to the point where $\vec{f}.\vec{d}$=0 (from now on referred to as the FD0 point) lying between the initial and trial point. This is advantageous as it it will lead to better estimates of the FD0 point, and will mean future steps are interpolating rather than extrapolating.

7.4.3 Estimating the Location of the FD0 Point

They may be more than one FD0 point. The FD0 we require is the closest one in the direction of the force vector. The direction of the force vector at a point on the PES will be in a direction that decreases the total energy of the system:

$$\vec{f} = -\frac{\partial E}{\partial \vec{x}} \tag{96}$$

Close to a minimum (or saddle point) the potential, and hence energy and the PES, varies quadratically with $\vec{x}$. This means the force varies linearly with respect to $\vec{x}$, and therefore with α. As we have two points on the PES, if we assume quadratic behaviour, we can estimate the FD0 point by linearly interpolating between the two points on a $\vec{f}.\vec{d}$ vs. α chart. Specifically, we calculate α_{calc} to place into 94:

$$\alpha_{\text{calc}} = \alpha_{\text{trial}} \left(\frac{\vec{f}.\vec{d}_1}{\vec{f}.\vec{d}_1 - \vec{f}.\vec{d}_2} \right) \tag{97}$$

The closer to the minimum or saddle point, the better the assumption that the PES is quadratic will be, and the more accurate this estimate will be.

There are two situations where this falls down quite badly. Firstly, at the start of a structural optimisation the structure could lie on a point on the PES far away from the quadratic region. Typically the CG algorithm will still result in the energy reducing, just not as efficiently as it theoretically could. For a perfectly quadratic PES, the CG will find the exact minimum in n steps, which means $n + 1$ force calls, where n is the dimensionality of the system.

The second situation causes more difficulties. A 'shoulder' is a region on the PES that has increasing $\vec{f}.\vec{d}$ in the search direction. It occurs when the energy is decreasing faster and faster as the minimum is approached. This is checked for in the algorithm and the system moved through this 'shoulder' region until $\vec{f}.\vec{d}$ starts decreasing in the search direction. Once it does, the linear interpolation strategy is resumed. To reduce the number of force calls, to move through the region initially, a trial point three times further away from the start point, as was originally chosen, is selected. Each failed attempt after this sees this factor doubling, to 6, then 12, then 24 etc. Further checks are in place to prevent the structure changing too radically within this process, and are detailed later in section 7.4.5.

7.4.4 Recursive Algorithm Details

The line minimiser will continue to choose points, and test the value of $\vec{f}.\vec{d}$ until it is less than a tenth of the initial value (at the very start of the line minimiser, not the

initial value of each linear interpolation), or less than the exit tolerance τ_{LM}. When an estimate of the FD0 point does not meet either of these criteria, another linear interpolation is performed to provide a new estimate of the FD0 point. Figure 22 shows exactly how this is done, based upon the values and positions of the previous initial, trial, and FD0 points.

7.4.5 Further Refinements and Checks

The PES of a molecule is never perfectly quadratic. Structures can be specified whose starting positions can give rise to large forces, such as when they are placed too close together. Certain molecules and structures can have PES that vary rapidly in some directions, and very slowly in others. These and other unforseen problems require checks to be built into this process, to ensure an atom is not moved too far. In small molecules, where atoms are being transferred, such as those in the Baker set, atoms on the edge of the system can 'escape' from the main system. An easy way to minimise the forces on an atom is to move it to infinity. By limiting the distance an atom can move, situations like this can be prevented. A list of such measures is now presented.

1. Upon entry to the line minimiser, it is ensured $\vec{f}.\vec{d} > 0$. This ensures $\vec{d}$ points in a direction that lowers the energy of the structure. If it is not the CG algorithm is initialised.

2. The trial step is always checked to be greater than a minimum value, currently set at 1.0×10^{-7} a.u. This is low enough to accomodate the small steps required for very small/accurate exit force tolerances near the end of a structural optimisation. It is also high enough to ensure numerical noise is orders of magnitude smaller than the changes the step produces in the force vector. If it is not it is set to be the minimum value.

3. The trial point at the start of the line minimiser is set such that the maximum shift of any component of any atom will be equal to half of the maximum shift

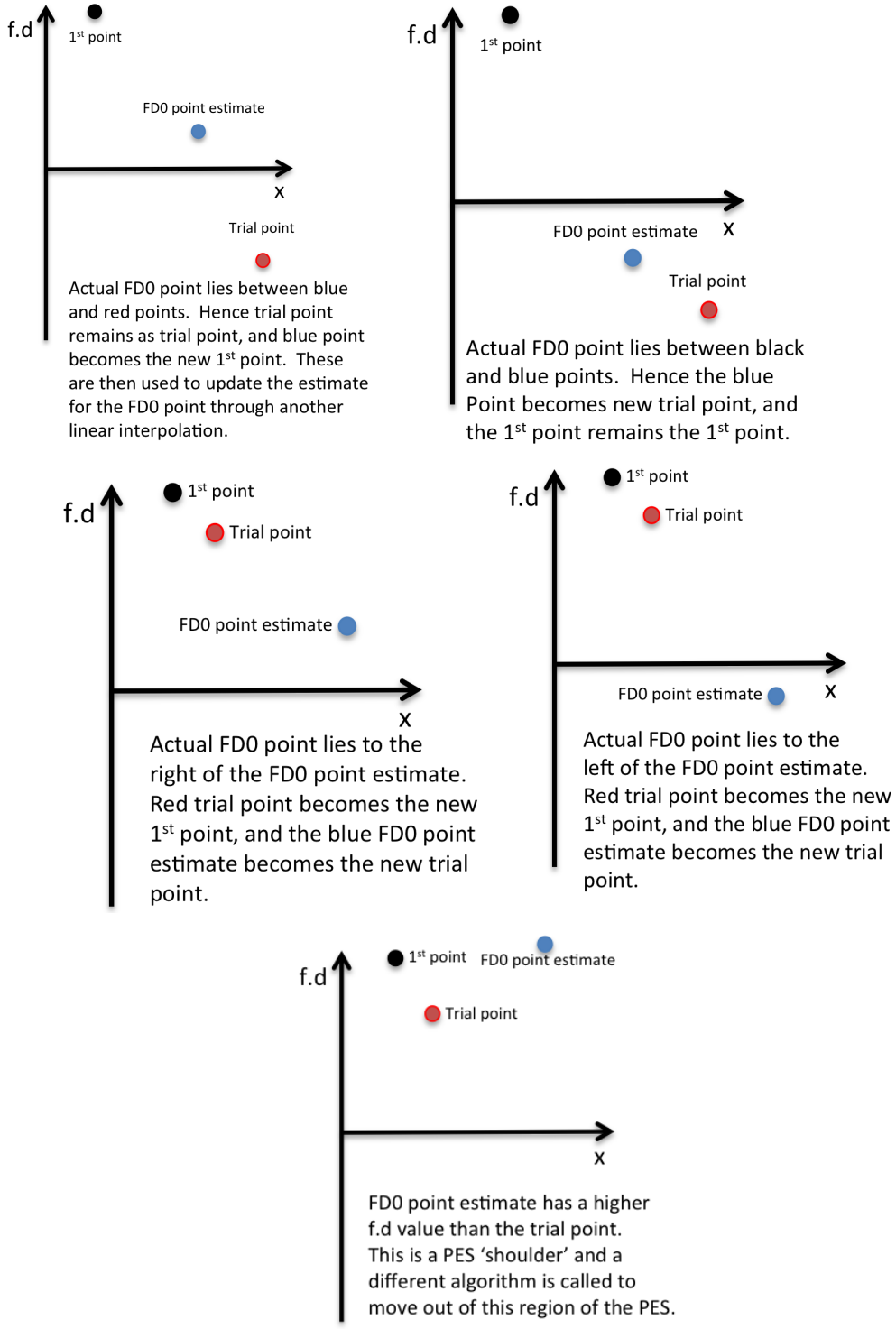


Figure 22: An illustration of how the recursive part of the line minimiser algorithm operates. The blue FD0 point estimate should lie on the x-axis. The actual value of $\vec{f}.\vec{d}$ may be above or below the tolerance band of width $2\tau_{\mathrm{LM}}$ around the x-axis, and another linear interpolation is performed. Five different scenarios are possible.

of any component of any atom observed in the previous line minimisation. After this is calculated, it is checked it is not beyond a maximum value, currently set to be 0.15 a.u., and if it is it is the step is shortened accordingly.

4. If the zero point does not lie between the initial and trial point, extrapolation is required to estimate the zero point, as opposed to interpolation if it does. Interpolation is much safer in terms of stability, as it does not allow large changes to happen to any part of the structure. Extra checks are thus performed for line minimisations that require extrapolation. The first of these is that the difference between $\vec{f}_1.\vec{d}$ and $\vec{f}_2.\vec{d}$ must be greater than a minimum value, currently set to 1.0×10^{-5}Ha/ a.u.. If this check is failed, a second trial point that satisfies this criteria is chosen.

5. The second of these type of checks is that $(\vec{f}_1.\vec{d})/(\vec{f}_2.\vec{d})$ is less than 0.99. This ensures not only an absolute minimum difference in the values of $\vec{f}.\vec{d}$ but a minimum relative difference as well. Both of these checks reduce the possibility of large erroneous jumps on the PES if the PES is highly non-quadratic. If this check is failed, a second trial point that satisfies this criteria is chosen.

6. At the start of the line minimisation the position of every atom is recorded. During the line minimisation the maximum change in any component of the position of any atom can change by a maximum of 0.3 a.u.. When this check fails, the line minimser exits with the most recent calculated structure.

One final task performed by the line minimiser is to let the CG algorithm know when a problem with the PES has been encountered, or the structure has moved too much. This means the new search direction should not be influenced by the previous ones. As the CG algorithm produces directions that are conjugate to all the previous ones, the new direction is influenced by the previous ones. Hence in these circumstances, the algorithm is reset, and the new search direction is given by the

force on the structure at that point on the PES, as was done in the first structural optimisation iteration.

7.4.6 Output from the Line Minimiser

A direct comparison of the previous method of line minimisation, which used both energies and forces, and this force based only line minimiser, is not a fair test as the previous method has access to more information, and would be expected to perform more efficiently. However it should be ensured that the reduction in efficiency is small. By running unfiltered structural optimisations using both methods, this can be done. The specific measures will be the number of force calls required for each method, the maximum component of force on the final structures, and the values of $\vec{f}.\vec{d}$ from the estimated FD0 points.

As a first test, a unit cell of 64 silicon atoms in the bulk configuration was taken as a starting point, then 3 atoms were moved 0.3% of the interatomic distance in the direction of each of the three lattice vectors. All other parameters were the same as in the silicon vacancy calculations in section 4.4, except the tolerance for the SCF process, which was set to three orders of magnitude lower. This was to ensure highly accurate forces were presented to both line minimisers, as forces smaller than 1.0×10^{-3}Ha/ a.u. require a more accurate SCF tolerance. If this isn't set, an element of randomness is introduced into the process, which would blur the distinction between the methods. Two calculations were then performed, both using the CG algorithm, firstly with the existing force and energy based line minimiser, and then with the new force only based one. The results are shown in table 40. The energies of the structures are identical to 7 decimal places, and the resulting structures were both of bulk silicon. The maximum force component is slightly higher for the force based only minimiser, but both are much lower than the initial value of 0.0035011 Ha/a.u..

The output from the force based line minimiser shows how much $\vec{f}.\vec{d}$ is reduced. In the output below, the first number indicates 1 as the initial point, 2 as the trial point,

Table 40: The maximum component of force on a bulk silicon system with 3 atoms moved 0.3% of the silicon-silicon interatomic distance, after a line minimisation step using two different line minimisers, one force and energy based, and the other a filtration-compliant method using just forces. Both methods used three force calls for the process. It can be seen the minimisers perform as well as each other. The initial maximum force component on the structure was 0.0035011 Ha/a.u..

Line Minimiser Type	Maximum Force Component (Ha/a.u.)	Energy of Structure (Ha)
Energy & Force	0.00018902	-253.44673
Force only	0.00019775	-253.44673

and 3 as the calculated point. The second number is α, i.e. how far in multiples of $\vec{d}$ has been moved, the third number the energy in Hartrees, and the last number is the value of $\vec{f}.\vec{d}$. The line minimiser requires a final value of $\vec{f}.\vec{d}$ 10 times less or more than the initial value. Here it has reduced the value by a factor of over 14,000, using the minimum number of force calls.

LINMIN: 1 0.0000000000 -253.4467068660 0.0035011240
LINMIN: 2 0.0038754819 -253.4467289004 0.0015340682
LINMIN: 3 0.0068978941 -253.4467341354 0.0000002420

As a second test, a unit cell of 64 silicon atoms in the bulk configuration were randomly shifted from their equilibrium positions, to a maximum of a tenth of a silicon-silicon bond length. Both line minimisers were again used in conjunction with the CG algorithm., The results of the structural optimisation are in table 41. In this instance, the force only based minimiser produces a lower maximum force component in fewer force calls than the its energy and force based counterpart. The differences are small enough to be negligible however - the conclusion is that the use of a force only based

Table 41: The maximum component of force on a randomised bulk silicon system after structural optimisation using CG, but with two different line minimisers, one force and energy based, and the other a filtration-compliant method using just forces. Each atom was moved up to a maximum of 10% of the silicon-silicon interatomic distance. The number of force calls required to reduce the force to this level is also provided. The initial maximum force component on the structure was 0.0601515 Ha/a.u.. Both line minimisers work as well as each other, with the force only based one producing slightly lower forces for slightly fewer force calls.

Line Minimiser Type	Maximum Force Component (Ha/a.u.)	Energy of Structure (Ha)	Number of Force Calls
Energy & Force	0.00022969	-253.44673	25
Force only	0.00019109	-253.44672	23

minimiser, when properly implemented, does not need to affect the efficiency of structural optimisation calculations.

The force only based line minimiser has been shown to be as effective as the force and energy based one in the two cases above. It is used not only in the Lanczos method, but also is an option for structural optimisation and NEB calculations. In both these cases it performs well for all structures it has encountered.

7.5 Results - Transition State Identification Using the Lanczos Method

To test the Lanzcos method against the NEB, outputs from previously completed NEB calculations were used. The structure of the highest energy image at, or near, the start of the NEB calculation was used as the starting structure for the Lanczos method. The resulting energies of the structures obtained from the two methods were compared, and used as a check that the same structures were obtained. A further

check on all the resulting structures from the Lanczos method was carried out to ensure the resulting structures were saddle points, by calculating the actual Hessian matrix, and ensuring it had one and only one negative eigenvalue. This check ensures a first order saddle point.

All of the NEB calculations were of diffusions of defects and self-diffusion in diamond [11, 35]. They were as follows:

- NCN - One saddle point in the process of diffusion of an A centre in diamond (two adjacent substitutional nitrogen atoms).
- V - Diffusion of a neutral vacancy in diamond.
- P1 - Saddle point associated with diffusion of single substitutional nitrogen atom in diamond.
- R2 - [100] split interstitial in diamond.
- N2VH - Reorientation in a vacancy decorated by two nitrogen atoms and a hydrogen atom.
- CE - Self diffusion, concerted exchange in diamond.

Table 42 displays the results for the identification of the transition states relating to these six diffusion processes. It can be seen that the differences in energy are very small, less than 1 meV in all cases. This would indicate the two methods have identified the same transition state. The final maximum force component for the Lanczos calculations are lower than seen for the NEB calculations, except in the case of N2VH, where both are extremely low. In every case, the Lanczos has arrived at the transition state found by the NEB, but using significantly fewer force calls. The difference in the efficiency of the two methods is displayed more convincingly in graphical format. This is provided in figure 23 for NCN, and figure 24 for CE. It is clear how effective it is to use the NEB to provide an estimate of the structure of a transition state, and then using the Lanczos method to perfect it.

Table 42: Lanczos vs. NEB results for transition state location in diamond diffusions and self-diffusions.

NEB Model	ΔE meV	Force calls NEB	Force calls Lanczos	Init Force mHa/a.u.	Final Force NEB mHa/a.u.	Final Force Lanczos mHa/a.u.
NCN	-0.55	884	12	5.445	0.994	0.055
V	-0.12	63	16	0.187	0.340	0.090
P1	0.09	476	15	0.300	1.294	0.048
R2	-0.93	150	60	1.266	1.410	0.085
N2VH	0.00	165	44	15.838	0.004	0.009
CE	-0.01	1440	22	8.981	0.400	0.057

Figure 23: Comparison of Lanczos and NEB method for identification of NCN saddle point. Maximum force component is shown against the number of force calls. Note the logarithmic scale for the maximum force component axis.

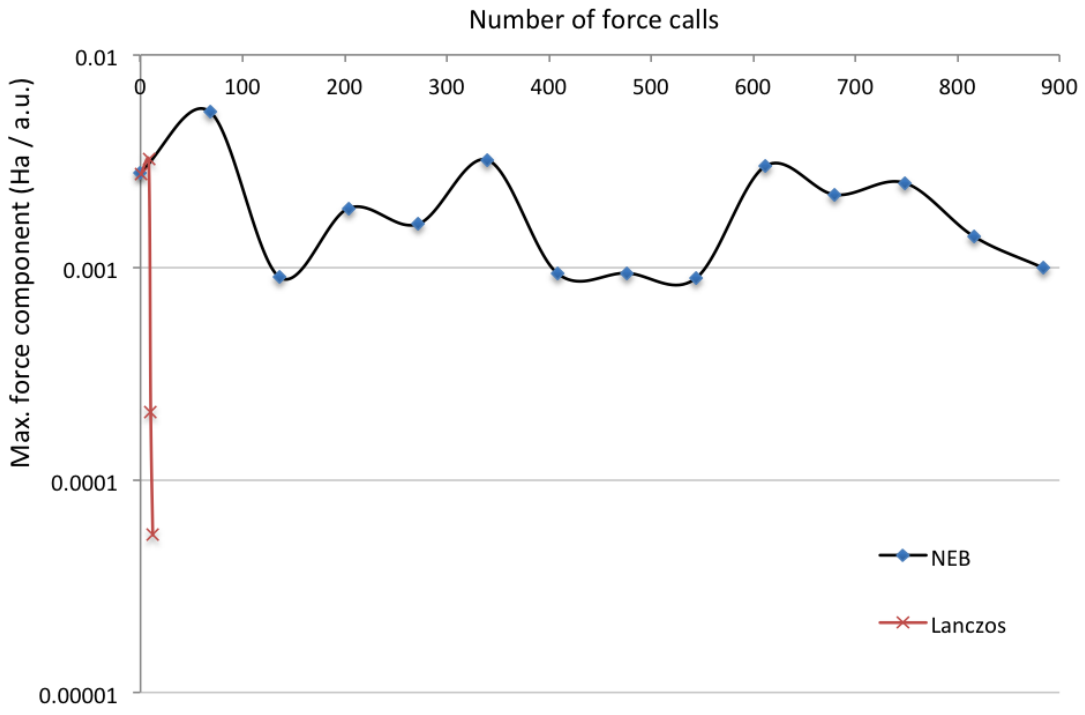
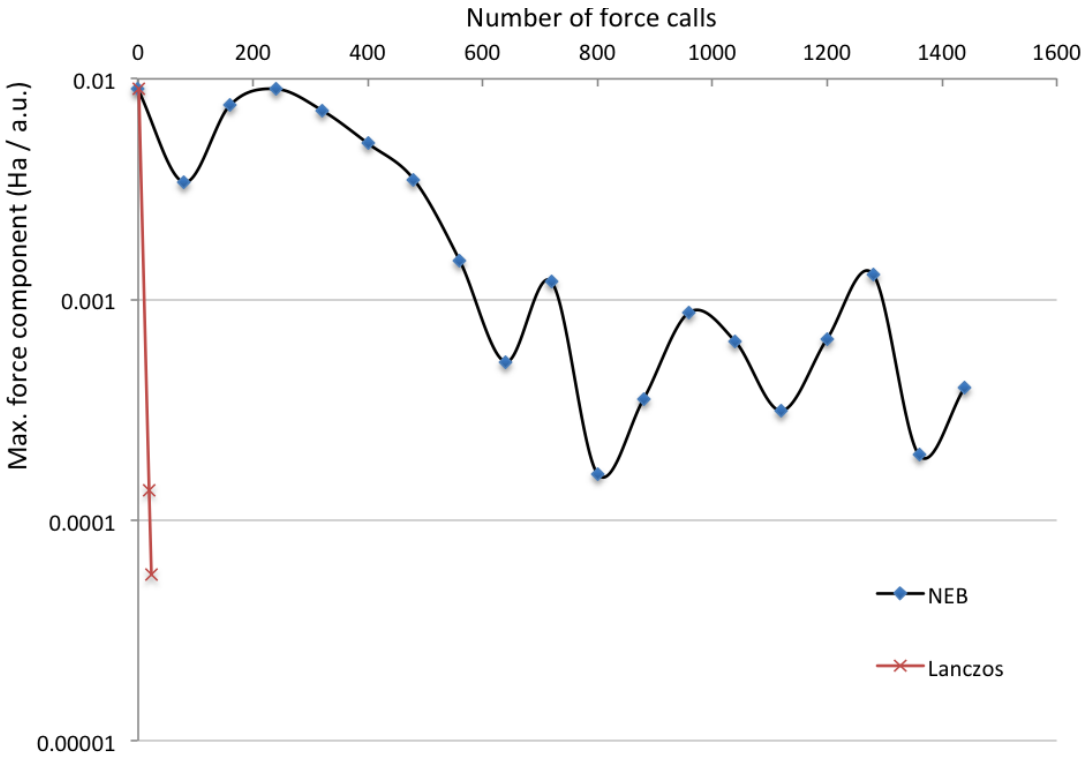


Figure 24: Comparison of Lanczos and NEB method for identification of CE in diamond saddle point. Maximum force component is shown against the number of force calls. Note the logarithmic scale for the maximum force component axis.



7.6 Advanced Features

The Lanczos method and the NEB work well together, so a further small, but highly useful, development was carried out. An option was added for NEB calculations, so that when the calculation was completed, the structure of the highest energy image, and the current uphill direction (identified as the direction of the elastic band at this image), were output in a file. The Lanczos method could then use this file, removing the need to create an initial random uphill direction. This has the dual effect of reducing the initial number of force calls required to calculate the initial uphill direction, and to ensure the starting uphill direction is the one related to the MEP of interest. The exit maximum component force tolerance for the NEB is set much higher, typically 10^{-1} to 10^{-2} Ha/a.u., so that the oscillatory behaviour of the NEB is not encountered. Results as to how effective this is are not presented, but would be of great interest. This is an ideal area for further work, and marks the end of this chapter, and thesis.

Chapter 8

CONCLUSIONS AND FURTHER WORK

The majority of this thesis has investigated the filtration algorithm, specifically its effect on the reduction of the time required for a SCF iteration, and the resulting accuracy of calculations. The standard filtration (SF) method was supplemented with the implementation of a technique referred to as advanced filtration (AF), where the number of primitive basis functions to be presented to the filtration process, is reduced, based on various criteria. SF and AF were applied to formation energy calculations in chapter 4, and structural optimisations in chapter 5.

8.1 Conclusions

Chapter 4 compared the effect of SF and AF on formation energy calculations for defects in silicon. For the formation of an ideal vacancy in silicon, an R_{cut} of 12 a.u. produced differences to the unfiltered calculation of less than 10 meV. AF quality was linked to a parameter τ. When AF was applied, it was a value of τ of 6 or above that was acceptable from an accuracy viewpoint. For the formation of a VO_2 defect in silicon from VO and O_i defects, a value of R_{cut} of 10 a.u. produced results within 10 meV of the unfiltered calculation, however a value of τ=10 was required to stay within this limit. To guarantee SF or AF results within 10 meV of the corresponding unfiltered result two options appear to be available. Either R_{cut}=12 a.u. with a low value of τ, 6 or 8, or R_{cut}=10 a.u. with a high value of τ, 10 or more. Both of these choices reduce the time required for an SCF iteration for systems of 216 atoms with high symmetry and Γ-point sampling of the Brillouin zone, or as low as 64 atom systems with a finer sampling grid or low degrees of symmetry. Any systems with more atoms, finer sampling grids, and/or lower symmetry will see SCF times reduced

significantly.

Chapter 5 started out applying either SF or AF to the entirety of a structural optimistion calculation. R_{cut} values of 10 a.u. produces changes in the formation energy calculations of the resulting stuctures that were too large, over 20 meV. When AF was applied, this effect was increased, and differences of up to 90 meV were witnessed. By simply changing the final total energy calculation to an unfiltered one, these differences were almost completely eliminated. This method ensures the efficiency savings of filtration are applied to the vast majority of a calculation. The overhead of one final unfiltered calculation is very small compared to the savings gained by using AF for all the preceeding calculations performed to determine the final structure. If a very large system was involved, the final energy could be performed using SF with a large enough value of R_{cut}, which would offer a speed-up factor close to the theoretical maximum. Forces appear to be more insensitive to filtration than energies.

The results of chapters 4 and 5 mean that optimisations of unit cells contaning 500 atoms or so can be completed at least an order of magnitude faster than before. This shows that filtration is not just a "specialist" technique, applicable to large unit cells of thousands of atoms, but can produce remarkable time savings for more standard calculations currently being peformed on cells of 200-700 atoms. Also, in practice the savings could often be many times this, as the specimen calculations done here frequently used Γ point sampling, whereas finer grids would often be employed on systems in practice.

Chapter 6 looked at three new developments to the filtration technique. These focus on determining the most efficient way of imposing a spatial cutoff to the filtered functions by selecting the functions which are to be retained during the filtration step. Previously, only one method to do this had been implemented, and development or testing of alternative strategies had not been considered. The new techniques were applied to formation energy calculations for four systems involving defects in silicon. A method named Radtrim, involving an inner sphere in which all functions centred

inside this are kept, and a selection method for functions outside of this sphere, proved to be the optimal method of choice. For a fixed number of kept functions, it produced formation energies closest to the unfiltered result. It also displayed the smallest variation in resulting formation energies when parameters were varied. During a structural optimisation, this small variation of energies and forces will clearly be advantageous in reproducing the PES of the unfiltered calculation.

The idea of an angular momentum based R_{cut} process failed to produce any improvement to the results. It had been supposed that smaller cutoffs with higher angular momenta would be acceptable, but this turned out to be incorrect. The conclusion was that the original strategy of including all functions sharing the same exponent was in fact optimal, although this is now known as a fact.

Chapter 7 detailed the implementation into AIMPRO of a technique for the identification of transition states, the Lanczos method. Previously, the standard method of choice was the Nudged Elastic Band (NEB) method, which suffered from two main problems. Firstly as it required multiple images to span the reaction path, the method is CPU intensive, requiring a large number of force calls. Also once the maximum force component reaches a certain level, the method struggles to reduce it further. It has difficulty optimising all the images at once. The Lanczos method showed it was possible to take the structure of the highest energy image from a NEB calculation, and reduce the force extremely quickly, using very few force calls. The use of the NEB to get a structure reasonably close to a transition state/saddle point, followed by the Lanczos method to quickly home in on the nearest saddle point, appears to be a highly efficient and useful combination. With a view to this, a tool to stop a NEB run, and then output a file to be used by the Lanczos method was developed. This file included both the structure and highest energy image, to ensure the Lanczos method headed towards the desired transition state.

8.2 Further Work

This work has established the radtrim method, as the most efficient method for imposing localisation on the filtered fuctions, and should be implemented as the standard choice for fltration calculations in future releases of the AIMPRO code. It would clearly be important to confirm that these conclusions also hold in dfferent materials, and some work to do this is already underway with encouraging results [REF PRB private comm]. It is also important to verify this conclusion when modelling derived quantities such as hyprfine couplings, infrared spectra, population analyses, dipole moments and so on. Partially completed work suggests this will be the case, but this needs to be definitvely established and published.

The work in this thesis has done a great deal of variation of parameters attached to filtration, but clearly this is not at all desirable in a project applying the technique to a problem in materials science. It is necessary to ensure that all parameters can be set automatically by the code as part of a run. Following an examination of the effect of the R_{sphere} parameter on calculations in other materials, the setting of this parameter could be optimised automatically, leaving τ as the remaining parameter pertaining to the localisation of filtered functions. This could remain as a type of 'quality' parameter to be set by the user, but it is still preferable for this to be done automatically. It would be an obvious further optimisation to gradually increase this parameter as part of the self consistent cycle — at present extremely accurate filtration is performed, even when a run is far from self consistency, and this imposes an unnecessary overhead. The value of τ could be increased from an approximate starting value until the energy is converged to a certain tolerance (which could have a default setting, but could also be over-ridden by the user as an energy convergence has a much more physically transparent meaning).

In structural optimisation, the initial optimisation steps do not require R_{cut} and τ to be set to high values. By the end of the calculation, if a very precise final structure

is needed, the reverse is true. By linking the maximum component of the force to the value of R_{cut}, and especially τ, both efficient and extremely accurate AF calculations could be performed transparently automatically without user intervention. In the spirit of the previous paragraph, the final value fo τ could be chosen to guarantee the default final force tolerance in an optimisation (e.g. 1 meV/Å) is safely achieved.

Moving on to the determination of saddle points, an investigation into the effect of providing the Lanczos with an uphill direction from a NEB calculation would be both interesting and useful. This has the potential to greatly speed up transtion state indentification, and to allow larger structures, and more complex reactions to be investigated.

By combining the Lanczos method, AF, and the three techniques mentioned above, filtered transition state identifications using the Lanczos method could be made to be extremely efficient, and accurate. This has the potential to open up structures for transition state identification, whose size previously precluded them from full DFT calculations, which is an exciting prospect.

References

[1] N. L. Allinger. Conformational analysis. 130. mm2. a hydrocarbon force field utilizing v1 and v2 torsional terms. *J. Am. Chem. Soc.*, 99:8127–8134, 1977.

[2] G. R. Bachelet, D. R. Hamann, and M. Schluter. Pseudopotentials that work: From h to pu. *Phys. Rev. B*, 26:4199–4228, 1982.

[3] J. Baker and F. Chan. The location of transition states: A comparison of cartesian, z-matrix, and natural internal coordinates. *J. Comp. Chem.*, 17:888–904, 1995.

[4] A. Baldereschi. Mean-value point in the brillouin zone. *Phys. Rev. B*, 7:5212, 1973.

[5] G. T. Barkema and N. Mousseau. Event-based relaxation of continuous disordered systems. *Phys. Rev. Lett.*, 77:4358, 1996.

[6] A. D. Becke. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A*, 6:3098–3100, 1988.

[7] P. E. Blochl. Projector augmented-wave method. *Phys. Rev. B*, 50:17953–17979, 1994.

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

[9] M. Born and J. R. Oppenheimer. Zur quantentheorie der molekeln. *Ann. Phys.*, 389:457–484, 1927.

[10] D. R. Bowler and M. J. Gillan. An efficient and robust technique for achieving self consistency in electronic structure calculations. *Chem. Phys. Lett.*, 325:473–476, 2000.

[11] S. J. Breuer and P. R. Briddon. Ab initio investigation of the native defects in diamond and self-diffusion. *Phys. Rev. B*, 51:6984, 1995.

[12] P. R. Briddon and Jones. R. Efficient iteration scheme for self-consistent pseudopotential calculations. *Phys. Status Solidi B*, 217:131–171, 2000.

[13] C. G. Broyden. The convergence of a class of double-rank minimization algorithms. *Journal of the Institute of Mathematics and Its Applications*, 6:76–90, 1970.

[14] J. M. Carr, S. A. Trygubenko, and D. J. Wales. Finding pathways between distant local minima. *J. Chem. Phys.*, 122:234903–234910, 2005.

[15] D. M. Ceperley and B. J. Alder. Ground state of the electron gas by a stochastic method. *Phys. Rev. Lett.*, 45:566–569, 1980.

[16] J. W. Chu, B. L. Trout, and B. R. Brooks. A super-linear minimization scheme for the nudged elastic band method. *J. Chem. Phys.*, 119:12708–12717, 2003.

[17] J. Coutinho, R. Jones, P. R. Briddon, and S. Oberg. Oxygen and dioxygen centers in si and ge: Density-functional calculations. *Phys. Rev. B*, 62:10824–10840, 2000.

[18] M. J. S. Dewar and W. Thiel. Ground states of molecules. 38. the mndo method. approximations and parameters. *J. Am. Chem. Soc.*, 99:4899–4907, 1977.

[19] O. Farkas and H. B. Schlegel. Methods for optimizing large molecules. ii. quadratic search. *J. Chem. Phys.*, 111:10806–10814, 1999.

[20] O Farkas and H. B. Schlegel. Methods for optimizing large molecules part iii. an improved algorithm for geometry optimization using direct inversion in the iterative subspace (gdiis). *Phys. Chem. Chem. Phys.*, 4:11–15, 2002.

[21] E. Fermi. Un metodo statistico per la determinazione di alcune prioprietà dell'atomo. *Rend. Accad. Naz. Lincei*, 6:602–607, 1927.

[22] R. Fletcher. A new approach to variable metric algorithms. *The Computer Journal*, 13:317–322, 1970.

[23] V. Fock. Näherungsmethode zur lösung des quantenmechanischen mehrkörperproblems. *Z. Phys.*, 61:126–148, 1930.

[24] D. Goldfarb. A family of variable metric updates derived by variational means. *Math. Comput.*, 24:23–26, 1970.

[25] J. P. Goss, M. J. Shaw, and P. R. Briddon. Marker-method calculations for electrical levels using gaussian-orbital basis sets. *Topics in Applied Physics*, 104:69–94, 2007.

[26] D. R. Hamann, M. Schluter, and C. Chiang. Norm-conserving pseudopotentials. *Phys. Rev. Lett.*, 43:1494–1497, 1979.

[27] B. Hammer, L. B. Hansen, and J. K. Norskov. improved adsorption energies within dft using revised pbe functionals. *Phys. Rev. B*, 59:7413, 1999.

[28] D. R. Hartree. The wave mechanics of an atom with a non-coulomb central field part i theory and methods. *Proc. Cambridge Philos. Soc.*, 24:89, 1927.

[29] C. Hartwigsen, S. Goedecker, and J. Hutter. Relativistic seperable dual-space gaussian pseudopotentials from h to rn. *Phys. Rev. B*, 58:3641–3662, 1998.

[30] W. J. Hehre. *Practical Strategies for Electronic Structure Calculation.* Wavefunction, 1995.

[31] G. Henkelman and H. Jonsson. A dimer method for finding saddle points on high dimensional potential surfaces using only first derivatives. *J. Chem. Phys.*, 111:7010–7022, 1999.

[32] G. Henkelman and H. Jonsson. Improved tangent estimate in the nudged elastic band method for nding minimum energy paths and saddle points. *J. Chem. Phys.*, 113:9978–9985, 2000.

[33] M. R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *J. Res. Natl. Bur. Stand.*, 49:409, 1952.

[34] N. D. M. Hine, P. D. Haynes, A. A. Mostofi, and M. C. Payne. Linear-scaling density-functional simulations of charged point defects in al2o3 using hierarchical sparse matrix algebra. *J. Chem. Phys.*, 133:114111(1)–114111(12), 2010.

[35] D. C. Hunt, D. J. Twitchen, M. E. Newton, J. M. Baker, T. R. Anthony, W. F. Banholzer, and S. S. Vagarali. Identification of the neutral carbon ¡100¿-split interstitial in diamond. *Phys. Rev. B*, 61:3863–3876, 2000.

[36] F. Jensen. *Introduction to Computational Chemistry*. Wiley, 2007.

[37] B. G. Johnson, M. W. P. Gill, and J. A. Pople. Preliminary results on the performance of a family of density functional methods. *J. Chem. Phys.*, 97:7846–7848, 1992.

[38] R. Jones and P. R. Briddon. The ab-initio cluster method and the dynamics of defects in semiconductors. In *Semiconductors and Semimetals*, volume 51A. Academic Press, Boston, 1998.

[39] G. P. Kerker. Non-singular atomic pseudopotentials for solid state applications. *J. Phys. C: Solid State Phys.*, 13:811–814, 1980.

[40] W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Phys. Rev.*, 140:A1133–A1138, 1965.

[41] D. B. Laks and C. G. Van de Walle. *Wide-band-gap Semiconductors: Doping limits in ZnSe*. North-Holland, 1992.

[42] W. K. Leung, R. J. Needs, and G. Rajagopal. Calculations of silicon self-interstitial defects. *Phys. Rev. Lett.*, 83:2351–2354, 1999.

[43] R. Malek and N. Mousseau. Dynamics of lennard-jones clusters: A characterization of the activation-relaxation technique. *Phys. Rev. E*, 62:7723–7728, 2000.

[44] P. E. Maslen, C. Ochsenfeld, C. A. White, M. S. Lee, and M. Head-Gordon. Locality and sparsity of ab initio one-particle density matrices and localized orbitals. *J. Phys. Chem.*, 102:2215–2222, 1998.

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

[46] N. Mousseau and G. T. Barkema. Traveling through potential energy surfaces of disordered materials: the activation-relaxation technique. *Phys. Rev. E*, 57:2419–2424, 1998.

[47] R. A. Olsen, G. J. Kroes, G. Henkelman, A. Arnaldsson, and H. Johnsson. Comparison of methods for finding saddle points without knowledge of the final states. *J. Chem. Phys.*, 121:9776–9792, 2004.

[48] J. P. Perdew, K. Burke, and M. Ernzerhof. Generalised gradient approximation made simple. *Phys. Rev. Lett.*, 77:3865–3868, 1996.

[49] J. P. Perdew and Y. Wang. Accurate and simple analytic representation of the electron-gas correlation-energy. *Phys. Rev. B*, 45:13244–13249, 1992.

[50] J. P. Perdew and A. Zunger. Self-interaction correction to density-functional approximations for many-electron systems. *Phys. Rev. B*, 23:5048–5079, 1981.

[51] N. Pinho, B. J. Coomer, J. P. Goss, and R. Jones. The tri-interstitial defect in si. *ENDEASD 2000*, 2000.

[52] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. *Numerical Recipes in FORTRAN 77*. Cambridge University Press, 1992.

[53] P. Pulay. Convergence acceleration of iterative sequences. the case of scf iteration. *Chem. Phys. Lett.*, 73:393–398, 1980.

[54] P. Pulay. Improved scf convergence acceleration. *J. Comput. Chem.*, 3:556–560, 1982.

[55] M. J. Rayson and P. R. Briddon. Rapid iterative method for electronic-structure eigenproblems using localised basis functions. *Comput. Phys. Commun.*, 178:128–134, 2008.

[56] M. J. Rayson and P. R. Briddon. Highly efficient method for kohn-sham density functional calculations of 500-10 000 atom systems. *Phys. Rev. B*, 80:205104, 2009.

[57] D. A. Richie, J. Kim, S. A. Barr, K. R. A. Hazzard, R. Hennig, and J. W. WIlkins. Complexity of small silicon self-interstitial defects. *Phys. Rev. Lett.*, 92:0444011–0444014, 2004.

[58] D. F. Shanno. A family of variable metric updates derived by variational means. *Math. Comput.*, 24:647–656, 1970.

[59] M. J. Shaw, P. R. Briddon, J. P. Goss, M. J. Rayson, A. Kerridge, A. H. Harker, and A. M. Stoneham. Importance of quantum tunneling in vacancy-hydrogen complexes in diamond. *Phys. Rev. Lett.*, 95(10):105502, 2005.

[60] S. Simdyankin. personal communication.

[61] J. Slater. A simplification of the hartree-fock method. *Phys. Rev.*, 81:385–390, 1951.

[62] J. Slater and G. Koster. Simplified lcao method for the periodic potential problem. *Phys. Rev.*, 94:1498–1524, 1954.

[63] J. C. Slater. Note on hartree's method. *Phys. Rev.*, 35:210–211, 1930.

[64] E. Teller. On the stability of molecules in the thomas–fermi theory. *Rev. Mod. Phys.*, 34:627–631, 1962.

[65] J. M. Thijssen. *Computational Physics*. Cambridge University Press, 2007.

[66] L. H. Thomas. The calculation of atomic fields. *Proc. Cambridge Philos. Soc.*, 23:542–548, 1927.

[67] U. von Barth and L. Hedin. A local exchange-correlation potential for the spin polarized case. *J. Phys. C*, 5:1629–1642, 1972.