Implementation and application of advanced density functionals

Michael Christopher Gibson

Implementation and Application of Advanced Density Functionals

Michael Christopher Gibson

A thesis submitted for the degree of

Doctor of Philosophy

Department of Physics

University of Durham

2006

Implementation and Application of Advanced Density Functionals

Michael Christopher Gibson

Abstract

Density functional theory (DFT) is a method of effectively solving the many-electron Schrödinger equation, enabling the properties of condensed matter systems to be calculated from first principles. With the commonly used local density approximation (LDA), and generalised gradient approximations (GGAs), to the exchange correlation functional, it is currently possible to perform calculations on systems containing several hundred atoms. The accuracy of such calculations depends on the system under study and on which particular properties one wishes to calculate. The use of more advanced functionals has the potential to improve accuracy, at the expense of greater computational demand. In this work we use the LDA to calculate certain properties of GaN, such as geometry, band structure, and surface properties, including the reconstruction of GaN surfaces under the presence of hydrogen. We then describe our computational implementation of advanced density functionals, including screened exchange (sX-LDA), Hartree-Fock (HF), and exact exchange (EXX), within an efficient, fully parallel, plane wave code. The implementation of sX-LDA and HF is used to calculate band structure properties of Si, GaN, and other simple semiconductors, and it is found that sX-LDA can improve results significantly beyond the LDA. We also derive and implement the theory that allows one to calculate directly the contribution to the stress tensor from exchange and correlation when using these functionals, and demonstrate this with some simple test cases. Finally, we introduce some new theoretical ideas that may pave the way for yet more accurate density functionals in the future.

Publications

The work presented here has contributed to the following publications:

- *Screened Exchange Stress Tensor in Density Functional Theory*, M. C. Gibson, S. Brand, S. J. Clark, Phys. Rev. B. **73**, 125120 (2006).

- *Screened Exchange Calculations of Semiconductor Band Structures*, M. C. Gibson, S. J. Clark, S. Brand, R. A. Abram, AIP Conf. Proc. **772**, 1125 (2005).

- *First-principles calculations of* 2×2 *reconstructions of* $GaN(0001)$ *surfaces involving N, Al, Ga, In, and As atoms*, V. Timon, S. Brand, S. J. Clark, M. C. Gibson, R. A. Abram, Phys. Rev. B **72**, 035327 (2005).

- *Defect Energy Levels in* HfO_2, *High-Dielectric-Constant Gate Oxide*, A. K. Xiong, J. Robertson, M. C. Gibson, S. J. Clark, App. Phys. Lett. **87**, 183505 (2005).

Declaration

The work presented here was undertaken within the Department of Physics at the University of Durham between August 2002 and August 2005. I confirm that no part of this work has previously been submitted for a degree at this or any other institution and, unless otherwise stated, it is the original work of the author.

Michael C. Gibson

May 2006

Acknowledgements

First and foremost I would like to thank my supervisor Stewart Clark for his help, advice, curry, and malt whisky over the past few years. Thanks also to my co-supervisor Stuart Brand, head of research group Richard Abram, and everyone else who has been involved with the Durham condensed matter theory group during my time here. In particular I would like to thank Ian Bolland for helping learn the basics of Linux when I arrived, Paul Tulip for his entertaining office rants, and Dom Jochym for some Linux-related tips that have proved very useful in writing this thesis. Outside of work I would like to thank my girlfriend and my family for their constant support and encouragement, without which the completion of this work would not have been possible.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this chapter we outline the theoretical framework that underpins the bulk of this work. We begin with a brief introduction to the concept of first principles calculations, and explain how the physics of condensed matter can be described by the many-electron Schrödinger equation. We then discuss how density functional theory (DFT)[1] represents, in principle, an alternative route to many-electron quantum mechanics for performing such calculations. We describe the Kohn-Sham method of dealing with DFT, and how this can be implemented efficiently with a plane wave basis set, using appropriate approximations (a more detailed account of these methods may be found in references [1, 2, 3]). We then discuss in more detail the issue of exchange and correlation in DFT. We also include a brief description of the CASTEP code, which we use to run all of the calculations in this work. Finally, after summarising this introductory chapter, we provide a chapter by chapter overview of the rest of the thesis.

[1] See Appendix A for a list of symbols and abbreviations used throughout this work.

1.1 Physics from First Principles

1.1.1 What are First Principles Calculations?

The world around us is made of *condensed matter*, i.e. matter whose energy is low enough that it has condensed to form stable systems of atoms and molecules, usually in solid or liquid phases. The large variety of ways in which these systems can take form leads to a rich diversity of physical phenomena that is practically endless in scope.

Because of this, approaching the field of condensed matter physics from a theoretical or computational angle can be a very challenging task to undertake. For the most part, the way this is done is to pick a particular macroscopic phenomenon, which has been well studied experimentally, and to build empirical, or semi-empirical, models to describe the experimentally observed results. This often provides a good understanding of the physics of the system under study, and it is often possible to interpolate or extrapolate these models in order to predict the behaviour of systems under conditions not yet tested experimentally. However, due to the complexity of condensed matter systems, and the difficulty in building accurate models, the predictive power of such an approach can be severely limited.

The *first principles* approach to condensed matter theory is entirely different from this. It starts from what we know about all condensed matter systems - that they are made of atoms, which in turn are made of a positively charged nucleus, and a number of negatively charged electrons. The interactions between atoms, such as chemical and molecular bonding, are determined by the interactions of their constituent electrons and nuclei. All of the physics of condensed matter systems arises ultimately from these basic interactions. If we can model these interactions accurately, then all of the complex physical phenomena that arise from them should emerge naturally in our calculations.

The physics that describes the interaction of electrons and nuclei that is relevant to most problems in condensed matter is actually relatively simple. There are only two different types of particle involved, and the behaviour of these particles is mostly

governed by basic quantum mechanics. What makes first principles calculations difficult is not so much the complexity of the physics, but rather the size of the problem in terms of a numerical formulation. The development of accurate and efficient theoretical and computational techniques for dealing with so many particles is therefore central to the ongoing research in this field.

1.1.2 The Many-Electron Schrödinger Equation

As we have discussed, condensed matter is made of positively charged nuclei and negatively charged electrons. Electrons behave as point-like particles and, to a very good approximation, the nuclei can be considered to be point-like particles also. The complete system of N electrons, and N_I nuclei[2], is described by the many-body Schrödinger equation[3][4]:

$$\hat{T}\Psi_{\text{MB}} + \hat{V}\Psi_{\text{MB}} = -i\frac{d\Psi_{\text{MB}}}{dt}, \tag{1.1}$$

where Ψ_{MB} is the many-body wavefunction, and $\hat{T}$ and $\hat{V}$ are the many-body kinetic energy and potential energy operators respectively. The many-body wavefunction contains the quantum *probability amplitude* for every possible configuration of electrons and nuclei, i.e.

$$\Psi_{\text{MB}} = \Psi_{\text{MB}}(\mathbf{r}_1, \mathbf{r}_2, \cdots \mathbf{r}_N, \mathbf{R}_{I_1}, \mathbf{R}_{I_2}, \cdots \mathbf{R}_{I_{N_I}}), \tag{1.2}$$

where the $\mathbf{r}_n$ are the coordinates of the electrons and the $\mathbf{R}_{I_n}$ are the coordinates of the nuclei. The coordinates $\mathbf{r}_n$ and $\mathbf{R}_{I_n}$ include spin degrees of freedom as well as position (see Appendix C). This is not especially important at this stage because spin does not enter the Hamiltonian, but will become more important later, in 1.2.2, when the issue will be discussed in more detail.

The many-body kinetic energy operator operates on the wavefunction as follows:

$$\hat{T}\Psi_{\text{MB}} = -\frac{1}{2}\left(\sum_n \nabla^2_{\mathbf{r}_n} + \sum_n \frac{\nabla^2_{\mathbf{R}_{I_n}}}{M_{I_n}}\right)\Psi_{\text{MB}}, \tag{1.3}$$

[2]The subscript I stands for "ion", as opposed to the more obvious "nucleus". The reason for this will become clear later in the context of pseudopotentials.

[3]In atomic units. For an explanation of the units used in this work see Appendix B

[4]This does not account for relativistic effects. In certain situations, relativistic effects must be included [4, 5, 6], but this will not be directly relevant to this work.

where the M_{I_n} are the masses of the nuclei. The operator, $\nabla^2_{\mathbf{r}_n}$, is the Laplacian operator for the spatial coordinates of the nth electron, i.e.

$$\nabla^2_{\mathbf{r}_n} = \frac{\partial^2}{\partial x_n^2} + \frac{\partial^2}{\partial y_n^2} + \frac{\partial^2}{\partial z_n^2}. \tag{1.4}$$

The many-body potential energy operator operates on the wavefunction as follows:

$$\hat{V}\Psi_{\text{MB}} = \left(\frac{1}{2} \sum_{n \neq m} \frac{1}{|\mathbf{r}_n - \mathbf{r}_m|} - \sum_{n,m} \frac{Z_{I_m}}{|\mathbf{r}_n - \mathbf{R}_{I_m}|} + \frac{1}{2} \sum_{n \neq m} \frac{Z_{I_n} Z_{I_m}}{|\mathbf{R}_{I_n} - \mathbf{R}_{I_m}|} \right) \Psi_{\text{MB}}, \tag{1.5}$$

where the Z_{I_n} are the charges of the nuclei.

Because the inverse masses of the nuclei are very small ($\sim 5.4 \times 10^{-4} m_e^{-1}$ in the case of hydrogen), we can use the Born-Oppenheimer, or adiabatic, approximation[7, 8], and treat the nuclei as classical particles that move on a time scale much longer than that of the electrons. This means that, as far as calculations on the electrons are concerned, the nuclei can be considered to be fixed in space.

The many-body problem is therefore reduced to the smaller problem of a system of electrons moving in some external potential, i.e. the potential created by the positively charged nuclei. The Schrödinger equation for this system is then

$$\hat{T}\Psi + \hat{V}\Psi = -i\frac{d\Psi}{dt}, \tag{1.6}$$

where Ψ is the many-electron wavefunction. This is the central object in electronic structure calculations, as it contains all the information about the system of electrons. It gives the probability amplitude for finding the system of electrons in a given configuration, i.e.

$$\Psi = \Psi(\mathbf{r}_1, \mathbf{r}_2, \cdots \mathbf{r}_N), \tag{1.7}$$

where the $\mathbf{r}_n$ are the coordinates of the electrons. Again, spin is included in the coordinates, $\mathbf{r}_n$, so that $\mathbf{r} = (x, y, z, \sigma)$, where σ is the spin coordinate, and can take the values of $\uparrow$ (spin-up) or $\downarrow$ (spin-down).

$\hat{T}$ is now the many-electron kinetic energy operator, acting on Ψ as

$$\hat{T}\Psi = -\frac{1}{2} \sum_n \nabla^2_{\mathbf{r}_n} \Psi. \tag{1.8}$$

$\hat{V}$ is the many-electron potential operator, which acts on Ψ as

$$\hat{V}\Psi = \left(\frac{1}{2}\sum_{n\neq m}\frac{1}{|\mathbf{r}_n - \mathbf{r}_m|} + \sum_n v_{ext}(\mathbf{r}_n)\right)\Psi, \tag{1.9}$$

where v_{ext} is the external potential in which the electrons are moving. For the system of electrons and nuclei this given by

$$v_{ext}(\mathbf{r}) = -\sum_n \frac{Z_{I_n}}{|\mathbf{r} - \mathbf{R}_{I_n}|}. \tag{1.10}$$

Solutions of the many-electron Schrödinger equation must satisfy the constraints of normalisation and exchange anti-symmetry. Normalisation simply ensures that the total probability for every possible configuration of electrons is equal to 1, i.e.

$$\langle\Psi|\Psi\rangle = \int d\mathbf{r}_1 \cdots \int d\mathbf{r}_N \Psi^*\Psi = 1, \tag{1.11}$$

while exchange anti-symmetry ensures that the wavefunction is anti-symmetric with respect to the exchange of any two electrons' coordinates (including spin), which must be the case for any system of identical fermions, i.e.

$$\Psi(\mathbf{r}_1, \mathbf{r}_2, \cdots \mathbf{r}_n, \cdots \mathbf{r}_m, \cdots \mathbf{r}_N) = -\Psi(\mathbf{r}_1, \mathbf{r}_2, \cdots \mathbf{r}_m, \cdots \mathbf{r}_n, \cdots \mathbf{r}_N). \tag{1.12}$$

For most purposes, we are only actually interested in the ground state of the electronic system. This is the lowest energy solution of the time-independent many-electron Schrödinger equation,

$$\hat{T}\Psi + \hat{V}\Psi = E\Psi, \tag{1.13}$$

where E is the ground state energy of the system of electrons. Calculations involving solution of this equation are known as *electronic structure* calculations.

Of course, we must not forget about the nuclei. Even though we are treating them as being fixed in space from the point of view of the electrons, we may well wish to study the evolution of a system on the longer time scales over which nuclear motion takes place. Also, we may wish to find the configuration of nuclei that gives the lowest *total energy* for the complete system, as this is the configuration that a real system will naturally tend to adopt at low temperature.

Assuming that we can solve the many-electron Schrödinger equation for a given external potential in order to obtain the electronic ground state energy, E, then we can consider this energy to be a function of the nuclear coordinates, i.e.

$$E = E(\mathbf{R}_{I_1}, \mathbf{R}_{I_2}, \cdots \mathbf{R}_{I_{N_I}}). \tag{1.14}$$

This energy includes the internal energy of the electrons and the energy due to the interaction of the electrons with the nuclei, but it does not include the energy due to the interactions of the nuclei with each other. This is the potential energy due to the mutual repulsion of the positively charged nuclei, and is given by

$$V_{I-I} = \frac{1}{2} \sum_{n \neq m} \frac{Z_{I_n} Z_{I_m}}{|\mathbf{R}_{I_n} - \mathbf{R}_{I_m}|}. \tag{1.15}$$

The total energy of the system is therefore

$$E_{TOT} = E + V_{I-I}. \tag{1.16}$$

This total energy is a function of the nuclear coordinates and governs the motion of the nuclei. So, within the Born-Oppenheimer approximation, and at temperatures for which the electrons can be considered to remain in their ground state, solution of the many-electron Schrödinger equation allows us, in principle, to predict the motion of the atoms of any condensed matter system, and also to calculate its lowest energy structure.

1.1.3 Extracting Quantities from the Many-Electron Wavefunction

We will now look at some important quantities that are stored in the many-electron wavefunction that will be of use in later sections. The first, and most important in the context of this work, is the *electron density*, $\rho(\mathbf{r})$, given by:

$$\rho(\mathbf{r}) = N \sum_{\sigma} \int d\mathbf{r}_2 \cdots \int d\mathbf{r}_N \Psi^*(\mathbf{r}, \mathbf{r}_2, \cdots \mathbf{r}_N) \Psi(\mathbf{r}, \mathbf{r}_2, \cdots \mathbf{r}_N). \tag{1.17}$$

This gives the probability density for finding an electron at position $\mathbf{r}$[5]. As well as this, we have the *electron pair density*, $\rho(\mathbf{r},\mathbf{r}')$, given by:

$$\rho(\mathbf{r},\mathbf{r}') = N(N-1)\sum_{\sigma\sigma'}\int d\mathbf{r}_3\cdots\int d\mathbf{r}_N\Psi^*(\mathbf{r},\mathbf{r}',\mathbf{r}_3,\cdots\mathbf{r}_N)\Psi(\mathbf{r},\mathbf{r}',\mathbf{r}_3,\cdots\mathbf{r}_N), \quad (1.18)$$

which is the probability density for simultaneously finding one electron at position $\mathbf{r}$ and another electron at position $\mathbf{r}'$. We also have the kinetic energy, T, given by

$$\begin{aligned} T &= \langle\Psi|\hat{T}|\Psi\rangle \\ &= -\frac{1}{2}\int d\mathbf{r}_1\cdots\int d\mathbf{r}_N\Psi^*\sum_n\nabla^2_{\mathbf{r}_n}\Psi, \end{aligned} \quad (1.19)$$

and the potential energy, V, given by

$$V = \langle\Psi|\hat{V}|\Psi\rangle, \quad (1.20)$$

which can be separated into two parts, the *external potential energy*, which is the potential energy due to the external potential, given by

$$\begin{aligned} V_{ext} &= \langle\Psi|\hat{V}_{ext}|\Psi\rangle \\ &= \int d\mathbf{r}_1\cdots\int d\mathbf{r}_N\Psi^*\left(\sum_n v_{ext}(\mathbf{r}_n)\right)\Psi \quad (1.21) \\ &= \int d\mathbf{r}v_{ext}(\mathbf{r})\rho(\mathbf{r}), \quad (1.22) \end{aligned}$$

and the *internal potential energy*, which is the potential energy due to the electron-electron repulsion, given by

$$\begin{aligned} V_{int} &= \langle\Psi|\hat{V}_{int}|\Psi\rangle \\ &= \frac{1}{2}\int d\mathbf{r}_1\cdots\int d\mathbf{r}_N\Psi^*\left(\sum_{n\neq m}\frac{1}{\mathbf{r}_n-\mathbf{r}_m}\right)\Psi, \end{aligned} \quad (1.23)$$

so that we have

$$V = V_{ext} + V_{int}. \quad (1.24)$$

[5]The sum over spin coordinates takes account of the fact that $\mathbf{r}$ only includes spin on the right hand side of this equation.

We can also define the *internal electronic energy*, F, as the sum of the kinetic energy and internal potential energy:

$$\begin{aligned} F &= \langle \Psi | \hat{F} | \Psi \rangle \\ &= \langle \Psi | \hat{T} + \hat{V}_{int} | \Psi \rangle \\ &= T + V_{int}. \end{aligned} \tag{1.25}$$

As we will see in the next section, it is the evaluation of the internal energy, F, that presents the main challenge in electronic structure calculations.

1.2 Basics of Density Functional Theory

Analytical solution of the the many-electron Schrödinger equation is not possible, and numerical solution, while perfectly possible in theory, is effectively impossible in practice for more than a handful of electrons due to the finite speed and memory of computers. In this section we introduce *density functional theory* (DFT) as a means of circumventing solution of the many-electron Schrödinger equation when calculating the ground state energy. This can, with appropriate approximations, lead to methods that are computationally feasible. Alternative methods also exist to DFT, some of which are more accurate than present DFT formulations, but at the expense of greater computational demand. These include quantum Monte-Carlo (QMC) [9, 10, 11] and configuration interaction (CI) methods [12]. However, the balance that present DFT methods strike between accuracy and computational efficiency mean that DFT is currently the most popular method of performing first principles calculations on extended systems.

1.2.1 The Hohenberg-Kohn Theorem

Density functional theory (DFT) is founded on the Hohenberg-Kohn theorem[13]. This comes in two parts, the first of which states that the ground

state energy of a system of electrons is a unique functional of the ground state density:

$$E_{GS} = E[\rho_{GS}]. \tag{1.26}$$

In fact all properties of the system, including excited state properties, are, in principle, exact functionals of the ground state density. The reason for this, as was proven by Hohenberg and Kohn, is that there is a one-to-one mapping between the ground state density and the external potential. If we happen to know the ground state density, then, in principle, we know the external potential, and if we know the external potential we can, again in principle, solve the many-electron Schrödinger equation and know everything about the system. Of course, this is not yet of any practical use, because the whole point of using DFT is so that we can avoid having to deal with the many-electron Schrödinger equation. Nevertheless, we are provided, at least in principle, with a means of finding the ground state energy for a given external potential. The internal electronic energy, F, of a system in its ground state can be expressed as

$$F = E - V_{ext}, \tag{1.27}$$

where V_{ext} is the external potential energy, given by

$$V_{ext} = \int d\mathbf{r} v_{ext}(\mathbf{r})\rho(\mathbf{r}). \tag{1.28}$$

Since E and v_{ext} are functionals of the density, it follows that F is also a functional of the density.

Supposing we now have an external potential and a ground state density, which may or may not be the ground state density corresponding to that potential, we can define the *variational energy*, E_{var}, as

$$E_{var}[v_{ext}, \rho] = F[\rho] + \int d\mathbf{r} v_{ext}(\mathbf{r})\rho(\mathbf{r}). \tag{1.29}$$

The true ground state density for $v_{ext}(\mathbf{r})$ is the density that minimises this energy - this is the second part of the Hohenberg-Kohn theorem. If we were able to calculate $F[\rho]$ for any given density, then we could perform a search to find the ground state density for any given external potential.

Such a search may be complicated by the fact that we have so far only defined the functional $F[\rho]$ for densities that correspond to the ground state of some external potential; such densities are described as being *V-representable.* It may be the case that, during a search, we would encounter densities that did not correspond to the ground state of any external potential. This problem can be overcome by extending the definition of $F[\rho]$ to include such densities, so long as E_{var} is still minimised by the correct ground state density. We now define $F[\rho]$ as the minimum internal electronic energy of any electronic wavefunction, ground state or otherwise, whose corresponding density is equal to ρ, i.e.

$$F[\rho] = \min_{\Psi \to \rho} \langle \Psi | \hat{F} | \Psi \rangle. \tag{1.30}$$

Essentially all densities, $\rho(\mathbf{r})$, that integrate to N, correspond to some N-electron wavefunction [14]; this property is described as *N-representability.* At this stage we are still no nearer to a practical method because exact evaluation of $F[\rho]$ would require us to solve the many-body Schrödinger equation. But, supposing we have a functional that is a good approximation to $F[\rho]$, but that can be evaluated in a practical manner, then a search should lead us to a good approximation to the ground state energy and density. This is the fundamental principle upon which all practical DFT calculation are founded.

1.2.2 The Kohn-Sham Method

The Kohn-Sham method [15] is a formulation of DFT that lends itself to finding good approximations to F. Essentially what it does is define a set of component energies that sum to give F, each of which has a clear physical origin, and some of which can be evaluated very easily. Only those components that cannot be easily evaluated are subject to approximation.

Central to the Kohn-Sham method is the introduction of a fictitious auxiliary system, which is intended in some way to mimic the true many-electron system that we are dealing with. This fictitious system is a set of particles whose properties are identical to those of electrons, except that the electron-electron repulsive interaction

is switched off. The particles move in some fictitious external potential, $\mu_{KS}(\mathbf{r})$, known as the *Kohn-Sham potential*, which is defined such that the system's ground state density equals $\rho(\mathbf{r})$ - the same density as the electronic system for which we want to evaluate $F[\rho]$. This assumes that the true ground state density actually is also the ground state density of a non-interacting system, a property described as *non- interacting μ-representability*, and while no proof exists that this is true of all true ground state densities, no examples exist to the contrary and the assumption is generally accepted as being reasonable.

Because there are no interactions between the particles, the ground state wavefunction of this system is far less complicated that of the true, interacting, system. In fact, we can write the ground state wavefunction explicitly in terms of simple single-particle wavefunctions. The only complication is that the full wavefunction, $\mathbf{\Psi}_S$, must still satisfy exchange anti-symmetry, and this can be achieved by placing single-particle wavefunctions in a *Slater determinant* [16], as follows:

$$\mathbf{\Psi}_S(\mathbf{r}_1, \cdots, \mathbf{r}_N) = \frac{1}{\sqrt{N!}} \begin{vmatrix} \psi_1(\mathbf{r}_1) & \cdots & \psi_1(\mathbf{r}_N) \\ \vdots & \ddots & \vdots \\ \psi_N(\mathbf{r}_1) & \cdots & \psi_N(\mathbf{r}_N) \end{vmatrix}, \tag{1.31}$$

where the $\psi_i(\mathbf{r})$ are the lowest N eigenstates satisfying the following Schrödinger-like equation:

$$-\frac{1}{2}\nabla^2\psi_i(\mathbf{r}) + \mu_{KS}(\mathbf{r})\psi_i(\mathbf{r}) = \varepsilon_i\psi_i(\mathbf{r}). \tag{1.32}$$

These single-particle wavefunctions are known as the *Kohn-Sham orbitals*. Exchange anti-symmetry is ensured by the property of determinants that swapping any pair of rows or columns simply causes the value of the determinant to switch sign.

The Kohn-Sham Slater determinant can be written in a more compact way by introducing a permutation operator, $P_n(i)$, where each n labels a different one of the $N!$ permutations of the numbers 1 to N. The particular order of the permutations is essentially arbitrary, but for convenience we define P_1 as the numbers in increasing order, i.e. $P_1(i) = i$, and require that each permutation can be generated from the previous one by swapping a pair of numbers. We can then write the Slater

determinant as follows:

$$\Psi_S = \frac{1}{\sqrt{N!}} \sum_{n=1}^{N!} (-1)^{n-1} \prod_{i=1}^{N} \psi_i(\mathbf{r}_{P_n(i)}). \tag{1.33}$$

At this stage, $\mathbf{r}$ implicitly contains both the position and spin coordinates of a particle. Within the standard Kohn-Sham theory, there is always a degeneracy between spin-up states and spin-down states. This means we can require, without loss of generality, that each Kohn-Sham orbital is an eigenstate of the $\hat{S}_z$ operator. We then have a set of spin-up orbitals and a set of spin-down orbitals. For each spin-up orbital there is a corresponding spin-down orbital with the same spatial wavefunction and eigenvalue (except in the case of odd N where the highest orbital will be unpaired).

Because of this spin-degeneracy, it is often convenient to deal only with the spatial components of the Kohn-Sham orbitals. We will use the notation $\phi_i(\mathbf{r})$ to represent the spatial component of the i'th spin-degenerate *pair* of Kohn-Sham orbitals. Hence for an N-particle system there will be either $N/2$ (for even N) or $N/2+1$ (for odd N) spatial-only Kohn-Sham orbitals, $\phi_i(\mathbf{r})$. Including spin explicitly now for clarity, the $\psi_i(\mathbf{r},\sigma)$, are related to the $\phi_i(\mathbf{r})$ by

$$\begin{aligned}
\psi_{2i-1}(\mathbf{r},\uparrow) &= \phi_i(\mathbf{r}), \\
\psi_{2i-1}(\mathbf{r},\downarrow) &= 0, \\
\psi_{2i}(\mathbf{r},\uparrow) &= 0, \\
\psi_{2i}(\mathbf{r},\downarrow) &= \phi_i(\mathbf{r}).
\end{aligned} \tag{1.34}$$

Now, by adopting the convention of having implicit factors of 2 where appropriate in summations, as detailed in Appendix C, most formulas expressed in terms of the $\phi_i(\mathbf{r})$ will actually appear to be identical to the equivalent formula expressed in term of the $\psi_i(\mathbf{r})$. For this reason, there is little point in repeating every formula in both ways. The spatial-only form will prove the more useful later in computational contexts, and for this reason we will choose to use this form exclusively in most of this work.

For example, the total density, $\rho(\mathbf{r})$, is given in terms of the orbitals, $\psi_i(\mathbf{r})$, by

$$\rho(\mathbf{r}) = \sum_i \psi_i^*(\mathbf{r})\psi_i(\mathbf{r}), \tag{1.35}$$

or, in terms of the spatial-only orbitals, $\phi_i(\mathbf{r})$, by

$$\rho(\mathbf{r}) = \sum_i \phi_i^*(\mathbf{r})\phi_i(\mathbf{r}), \tag{1.36}$$

as shown in Appendix D.

Because the Kohn-Sham Slater determinant is a many-body wavefunction satisfying exchange anti-symmetry, and produces the same density as the true many-electron wavefunction, it might be considered as a reasonable starting point for extracting physical properties that contribute to F. For the true many-electron wavefunction, we can obtain F by taking the expectation value of its associated operator:

$$\begin{aligned} F &= \langle \Psi | \hat{F} | \Psi \rangle \\ &= \langle \Psi | \hat{T} + \hat{V}_{int} | \Psi \rangle. \end{aligned} \tag{1.37}$$

We could therefore obtain a first guess at F by taking the expectation value of its operator for the Kohn-Sham Slater determinant:

$$\begin{aligned} F_S &= \langle \Psi_S | \hat{F} | \Psi_S \rangle \\ &= \langle \Psi_S | \hat{T} + \hat{V}_{int} | \Psi_\mathbf{S} \rangle \\ &= T_S + V_{int}^{(S)}. \end{aligned} \tag{1.38}$$

Here, T_S, is the *non-interacting kinetic energy* and $V_{int}^{(S)}$ is the *non-interacting internal potential energy.* Derivations of the following formulas for these terms are given in Appendix D; here we will simply quote them.

The non-interacting kinetic energy, T_S, is given in terms of the Kohn-Sham orbitals by

$$T_S = -\frac{1}{2} \sum_i \int d\mathbf{r} \phi_i^*(\mathbf{r}) \nabla^2 \phi_i(\mathbf{r}). \tag{1.39}$$

The non-interacting internal potential energy, $V_{int}^{(S)}$, is given by

$$V_{int}^{(S)} = \frac{1}{2}\int d\mathbf{r}\int d\mathbf{r}'\frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} - \frac{1}{2}\sum_{ij}\int d\mathbf{r}\int d\mathbf{r}'\frac{\phi_i^*(\mathbf{r})\phi_i(\mathbf{r}')\phi_j^*(\mathbf{r}')\phi_j(\mathbf{r})}{|\mathbf{r}-\mathbf{r}'|}. \quad (1.40)$$

This is usually separated into the *Hartree energy*, V_H, and *exchange energy*, V_X, so that we have

$$V_{int}^{(S)} = V_H + V_X, \quad (1.41)$$

where

$$V_H = \frac{1}{2}\int d\mathbf{r}\int d\mathbf{r}'\frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}, \quad (1.42)$$

and

$$V_X = -\frac{1}{2}\sum_{ij}\int d\mathbf{r}\int d\mathbf{r}'\frac{\phi_i^*(\mathbf{r})\phi_i(\mathbf{r}')\phi_j^*(\mathbf{r}')\phi_j(\mathbf{r})}{|\mathbf{r}-\mathbf{r}'|}. \quad (1.43)$$

The Hartree energy, V_H, is equal to the Coulombic self-energy of a stationary, non-quantised, distribution of electric charge of density $\rho(\mathbf{r})$. The name comes from *Hartree theory*, which is a predecessor of DFT [17].

The exchange energy, V_X, then accounts for the quantised nature of the charge, and the fact that the wavefunction is anti-symmetric with respect to the exchange of any two particles' coordinates.

The difference between F_S and the true internal energy, F, is called the *correlation energy*, E_C:

$$E_C = F - F_S. \quad (1.44)$$

This accounts for the fact that in the interacting system the electrons will tend to avoid each other due to their mutual repulsion. To understand how the mutual repulsion of the electrons changes the energy, we can imagine taking the non-interacting system and then switching on the mutual repulsion, while keeping the density fixed and keeping the system in its ground state. The physical effect of the repulsion is that the electrons will tend to avoid each other at close range more than they would in the non-interacting system. In terms of energy, this has two effects:

1. Because any given pair of electrons are less likely to be found near to each other, the internal potential energy is reduced.

2. Because the wavefunction must change in order to describe this greater mutual avoidance, the kinetic energy is increased.

Hence, the correlation energy has both a potential and a kinetic component:

$$E_C = T_C + V_C, \tag{1.45}$$

where

$$T_C = T - T_S, \tag{1.46}$$

and

$$V_C = V_{int} - V_H - V_X. \tag{1.47}$$

Correlation properties can be described as purely many-body properties and they can only be calculated exactly if we solve the many-body Schrödinger equation. In practical calculations, therefore, the correlation energy is always approximated to some degree. Also, while easily accessible in principle, the exchange energy is considerably more expensive to evaluate computationally than the Hartree and non-interacting kinetic energies. It is therefore very common to approximate the exchange energy as well as the correlation energy. This leads to the usual grouping together of the exchange and correlation energies into the *exchange-correlation energy*, E_{XC}, given by

$$E_{XC} = V_X + E_C \tag{1.48}$$

$$= F - T_S - V_H. \tag{1.49}$$

The standard exact expression for F within the Kohn-Sham framework is thus

$$F = T_S + V_H + E_{XC}, \tag{1.50}$$

and the variational energy for a given external potential and density is

$$E_{var}[v_{ext}, \rho] = \int d\mathbf{r} v_{ext}(\mathbf{r})\rho(\mathbf{r}) + T_S[\rho] + V_H[\rho] + E_{XC}[\rho]. \tag{1.51}$$

Practical Kohn-Sham calculations involve searching for the density that minimises this functional.

It should be noted that, although the explicit formulas for T_S and V_X are written in terms of the Kohn-Sham orbitals, they are still functionals of the density because the Kohn-Sham orbitals are themselves functionals of the density.

1.2.3 Approximations for E_{XC}

As we have mentioned, the correlation energy is always approximated in practical calculations and the exchange energy, while obtainable exactly, is usually approximated also. Several approximations exist for the exchange-correlation functional, $E_{XC}[\rho]$, of varying accuracy and computational cost. Some of these approximations are so-called "empirical", or "semi-empirical", functionals, such as B3LYP [18], in which some of the information used to define them is derived from experimental results on particular materials or chemicals. Other approximations fall into the category of "non-empirical" functionals, which are defined purely from the results of first principles calculations.

While popular within the organic chemistry community, empirical, or semi-empirical, functionals tend only to out-perform non-empirical functionals of similar cost for the particular class of organic chemicals for which they are defined. If an element is present that was not present in the "training set", such functionals can perform very poorly [19].

Functionals that are derived from first principles tend to be much more versatile. Much of the motivation for doing first principles calculations is in the ability to predict the results of experiments on complex materials, often in extreme conditions, without access to any experimental data other than the values of the fundamental constants. As soon as one introduces empirical or semi-empirical functionals, the calculations can no longer be considered to be "first principles", and are unlikely to possess the same degree of predictive power. In this work we will only be dealing with pure, non-empirical, functionals.

The simplest, and most commonly used, exchange-correlation functional is the *local density approximation* (LDA)[15]. The LDA, like most functionals, involves defining the *exchange-correlation energy per electron*, $\varepsilon_{XC}(\mathbf{r})$. The precise definition of $\varepsilon_{XC}(\mathbf{r})$, and its contributing components, will be discussed later in 1.2.6; for now we simply need to know that the total exchange-correlation energy can be obtained via

$$E_{XC} = \int d\mathbf{r} \varepsilon_{XC}(\mathbf{r}) \rho(\mathbf{r}). \tag{1.52}$$

The exact value of $\varepsilon_{XC}(\mathbf{r})$ is a functional of the density at every point in space. The LDA makes the apparently drastic approximation that the value of $\varepsilon_{XC}(\mathbf{r})$ is simply a function of the density at the same point, $\mathbf{r}$. Any function could in principle be chosen here, but the one that is most clearly justified from first principles is as follows:

$$\varepsilon_{XC}^{(LDA)}(\mathbf{r}) = \varepsilon_{XC}^{HEG}\left(\rho(\mathbf{r})\right), \tag{1.53}$$

where $\varepsilon_{XC}^{HEG}(\rho)$ is the exchange-correlation energy of a *homogeneous electron gas* (HEG) of density ρ. An HEG is a system of electrons of uniform density in its ground state; such a system can be uniquely specified by its density. Defining the LDA in this way means that the approximation becomes exact in the limit of very slowly varying densities.

Of course, if we are to use the LDA in practical calculations, we need to know the actual numerical value of $\varepsilon_{XC}^{HEG}(\rho)$ for all densities. As there is no exact analytical form for this, we have to rely on interpolation of known numerical results from quantum Monte Carlo (QMC) calculations, such as those performed by Ceperley and Alder [20]. These were used by Perdew et. al. to parameterise numerical formulas for $\varepsilon_{XC}(\rho)$ [21, 22] that are accurate for a wide range of values of ρ.

Another group of simple approximations for $E_{XC}[\rho]$ are generalised gradient approximations (GGAs) [23, 21, 22, 24, 25]. These can be thought of as an extension to the idea of the LDA, in which not only the local density, $\rho(\mathbf{r})$, but also the local density gradient, $|\nabla\rho(\mathbf{r})|$, is used as input to a function for $\varepsilon_{XC}(\mathbf{r})$, i.e.

$$E_{XC}^{(GGA)}[\rho] = \int d\mathbf{r}\rho(\mathbf{r})\varepsilon_{XC}^{(GGA)}\left(\rho(\mathbf{r}), |\nabla\rho(\mathbf{r})|\right). \tag{1.54}$$

Unlike $\varepsilon_{XC}^{(LDA)}(\rho)$, the function $\varepsilon_{XC}^{(GGA)}(\rho)$ lacks a uniquely justifiable form. Often the form is chosen to satisfy various physical constraints, however it is not possible to satisfy every constraint simultaneously and so the form is often chosen according to the nature of the system under study [26]. Within the condensed matter physics community, the most commonly used GGAs are PW91 and PBE [22, 25, 27].

Other, more advanced, functionals beyond the LDA and GGAs also exist, and will be discussed in Chapter 3.

1.2.4 Physical Meaning of $\mu_{KS}(\mathbf{r})$

So far, all we have said about the Kohn-Sham potential $\mu_{KS}(\mathbf{r})$ is that it is the local potential which causes the non-interacting system to adopt a ground state density of $\rho(\mathbf{r})$. We will now show how it can also be defined in terms of functional derivatives, separated into various components of different physical origin, in a similar way to how the variational energy, E_{var}, is separated in Equation (1.51). We start by noting that, since the ground state density, $\rho(\mathbf{r})$, is the density that minimises the variational energy, this energy must be stationary with respect to small changes in the density, i.e.

$$\frac{\delta E_{var}}{\delta \rho(\mathbf{r})} = 0. \tag{1.55}$$

Here, the change in density, $\delta\rho(\mathbf{r})$, is restricted so as to be charge-conserving, i.e.

$$\int d\mathbf{r}\delta\rho(\mathbf{r}) = 0, \tag{1.56}$$

so that the total number of particles, N, remains fixed.

Now, we note that $\rho(\mathbf{r})$ is also the ground state density of the non-interacting system, and must therefore minimise some fictitious energy, E_{fic}, given by

$$E_{fic}[\rho] = T_S[\rho] + \int d\mathbf{r}\mu_{KS}(\mathbf{r})\rho(\mathbf{r}). \tag{1.57}$$

This fictitious energy must also be stationary with respect to small, charge-conserving changes in the density, i.e.

$$\frac{\delta E_{fic}}{\delta \rho(\mathbf{r})} = 0 \tag{1.58}$$

$$\Rightarrow \frac{\delta T_S}{\delta \rho(\mathbf{r})} + \mu_{KS}(\mathbf{r}) = 0$$

$$\Rightarrow \mu_{KS}(\mathbf{r}) = -\frac{\delta T_S}{\delta \rho(\mathbf{r})}. \tag{1.59}$$

If we now take Equation (1.55), and substitute the form for E_{var} of Equation (1.51), we have

$$\begin{aligned}
\frac{\delta}{\delta\rho(\mathbf{r})}\left(\int d\mathbf{r}' v_{ext}(\mathbf{r}')\rho(\mathbf{r}') + T_S[\rho] + V_H[\rho] + E_{XC}[\rho]\right) &= 0, \\
\Rightarrow v_{ext}(\mathbf{r}) + \frac{\delta T_S}{\delta\rho(\mathbf{r})} + \frac{\delta V_H}{\delta\rho(\mathbf{r})} + \frac{\delta E_{XC}}{\delta\rho(\mathbf{r})} &= 0, \\
\Rightarrow v_{ext}(\mathbf{r}) + \frac{\delta V_H}{\delta\rho(\mathbf{r})} + \frac{\delta E_{XC}}{\delta\rho(\mathbf{r})} &= -\frac{\delta T_S}{\delta\rho(\mathbf{r})}. \qquad (1.60)
\end{aligned}$$

Comparing this with Equation (1.59), we arrive at

$$\mu_{KS}(\mathbf{r}) = v_{ext}(\mathbf{r}) + \frac{\delta V_H}{\delta\rho(\mathbf{r})} + \frac{\delta E_{XC}}{\delta\rho(\mathbf{r})}, \qquad (1.61)$$

which defines the Kohn-Sham potential in terms of functional derivatives of the various contributions to the total energy. We can also write this as

$$\mu_{KS}(\mathbf{r}) = v_{ext}(\mathbf{r}) + \mu_H(\mathbf{r}) + \mu_{XC}(\mathbf{r}), \qquad (1.62)$$

where $\mu_H(\mathbf{r})$ is the *Hartree potential*, given by

$$\begin{aligned}
\mu_H(\mathbf{r}) &= \frac{\delta V_H}{\delta\rho(\mathbf{r})} \qquad (1.63) \\
&= \frac{\delta}{\delta\rho(\mathbf{r})}\left(\frac{1}{2}\int d\mathbf{r}' \int d\mathbf{r}'' \frac{\rho(\mathbf{r}')\rho(\mathbf{r}'')}{|\mathbf{r}' - \mathbf{r}''|}\right) \\
&= \int d\mathbf{r}' \frac{\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}, \qquad (1.64)
\end{aligned}$$

and $\mu_{XC}(\mathbf{r})$ is the *exchange-correlation potential*, given by

$$\mu_{XC}(\mathbf{r}) = \frac{\delta E_{XC}}{\delta\rho(\mathbf{r})}. \qquad (1.65)$$

How we evaluate this functional derivative for $\mu_{XC}(\mathbf{r})$ depends entirely on the choice of approximation for $E_{XC}[\rho]$. For advanced functionals, as we will see later, it is often the most challenging task associated with a DFT calculation. For simple functionals, however, things are relatively simple; in the LDA, for example, we have

$$\begin{aligned}
\mu_{XC}^{(LDA)}(\mathbf{r}) &= \frac{\delta}{\delta\rho(\mathbf{r})}\left(\int d\mathbf{r}' \varepsilon_{XC}^{HEG}\left(\rho(\mathbf{r}')\right)\rho(\mathbf{r}')\right) \\
&= \varepsilon_{XC}^{HEG}(\mathbf{r}) + \left.\frac{d\varepsilon_{XC}^{HEG}}{d\rho}\right|_{\rho=\rho(\mathbf{r})}\rho(\mathbf{r}). \qquad (1.66)
\end{aligned}$$

Forms for $d\varepsilon_{XC}^{HEG}/d\rho$ were derived by Perdew et. al. that correspond to the forms for $\varepsilon_{XC}^{HEG}(\rho)$ [21, 22].

1.2.5 Minimisation within the Kohn-Sham Method

The second part of the Hohenberg-Kohn theorem tells us that, for a given external potential, we can obtain the ground state energy and density by minimising the variational energy of Equation (1.29). In principle, if we had a practical means of directly evaluating $E_{var}[v_{ext}, \rho]$ and its gradient with respect to the density, this would be a matter of a simple search. In practice, however, this is not the case, mainly due to the fact that, even if we have a simple approximation for $E_{XC}[\rho]$, the non-interacting kinetic energy, $T_S[\rho]$, is not a *direct* functional of the density. There are several ways in which one can minimise the energy in a practical manner. The simplest way is just to perform an iterative cycle, which we will now describe.

We have a direct relationship between the ground state density, $\rho(\mathbf{r})$, and the Kohn-Sham potential, $\mu_{KS}(\mathbf{r})$, as given by Equation (1.61). Because this relationship rests on the total energy being stationary with respect to the density, it only holds true for the ground state density. However it still allows us to associate some local potential, $\mu_{loc}(\mathbf{r})$ with *any* given density, $\rho(\mathbf{r})$, and, for simple exchange-correlation functionals at least, evaluate that potential directly from the density in a practical manner. We can represent this procedure as follows:

$$\rho(\mathbf{r}) \rightarrow \mu_{loc}(\mathbf{r}). \tag{1.67}$$

As well as this, for any given Kohn-Sham potential, we can, in principle, solve Equation (1.32) to obtain an associated set of orbitals, $\{\phi_i(\mathbf{r})\}$. From these orbitals, we can then evaluate the associated density via Equation (1.36). Hence, assuming we can solve Equation (1.32), we also have a practical means of evaluating an associated density from a given local potential, which we can represent as follows:

$$\mu_{loc}(\mathbf{r}) \rightarrow \rho(\mathbf{r}). \tag{1.68}$$

With the two procedures above, we could take a density, use it to evaluate a potential, and then use that potential to re-evaluate a density. If this re-evaluated

density is equal to the original density, then it must be the ground state density of the system, because this is only the density for which the potential generated is the correct Kohn-Sham potential. This property of the ground state density is called *self-consistency.* We can cycle around an iterative loop, in which an initial "trial density" is iteratively processed via (1.67) and (1.68) until it becomes self-consistent. Performing such a calculation requires us to be able to solve Equation (1.32) for a given local potential. This potential will have been generated from a density via Equation (1.61), so the set of orbitals, $\{\phi_i(\mathbf{r})\}$, are the solutions of

$$-\frac{1}{2}\nabla^2\phi_i(\mathbf{r}) + v_{ext}(\mathbf{r})\phi_i(\mathbf{r}) + v_H[\rho](\mathbf{r})\phi_i(\mathbf{r}) + \mu_{XC}[\rho](\mathbf{r})\phi_i(\mathbf{r}) = \varepsilon_i\phi_i(\mathbf{r}), \qquad (1.69)$$

which are known as the *Kohn-Sham equations.* In minimising the electronic energy, we are seeking the self-consistent solution of these equations. The simple iterative procedure just described is one way of doing this. More efficient procedures exist that involve minimising the energy directly with respect to the orbitals, under the constraint of orthonormalisation. This requires us to be able to calculate the gradient of the energy with respect to the orbitals, i.e.

$$\begin{aligned}
\frac{\delta E}{\delta\phi^*_{\mathbf{ik}}(\mathbf{r})} &= \frac{\delta(E-T_S)}{\delta\phi^*_{\mathbf{ik}}(\mathbf{r})} + \frac{\delta T_S}{\delta\phi^*_{\mathbf{ik}}(\mathbf{r})} \\
&= \int d\mathbf{r}' \frac{\delta(E-T_S)}{\delta\rho(\mathbf{r}')}\frac{\delta\rho(\mathbf{r}')}{\delta\phi^*_{\mathbf{ik}}(\mathbf{r})} + \frac{\delta T_S}{\delta\phi^*_{\mathbf{ik}}(\mathbf{r})} \\
&= \int d\mathbf{r}' \mu_{KS}(\mathbf{r}')\frac{\delta\rho(\mathbf{r}')}{\delta\phi^*_{\mathbf{ik}}(\mathbf{r})} + \frac{\delta T_S}{\delta\phi^*_{\mathbf{ik}}(\mathbf{r})} \\
&= \int d\mathbf{r}' \mu_{KS}(\mathbf{r}')\frac{\delta}{\delta\phi^*_{\mathbf{ik}}(\mathbf{r})}\sum_{i'\mathbf{k}'}\phi^*_{i'\mathbf{k}'}(\mathbf{r}')\phi_{i'\mathbf{k}'}(\mathbf{r}') \\
&\quad + \frac{\delta}{\delta\phi^*_{\mathbf{ik}}(\mathbf{r})}\sum_{i'\mathbf{k}'}\int d\mathbf{r}'\phi^*_{i'\mathbf{k}'}(\mathbf{r}')\left(\frac{\nabla^2}{2}\right)\phi_{i'\mathbf{k}'}(\mathbf{r}') \\
&= 2\mu_{KS}(\mathbf{r})\phi_{\mathbf{ik}}(\mathbf{r}) + \nabla^2\phi_{\mathbf{ik}}(\mathbf{r}), \qquad (1.70)
\end{aligned}$$

where the factor of 2 in the last line arises from the implicit factor of 2 in the summations (see Appendix C). We will discuss the minimisation procedure in greater detail in the context of the reciprocal space representation.

1.2.6 The Exchange-Correlation Hole

Within the Kohn-Sham framework, the exchange-correlation energy may be expressed as

$$E_{XC}[\rho] = F[\rho] - (T_S[\rho] + V_H[\rho]). \tag{1.71}$$

So we can view E_{XC} as the difference between the true internal electronic energy, F, and an approximation to it in which the kinetic energy is that of a non-interacting system, and the potential energy is that of a non-quantised distribution of charge. A deeper understanding of the meaning of exchange and correlation can be achieved by examining in more detail how the true system differs from this approximate picture.

Firstly, let us separate E_{XC} into its various components, and then re-group them as follows:

$$\begin{aligned} E_{XC} &= V_X + E_C \\ &= V_X + T_C + V_C \\ &= V_{XC} + T_C. \end{aligned} \tag{1.72}$$

So we have now defined the *exchange-correlation potential energy*, V_{XC}, which can also be expressed as

$$V_{XC} = V_{int} - V_H. \tag{1.73}$$

We already know that V_{int} can be obtained by taking the expectation value of $\hat{V}_{int}$ for the many-electron wavefunction, and that V_H can be obtained by essentially taking a double integral of the density as in Equation (1.42). With reference to that equation, the Hartree energy can be expressed as follows:

$$V_H = \int d\mathbf{r} v_H(\mathbf{r})\rho(\mathbf{r}), \tag{1.74}$$

where $v_H(\mathbf{r})$ is the Hartree energy per electron, given by

$$v_H(\mathbf{r}) = \frac{1}{2}\int d\mathbf{r}' \frac{\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}. \tag{1.75}$$

This is also equal to half the Hartree potential, i.e. $v_H(\mathbf{r}) = \frac{1}{2}\mu_H(\mathbf{r})$. If an electron is found at position $\mathbf{r}$, within a non-quantised charge distribution, it can be considered to have a potential energy of $v_H(\mathbf{r})$ due to its interaction with that charge

distribution. This is consistent with the Hartree energy being equal to the integral of $v_H(\mathbf{r})$ times the density.

We can look to define a similar quantity associated with the true internal potential energy, V_{int}, such that

$$V_{int} = \int d\mathbf{r} v_{int}(\mathbf{r})\rho(\mathbf{r}), \tag{1.76}$$

where $v_{int}(\mathbf{r})$ is the *internal potential energy per electron.* Again, this can be considered to arise from the interaction of an electron at $\mathbf{r}$ with a distribution of charge, but now matters are complicated because, in the real system, an electron at $\mathbf{r}$ will "see" a distribution of charge that is affected by the fact that this electron is at $\mathbf{r}$. That is, for every position in space, $\mathbf{r}$, there is a *conditional density,* $\rho(\mathbf{r}'|\mathbf{r})$, that will be seen by an electron at $\mathbf{r}$. The internal potential energy per electron is therefore given by

$$v_{int}(\mathbf{r}) = \frac{1}{2}\int d\mathbf{r}' \frac{\rho(\mathbf{r}'|\mathbf{r})}{|\mathbf{r}-\mathbf{r}'|}. \tag{1.77}$$

The conditional density is related to the electron pair density, $\rho(\mathbf{r},\mathbf{r}')$, which was defined in Section 1.1; we have

$$\rho(\mathbf{r}'|\mathbf{r}) = \frac{\rho(\mathbf{r},\mathbf{r}')}{\rho(\mathbf{r})}. \tag{1.78}$$

Now, since V_{XC} is simply the difference between V_{int} and V_H we can define the *exchange-correlation potential energy per electron,* $v_{XC}(\mathbf{r})$, as follows:

$$v_{XC}(\mathbf{r}) = v_{int}(\mathbf{r}) - v_H(\mathbf{r}). \tag{1.79}$$

Substituting the above equations for $v_{int}(\mathbf{r})$ and $v_H(\mathbf{r})$, we have

$$\begin{aligned} v_{XC}(\mathbf{r}) &= \frac{1}{2}\int d\mathbf{r}' \frac{\rho(\mathbf{r}'|\mathbf{r})}{|\mathbf{r}-\mathbf{r}'|} - \frac{1}{2}\int d\mathbf{r}' \frac{\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} \\ &= \frac{1}{2}\int d\mathbf{r}' \frac{\rho(\mathbf{r}'|\mathbf{r}) - \rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}. \end{aligned} \tag{1.80}$$

Examining this equation we see that this has a similar integral form to the equations for $v_{int}(\mathbf{r})$ and $v_H(\mathbf{r})$, where the numerator of the integrand is now the *exchange-correlation hole,* $h_{XC}(\mathbf{r}'|\mathbf{r})$, given by

$$h_{XC}(\mathbf{r}'|\mathbf{r}) = \rho(\mathbf{r}'|\mathbf{r}) - \rho(\mathbf{r}'). \tag{1.81}$$

This is the change in the electron density caused by the presence of an electron at $\mathbf{r}$. Every electron in the system is effectively surrounded by such an exchange-correlation hole, which results in a reduction in its potential energy.

The exchange-correlation hole can be considered to be the sum of an *exchange hole*, $h_X(\mathbf{r}'|\mathbf{r})$, and a *correlation hole*, $h_C(\mathbf{r}'|\mathbf{r})$, which are components of $h_{XC}(\mathbf{r}'|\mathbf{r})$ that can be ascribed to exchange and correlation respectively. The exchange hole can be defined purely in terms of the non-interacting system, i.e.

$$h_X(\mathbf{r}'|\mathbf{r}) = \rho_S(\mathbf{r}'|\mathbf{r}) - \rho(\mathbf{r}'), \tag{1.82}$$

where $\rho_S(\mathbf{r}'|\mathbf{r})$ is the conditional density of the non-interacting system, given by

$$\rho_S(\mathbf{r}'|\mathbf{r}) = \rho(\mathbf{r}') - \frac{|\sum_i \phi_i^*(\mathbf{r})\phi_i(\mathbf{r}')|^2}{\rho(\mathbf{r})}, \tag{1.83}$$

(see Appendix D), and the correlation hole is then simply given by

$$h_C(\mathbf{r}'|\mathbf{r}) = h_{XC}(\mathbf{r}'|\mathbf{r}) - h_X(\mathbf{r}'|\mathbf{r}). \tag{1.84}$$

We can also define the *exchange energy per electron*, and the *correlation potential energy per electron*, respectively as

$$v_X(\mathbf{r}) = \frac{1}{2}\int d\mathbf{r}' \frac{h_X(\mathbf{r}'|\mathbf{r})}{|\mathbf{r}-\mathbf{r}'|}, \tag{1.85}$$

and

$$v_C(\mathbf{r}) = \frac{1}{2}\int d\mathbf{r}' \frac{h_C(\mathbf{r}'|\mathbf{r})}{|\mathbf{r}-\mathbf{r}'|}. \tag{1.86}$$

Up to this point the non-interacting system and the true, interacting, many-electron system have been thought of as being entirely separate objects, connected only by the fact that they have the same density. But, it is actually possible to view both systems simply as particular instances of a continuous set of systems, each defined by a single parameter, λ, which is the electron-electron *coupling constant*. Each system is described by a many-electron wavefunction, Ψ_λ, which is the ground state of the following many-electron Schrödinger equation:

$$\hat{T}\Psi + \hat{V}_\lambda[\rho]\Psi = -i\frac{d\Psi}{dt}, \tag{1.87}$$

where $\hat{V}_\lambda[\rho]$ is a modified version of the standard potential operator, defined such that the electron-electron repulsive interaction is proportional to λ, so that we have

$$\hat{V}_\lambda[\rho]\Psi_\lambda = \left(\frac{1}{2}\sum_{n\neq m}\frac{\lambda}{|\mathbf{r}_n - \mathbf{r}_m|} + \sum_n \mu_\lambda[\rho](\mathbf{r}_n)\right)\Psi_\lambda, \tag{1.88}$$

and where $\mu_\lambda[\rho](\mathbf{r})$ is a fictitious external potential defined such that the ground state density is equal to $\rho(\mathbf{r})$ for all values of λ. Of particular interest are those values of λ in the range $0 \leq \lambda \leq 1$ as these generate a continuous range of systems going from the non-interacting system ($\lambda = 0$) to the true interacting system ($\lambda = 1$). This continuous path between the two systems is called the *adiabatic connection.*

The adiabatic connection allows us to relate the correlation kinetic energy, T_C, to the correlation hole, $h_C(\mathbf{r}'|\mathbf{r})$. Because the ground state wavefunction minimises the total electronic energy with respect to all degrees of freedom that are not constrained (by normalisation, exchange anti-symmetry, etc.), a small change in any wavefunction-related quantity, at fixed λ, can produce no first order change in the energy, so long as it preserves the constraints. The correlation hole is directly related to the wavefunction, and hence a small change in the correlation hole will not change the total energy. Consider a small change, $\delta h_C(\mathbf{r}'|\mathbf{r})$, in the correlation hole that *does not change* the density, and that is made without changing the value of λ. The resulting change in the total energy is zero, and, because we have required that the density remains fixed, there is no change in the external, Hartree, non-interacting kinetic, or exchange energy. We thus have

$$\begin{aligned} d(V_C + T_C) &= 0 \\ \Rightarrow dT_C &= -dV_C. \end{aligned} \tag{1.89}$$

This immediately provides a link between the kinetic and potential parts of the correlation energy. We can now define the *correlation kinetic energy per electron,* $t_C(\mathbf{r})$, such that

$$\delta t_C(\mathbf{r}) = -\delta v_C(\mathbf{r}), \tag{1.90}$$

for small changes in the correlation hole, $\delta h_C(\mathbf{r}'|\mathbf{r})$, at fixed λ, that do not change the total density, $\rho(\mathbf{r})$. Now consider a small step, $d\lambda$, along the adiabatic connection.

This results in a change in the correlation potential energy per electron that has two components - a component due to the increased numerator in the Coulomb potential, and a component due to the change in shape of the correlation hole, i.e.

$$\delta v_C(\mathbf{r}) = \frac{d\lambda}{2}\int d\mathbf{r}' \frac{h_C^{(\lambda)}(\mathbf{r}'|\mathbf{r})}{|\mathbf{r}-\mathbf{r}'|} + \frac{\lambda}{2}\int d\mathbf{r}' \frac{\delta h_C^{(\lambda)}(\mathbf{r}'|\mathbf{r})}{|\mathbf{r}-\mathbf{r}'|}. \tag{1.91}$$

The second term is a change in the correlation potential energy per electron *at fixed* λ, and hence we can use Equation (1.90) to write

$$\begin{aligned} \delta v_C(\mathbf{r}) &= \frac{d\lambda}{2}\int d\mathbf{r}' \frac{h_C^{(\lambda)}(\mathbf{r}'|\mathbf{r})}{|\mathbf{r}-\mathbf{r}'|} - \delta t_C(\mathbf{r}), \\ \Rightarrow \delta\varepsilon_C(\mathbf{r}) &= \frac{d\lambda}{2}\int d\mathbf{r}' \frac{h_C^{(\lambda)}(\mathbf{r}'|\mathbf{r})}{|\mathbf{r}-\mathbf{r}'|}, \end{aligned} \tag{1.92}$$

where $\varepsilon_C(\mathbf{r})$ is the total correlation energy per electron. Integrating this between the limits of $\lambda = 0$ and $\lambda = 1$ yields

$$\varepsilon_C(\mathbf{r}) = \int_{\lambda=0}^{\lambda=1} d\lambda \frac{1}{2}\int d\mathbf{r}' \frac{h_C^{(\lambda)}(\mathbf{r}'|\mathbf{r})}{|\mathbf{r}-\mathbf{r}'|} \tag{1.93}$$

$$= \frac{1}{2}\int d\mathbf{r}' \frac{\bar{h}_C(\mathbf{r}'|\mathbf{r})}{|\mathbf{r}-\mathbf{r}'|}, \tag{1.94}$$

where $\bar{h}_C(\mathbf{r}'|\mathbf{r})$ is the *coupling constant averaged correlation hole.* This is essentially a modified version of the correlation hole that yields the total correlation energy per electron, rather than just the potential part. Also, since the exchange hole does not depend on λ, we can also define the *coupling constant averaged exchange-correlation hole*, $\bar{h}_{XC}(\mathbf{r}'|\mathbf{r})$, as

$$\bar{h}_{XC}(\mathbf{r}'|\mathbf{r}) = \int_{\lambda=0}^{\lambda=1} d\lambda h_{XC}^{(\lambda)}(\mathbf{r}'|\mathbf{r}), \tag{1.95}$$

which yields the total exchange-correlation energy per electron via

$$\varepsilon_{XC}(\mathbf{r}) = \frac{1}{2}\int d\mathbf{r}' \frac{\bar{h}_{XC}(\mathbf{r}'|\mathbf{r})}{|\mathbf{r}-\mathbf{r}'|}. \tag{1.96}$$

This coupling-constant averaged exchange-correlation hole, while physically less meaningful, is often more useful when examining exchange and correlation effects, and when defining new exchange-correlation functionals. The concept of the exchange-correlation hole will be revisited later, particularly in our discussion of directions for future research in Chapter 7.

1.3 The Plane Wave Pseudopotential Approach

In order to use DFT for practical calculations on real systems, we need to solve the Kohn-Sham equations numerically with a computer, which means that the problem must be cast in a finite manner. Furthermore, it is advantageous to cast the problem in a way that is computationally efficient, and that allows the numerical accuracy to be controlled in a sensible way.

In all of the calculations in this work we will use the *plane wave pseudopotential* approach to solving the Kohn-Sham equations. This involves using a *plane wave basis set* to represent the orbitals, and *pseudopotentials* to represent the nuclei and core electrons. In this section, we will describe this plane wave pseudopotential approach.

Alternative approaches to the plane wave pseudopotential exist. These involve using basis functions that are localised around individual atoms [28, 29, 30]. While cheaper computationally, they suffer from the problem that the basis set is incomplete and so it is often unclear whether or not a given calculation is truly converged with respect to the basis.

1.3.1 Periodic Boundary Conditions

A plane wave basis set must be used in conjunction with *periodic boundary conditions*. This requires that the external potential, and hence the ground state density, be periodic in space, i.e.

$$v_{ext}(\mathbf{r} + \mathbf{R}) = v_{ext}(\mathbf{r}), \tag{1.97}$$

and

$$\rho(\mathbf{r} + \mathbf{R}) = \rho(\mathbf{r}), \tag{1.98}$$

where $\mathbf{R}$ is any *real lattice vector*, defined by

$$\mathbf{R} = l\mathbf{a} + m\mathbf{b} + n\mathbf{c}, \tag{1.99}$$

where l, m, and n can each take any integer value, and $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$ are vectors defining 3 edges of the parallelepiped that forms the *unit cell*. In a similar way to

the real lattice vectors, the *reciprocal lattice vectors* are defined by

$$\mathbf{G} = l\mathbf{a}^* + m\mathbf{b}^* + n\mathbf{c}^*, \tag{1.100}$$

where the vectors $\mathbf{a}^*$, $\mathbf{b}^*$, and $\mathbf{c}^*$ are related to $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$ by

$$\mathbf{a}^* = 2\pi\frac{\mathbf{b}\times\mathbf{c}}{\Omega}, \tag{1.101}$$

$$\mathbf{b}^* = 2\pi\frac{\mathbf{c}\times\mathbf{a}}{\Omega}, \tag{1.102}$$

$$\mathbf{c}^* = 2\pi\frac{\mathbf{a}\times\mathbf{b}}{\Omega}, \tag{1.103}$$

and Ω is the volume of the cell, given by

$$\Omega = |\mathbf{a}\cdot(\mathbf{b}\times\mathbf{c})|. \tag{1.104}$$

1.3.2 Removing Infinities with Periodic Boundary Conditions

The use of periodic boundaries appears to present a number of problems associated with the now infinite size of the system. For example, we have previously said that the total number of electrons, N, is equal to the integral of the density, $\rho(\mathbf{r})$, over all space, i.e.

$$N = \int d\mathbf{r}\rho(\mathbf{r}). \tag{1.105}$$

Clearly, this would integrate to infinity under periodic boundary conditions. However, we can avoid such problems by adopting the following two conventions:

1. Any integral in which the integrand is cell-periodic is taken over the extent of one unit cell only.

2. Any integral in which the integrand is not cell periodic is taken over all space.

In the case of the Hartree energy, we have a further problem because while one of the integrals has a periodic integrand, the other does not due to the $1/|\mathbf{r}-\mathbf{r}'|$ Coulomb factor. The integral of this term over all space diverges, which would still

lead to the Hartree energy being infinite. If we consider the electronic system alone, each electron is effectively interacting with an infinitely large distribution of negative charge, leading to an infinite energy. However, we also know that there is an equal amount of positive charge per unit cell due to the atomic nuclei, making the system charge neutral on average. Hence the average negative charge of the electrons is cancelled by the average positive charge of the nuclei - we only need to consider differences in the local charge density relative to the average. When using periodic boundary conditions, therefore, the correct equation for the Hartree energy is

$$V_H = \frac{1}{2}\int d\mathbf{r}\int d\mathbf{r}'\frac{\rho(\mathbf{r})(\rho(\mathbf{r}')-\langle\rho\rangle)}{|\mathbf{r}-\mathbf{r}'|}, \tag{1.106}$$

where $\langle\rho\rangle$ is the average density of the system. This term is usually not explicitly written down, but should always be taken to be present.

A similar problem arises when evaluating the external potential from the nuclear charges and when evaluating the contribution of nuclear-nuclear repulsive interaction to the total energy. Again, only differences from the average charge density need to be considered. The means by which the nuclear-nuclear repulsive interaction V_{I-I} is dealt with is described in reference [2].

1.3.3 Kohn-Sham Orbitals with Periodic Boundary Conditions

While the potential and ground state density of a periodic system satisfy periodic boundary conditions, the Kohn-Sham orbitals do not necessarily have to. However, the contribution to the density from a given orbital, given by $\phi(\mathbf{r})^*\phi(\mathbf{r})$, does have to satisfy periodic boundary conditions and so the orbital must have some degree of periodicity. In order to produce the periodic density contribution, its magnitude, $|\phi(\mathbf{r})|$, must be periodic. Further, in order to satisfy the Kohn-Sham equations for a periodic Kohn-Sham potential, we must have

$$\phi(\mathbf{r}+\mathbf{R}) = \phi(\mathbf{r})e^{i\mathbf{k}\cdot\mathbf{R}}, \tag{1.107}$$

where $\mathbf{R}$ is any real lattice vector and $\mathbf{k}$ is a vector that is constant for a given orbital. This leads to *Bloch's theorem* for Kohn-Sham orbitals [31]

$$\phi(\mathbf{r}) = u(\mathbf{r})e^{i\mathbf{k}\cdot\mathbf{r}}, \tag{1.108}$$

where u is a periodic function satisfying

$$u(\mathbf{r}+\mathbf{R}) = u(\mathbf{r}). \tag{1.109}$$

Now, the Bloch wave vector, $\mathbf{k}$, is not uniquely defined for a particular orbital because we can always make the transformation

$$u(\mathbf{r}) \rightarrow u(\mathbf{r})e^{-i\mathbf{G}\cdot\mathbf{r}}, \tag{1.110}$$

and

$$\mathbf{k} \rightarrow \mathbf{k}+\mathbf{G}, \tag{1.111}$$

where $\mathbf{G}$ is any reciprocal lattice vector, without changing the periodicity of $u(\mathbf{r})$, and without changing the value of $\phi(\mathbf{r})$ at any point. However, we can always require that $\mathbf{k}$ lies within the *Brillouin zone*, i.e.

$$\begin{aligned} -\tfrac{1}{2} &< k_l \leq +\tfrac{1}{2}, \\ -\tfrac{1}{2} &< k_m \leq +\tfrac{1}{2}, \\ -\tfrac{1}{2} &< k_n \leq +\tfrac{1}{2}, \end{aligned} \tag{1.112}$$

where

$$\mathbf{k} = k_l\mathbf{a}^* + k_m\mathbf{b}^* + k_n\mathbf{c}^*. \tag{1.113}$$

This ensures that there is only one value of $\mathbf{k}$ for any given orbital, $\phi(\mathbf{r})$.

Now, we define the normalisation condition for a Kohn-Sham orbital such that

$$\int d\mathbf{r}\phi^*(\mathbf{r})\phi(\mathbf{r}) = 1. \tag{1.114}$$

The Bloch wave vector, $\mathbf{k}$, of a given orbital could lie anywhere within the Brillouin zone. For a general system, there should, in principle be a set orbitals for every possible value of $\mathbf{k}$. In practice, we only ever deal with a finite set of $\mathbf{k}$-points,

usually distributed evenly throughout the Brillouin zone according to a Monkhorst-Pack scheme [32] (see 1.3.8). When using periodic boundary conditions, we index the orbitals by $\mathbf{k}$, as well as i. There will be, on average, N orbitals per $\mathbf{k}$-point.

Supposing there are N_k such $\mathbf{k}$-points, then each $\mathbf{k}$-point is given a "weighting" of $1/N_k$. This means that we still have the correct normalisation with a complete set of orbitals for each $\mathbf{k}$-point, so, for example, the density is given by

$$\rho(\mathbf{r}) = \sum_{\mathbf{k}} \sum_{i} \phi^*_{i\mathbf{k}}(\mathbf{r})\phi_{i\mathbf{k}}(\mathbf{r}), \tag{1.115}$$

where there is an implicit prefactor of $1/N_k$ before the summation, as discussed in Appendix C.

Whereas before, the Kohn-Sham orbitals were defined as the N eigenstates of the Kohn-Sham equations of lowest eigenvalue, they must now be defined as the eigenstates of lowest eigenvalue such that the total density integrates to N over 1 cell. There can therefore be different numbers of orbitals on different $\mathbf{k}$-points if this results in the lowest set of eigenvalues. Also, while there must be an integer number of orbitals on each $\mathbf{k}$-point, the average number of electrons per cell, N, can, in principle, be fractional.

1.3.4 Plane Waves and Reciprocal Space

Any continuous periodic orbital, $\phi_{i\mathbf{k}}(\mathbf{r})$, with Bloch wave vector $\mathbf{k}$, may be written as

$$\phi_{i\mathbf{k}}(\mathbf{r}) = \frac{1}{\sqrt{\Omega}} \sum_{\mathbf{G}} c_{i\mathbf{k}}(\mathbf{G}) e^{i(\mathbf{k}+\mathbf{G})\cdot\mathbf{r}}, \tag{1.116}$$

$$= \mathrm{FT}^{-1}[c_{i\mathbf{k}}(\mathbf{G})], \tag{1.117}$$

where the $c_{i\mathbf{k}}(\mathbf{G})$ are a set of complex coefficients that constitute the reciprocal space representation of the orbital, and the $\mathbf{G}$ are reciprocal lattice vectors. The functions $e^{i(\mathbf{k}+\mathbf{G})\cdot\mathbf{r}}$ are plane waves, and are the basis functions in this representation. The pre-factor of $1/\sqrt{\Omega}$ preserves the normalisation of the wavefunction, so we have

$$\int d\mathbf{r} \phi^*_{i\mathbf{k}}(\mathbf{r})\phi_{i\mathbf{k}}(\mathbf{r}) = 1, \tag{1.118}$$

and

$$\sum_{\mathbf{G}} c^*_{i\mathbf{k}}(\mathbf{G}) c_{i\mathbf{k}}(\mathbf{G}) = 1. \tag{1.119}$$

Equation (1.116) is an inverse Fourier transform of an orbital from reciprocal to real space. The corresponding Fourier transform, that transforms from real to reciprocal space, is:

$$c_{i\mathbf{k}}(\mathbf{G}) = \mathrm{FT}[\phi_{i\mathbf{k}}(\mathbf{r})] \tag{1.120}$$

$$= \frac{1}{\sqrt{\Omega}} \int d\mathbf{r} \phi_{i\mathbf{k}}(\mathbf{r}) e^{-i(\mathbf{k}+\mathbf{G})\cdot\mathbf{r}}. \tag{1.121}$$

As well as representing the Kohn-Sham orbitals in reciprocal space, it is also often useful to represent other quantities, such as densities and potentials, in this way also. We define the reciprocal representation of the density, $\rho(\mathbf{G})$, as follows:

$$\rho(\mathbf{r}) = \frac{1}{\Omega} \sum_{\mathbf{G}} \rho(\mathbf{G}) e^{i\mathbf{G}\cdot\mathbf{r}}, \tag{1.122}$$

and

$$\rho(\mathbf{G}) = \int d\mathbf{r} \rho(\mathbf{r}) e^{-i\mathbf{G}\cdot\mathbf{r}}. \tag{1.123}$$

Again, there is a combined prefactor of $1/\Omega$ between the Fourier transform and its inverse. The choice of where this factor is placed is essentially arbitrary, but placing it in this way means that the reciprocal space coefficients, $\rho(\mathbf{G})$, are constant with respect to a charge-conserving scaling of space. This is very convenient when, for example, evaluating contributions to the stress tensor as we will see later.

The Fourier transforms of Equations (1.122) and (1.123) apply to electron densities. Throughout this work the same relations will apply to all scalar fields, such as potentials, unless otherwise stated.

1.3.5 Evaluating Quantities in Reciprocal Space

Within the Kohn-Sham framework, the non-interacting kinetic energy, T_S, the Hartree energy, E_H, and Hartree potential, $\mu_H(\mathbf{r})$, are more easily evaluated in reciprocal space rather than real space. If we write an orbital as a sum of plane

waves and apply the single-particle kinetic energy operator, we have

$$\begin{aligned} -\frac{1}{2}\nabla^2\phi_{i\mathbf{k}}(\mathbf{r}) &= -\frac{1}{2}\nabla^2\left(\frac{1}{\sqrt{\Omega}}\sum_{\mathbf{G}} c_{i\mathbf{k}}(\mathbf{G})e^{i(\mathbf{k}+\mathbf{G})\cdot\mathbf{r}}\right) \\ &= \frac{1}{2\sqrt{\Omega}}\sum_{\mathbf{G}}(\mathbf{k}+\mathbf{G})^2 c_{i\mathbf{k}}(\mathbf{G})e^{i(\mathbf{k}+\mathbf{G})\cdot\mathbf{r}}. \end{aligned} \tag{1.124}$$

The effect of the kinetic energy operator in reciprocal space is thus to multiply each plane wave coefficient by $\frac{1}{2}$ times the square of its wave vector. This leads to a simple expression for the non-interacting kinetic energy:

$$T_S = \frac{1}{2}\sum_{i\mathbf{k}}\sum_{\mathbf{G}} c^*_{i\mathbf{k}}(\mathbf{G})c_{i\mathbf{k}}(\mathbf{G})(\mathbf{k}+\mathbf{G})^2. \tag{1.125}$$

The Hartree potential can be related to the density via Poisson's equation:

$$-\nabla^2\mu_H(\mathbf{r}) = \rho(\mathbf{r}) - \langle\rho\rangle. \tag{1.126}$$

Writing the Hartree potential and the density in their reciprocal space representations this becomes

$$-\nabla^2\left(\frac{1}{\Omega}\sum_{\mathbf{G}\neq\mathbf{0}}\mu_H(\mathbf{G})e^{i\mathbf{G}\cdot\mathbf{r}}\right) = \left(\frac{1}{\Omega}\sum_{\mathbf{G}\neq\mathbf{0}}\rho(\mathbf{G})e^{i\mathbf{G}\cdot\mathbf{r}}\right). \tag{1.127}$$

Note that the subtraction of the average density in Equation (1.126) is accounted for by the exclusion of the $\mathbf{G} = \mathbf{0}$ term in the reciprocal representation. We now apply the ∇^2 operator to $\mu_H(\mathbf{G})$ and equate exponential coefficients as follows:

$$\begin{aligned} \frac{1}{\Omega}\sum_{\mathbf{G}\neq\mathbf{0}}\mu_H(\mathbf{G})\mathbf{G}^2e^{i\mathbf{G}\cdot\mathbf{r}} &= \frac{1}{\Omega}\sum_{\mathbf{G}\neq\mathbf{0}}\rho(\mathbf{G})e^{i\mathbf{G}\cdot\mathbf{r}}, \\ \Rightarrow \mu_H(\mathbf{G})\mathbf{G}^2 &= \rho(\mathbf{G}), \\ \Rightarrow \mu_H(\mathbf{G}) &= \frac{\rho(\mathbf{G})}{\mathbf{G}^2}. \end{aligned} \tag{1.128}$$

This is the expression for the Hartree potential in reciprocal space. It excludes the $\mathbf{G} = \mathbf{0}$ term, which can be set to zero as it exactly cancels the equivalent term from the positive nuclear charge in a neutral system.

The Hartree energy in real space is given by

$$\begin{aligned} V_H &= \frac{1}{2}\int d\mathbf{r}\mu_H(\mathbf{r})\rho(\mathbf{r}) \\ &= \frac{1}{2}\int d\mathbf{r}\mu^*_H(\mathbf{r})\rho(\mathbf{r}), \end{aligned} \tag{1.129}$$

where the conjugate on $\mu_H^*(\mathbf{r})$ has no effect since it is a real field. Writing each term in its reciprocal space representation gives

$$\begin{aligned} V_H &= \frac{1}{2}\int d\mathbf{r}\left(\frac{1}{\Omega}\sum_{\mathbf{G}\neq 0}\mu_H^*(\mathbf{G})e^{-i\mathbf{G}\cdot\mathbf{r}}\right)\left(\frac{1}{\Omega}\sum_{\mathbf{G}'\neq 0}\rho(\mathbf{G}')e^{i\mathbf{G}'\cdot\mathbf{r}}\right) \\ &= \frac{1}{2\Omega^2}\sum_{\mathbf{G},\mathbf{G}'}\mu_H^*(\mathbf{G})\rho(\mathbf{G}')\int d\mathbf{r}e^{i(\mathbf{G}'-\mathbf{G})\cdot\mathbf{r}} \\ &= \frac{1}{2\Omega}\sum_{\mathbf{G}}\mu_H^*(\mathbf{G})\rho(\mathbf{G}). \end{aligned} \tag{1.130}$$

If we then substitute $\mu_H(\mathbf{G})$ for its form as given in Equation (1.128), we obtain

$$V_H = \frac{1}{2\Omega}\sum_{\mathbf{G}}\rho^*(\mathbf{G})\rho(\mathbf{G}), \tag{1.131}$$

which gives the Hartree energy in reciprocal space.

1.3.6 Operators in Reciprocal Space

Any single-particle quantum operator, $\hat{O}$, can be represented in reciprocal space as a matrix, $O(\mathbf{G},\mathbf{G}')$, which acts on an orbital represented by the coefficients, $c_{i\mathbf{k}}(\mathbf{G})$, as follows:

$$\hat{O}c_{i\mathbf{k}}(\mathbf{G}) = \sum_{\mathbf{G}'}O(\mathbf{G},\mathbf{G}')c_{i\mathbf{k}}(\mathbf{G}'). \tag{1.132}$$

We have just seen that the kinetic energy operator, $\hat{T}$, has the effect of multiplying each coefficient by $\frac{1}{2}|\mathbf{k}+\mathbf{G}|^2$, and so its matrix representation is

$$T(\mathbf{G},\mathbf{G}') = \frac{1}{2}|\mathbf{k}+\mathbf{G}|^2\delta_{\mathbf{G},\mathbf{G}'}. \tag{1.133}$$

For a local potential operator, $\mu(\mathbf{r})$, such as the Hartree potential, we require that

$$\begin{aligned} \sum_{\mathbf{G}'}\mu(\mathbf{G},\mathbf{G}')c_{i\mathbf{k}}(\mathbf{G}') &= \mathrm{FT}[\mu(\mathbf{r})\phi_{i\mathbf{k}}(\mathbf{r})] \\ &= \frac{1}{\sqrt{\Omega}}\int d\mathbf{r}\mu(\mathbf{r})\phi_{i\mathbf{k}}(\mathbf{r})e^{-i(\mathbf{k}+\mathbf{G})\cdot\mathbf{r}}. \end{aligned} \tag{1.134}$$

Substituting the reciprocal space representations of $\mu(\mathbf{r})$ and $\phi_{\mathbf{ik}}(\mathbf{r})$, we have

$$\begin{aligned}\sum_{\mathbf{G}'}\mu(\mathbf{G},\mathbf{G}')c_{\mathbf{ik}}(\mathbf{G}') &= \frac{1}{\sqrt{\Omega}}\int d\mathbf{r}\frac{1}{\Omega}\sum_{\mathbf{G}''}\mu(\mathbf{G}'')e^{i\mathbf{G}''\cdot\mathbf{r}} \\ &\quad\times\frac{1}{\sqrt{\Omega}}\sum_{\mathbf{G}'''}c_{\mathbf{ik}}(\mathbf{G}''')e^{i(\mathbf{k}+\mathbf{G}''')\cdot\mathbf{r}}e^{-i(\mathbf{k}+\mathbf{G})\cdot\mathbf{r}} \\ &= \frac{1}{\Omega^2}\sum_{\mathbf{G}''\mathbf{G}'''}\mu(\mathbf{G}'')c_{\mathbf{ik}}(\mathbf{G}''')\int d\mathbf{r}e^{i(\mathbf{G}''+\mathbf{G}'''-\mathbf{G})\cdot\mathbf{r}} \\ &= \frac{1}{\Omega}\sum_{\mathbf{G}'''}\mu(\mathbf{G}-\mathbf{G}''')c_{\mathbf{ik}}(\mathbf{G}'''). \qquad (1.135)\end{aligned}$$

Hence, by comparing terms, we see that the matrix representation of a local potential is

$$\mu(\mathbf{G},\mathbf{G}') = \frac{1}{\Omega}\mu(\mathbf{G}-\mathbf{G}'). \qquad (1.136)$$

1.3.7 Convergence of the Plane Wave Basis Set

The complete set of reciprocal lattice vectors, **G**, is infinite, which means that evaluation of a sum over all such vectors would take infinitely long to compute. Fortunately, in realistic systems, the orbitals and densities tend to become smoothly varying at small scales - meaning that their plane wave components become negligible for large **G**-vectors. We can take advantage of this fact by truncating the set of **G**-vectors so that we exclude those with magnitudes larger than some cut-off radius.

The cut-off radius for Kohn-Sham orbitals, G_{cut}, is usually defined in terms of its corresponding kinetic energy, E_{cut}:

$$G_{cut} = \sqrt{2E_{cut}}. \qquad (1.137)$$

The cut-off radius for densities and potentials should be double that for orbitals.

The cut-off energy that is appropriate for a given calculation is not usually known in advance, as it depends very much on the system in question, and on which quantities one wishes to calculate. However, establishing what the appropriate cut-off energy is

essentially just a matter of increasing it until the result stops changing - a procedure known as a *convergence test.* This feature of a plane wave basis set - i.e. that its accuracy can be controlled by a single parameter - is a major advantage over, for example, localised basis sets [28, 29, 30].

One can usually be quite confident, once a result is converged with respect to the cut-off energy, that there is no unrepresented Hilbert space that would significantly affect that result. It is however, important to bear in mind that quantities related to derivatives of the energy, such as force and stress, may require a slightly higher cut off energy than the energy itself, depending on the desired accuracy.

1.3.8 Monkhorst-Pack Grids

As mentioned previously, in practical calculations, we only ever use a finite number of **k**-points to sample the Brillouin zone. This is in order that the calculation remains finite, and is justified so long as the orbitals vary smoothly with respect to **k**. A Monkhorst-Pack grid [32] is an unbiased method of choosing a set of **k**-points for sampling the Brillouin zone. In fractional coordinates, it is a rectangular grid of points of dimensions $M_x \times M_y \times M_z$, spaced evenly throughout the Brillouin zone. The larger the dimensions of the grid, the finer and more accurate will be the sampling. Much like the cut-off energy, the size of grid required depends on the system under study, but the appropriate size can be established by means of a convergence test.

1.3.9 Introduction to Pseudopotentials

One of the main advantages of using a plane wave basis set is that its accuracy can be easily controlled. This is related to the fact that, when using such a basis set, we are making no assumptions about the final shape of the orbitals, other than that there is some scale below which they become smoothly varying. However, this also leads to a major disadvantage of using a plane wave basis set, which is that the size of the basis set required for a given system is often far larger than would be

required with a localised basis set. This is because, in condensed matter systems, the orbitals tend to oscillate very rapidly in the vicinity of atomic nuclei, and are much more smoothly varying elsewhere. In order to describe this rapid oscillation we must set a very large cut-off energy, so that we include plane waves with very short wavelengths. But, since most of the space in the cell does not contain rapidly oscillating orbitals, most of the computational expense associated with all these plane waves effectively goes to waste. A localised basis set can be tailored such that the basis functions themselves are rapidly oscillating in the vicinity of atomic nuclei and more smoothly oscillating elsewhere, so that the total number of basis functions required for the system is far smaller.

The use of *pseudopotentials* [33, 34, 35], in conjunction with plane waves, can dramatically reduce the magnitude of this problem. To understand what a pseudopotential does, we note the following two facts about orbitals in condensed matter systems:

1. Lower energy orbitals can often be considered represent *core electrons*. These are electrons that are well localised around an atomic nucleus and whose properties do not change significantly with the atom's "chemical environment".

2. Orbitals representing electrons that are not core electrons oscillate very rapidly in the vicinity of atomic nuclei, but most of this oscillation can be put down to the fact that they have to be orthogonal to the core electrons.

A pseudopotential essentially changes part of what the outer, or *valence*, electrons "see". The core electrons, and the potential due to the bare nuclear charge, are replaced by a fictitious potential that is defined such that the behaviour of the valence electrons is not affected outside of some cut-off radius from the nucleus. So long as this radius is not so large that it overlaps regions of space that are involved in chemical bonding, the pseudopotential approximation should not significantly alter the inter-atomic interactions that govern the behaviour of condensed matter.

Using pseudopotentials reduces the computational cost of a calculation in three ways:

1. By effectively removing core electrons from the calculation, the number of Kohn-Sham orbitals is reduced. This reduces the memory required to store the orbitals, the time required to evaluate orbital-dependant quantities, and the time required to orthonormalise a set of orbitals.

2. Because there are no core-electrons to which valence electrons must be orthogonal, there is less oscillation of the corresponding orbitals in the vicinity of the nucleus. This means that a lower cut-off energy can be used to represent the orbitals, resulting in lower memory requirements and greater speed. This lowering of the cut-off energy is typically a few orders of magnitude resulting in massive gains in efficiency.

3. Because the pseudopotential is not uniquely defined for a particular element, we can optimise the shape of the potential so as to give as low a required cut-off energy as possible. Again, this reduces memory and increases speed.

Because we only explicitly treat the valence electrons in a calculation when using pseudopotentials, we tend to think of the system as being made of electrons and ions rather that electrons and nuclei. This, incidentally, is why we used the more general subscript of I rather that N when referring to nuclei in Section 1.1.

1.3.10 Basic Pseudopotential Theory

Consider a single isolated atom, with atomic number Z. There are N $(= Z)$ electrons, moving in an external potential given by

$$v_{ext}(r) = \frac{Z}{r}. \tag{1.138}$$

Applying Kohn-Sham DFT to this system will result in a set of N Kohn-Sham orbitals, $\psi_i(\mathbf{r})$, a corresponding density, $\rho(\mathbf{r})$, and a Kohn-Sham potential, $\mu_{KS}(\mathbf{r})$.

In order to create a pseudopotential for this atom, we must first specify which orbitals are to be considered core, and which are to be considered valence, and also specify the cut-off radius, r_c. In most cases, all the electrons that are in "closed shells" are considered core, while the remainder are considered valence. In general,

the pseudopotential $\hat{v}_{PS}$ is *non-local*, in that there is a separate local potential, $v_{PS}^{(l)}(r)$, acting on each angular momentum component, l, of a given orbital.

If we apply Kohn-Sham DFT to the atom, with the external potential, $v_{ext}(r)$, now replaced with the pseudopotential, $\hat{v}_{PS}$, and with only valence electrons present, the resulting pseudo-orbitals, $\psi_{PS}(\mathbf{r})$, must satisfy the following requirements:

1. Each pseudo-orbital, $\psi_{PS}(\mathbf{r})$, must equal the corresponding orbital from the all-electron calculation, $\psi_{AE}(\mathbf{r})$, for all points $\mathbf{r}$ that lie outside the cut-off radius.

2. The eigenvalue of each pseudo-orbital must equal the eigenvalue of the corresponding all-electron orbital.

3. The first and second derivatives of each $\psi_{PS}(\mathbf{r})$ must equal those of the corresponding $\psi_{AE}(\mathbf{r})$ at the cut-off radius.

4. There must be no radial nodes of the pseudo-orbitals inside the cut-off radius.

Implicit in (1) above is the requirement that the total electronic charge of the valence electrons inside the cut-off radius is equal for both the pseudo- and all-electron orbitals. This is because in standard Kohn-Sham theory, each orbital is normalised to 1. Pseudopotentials in which this condition is respected are referred to as *norm-conserving* pseudopotentials. A class of pseudopotentials, called *ultrasoft* pseudopotentials [36], also exist in which this condition is relaxed, allowing a lower plane wave cut-off energy, but such potentials will not be used in this work. Now, any pseudopotential can be chosen so long as it satisfies the above conditions, and the particular form is usually chosen so as to make the pseudo-orbitals as smooth as possible to minimise the required plane wave cut-off energy. In the basic non-local form described above, in which there is a separate local potential for each angular momentum component, the pseudopotential acts on an orbital as follows:

$$\hat{v}_{PS}\psi(\mathbf{r}) = \sum_{lm} v_{PS}^{(l)}(r) Y_{lm}(\theta,\phi)\langle Y_{lm}|\psi(\mathbf{r})\rangle, \tag{1.139}$$

where the $Y_{lm}(\theta,\phi)$ are spherical harmonics. When implemented within a plane wave basis, applying such a pseudopotential would require a double-loop over **G**-vectors

of the general form

$$\hat{v}_{PS}c_{\mathbf{ik}}(\mathbf{G}) = \sum_{\mathbf{G}'} v_{PS}(\mathbf{G},\mathbf{G}')c_{\mathbf{ik}}(\mathbf{G}'), \tag{1.140}$$

which would lead to unfavourable scaling with system size. This problem can be overcome by using *Kleinman-Bylander* pseudopotentials [37], in which each angular component of the pseudopotential is separated into a local and a non-local component as follows:

$$v_{PS}^{(l)}(r) = v_{PS}^{LOC}(r) + \delta v_{PS}^{(l)}(r). \tag{1.141}$$

The non-local component, $\delta v_{PS}^{(l)}$, is then approximated as

$$\delta v_{PS}^{(l)}(r) = \frac{|\delta v_{PS}^{(l)}(r)\psi_l^0(r)\rangle\langle\psi_l^0(r)\delta v_{PS}^{(l)}(r)|}{\langle\psi_l^0(r)|\delta v_{PS}^{(l)}(r)\psi_l^0(r)\rangle}, \tag{1.142}$$

where the $\psi_l^0(r)$ are the pseudo-orbitals for the atomic system. Constructing the pseudopotentials in this way reduces the computational costs so this part of the calculation scales linearly with the number of plane waves.

1.3.11 Disadvantages of Pseudopotentials

Although providing enormous benefits in terms of speed and memory, the use of pseudopotentials is not without its drawbacks.

The most serious drawback of using pseudopotentials is that we are almost invariably drawn away from the safe ground of purely first principles calculations. This is because while, in theory, the shape of the pseudopotential should not affect the chemical behaviour of a system, this is only actually the case if the cut-off radius for the potential is sufficiently small. Often, larger cut-off radii are chosen, as this allows us to use a lower plane wave cut-off energy. The shape of the pseudopotential is then chosen so as to still give "good" results. The problem is that what is often meant by "good results" is actually "results that are in good agreement with experiment". As soon as one makes a choice of pseudopotential that is in any way based on experimental results then the calculation cannot be described as being purely first principles.

Within the bounds of first principles calculations, results may be considered to be "good" if they are in close agreement with all-electron calculations, i.e. calculations

not involving pseudopotentials. However, even pseudopotentials that are optimised in this way tend only to perform well when the atoms are in a similar "chemical environment" to the one for which they were tailored. The "transferability" of a pseudopotential, i.e. how well it performs in differing chemical environments, can only be reliably improved by reducing its cut-off radius. It should also be noted that ultrasoft pseudopotentials tend to be more transferable that norm-conserving ones [36].

There are many other issues related to pseudopotentials, which may or may not be of importance depending on the system under study, and on which properties one wishes to calculate. For example, deciding which electrons to treat as 'core' and which to treat as 'valence', or whether or not relativistic effects should be included. In the end, there is such a multitude of adjustable parameters and degrees of freedom available in the generation of pseudopotentials that even the detailed study of one pseudopotential for one individual element could be a work in its own right. For this reason, we are often forced simply to accept that the use of pseudopotentials may incur a loss of accuracy of a similar order of magnitude to that incurred, for example, by using the LDA to treat exchange and correlation, and that further, the results of calculations in which different pseudopotentials have been used should not necessarily be expected to agree with each other. The gains in efficiency that pseudopotentials afford simply make them a necessary evil.

1.4 Solving the Kohn-Sham Equations

Using the matrix representations of operators in reciprocal space, the Kohn-Sham equations (1.69) can be written as

$$\sum_{\mathbf{G}'} \left(\frac{1}{2}(\mathbf{k}+\mathbf{G})^2 \delta_{\mathbf{G}\mathbf{G}'} + v_{ext}(\mathbf{G}-\mathbf{G}') + \mu_H(\mathbf{G}-\mathbf{G}') + \mu_{XC}(\mathbf{G}-\mathbf{G}') \right) c_{i\mathbf{k}}(\mathbf{G}') = \varepsilon_{i\mathbf{k}} c_{i\mathbf{k}}(\mathbf{G}). \quad (1.143)$$

With a finite basis set, this is a numerically solvable matrix eigenproblem, with the bracketed terms forming the matrix elements, and the plane wave coefficients

forming the vector elements. However, due to the very large number of plane waves we are dealing with ($\sim 10^5$ in a typical calculation), direct matrix diagonalisation is very expensive. It is also very wasteful, because we end up with as many eigenstates as there are plane waves, when we only actually require the lowest $\sim N/2$ eigenstates. In general, the solution obtained will not be self-consistent, and we would have to perform several iterative cycles in order to minimise the energy.

An alternative approach to the simple iterative loop combined with matrix diagonalisation is to minimise the energy *directly* with respect to the Kohn-Sham orbitals themselves. To do this efficiently requires us to be able to calculate the *gradient* of the energy with respect to the orbitals, discussed in Section 1.2.5. In reciprocal space, Equation (1.70) becomes

$$\frac{\delta E}{\delta c^*_{\mathbf{ik}}(\mathbf{G})} =$$

$$2\sum_{\mathbf{G}'}\left(\frac{1}{2}(\mathbf{k}+\mathbf{G})^2\delta_{\mathbf{GG}'} + v_{ext}(\mathbf{G}-\mathbf{G}') + \mu_H(\mathbf{G}-\mathbf{G}') + \mu_{XC}(\mathbf{G}-\mathbf{G}')\right)c_{\mathbf{ik}}(\mathbf{G}'), \tag{1.144}$$

i.e. we can obtain the gradient by applying the Kohn-Sham Hamiltonian to the orbitals. This allows us to use methods such as *steepest descents* or *conjugate gradients* [38] to minimise the energy and obtain a self-consistent solution of the Kohn-Sham equations.

1.5 The Kohn-Sham Band Structure

Supposing we have a self-consistent solution of the Kohn-Sham equations, then we have $\sim N/2$ orbitals for each of the N_k **k**-points used to sample the Brillouin zone. These orbitals are solutions of a single-particle Schrödinger-like equation in which the local potential is the Kohn-Sham potential, $\mu_{KS}(\mathbf{r})$. Now, having obtained a self-consistent solution, we can also solve this equation for **k**-points other than those in the original set, and look for solutions other than just the lowest $\sim N/2$ eigenstates. The complete set of eigenvalues for each **k**-point in the Brillouin zone forms the *Kohn-Sham band structure.* This band structure is often assumed to approximate the true band structure of the interacting system. However, there is

no reason to believe that it would exactly equal the true band structure, even if we had an exact exchange-correlation functional. How closely the Kohn-Sham band structure approximates the true band structure, and how this depends on the choice of functional, will be the subject of discussion in later chapters.

1.6 Extensions of Kohn-Sham Theory

We have outlined the essential aspects of the standard Kohn-Sham formalism, however a number of extensions of this formalism exist that improve the description of some systems. The two most common extensions are fractional occupancies and spin-dependent DFT, which we will now briefly describe.

1.6.1 Fractional Occupancies

In standard Kohn-Sham theory the orbitals are essentially defined as the lowest eigenstates of the self-consistent Kohn-Sham Hamiltonian. Higher eigenstates may be calculated as part of a band structure calculation, but are otherwise not involved in the calculation of ground state properties. However, it is possible to re-define the set of orbitals to include *all* eigenstates of the Hamiltonian, with each orbital assigned an *occupancy*, $f_{\mathbf{ik}}$, of between 0 and 1. Standard Kohn-Sham theory is then the special case in which original, lowest, orbitals have an occupancy of 1 and all higher orbitals have an occupancy of 0. When evaluating quantities in terms of a sum over orbital functionals, each term is now weighted by the occupancy of the orbital, e.g. the density is now given by

$$\rho(\mathbf{r}) = \sum_{\mathbf{ik}} f_{\mathbf{ik}} \phi^*_{\mathbf{ik}}(\mathbf{r}) \phi_{\mathbf{ik}}(\mathbf{r}). \tag{1.145}$$

1.6.2 Spin-Dependent DFT

While an exact DFT calculation is guaranteed to return the correct ground state and energy of any system, it would not necessarily tell us anything about the spin of

the electrons. In standard Kohn-Sham theory, because there is no spin component in the Kohn-Sham potential, the numbers of spin-up and spin-down orbitals never differ by more than one. Therefore, even if the system is highly spin-polarised, the Kohn-Sham orbitals do not reflect this fact. Spin-dependent DFT is a generalisation of the standard Kohn-Sham formalism that allows us to deal more sensibly with systems that have collinear spin polarisation. This involves adding a spin-index to both the density and the Kohn-Sham potential, one for spin-up electrons and one for spin-down electrons, i.e.

$$\rho(\mathbf{r}) \rightarrow \rho(\mathbf{r}, \sigma), \tag{1.146}$$

and,

$$\mu_{KS}(\mathbf{r}) \rightarrow \mu_{KS}(\mathbf{r}, \sigma), \tag{1.147}$$

with the density now related to the Kohn-Sham orbitals via

$$\rho(\mathbf{r}, \sigma) = \sum_{i\mathbf{k}} f_{i\mathbf{k},\sigma} \phi^*_{i\mathbf{k}}(\mathbf{r}) \phi_{i\mathbf{k}}(\mathbf{r}). \tag{1.148}$$

The spin polarisation of a system emerges via the exchange-correlation functional, for example, the LDA can be extended to the *local spin density approximation* (LSDA) [15].

1.7 The CASTEP Code

Much of this work has involved the computational implementation of non-local exchange-correlation functionals. This implementation has been built into the existing CASTEP code [3, 39], which is one of the leading codes in the field of electronic structure calculations. All DFT calculations in this work were performed using CASTEP. In order to describe the implementation of the non-local functionals in later chapters, it is necessary first to explain some of the features of the basic CASTEP implementation.

1.7.1 Algorithms and Tools

A range of computational tools are available that aid in the implementation of different functionalities. They include lists of physical constants, basic numerical algorithms, and file-handling capabilities for the input and output of data. Of particular importance for computationally intensive work are fast Fourier transforms (FFTs). Implementations of the FFT algorithm [38] are in place for transforming data between real space and reciprocal space representations. This is one of the most commonly applied, and most computationally intensive operations that is used in these calculations and as such, the implementation of the algorithm has been designed to be as efficient as possible. Tools are also available for performing efficient calculations in parallel on several processors; this is discussed further in 1.7.6.

1.7.2 Elements of a DFT Calculation

Numerical representations are defined of physical objects that are used in DFT calculations, i.e. the unit cell, potentials, wavefunctions, and electron densities. Also defined numerically are the plane wave basis set, and the reciprocal and real space grids on which data associated with potentials, wavefunctions, and densities is represented.

Data associated with the unit cell includes the real- and reciprocal-space lattice vectors, the positions of the atoms within the cell and the set of **k**-points used for Brillouin zone integration.

Most data associated with the plane wave basis set involves defining mappings between the logical indices, used internally by the computer, and the physical coordinates that have meaning in terms of the real system that is being modelled. For example, if a field is defined on the real space grid, it will be stored in a simple 1-dimensional array and a mapping array is defined that translates the indices of that array into coordinates in real space.

The three basic types of field that we deal with in a calculation are densities, potentials, and orbitals. Common operations involving these objects include applying

a potential to a set of orbitals, calculating a density from a set of orbitals, and calculating a potential from a given density. For efficient implementation of such operations data is often Fourier transformed between real and reciprocal space.

1.7.3 Functionality

The primary functionality of any DFT implementation is the ability to minimise the energy in order to obtain the electronic ground state of the system. The implementation of this functionality is based largely on the methods discussed in Section 1.4.

Other important functionality, which we will be using throughout this work are the calculation of geometries and band structures. *Geometry optimisation* is the process of searching for the configuration of atomic positions that minimises the total energy. This is important as it represents the atomic geometry the system would be expected to adopt naturally at zero temperature. In order to perform a geometry optimisation efficiently it is necessary to calculate the forces acting on each atom when the system in a given configuration. Although it is always possible to do this by calculating finite differences, this is expensive as it would require $\sim 3N$ calculations of the total energy. In order to avoid having to do this we can use the *Hellmann-Feynman theorem* [40, 41]. This essentially says that the derivative of the electronic energy with respect to some external parameter, such as an atomic position, is equal to the corresponding *partial derivative* in which the electronic wavefunction is held constant. This means that we can calculate the forces on the atoms without recalculating the electronic structure. With the forces known, we can move the atoms in a "downhill" direction closer to the local energy minimum. Successive iterations of this process will result in the system arriving at a geometry at which the forces are zero, and the energy is minimised. For extended systems, we also have to calculate the stress on the unit cell[6], and adjust the lattice parameters accordingly in a similar way to how we move the atoms in response to the forces in order to find the energy minimum.

[6]the way this is done is described later in Chapter 6.

Band structure calculations are carried out after a self-consistent (SCF) solution of the Kohn-Sham equations has been obtained. The Kohn-Sham potential is calculated from the ground state density, and then the corresponding eigenstates and eigenvalues, including conduction bands, are calculated for the set of band structure **k**-points. In general these will be different to the SCF **k**-points and so the band structure calculation is rarely fully self-consistent. It is also possible to calculate the band structure using a different exchange-correlation functional to the one used in the SCF stage of the calculation; this will be discussed further in Chapters 4 and 5.

1.7.4 Cell Symmetry

If there are geometric symmetries in the structure of the unit cell then this can be utilised in order to reduce computational requirements. We can take advantage of the fact that if two **k**-points are related by a symmetry operation, then the orbitals on those **k**-points are also related. Any symmetry operation, S, can be described by a rotation followed by a translation; for a point in real space we have

$$S\mathbf{r} = \mathbf{M}\mathbf{r} + \mathbf{t}, \tag{1.149}$$

where $\mathbf{M}$ is a matrix representing the rotational component of S, and $\mathbf{t}$ is a vector representing the translational component. For an orbital, $\phi_{i\mathbf{k}}(\mathbf{r})$, we have

$$S\phi_{i\mathbf{k}}(\mathbf{r}) = \phi_{i\mathbf{k}}(S^{-1}\mathbf{r}). \tag{1.150}$$

If the orbital is represented in reciprocal space, we therefore have

$$\begin{aligned} S\phi_{i\mathbf{k}}(\mathbf{r}) &= \frac{1}{\sqrt{\Omega}}\sum_{\mathbf{G}} c_{i\mathbf{k}}(\mathbf{G}) e^{i(\mathbf{k}+\mathbf{G})\cdot(\mathbf{M}^{-1}(\mathbf{r}-\mathbf{t}))} \\ &= \frac{1}{\sqrt{\Omega}}\sum_{\mathbf{G}} c_{i\mathbf{k}}(\mathbf{G}) e^{i(\mathbf{k}+\mathbf{G})\cdot(\mathbf{M}^{-1}(\mathbf{r}))} e^{i(\mathbf{k}+\mathbf{G})\cdot(\mathbf{M}^{-1}(-\mathbf{t}))} \\ &= \frac{1}{\sqrt{\Omega}}\sum_{\mathbf{G}} c_{i\mathbf{k}}(\mathbf{G}) e^{i(\mathbf{M}(\mathbf{k}+\mathbf{G}))\cdot\mathbf{r}} e^{i(\mathbf{k}+\mathbf{G})\cdot(\mathbf{M}^{-1}(-\mathbf{t}))}. \end{aligned} \tag{1.151}$$

Comparing coefficients, this leads to

$$Sc_{i\mathbf{k}}(\mathbf{G}) = c_{i\mathbf{M}^{-1}(\mathbf{k})}(\mathbf{M}^{-1}\mathbf{G}) e^{-i(\mathbf{k}+\mathbf{G})\cdot\mathbf{t}}. \tag{1.152}$$

Hence two $\mathbf{k}$ points are related by the operation, S, if they are mapped to and from each other by $\mathbf{M}$ and $\mathbf{M}^{-1}$. The reciprocal space coefficients are then the same, except for a phase factor. Because of this, it is possible to perform calculations using a reduced set of $\mathbf{k}$-points, defined such that no two $\mathbf{k}$-points are related by symmetry. Calculation of orbital dependent quantities then requires a weighting to be associated with each $\mathbf{k}$-point proportional to the number of related points in full set.

1.7.5 Time-Reversal Symmetry

Even in systems that have no geometric symmetries, the size of the $\mathbf{k}$-point set can still, in general, be reduced by $\sim \frac{1}{2}$ due to the "time-reversal symmetry" inherent in the Kohn-Sham equations. For any solution, $\phi_{i\mathbf{k}}(\mathbf{r})$, which by Bloch's theorem we can write as

$$\phi_{i\mathbf{k}}(\mathbf{r}) = u_{i\mathbf{k}}(\mathbf{r})e^{i\mathbf{k}\cdot\mathbf{r}}, \tag{1.153}$$

we can take the complex conjugate and obtain another solution, $\phi_{i-\mathbf{k}}(\mathbf{r})$,

$$\phi_{i-\mathbf{k}}(\mathbf{r}) = u^*_{i\mathbf{k}}(\mathbf{r})e^{-i\mathbf{k}\cdot\mathbf{r}}. \tag{1.154}$$

As this has the same contribution to the density as $\phi_{i\mathbf{k}}(\mathbf{r})$, in practice we only need to treat one of these $\mathbf{k}$-points explicitly and hence we can reduce the size of the $\mathbf{k}$-point set. Of course this does not apply to the Γ-point, i.e. $\mathbf{k} = \mathbf{0}$, and so the reduction is size is not necessarily exactly $\frac{1}{2}$.

1.7.6 Parallelism

Running in parallel generally increases the speed of a given calculation because the work is distributed between processors and so each processor has a smaller amount of work to do. It also generally reduces the memory requirements per processor for the same reason. It is rarely the case, however, that the speed-up is simply proportional to the number of processors. This is because there are "overheads" involved whenever data is transferred between processors. How close we get to the

limit of linear scaling with number of processors depends on how often this inter-processor communication has to take place.

Parallelisation can involve distribution of **k**-points, **G**-vectors, or both. The **k**-points and/or **G**-vectors are divided roughly evenly into groups, with each processor belonging to precisely 1 **k**-point group and 1 **G**-vector group. Where possible, distribution by **k**-points is preferable to distribution by **G**-vectors as there is less "interaction" between **k**-points, requiring less communication between processors.

1.8 Summary and Outline of Chapters

In summary, we have described the basic theory and methods that will be used in most of this work. We have seen how the Kohn-Sham formulation of DFT can be realised in practice by using a plane wave basis set, pseudopotentials, and simple approximations to the exchange-correlation functionals. The concept of the reciprocal space representation will be of particular importance in later chapters when we describe the theory and implementation of more advanced exchange-correlation functionals. We have also described briefly the structure and workings of the CASTEP code, a basic understanding of which will be useful later, particularly in Chapter 4, when we describe the computational implementation of new exchange-correlation functionals within this code.

Having described the basic methods that are used for our calculations, we proceed in Chapter 2 to calculate various properties of GaN from first principles, including geometric and energetic properties, band structures, and surface reconstructions. It is found that some of the calculated properties are in good agreement with experiment, but certain properties, most notably the electronic band gap, disagree with experiment quite markedly. The problem stems mainly from the use of the LDA to describe exchange and correlation effects, and this motivates us to consider using alternative, more advanced, exchange correlation functionals. In Chapter 3, we describe a number of such functionals, focusing in particular on screened exchange (sX-LDA), Hartree-Fock (HF), and exact exchange(EXX). Then in Chapter

4 we describe how we have implemented these functionals within CASTEP, and in Chapter 5 we present the results of band structure calculations on a number of semiconductors using sX-LDA and HF. We also discuss in some detail the reasons for the performance of the various functionals in calculating the electronic band gap. Aside from band structure calculations, another potential application of advanced exchange correlation functionals is in the calculation of geometries and molecular dynamics where the size of the unit cell may vary; this requires the calculation of the stress tensor. In Chapter 6 we derive and implement the theory that allows us to calculate the stress tensor when using the exchange-correlation functionals sX-LDA, HF, and EXX, and verify the theory and implementation by performing simple tests on silicon. Finally in Chapter 7 we summarise the work and its major conclusions, and discuss some possible directions for future work in this area.

Chapter 2

LDA Calculations on GaN and GaN Surfaces

In this chapter we briefly review the history of GaN and the wide range of technological applications for which it is used today. We also discuss modern growth techniques that are used to produce the material and some of the methods used to model this growth theoretically. More detailed accounts of these subjects may be found in references [42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53]. We then describe in detail a series of LDA calculations on the bulk material, including calculations of its geometry, formation energy, and band structure. After this, we discuss the theory of surface energetics, which allows us to compare surfaces with differing numbers of atoms, and which can be used to predict the surface reconstructions that will be present under various growth conditions. We then apply this theory to the calculation of reconstructions of GaN surfaces in the presence of hydrogen. We conclude by discussing the success or otherwise of the LDA in these calculations, in particular with regard to its underestimation of the band-gap.

2.1 About GaN

2.1.1 GaN and GaN-based Technology

Gallium nitride (GaN) is binary semiconductor compound. Its natural structure under standard conditions is the hexagonal wurtzite structure, but can also be produced with a cubic zinc blende structure. Both of these structures are described in more detail in 2.2.1. GaN was first produced in the 1930s, but was not able to be produced in large enough crystals for technological applications. The development of vapour phase epitaxy (VPE) [42, 43, 44] techniques led to an increase in quality of production from the 1960s onward. With modern growth techniques, discussed in 2.1.2, GaN based devices are now routinely manufactured for technological applications.

GaN, and GaN alloys, have several useful properties in terms of technological applications. The most important property of GaN is wide electronic band gap ($\sim$ 3.5eV [54]), which means it can be used to make short-wavelength opto-electronic devices such as light-emitting diodes (LEDs), laser diodes (LDs), and photo-detectors. Another important property is its high melting point ($\sim$ 1700°C [44]), which along with the wide band gap, make it a candidate for use in high temperature electronic applications.

2.1.2 Modern Growth Techniques

The growth of GaN is an area of ongoing development, and can involve a wide range of complicated techniques and processes. Most modern techniques, however, involve either metal-organic vapour phase epitaxy (MOVPE) or molecular beam epitaxy (MBE) [42, 43, 44]. In both of these methods, the material is built up layer by layer from a substrate, such as sapphire, usually separated with a buffer layer of AlN to rectify the lattice mismatch.

In MOVPE, chemical compounds in vapour phase are passed over the hot surface, where they react to form the bulk GaN plus waste products in vapour phase. In

order for the reaction to occur it is necessary to heat the surface to around 1000°C. Typical sources for Ga and N are trimethylgallium (TMGa) and ammonia (NH_3). In MBE, beams of atoms and molecules are directed at the surface where again they react to form the bulk material. The source of Ga atoms is now bulk Ga, while the source of N is, again, ammonia.

2.1.3 Modelling Growth Conditions

An important factor influencing the growth of GaN and other semiconducting materials is the atomic structure of the surface. This can, for example, affect the mobility of atoms on the surface, which can in turn affect the quality of the crystals produced. It can also affect the ease with which dopant atoms can become incorporated into the material. While in-situ experimental techniques such as reflection high energy electron diffraction (RHEED) can give some indication as to the surface structure during growth, it has been found that theoretical studies can provide much more detailed information as to the precise atomic configurations under different growth conditions. A number of such studies can be found in the references [46, 47, 48, 49, 50, 51, 52, 53]. In terms of first principles calculations using the density functional methods described in Chapter 1, the basic procedure is to set up a *surface supercell*, and perform geometry optimisation calculations to obtain the relative energies of the different possible *reconstructions* of the surface under study. The detailed procedure for these calculations will be explained further in sections 2.3 and 2.4 when we carry out calculations of surface reconstructions in the presence of H.

2.2 Calculations on Bulk GaN

2.2.1 The Unit Cell

The primitive GaN unit cell contains 4 atoms, in the case of the wurtzite structure (space group $P6_3mc$), and 2 atoms, in the case of the zinc blende structure (space

group F$\bar{4}$3m). There are several equivalent ways to define the unit cells. For the purposes of these initial calculations we will define the structures as follows:

The shape of the wurtzite cell is a vertically oriented prism, with the base defined by the primitive lattice vectors, **a**, and **b**, which are of equal length and are separated by an angle of 60^o; **a** and **b** both lie in the horizontal xy-plane. The height of the cell is defined by the vector, **c**, which is oriented vertically at 90^o to both **a** and **b**. In the "ideal" wurtzite structure c is related to a by $c = 2\sqrt{\frac{2}{3}}a$; this is not necessarily the case in the real structure, as we will discuss in a moment.

To specify the positions of atoms within the cell we usually use *fractional coordinates* for convenience. If a point in space, **r**, has Cartesian coordinates, (x, y, z), then its fractional coordinates, $[x', y', z']$, are defined such that

$$\mathbf{r} = x'\mathbf{a} + y'\mathbf{b} + z'\mathbf{c}. \tag{2.1}$$

Note that we write fractional coordinates in square brackets to distinguish them from Cartesian coordinates.

The Ga atoms are positioned such that one is at the origin, $[0, 0, 0]$, and the other is at $[\frac{1}{3}, \frac{1}{3}, \frac{1}{2}]$. The N atoms are positioned directly above the Ga atoms. In the "ideal" wurtzite structure, these are at $[0, 0, \frac{3}{8}]$ and $[\frac{1}{3}, \frac{1}{3}, \frac{7}{8}]$, so that the length of each Ga-N bond is the same if $c = 2\sqrt{\frac{2}{3}}a$; a graphical representation of the ideal wurtzite cell is shown in Figure 2.1.

However, in terms of cell symmetry, the vertical Ga-N bonds are not related to the diagonally oriented Ga-N bonds. Because of this, there is no *a priory* reason to expect these two sets of bonds to be the same length. There are therefore two extra degrees of freedom compared to the ideal structure - the length of the lattice vector, **c**, relative to **a** and **b**, and the vertical position of the N-atoms, relative to the Ga-atoms.

The deviation of the atomic coordinates from the ideal structure can be described in terms of a parameter, d, such that the positions of the N-atoms are given by $[0, 0, \frac{3}{8} + d]$ and $[\frac{1}{3}, \frac{1}{3}, \frac{7}{8} + d]$.

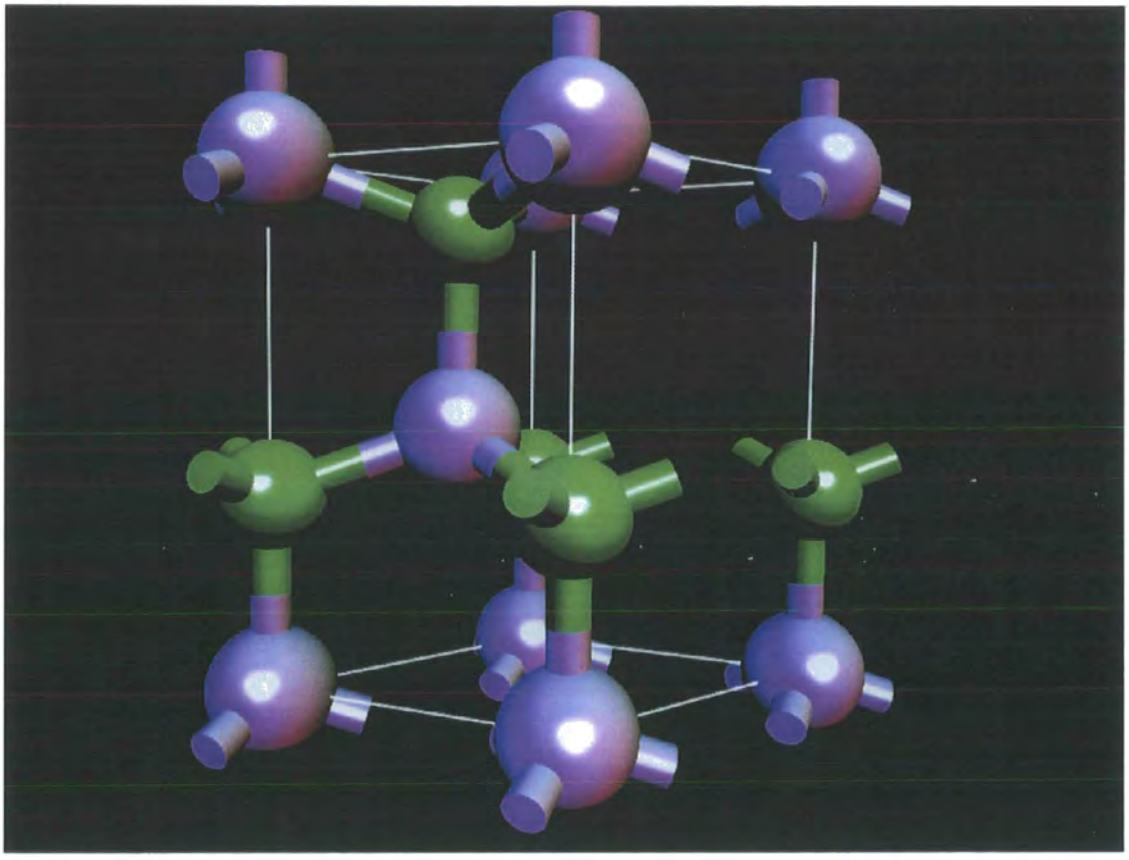


Figure 2.1: Primitive unit cell of wurtzite GaN. Ga atoms are represented by large grey spheres, and N atoms by smaller green spheres.

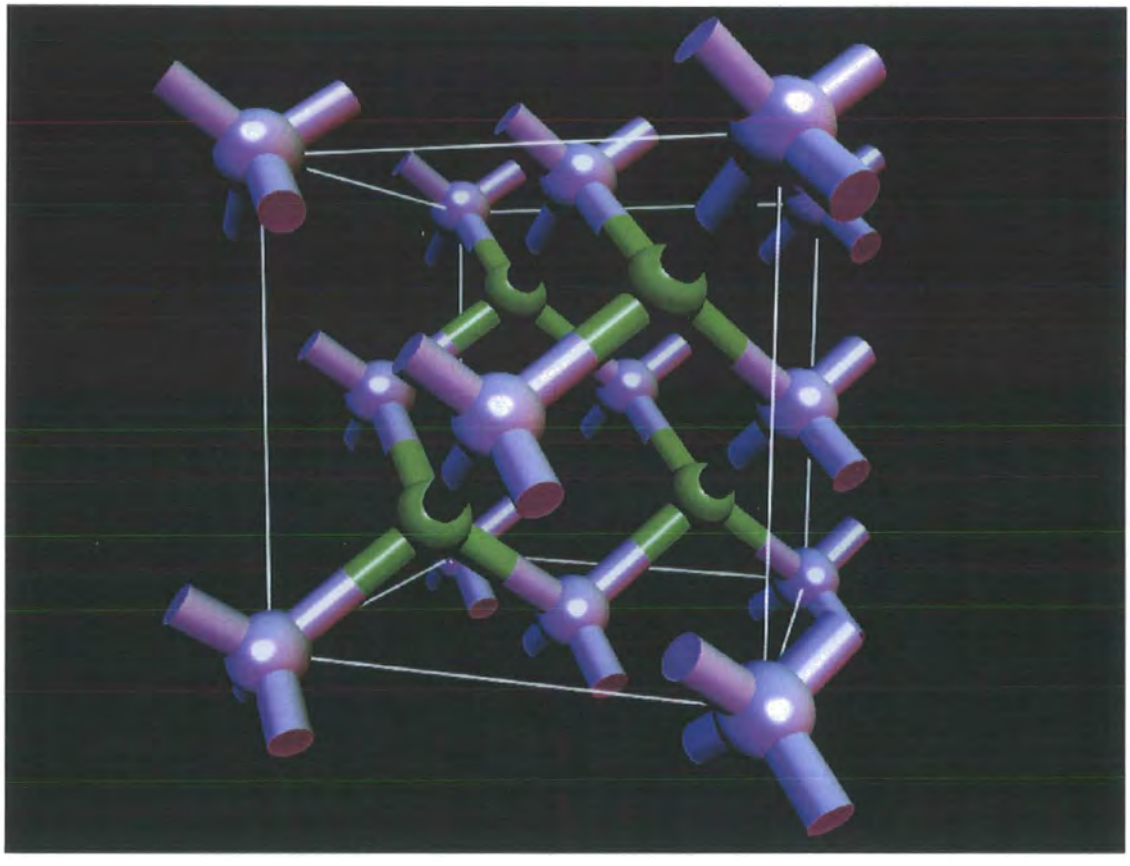


Figure 2.2: 8-atom cubic cell of zinc blende GaN. Ga atoms are represented by large grey spheres, and N atoms by smaller green spheres.

The shape of the primitive 2-atom zinc blende cell is an equal-sided parallelepiped that can be most easily visualised with reference to a larger, 8-atom cubic cell, as shown graphically in Figure 2.2. This cubic cell has Ga-atoms at the origin and in the centre of each of the three faces that touch the origin. For each Ga-atom, there is a N-atom at a displacement of $[\frac{1}{4}, \frac{1}{4}, \frac{1}{4}]$ away from it. The lattice vectors defining the primitive cell are the three vectors going from the origin to the centre of the three faces where the Ga-atoms are. These vectors are of equal length and are separated from each other by angles of 60^o. The three Ga-atoms on the faces of the cube are not in the primitive cell as they are simply the periodic repetitions of the atom at the origin. The primitive cell thus contains a Ga-atom at $[0, 0, 0]$ and a N-atom at $[\frac{1}{4}, \frac{1}{4}, \frac{1}{4}]$.

2.2.2 Choice of Pseudopotentials

We have found from our own tests, as has also been noted by others [46, 47, 48, 49, 50, 51, 52, 53], that for the GaN surface calculations covered in this work, we need to use norm-conserving pseudopotentials, and include the Ga *d*-electrons as valence. The ultrasoft pseudopotentials that we have tested fail to adequately describe the N≡N triple bond in the N_2 molecule, which we need to model in order to compare surfaces with different numbers of N-atoms. For all the calculations in this work we use the norm-conserving pseudopotentials from the standard set available with CASTEP, generated by Lee [55, 56].

2.2.3 Convergence of the Plane Wave Basis Set

With a given structure and set of pseudopotentials in place, the first task in any plane wave calculation should be to choose an appropriate cut-off energy for the basis set. This is done by means of a *convergence test*, in which we perform a series of calculations, using increasing cut-off energy, and monitor the convergence of a given quantity (usually the total energy) towards its large cut-off limit.

For these purposes, we only need to use 1 **k**-point to sample the Brillouin zone. This is because we are only seeking to determine a suitable cut-off energy for convergence

rather than a final result. However, we choose to give the **k**-point an offset of $[\frac{1}{2}, \frac{1}{2}, \frac{1}{2}]$ as this is known to give a more accurate answer at no extra computational expense (as is found later in 2.2.4 and is also noted in references [57, 58]). We set up the wurtzite and zinc blende cells with the atoms positioned in the ideal structure, as described earlier, and the lattice parameters set near to the experimental values [54]. Note that this is done simply for the purposes of the convergence test - it does not result in a semi-empirical calculation as both the atomic coordinates and lattice vectors will be allowed to vary later.

We run a series of total energy calculations using CASTEP, for a range of cut-off energies, the results of which are plotted in Figure 2.3; we find that the calculated value of E converges towards a value of around -3935.198eV. As a general rule in electronic structure, we would like to know energies to an accuracy of within 0.01eV per atom, which in this case is safely reached with a cut-off of around 1200eV.

What we have established here is an appropriate cut-off energy for use in calculations in which we want to know the total energy of the system as an absolute quantity. However, in almost all problems in physics, what is important is not the absolute energy, but rather the *difference* in energy between alternative configurations. In our study of GaN, we are looking to determine properties that are essentially chemical in nature. Chemical properties are mostly determined by the electronic structure in the regions of space *between* the atoms, rather than in the space within the atomic cores. This fact has already been made use of in the pseudopotential approximation, and, as we will now see, can also allow us to use a lower cut-off energy than is suggested by this first convergence test.

We now perform a second convergence test, which looks at the convergence of the *difference* in energy between bulk GaN and a system in which the atoms are in a very different chemical environment. A sensible choice for such a system is the case of isolated atoms, because this is a very different chemical environment to anything involving bonding. To calculate the energy of the isolated atoms, we place them in a 4Å cubic cell (possibly not large enough to completely eliminate interactions between neighbouring cells, but adequate for the present purposes). Since isolated atoms tend to be spin polarised, we treat them using spin-dependent DFT, with

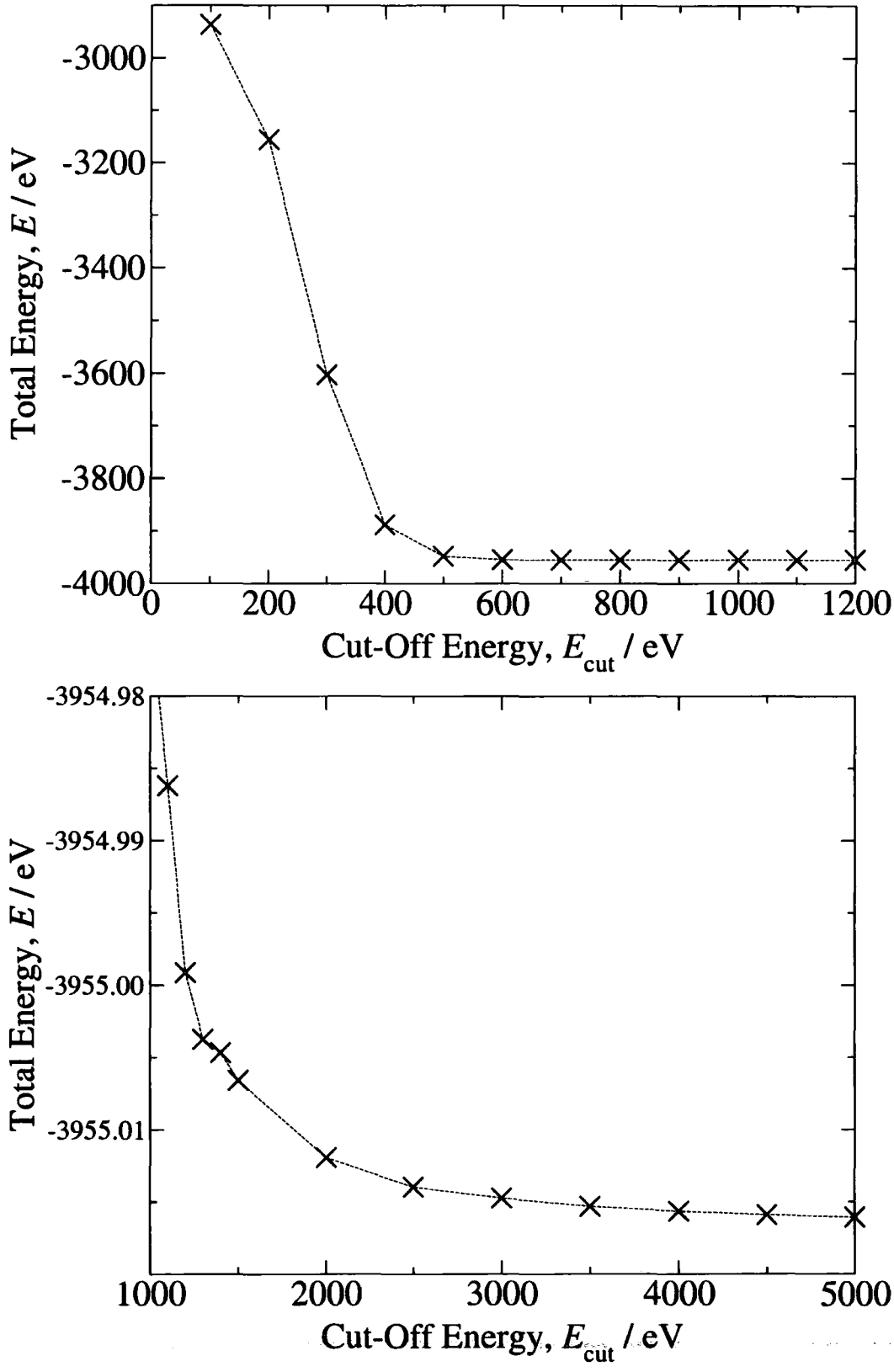


Figure 2.3: Convergence of the total energy, E, per 4-atom cell, with respect to the plane wave cut-off energy, E_{cut}, in wurtzite GaN, using a single **k** -point to sample the Brillouin zone. The two graphs display the same data at different scales of magnification.

the spin configuration determined by Hund's rules. The results of this convergence test are plotted in Figure 2.4. We now see that the energy difference converges to within the acceptable tolerance of 0.01eV per atom far more rapidly than did the total energy. For the calculations we will be doing, a cut-off energy of 800eV ought to be more than sufficient.

2.2.4 Convergence of the k-point Set

Now that we have established the cut-off energy, using a single **k**-point, we must perform a similar test to determine the appropriate number of **k**-points to use. Again, we perform a series of calculations, monitoring the convergence of the total energy, this time as we increase the density of the **k**-point grid, keeping the cut-off energy fixed at 800eV.

For convenience, we start with the ZB structure only, as this has reciprocal lattice vectors all of equal length, meaning that we can set $M_x = M_y = M_z$ and thus control the **k**-point density with only 1 free parameter. For each **k**-point density, we will consider two types of offset - *on-origin*, in which the central **k**-point is located on the origin, and *off-origin*, in which the grid is offset such that the origin is exactly in the centre of the cube formed by the 8 **k**-points nearest to it (in fractional coordinates). The results of this convergence test are shown in Figure 2.5; we see that, for off-origin grids, the **k**-point set is converged to within an acceptable tolerance when it has dimensions of $3 \times 3 \times 3$. This is not the case for on-origin grids, which appear to converge less rapidly, suggesting that one should avoid choosing such grids where possible. By default, in CASTEP, grids with odd dimensions are on-origin, while grids with even dimensions are off-origin. This can lead to a zig-zagging of the convergence plot, which may lead inexperienced users to doubt whether even a $4 \times 4 \times 4$ grid is really converged, when a $5 \times 5 \times 5$ grid, by default, gives a very different result. By plotting both the on-origin and off-origin results, as we have done, it becomes clear that a $3 \times 3 \times 3$ grid is actually quite adequate, so long as it is off-origin. This issue is discussed in greater detail in references [57, 58].

We now consider the WZ structure, in which $\mathbf{c}^*$ is shorter than $\mathbf{a}^*$ and $\mathbf{b}^*$, by

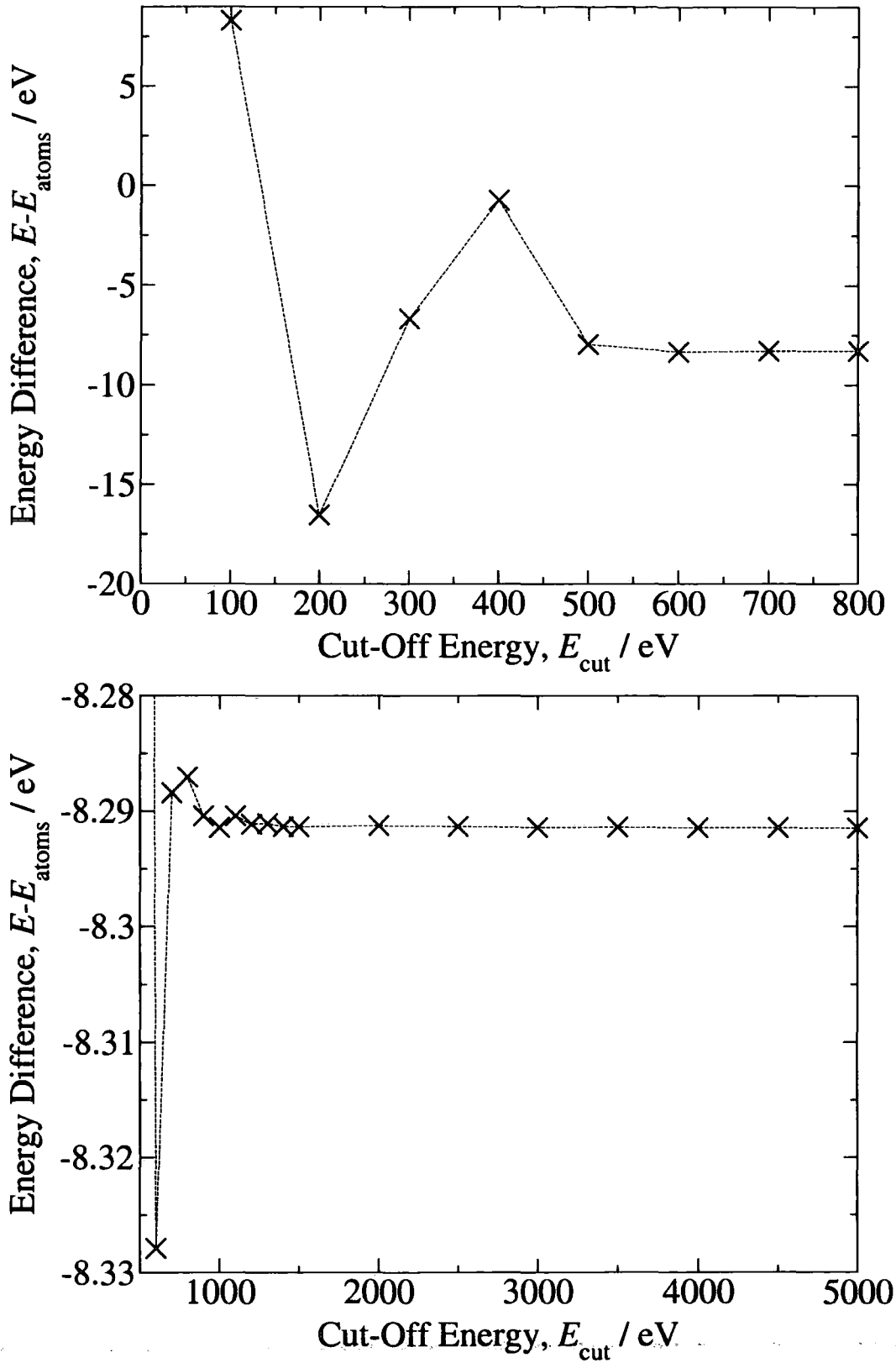


Figure 2.4: Convergence of the difference in energy, $E - E_{\text{atoms}}$, per 4-atom cell, between wurtzite GaN and its constituent atoms, with respect to the plane wave cut-off energy, E_{cut}, using a single k-point to sample the Brillouin zone. The two graphs display the same data at different scales of magnification.

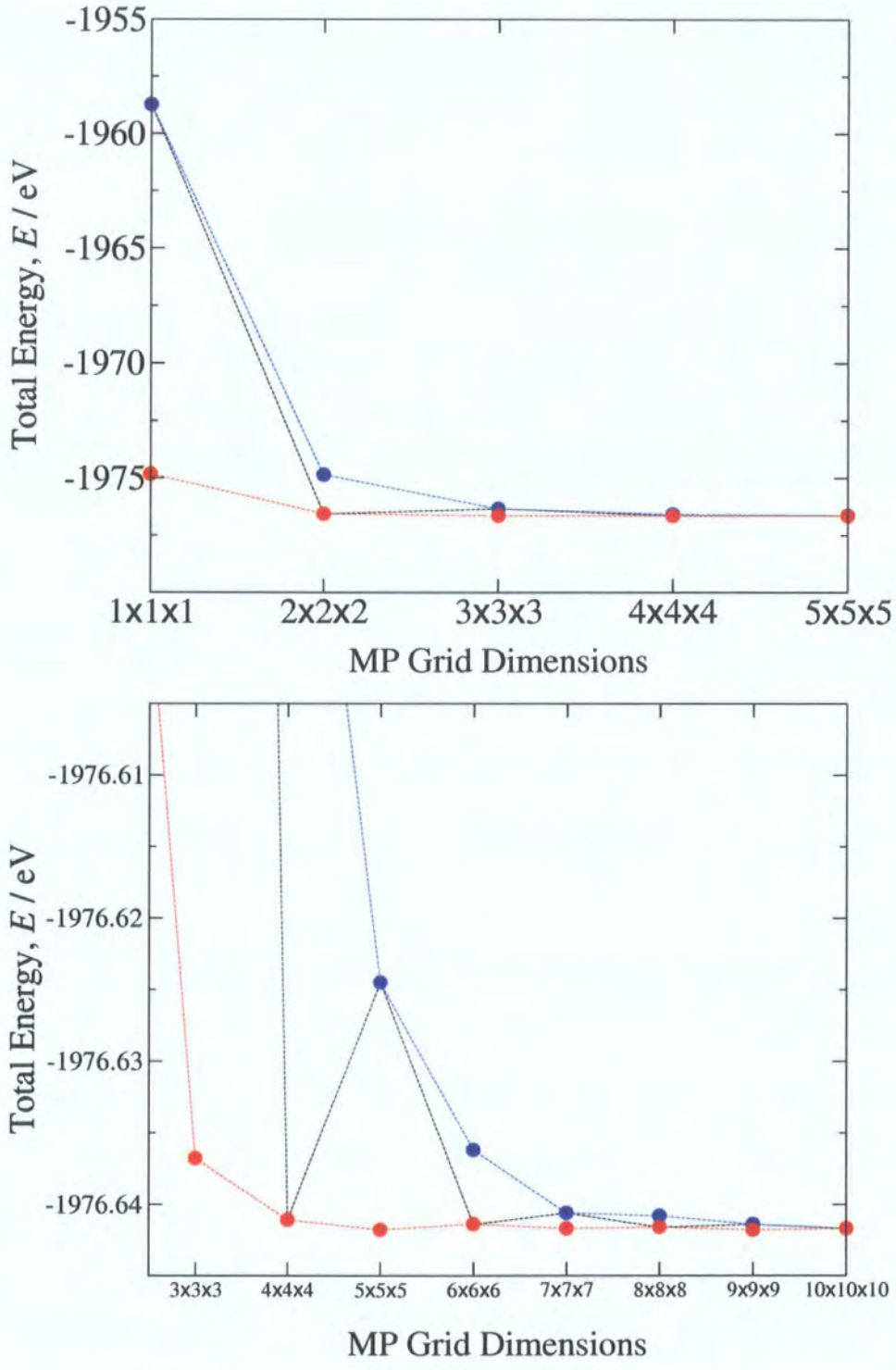


Figure 2.5: Convergence of the total energy, E, per 2-atom cell, in zinc blende GaN, as the size of the Monkhorst-Pack **k**-point grid is increased. The blue line represents grids that are on-origin, while the red line represents grids that are off-origin. The black line represents the CASTEP default, which is on-origin for odd grids and off-origin for even grids.

a factor of $\frac{1}{2}\sqrt{\frac{3}{2}} \approx 0.612$ in the ideal structure. This means that for a roughly even distribution of **k**-points, we might be able to set M_z to less than M_x and M_y. Because the base of the Brillouin zone is not rectangular we should not simply set $M_z \approx \frac{1}{2}\sqrt{\frac{3}{2}}M_x$. We have to consider the 2-dimensional density of **k**-points in the horizontal plane; the area of the base is $|\mathbf{a}^* \times \mathbf{b}^*| = \frac{\sqrt{3}}{2a^2}$ and hence the **k**-point density in the horizontal plane is $\frac{2M_x^2a^2}{\sqrt{3}}$. For an equal effective 1-dimensional sampling density in each of the 3 Cartesian directions, this would be the square of of that sampling density. Because the $\mathbf{c}^*$-direction is normal to the horizontal plane, we thus require

$$\begin{aligned} M_z^2c^2 &\approx \frac{2M_x^2a^2}{\sqrt{3}}, \\ \Rightarrow M_z^2 &\approx \frac{2M_x^2a^2}{c^2\sqrt{3}}, \\ &\approx \frac{\sqrt{3}}{4}M_x^2, \\ \Rightarrow M_z &\approx \frac{\sqrt[4]{3}}{2}M_x \approx 0.658M_x. \end{aligned} \tag{2.2}$$

To test for convergence, we set a real target value, m_x, as the control parameter, and then select the nearest integer to m_x as the value for M_x, and the nearest integer to $\frac{\sqrt[4]{3}}{2}m_x$ as the value for M_z. This ensures that the **k**-point grid will remain roughly evenly distributed. The issue of origin offsets is also complicated now, because it is possible for the grid to be, for example, off-origin in the horizontal plane but on-origin in the vertical direction, or vice versa. For clarity, for each set of grid dimensions, we consider the fully on-origin case, the fully off-origin case, and the the two mixed cases. The results are plotted in Figure 2.2.4.

For the wurtzite GaN structure, we see that the energy is adequately converged with a $4 \times 4 \times 3$ Monkhorst-Pack grid, with the CASTEP default offsets, i.e. off-origin in the horizontal plane and on-origin in the vertical direction.

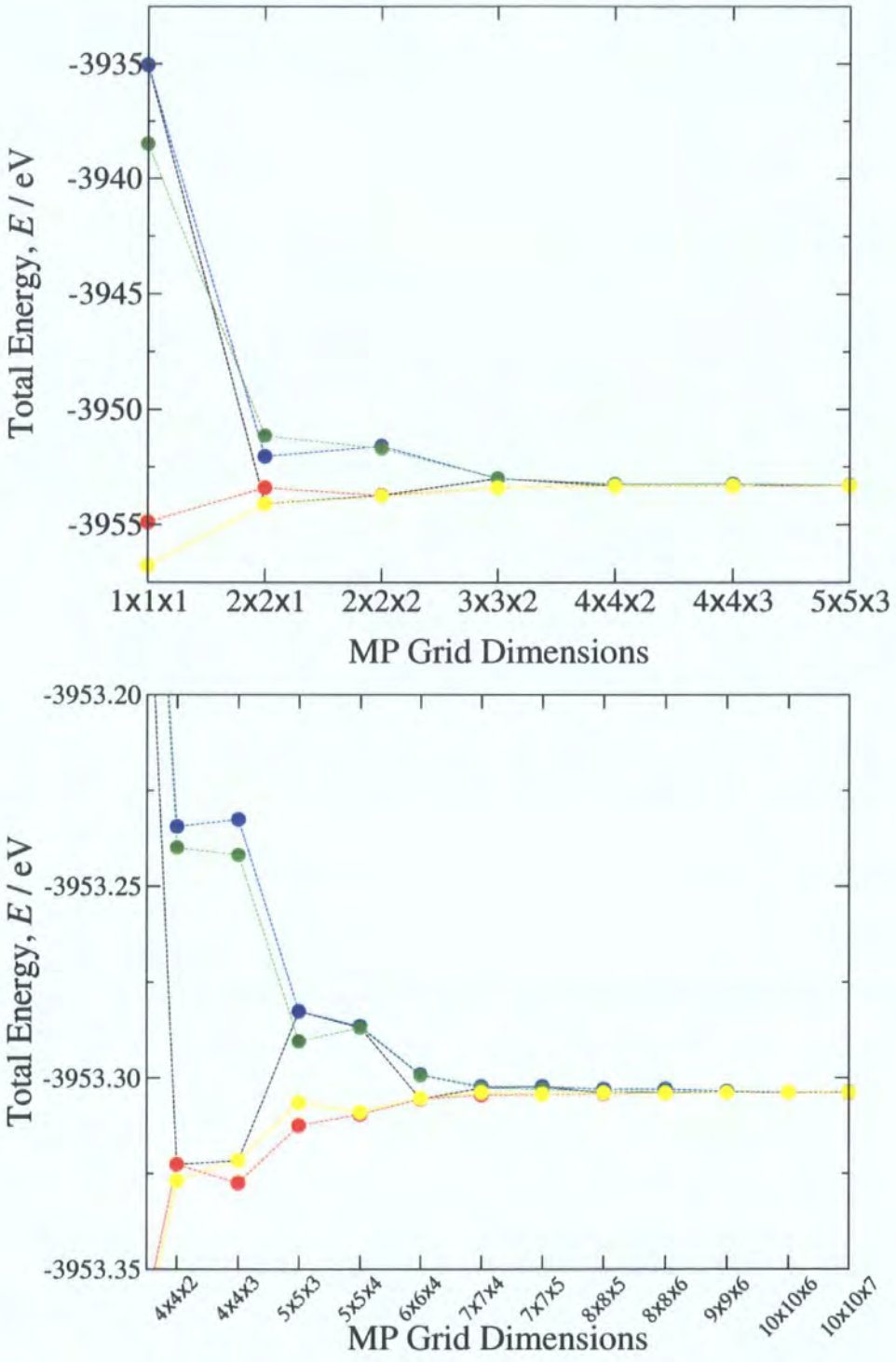


Figure 2.6: Convergence of the total energy, E, per 4-atom cell, in wurtzite GaN, as the size of the Monkhorst-Pack k-point grid is increased. The blue line represents grids that are fully on-origin, the red line represents grids that are fully off-origin, the green line represents grids that are on-origin in the horizontal plane but off-origin in the vertical direction, while the yellow line represents grids that are off-origin in the horizontal plane but on-origin in the vertical direction. The black line represents the CASTEP default, which is on-origin for odd grids and off-origin for even grids.

2.2.5 Geometry Optimisation

As mentioned earlier, the geometries of the cells have only been defined approximately. For the two structures, we must now find the geometry that minimises the total energy by performing a geometry optimisation calculation as described in 1.7.3. We specify tolerances for iterative convergence of the geometry such that the atomic positions are converged to within ~ 0.001Å. For each total energy calculation we use an 800eV plane wave cut-off and the Monkhorst-Pack grids as determined in the convergence tests. We obtain geometric parameters for wurtzite of $a = 3.18$Å, $c = 5.18$Å, $d = 3.03 \times 10^{-4}$, and for zinc blende of $a = 4.50$Å.

2.2.6 Energetics

We now have the correct geometry for GaN (when using the LDA with these pseudopotentials). We also have the total energy of the system per unit cell. However, the total energy by itself is not a particularly important quantity. What is more important is difference in energy between the GaN structure and other possible configurations of the same atoms. For example, the *cohesive energy* of the structure is the energy that would be required to pull the structure apart into its individual atoms. This essentially tells us how well bound the structure is, and is related to the material's strength, and melting/boiling points. To calculate the cohesive energy we must calculate the energy of an isolated Ga atom and an isolated N atom, using spin-dependent DFT to describe the spin-polarised nature of these systems. This requires us to use a large enough unit cell that the atom does not interact with the atoms in the neighbouring cells. Again, this can be done by means of convergence tests. For the Ga atom, we use a 12Å cubic cell, which gives an energy of -1701.05eV, and for the N atom we use a 7Å cubic cell, which gives and energy of -265.17eV.

Now that we have established the energy of isolated Ga and N atoms, we can calculate the cohesive energy of GaN. We will define this as the cohesive energy per GaN dimer:

$$E_B = E_{\mathrm{GaN}} - E_{\mathrm{Ga}-atom} - E_{\mathrm{N}-atom}, \tag{2.3}$$

The energy per dimer of the wurtzite structure is -1976.67eV, while the energy per dimer of the zinc blende structure is -1976.65eV. This gives a cohesive energy of 10.46eV for the wurtzite structure and 10.43eV for the zinc blende structure. The experimental value may be in a certain amount of doubt, as we will discuss in a moment in the context of the formation energy, but the tabulated value for the wurtzite structure is 8.96eV [59]. This is much less than the value we have just calculated with the LDA. The LDA is therefore *overbinding* the structure, which is a systematic failing of this functional [60].

Another important energetic quantity is the *formation energy.* This is the change in energy when the structure is formed from its constituent elements in their natural state. That is metallic gallium (α-Ga), and nitrogen gas (N_2). The formation energy per GaN dimer is given by

$$E_{form} = E_{\mathrm{GaN}} - E_{\alpha-\mathrm{Ga}} - \frac{1}{2}E_{\mathrm{N_2}}. \tag{2.4}$$

Again, to calculate this quantity, we must perform separate calculations on α-Ga and N_2, including convergence tests for **k**-points and cell size. We obtain an energy of -1704.52eV per atom of α-Ga and an energy of -540.90eV for the N_2 molecule. This gives a formation energy for wurtzite GaN of -1.70eV and for zinc blende GaN of -1.68eV.

While quoted experimental values vary, commonly quoted values are generally in the region of -1.2eV [61]. This appears to suggest that the LDA is overestimating the formation energy. However, more recent experiments have put the value at around -1.6eV [62]. Much LDA work in the literature employs pseudopotentials that give values in the region of -1.2eV, despite the fact that all-electron calculations have shown that the correct LDA value is actually around -1.56eV [63]. It is likely that pseudopotentials were selected in order to reproduce what were believed to be the correct experimental results, rather than all-electron results. Our calculated value of -1.70eV is much closer to all-electron results, and also to recent experiments [62], than that used in in most other work.

2.2.7 Band Structure Calculations

We now calculate the Kohn-Sham band structure with the LDA, as described in 1.7.3, for both the wurtzite and zinc blende GaN structures. The shape of the first Brillouin zone for these structures is shown in Figures 2.7 and 2.8. The full band structure is a 3-dimensional function, which is difficult to represent graphically (and expensive to calculate in high resolution). The usual way to present band structures is as a function of a 1-dimensional path through reciprocal space made up of straight lines connecting points of high symmetry. These points are marked on Figures 2.7 and 2.8.

The calculated band structures are shown in Figures 2.9 and 2.10. The important features of the band structures can be summarised by tabulating the eigenvalues of the highest valence and lowest conduction band eigenvalues at the symmetry points, relative to the valence band maximum; these are shown in Tables 2.1 and 2.2.

According to these LDA calculations, wurtzite GaN has a direct band gap of 1.86eV, and zinc blende GaN has a slightly lower direct band gap of 1.70eV. These values are significantly lower than the experimentally measured value for wurtzite GaN of $\sim$ 3.5eV [54] (The gap for zinc blende GaN has not been reliably measured but is believed to be in the order of a few tenths of an eV lower than that of the wurtzite structure at around 3.3eV [64]). This is an example of the "band gap problem", which is a general property of the LDA functional and can only be rectified in a credible manner by using more advanced exchange-correlation functionals. This issue is dealt with in the remaining chapters of this work.

2.3 Theory of Surface Energetics

In performing first principles calculations on GaN surfaces, our aim is to compare surfaces with different *reconstructions*, to see which is energetically the most favourable in different conditions. A surface reconstruction is any atomic configuration of the surface that is different to what we would get if we were to imagine

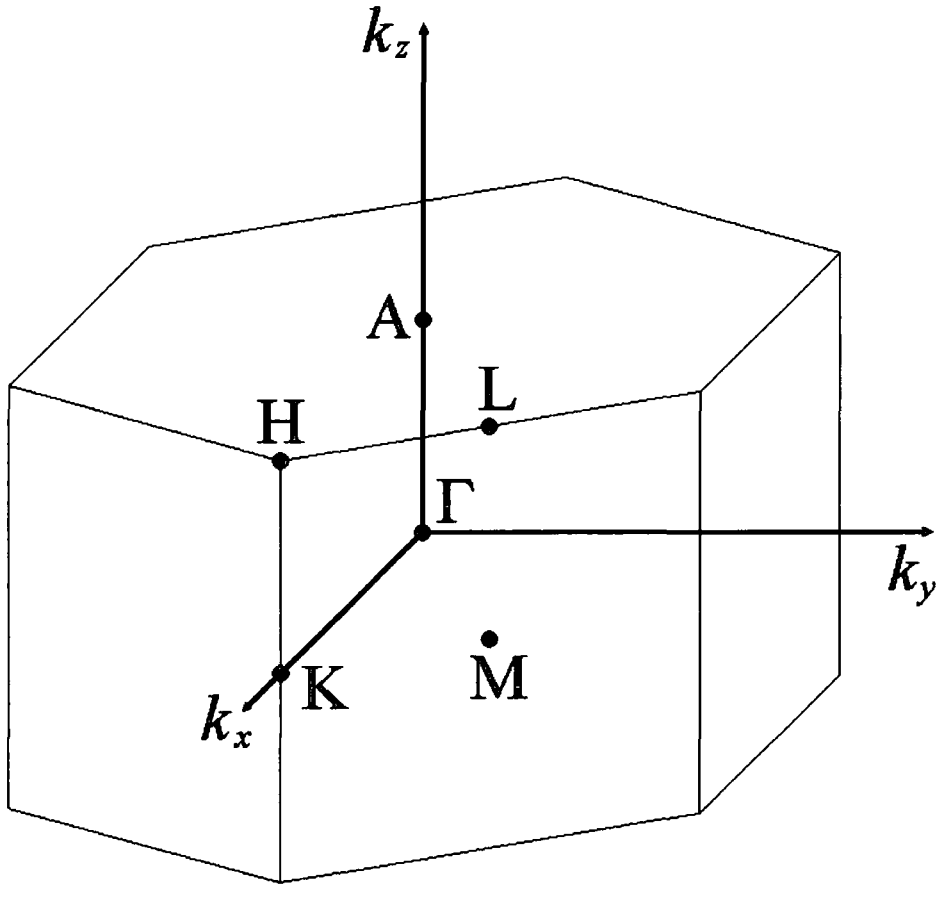


Figure 2.7: Brillouin zone for hexagonal lattices, such as the wurtzite structure, with the points of high symmetry labelled.

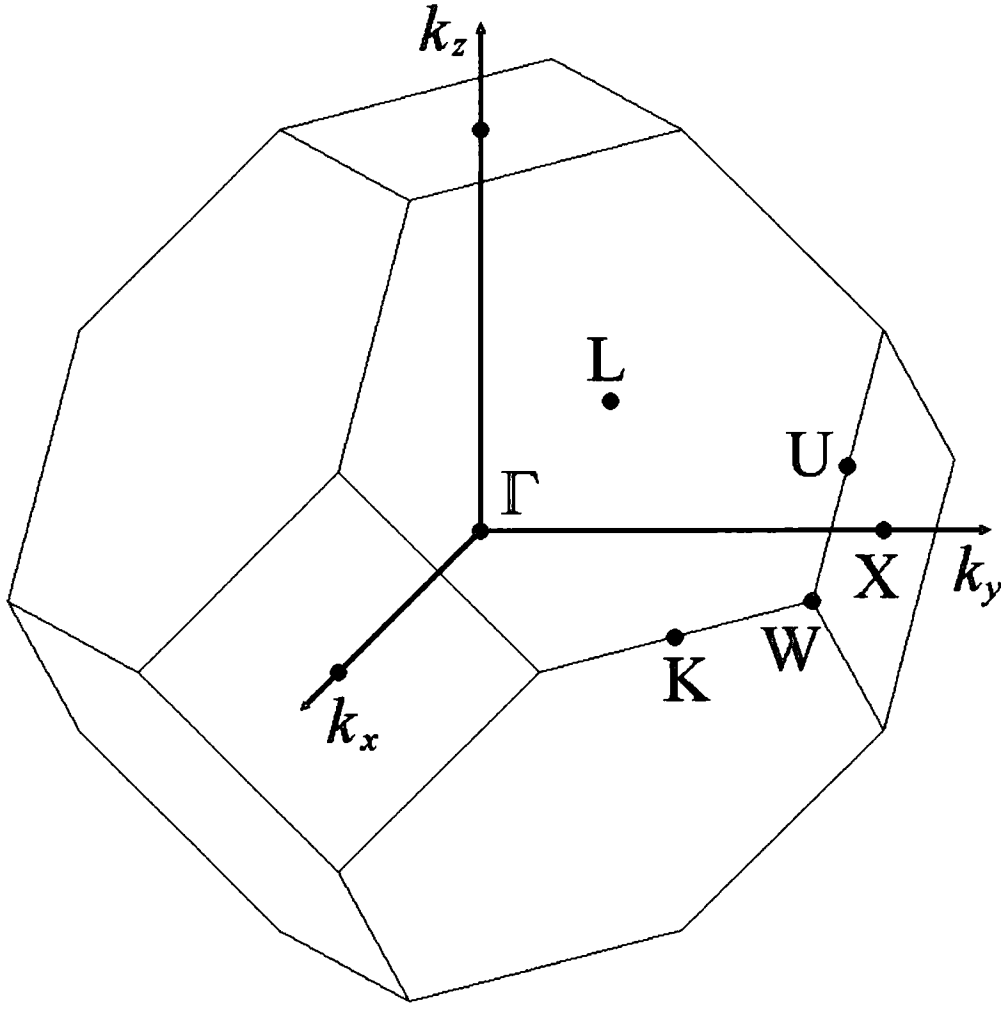


Figure 2.8: Brillouin zone for face centred cubic lattices, such as the zinc blende structure, with the points of high symmetry labelled.

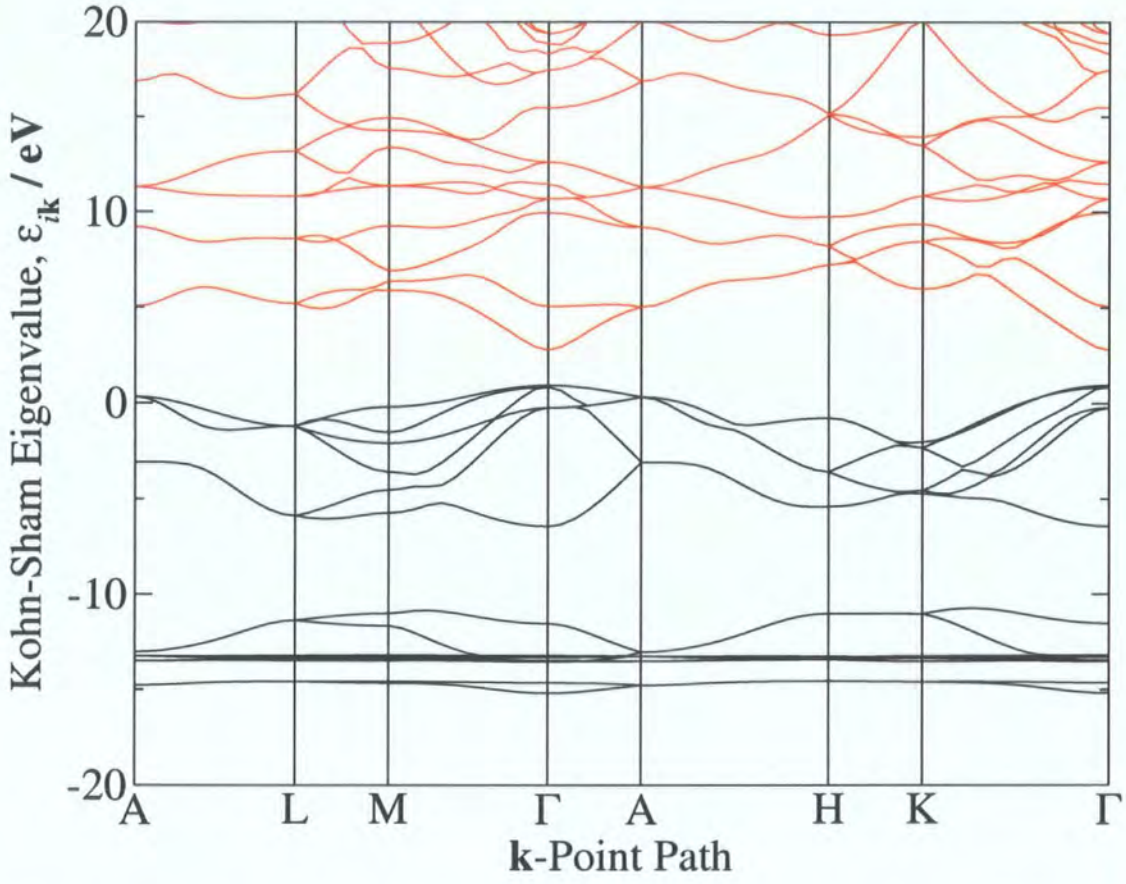


Figure 2.9: Kohn-Sham band structure of wurtzite GaN calculated using the LDA. Black lines indicate occupied valence bands, while red lines indicate unoccupied conduction bands.

Symmetry Point	V.B. Max. / eV	C.B. Min. / eV
A	-0.61	4.10
L	-2.11	4.26
M	-1.12	4.96
Γ	0.00	1.86
H	-1.70	6.31
K	-2.95	5.04

Table 2.1: Eigenvalues of the highest valence band and lowest conduction band at the symmetry points in wurtzite GaN, calculated with the LDA.

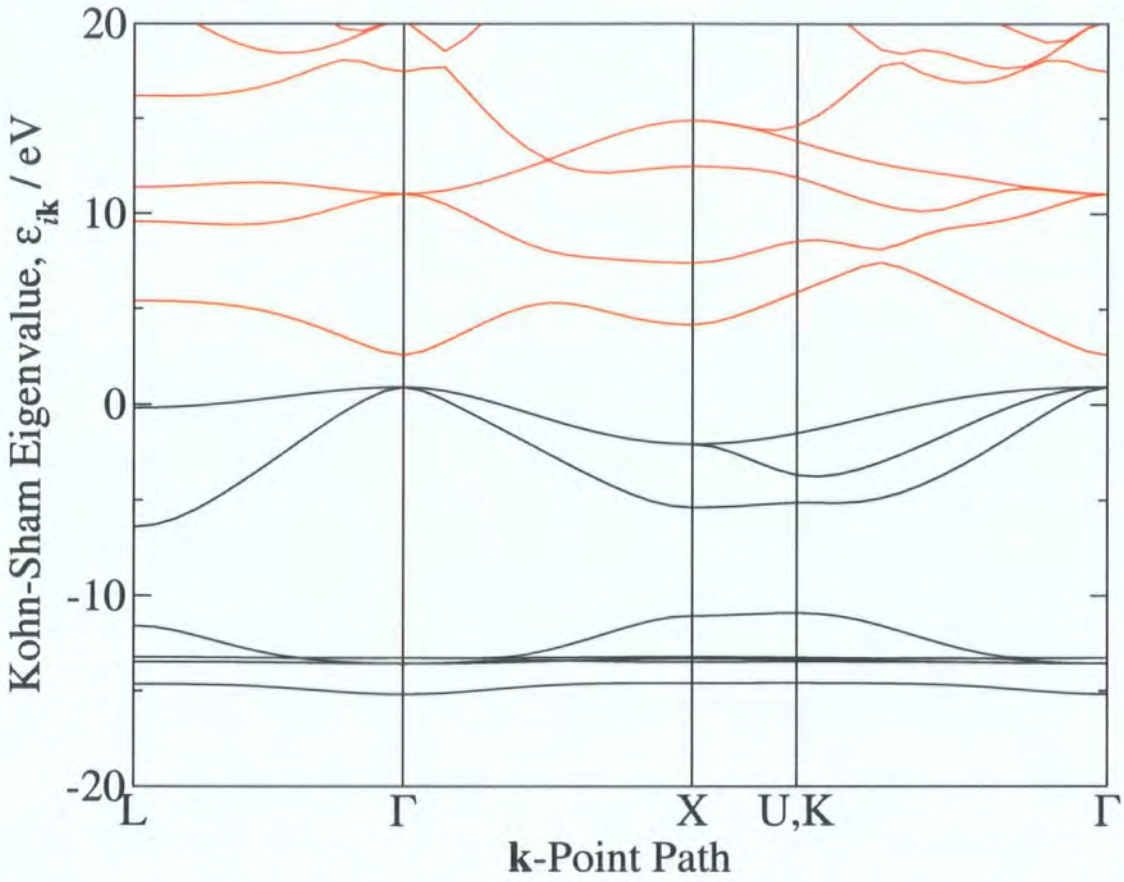


Figure 2.10: Kohn-Sham band structure of zinc blende GaN calculated using the LDA. Black lines indicate occupied valence bands, while red lines indicate unoccupied conduction bands.

Symmetry Point	V.B. Max. / eV	C.B. Min. / eV
L	-1.02	4.49
Γ	0.00	1.70
X	-2.99	3.28
U,K	-2.42	4.98

Table 2.2: Eigenvalues of the highest valence band and lowest conduction band at the symmetry points in zinc blende GaN, calculated with the LDA.

simply cleaving the bulk material. In experimental growth situations, various parameters can be adjusted, each of which can change the energetic favourability of different surfaces. This means that the phase diagram representing the particular reconstruction that is energetically favourable in a particular set of conditions is, in principle, a multi-dimensional object. However, Van de Walle and Neugebauer have shown [52] that the temperature and the various partial pressures can effectively be collapsed onto a smaller set of *chemical potentials*, the definition of which we discuss in a moment.

In most first principles calculations, we are effectively studying a system at zero temperature because we are not considering the kinetic energy of the atoms. For systems with a fixed number of atoms of each species, therefore, we obtain the atomic configuration simply by minimising the total energy. However, if we want to compare systems with different numbers of atoms, we must instead define some *free energy*, F, that is to be minimised. This free energy includes terms related to the number of atoms of each species, i.e.

$$F = E_{TOT} - \sum_X \mu_X n_X, \tag{2.5}$$

where n_X is the number of atoms in the system of species X, and μ_X is the chemical potential for atoms of that species. The chemical potential is related to external environmental conditions, such as the temperature, pressure, and concentration of atomic or molecular species containing X. It cannot be determined from first principles calculations - all we can obtain from first principles calculations is the free energy of a system *as a function of the chemical potential.* However, we can use physical arguments to determine the *range* of chemical potential over which a surface could ever be stable.

For example, consider the case of a GaN surface that is in thermal equilibrium with its environment. Thermal equilibrium requires that the addition of a complete extra layer of GaN to this surface cannot change its free energy - if it did then the surface would either grow or retract. This restriction can be expressed mathematically as

$$E_{\mathrm{GaN}} - \mu_{\mathrm{Ga}} - \mu_{\mathrm{N}} = 0, \tag{2.6}$$

i.e. for a GaN surface in thermal equilibrium with its environment, the sum of the Ga and N chemical potentials must equal the total energy per GaN dimer in the bulk material. This means, if there are no other atomic species present, the number of free environmental parameters affecting the surface can be reduced from 2 to 1, as μ_N is directly related to μ_{Ga}, and vice versa, i.e.

$$\mu_N = E_{GaN} - \mu_{Ga}. \tag{2.7}$$

We will choose to use μ_{Ga} as our variable parameter. There is a limited range of values that μ_{Ga} can actually take - outside of this range the system would become unstable as it would be energetically favourable to form either bulk α-Ga or N_2 molecules. In order for these substances not to form we must have

$$\frac{1}{2}E_{N_2} - \mu_N > 0, \tag{2.8}$$

and

$$E_{\alpha-Ga} - \mu_{Ga} > 0. \tag{2.9}$$

Using the relationship between μ_{Ga} and μ_N, these can be written as

$$E_{GaN} - \frac{1}{2}E_{N_2} < \mu_{Ga} < E_{\alpha-Ga}, \tag{2.10}$$

which defines the allowed range of values for μ_{Ga}. Note that the size of this range is equal to the formation energy of GaN, so the value of this quantity is of significance here.

2.4 GaN Surface Calculations

2.4.1 Surface Supercells

In order to perform calculations on surfaces within a plane wave periodic framework, we need to define a *surface supercell.* This is because a surface, although it may be periodic in the horizontal direction, is not periodic in the vertical. We have to define a periodic system that emulates the non-periodic system that we wish to study. This is done by using a "slab" of material that is thick enough that its surface behaves

in the same way as the surface of an infinitely thick slab. The slab is separated from its periodic repetitions by a region of vacuum. This vacuum region must again be sufficiently large that the surface behaves in the same way as it would with an infinitely large vacuum region. This is illustrated in Figure 2.11

The calculations on this work are on the (0001) surface (labelled according to the crystallographic index system [65]) as this is the surface most commonly used for growth. The clean surface is created by cleaving the crystal along a plane lying perpendicular to the c-direction, passing through the vertically oriented Ga-N bonds.

In creating a surface in a periodic cell, we inevitably create a second surface on the other side of the slab. We usually want this second surface to have as little interaction as possible with the "active" surface that we wish to study. To this end the dangling bonds on this second surface are usually "passivated" by attaching hydrogen atoms to them. Also, it is common practice to fix the positions of atoms to their positions in bulk material a certain depth into the surface. This is to prevent relaxation of atoms near to the other surface, which could have an undesired affect on the surface under study, and also speeds up the geometry optimisation process by having fewer degrees of freedom.

In order to determine the appropriate slab thickness and vacuum region size it is necessary to perform convergence tests of the energy with respect to these parameters. In order to obtain energy differences between different surface reconstructions that are converged with an accuracy of ~ 0.01eV per atom, we use a slab thickness of 6 GaN bilayers, a vacuum region of $\sim$ 14Å (for the clean surface), and we fix atoms lying below the highest two bilayers of material.

Surface reconstructions may alter the periodicity of the surface. A clean, unrelaxed, surface, i.e. one in which the atoms are in the same positions as they would be in the bulk material, is defined to have 1×1 periodicity. Reconstructions may increase the size of the primitive unit cell that can be used to define the periodic surface structure. The shape of the new unit cell can usually be defined in terms of the 1×1 cell. For example, most of the reconstructions we will consider have 2×2 periodicity - i.e. a 2×2 repetition of the shape of the 1×1 cell.

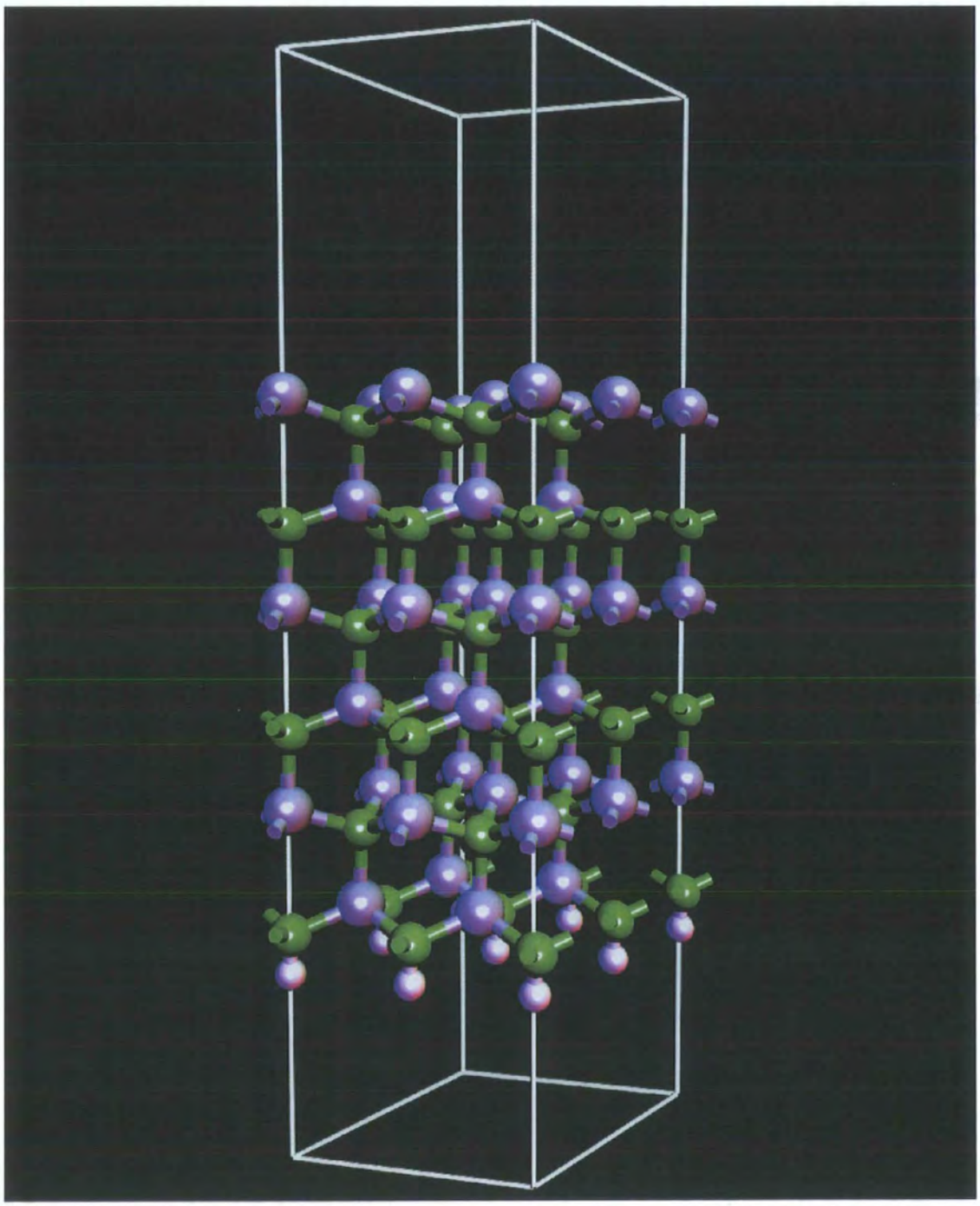


Figure 2.11: 2×2 Surface supercell for a clean wurtzite GaN (0001) surface.

2.4.2 Reconstructions in the Presence of Hydrogen

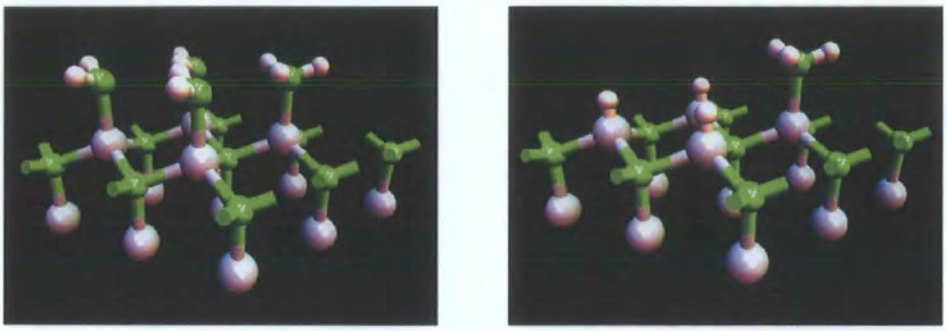
Following the work of reference [52], we consider eight different reconstructions which may occur in the presence of hydrogen, as shown in Figure 2.12. Seven of these have 2×2 periodicity, while one has $\sqrt{3} \times \sqrt{3}$ periodicity. Most of these reconstructions involve the additions of combinations of adatoms and chemical groups onto the clean surface. The $\sqrt{3} \times \sqrt{3}$ reconstruction involves the addition of a bilayer of Ga, which represents the laterally contracted bilayer model [50], that is believed to describe the surface under Ga-rich conditions.

2.4.3 Results: Phase Diagram

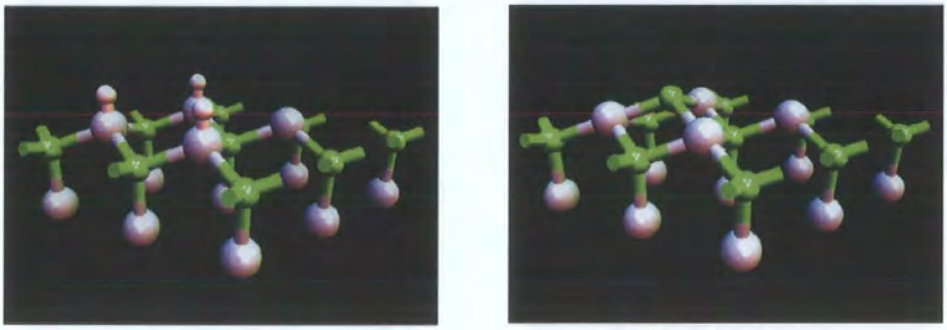
We have performed geometry optimisation on all of the reconstructions under consideration, obtaining the total energy of the relaxed structure in each case. Using Equation (2.5), we obtain the phase diagram shown in Figure 2.13.

This phase diagram is 2-dimensional, reflecting the fact that we effectively have two free parameters, which are the chemical potentials, μ_{H} and μ_{Ga}. The higher the chemical potential for a given element, the more "rich" the environment is said to be in that element. Hence, as we move up towards the top of the diagram, the environment is becoming more H-rich, while as we move downward, the environment is becoming H-poor. Likewise, the environment becomes more Ga-rich as we move towards the right of the diagram, and Ga-poor as we move to the left. Because of the relationship between μ_{N} and μ_{Ga}, Ga-rich conditions correspond to N-poor conditions and vice versa, hence the left of the diagram represents N-rich conditions, while the right represents N-poor conditions.

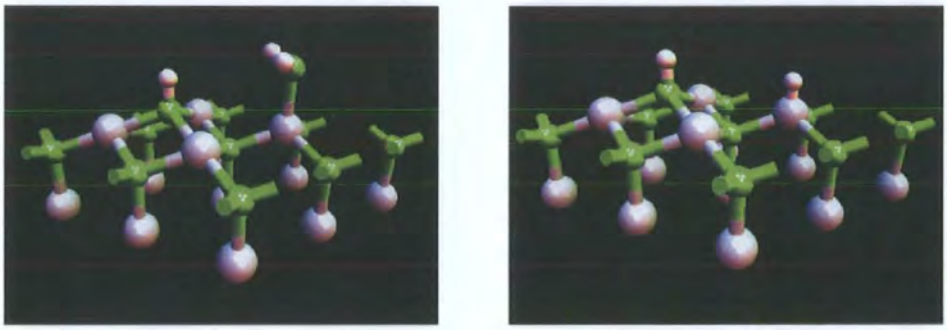
This explains the locations various reconstructions in the phase diagram. The top half is dominated by the $NH_3 + 3NH_2$ reconstruction; this is the reconstruction that contains the largest amount of hydrogen, corresponding to the H-rich conditions in this region. As we move downward towards more H-poor conditions, the reconstructions tend to contain fewer H atoms, with the reconstructions that contain no hydrogen at all occurring at the bottom. In a similar way, as we move left towards



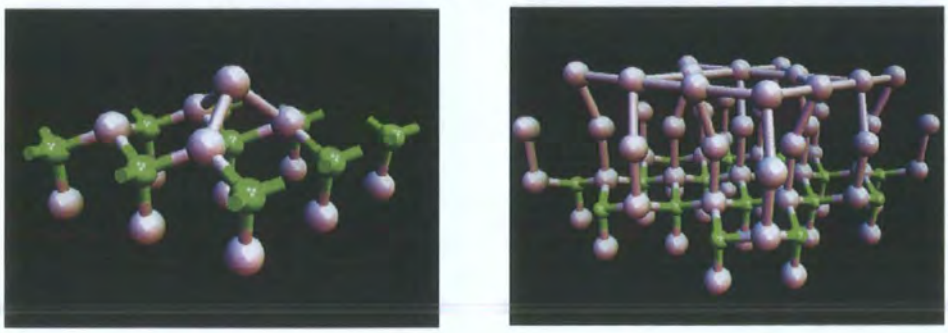
$NH_3 + 3NH_2$ $NH_3 + 3Ga\text{-}H$



$3Ga\text{-}H$ N_{ad}



$N_{ad}\text{-}H + NH_2$ $N_{ad}\text{-}H + Ga\text{-}H$



Ga_{ad} $Ga_{bilayer}$

Figure 2.12: Reconstructions of the GaN (0001) surface in the presence of H. Ga atoms are represented by large grey spheres, N atoms by green spheres, and H atoms by small white spheres.

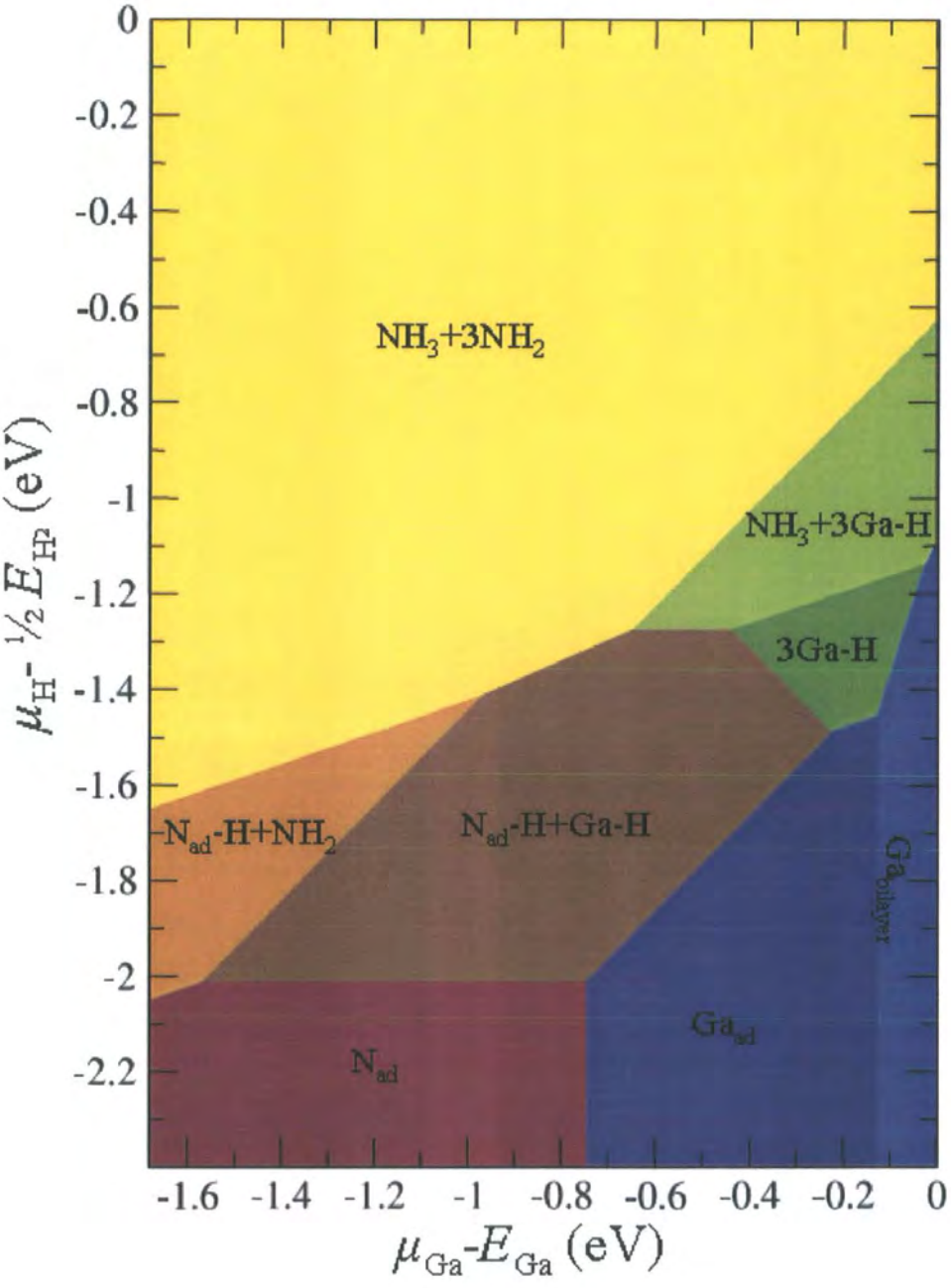


Figure 2.13: Surface phase diagram for reconstructions of the (0001) surface of wurtzite GaN in the presence of hydrogen, calculated using the LDA, as a function of the Ga and H chemical potentials.

more N-rich conditions we tend to encounter reconstructions with larger numbers of N-atoms, while as we move right towards Ga-rich conditions we encounter reconstructions with larger numbers of Ga atoms.

The boundaries of the diagram are determined by the allowed ranges of chemical potentials, as discussed earlier. At left-hand boundary, the conditions become so N-rich that N_2 molecules would spontaneously form, while at the right hand boundary the conditions become so Ga-rich that bulk α-Ga would form. The beginnings of this can be seen in the presence of the Ga-bilayer reconstruction, which is like a thin layer of metallic Ga. At the upper boundary, conditions become so H-rich that H_2 molecules would spontaneously form, however there is no lower limit to the H chemical potential (the lower boundary of the diagram is set arbitrarily). The bottom of the diagram essentially corresponds to conditions in which no hydrogen is present.

In terms of comparison with experiment, it is believed [52] that the transition between the N_{ad}-H + Ga-H and 3Ga-H reconstructions corresponds to a transition observed in growth experiments [66]. This is an example of how the combination of experiment and first principles calculations have been able to establish the details of atomic structures of different surface reconstructions during growth.

2.5 Summary and Conclusions

In this chapter we have described some of the background motivation for studying GaN, and carried out a series of calculations on this material. We established the appropriate basis set parameters for such calculations via a series of convergence tests, and calculated the LDA geometry of both the wurtzite and zinc blende GaN structures. The full LDA band structure was then calculated and compared to available experimental data. It was found that the LDA severely underestimates the band gap; this is a general property of the LDA, and so in order to obtain improved results from Kohn-Sham DFT we must consider using more advanced functionals.

As well as the bulk geometry and band structure, we also carried out calculations of surface reconstructions in the presence of hydrogen. This allowed us to produce a phase diagram for the different reconstructions as a function of the chemical potentials. This showed how the reconstructions may change depending on growth conditions, which may ultimately affect how the crystal grows in terms of the quality of the material produced. The phase diagram agrees closely with the work of other groups [52], however the range of the chemical potential is much wider in our results due to the larger energy of formation of GaN that we calculate. Our value for the formation energy seems to agree more closely with both all-electron calculations [63] and recent experiments [62].

In general, the LDA performs well in terms of geometric properties, but less well in terms of binding energies, and band structures. While the successes of the LDA can largely be ascribed to the fact that it satisfies the sum rule [67], its failures lie in the fact that the self-interaction correction in the Hartree energy is not fully corrected and that it does not describe the discontinuity in the exchange-correlation potential. These issues are discussed in more detail later in Section 5.3.

The inability of the LDA to predict band structures correctly provides the main motivation for the remainder of this work, which is concerned with the implementation and application of advanced functionals beyond the LDA.

Chapter 3

Theory of Non-Local Functionals

All the calculations we have performed so far have used the LDA to treat exchange and correlation. While the LDA was found to perform well in the calculation of geometric properties, and certain energetic properties, it failed to give an adequate description of the band structure, underestimating the band gap of GaN quite severely. In general, the LDA is known to systematically underestimate band-gaps, underestimate bond-lengths and lattice constants, and overestimate binding energies. Use of GGAs, while improving results in certain cases, does not perform much better than the LDA overall [26]. Band gaps are still systematically underestimated, while bond lengths and lattice constants are now overestimated, and binding energies underestimated. The reasons for these problems stem mostly from the *local* nature of the functionals, as will be discussed later in Section 5.3, and the only way to fix them, within a DFT approach, is therefore to use more advanced functionals that incorporate *non-local* information. In this chapter we describe a number of such non-local exchange-correlation functionals, paying particular attention to the three functionals we have implemented computationally (see Chapter 4), namely screened exchange (sX-LDA), Hartree-Fock (HF), and exact exchange (EXX), but also describing briefly some other functionals that are also growing in popularity.

3.1 sX-LDA and HF

Screened exchange (sX-LDA) and Hartree-Fock (HF) are closely related functionals [68, 69, 70, 71]. HF was originally developed before DFT [72, 16], and is in widespread use in the organic chemistry community. In the context of this work, it can be considered to be an implicit density functional within a generalised Kohn-Sham (GKS) framework. Both HF and sX-LDA involve a generalisation of the basic Kohn-Sham formalism so that the exchange-correlation potential is a *non-local* operator. This means that the GKS orbitals are no longer eigenstates of a Hamiltonian with a purely local potential.

The standard Kohn-Sham orbitals are the lowest eigenvalue solutions of an equation of the form,

$$-\frac{1}{2}\nabla^2\phi_i(\mathbf{r}) + \mu^{loc}(\mathbf{r})\phi_i(\mathbf{r}) = \varepsilon_i\phi_i(\mathbf{r}), \tag{3.1}$$

where the local potential, $\mu^{loc}(\mathbf{r})$, is defined such that the orbitals generate the correct ground state density, $\rho(\mathbf{r})$. The GKS orbitals are the lowest eigenvalue solutions of an equation of the form

$$-\frac{1}{2}\nabla^2\phi_i(\mathbf{r}) + \mu^{loc}(\mathbf{r})\phi_i(\mathbf{r}) + \int d\mathbf{r}' V^{NL}(\mathbf{r},\mathbf{r}')\phi_i(\mathbf{r}') = \varepsilon_i\phi_i(\mathbf{r}), \tag{3.2}$$

where $V^{NL}(\mathbf{r},\mathbf{r}')$ is a non-local *integral operator*, and, importantly is a *direct functional* of the orbitals. The local potential, $\mu^{loc}(\mathbf{r})$, is again defined such that the orbitals generate the correct ground state density, $\rho(\mathbf{r})$. How $V^{NL}(\mathbf{r},\mathbf{r}')$ is calculated from the orbitals depends on whether we are using HF or sX-LDA, as we shall see in a moment.

3.1.1 Definition of the Energy and Potential

We begin with the original definition of the exchange energy in terms of the Kohn-Sham orbitals, i.e.

$$V_X = -\frac{1}{2}\sum_{\mathbf{ikjq}}\int\int d\mathbf{r}d\mathbf{r}'\frac{\phi^*_{\mathbf{ik}}(\mathbf{r})\phi_{\mathbf{ik}}(\mathbf{r}')\phi^*_{\mathbf{jq}}(\mathbf{r}')\phi_{\mathbf{jq}}(\mathbf{r})}{|\mathbf{r}-\mathbf{r}'|}, \tag{3.3}$$

where j and $\mathbf{q}$ label bands and $\mathbf{k}$-points in the same way that i and $\mathbf{k}$ do. In HF, this is used to calculate the exchange energy, with the orbitals now being GKS orbitals

rather than KS orbitals. There is no treatment of correlation within standard HF so the total energy is given by

$$E_{TOT}^{HF} = T_S + V_H + V_{ext} + V_{I-I} + V_X, \tag{3.4}$$

where T_S is now calculated from the GKS orbitals.

While HF essentially treats exchange exactly, but does not treat correlation at all, sX-LDA attempts to incorporate some of the effects of correlation into Equation (3.3). One of the effects of correlation is effectively to screen the effect of exchange at long range. This can be achieved in a simple manner by multiplying the integrand of the exchange energy by a factor that decays exponentially with increasing electron-electron separation [69], i.e.

$$E_{XC}^{NL} = -\frac{1}{2}\sum_{\mathbf{ikjq}} \int\int d\mathbf{r}d\mathbf{r}' \frac{\phi_{\mathbf{ik}}^*(\mathbf{r})\phi_{\mathbf{ik}}(\mathbf{r}')\phi_{\mathbf{jq}}^*(\mathbf{r}')\phi_{\mathbf{jq}}(\mathbf{r})}{|\mathbf{r}-\mathbf{r}'|} e^{-k_s|\mathbf{r}-\mathbf{r}'|}. \tag{3.5}$$

Where k_s is the *reciprocal screening length.* The details of how the value k_s can be determined from first principles will be discussed later in 3.1.3.

In sX-LDA, there is also a local contribution to the exchange-correlation energy, in the spirit of the LDA, where the local exchange-correlation energy per electron is now given by

$$\varepsilon_{XC}^{loc}(\mathbf{r}) = \varepsilon_{XC}^{HEG}\left(\rho(\mathbf{r})\right) - \varepsilon_{XC}^{nlHEG}\left(\rho(\mathbf{r})\right), \tag{3.6}$$

where $\varepsilon_{XC}^{nlHEG}(\rho)$ is the non-local exchange-correlation energy per electron evaluated in an homogeneous electron gas of density ρ. The calculation of this function will be discussed later. The total energy functional for screened exchange is:

$$E_{TOT}^{SX-LDA} = T_S + V_H + V_{ext} + V_{I-I} + E_{XC}^{NL} + E_{XC}^{LOC}. \tag{3.7}$$

HF can almost be viewed as a special case of sX-LDA, i.e. the case of $k_s = 0$, except that we don't include correlation. For this reason we will focus most of this section on the screened exchange functional, as everything we say about sX-LDA will apply equally to HF.

Now, as we have mentioned, the GKS orbitals are eigenstates of a Hamiltonian for which there is a non-local integral operator component in the potential. This

potential operator is defined such that the sum of its expectation values yields the non-local exchange-correlation energy, E_{XC}^{NL}, i.e.

$$\begin{aligned} E_{XC}^{NL} &= \sum_{\mathbf{ik}} \langle i\mathbf{k}|\hat{V}_{XC}^{NL}|i\mathbf{k}\rangle \\ &= \sum_{\mathbf{ik}} \int d\mathbf{r}\phi_{\mathbf{ik}}^{*}(\mathbf{r}) \int d\mathbf{r}' V_{XC}^{NL}(\mathbf{r},\mathbf{r}')\phi_{\mathbf{ik}}(\mathbf{r}'). \end{aligned} \tag{3.8}$$

Comparing this with Equation (3.5) we see that the potential is given by

$$V_{XC}^{NL}(\mathbf{r},\mathbf{r}') = -\frac{1}{2}\sum_{j\mathbf{q}} \frac{\phi_{j\mathbf{q}}(\mathbf{r})\phi_{j\mathbf{q}}^{*}(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} e^{-k_s|\mathbf{r}-\mathbf{r}'|}. \tag{3.9}$$

3.1.2 Minimisation within the GKS Framework

Minimisation of the total energy in HF or sX-LDA calculations proceeds in much the same way as in the standard Kohn-Sham framework. A self-consistent cycle can still be defined, the only difference being that part of the potential now depends directly on the orbitals, rather than on the density, i.e. we are looking for solutions of

$$\begin{aligned} -\frac{1}{2}\nabla^2\phi_i(\mathbf{r}) + v_{ext}(\mathbf{r})\phi_i(\mathbf{r}) + v_H[\rho](\mathbf{r})\phi_i(\mathbf{r}) + \int d\mathbf{r}' V_{XC}^{NL}(\mathbf{r},\mathbf{r}')\phi_i(\mathbf{r}') \\ +\mu_{XC}^{LOC}[\rho](\mathbf{r})\phi_i(\mathbf{r}) = \varepsilon_i\phi_i(\mathbf{r}), \end{aligned} \tag{3.10}$$

which are the *GKS equations*. There is a slight complication here in that the non-local potential, $V_{XC}^{NL}(\mathbf{r},\mathbf{r}')$, depends on the orbitals, but it is not clear whether these should remain as the orbitals *prior* to solution of (3.10) within one cycle, or whether a solution should be sought in which $V_{XC}^{NL}(\mathbf{r},\mathbf{r}')$ is consistent with the final set of orbitals in that cycle. Of course, we will still end up with the same self-consistent solution either way, but which way we choose may well affect how many cycles are needed before we reach self-consistency. This issue will be discussed further in the context of the reciprocal space representation of these functionals in 3.1.4.

3.1.3 Screening Constants and the HEG

The screening constant, k_s, determines the effective range of the non-local exchange interaction, and can, in principle, be set anywhere in the range $[0,\infty)$. A value of

0 is equivalent to using HF with LDA treatment of correlation, while a value of infinity is equivalent to using the LDA alone. Ideally we would like to establish a value of k_s to use from first principles, and this can be done by using its relation to the density in the Thomas-Fermi model [73, 74, 75], i.e.

$$\begin{aligned} k_s &= 2\sqrt{\frac{k_F}{\pi}} \\ &= \frac{2}{\sqrt{\pi}}(3\pi^2\rho)^{\frac{1}{6}}. \end{aligned} \tag{3.11}$$

For an efficient implementation, the value of k_s must be constant for a given calculation, although it may be allowed to depend on, for example, the average density of the system under study. Alternatively, we can set a universal value for k_s by relating it to the "natural" density of the homogeneous electron gas, i.e. the density that minimises the total energy per electron. Using Perdew's parameterisation of the correlation energy [22], this gives a numerical value of k_s of about 0.764Bohr^{-1}.

The homogeneous electron gas is also used to parameterise the local exchange-correlation energy per electron, $\varepsilon_{XC}^{LOC}(\rho)$. This is given by

$$\varepsilon_{XC}^{LOC}(\rho) = \varepsilon_{XC}^{HEG}(\rho) - \varepsilon_{XC}^{NL,HEG}(\rho), \tag{3.12}$$

the total exchange-correlation energy per electron, $\varepsilon_{XC}^{HEG}(\rho)$, is the same as in the LDA. The non-local part, $\varepsilon_{XC}^{NL,HEG}(\rho)$, is the non-local exchange-correlation energy per electron we would have if we were to apply Equation (3.5) to the HEG. This related to the pure exchange energy per electron, $v_X^{HEG}(\rho)$, by

$$\varepsilon_{XC}^{NL,HEG}(\rho) = v_X^{HEG}(\rho)F\left(\gamma(\rho)\right), \tag{3.13}$$

where $\gamma(\rho) = k_s/k_F(\rho)$, and $F(\gamma)$ is given by [76]

$$F(\gamma) = 1 - \frac{4}{3}\gamma \arctan\left(\frac{2}{\gamma}\right) - \frac{\gamma^2}{6}\left[1 - \left(\frac{\gamma^2}{4} + 3\right)\ln\left(1 + \frac{4}{\gamma^2}\right)\right]. \tag{3.14}$$

3.1.4 HF and sX-LDA in Reciprocal Space

As with the standard Kohn-Sham minimisation procedure, the GKS procedure can be performed in reciprocal space, leading to gains in computational efficiency. The

GKS equations in reciprocal space read

$$\sum_{\mathbf{G}'}\Big[\frac{1}{2}(\mathbf{k}+\mathbf{G})^2\delta_{\mathbf{G}\mathbf{G}'}+v_{ext}(\mathbf{G}-\mathbf{G}')+\mu_H(\mathbf{G}-\mathbf{G}') \\ +V^{NL}_{\mathbf{k}XC}(\mathbf{G},\mathbf{G}')+\mu^{LOC}_{XC}(\mathbf{G}-\mathbf{G}')\Big]c_{\mathbf{i}\mathbf{k}}(\mathbf{G}') = \varepsilon_{\mathbf{i}\mathbf{k}}c_{\mathbf{i}\mathbf{k}}(\mathbf{G}), \tag{3.15}$$

where $V^{NL}_{\mathbf{k}XC}(\mathbf{G},\mathbf{G}')$ is the reciprocal space representation of the non-local potential operator. This can be obtained from the real space representation as follows:

$$\begin{aligned} V^{NL}_{\mathbf{k}XC}(\mathbf{G},\mathbf{G}') &= \mathrm{FT}\left[V^{NL}_{XC}(\mathbf{r},\mathbf{r}')\right]\mathrm{FT}^{-1} \\ &= \frac{1}{\Omega}\int d\mathbf{r}e^{-i(\mathbf{k}+\mathbf{G})\cdot\mathbf{r}}\int d\mathbf{r}'V^{NL}_{XC}(\mathbf{r},\mathbf{r}')e^{i(\mathbf{k}+\mathbf{G}')\cdot\mathbf{r}'} \\ &= -\frac{1}{2\Omega}\sum_{j\mathbf{q}}\int d\mathbf{r}e^{-i(\mathbf{k}+\mathbf{G})\cdot\mathbf{r}}\phi_{j\mathbf{q}}(\mathbf{r})\int d\mathbf{r}'\frac{\phi^*_{j\mathbf{q}}(\mathbf{r}')e^{-k_s|\mathbf{r}-\mathbf{r}'|}}{|\mathbf{r}-\mathbf{r}'|}e^{i(\mathbf{k}+\mathbf{G}')\cdot\mathbf{r}'}. \end{aligned} \tag{3.16}$$

Taking the second integral, and substituting the reciprocal space representation of $\phi^*_{j\mathbf{q}}(\mathbf{r}')$, we have

$$\int d\mathbf{r}'\frac{\phi^*_{j\mathbf{q}}(\mathbf{r}')e^{-k_s|\mathbf{r}-\mathbf{r}'|}}{|\mathbf{r}-\mathbf{r}'|}e^{i(\mathbf{k}+\mathbf{G}')\cdot\mathbf{r}'} = \\ \frac{1}{\sqrt{\Omega}}\int d\mathbf{r}'\sum_{\mathbf{G}''}c^*_{j\mathbf{q}}(\mathbf{G}'')e^{-i(\mathbf{q}+\mathbf{G}'')\cdot\mathbf{r}'}\frac{e^{-k_s|\mathbf{r}-\mathbf{r}'|}}{|\mathbf{r}-\mathbf{r}'|}e^{i(\mathbf{k}+\mathbf{G}')\cdot\mathbf{r}'} \\ = \frac{1}{\sqrt{\Omega}}\sum_{\mathbf{G}''}c^*_{j\mathbf{q}}(\mathbf{G}'')\int d\mathbf{r}'\frac{e^{-k_s|\mathbf{r}-\mathbf{r}'|}}{|\mathbf{r}-\mathbf{r}'|}e^{i(\mathbf{k}-\mathbf{q}+\mathbf{G}'-\mathbf{G}'')\cdot\mathbf{r}'} \\ = \frac{4\pi}{\sqrt{\Omega}}\sum_{\mathbf{G}''}\frac{c^*_{j\mathbf{q}}(\mathbf{G}'')}{|\mathbf{k}-\mathbf{q}+\mathbf{G}'-\mathbf{G}''|^2+k_s^2}e^{i(\mathbf{k}-\mathbf{q}+\mathbf{G}'-\mathbf{G}'')\cdot\mathbf{r}}, \tag{3.17}$$

where we have used the following identity:

$$\int d\mathbf{r}\frac{e^{-k|\mathbf{r}|+i\mathbf{K}\cdot\mathbf{r}}}{|\mathbf{r}|} \equiv \frac{4\pi}{|\mathbf{K}|^2+k^2}. \tag{3.18}$$

Putting (3.17) back into Equation (3.16), we have

$$\begin{aligned} V^{NL}_{\mathbf{k}XC}(\mathbf{G},\mathbf{G}') &= -\frac{2\pi}{\Omega^{\frac{3}{2}}}\sum_{j\mathbf{q}}\int d\mathbf{r} e^{-i(\mathbf{k}+\mathbf{G})\cdot\mathbf{r}}\phi_{j\mathbf{q}}(\mathbf{r}) \\ &\quad \times \sum_{\mathbf{G}''}\frac{c^*_{j\mathbf{q}}(\mathbf{G}'')}{|\mathbf{k}-\mathbf{q}+\mathbf{G}'-\mathbf{G}''|^2+k_s^2}e^{i(\mathbf{k}-\mathbf{q}+\mathbf{G}'-\mathbf{G}'')\cdot\mathbf{r}} \\ &= -\frac{2\pi}{\Omega^{\frac{3}{2}}}\sum_{j\mathbf{q}}\sum_{\mathbf{G}''}\frac{c^*_{j\mathbf{q}}(\mathbf{G}'')}{|\mathbf{k}-\mathbf{q}+\mathbf{G}'-\mathbf{G}''|^2+k_s^2} \\ &\quad \times \int d\mathbf{r}\phi_{j\mathbf{q}}(\mathbf{r})e^{-i(\mathbf{q}+\mathbf{G}-\mathbf{G}'+\mathbf{G}'')}. \end{aligned} \tag{3.19}$$

Recognising the integral here as being essentially the Fourier transform of $\phi_{j\mathbf{q}}(\mathbf{r})$, we have

$$\begin{aligned} V^{NL}_{\mathbf{k}XC}(\mathbf{G},\mathbf{G}') &= -\frac{2\pi}{\Omega}\sum_{j\mathbf{q}}\sum_{\mathbf{G}''}\frac{c^*_{j\mathbf{q}}(\mathbf{G}'')c_{j\mathbf{q}}(\mathbf{G}-\mathbf{G}'+\mathbf{G}'')}{|\mathbf{k}-\mathbf{q}+\mathbf{G}'-\mathbf{G}''|^2+k_s^2} \\ &= -\frac{2\pi}{\Omega}\sum_{j\mathbf{q}}\sum_{\mathbf{G}''}\frac{c^*_{j\mathbf{q}}(\mathbf{G}'+\mathbf{G}'')c_{j\mathbf{q}}(\mathbf{G}+\mathbf{G}'')}{|\mathbf{q}-\mathbf{k}+\mathbf{G}''|^2+k_s^2}, \end{aligned} \tag{3.20}$$

which is the standard expression for $V^{NL}_{\mathbf{k}XC}(\mathbf{G},\mathbf{G}')$ in reciprocal space.

Since the non-local part of the exchange-correlation energy, E^{NL}_{XC}, is the expectation value of this operator, we can obtain an expression for E^{NL}_{XC} in reciprocal space as follows:

$$\begin{aligned} E^{NL}_{XC} &= \sum_{i\mathbf{k}}\langle i\mathbf{k}|\hat{V}^{NL}_{XC}|i\mathbf{k}\rangle \\ &= \sum_{i\mathbf{k}}\sum_{\mathbf{G}}c^*_{i\mathbf{k}}(\mathbf{G})\sum_{\mathbf{G}'}V^{NL}_{\mathbf{k}XC}(\mathbf{G},\mathbf{G}')c_{i\mathbf{k}}(\mathbf{G}') \\ &= -\frac{2\pi}{\Omega}\sum_{i\mathbf{k}j\mathbf{q}}\sum_{\mathbf{G}\mathbf{G}'\mathbf{G}''}\frac{c^*_{i\mathbf{k}}(\mathbf{G})c_{i\mathbf{k}}(\mathbf{G}')c^*_{j\mathbf{q}}(\mathbf{G}'+\mathbf{G}'')c_{j\mathbf{q}}(\mathbf{G}+\mathbf{G}'')}{|\mathbf{q}-\mathbf{k}+\mathbf{G}''|^2+k_s^2}, \end{aligned} \tag{3.21}$$

which is the standard expression for E^{NL}_{XC} in reciprocal space.

Solution of the GKS equations can be carried out in a similar way to the solution of the standard Kohn-Sham equations, i.e. by either diagonalising the matrix directly, or by minimising the expectation values with respect to the orbital coefficients.

If we were to diagonalise the matrix, then we would have to construct the entire matrix, with each element including a component from $V^{NL}_{\mathbf{k}XC}(\mathbf{G},\mathbf{G}')$. So, effectively we would be calculating the entire non-local exchange-correlation operator. As can be seen, if we were to calculate this object following Equation (3.20), we would require a triple-loop over **G**-vectors for each band and **k**-point in the system. Hence we would expect the calculation to scale as $N_p^3 N_b N_k$, which would put severe limits on the size of system we could reasonably expect to deal with. Fortunately there does exist a more efficient formulation for solving the GKS equations, and calculating the energy, as we will now describe.

3.1.5 Efficient Procedure for HF and sX-LDA

The following method is based on the work of Chawla and Voth [71], however we have extended it to include multiple **k**-points, and screening. We start with the standard expression for E^{NL}_{XC} in reciprocal space, as given by Equation (3.21), and note that it can be re-written in the following form:

$$E^{NL}_{XC} = -\frac{2\pi}{\Omega}\sum_{\mathbf{ikjq}}\sum_{\mathbf{G}''}\frac{\sum_{\mathbf{G}} c^*_{\mathbf{ik}}(\mathbf{G})c_{\mathbf{jq}}(\mathbf{G}+\mathbf{G}'')\sum_{\mathbf{G}'} c_{\mathbf{ik}}(\mathbf{G}')c^*_{\mathbf{jq}}(\mathbf{G}'+\mathbf{G}'')}{|\mathbf{q}-\mathbf{k}+\mathbf{G}''|^2+k_s^2}. \tag{3.22}$$

We then define the correlation function, $C_{\mathbf{jqik}}(\mathbf{G})$, as

$$C_{\mathbf{jqik}}(\mathbf{G}) = \sum_{\mathbf{G}'} c^*_{\mathbf{ik}}(\mathbf{G}')c_{\mathbf{jq}}(\mathbf{G}+\mathbf{G}'), \tag{3.23}$$

which allows us to write Equation (3.22) as

$$E^{NL}_{XC} = -\frac{2\pi}{\Omega}\sum_{\mathbf{ikjq}}\sum_{\mathbf{G}}\frac{|C_{\mathbf{jqik}}(\mathbf{G})|^2}{|\mathbf{q}-\mathbf{k}+\mathbf{G}|^2+k_s^2}. \tag{3.24}$$

Now, if we take Equation (3.23) and substitute in the real space representations of $c^*_{\mathbf{ik}}(\mathbf{G}')$ and $c_{\mathbf{jq}}(\mathbf{G}+\mathbf{G}')$, we have

$$
\begin{aligned}
C_{\mathbf{jqik}}(\mathbf{G}) &= \frac{1}{\Omega}\sum_{\mathbf{G}'}\int d\mathbf{r}\phi^*_{\mathbf{ik}}(\mathbf{r})e^{i(\mathbf{k}+\mathbf{G}')\cdot\mathbf{r}}\int d\mathbf{r}'\phi_{\mathbf{jq}}(\mathbf{r}')e^{-i(\mathbf{q}+\mathbf{G}+\mathbf{G}')\cdot\mathbf{r}'} \\
&= \frac{1}{\Omega}\int d\mathbf{r}\phi^*_{\mathbf{ik}}(\mathbf{r})e^{i\mathbf{k}\cdot\mathbf{r}}\int d\mathbf{r}'\phi_{\mathbf{jq}}(\mathbf{r}')e^{-i(\mathbf{q}+\mathbf{G})\cdot\mathbf{r}'}\sum_{\mathbf{G}'}e^{i\mathbf{G}'\cdot(\mathbf{r}-\mathbf{r}')} \\
&= \int d\mathbf{r}\phi^*_{\mathbf{ik}}(\mathbf{r})e^{i\mathbf{k}\cdot\mathbf{r}}\int d\mathbf{r}'\phi_{\mathbf{jq}}(\mathbf{r}')e^{-i(\mathbf{q}+\mathbf{G})\cdot\mathbf{r}'}\delta(\mathbf{r}-\mathbf{r}') \\
&= \int d\mathbf{r}\phi^*_{\mathbf{ik}}(\mathbf{r})\phi_{\mathbf{jq}}(\mathbf{r})e^{-i(\mathbf{q}-\mathbf{k}+\mathbf{G})\cdot\mathbf{r}}, \qquad (3.25)
\end{aligned}
$$

which is essentially just the Fourier transform of a product of orbitals, i.e.

$$C_{\mathbf{jqik}}(\mathbf{G}) = \sqrt{\Omega}\mathrm{FT}[\phi^*_{\mathbf{ik}}(\mathbf{r})\phi_{\mathbf{jq}}(\mathbf{r})]. \qquad (3.26)$$

This means that in order to evaluate $C_{\mathbf{jqik}}(\mathbf{G})$ for given a set of orbitals in reciprocal space, $c_{\mathbf{ik}}(\mathbf{G})$, we inverse-Fourier transform the orbitals to real space, multiply each orbital with every other orbital, and Fourier-transform these products back to reciprocal space according to Equation (3.26). We then insert the $C_{\mathbf{jqik}}(\mathbf{G})$ into Equation (3.24) to calculate E_{XC}^{NL}. By making use of fast Fourier transforms (FFTs), each Fourier transform scales as $N_p\log(N_p)$, rather than N_p^2, which increases the speed of the calculation significantly. The most expensive part of the process is Fourier transforming the products; and we need to perform $N_k^2N_b^2$ Fourier transforms, so the overall scaling for the calculation of E_{XC}^{NL} is $N_p\log(N_p)N_k^2N_b^2$.

So we now have an efficient method of calculating the non-local exchange-correlation energy, but in order to solve the GKS equations, we also need either to evaluate and diagonalise the full Hamiltonian matrix, including the contribution from $V_{XC}^{NL}(\mathbf{G},\mathbf{G}')$, or minimise the eigenvalue sum by evaluating its *gradient* with respect to the orbital coefficients, including the contribution from E_{XC}^{NL}. If we wish to avoid the unfavourable scaling involved in calculating $V_{XC}^{NL}(\mathbf{G},\mathbf{G}')$ we must look to the latter approach. We will start by considering the gradient in real space, i.e.

$$
\begin{aligned}
\frac{\delta E_{XC}^{NL}}{\delta\phi^*_{\mathbf{ik}}(\mathbf{r})} &= -\frac{1}{2}\frac{\delta}{\delta\phi^*_{\mathbf{ik}}(\mathbf{r})}\sum_{\mathbf{i'k'jq}}\int\int d\mathbf{r}'d\mathbf{r}''\frac{\phi^*_{\mathbf{i'k'}}(\mathbf{r}')\phi_{\mathbf{i'k'}}(\mathbf{r}'')\phi^*_{\mathbf{jq}}(\mathbf{r}'')\phi_{\mathbf{jq}}(\mathbf{r}')}{|\mathbf{r}'-\mathbf{r}''|}e^{-k_s|\mathbf{r}'-\mathbf{r}''|} \\
&= -\sum_{\mathbf{jq}}\phi_{\mathbf{jq}}(\mathbf{r})\int d\mathbf{r}'\frac{\phi_{\mathbf{ik}}(\mathbf{r}')\phi^*_{\mathbf{jq}}(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}e^{-k_s|\mathbf{r}-\mathbf{r}'|}. \qquad (3.27)
\end{aligned}
$$

Taking the integral, and substituting the reciprocal space representations of $\phi_{\mathbf{ik}}(\mathbf{r}')$ and $\phi^*_{\mathbf{jq}}(\mathbf{r}')$, we have

$$\int d\mathbf{r}' \frac{\phi_{\mathbf{ik}}(\mathbf{r}')\phi^*_{\mathbf{jq}}(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} e^{-k_s|\mathbf{r}-\mathbf{r}'|} =$$

$$\frac{1}{\Omega}\int d\mathbf{r}' \sum_{\mathbf{G}} c_{\mathbf{ik}}(\mathbf{G}) e^{i(\mathbf{k}+\mathbf{G})\cdot\mathbf{r}'} \sum_{\mathbf{G}'} c^*_{\mathbf{jq}}(\mathbf{G}') e^{-i(\mathbf{q}+\mathbf{G}')\cdot\mathbf{r}'} \frac{e^{-k_s|\mathbf{r}-\mathbf{r}'|}}{|\mathbf{r}-\mathbf{r}'|}$$

$$= \frac{1}{\Omega}\sum_{\mathbf{G}\mathbf{G}'} c_{\mathbf{ik}}(\mathbf{G}) c^*_{\mathbf{jq}}(\mathbf{G}') \int d\mathbf{r}' \frac{e^{-k_s|\mathbf{r}-\mathbf{r}'|}}{|\mathbf{r}-\mathbf{r}'|} e^{i(\mathbf{k}-\mathbf{q}+\mathbf{G}-\mathbf{G}')\cdot\mathbf{r}'},$$

which, using the identity of Equation (3.18), becomes

$$\frac{4\pi}{\Omega}\sum_{\mathbf{G}\mathbf{G}'} \frac{c_{\mathbf{ik}}(\mathbf{G}) c^*_{\mathbf{jq}}(\mathbf{G}')}{|\mathbf{k}-\mathbf{q}+\mathbf{G}-\mathbf{G}'|^2 + k_s^2} e^{i(\mathbf{k}-\mathbf{q}+\mathbf{G}-\mathbf{G}')\cdot\mathbf{r}}$$

$$= \frac{4\pi}{\Omega}\sum_{\mathbf{G}\mathbf{G}'} \frac{c_{\mathbf{ik}}(\mathbf{G}) c^*_{\mathbf{jq}}(\mathbf{G}+\mathbf{G}')}{|\mathbf{q}-\mathbf{k}+\mathbf{G}'|^2 + k_s^2} e^{i(\mathbf{k}-\mathbf{q}-\mathbf{G}')\cdot\mathbf{r}}$$

$$= \frac{4\pi}{\Omega}\sum_{\mathbf{G}'} \frac{C^*_{\mathbf{jqik}}(\mathbf{G}')}{|\mathbf{q}-\mathbf{k}+\mathbf{G}'|^2 + k_s^2} e^{i(\mathbf{k}-\mathbf{q}-\mathbf{G}')\cdot\mathbf{r}}$$

$$= \frac{4\pi}{\sqrt{\Omega}}\mathrm{FT}^{-1}\left[\frac{C^*_{\mathbf{jqik}}(\mathbf{G})}{|\mathbf{q}-\mathbf{k}+\mathbf{G}|^2 + k_s^2}\right] = f_{\mathbf{jqik}}(\mathbf{r}), \tag{3.28}$$

which defines a set of real space functions, $f_{\mathbf{jqik}}(\mathbf{r})$. Putting this back into Equation (3.27), we have

$$\frac{\delta E^{NL}_{XC}}{\delta \phi^*_{\mathbf{ik}}(\mathbf{r})} = -\sum_{\mathbf{jq}} \phi_{\mathbf{jq}}(\mathbf{r}) f_{\mathbf{jqik}}(\mathbf{r}). \tag{3.29}$$

Finally, we transform this into reciprocal space to obtain

$$\frac{\delta E^{NL}_{XC}}{\delta c^*_{\mathbf{ik}}(\mathbf{G})} = \int d\mathbf{r} \frac{\delta E^{NL}_{XC}}{\delta \phi^*_{\mathbf{ik}}(\mathbf{r})} e^{-i(\mathbf{k}+\mathbf{G})\cdot\mathbf{r}}$$

$$= \sqrt{\Omega}\mathrm{FT}\left[-\sum_{\mathbf{jq}} \phi_{\mathbf{jq}}(\mathbf{r}) f_{\mathbf{jqik}}(\mathbf{r})\right]. \tag{3.30}$$

This means that in order to calculate the gradient of E^{NL}_{XC} with respect to the orbital coefficients in reciprocal space, we basically have to inverse-Fourier transform each $C_{\mathbf{jqik}}(\mathbf{G})$, multiply by an orbital in real space, and then Fourier-transform back. The most expensive part of this procedure is the inverse-Fourier transform, which

must be done $N_b^2 N_k^2$ times. Hence, as was the case in the calculation of E_{XC}^{NL}, the calculation of its gradient also scales as $N_p \log(N_p) N_b^2 N_k^2$.

We therefore see that by formulating the theory in this way it should be possible to implement these functionals in a much more efficient manner than we might first suppose.

3.2 Exact Exchange

We will now describe the exact exchange (EXX) functional [77, 78, 79]. A common misconception amongst non-specialists is that HF and EXX are the same thing. It is true, of course, that both functionals treat exchange "exactly" in the sense that the exchange energy is given in terms of the orbitals by Equation (3.3). The difference arises from the fact the orbitals themselves are different - in EXX the orbitals are eigenstates of a Hamiltonian with a *local* potential, while in HF they are eigenstates of a Hamiltonian with a *non-local* potential. This leads to very different results, notably in the calculation of band structures, where, as we will see in Chapter 4, HF performs disastrously while EXX performs very well indeed [79]. EXX also includes treatment of correlation in the style of the LDA, which is not the case in standard HF. Although LDA correlation can be added to HF, this does not improve the band structure problem, the reason for which will become clear in the discussion in Section 5.3.

3.2.1 Definition of the Exact Exchange Potential

EXX is essentially standard Kohn-Sham theory, in which the exchange energy is defined exactly in terms of the Kohn-Sham orbitals. In standard Kohn-Sham theory the exchange potential is defined as

$$\mu_X(\mathbf{r}) = \frac{\delta V_X}{\delta \rho(\mathbf{r})}. \tag{3.31}$$

In the LDA or GGA, for example, V_X is a direct functional of the density, making evaluation of this functional derivative relatively straightforward. In EXX, however,

this is not the case. While V_X is still certainly a functional of the the density, because there is a one-to-one mapping between the density and the orbitals, there is no simple mathematical expression that gives the orbitals, and hence V_X, in terms of the density.

The procedure for calculating the exact exchange potential is a rather recent development [77, 78, 79]. It is based on the chain rule for functional derivatives, which leads to the following expression for $\mu_X(\mathbf{r})$ in real space:

$$\mu_X(\mathbf{r}) = \frac{\delta V_X}{\delta \rho(\mathbf{r})} = \sum_{v\mathbf{k}} \int d\mathbf{r}' \int d\mathbf{r}'' \left(\frac{\delta V_X}{\delta \phi_{v\mathbf{k}}(\mathbf{r}')} \frac{\delta \phi_{v\mathbf{k}}(\mathbf{r}')}{\delta \mu_{KS}(\mathbf{r}'')} + \text{c.c.} \right) \frac{\delta \mu_{KS}(\mathbf{r}'')}{\delta \rho(\mathbf{r})}. \tag{3.32}$$

Note that we are now using the subscript v to index the orbitals to distinguish them as valence band orbitals as opposed to conduction band orbitals, which we will also have to consider in a moment. There are three different functional derivatives on the right-hand side. The first one is quite straightforward, as it is simply the derivative of the exchange energy with respect to the orbitals, i.e.

$$\begin{aligned} \frac{\delta V_X}{\delta \phi_{v\mathbf{k}}(\mathbf{r}')} &= -\frac{1}{2} \frac{\delta}{\delta \phi_{v\mathbf{k}}(\mathbf{r}')} \sum_{v\mathbf{k}u\mathbf{q}} \int \int d\mathbf{r}''' d\mathbf{r}'''' \frac{\phi^*_{v\mathbf{k}}(\mathbf{r}''') \phi_{v\mathbf{k}}(\mathbf{r}'''') \phi^*_{u\mathbf{q}}(\mathbf{r}'''') \phi_{u\mathbf{q}}(\mathbf{r}''')}{|\mathbf{r}''' - \mathbf{r}''''|} \\ &= -\sum_{u\mathbf{q}} \int d\mathbf{r}''' \frac{\phi^*_{v\mathbf{k}}(\mathbf{r}''') \phi^*_{u\mathbf{q}}(\mathbf{r}') \phi_{u\mathbf{q}}(\mathbf{r}''')}{|\mathbf{r}' - \mathbf{r}'''|}. \end{aligned} \tag{3.33}$$

The second one is the functional derivative of each orbital with respect to the Kohn-Sham potential. This is essentially the first order response of the orbitals, $\delta\phi_{v\mathbf{k}}(\mathbf{r}')$, caused by a small change in the potential $\delta\mu_{KS}(\mathbf{r}'')$, and hence we can employ first order perturbation theory. This involves an expansion of $\delta\phi_{v\mathbf{k}}(\mathbf{r}')$ in terms of the *complete* set of eigenstates of the Kohn-Sham Hamiltonian, i.e.

$$\delta \phi_{v\mathbf{k}}(\mathbf{r}') = \sum_{n'\mathbf{k}' \neq v\mathbf{k}} \phi_{n'\mathbf{k}'}(\mathbf{r}') \frac{\int d\mathbf{r}'' \phi^*_{n'\mathbf{k}'}(\mathbf{r}'') \delta \mu_{KS}(\mathbf{r}'') \phi_{v\mathbf{k}}(\mathbf{r}'')}{\varepsilon_{v\mathbf{k}} - \varepsilon_{n'\mathbf{k}'}}, \tag{3.34}$$

where the sum over n' includes all eigenstates of the Kohn-Sham Hamiltonian, i.e. both valence *and* conduction bands. The equation for $\delta\phi_{v\mathbf{k}}(\mathbf{r}')/\delta\mu_{KS}(\mathbf{r}'')$ is thus

$$\frac{\delta \phi_{v\mathbf{k}}(\mathbf{r}')}{\delta \mu_{KS}(\mathbf{r}'')} = \sum_{n'\mathbf{k}' \neq v\mathbf{k}} \phi_{n'\mathbf{k}'}(\mathbf{r}') \frac{\phi^*_{n'\mathbf{k}'}(\mathbf{r}'') \phi_{v\mathbf{k}}(\mathbf{r}'')}{\varepsilon_{v\mathbf{k}} - \varepsilon_{n'\mathbf{k}'}}, \tag{3.35}$$

Combining the first two functional derivatives, we have

$$\int d\mathbf{r}' \frac{\delta V_X}{\delta\phi_{v\mathbf{k}}(\mathbf{r}')}\frac{\delta\phi_{v\mathbf{k}}(\mathbf{r}')}{\delta\mu_{KS}(\mathbf{r}'')} =$$

$$-\int d\mathbf{r}' \sum_{u\mathbf{q}} \int d\mathbf{r}''' \frac{\phi^*_{v\mathbf{k}}(\mathbf{r}''')\phi^*_{u\mathbf{q}}(\mathbf{r}')\phi_{u\mathbf{q}}(\mathbf{r}''')}{|\mathbf{r}' - \mathbf{r}'''|} \sum_{n'\mathbf{k}'\neq v\mathbf{k}} \phi_{n'\mathbf{k}'}(\mathbf{r}')\frac{\phi^*_{n'\mathbf{k}'}(\mathbf{r}'')\phi_{v\mathbf{k}}(\mathbf{r}'')}{\varepsilon_{v\mathbf{k}} - \varepsilon_{n'\mathbf{k}'}}$$

$$= 2\int d\mathbf{r}' \int d\mathbf{r}''' \phi^*_{v\mathbf{k}}(\mathbf{r}''')V_X^{NL}(\mathbf{r}''', \mathbf{r}') \sum_{n'\mathbf{k}'\neq v\mathbf{k}} \phi_{n'\mathbf{k}'}(\mathbf{r}')\frac{\phi^*_{n'\mathbf{k}'}(\mathbf{r}'')\phi_{v\mathbf{k}}(\mathbf{r}'')}{\varepsilon_{v\mathbf{k}} - \varepsilon_{n'\mathbf{k}'}}$$

$$= 2\sum_{n'\mathbf{k}'\neq v\mathbf{k}} \langle v\mathbf{k}|\hat{V}_X^{NL}|n'\mathbf{k}'\rangle \frac{\phi^*_{n'\mathbf{k}'}(\mathbf{r}'')\phi_{v\mathbf{k}}(\mathbf{r}'')}{\varepsilon_{v\mathbf{k}} - \varepsilon_{n'\mathbf{k}'}}. \tag{3.36}$$

We now insert this into Equation (3.32). Due to the fact that we are summing over v and $\mathbf{k}$, and adding the complex conjugate, the presence of the factor $1/(\varepsilon_{v\mathbf{k}} - \varepsilon_{n'\mathbf{k}'})$ means that all terms in which $n', \mathbf{k}'$ is a valence band cancel with the corresponding term in which these indices swapped with v and $\mathbf{k}$. Hence the sum over n' can be replaced with a sum over conduction bands, c, only. Equation (3.32) therefore becomes

$$\mu_X(\mathbf{r}) = 2\int d\mathbf{r}' \sum_{vc\mathbf{k}} \left(\langle v\mathbf{k}|\hat{V}_X^{NL}|c\mathbf{k}\rangle \frac{\phi^*_{c\mathbf{k}}(\mathbf{r}')\phi_{v\mathbf{k}}(\mathbf{r}')}{\varepsilon_{v\mathbf{k}} - \varepsilon_{c\mathbf{k}}} + \text{c.c.} \right) \frac{\delta\mu_{KS}(\mathbf{r}')}{\delta\rho(\mathbf{r})}. \tag{3.37}$$

The third and final functional derivative is the change in the Kohn-Sham potential resulting from a small change in the density. It is best dealt with by considering its inverse, which is the linear response matrix for non-interacting particles, $\chi_0(\mathbf{r}, \mathbf{r}')$, i.e.

$$\chi_0(\mathbf{r}, \mathbf{r}') = \frac{\delta\rho(\mathbf{r})}{\delta\mu_{KS}(\mathbf{r}')}. \tag{3.38}$$

Again using first order perturbation theory, this is given in terms of the Kohn-Sham orbitals by,

$$\chi_0(\mathbf{r}, \mathbf{r}') = 2\sum_{vc\mathbf{k}} \frac{\phi^*_{v\mathbf{k}}(\mathbf{r})\phi_{c\mathbf{k}}(\mathbf{r})\phi^*_{c\mathbf{k}}(\mathbf{r}')\phi_{v\mathbf{k}}(\mathbf{r}') + \text{c.c.}}{\varepsilon_{v\mathbf{k}} - \varepsilon_{c\mathbf{k}}}. \tag{3.39}$$

There is, however, a problem inverting χ_0 in order to use Equation (3.37), because a rigid shift in $\mu_{KS}(\mathbf{r})$ has no effect on the orbitals and hence the density, meaning that the matrix is singular and therefore has no inverse. This problem can be avoided by transforming the equations into reciprocal space. We can then simply exclude the $\mathbf{G} = \mathbf{0}$ components of $\mu_{KS}(\mathbf{G})$ from the theory, as these do not affect either the

energy or the density. Using the definition of the functions, $C_{j\mathbf{q}i\mathbf{k}}(\mathbf{G})$ in terms of the Fourier transform of a product of orbitals, given in Equation (3.26), we have

$$\chi_0(\mathbf{G},\mathbf{G}') = \frac{4}{\Omega}\sum_{vc\mathbf{k}} \frac{C_{c\mathbf{k}v\mathbf{k}}(\mathbf{G})C^*_{c\mathbf{k}v\mathbf{k}}(\mathbf{G}')}{\varepsilon_{v\mathbf{k}} - \varepsilon_{c\mathbf{k}}}. \tag{3.40}$$

The matrix $\tilde{\chi}_0(\mathbf{G},\mathbf{G}')$ is then defined as the sub-matrix of $\chi_0(\mathbf{G},\mathbf{G}')$ that excludes the row and column corresponding to $\mathbf{G} = 0$ and $\mathbf{G}' = 0$ respectively. Now transforming Equation (3.37) we arrive at

$$\mu_X(\mathbf{G}) = 2\sum_{\mathbf{G}'\neq\mathbf{0}} [E(\mathbf{G}') + E^*(-\mathbf{G}')]\tilde{\chi}_0^{-1}(\mathbf{G},\mathbf{G}'), \tag{3.41}$$

where the function, $E(\mathbf{G})$, is given by

$$E(\mathbf{G}) = \frac{1}{\sqrt{\Omega}}\sum_{vc\mathbf{k}} \langle v\mathbf{k}|\hat{V}_X^{NL}|c\mathbf{k}\rangle \frac{C_{v\mathbf{k}c\mathbf{k}}(\mathbf{G})}{\varepsilon_{v\mathbf{k}} - \varepsilon_{c\mathbf{k}}}. \tag{3.42}$$

These equations allow us, in principle, to calculate the exact exchange potential for a given set of Kohn-Sham orbitals and eigenvalues. However it should be mentioned that this method assumes that the occupancy of each orbital is fixed with respect to small changes in the Kohn-Sham potential, which may cause problems when applying this method to metallic systems.

3.2.2 The OEP Method

Originally developed for the treatment of spherically symmetric systems [80, 81], the *optimised effective potential* (OEP) method involves performing Kohn-Sham DFT calculations with a strictly *local* exchange-correlation potential, when the exchange-correlation energy is defined explicitly in terms of the orbitals. Combining the OEP with Equation (3.3) for the exact exchange energy provides a practical means of implementing EXX.

For the purposes of treating extended systems, the OEP method essentially involves minimising the total electronic energy with respect to the Kohn-Sham potential, $\mu_{KS}(\mathbf{r})$, rather than the Kohn-Sham orbitals as is done in standard calculations, or sX-LDA/HF calculations. As with these calculations, the effect of the procedure is still effectively to minimise the energy with respect to the density.

Explicit minimisation with respect to the potential can proceed either by employing an iterative self-consistent loop, or a standard numerical minimisation procedure such as conjugate gradients [38]. In the former case, all that is required beyond an LDA calculation is the calculation of the exchange potential in terms of the orbitals as discussed previously. In the latter case, however, we need to be able to calculate the gradient of the energy with respect to the potential, rather than the density, and this applies both to orbital based components (e.g. V_X, T_S), and density based components (e.g. V_H, V_C). For all quantities, Q, associated with any of the functionals used in this work, we are already able to calculate at least one of the derivatives

$$\frac{\delta Q}{\delta c^*_{\mathbf{ik}}(\mathbf{G})},$$

or

$$\frac{\delta Q}{\delta \rho(\mathbf{r})}.$$

For orbital derivatives, these can be converted into potential derivatives by analogy with the theory of exact exchange of the previous section. There, we saw that the derivative of the exchange energy with respect to the Kohn-Sham potential, $\delta V_X/\delta \mu_{KS}(\mathbf{r})$, can be expressed in reciprocal space as

$$E(\mathbf{G}) = \frac{1}{\sqrt{\Omega}} \sum_{v\mathbf{ck}} \langle v\mathbf{k}|\hat{V}_X^{NL}|c\mathbf{k}\rangle \frac{C_{v\mathbf{kck}}(\mathbf{G})}{\varepsilon_{v\mathbf{k}} - \varepsilon_{c\mathbf{k}}}. \tag{3.43}$$

The procedure used to arrive at this equation is applicable for any physical operator, $\hat{Q}$, i.e. the derivative $\delta Q/\delta \mu_{KS}(\mathbf{r})$ can be expressed in reciprocal space as

$$\frac{\delta Q}{\delta \mu_{KS}(\mathbf{G})} = \frac{1}{\sqrt{\Omega}} \sum_{v\mathbf{ck}} \langle v\mathbf{k}|\hat{Q}|c\mathbf{k}\rangle \frac{C_{v\mathbf{kck}}(\mathbf{G})}{\varepsilon_{v\mathbf{k}} - \varepsilon_{c\mathbf{k}}}. \tag{3.44}$$

So, if we have $\delta Q/\delta c^*_{\mathbf{ik}}(\mathbf{G})$, which is the conjugate of $\langle v\mathbf{k}|\hat{Q}|$, then all the ingredients are in place to calculate $\delta Q/\delta \mu_{KS}(\mathbf{r})$, which is needed for an OEP minimisation.

For density derivatives, we can apply the chain rule to obtain

$$\begin{aligned} \frac{\delta Q}{\delta \mu_{KS}(\mathbf{r})} &= \int d\mathbf{r}' \frac{\delta Q}{\delta \rho(\mathbf{r}')} \frac{\delta \rho(\mathbf{r}')}{\delta \mu_{KS}(\mathbf{r})} \\ &= \int d\mathbf{r}' \frac{\delta Q}{\delta \rho(\mathbf{r}')} \chi_0(\mathbf{r}', \mathbf{r}). \end{aligned} \tag{3.45}$$

In reciprocal space we therefore have

$$\frac{\delta Q}{\delta \mu_{KS}(\mathbf{G})} = \sum_{\mathbf{G}'} \frac{\delta Q}{\delta \rho(\mathbf{G}')} \chi_0(\mathbf{G}', \mathbf{G}), \tag{3.46}$$

so that now, if we have $\delta Q/\delta\rho(\mathbf{r})$, then we have all the ingredients in place to calculate $\delta Q/\delta\mu_{KS}(\mathbf{r})$, which is also needed for an OEP minimisation.

Together, the ability to evaluate $\delta Q/\delta\mu_{KS}(\mathbf{r})$ given either $\delta Q/\delta c^*_{\mathbf{ik}}(\mathbf{G})$ or $\delta Q/\delta\rho(\mathbf{r})$ mean that we can calculate the derivative with respect to the Kohn-Sham potential for any quantity in the Kohn-Sham framework. It is therefore possible to implement the OEP method using minimisation schemes such as conjugate gradients.

3.3 Improving Brillouin Zone Integration

The use of a relatively coarse grid for Brillouin zone sampling in practical calculations is justified when the integrands contributing towards the total energy are smoothly varying as a function of $\mathbf{k}$. In the case of sX-LDA, HF, and EXX, however, we also have an integration over $\mathbf{q}$ that is not necessarily as smoothly varying as the other integrands involved in a standard calculation. In particular, when no screening is present, there is a singularity where $\mathbf{q} = \mathbf{k}$ and $\mathbf{G} = \mathbf{0}$. In general, the form $1/|\mathbf{q}-\mathbf{k}+\mathbf{G}|^2$ is likely to require finer sampling in the region where $|\mathbf{q}-\mathbf{k}|$ is small. We will now discuss two possible approaches to tackling this problem, the first of which was introduced by Gygi and Baldereschi [68], and is aimed at removing the singularity from the summation, and the second of which has been developed by ourselves, and involves integrating the $1/|\mathbf{q}-\mathbf{k}+\mathbf{G}|^2$ part on a much finer grid than the orbital-dependant part.

3.3.1 The Divergence Correction

The equation for the exchange energy in reciprocal space contains a singularity where $\mathbf{q} = \mathbf{k}$ and $\mathbf{G} = \mathbf{0}$, and the summand for the Brillouin zone sampling diverges as $\mathbf{q} \to \mathbf{k}$ when $\mathbf{G} = \mathbf{0}$. This means that the assumption that the summand is smoothly varying, which we use to justify the coarse sampling in standard calculations, is no

longer valid. Without correction, this would force us to use much finer **k**-point sampling for HF and EXX calculations.

However, a method exists of correcting this problem following the work of Gygi and Baldereschi [68], that involves accurate numerical integration of the divergent component of the integrand. We start with the equation for the exchange energy as used in efficient implementations, i.e.

$$V_X = -\frac{2\pi}{\Omega}\sum_{\mathbf{ikjq}}\sum_{\mathbf{G}}\frac{|C_{\mathbf{jqik}}(\mathbf{G})|^2}{|\mathbf{q}-\mathbf{k}+\mathbf{G}|^2}. \tag{3.47}$$

If we were sampling the Brillouin zone with arbitrarily fine precision, because the wavefunction coefficients are smoothly varying with **k**, we would expect the summand, in the case of $\mathbf{G} = \mathbf{0}$, to tend towards

$$\begin{aligned}&\sum_{ij}\frac{|C_{\mathbf{jkik}}(\mathbf{0})|^2}{|\mathbf{q}-\mathbf{k}|^2}\\ &=\sum_{ij}\frac{|\delta_{ij}|^2}{|\mathbf{q}-\mathbf{k}|^2}\\ &=\frac{N(\mathbf{k})}{|\mathbf{q}-\mathbf{k}|^2},\end{aligned}$$

as $\mathbf{q} \to \mathbf{k}$ (where $N(\mathbf{k})$ is the number of orbitals on each **k**-point). This approximate non-dependence of $C_{\mathbf{jqik}}(\mathbf{0})$ on **q** in the region near $\mathbf{q} = \mathbf{k}$ means that we can correct for the divergence by adding an extra term to our expression for the exchange energy so that we have

$$\begin{aligned}V_X \;=\; &-\frac{2\pi}{\Omega}\sum_{\mathbf{ikjq}}\sum_{\mathbf{G}}\frac{|C_{\mathbf{jqik}}(\mathbf{G})|^2}{|\mathbf{q}-\mathbf{k}+\mathbf{G}|^2}\\ &+\frac{2\pi}{\Omega}N(\mathbf{k})\sum_{\mathbf{k}}\left(\sum_{\mathbf{q}}F(\mathbf{q}-\mathbf{k})-\Omega\int_{B.Z.}d\mathbf{q}F(\mathbf{q}-\mathbf{k})\right).\end{aligned} \tag{3.48}$$

Where $F(\mathbf{p})$ is a Brillouin zone periodic function that has the same divergence as the original equation for E_X, i.e. $F(\mathbf{p}) \to 1/p^2$ as $p \to 0$, and is smoothly varying away from the singularity. The additional term in Equation (3.48) is essentially the difference between summing the divergent function on a discrete grid and integrating it. The term therefore tends to zero in the limit of an infinitely fine grid. By including the term in our expression for V_X we effectively remove the overall divergence of the

summations, which allows us to use a relatively coarse grid to sample the Brillouin zone.

Other than having the properties just mentioned, the choice of form for $F(\mathbf{p})$ is essentially arbitrary. One possible choice is

$$F(\mathbf{p}) = \sum_{\mathbf{G}} \frac{S(|\mathbf{p}-\mathbf{G}|)}{|\mathbf{p}-\mathbf{G}|^2}, \tag{3.49}$$

where S is a sinusoidal envelope function of the following form:

$$\begin{aligned} S(x)_{|x|\leq w} &= \frac{1}{2}(\cos(\frac{\pi x}{w})+1), \\ S(x)_{|x|>w} &= 0. \end{aligned} \tag{3.50}$$

The larger the width of this envelope, w, the smoother F is away from the singularity; w should therefore be set substantially larger than the typical separation of the **k**-points. The advantage of choosing this form for $F(\mathbf{p})$ is that there is only a small number of non-zero terms in the sum over **G**-vectors.

Because we have changed our expression for E_X, we should now consider the effect this has on quantities that are derived from E_X, specifically the functional derivative of the exchange energy with respect to the orbital coefficients, $\delta E_X/\delta c^*_{\mathbf{ik}}(\mathbf{G})$, as used in HF, and the local exchange potential, $\mu_X(\mathbf{r})$, as used in EXX. Inspecting its form we see that, so long as the normalisation of the orbitals is preserved, the correction does not depend on the orbital coefficients. However there is still a contribution to $\delta E_X/\delta c^*_{\mathbf{ik}}(\mathbf{G})$ that is proportional to the function $c^*_{\mathbf{ik}}(\mathbf{G})$, i.e.

$$\frac{\delta}{\delta c^*_{\mathbf{ik}}(\mathbf{G})}\left[\frac{2\pi}{\Omega}\sum_{i'\mathbf{k}'}|C_{i'\mathbf{k}'i'\mathbf{k}'}(\mathbf{0})|^2\left(\sum_{\mathbf{q}}F(\mathbf{q}-\mathbf{k}')-\Omega\int_{B.Z.}d\mathbf{q}F(\mathbf{q}-\mathbf{k}')\right)\right]$$

$$= \frac{2\pi}{\Omega}\frac{\delta}{\delta c^*_{\mathbf{ik}}(\mathbf{G})}\sum_{i'\mathbf{k}'}\sum_{\mathbf{G}'\mathbf{G}''}c_{i'\mathbf{k}'}(\mathbf{G}')c^*_{i'\mathbf{k}'}(\mathbf{G}')c^*_{i'\mathbf{k}'}(\mathbf{G}'')c_{i'\mathbf{k}'}(\mathbf{G}'')$$

$$\left(\sum_{\mathbf{q}}F(\mathbf{q}-\mathbf{k}')-\Omega\int_{B.Z.}d\mathbf{q}F(\mathbf{q}-\mathbf{k}')\right)$$

$$= \frac{4\pi}{\Omega}\sum_{\mathbf{G}'}c_{\mathbf{ik}}(\mathbf{G}')c^*_{\mathbf{ik}}(\mathbf{G}')c_{\mathbf{ik}}(\mathbf{G})\left(\sum_{\mathbf{q}}F(\mathbf{q}-\mathbf{k})-\Omega\int_{B.Z.}d\mathbf{q}F(\mathbf{q}-\mathbf{k})\right)$$

$$= \frac{4\pi}{\Omega}c_{\mathbf{ik}}(\mathbf{G})\left(\sum_{\mathbf{q}}F(\mathbf{q}-\mathbf{k})-\Omega\int_{B.Z.}d\mathbf{q}F(\mathbf{q}-\mathbf{k})\right). \tag{3.51}$$

This does not alter the search direction but does ensure that the inner product of $\delta E_X/\delta c^*_{\mathbf{ik}}(\mathbf{G})$ and $c^*_{\mathbf{ik}}(\mathbf{G})$ yields the corrected eigenvalues. Because of this non-dependence of the correction on the orbitals, the ground state orbitals and density are not affected. Similarly, in EXX calculations, the correction alters the total energy but not the exchange potential.

3.3.2 Parallelepiped Integration

An alternative way of dealing with the $1/|\mathbf{q}-\mathbf{k}+\mathbf{G}|^2$ factor in the equation for the exchange energy is to integrate this term over an extended region of space surrounding each $\mathbf{q}$-point. Because each $\mathbf{q}$-point can be considered to represent a parallelepiped of reciprocal space, the simplest way to do this is to replace the $1/|\mathbf{q}-\mathbf{k}+\mathbf{G}|^2$ with the average value of this term over the extent of the parallelepiped surrounding $\mathbf{q}$, so that we have

$$V_X = -\frac{2\pi}{\Omega}\sum_{i\mathbf{k}j\mathbf{q}}\sum_{\mathbf{G}}|C_{j\mathbf{q}i\mathbf{k}}(\mathbf{G})|^2 I(\mathbf{k}-\mathbf{q}+\mathbf{G}), \tag{3.52}$$

where

$$I(\mathbf{P}) = \frac{1}{V_{\mathrm{PPD}}}\int_{\mathrm{PPD}(\mathbf{P})}d\mathbf{P}\frac{1}{|\mathbf{P}|^2}, \tag{3.53}$$

and where V_{PPD} is the volume of each parallelepiped in reciprocal space. In doing this we are effectively assuming that the orbital coefficients do not change significantly over the extent of one parallelepiped. This is reasonable so long as the **k**-point density is adequately converged.

3.4 Other Non-Local Functionals

A number of other non-local functionals also exist, both semi-empirical and non-empirical. Here we will briefly describe some other non-empirical functionals that are growing in use. This is by no means an exhaustive list, but is intended to give a flavour of some alternative directions that are being pursued other than the orbital functional approach that we focus on in this work.

3.4.1 WDA

The *weighted density approximation* (WDA) [82, 83, 84] is an exchange-correlation functional that involves estimating the shape of the coupling constant averaged exchange-correlation hole, $\bar{h}_{XC}(\mathbf{r}|\mathbf{r}')$. It is based on the assumption that the hole surrounding an electron at $\mathbf{r}'$ can be written as the product of the density at $\mathbf{r}$ and some radial function, G^{WDA}, centred on $\mathbf{r}'$. In its most general form, the hole is given by

$$\bar{h}_{XC}(\mathbf{r}|\mathbf{r}') = \rho(\mathbf{r}) G^{WDA}\left(|\mathbf{r}-\mathbf{r}'|, \tilde{\rho}(\mathbf{r}')\right), \tag{3.54}$$

where $\tilde{\rho}(\mathbf{r}')$ is the "weighted density", and is defined such that the hole integrates to exactly -1. With the hole shape determined for every $\mathbf{r}$, $\mathbf{r}'$, the exchange correlation energy is obtained via Equation (1.96).

For reasons of efficiency, it is convenient to require G^{WDA} to have the general form

$$G^{WDA}\left(R, \tilde{\rho}(\mathbf{r})\right) = \alpha\left(\tilde{\rho}(\mathbf{r})\right) f\left(\frac{R}{\beta\left(\tilde{\rho}(\mathbf{r})\right)}\right), \tag{3.55}$$

which means that it essentially always has the same basic shape, determined by the *pair correlation function*, f, just scaled either horizontally or vertically according to

the weighted density at $\mathbf{r}$. The parameters α and β are determined such so that the sum rule is satisfied and the exact exchange-correlation energy density is obtained for a homogeneous electron gas.

It has been found that the WDA can give good structural parameters and band gaps [26, 85], but that the band gaps obtained depend strongly on the choice of pair correlation function. There is no obvious way of defining a unique form for f from first principles, although it can be required to obey known physical limits [26], and it is currently unclear whether the good band gaps obtained are a result of the inherent quality of the WDA method itself, or simply a result of choosing a pair correlation functional that is not physically justified. It should of course be noted that the exact Kohn-Sham band gap is not necessarily close to the experimental one, and so a failure of a WDA to replicate the experimental band gap does not necessarily mean that the functional itself is inaccurate.

3.4.2 Meta-GGA and Hyper-GGA

The "meta-GGA" (MGGA) [86, 87, 88] it is essentially an extension of the GGA in which the non-interacting kinetic energy density is used as input to the functional as well as the electron density and its gradient. The spin-independent form of this functional is thus

$$E_{XC}^{MGGA}[\rho] = \int d\mathbf{r}\rho(\mathbf{r})\varepsilon_{XC}(\rho(\mathbf{r}), \nabla\rho(\mathbf{r}), \tau_S(\mathbf{r})), \tag{3.56}$$

where $\tau_S(\mathbf{r})$ is the non-interacting kinetic energy density defined, at least for these purposes[1], as

$$\tau_S(\mathbf{r}) = \frac{1}{2}\sum_{\mathbf{ik}} |\nabla\phi_{\mathbf{ik}}(\mathbf{r})|^2. \tag{3.57}$$

Implementation of such a functional is likely to be expensive however, as calculation of the exchange-correlation potential will require the evaluation of the functional derivative,

$$\frac{\delta\tau_S(\mathbf{r}')}{\delta\rho(\mathbf{r})},$$

[1]This is not the only way to define the kinetic energy density - for a good discussion of this issue see [89].

which would require methods related to those used in EXX.

Perdew et. al. also propose a "hyper-GGA" in which the exact exchange energy density is also included as an ingredient to the functional. This is intended to provide an accurate treatment of correlation, beyond the LDA or GGA, when using EXX to treat exchange.

3.5 Summary

In summary, in this chapter we have described a number of exchange-correlation functionals that go beyond either the LDA or GGAs in the sense that they incorporate non-local information. We described the functionals sX-LDA and HF, and how they can be cast in an manner that will facilitate efficient computational implementation. We also described the EXX functional, and the OEP procedure that is required in order to perform self-consistent calculations with EXX. In the next chapter, we describe how these functionals have been implemented computationally.

Chapter 4

Computational Implementation of HF, sX-LDA, and EXX

In this chapter we describe our computational implementation of HF, sX-LDA, and EXX within the CASTEP code. We start with a brief introduction explaining the motivation for carrying out this implementation, and an overview of its functionality. We then describe the implementation details of the new functionality required for calculations using non-local functionals. We describe how both symmetry and parallelisation are used, and how we deal with spin-dependent DFT, and calculations involving fractional band occupancies. We also present results of performance tests assessing the efficiency of the code, and its scaling properties with respect to basis-size, number of **k**-points, and number of parallel processors.

4.1 Introduction

As with all practical electronic structure methods, the non-local exchange correlation functionals described in the previous chapter must be implemented in a computer code before they can actually be used. While other implementations do already exist, implementation within an advanced code such as CASTEP has several advantages, including

- Well tested and optimised code is already in place at a lower-level in a modular structure, which can be accessed and used by the new code. This saves a lot of development time.
- A wide range of functionality exists at a higher level in the modular structure, which can be readily combined with the non-local functionals.
- Adherence to good programming practises must be kept, which makes for more stable and reliable code in the long run. Also, the large user-base, and regular testing by the distributors [90], mean bugs can be quickly identified and fixed.

Our implementation allows calculations to be run in parallel, with distribution of both **k**-points and/or **G**-vectors across multiple processors with distributed memory. It can make use of crystal symmetries to reduce the size of calculations, and can deal with spin-polarised systems, and systems with variable band-occupancies.

The description of the code given here is intended to give the reader a basic idea of the algorithms implemented in relation to non-local functionals, rather than to provide a detailed account of every part of the implementation. We have avoided listing actual computer code, preferring to describe what is happening algorithmically, with reference to data structures and parallel distribution. Some references are also made to other parts of the CASTEP code, which was described in Chapter 1. The implementation is complicated by symmetry, parallelisation, and other features, but we first give an account of the basic code without discussing these aspects in detail. The changes to the code that these features involve are then described later.

4.2 Preparation of Basis Set Data

Data related to the plane wave basis set and grids needs to be generated and stored at the beginning of a calculation, so that it can be used by other parts of the implementation later. A large amount of this data is of a general nature, and some of it is already available in the existing implementation of local functionals. Other

data relates more specifically to sX-LDA and HF calculations, in which we effectively have to implement the double-sum over **k**-points in Equations (3.24) and (3.30).

Because they contain terms involving $\mathbf{q} - \mathbf{k}$, these equations in general break the symmetries between different **k**-points, meaning one of the summations should be over the full set, while the other summation can still exploit symmetries. For this reason, in the context of the computational implementation, we consider these two summations separately and label the full set, that is not symmetry-reduced, as **q**-points, and the set that may be reduced by symmetry as **k**-points. Note that even if there are no crystal symmetries present, the size of the **k**-point set can still generally be reduced by a factor of $\frac{1}{2}$ due to time-reversal symmetry.

This distinction between **k**-points and **q**-points becomes particularly important in band structure calculations, where the **k**-points for which the band structure is to be calculated may be entirely unrelated to the **q**-points from which the potential is defined.

4.2.1 Initialisation of Basis Data

Initialisation, i.e. allocation and generation, of the basis-related data is separated into four successive stages. The main reason for this is that certain items of data may need to be processed by other modules before the generation of other items of data can proceed. Another reason is so that certain items of data can be re-initialised without the need to re-initialise everything.

The first part of the initialisation procedure deals with data related to the plane wave basis set for the **k**-points alone, that are not stored in standard local functional calculations. This include an array that maps a plane wave's reciprocal space grid coordinate to its logical index in the basis set. This complements the mapping from index to coordinate that is already part of the existing implementation for local functionals.

The second part of the initialisation procedure deals with **q**-points. Most of the data that is generated is related to the plane wave basis for orbitals stored on the

q-points. For example, we already store the reciprocal space coordinates of each basis function - this must also be done for the **q**-point basis. Also, as is done for the **k**-points in the first part of the initialisation procedure, extra arrays need to be generated, for example, that map a coordinate to an array index. We also generate look-up arrays that store the index of the **k**-point (in a self-consistent calculation) that corresponds to each **q**-point, along with the phase factor that may be needed as discussed shortly in the context of symmetry.

The third part of the initialisation procedure deals with the relationship between the **k**-points and the **q**-points. In all of the equations related to the non-local functionals in this work, **k**-points and **q**-points appear together as the difference $(\mathbf{q}-\mathbf{k})$. We therefore define a set of "$\mathbf{q}-\mathbf{k}$" points, i.e. the set of all points that can be expressed as the difference between a **q**-point and a **k**-point. This set is generated and stored, along with a look-up array that stores the index of the $\mathbf{q}-\mathbf{k}$ point corresponding to every possible pair of **q**-points and **k**-points.

The fourth part of the initialisation procedure involves calculating the reciprocal Coulomb factor, $1/(|\mathbf{q}-\mathbf{k}+\mathbf{G}|^2+k_s^2)$, that appears in Equation (3.24), for every possible value of $(\mathbf{q}-\mathbf{k})$ and **G**. If we are not using the divergence correction or parallelepiped integration then this is simply a case of evaluating each factor in turn. If we are using parallelepiped integration then we perform an accurate numerical integration of the Coulomb factor over the parallelepiped surrounding each point $\mathbf{q}-\mathbf{k}+\mathbf{G}$. If instead we are using the divergence correction, then we evaluate the Coulomb factor for each point, and also evaluate the divergence correction for each **k**-point.

4.3 Elements of Non-Local Functional Calculations

We now explain how we calculate the non-local exchange-correlation energy, E_{XC}^{NL}, for a given set of orbitals, and how we apply the non-local exchange-correlation potential operator, $\hat{V}_{XC}^{NL}$, to a set of orbitals to obtain the functional derivative of E_{XC}^{NL} with respect to the orbital coefficients.

4.3.1 Preparing the Data

Firstly, we take a set of orbitals corresponding to the set of **k**-points and generate a set of orbitals corresponding to the **q**-points in both reciprocal and real space. This is complicated when we are exploiting crystal symmetries, as will be discussed later, but for a normal calculation it is relatively straight forward. Essentially the orbitals are just directly copied, and then Fourier transformed to real space band by band. The only complication is that the **q**-points may include points in the negative hemisphere (i.e. points that are not in the **k**-point set due to time-reversal symmetry) - for these points we must take the conjugate of each coefficient, which effectively applies the time-reversal operation to the orbital.

Secondly, if required, we calculate the non-local exchange energy of the system. This involves calculating the expectation values of the non-local operator, which are essentially the individual terms in the summation over **k**-points, **q**-points and bands in Equation 3.24, as we will describe shortly, and then summing them to obtain the total non-local energy.

This step is not always necessary because sometimes we are only aiming to calculate the derivative of the energy with respect to the orbitals. The **q**-point orbitals still have to be generated though, so some preparation is always necessary before proceeding with the calculation.

4.3.2 Expectation Values of the Non-Local Operator

Given a set of orbitals, we evaluate the expectation value of the non-local exchange-correlation operator $\hat{V}_{XC}^{NL}$ for each orbital. From Equation (3.24), we see that the expectation value $\varepsilon_{\mathbf{ik}}^{NLXC}$ is given by

$$\varepsilon_{\mathbf{ik}}^{NLXC} = -\frac{2\pi}{\Omega}\sum_{j\mathbf{q}}\sum_{\mathbf{G}}\frac{|C_{j\mathbf{qik}}(\mathbf{G})|^2}{|\mathbf{q}-\mathbf{k}+\mathbf{G}|^2+k_s^2}. \tag{4.1}$$

This, of course, requires us to calculate the functions $C_{j\mathbf{qik}}(\mathbf{G})$, which in turn requires us to transform the orbitals into real space. For each band and **k**-point the procedure is essentially as follows:

1. Transform the coefficients $c_{\mathbf{ik}}(\mathbf{G})$ into the real space function $\phi_{\mathbf{ik}}(\mathbf{r})$.

2. Loop over bands and $\mathbf{q}$-points.

3. Following Equation (3.26), multiply the conjugate of $\phi_{\mathbf{ik}}(\mathbf{r})$ by $\phi_{j\mathbf{q}}(\mathbf{r})$, and transform to reciprocal space to obtain $C_{j\mathbf{qik}}(\mathbf{G})$.

4. Sum the product of $|C_{j\mathbf{qik}}(\mathbf{G})|^2$ with the reciprocal Coulomb factors, $1/(|\mathbf{q}-\mathbf{k}+\mathbf{G}|^2+k_s^2)$ to obtain the contribution to the expectation value from this band and $\mathbf{q}$-point.

This completes the basic calculation of the expectation values. However, if we are using the divergence correction, then, with reference to Equation (3.48), we must also add a term to each expectation value of

$$\frac{2\pi}{\Omega}\left(\sum_{\mathbf{q}} F(\mathbf{q}-\mathbf{k}) - \Omega \int_{B.Z.} d\mathbf{q} F(\mathbf{q}-\mathbf{k})\right).$$

Note prefactor of $2\pi/\Omega$ does not need to be included until the end of the routine, as including it in inner loops would only slow down the calculation.

4.3.3 Applying the Non-Local Operator

We take a set of orbitals and apply the the non-local exchange-correlation operator $\hat{V}_{XC}^{NL}$ to each orbital. The result of applying this operator is also equal to the functional derivative of the non-local exchange-correlation energy, E_{XC}^{NL}, with respect to the orbital coefficients. From Equation (3.30) we see that this is given by

$$\frac{\delta E_{XC}^{NL}}{\delta c_{\mathbf{ik}}^*(\mathbf{G})} = \sqrt{\Omega}\mathrm{FT}\left[-\sum_{j\mathbf{q}} \phi_{j\mathbf{q}}(\mathbf{r}) f_{j\mathbf{qik}}(\mathbf{r})\right], \tag{4.2}$$

where the functions, $f_{j\mathbf{qik}}(\mathbf{r})$, are given by

$$f_{j\mathbf{qik}}(\mathbf{r}) = \frac{4\pi}{\sqrt{\Omega}}\mathrm{FT}^{-1}\left[\frac{C_{j\mathbf{qik}}^*(\mathbf{G})}{|\mathbf{q}-\mathbf{k}+\mathbf{G}|^2+k_s^2}\right]. \tag{4.3}$$

So, in order to perform this operation, we start in essentially the same way as when calculating the expectation values up to the point at which we have calculated $C_{j\mathbf{qik}}(\mathbf{G})$. We then multiply the conjugate of this with the Coulomb factor,

$1/(|\mathbf{q}-\mathbf{k}+\mathbf{G}|^2+k_s^2)$, and transform to real space to obtain $f_{j\mathbf{q}i\mathbf{k}}(\mathbf{r})$. This is then multiplied by $\phi_{j\mathbf{q}}(\mathbf{r})$ and added to a running sum over bands and **q**-points.

After all the bands and **q**-points have been looped over, the sum is transformed into reciprocal space to obtain $\delta E_{XC}^{NL}/\delta c_{i\mathbf{k}}^{*}(\mathbf{G})$ for the current band and **k**-point.

If we are using the divergence correction, then there is also a component proportional to $c_{i\mathbf{k}}(\mathbf{G})$ that also needs to be added, i.e.

$$\frac{2\pi}{\Omega}c_{i\mathbf{k}}(\mathbf{G})\left(\sum_{\mathbf{q}}F(\mathbf{q}-\mathbf{k})-\Omega\int_{B.Z.}d\mathbf{q}F(\mathbf{q}-\mathbf{k})\right). \tag{4.4}$$

4.4 Additional Considerations

4.4.1 Defining q-points

The set of **q**-points is generated in a similar way to the **k**-points in a standard calculation with local functionals. We assume that a Monkhorst-Pack grid is being used to define the **k**-points, and, for a self-consistent calculation, the same grid must be used to define the **q**-points. Depending on how the calculation has been set up, we may not know explicitly what the dimensions and offset of the Monkhorst-Pack grid are - we may simply have a symmetry reduced set of **k**-points. We have therefore devised an algorithm that can "detect" whether a given set of points correspond to a Monkhorst-Pack grid, and, if so, what the size and offset of the grid is.

With the size and offset of the grid established, the **q**-point is generated. As well as this set of points, arrays are also needed that map each **q**-point to its corresponding **k**-point. The **k**-point set is almost always reduced by time-reversal symmetry, and so some **q**-points are related to their corresponding **k**-points via a time-reversal operation. If this is the case, then the conjugate must be taken when copying wavefunctions, so we create a logical array that indicates whether this is true for each **q**-point.

4.4.2 Electronic Minimisation and Band Structure

Application of the non-local exchange-correlation potential, and calculation of its expectation values, must now be included where necessary in the electronic minimisation algorithms. This is also the case when performing band structure calculations. In these calculations the various components of the Kohn-Sham potential are applied in turn. According to whether or not a calculation involving non-local functionals is being performed, we may now also calculate and apply the non-local exchange potential. Other than adding an extra component to both the energy and search direction, the electronic minimisation and band structure algorithms are essentially unchanged.

In band structure calculations the situation can be complicated by the fact that a different exchange-correlation functional may be being used for the band structure from that used for the preceding self-consistent calculation. For example, as we will discuss in greater detail in Chapter 5, an sX-LDA band structure calculation is often performed following a self-consistent LDA calculation. Matters are further complicated by the fact that the **k**-point set for the band structure may be completely unrelated to that for the self-consistent calculation. This means that the **q**-points, which come from the self-consistent calculation, are different to the **k**-points, which define the band structure path.

4.5 Parallelisation, Symmetry, and Other Issues

Up to this point, we have kept the description of the implementation relatively simple, deliberately omitting aspects related to parallelisation, the use of symmetry, and issues such as spin-polarisation and fractional occupancies. While the spin-polarisation and fractional occupancies are rather trivial, parallelisation and symmetry are crucial both to the structure of the code, and to its applicability to large systems. We now describe the methods used in extending the implementation to deal with these issues.

4.5.1 Symmetry

We will deal with symmetry first, as there are parts of the discussion on parallelisation that refer to symmetry-related issues. In a standard calculation, the use of symmetry has two main advantages. Firstly, it reduces the number of **k**-points that need to be explicitly dealt with by the computer, resulting in increased speed and reduced memory requirements. Secondly, in certain situations, it causes the effective **k**-point set to be larger than the original Monkhorst-Pack grid, which may improve convergence, and ensures that symmetry-related degeneracies are exactly satisfied. In an NLXC calculation, with the present implementation, use of symmetry has the first advantage but *not* the second. This is because, unlike in a standard calculation, we have a **q**-point set that is not symmetry reduced. Expansion of this set by symmetry would actually cause the calculation to be slower and require more memory. Nevertheless, reduction of the **k**-point set can still be highly beneficial in terms of performance for self-consistent calculations.

The use of symmetry is based on relationships between orbitals on symmetry related **k**-points, as described in 1.7.4. Essentially, if two **k**-points are related by a symmetry operation then any orbitals on those **k**-points are also related by the same operation. If the operation includes a translational component in real space, this becomes a phase factor in reciprocal space according to Equation (1.152). The symmetry reduced **k**-point set only includes points that are not related by symmetry. Each **q**-point must be related to one of the **k**-points by at least one symmetry operation.

As described previously the **q**-point basis is generated with reference to the **k**-point basis. In order to do this with a symmetry reduced **k**-point set, we must make use of the symmetry operations relating **k**-points and **q**-points. Applying the appropriate symmetry operation to the reciprocal coordinate of a plane wave in the **k**-point basis will yield the coordinate of the corresponding plane wave in the **q**-point basis. Doing this for each plane wave and **q**-point will generate the **q**-point basis and an array mapping plane waves in the **q**-point basis to corresponding plane waves in the **k**-point basis. Also, for symmetry operations with translational components, we calculate and store the phase factor that will be needed when mapping orbital

coefficients between **k**-point and **q**-points.

Symmetry only directly affects the preparation stage for applying $\hat{V}_{XC}^{NL}$ where orbital data is mapped from **k**-points onto **q**-points. The mapping between plane waves in the **k**-point basis and plane waves in the **q**-point basis is used, along with the phase factor for translational symmetries. Again, if the time-reversal operation is involved, the conjugate of the data is taken.

4.5.2 Parallelisation

The ability to run in parallel, with distribution by **k**-points and/or **G**-vectors, is probably the most important feature of the implementation. From a developer's point of view, it is also the most complicated, both in terms of writing code and in de-bugging. In Chapter 1 we outlined the essential strategy of parallelisation in a plane wave code such as CASTEP. The **k**-points are shared between **k**-point groups of processors, and the **G**-vectors are shared between **G**-vector groups of processors; each processor is a member of one **k**-point group and one **G**-vector group. Data is exchanged between processors, and the code is structured so as to minimise the required number of such calls due the associated cost of latency and data transfer.

Many of the arrays that are generated have to deal with the possibility of the data being distributed over several processors. Also, many of the arrays that are already generated by the existing implementation for local functionals module need to be "gathered" onto each node for an efficient parallel implementation of the NLXC functionals to be possible. For example, the array that stores the reciprocal space coordinates of each of the plane waves in the basis set is distributed across processors. We need to generate a version of that array that contains all the information from every **G**-vector group; this is done by gathering the data onto an array that has an extra index to indicate the **G**-vector group on which the data is stored in the distributed array. The inverse of this mapping now needs to include not only the logical index corresponding to each coordinate, but also the **G**-vector group on which that point is stored. Similarly, the mapping between **q**-points and equivalent **k**-points needs to include the **k**-point group of the **k**-point as well as its index.

Arrays such as the one that stores $C_{j\mathbf{q}\mathbf{i}\mathbf{k}}(\mathbf{G})$ can be distributed by **G**-vectors and/or **k**-points. This means that they take up less memory per processor, and loops over **k**-points and **G**-vectors only have to deal with those points that are on the local processor, which means that they take less time to complete. If the purpose of a loop is to evaluate a sum, then each processor only has part of the sum at the end of the loop. We complete the sum by adding together the components from each processor, and sharing the result between all processors. Fourier transforms are also faster when running with **G**-vector parallelisation, and these can be handled by the existing implementation for local functionals.

The **q**-point data is not distributed. This because it all needs to be accessed rapidly by each processor in inner-loops. While distribution would have been possible in principle, it would not have been practical in terms of efficiency; instead, we accept the need to use more memory for a non-local functional calculation that a standard calculation.

4.5.3 Spin-Polarised Systems and Fractional Occupancies

Extension of the code to be able to treat systems with collinear spin polarisation, using spin-dependant DFT, and systems with fractional occupancies is relatively straightforward. Because there is zero exchange interaction between unlike spins, the calculation of expectation values, and application of the potential, for the spin-up orbitals are independent of the same procedures for the spin-down orbitals. Hence, the extension to spin-polarised systems essentially involves simply placing these procedures within an outer loop over spin indices. We also have to consider the possibility of having a spin-dependent screening constant. This means that the Coulomb factors need to have an extra spin dimension. As discussed in 1.6.1, when dealing with systems with fractional occupancies, the calculation of orbital dependent quantities must include a weighting for each orbital according to its occupancy. These extensions of the theory can be expressed, for example, by re-writing Equation (3.24) as follows,

$$E_{XC}^{NL} = -\frac{2\pi}{\Omega}\sum_{\sigma}\sum_{\mathbf{i}\mathbf{k}j\mathbf{q}}\sum_{\mathbf{G}} f_{\mathbf{i}\mathbf{k},\sigma} f_{j\mathbf{q},\sigma} \frac{|C_{j\mathbf{q}\mathbf{i}\mathbf{k},\sigma}(\mathbf{G})|^2}{|\mathbf{q}-\mathbf{k}+\mathbf{G}|^2 + k_{s\sigma}^2}, \tag{4.5}$$

where σ is the spin index, and the $f_{i\mathbf{k},\sigma}$ are the fractional occupancies of the orbitals.

4.6 Performance Tests

Having described the computational implementation in terms of the algorithmic procedures, we now demonstrate some of the features of the code, including the scaling properties with respect to basis size, and the ability of parallelisation and harnessing of symmetry to reduce the computational cost of a calculation.

4.6.1 Scaling With Basis Size

We first investigate the performance of the code as a function of the size of the plane wave basis set used. We perform a series of total energy calculations, using sX-LDA, on a 2-atom primitive cell of silicon, using a single **k**-point at [0.5,0.5,0.5] to sample the Brillouin zone (although this means the calculation is not converged with respect to **k**-point sampling, this is not important when simply evaluating scaling properties of the implementation). The plane wave cut-off energy is varied between 200eV and 2800eV in steps of 200eV. The speed of the calculation is determined in terms of the average time for one conjugate gradients line search during the total energy minimisation. The results are shown in Figure 4.1. As discussed earlier in 3.1.5, we would expect this calculation to scale as $N_p \log(N_p)$. By fitting the results to the equivalent form as a function of cut-off energy, i.e. $E_{cut}^{\frac{3}{2}} \log(E_{cut})$ we see that the calculation does indeed scale roughly as expected, however the time per SCF cycle seems to go up in large jumps rather than increasing smoothly with cut-off energy. This can be explained in terms of the size of the full grid used for the FFTs, which, due to the nature of the FFT algorithm, increases in relatively large steps. Also shown in the figure are LDA results for comparison. For calculations of this size, an LDA calculation also scales roughly as $N_p \log(N_p)$, but the prefactor is about an order of magnitude smaller.

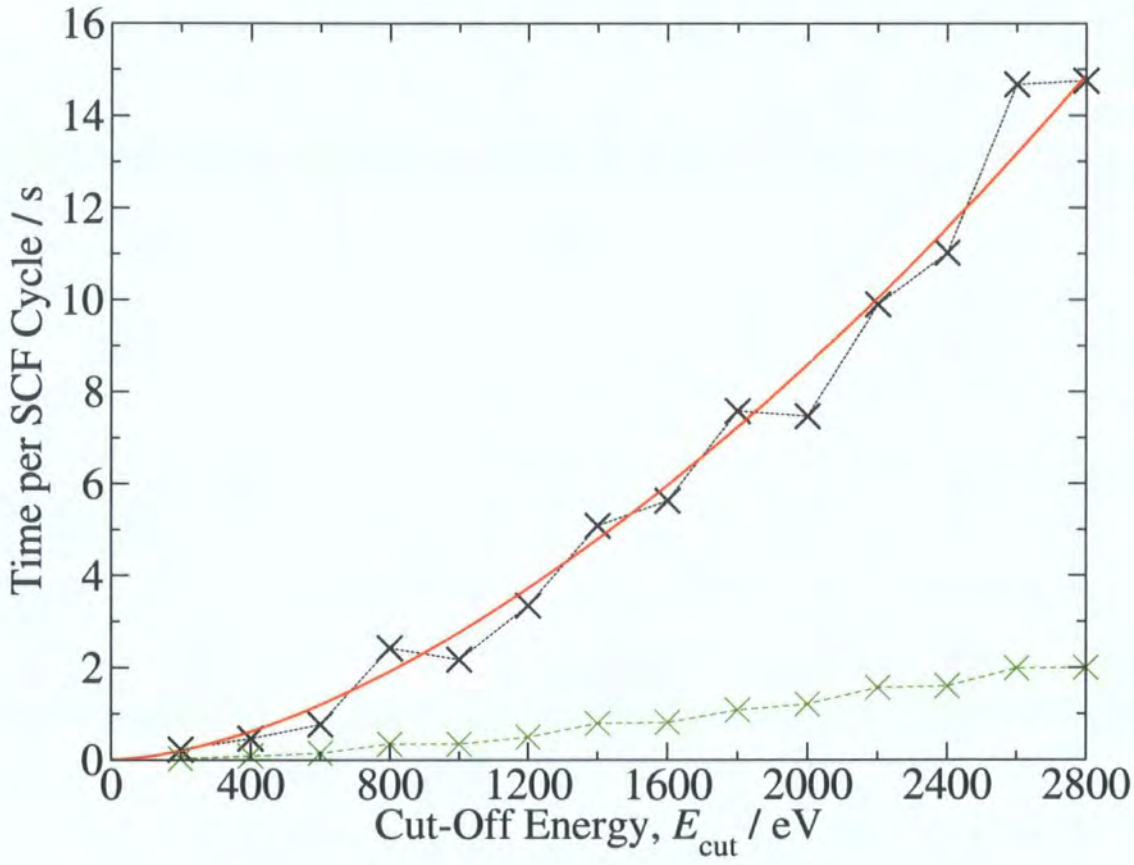


Figure 4.1: Average time per SCF cycle in a typical sX-LDA calculation as a function of the plane wave cut-off energy, E_{cut} (black crosses), and best fit curve of the predicted form of $\sim E_{cut}^{\frac{3}{2}} \log(E_{cut})$ (red line). Also shown (in green) are the results from an equivalent calculation using the LDA.

4.6.2 Scaling with k-points and Symmetry

We now look at the performance of the code as a function of the number of **k**-points used to sample the Brillouin zone. Again we perform a series of total energy calculations on a 2-atom primitive cell of silicon, this time fixing the cut-off energy at 350eV. The number of **k**-points is varied by using off-origin Monkhorst-Pack grids of dimesionalities between $1 \times 1 \times 1$ and $5 \times 5 \times 5$, which, taking into account time-reversal symmetry, result in **k**-point set sizes of 1, 4, 14, 32, and 63. We also perform the same calculations but harnessing crystal symmetries to reduce the number of **k**-points. With symmetry present the sizes of the reduced **k**-point sets are 1, 2, 6, 10, and 19, which should result in a speed-up proportional to this reduction. The results of these calculations are shown in Figure 4.2. As discussed in 3.1.5, without symmetry we expect the calculation to scale roughly as N_k^2, which we confirm by fitting the results to a function of this form. With symmetry harnessed, we also observe approximately the expected speed-up, i.e. proportional to the ratio of unsymmetrised to symmetrised **k**-points.

4.6.3 Parallelisation by k-points

The ability to run code in parallel is one of the main features of this implementation, so it is essential to check that the parallel aspects of the code actually result in significant increases in speed. We first look at parallelisation by **k**-points. We use a primitive silicon cell with a $4 \times 4 \times 4$ Monkhorst-Pack grid and a 350eV cut-off. Taking into account time-reversal symmetry we have 32 **k**-points, which can be distributed over a maximum of 32 processors. We run the calculation first using a single processor, then repeat it with the **k**-points distributed across 2, 4, 8, 16, and 32 processors. We determine the speed of the calculation in terms of the average number of SCF cycles per second during the main part of the calculation. The results are shown in Figure 4.3. We see that the calculation scales almost linearly with the number of processors, which shows that the parallelisation is working very effectively.

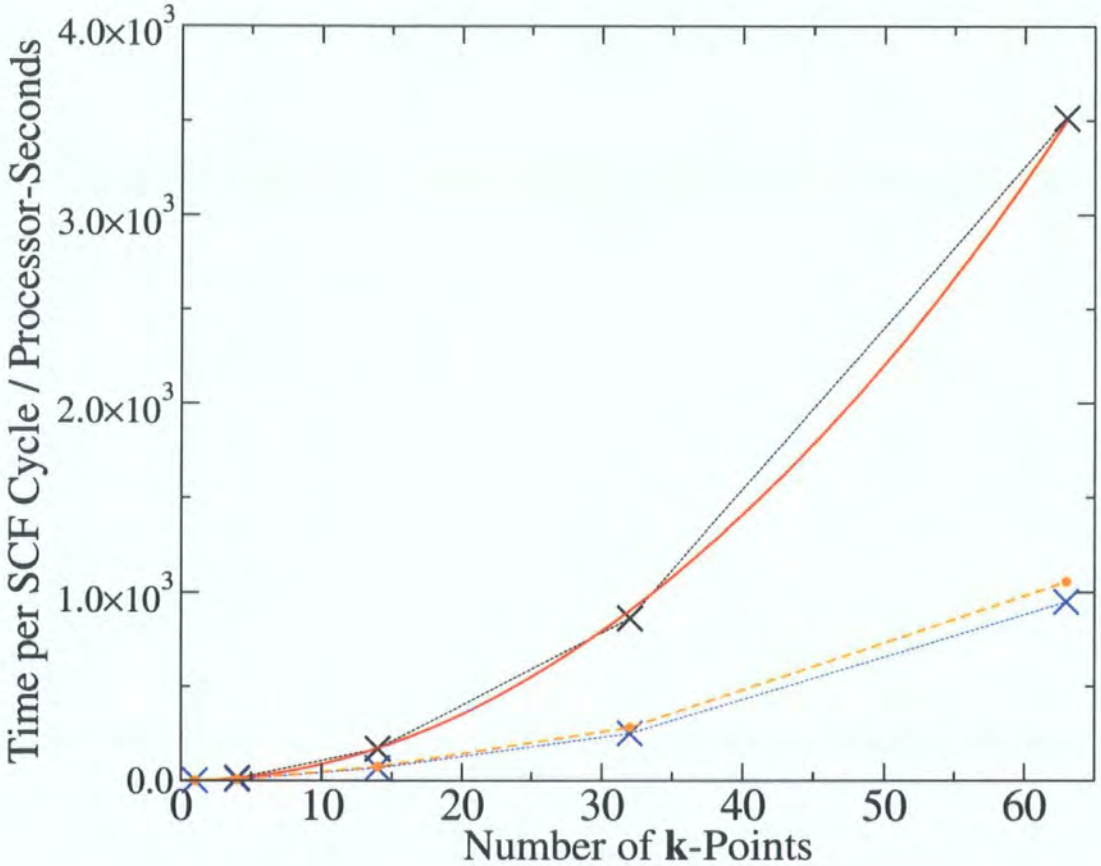


Figure 4.2: Average processor time per SCF cycle in a typical sX-LDA calculation as a function of the number of **k**-points (black crosses), and best fit curve of the predicted form of $\sim N_k^2$. Also shown are the times when symmetry is harnessed (blue crosses), along with the predicted times based on the best fit curve and the ratio of reduced to full **k**-points (orange dots).

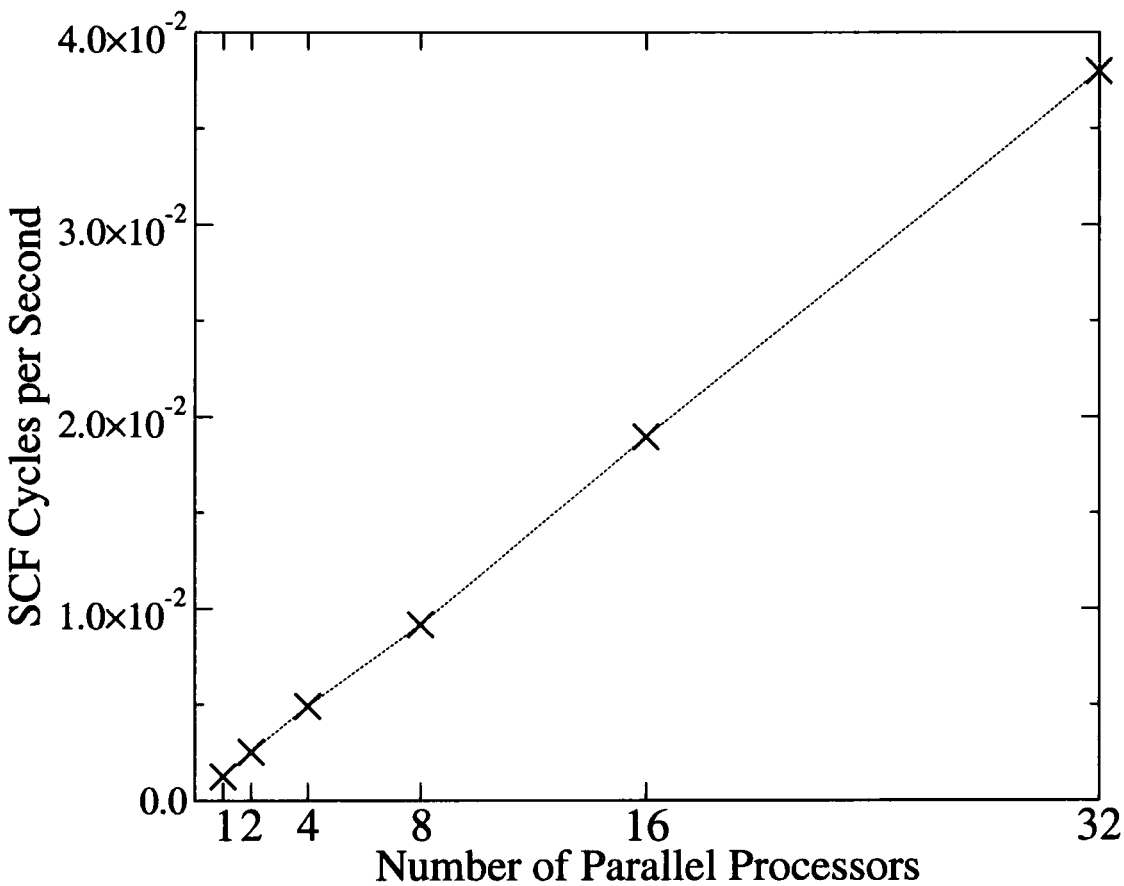


Figure 4.3: Average number of SCF cycles per second in a typical sX-LDA calculation as a function of the number of processors over which **k**-points are distributed.

4.6.4 Parallelisation by G-vectors

We now carry out a similar test for parallelisation by G-vectors. G-vector parallelisation is likely to be most effective for larger calculations so for this test we use an 8-atom cubic silicon cell and set the cut-off energy to 2000eV. A single **k**-point at [0.5,0.5,0.5] is used to sample the Brillouin zone. Again, we run the calculation on a single processor, and then choose to repeat it with the G-vectors distributed across 2, 4, 8, 16, and 32 processors (although we are not actually restricted to factors of 32 in this case). The results are shown in Figure 4.4. We see that a significant increase in speed can be achieved by such a parallelisation strategy, but it is not linear as it was in the case of **k**-points. Of particular note is the fact that the speed actually *decreases* as we go from 1 to 2 processors. The reason that parallelisation by G-vectors is not as good as parallelisation by **k**-points is that it requires much more inter-processor communication due to the Fourier transforms, and hence incurs many more latency and data transfer related overheads. For this reason, **k**-point parallelisation should always be chosen over G-vector parallelisation when the option is available.

4.7 EXX and the OEP Method

Implementation of EXX involves a certain amount of extra development beyond the non-local functional implementation described so far. This includes the calculation of the functional derivatives described in 3.2.2, and the implementation of the basic OEP procedure itself.

4.7.1 Calculating Functional Derivatives

In Section 3.2, we discussed the theory of the local EXX potential, and the functional derivatives of quantities with respect to the Kohn-Sham potential required for the OEP method. The calculation of the local EXX potential involves calculation of the functional derivative of the exchange energy, with respect to the density, but

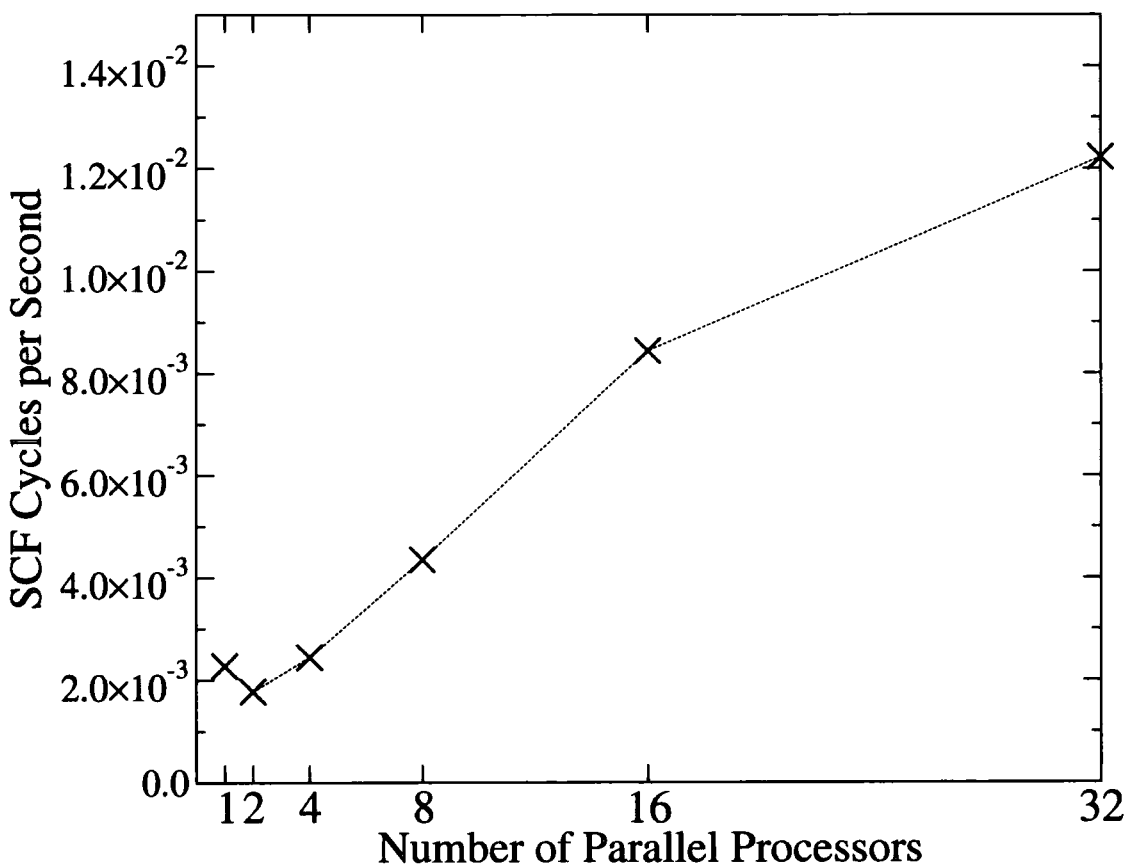


Figure 4.4: Average number of SCF cycles per second in a typical sX-LDA calculation as a function of the number of processors over which **G**-vectors are distributed.

the procedure can be applied to any operator. In order to implement EXX and the OEP method, we need to implement procedures for carrying out the following three tasks for a given operator, $\hat{Q}$, pertaining to a physical quantity, Q:

- Given a set of Kohn-Sham orbitals, $\{\phi_{\mathbf{ik}}(\mathbf{r})\}$, associated eigenvalues, $\{\varepsilon_{\mathbf{ik}}\}$, and the result of applying the operator, $\hat{Q}$, to those orbitals, i.e. $\{\hat{Q}|\phi_{\mathbf{ik}}(\mathbf{r})\rangle\}$, calculate the functional derivative of Q with respect to the density, i.e. $\delta Q/\delta\rho(\mathbf{r})$.

- Given a set of Kohn-Sham orbitals, $\{\phi_{\mathbf{ik}}(\mathbf{r})\}$, associated eigenvalues, $\{\varepsilon_{\mathbf{ik}}\}$, and the result of applying the operator, $\hat{Q}$, to those orbitals, i.e. $\{\hat{Q}|\phi_{\mathbf{ik}}(\mathbf{r})\rangle\}$, calculate the functional derivative of Q with respect to the Kohn-Sham potential, i.e. $\delta Q/\delta\mu_{KS}(\mathbf{r})$.

- Given a set of Kohn-Sham orbitals, $\{\phi_{\mathbf{ik}}(\mathbf{r})\}$, associated eigenvalues, $\{\varepsilon_{\mathbf{ik}}\}$, and the functional derivative of Q with respect to the density, $\delta Q/\delta\rho(\mathbf{r})$, calculate the functional derivative of Q with respect to the Kohn-Sham potential, i.e. $\delta Q/\delta\mu_{KS}(\mathbf{r})$.

To calculate $\delta Q/\delta\rho(\mathbf{r})$, we essentially follow the procedure for calculating the local EXX potential, as described in 3.2, except that the non-local exact exchange operator, $\hat{V}_X^{NL}(\mathbf{r},\mathbf{r}')$, is replaced by the general operator, $\hat{Q}$, and the exchange energy, V_X, is replaced by the general physical quantity, Q. The reason for this generalisation is to facilitate the implementation of new functionals in the future that may need to make use of this type of procedure, such as the meta-GGA, described in 3.4.2. The orbitals, eigenvalues, and $\{\hat{Q}|\phi_{\mathbf{ik}}(\mathbf{r})\rangle\}$ data, are calculated in reciprocal space; the orbitals and eigenvalues must include the full set of valence and conduction bands, but the $\{\hat{Q}|\phi_{\mathbf{ik}}(\mathbf{r})\rangle\}$ data only needs to include the valence bands. This is important in terms of efficiency, as the calculation of the $\{\hat{Q}|\phi_{\mathbf{ik}}(\mathbf{r})\rangle\}$ may be expensive. The main purpose of this algorithm is essentially to solve Equation (3.41), which means we must compute the linear response matrix, $\chi_0(\mathbf{G},\mathbf{G}')$, and the derivatives $\delta Q/\delta\mu_{KS}(\mathbf{G})$. Solution of this equation is very expensive, and we are forced to consider ways of maximising the efficiency of the procedure. By using a reduced grid of points, lying within a sphere of limited radius, rather than the full grid, to represent

$\chi_0(\mathbf{G},\mathbf{G}')$, and $\delta Q/\delta\mu_{KS}(\mathbf{G})$, we can increase the efficiency dramatically. From our preliminary tests, and also from the work of others [79], it seems that using a radius equal to the plane wave cut-off radius does not significantly affect the accuracy of the calculation. The functions $C_{v\mathbf{k}c\mathbf{k}}(\mathbf{G})$ are calculated using the same efficient procedure described in 3.1.5, and then transformed onto the reduced grid to calculate $\chi_0(\mathbf{G},\mathbf{G}')$, and $\delta Q/\delta\mu_{KS}(\mathbf{G})$. A numerical routine is used to solve Equation (3.41), and the resulting functional derivative is then transformed back onto the full grid. Finally, the derivative is Fourier transformed to real space as required.

We have implemented Equation (3.46) by calculating the linear response matrix, $\chi_0(\mathbf{G},\mathbf{G}')$, and applying it to the Fourier transform of the input derivative, $\delta Q/\delta\rho(\mathbf{r})$. The calculation of $\chi_0(\mathbf{G},\mathbf{G}')$ is done in the same way as it is in the procedure for calculating $\delta Q/\delta\rho(\mathbf{r})$, and again, this is more efficient if we use a reduced grid. After applying the matrix the resulting derivative, $\delta Q/\delta\mu_{KS}(\mathbf{G})$, is Fourier transformed to real space as required.

4.7.2 Minimisation with the OEP Method

In calculations involving the EXX functional, the minimisation of the electronic energy must proceed via the OEP method rather than the existing electronic minimisation algorithms for local functionals. It is therefore necessary to implement the OEP method, described in 3.2.2, for using the Kohn-Sham potential, rather than the Kohn-Sham orbitals, as the object with respect to which the energy is minimised.

As discussed in 3.2.2, there are two different ways of solving the Kohn-Sham equations with the OEP method. The first is a basic self-consistent iterative cycle in which a trial potential is used to generate orbitals and a density, which are used to calculate a new potential, and so on. The second is a more sophisticated search in which the energy is minimised with respect to the Kohn-Sham potential using a numerical algorithm such as conjugate gradients. We have implemented both of these approaches. The conjugate gradients method requires many evaluations of the total energy during the line minimisation stages, but these are relatively cheap as they do not require the functional derivative to be evaluated more than once per

line search. Also, when calculating search directions, it is only necessary to calculate derivatives with respect to the potential, rather than the density, which is a more expensive operation. The basic iterative approach requires only a small number of calculations of the total energy, but each one involves diagonalisation of the full Kohn-Sham Hamiltonian in order to obtain the full set of valence and conduction bands; in addition to this, $\delta Q/\delta\rho(\mathbf{r})$ needs to be calculated at each iterative step, which is expensive. At the present stage of development, it is not clear which of the two approaches is the most efficient, and this may well depend on the size and nature of the system under study, but the simple iterative approach has proven to be the more stable close to the minimum. It may be that the best approach would be to use the conjugate gradients method to get close to the minimum and then use the iterative procedure for the last few steps. Once the minimisation procedure is complete, we calculate the local exchange potential, as described in Section 3.2. This allows the local potential to be used in, for example, band structure calculations that may follow this self-consistent minimisation stage.

In band structure calculations we may wish to perform the calculation non-self-consistently, using a local functional such as the LDA for the self-consistent energy minimisation, and EXX for the band structure. In such circumstances we can evaluate the exact local exchange potential simply given the ground state density from the self-consistent calculation. The orbitals and eigenvalues that are used in Equations (3.41) and (3.42) are those corresponding to the Kohn-Sham potential calculated with the local functional. We cannot simply use the values obtained at the end of the self-consistent calculation, however, because we now need the complete set of valence and conduction bands. The full set of orbitals thus have to be obtained by this by calculating the Kohn-Sham potential with the local functional and then diagonalising the full Hamiltonian. With the local exchange potential evaluated, it can then be used to replace the exchange potential from the local functional, and the band structure can be calculated using the existing implementation for local functionals.

4.8 Summary and Conclusions

In summary, in this chapter we have described the computational implementation of non-local exchange-correlation functionals sX-LDA and HF, using the efficient formulation described in Chapter 3. This included a description of how the implementation can make use of cell symmetry and be run in parallel on multiple computer processors. We tested the scaling properties of the implementation as a function of cut-off energy, number of **k**-points, and number of parallel processors. The results of these tests agreed with the prediction that the speed of a calculation should scale as $N_p \log(N_p) N_k^2 N_b^2$. This means that in terms of the size of a system, measured by the number of atoms present, N_a, we would expect an sX-LDA or HF calculation to scale as $N_a^3 \log(N_a)$. This compares to an LDA calculation, which ultimately scales as N_a^3 in a standard plane wave based implementation. We also found that by making use of symmetry we can increase the speed of a calculation, as expected, in proportion to the ratio of the size of the symmetry-reduced **k**-point set to the size of the unreduced set. When running in parallel, we found that distribution of **k**-points increased the speed of a calculation approximately linearly as a function of the number of processors. Distribution of **G**-vectors can also result in a significant speed-up, but the scaling is not as favourable as it is in the case of **k**-points. This is due to the much greater amount of inter-processor communication involved when distributing **G**-vectors. Finally, we discussed the implementation of EXX via the OEP method as described in Chapter 3. While the implementation is essentially complete, further optimisation will be necessary before using this implementation for practical calculations.

With the implementation of sX-LDA and HF fully working and tested, in the next chapter we proceed to use this implementation to calculate the band structures of semiconductors with these functionals.

Chapter 5

Semiconductor Band Structures from Non-Local Functionals

Band structures are an important property of semiconductor materials as they ultimately determine most of the electronic and optical properties that make these materials technologically useful. In this chapter we use the sX-LDA and HF functionals implemented in Chapter 4 to calculate the band structure of a number of semiconducting materials. We begin with the calculation of complete band structures for Si and GaN, and compare the performance of the functionals against both LDA results and to experiment. We then calculate the band structure at the main symmetry points of a large number of group IV and III-V semiconductors, again comparing with LDA and experiment. Finally we conclude with a discussion of the reasons for the variation in performance of the different functionals.

5.1 Full Band Structures for Silicon and GaN

5.1.1 Preliminaries

In Chapter 2, we established the appropriate cut-off energy and k-point grid for GaN and calculated the LDA geometry of both the wurtzite and zinc blende structures. For the band structure calculations in this chapter, we will use the same parameters

and the same LDA geometry irrespective of the functional used. This is currently common practice for such calculations due to the extra computational expense of calculating the geometry with non-local functionals. For calculations on Si, we follow a similar procedure to that of Chapter 2 to establish the appropriate parameters and LDA geometry. Using a primitive 2-atom cell, we find that energy differences are converged to within 0.01eV per atom with a cut-off energy of 350eV and a $3 \times 3 \times 3$ off-origin Monkhorst-Pack grid. Running a geometry optimisation we obtain a lattice parameter of 5.39Å, which is close to the experimentally measured value of 5.43Å[54].

5.1.2 Silicon

We begin by calculating the band structure of silicon using, sX-LDA and HF, and comparing with LDA results. The calculated band structures are shown in Figures 5.1, and 5.2. In the non-local calculations we calculate the band structure after both an LDA self-consistent minimisation, and a self-consistent minimisation with the same functional as used to calculate the band structure. From the data we find that for the sX-LDA calculations, we can perform the SCF part of the calculation with the LDA without significantly altering the results. Because of the gain in efficiency afforded by doing this, all sX-LDA band structures in this work will performed in the same way. Unfortunately this does not work so well when using HF so all HF calculations are performed fully self-consistently.

We see that Si has an indirect band gap, with the conduction band minimum about $\frac{4}{5}$ of the way along the $\Gamma - X$ line. The LDA gives a value for this band gap of 0.48eV, while sX-LDA gives 0.97eV, and HF gives a value of 4.78eV. The experimentally measured value for this band gap is 1.12eV [54]. We see, therefore, that the LDA underestimates the gap, while HF grossly overestimates it. sX-LDA, on the other hand, gives a result reasonably close to experiment. The reasons for this are discussed in some detail in Section 5.3

Closely related to the band structure of a material is the *density of states* (DOS). This is an important quantity, particularly in terms of optical properties, as it affects

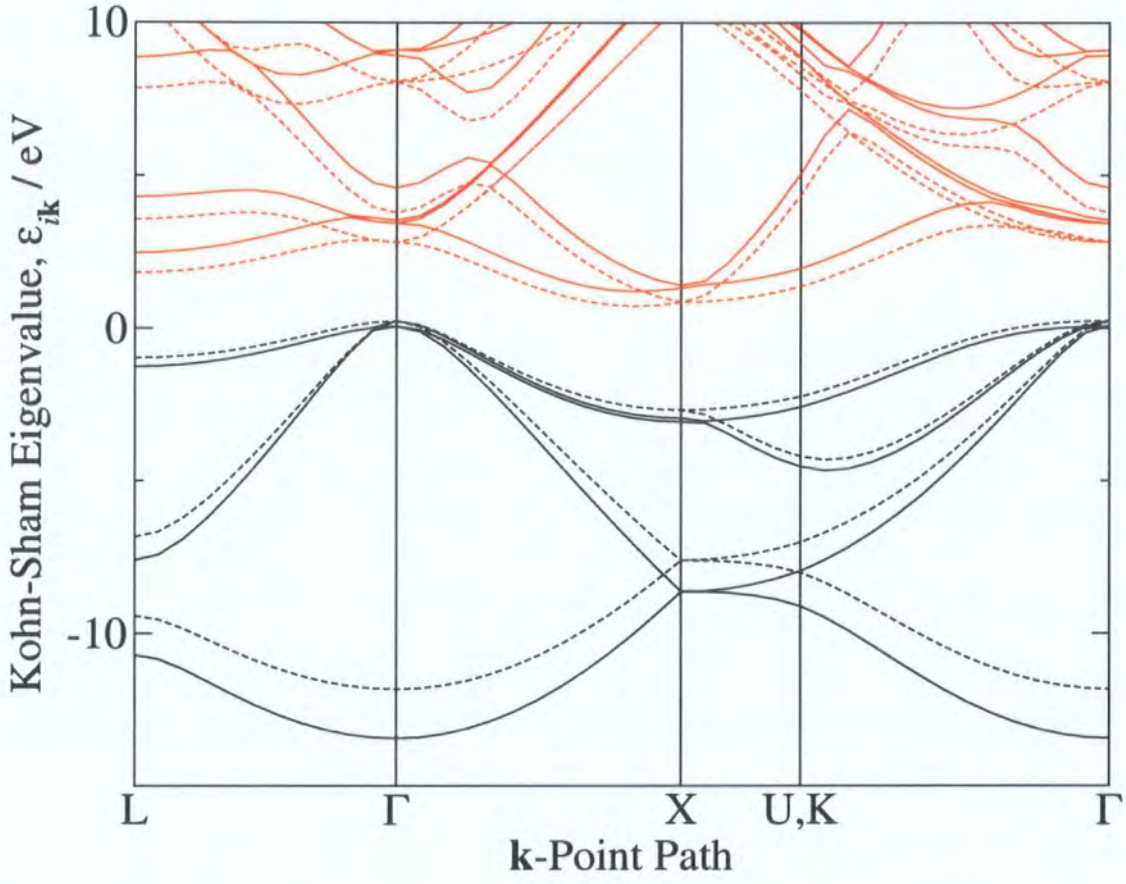


Figure 5.1: Kohn-Sham band structure of silicon calculated using sX-LDA. Black lines indicate occupied valence bands, while red lines indicate unoccupied conduction bands. Dashed lines show results calculated with the LDA with all eigenvalues shifted such that the valence band maximum equals that of the sX-LDA calculation.

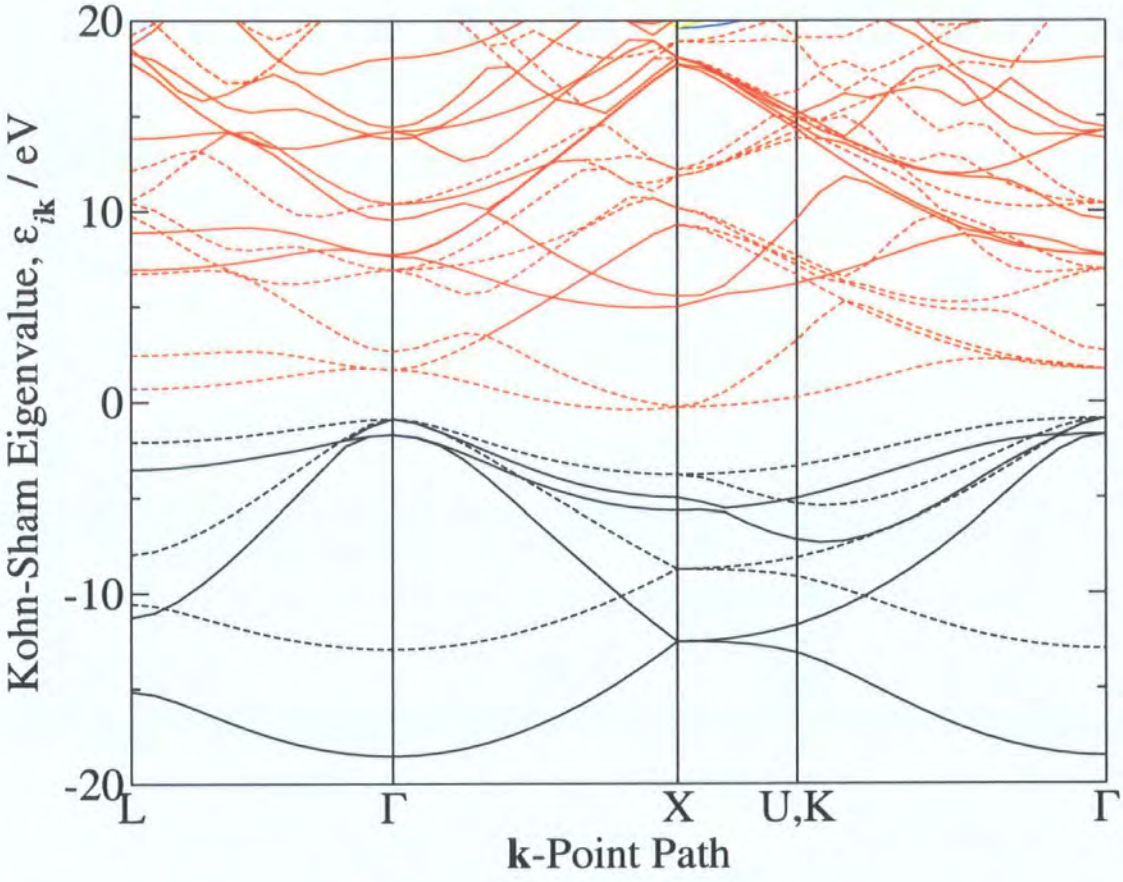


Figure 5.2: Kohn-Sham band structure of silicon calculated using HF. Black lines indicate occupied valence bands, while red lines indicate unoccupied conduction bands. Dashed lines show results calculated with the LDA with all eigenvalues shifted such that the valence band maximum equals that of the sX-LDA calculation.

the rate of absorption or emission of photons of a given energy. In terms of a Kohn-Sham DFT picture, the density of states, $D(\varepsilon)$, essentially tells us how many orbitals there are of eigenvalue ε; it is defined as follows,

$$D(\varepsilon) = \sum_{n\mathbf{k}} \delta(\varepsilon_{i\mathbf{k}} - \varepsilon), \tag{5.1}$$

where the index n runs over both valence and conduction bands. We have calculated the DOS in Si for each of the functionals LDA, sX-LDA, and HF. In these calculations we must perform a full 3D band structure calculation, sampling the whole of the Brillouin zone. Because of the finite sampling, the resulting DOS results tend to appear spikey, even when using very fine grids. We use an off-origin Monkhorst-Pack grid of dimension $8 \times 8 \times 8$, much finer than the grid used for the self-consistent total energy calculation. We also choose to smear each of the δ-functions in 5.1 with a Gaussian function of width 0.4eV. The calculated DOS for Si using the LDA, sX-LDA, and HF are shown in Figures 5.3, 5.4, and 5.5 respectively. The Gaussian smearing has the effect of masking the precise locations band edges, however the band gaps, of varying widths depending on the functional, are still clearly identifiable. We also note that the use of the different functionals has little effect on the overall shape of the DOS, however, particularly in the case of HF we see that the DOS is stretched over a wider range of eigenvalues that the LDA.

We have also calculated 2-dimensional plots of charge densities for the highest valence and lowest conduction orbitals on the Γ-point, when using each of the non-local functionals sX-LDA and HF, and compared them to LDA results. These are shown in Figures 5.6 and 5.7. From these plots we see that the main differences lie in the shape of the conduction band orbitals, rather than the valence band orbitals. This suggests that the differences in eigenvalue arise more from the change in the shape of the conduction bands than the valence band. However, we know from the fact that we cannot use the LDA self-consistent ground state for a HF calculation, that there must still be some important differences in the valence bands also.

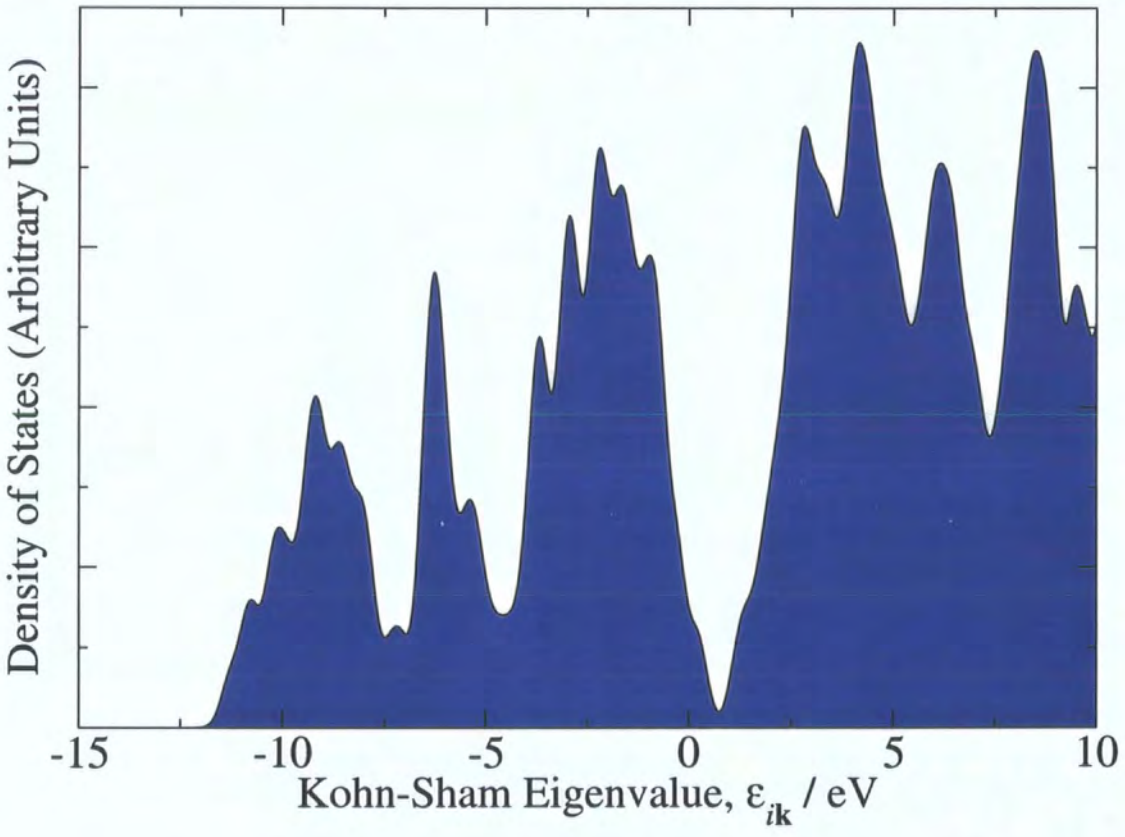


Figure 5.3: Electronic density of states (DOS) in Si, calculated using the LDA (the Brillouin zone is sampled with an $8 \times 8 \times 8$ MP grid and the graph is smoothed by Gaussian smearing of width 0.4eV).

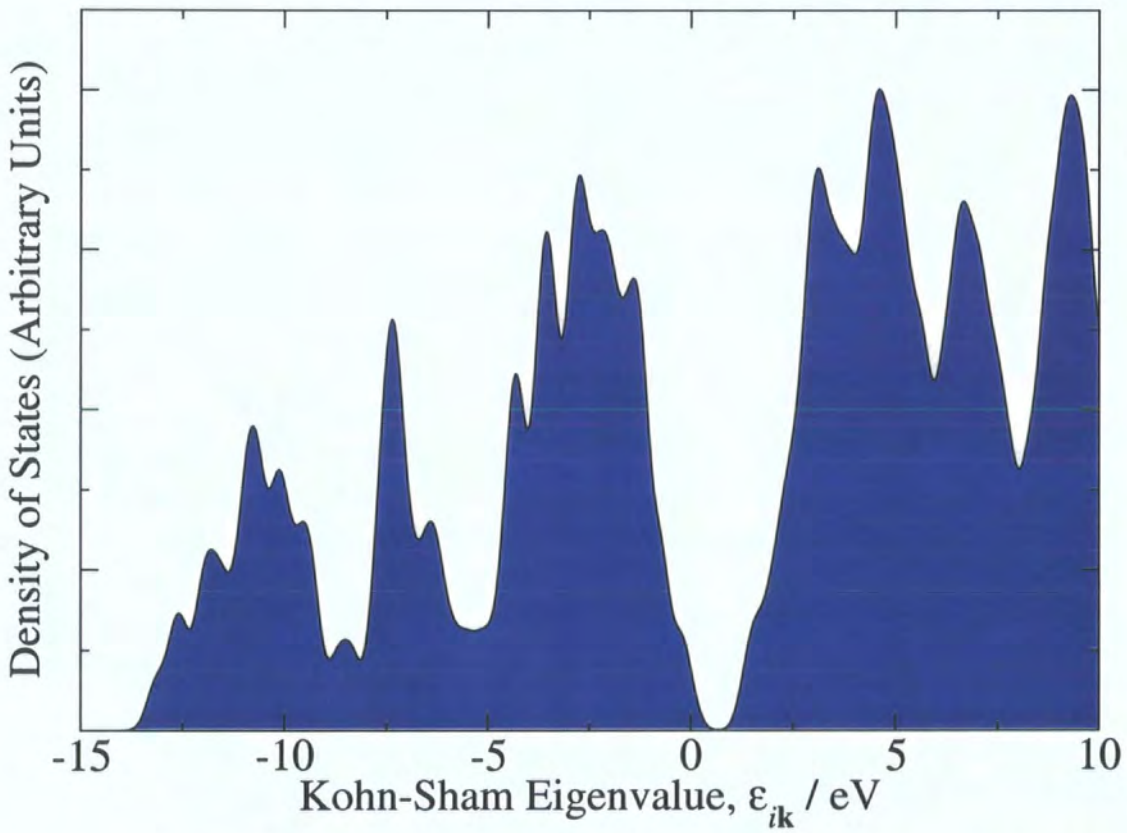


Figure 5.4: Electronic density of states (DOS) in Si, calculated using sX-LDA (the Brillouin zone is sampled with an $8 \times 8 \times 8$ MP grid and the graph is smoothed by Gaussian smearing of width 0.4eV).

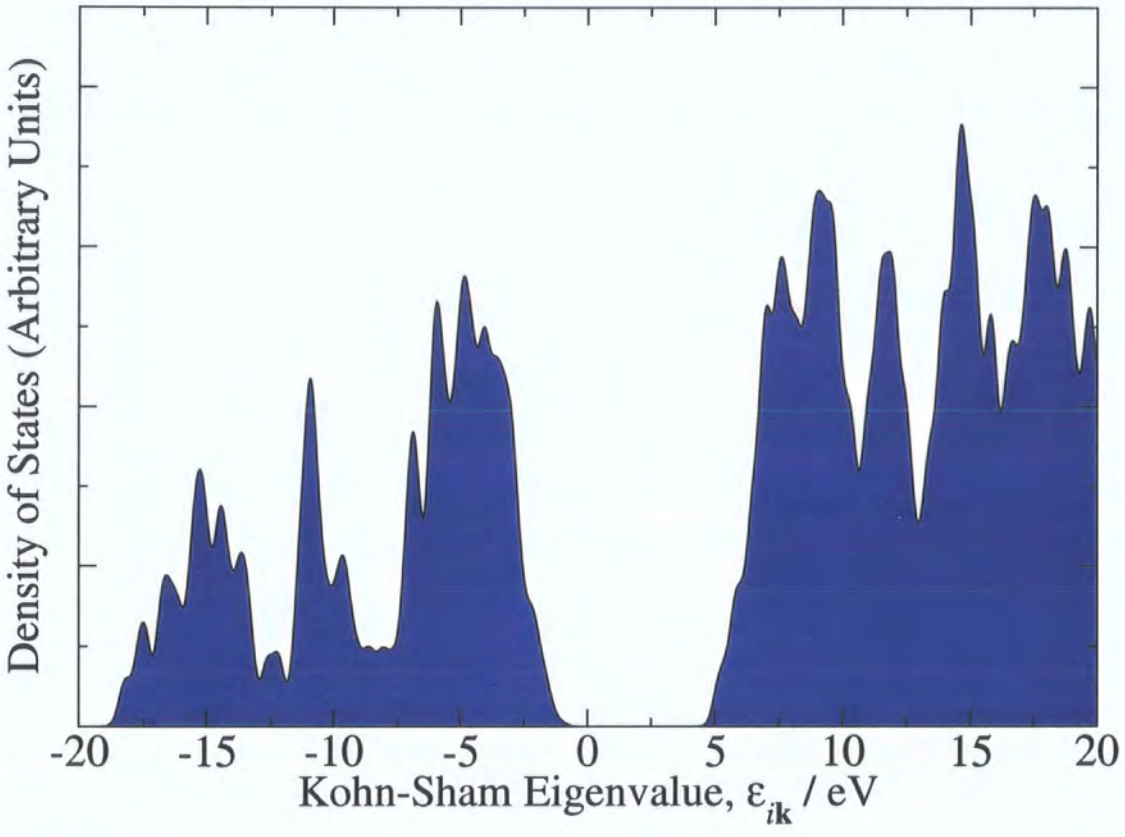


Figure 5.5: Electronic density of states (DOS) in Si, calculated using HF (the Brillouin zone is sampled with an $8 \times 8 \times 8$ MP grid and the graph is smoothed by Gaussian smearing of width 0.4eV).

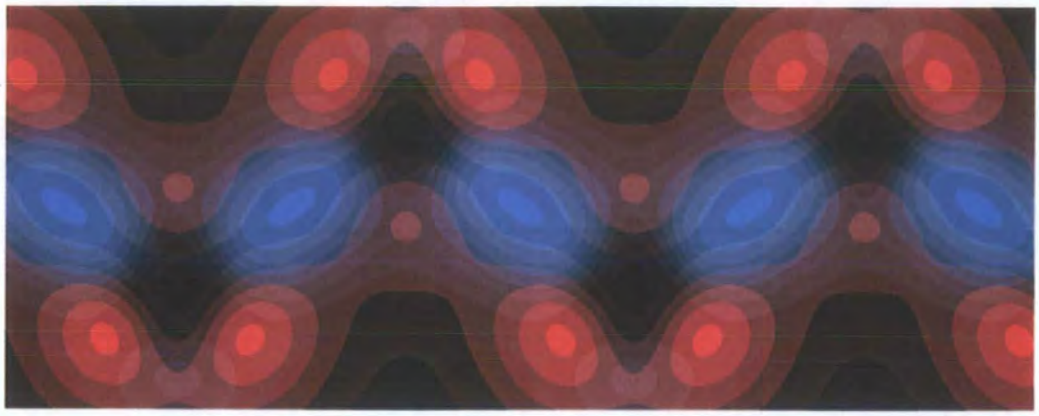


Figure 5.6: Plot of the electron density of the highest valence bands (blue regions) and lowest conduction bands (red regions) on the Γ-point in Si, calculated with the LDA.

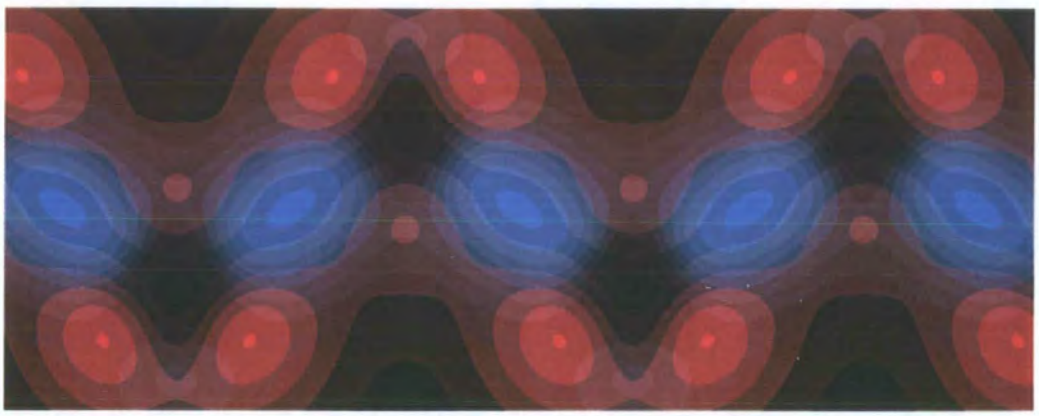


Figure 5.7: Plot of the electron density of the highest valence bands (blue regions) and lowest conduction bands (red regions) on the Γ-point in Si, calculated with sX-LDA.

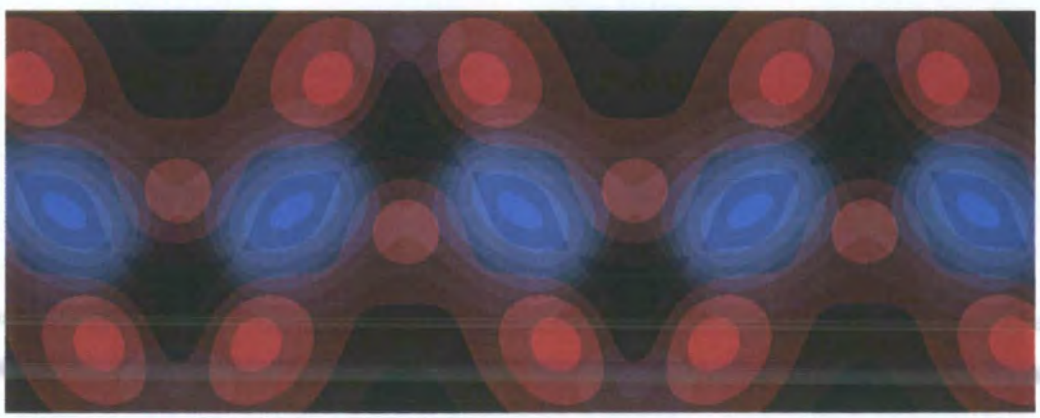


Figure 5.8: Plot of the electron density of the highest valence bands (blue regions) and lowest conduction bands (red regions) on the Γ-point in Si, calculated with HF.

5.1.3 GaN

As part of the study on GaN in Chapter 2, we calculated the band structure of both the wurtzite and zinc blende structures using the LDA, and found that the band gap was underestimated compared to experiment. We now perform band structure calculations using sX-LDA and HF, the results of which are shown in Figures 5.9 and 5.11. Both GaN structures have direct band gaps at Γ, with the calculated values being 2.66eV (WZ), 2.33eV (ZB) with sX-LDA and 9.66eV (WZ), 9.02eV (ZB) with HF. This compares with values of 1.86eV (WZ), 1.69eV (ZB), calculated in Chapter 2 with the LDA. The experimentally measured value for the wurtzite structure is around 3.5eV [54], and for the zinc blende structure the gap is believed to be in the region of 3.3eV [64]. Again, we see that the LDA underestimates the gap, while HF grossly overestimates it. In this case sX-LDA gives the closest result to experiment of the three functionals, but it is not as successful as it was in the case of silicon.

5.2 Other Group IV and III-V Semiconductors

In this section extend the range of materials studied to include most of the other group IV and III-V semiconductors; C (diamond), SiC, Ge, AlN, AlP, AlAs, AlSb, GaP, GaAs, GaSb, InN, InP, InAs, InSb. C and Ge have a cubic diamond structure, while AlN and InN have a wurtzite structure; all the other compounds studied have a zinc blende structure. As well as these materials, for completeness, we also include the results already established previously for Si and the two GaN structures.

5.2.1 Preliminaries

As with Si in the previous section, for each of the semiconducting materials studied here we must first establish appropriate cut-off energies and k-point grids and then calculate the LDA geometries. The parameters and geometries for the diamond and zinc blende structures are shown in Table 5.1, while those for the wurtzite

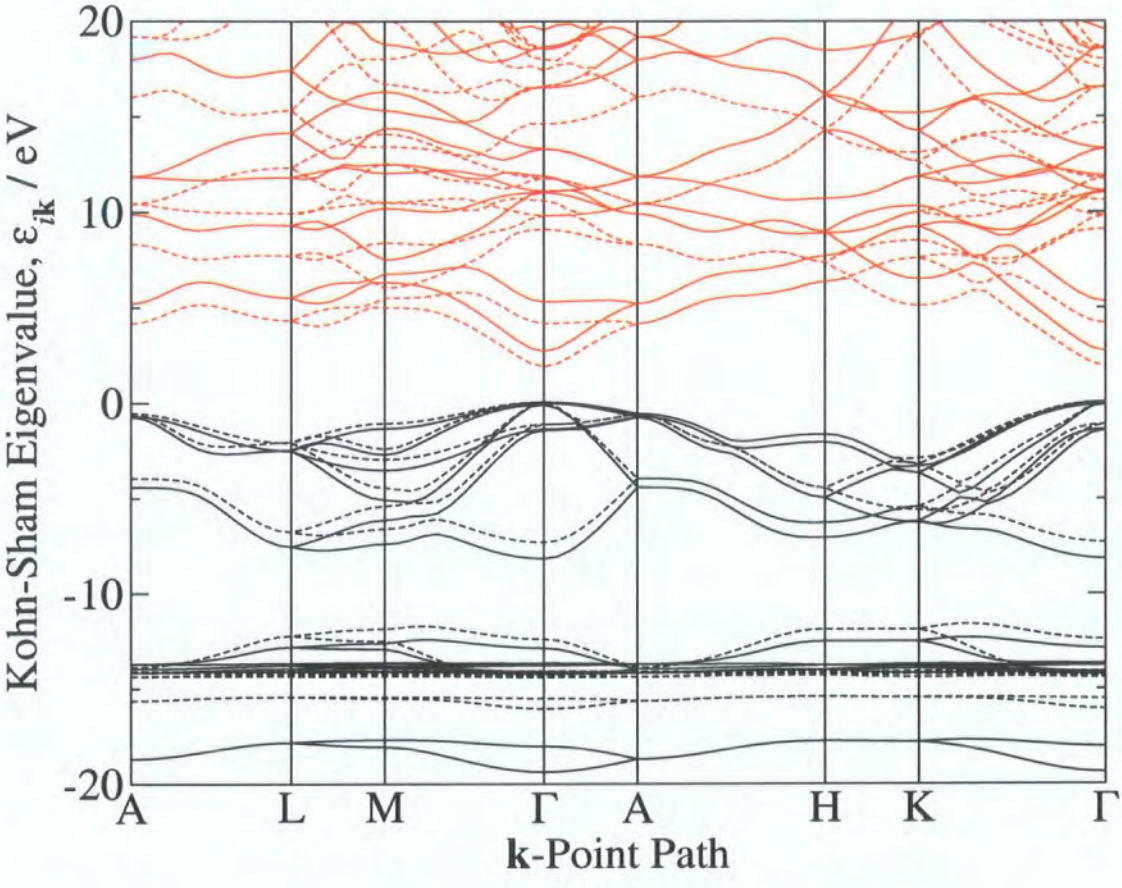


Figure 5.9: Kohn-Sham band structure of wurtzite GaN calculated using sX-LDA. Black lines indicate occupied valence bands, while red lines indicate unoccupied conduction bands. Dashed lines show results calculated with the LDA with all eigenvalues shifted such that the valence band maximum equals that of the sX-LDA calculation.

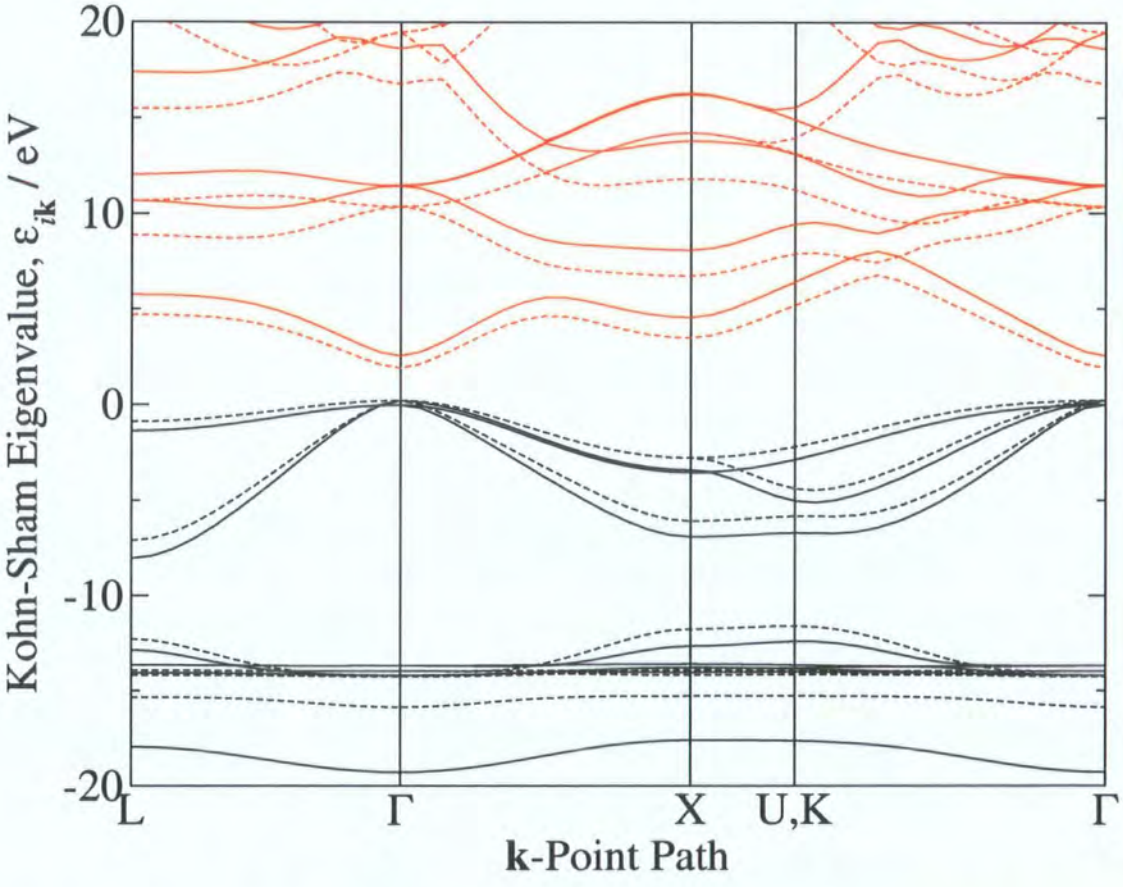


Figure 5.10: Kohn-Sham band structure of zinc blende GaN calculated using sX-LDA. Black lines indicate occupied valence bands, while red lines indicate unoccupied conduction bands. Dashed lines show results calculated with the LDA with all eigenvalues shifted such that the valence band maximum equals that of the sX-LDA calculation.

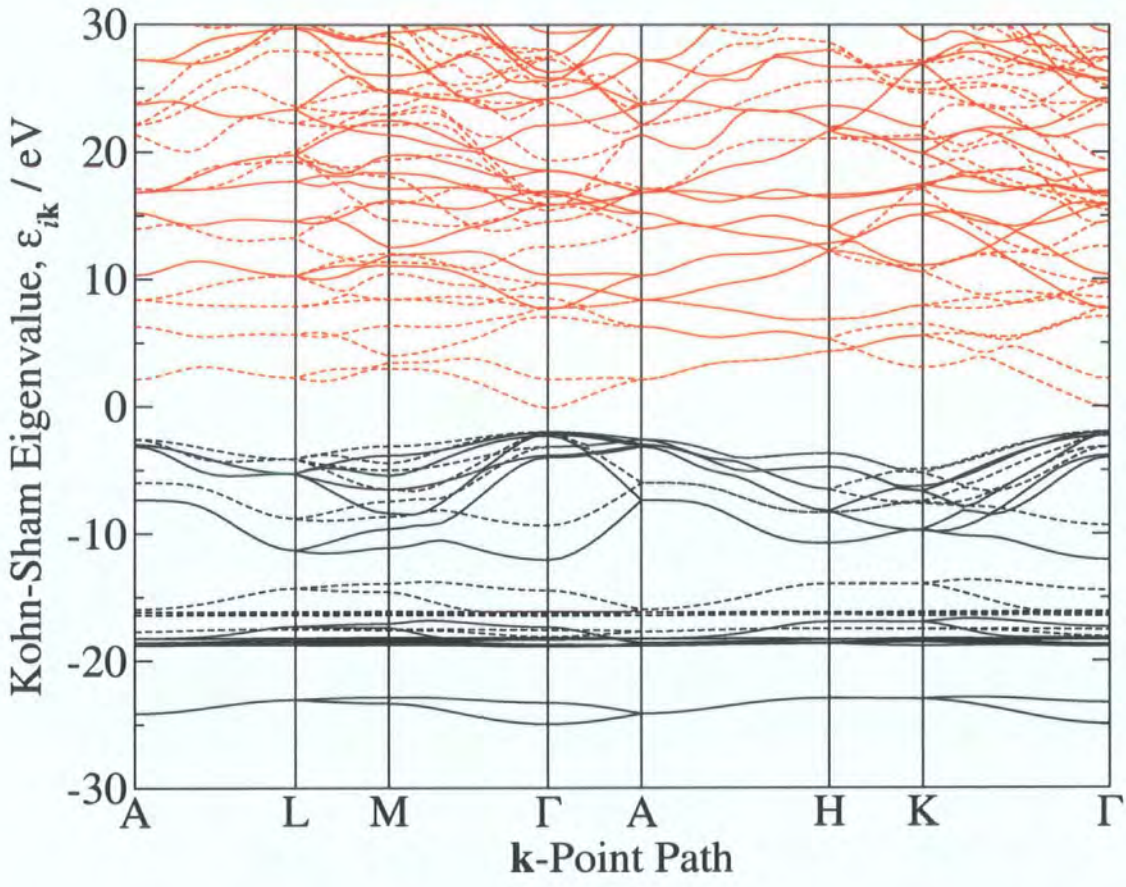


Figure 5.11: Kohn-Sham band structure of wurtzite GaN calculated using HF. Black lines indicate occupied valence bands, while red lines indicate unoccupied conduction bands. Dashed lines show results calculated with the LDA with all eigenvalues shifted such that the valence band maximum equals that of the sX-LDA calculation.

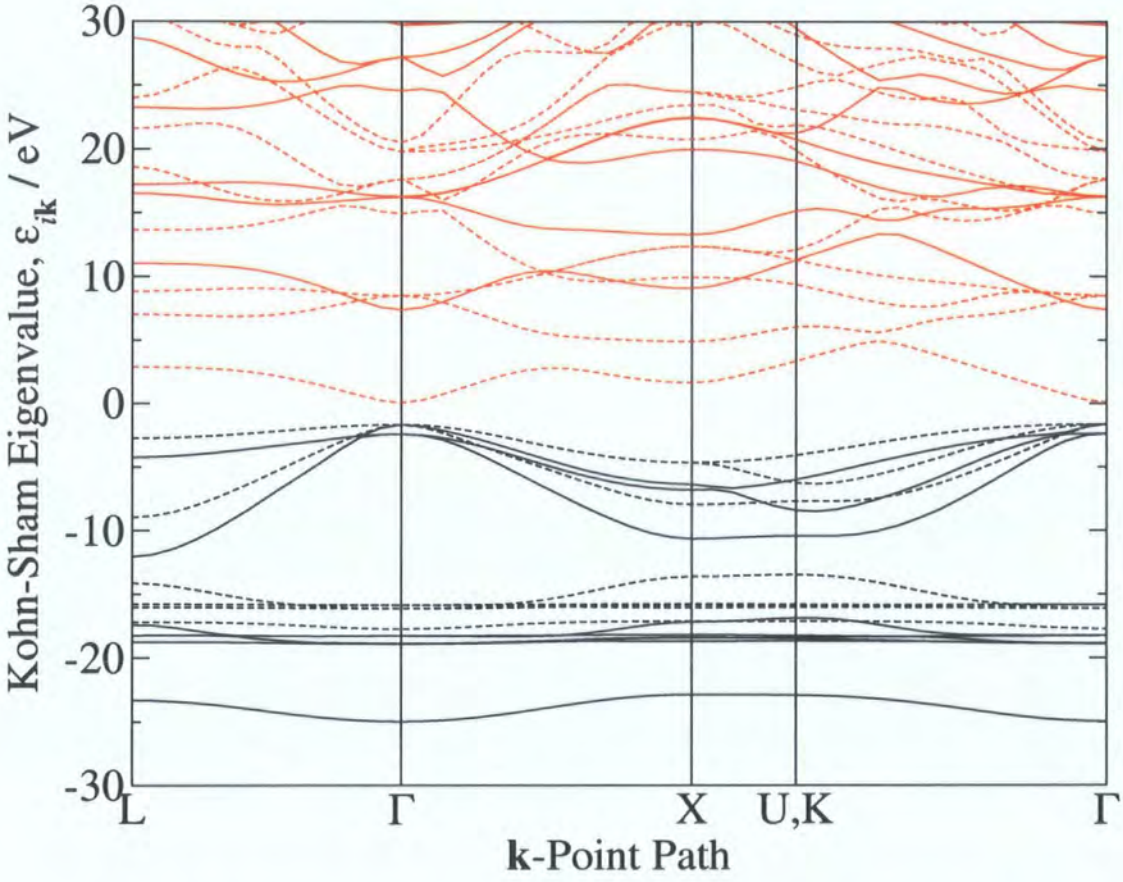


Figure 5.12: Kohn-Sham band structure of zinc blende GaN calculated using HF. Black lines indicate occupied valence bands, while red lines indicate unoccupied conduction bands. Dashed lines show results calculated with the LDA with all eigenvalues shifted such that the valence band maximum equals that of the sX-LDA calculation.

Material	E_{cut}/eV	MP Grid	a/Å(LDA)	a/Å(Exp.)
C	650	3 × 3 × 3	3.53	3.56
Si	350	3 × 3 × 3	5.39	5.43
Ge	750	4 × 4 × 4	5.59	5.66
SiC	600	3 × 3 × 3	4.32	4.36
AlP	500	3 × 3 × 3	5.43	5.46
AlAs	300	3 × 3 × 3	5.62	5.66
AlSb	300	3 × 3 × 3	6.05	6.14
GaN (ZB)	800	3 × 3 × 3	4.50	4.50
GaP	800	3 × 3 × 3	5.40	5.45
GaAs	700	3 × 3 × 3	5.58	5.65
GaSb	700	3 × 3 × 3	5.92	6.10
InP	500	3 × 3 × 3	5.92	5.87
InAs	300	3 × 3 × 3	6.10	6.06
InSb	300	3 × 3 × 3	6.44	6.48

Table 5.1: Parameters for calculations on semiconductors with diamond or zinc blende structures. All MP grids are off-origin. All experimental data from reference [54], except ZB GaN [64].

Material	E_{cut}/eV	MP Grid	a, c/Å(LDA)		a, c/Å(Exp.)	
AlN	600	4 × 4 × 3	3.08	4.96	3.11	4.98
GaN (WZ)	800	4 × 4 × 3	3.18	5.18	3.19	5.19
InN	600	4 × 4 × 3	3.36	5.40	3.54	5.70

Table 5.2: Parameters for calculations on semiconductors with wurtzite structures. All MP grids are off-origin in the horizontal plane and on-origin in the vertical direction. All experimental data from reference [54].

Species	Valence Electrons
C	$2s$, $2p$
N	$2s$, $2p$
Al	$3s$, $3p$
Si	$3s$, $3p$
P	$3s$, $3p$
Ga	$4s$, $3d$, $4p$
Ge	$4s$, $3d$, $4p$
As	$4s$, $4p$
In	$5s$, $4d$, $5p$
Sb	$5s$, $5p$

Table 5.3: Electrons treated as valence for the pseudopotentials of each atomic species used in this work.

structures are shown in Table 5.2. The experimentally measured lattice constants are also shown for comparison. For the Ge and In pseudopotentials, we find that the band structures are very sensitive to whether the d-electrons are treated as core or valence, and so we treat these d-electrons as valence in all of our calculations. A summary of which electrons are treated as valence for each atomic species is given in Table 5.3.

5.2.2 Tabulated Eigenvalues

With the calculation parameters established for each material we proceed to calculate the Kohn-Sham eigenvalue spectrum at the main symmetry points. For cubic and zinc blende structures these are L, Γ, X, and U/K, while for wurtzite structures these are A, L, M, Γ, H, and K. To present the results in a sensible manner, we tabulate the highest valence band eigenvalue and lowest conduction band eigenvalue at each point *relative* to the valence band maximum. The results for the zinc blende structures are shown in table 5.4, while those for the wurtzite structures are shown in table 5.5. We see that, as was the case in Si and GaN, the general trend is that

as we go from the LDA, to sX-LDA, to HF, the eigenvalues of the valence bands are lowered, while those of the conduction band are raised. Again, the reasons for this will be discussed in Section 5.3.

5.2.3 Band Gaps

In most cases, the band gap of each of the materials studied can be read off from Tables 5.4 and 5.5. This is not the case for materials in which the conduction band minimum does not lie on a point of symmetry. In these cases we must perform a continuous band structure calculation along the path on which the minimum lies. Table 5.6 shows the band gaps for all of the materials studied in this work, calculated with the functionals LDA, sX-LDA, and HF, as well as the experimentally measured values (from reference [54], except zinc blende GaN [64] and InN [91]).

To visualise the performance of the various functionals more clearly, and to compare with experiment, we now plot the calculated gaps as a function of the experimentally measured values. The deviation of the calculated gaps from experiment can then be readily seen in terms of their distance from the $y = x$ line; this graph is shown in Figure 5.13.

Inspecting this graph we see very clearly how the calculated gaps depend on the choice of functional. In almost all cases, the LDA underestimates the gap substantially, while HF overestimates the gap massively. In most cases sX-LDA is much closer to the experimental results than the LDA, but still tends to underestimate the gap somewhat. The reasons for these general patterns are discussed in Section 5.3.

There are one or two exceptions to these general patterns, however. Notably, the antimony compounds GaSb and InSb display overestimated gaps, even when using the LDA; sX-LDA overestimated their gaps further still. This may be due to the fact that we have not included d-electrons as valence in the Sb pseudopotential, but an Sb pseudopotential including d-electrons is not currently available to us.

Material	Sym. Pt.	LDA Results		sX-LDA Results		HF Results	
		V.B. Max. / eV	C.B. Min. / eV	V.B. Max. / eV	C.B. Min. / eV	V.B. Max. / eV	C.B. Min. / eV
C	L	-2.85	8.53	-3.49	10.41	-5.07	18.13
	Γ	0.00	5.64	0.00	7.19	0.00	14.05
	X	-6.43	4.75	-7.32	5.84	-8.79	12.51
	K	-5.47	5.67	-6.41	7.01	-8.28	14.30
Si	L	-1.21	1.58	-1.50	2.22	-2.66	7.82
	Γ	0.00	2.57	0.00	3.19	0.00	8.52
	X	-2.91	0.61	-3.18	1.07	-4.11	5.87
	K	-2.47	1.12	-2.82	1.72	-4.16	7.07
Ge	L	-1.63	0.40	-1.84	0.98	-2.73	6.46
	Γ	0.00	0.40	0.00	0.78	0.00	6.45
	X	-3.72	1.64	-1.75	1.07	-4.85	6.76
	K	-2.98	1.32	-3.26	2.11	-4.43	7.47
SiC	L	-1.06	5.58	-1.45	6.77	-2.45	13.87
	Γ	0.00	6.58	0.00	7.79	0.00	14.89
	X	-3.24	1.27	-3.69	2.12	-4.54	8.26
	K	-2.63	3.05	-3.16	4.07	-4.38	10.66
AlP	L	-0.91	2.63	-1.12	3.44	-2.02	9.32
	Γ	0.00	3.04	0.00	3.85	0.00	9.76
	X	-2.37	1.44	-2.56	2.12	-3.35	7.39
	K	-1.97	2.32	-2.22	3.08	-3.29	8.70
AlAs	L	-0.95	2.13	-1.15	2.78	-2.15	8.23
	Γ	0.00	2.19	0.00	2.83	0.00	8.21
	X	-2.40	1.38	-2.59	1.94	-3.44	6.85
	K	-2.01	2.20	-2.25	2.82	-3.41	8.08

table continues next page....

....table continued from previous page

Material	Sym. Pt.	LDA Results		sX-LDA Results		HF Results	
		V.B. Max. / eV	C.B. Min. / eV	V.B. Max. / eV	C.B. Min. / eV	V.B. Max. / eV	C.B. Min. / eV
AlSb	L	-1.02	1.81	-1.20	2.32	-2.19	7.43
	Γ	0.00	2.57	0.00	3.16	0.00	8.27
	X	-2.41	1.42	-2.59	1.84	-3.45	6.30
	K	-2.04	1.85	-2.25	2.33	-3.42	7.22
GaN	L	-1.11	4.49	-1.60	5.54	-2.61	12.66
	Γ	0.00	1.70	0.00	2.33	0.00	9.02
	X	-2.99	3.28	-3.62	4.35	-4.71	10.70
	K	-2.42	4.98	-3.09	6.19	-4.34	12.92
GaP	L	-1.26	1.61	-1.64	2.19	-2.84	7.90
	Γ	0.00	1.81	0.00	2.33	0.00	7.99
	X	-2.95	1.48	-3.37	2.02	-4.49	7.09
	K	-2.49	2.07	-2.97	2.70	-4.38	8.21
GaAs	L	-1.29	1.13	-1.65	1.58	-2.95	6.85
	Γ	0.00	0.97	0.00	1.32	0.00	6.42
	X	-2.97	1.37	-3.35	1.80	-4.53	6.46
	K	-2.51	1.91	-2.96	2.41	-4.46	7.54
GaSb	L	-1.38	1.06	-1.67	1.48	-2.91	6.56
	Γ	0.00	1.50	0.00	1.91	0.00	7.01
	X	-3.01	0.90	-3.34	1.27	-4.48	5.77
	K	-2.57	1.52	-2.94	1.96	-4.40	6.82
InP	L	-0.95	1.58	-1.26	2.08	-2.30	7.51
	Γ	0.00	0.73	0.00	1.10	0.00	6.34
	X	-2.28	1.60	-2.63	2.07	-3.66	6.94
	K	-1.90	2.31	-2.30	2.86	-3.54	8.10

table continues next page....

....table continued from previous page

Material	Sym. Pt.	LDA Results		sX-LDA Results		HF Results	
		V.B. Max. / eV	C.B. Min. / eV	V.B. Max. / eV	C.B. Min. / eV	V.B. Max. / eV	C.B. Min. / eV
InAs	L	-0.96	1.19	-1.26	1.57	-2.40	6.58
	Γ	0.00	0.15	0.00	0.42	0.00	5.10
	X	-2.27	1.49	-2.59	1.86	-3.68	6.36
	K	-1.90	2.14	-2.27	2.59	-3.60	7.47
InSb	L	-1.02	1.22	-1.26	1.57	-2.33	6.48
	Γ	0.00	1.09	0.00	1.41	0.00	6.33
	X	-2.30	1.53	-2.58	1.86	-3.61	6.20
	K	-1.95	1.88	-2.26	2.25	-3.52	6.98

Table 5.4: Kohn-Sham eigenvalues of the highest valence band and lowest conduction bands (relative to the valence band maximum) at the main symmetry points of some diamond/zinc blende semiconductors, calculated with the LDA, sX-LDA, and HF functionals.

Another interesting result is the case of InN, a material whose band gap has been the subject of substantial revision in recent years [91]. Including *d*-electrons as valence results in a gap of zero when using the LDA, whereas using sX-LDA raises this to around 0.4eV. This is still somewhat lower than the experimental value, however.

Our results are consistent with those other groups such as the all-electron calculations of Asahi et. al. [92] and Geller et. al. [93].

Material	Sym. Pt.	LDA Results		sX-LDA Results		HF Results	
		V.B. Max. / eV	C.B. Min. / eV	V.B. Max. / eV	C.B. Min. / eV	V.B. Max. / eV	C.B. Min. / eV
AlN	A	-0.46	6.28	-0.48	8.07	-0.61	15.76
	L	-1.59	5.23	-1.70	7.04	-2.13	14.27
	M	-0.72	5.82	-0.76	7.68	-0.97	14.86
	Γ	0.00	4.11	0.00	5.66	0.00	13.19
	H	-1.03	7.61	-1.12	9.62	-1.44	17.21
	K	-2.64	4.94	-2.79	6.81	-3.46	13.86
GaN	A	-0.61	4.10	-0.75	5.15	-1.00	12.29
	L	-2.11	4.26	-2.53	5.45	-3.25	12.27
	M	-1.12	4.96	-1.40	6.26	-1.86	13.07
	Γ	0.00	1.86	0.00	2.66	0.00	9.66
	H	-1.70	6.31	-2.12	7.65	-2.78	14.83
	K	-2.95	5.04	-3.37	6.43	-4.28	13.08
InN	A	-0.44	2.37	-0.55	3.10	-0.83	9.44
	L	-1.55	3.31	-1.88	4.29	-2.58	10.52
	M	-0.81	3.87	-1.03	4.85	-1.50	11.45
	Γ	0.00	0.00	0.00	0.40	0.00	6.49
	H	-1.25	5.18	-1.58	6.29	-2.24	12.86
	K	-2.07	4.60	-2.37	5.82	-3.23	11.93

Table 5.5: Kohn-Sham eigenvalues of the highest valence band and lowest conduction bands (relative to the valence band maximum) at the main symmetry points of some wurtzite semiconductors, calculated with the LDA, sX-LDA, and HF functionals.

Material	Band Gaps			
	LDA	sX-LDA	HF	Exp.
C	4.15	5.27	12.07	5.50
Si	0.48	0.97	4.78	1.12
Ge	0.40	0.78	6.45	0.66
SiC	1.27	2.12	8.26	2.42
AlN	4.11	5.66	13.19	6.2
AlP	1.44	2.12	7.39	2.51
AlAs	1.38	1.94	6.85	2.15
AlSb	1.33	1.74	6.26	1.62
GaN(WZ)	1.86	2.66	9.66	3.5
GaN(ZB)	1.70	2.33	9.02	3.3
GaP	1.43	1.96	7.04	2.27
GaAs	0.97	1.32	6.42	1.42
GaSb	0.90	1.27	5.77	0.75
InN	0.00	0.40	6.49	0.7
InP	0.73	1.10	6.34	1.34
InAs	0.15	0.42	5.10	0.35
InSb	1.09	1.41	6.20	0.23

Table 5.6: Band gaps of a range of group IV and III-V semiconductors calculated with the LDA, sX-LDA, and HF functionals. Experimental data is also tabulated for comparison (from reference [54] except zinc blende GaN [64] and InN [91]).

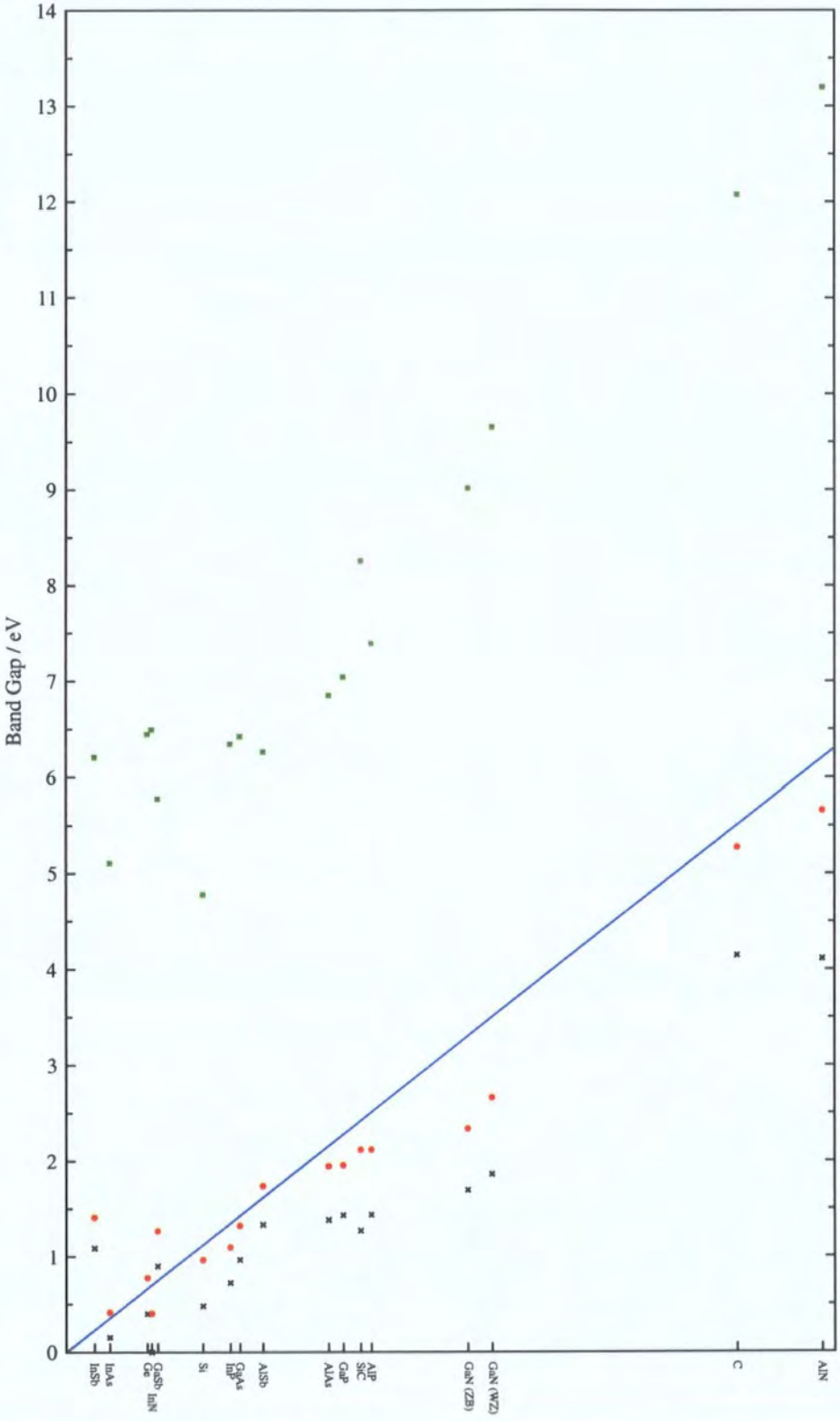


Figure 5.13: Band gaps of a range of group IV and III-V semiconductors calculated with the LDA (black crosses), sX-LDA (red dots), and HF (green squares). Experimental data is represented by the blue line (from reference [54] except zinc blende GaN [64] and InN [91]).

5.3 Discussion of Results

We have seen that, in terms of band structure calculations, the LDA tends to underestimate band gaps, HF always overestimates band gaps, while sX-LDA tends to underestimate gaps somewhat, but to a far lesser extent than the LDA. We will now discuss some of the reasons for the variation in performance of these different functionals.

We should begin with the actual definition of the band gap, so we can understand what it is that we are actually aiming for. The band gap, E_g, is the minimum energy of an electronic excitation in which there is no interaction between the promoted electron and the hole that it leaves behind. E_g can thus be defined in terms of an $N+1$ electron system and a separate $N-1$ electron system as follows

$$\begin{aligned} E_g &= [E(N+1) - E(N)] + [E(N-1) - E(N)] \\ &= E(N+1) + E(N-1) - 2E(N). \end{aligned} \tag{5.2}$$

In terms of the Kohn-Sham description of the system, the band gap is usually taken to be the difference between the eigenvalues of the conduction band minimum and the valence band maximum. This is based on the interpretation of the Kohn-Sham eigenvalues as the functional derivative of the energy with respect to the occupancy of the corresponding orbitals. This would be the case if $(E[\rho] - T_S[\rho])$ was a well behaved functional of the density, but unfortunately, as was discussed for example by Sham and Schlüter [94, 95], there are discontinuities in its first derivative at integer values of N for changes in $\rho(r)$ that do not conserve charge. It is the size of the discontinuity that determines the difference between the "exact" Kohn-Sham band gap (i.e. the gap we would obtain from an exact exchange-correlation functional), and the true band gap as defined in Equation (5.2). It would, in principle, be possible to calculate the true band gap with an exact exchange-correlation functional by performing three separate ground state energy calculations on $(N-1)$, N, and $(N+1)$ electron systems. This would then tell us precisely the size of the error in interpreting the Kohn-Sham band gap as the true band gap.

We will now discuss the failure of LDA calculations to give the correct band gap.

Supposing we were to perform a series of total energy calculations with the LDA on $(N-1)$, N, and $(N+1)$ electron systems, we could use Equation (5.2) to calculate the "true" band gap for the LDA. But, if we look at the functional $(E[\rho]-T_S[\rho])$ for the LDA we see that it it does not have the discontinuities that are present in the exact functional, and that therefore the "true" LDA gap must be the same as the Kohn-Sham LDA gap. This means that the underestimation of the band gap with the LDA is not simply due to the incorrect interpretation of the Kohn-Sham eigenvalues, but is actually a problem with the way the LDA treats exchange and correlation. The functional $E_{XC}^{LDA}[\rho]$, being a direct, well behaved, functional of the density, is not sensitive to particle number and hence does not deal correctly with issues such as the self-interaction of electrons that is present in the Hartree energy. While approximate methods exist of correcting the sefl-interaction in the LDA [96], these tend to require a localised basis set and a somewhat artificial distinction between orbitals localised mostly same atom and those localised mostly on different atoms.

The Hartree-Fock method can be viewed as an improved way of dealing with exchange and correlation in the sense that, by using the exact definition of the exchange energy in terms of the orbitals, the self-interaction of the electrons in the Hartree energy is fully removed. However, this is only true for valence band orbitals - the self-interaction for conduction band orbitals is completely uncorrected. This explains the very large band gaps obtained from HF calculations; the energy of the valence electrons is lowered by the removal of the self-interaction energy, but the energy of the conduction bands is increased, relative to the LDA, as there is no correction whatsoever to their self-interaction energy.

Screened exchange can be viewed, in a sense, as a compromise between the LDA and HF approaches. Some of the improved treatment of exchange and correlation in the HF method is included, i.e. there is effectively non-local information about the density in the functional. Correlation is also incorporated into the non-local part by screening the exchange interaction at long range. This means that the non-local exchange-correlation energy is much lower that the exchange energy in HF, so the problem of the lack of self-interaction correction for conduction bands suffered by HF is much less pronounced in sX-LDA. Also, any self-interaction corrections

present in the local component of the exchange-correlation potential apply equally to both valence and conduction bands, as they should. One problem that seems to be suffered by screened exchange is related to the lattice parameters that are used. In some of our earlier calculations [97] in which the, usually larger, experimentally measured lattice parameters are used, the band gaps tended to be somewhat wider, and closer to experiment than the ones obtained in this work using LDA lattice parameters. This problem is unlikely to be fixed by using sX-LDA lattice parameters either, since, as we will see in the next chapter, sX-LDA seems to give lattice parameters even lower than the LDA.

If we look at the actual eigenvalues calculated with the various functionals the general trend is that, as we go from the LDA to sX-LDA to HF, the valence band eigenvalues are lowered, while the conduction band eigenvalues are raised. This can be understood in terms of the above discussion about the self-interaction energy. It raises the question though of what happens to the energies of defect states that lie somewhere within the band gap. When using local functionals, it is possible to artificially add a rigid shift to conduction band eigenvalues, but the question of what to do with defect sates is unclear. Our implementation of sX-LDA has recently been used to calculate the energies of defect states in HfO_2, where it was found to give good results for the positioning of defect levels within the gap [98] unlike treatments that simply combine the LDA with a rigid shift.

Although we have not performed EXX calculations ourselves, this functional has been reported to give band gaps in very good agreement with experiment [79]. This can be attributed to the fact that EXX correctly cancels the self-interaction for both valence and conduction bands.

In conclusion, we have shown that our implementation of the functionals sX-LDA and HF can be used to calculate the band structures of semiconductors. The general pattern is that the LDA underestimates band gaps, HF overestimates band gaps substantially, and sX-LDA underestimates gaps somewhat, but the results are usually much closer to experiment than the LDA. These patterns can be explained in terms of the discontinuity in the derivative of the exchange-correlation functional, and the extent to which the self-interaction energy is corrected with the various functionals.

Chapter 6

Calculating Stress with Non-Local Functionals

The main application of non-local functionals so far, at least in the context of condensed matter physics in extended systems, has been the calculation of semiconductor band structures. However, the applications of DFT in general are far wider than this. A very important application, as we saw for example in Chapter 2, is the calculation of cell geometries. Related to this is *molecular dynamics* in which DFT is combined with Newtonian mechanics to model the motion of the atoms. In both of these applications, we usually need to consider changes not only in the atomic positions, but also in the size and shape of the unit cell. In particular, we need to know how such changes affect the total energy of the system, and this requires us to calculate the *stress tensor*, $\sigma_{\alpha\beta}$. This tensor can always be evaluated numerically by taking finite differences, but this would require several evaluations of the total energy. Far more efficient methods already exist for evaluating $\sigma_{\alpha\beta}$ when using simple exchange-correlation functionals such as the LDA, but, to the author's knowledge, no work has been done so far on the calculation of stress when using non-local functionals such as sX-LDA, HF, and EXX. In this chapter we introduce the theory, derived by ourselves [99], which allows us to calculate the contribution to $\sigma_{\alpha\beta}$ from non-local exchange and correlation effects when using sX-LDA, HF, and EXX. We also describe how this has been implemented as an addition to the code

described in Chapter 4, and apply it to a number of test cases.

6.1 Theory

6.1.1 The Stress Tensor

In the context of a cell periodic system, stress and strain are properties related to changes in the size and shape of the unit cell. The *unsymmetrised strain tensor*, $\epsilon_{\alpha\beta}$, is defined in terms of a scaling of space, i.e.

$$\mathbf{r}_\alpha \rightarrow \sum_{\alpha\beta}(\delta_{\alpha\beta} + \epsilon_{\alpha\beta})\mathbf{r}_\beta, \tag{6.1}$$

where $\mathbf{r}$ is any position vector in real space. Under such a transformation, the lattice vectors of the unit cell will change, as will the positions of the atoms, however the *fractional coordinates* of the atoms will remain fixed. In reciprocal space, the scaling is given by

$$\mathbf{K}_\alpha \rightarrow \sum_{\alpha\beta}(\delta_{\alpha\beta} - \epsilon_{\alpha\beta})\mathbf{K}_\beta, \tag{6.2}$$

for a general reciprocal space vector, $\mathbf{K}$.

If the structure is under stress, then there will be a first order change in its internal energy in response to a first order change in the strain tensor. The *stress tensor*, $\sigma_{\alpha\beta}$, is defined as

$$\sigma_{\alpha\beta} = -\frac{1}{\Omega}\frac{\partial E_{TOT}}{\partial \epsilon_{\alpha\beta}}. \tag{6.3}$$

Hence, in order to calculate the stress tensor we essentially need to be able to differentiate the total energy per cell with respect to either the real or reciprocal space lattice vectors.

6.1.2 Stress in the Kohn-Sham Framework

The theory of stress within the Kohn-Sham framework, when using local functionals, is well established [100, 101, 102]; here, we briefly describe it. In the standard Kohn-Sham framework, the total energy, E_{TOT}, is given by

$$E_{TOT} = T_S + V_{ext} + V_H + E_{XC} + V_{I-I}, \tag{6.4}$$

as described in Chapter 1. The stress tensor can therefore be expressed as

$$\begin{aligned}\sigma_{\alpha\beta} &= \frac{1}{\Omega}\frac{\partial}{\partial\epsilon_{\alpha\beta}}(T_S + V_{ext} + V_H + E_{XC} + V_{I-I}) \\ &= \frac{1}{\Omega}\frac{\partial T_S}{\partial\epsilon_{\alpha\beta}} + \frac{1}{\Omega}\frac{\partial V_{ext}}{\partial\epsilon_{\alpha\beta}} + \frac{1}{\Omega}\frac{\partial V_H}{\partial\epsilon_{\alpha\beta}} + \frac{1}{\Omega}\frac{\partial E_{XC}}{\partial\epsilon_{\alpha\beta}} + \frac{1}{\Omega}\frac{\partial V_{I-I}}{\partial\epsilon_{\alpha\beta}}.\end{aligned} \tag{6.5}$$

The terms T_S, V_{ext}, V_H, and E_{XC} all depend, either explicitly or implicitly, on the orbital coefficients, $\{c_{\mathbf{ik}}(\mathbf{G})\}$. A first order change in the strain tensor, $\epsilon_{\alpha\beta}$ does in general result in a first order change in the coefficients, and there is a resulting component of the change in T_S, V_{ext}, V_H, and E_{XC} that can be ascribed to changes in these coefficients. However, the sum of these four terms is the total electronic energy, E, which is minimised in the ground state and is therefore *stationary* with respect to the coefficients. We can therefore ignore changes in T_S, V_{ext}, V_H, and E_{XC} due to the changing coefficients, because they will sum to zero. This simplifies the expressions for the different contributions to the stress. We are effectively treating the coefficients as being constant with respect to small changes in the size and shape of the unit cell. This will only work, however, if we have the correct normalisation conditions on $c_{\mathbf{ik}}(\mathbf{G})$ and $\rho(\mathbf{G})$ so that normalisation is preserved when the volume of the cell changes. This has been achieved with the appropriate positioning of the reciprocal volume factors in the definitions of the Fourier transforms in Equations (1.116), (1.120), and (1.122) of Chapter 1.

6.1.3 Non-Local Functionals

As we have just discussed, the contribution to the stress tensor from a given term in the Kohn-Sham energy sum can be obtained by taking the partial derivative of that term with respect to the strain tensor. In order to do this for the non-local exchange-correlation energy in sX-LDA, HF, and EXX, we make use of two important results, which are

$$\frac{\partial\Omega}{\partial\epsilon_{\alpha\beta}} = \Omega\delta_{\alpha\beta}, \tag{6.6}$$

and

$$\frac{\partial f(|\mathbf{K}|)}{\partial\epsilon_{\alpha\beta}} = -f'(|\mathbf{K}|)\frac{\mathbf{K}_\alpha\mathbf{K}_\beta}{|\mathbf{K}|}. \tag{6.7}$$

Applying these to the non-local exchange-correlation energy we thus have

$$\begin{aligned}
\sigma_{\alpha\beta}^{NLXC} &= \frac{1}{\Omega}\frac{\partial}{\partial\epsilon_{\alpha\beta}}E_{XC}^{NL} \\
&= -\frac{2}{\pi\Omega}\frac{\partial}{\partial\epsilon_{\alpha\beta}}\left[\frac{1}{\Omega}\sum_{\mathbf{ikjq}}\sum_{\mathbf{G}}\frac{|C_{\mathbf{jqik}}(\mathbf{G})|^2}{|\mathbf{q}-\mathbf{k}+\mathbf{G}|^2+k_s^2}\right] \\
&= -\frac{2}{\pi\Omega^2}\sum_{\mathbf{ikjq}}\sum_{\mathbf{G}}\frac{|C_{\mathbf{jqik}}(\mathbf{G})|^2}{|\mathbf{q}-\mathbf{k}+\mathbf{G}|^2+k_s^2} \\
&\quad\times\left(\frac{2(\mathbf{q}-\mathbf{k}+\mathbf{G})_\alpha(\mathbf{q}-\mathbf{k}+\mathbf{G})_\beta}{|\mathbf{q}-\mathbf{k}+\mathbf{G}|^2+k_s^2}-\delta_{\alpha\beta}\right), \qquad (6.8)
\end{aligned}$$

which is the contribution to the stress from the non-local exchange correlation energy (where k_s is zero in the case of HF and EXX).

In practical applications, we may be using the divergence correction or parallelepiped integration to deal with the singularity, as discussed in Section 3.3. As this changes the expression for the energy, it will also affect the calculation of the stress tensor. In the case of parallelepiped integration the integrals simply replace the values of $1/(|\mathbf{q}-\mathbf{k}+\mathbf{G}|^2)$. However, if we are using the divergence correction then we must consider derivatives of the extra term in the energy with respect to the strain tensor, i.e. there will be an extra contribution to the stress of

$$\begin{aligned}
&\frac{1}{\Omega}\frac{\partial}{\partial\epsilon_{\alpha\beta}}\left[\frac{2\pi}{\Omega}\sum\mathbf{k}N(\mathbf{k})\left(\sum_{\mathbf{q}}F(\mathbf{q}-\mathbf{k})-\Omega\int_{B.Z.}d\mathbf{q}F(\mathbf{q}-\mathbf{k})\right)\right] \\
&= \frac{2\pi}{\Omega}\sum_{\mathbf{k}}N(\mathbf{k})\frac{\partial}{\partial\epsilon_{\alpha\beta}}\left(\frac{1}{\Omega}(\sum_{\mathbf{q}}F(\mathbf{q}-\mathbf{k})-\Omega\int_{B.Z.}d\mathbf{q}F(\mathbf{q}-\mathbf{k})\right).
\end{aligned}$$

Assuming we are using the form for F given in Equation (3.49), and that the width of the envelope, w, is constant, then the integral is constant with respect to the lattice vectors and therefore makes no contribution to the stress. The sum can be differentiated as follows:

$$\frac{\partial}{\partial\epsilon_{\alpha\beta}}\left(\frac{1}{\Omega}\sum_{\mathbf{q}}F(\mathbf{q}-\mathbf{k})\right)=\frac{\partial}{\partial\epsilon_{\alpha\beta}}\left(\frac{1}{\Omega}\sum_{\mathbf{q}}\sum_{\mathbf{G}}\frac{S(|\mathbf{q}-\mathbf{k}-\mathbf{G}|)}{|\mathbf{q}-\mathbf{k}-\mathbf{G}|^2}\right)$$

$$\begin{aligned}=\ & \frac{-1}{\Omega|\mathbf{q}-\mathbf{k}-\mathbf{G}|^2}\sum_{\mathbf{q}}\sum_{\mathbf{G}}\Big(S(|\mathbf{q}-\mathbf{k}-\mathbf{G}|)\delta_{\alpha\beta}\\ & +S'(|\mathbf{q}-\mathbf{k}-\mathbf{G}|)\frac{(\mathbf{q}-\mathbf{k}-\mathbf{G})_\alpha(\mathbf{q}-\mathbf{k}-\mathbf{G})_\beta}{|\mathbf{q}-\mathbf{k}-\mathbf{G}|}\\ & -S(|\mathbf{q}-\mathbf{k}-\mathbf{G}|)\frac{2(\mathbf{q}-\mathbf{k}-\mathbf{G})_\alpha(\mathbf{q}-\mathbf{k}-\mathbf{G})_\beta}{|\mathbf{q}-\mathbf{k}-\mathbf{G}|^2}\Big).\end{aligned} \tag{6.9}$$

This requires us to be able to differentiate the envelope function, $S(x)$. In the case of the simple form of Equation (3.50), we have

$$\begin{aligned}S'(x)_{|x|\leq w} &= -\frac{\pi}{w}\sin(\frac{\pi x}{w}),\\ S'(x)_{|x|>w} &= 0.\end{aligned} \tag{6.10}$$

6.2 Implementation and Tests

6.2.1 Implementation in CASTEP

The calculation of $\sigma_{\alpha\beta}^{NLXC}$ has been implemented in CASTEP along with the main NLXC code described in Chapter 4. The implementation follows the same procedure described in Chapter 4 for evaluating the functions $C_{j\mathbf{q}i\mathbf{k}}(\mathbf{G})$, and making use of the list of values of $1/(|\mathbf{q}-\mathbf{k}+\mathbf{G}|^2+k_s^2)$ stored in memory. The computational cost of calculating the stress is similar to a single evaluation of the total energy. However, the stress only needs to be calculated once, after the completion of the electronic minimisation procedure, so the additional cost of calculating the stress is insignificant.

If we are using parallelepiped integration to deal with the singularity then we must replace the values of $1/(|\mathbf{q}-\mathbf{k}+\mathbf{G}|^2)$ with the appropriate integral from Equation

(3.53). If we are using the divergence correction then we need to calculate the correction to the stress tensor as well as the correction to the eigenvalues, as described in the previous section.

6.2.2 Tests on Silicon

As a test of the implementation, we set up an 8-atom cubic silicon cell, and apply both hydrostatic and shear strain to the structure.

This is done by performing a series of sX-LDA total energy calculations with varying lattice vectors, using a 350eV plane wave cut-off energy, and a $4 \times 4 \times 4$ Monkhorst-Pack grid.

In the case of hydrostatic strain, all three lattice vectors remain equal to each other in length, i.e. $(a = b = c)$, with the lattice constant, a, varying between 5.19Å and 5.47Å in steps of 0.02Å.

At each value of a, we calculate both the total energy, the total stress, and the non-local exchange-correlation contribution to the stress. The results of these calculations are shown in Figure 6.1. We would expect that, if the theory and implementation are working correctly, the diagonal components of the total total stress should equal zero at the point where the total energy is minimised. We see from the figure that this is the case[1], and, since the non-local exchange-correlation contribution is significant, we can conclude that the calculation of the stress is consistent with the calculation of the total energy.

In the case of shear strain we vary the x-component of the lattice vector **b**, while keeping **a** and **c** fixed. The shear strain, $\Delta x/y$, is varied between -0.35 and $+0.35$. At each value of $\Delta x/y$, we again calculate both the total energy, the total stress, and the non-local exchange-correlation contribution to the stress. The results of these calculations are shown in Figure 6.2. This time, we expect the off-diagonal component of the stress tensor corresponding to the applied shear strain to equal zero at

[1]The small discrepancy can be attributed to the finite basis set, and is of a similar magnitude to that found when using the LDA

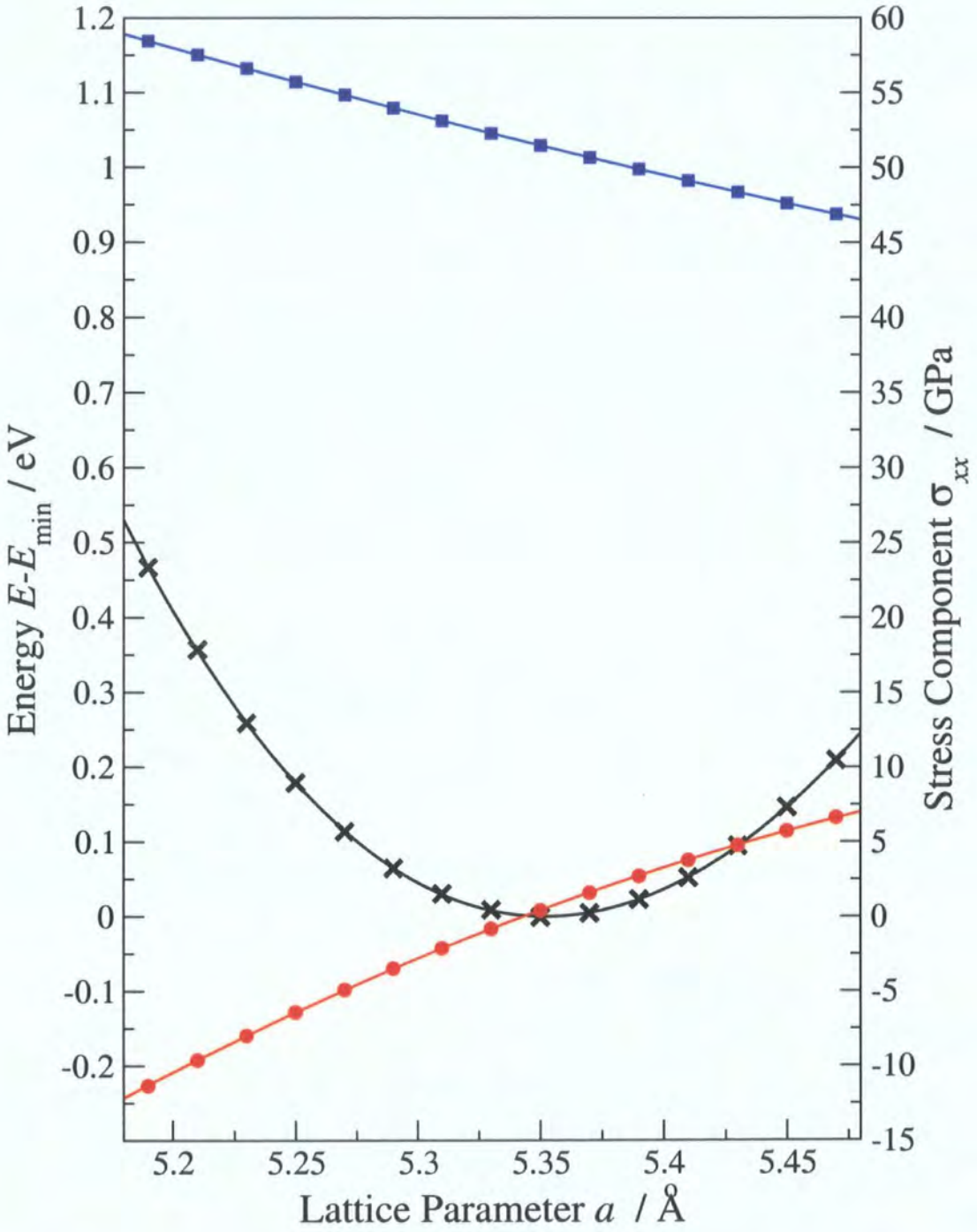


Figure 6.1: Total energy per 8-atom unit cell, calculated using screened exchange, relative to the energy at equilibrium lattice constant (black line), along with the diagonal component of the total stress tensor, σ_{xx} (red line), and the contribution to this stress component from the non-local screened exchange term, σ_{xx}^{NLXC} (blue line), as a function of the lattice parameter a $(= b = c)$ under hydrostatic strain.

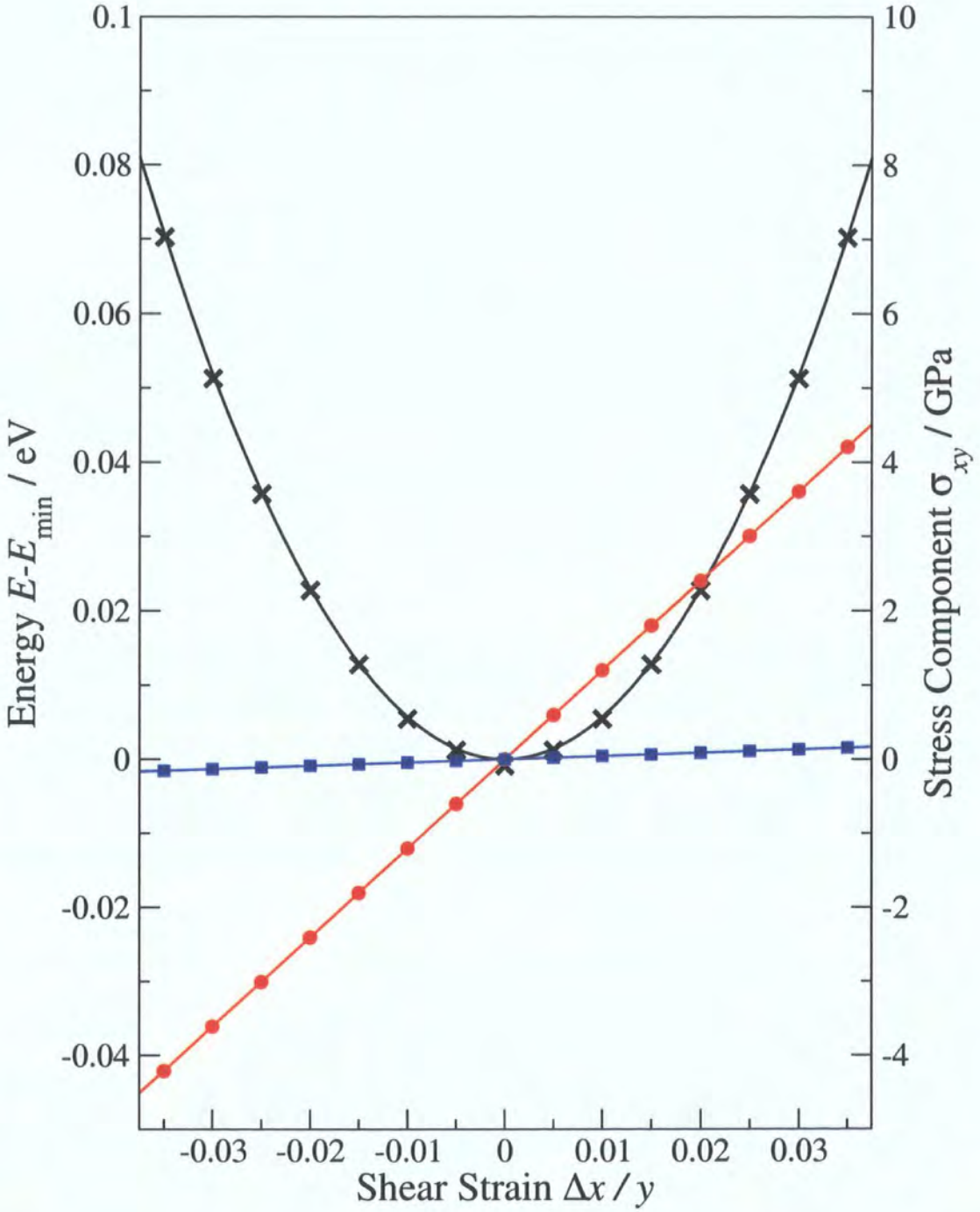


Figure 6.2: Total energy per 8-atom unit cell, calculated using screened exchange, relative to the energy at equilibrium lattice constant (black line), along with the off-diagonal component of the total stress tensor, σ_{xy} (red line), and the contribution to this stress component from the non-local screened exchange term, σ_{xy}^{NLXC} (blue line), as a function of the shear strain, $\Delta x/y$.

the point at which the total energy is minimised, although this result is rather trivial due to symmetry. We do observe, however, that the non-local exchange-correlation contribution to the stress is very small for these off-diagonal components, but is not zero. This differs from the case of the LDA in which off-diagonal contributions are always zero.

6.3 Summary and Conclusions

We have successfully derived and implemented the theory for calculating stress when using non-local orbital-based functionals such as sX-LDA, HF, and EXX, which will allow efficient geometry optimisation and variable-cell molecular dynamics when using these functionals. The computational expense of the method is insignificant compared to the cost of a self-consistent total energy calculation.

It is worth noting that the lattice parameter obtained for silicon using sX-LDA is lower than both the LDA and experimental values [54]. This may be a property of the functional itself, after all sX-LDA has not been extensively tested as to its accuracy when calculating geometries. It may also be due to the fact that we are using LDA pseudopotentials which may not be appropriate for screened exchange calculations. However the question of how best to define a pseudopotential appropriate for screened exchange remains open.

It is likely that most, if not all, major developments in exchange-correlation functionals in the future will use EXX to treat exchange. Any such calculations involving efficient geometry optimisation, or variable-cell molecular dynamics, will require calculation of the exchange contribution to the stress tensor. The future scope of this work should therefore be extensive.

Chapter 7

Conclusions and Avenues for Further Work

In this chapter we summarise the conclusions of the previous chapters. We then explore some possible avenues for further work related to the general area of exchange-correlation functionals in DFT.

7.1 Summary of Conclusions

This work has centred on the problem of treating exchange and correlation in DFT calculations. In Chapter 1 we outlined the basic theory of Kohn-Sham DFT as a means of effectively solving the many-electron Schrödinger equation that describes most of the physics of condensed matter. We then described how this theory can be implemented with a plane wave pseudopotential approach, using as an example the CASTEP code, which we have used for all the calculations in this work. The issue of exchange and correlation was discussed in some detail, although at this stage the only exchange-correlation functionals mentioned were the basic local functionals, namely the LDA and GGAs.

In Chapter 2, we applied this theory, using the LDA, to calculate a number of properties of GaN. These included bulk geometric and energetic properties, surface

reconstructions in the presence of hydrogen, and electronic band structures. It was found that the performance of the LDA depended on which properties were being calculated. While the calculated geometric properties were very accurate, the functional performed very poorly in the calculation of band structures in particular. This provided the motivation for us to consider using more advanced, non-local, exchange-correlation functionals, which were to be the subject of the remainder of the work.

We began this in Chapter 3 by describing the non-local functionals sX-LDA and HF. These functionals are more advanced than the LDA because they incorporate non-local information by treating exchange effects explicitly in terms of the Kohn-Sham orbitals. They involve the re-definition of the Kohn-Sham potential to include a non-local operator component in the exchange-correlation potential. After describing the basic theory of these functionals, we proceeded to show how the theory can be cast in a manner that lends itself to efficient computational implementation. We also then described the EXX functional, which, like HF, defines the exchange energy exactly in terms of the orbitals, but maintains the strictly local nature of the potential in the standard Kohn-Sham framework. As well as describing the definition of the EXX functional, we also described the OEP method that is required in order to perform EXX calculations.

Having explained the theory of the non-local functionals, we proceeded in Chapter 4 to describe our computational implementation of sX-LDA and HF, and also of EXX via the OEP method. This included a description of how cell symmetry and multi-processor parallelisation have been utilised in order to improve the efficiency of the calculations. We presented the results of performance tests that confirmed that sX-LDA and HF calculations scaled as expected with respect to cut-off energy and number of **k**-points, and also that the utilisation of cell symmetry and parallelisation increases the speed of calculations significantly. We found that distribution of **k**-points results in approximately linear scaling with number of processors, whereas distribution of **G**-vectors is less favourable due to the larger amount of inter-processor communication entailed, but can still result in significant increases in speed. We described the implementation of EXX via the OEP method; while

this implementation is essentially complete, further optimisation will be necessary before it can be used for practical calculations.

With the implementation of sX-LDA and HF in place, we applied these functionals to the calculation of semiconductor band structures. Beginning with Si and GaN, we calculated the band structures along continuous paths between points of symmetry in the Brillouin zone. We found that, while the LDA underestimated the band gap significantly, sX-LDA gave gaps significantly closer to experimentally measured values; HF performed very badly, overestimating the band gaps severely. As well as the band structure, we also calculated the density of states of Si and examined 2D plots or orbital charge densities when using these non-local functionals and compared them to LDA results. We found that the density of states was similar in appearance for all three functionals, but displaying the larger gaps from the non-local functionals, and, in the case of HF, stretched over a larger range of energy. The orbital charge density for the highest valence band is not visibly different from the LDA for either of the non-local functionals, but there is a clear difference for the lowest conduction band, particularly in the case of HF. To widen the range of materials studied, we proceeded to calculate the band structures of most of the other group IV and III-V compounds. From the results we established the general trend that the LDA significantly underestimates band gaps, while HF severely overestimates them; sX-LDA tends to give results much closer to experiment than the LDA, but still seems to underestimate band gaps to some extent. We concluded Chapter 5 with a discussion of the reasons for the successes and failures of the different functionals in predicting band structures.

Aside from the calculation of band structures, which is currently the most common application of non-local functionals, a very important application of DFT is in the calculation of geometries and molecular dynamics. In calculations in which the unit cell is allowed to vary, this requires calculation of the stress tensor. In Chapter 6, we derived the theory that allows us to calculate the contribution to the stress tensor from non-local functionals such as sX-LDA, HF, and EXX, which will allow variable-cell geometry optimisation and molecular dynamics calculations to be preformed when using these functionals. We also described how this has been

implemented as an extension to the work described in Chapter 4. As a test of this theory, we calculated the stress on a Si crystal as both hydrostatic and shear strains were applied. We found that our calculation of the stress tensor was consistent with the variation of the total energy with lattice parameter, which confirmed that the calculation of the stress tensor had been successful. This work will become increasingly important as the use of non-local functionals becomes more popular, and the calculation of stress is routinely performed.

We have now summarised the work and conclusions of this thesis. Our overall conclusion is that the use of non-local exchange-correlation functionals can result in improvements local functionals, but that there certainly remains room for further advancement in this area. For the remainder of this chapter, therefore, we will discuss some novel ideas concerning possible future advancements in the field of non-local exchange correlation functionals.

7.2 Simple Correlation Functionals

7.2.1 Combinations of Input

In Section 3.4.2 we discussed the meta-GGA, and Perdew's proposed hyper-GGA, both of which involve extra input to the exchange-correlation functional beyond the local density and gradient. We propose a direction of investigation involving various combinations of input, that do not necessarily involve the density or gradient. These are intended to be combined with EXX treatment of exchange will hence be referred to simply as "correlation functionals" rather than "exchange-correlation functionals".

In the standard formulation of EXX, correlation is treated by means of the LDA, in which the input is the local density, and the correlation energy per electron is derived from exact results for the homogeneous electron gas (HEG). But why use the local density as input? Why not use exchange energy per electron instead? After all, the exchange energy density is readily available in an EXX calculation, and contains

non-local information about the density. Further, the standard HEG data can easily be manipulated to produce an unambiguous parameterisation of such a functional. That is, for a given value of ε_X, there is a corresponding HEG density ρ, and for a given ρ there a corresponding HEG correlation energy per electron, ε_C. This simple functional can be written as follows:

$$E_C[\rho] = \int d\mathbf{r}\rho(\mathbf{r})\varepsilon_C^{HEG}\left(\rho^{HEG}\left(\varepsilon_X(\mathbf{r})\right)\right). \tag{7.1}$$

In fact, any readily available local quantity could, in principle, be used as input to such a functional. The only conditions are that there is a one-to-one relationship between the quantity and the density in a HEG, and that for every value of the quantity in the system under study, there exists a corresponding HEG density.

The advantage of having only a single local quantity as input is that the functional is unambiguously parameterised by HEG data. Of course, by having several inputs, more information is available, and a better functional should result, depending on the parameterisation. For Perdew's hyper-GGA, the density, gradient, non-interacting kinetic energy density, and exchange energy density are all used as input. A simpler version of this would exclude the density gradient, producing a functional of the form

$$E_C[\rho] = \int d\mathbf{r}\rho(\mathbf{r})\varepsilon_C(\rho, t_S, \varepsilon_X). \tag{7.2}$$

The problem with such functionals is parameterisation. Different parameterisations can always be chosen in order to satisfy different sets of physical constraints. In the next section we describe a possible method of parameterisation ideal for functionals with three local inputs.

7.2.2 Parameterisation from Sinusoidal Electron Gas

We can view the LDA as being defined such that exact results must be obtained for the simplest extended system imaginable, i.e. the homogeneous electron gas. For functionals with more than one quantity as input, we could extend this requirement so that the results have to be exact for what is arguably the second simplest extended system imaginable - a system with a sinusoidally varying density. A sinusoidal

density is defined by three independent parameters, namely the average density, ρ_0, the spatial frequency of the density oscillation, k, and the amplitude of the oscillation, A, i.e.

$$\rho(x) = \rho_0 + A\sin(kx). \tag{7.3}$$

It may therefore be possible to define a correlation functional with three local inputs that is exact for all sinusoidal densities.

However, if we are ever to use such a functional in actual calculations we would require accurate QMC data for sinusoidal densities. QMC work has been done [103] on systems with sinusoidally varying external potentials, but there is no simple way of knowing what the external potential is that corresponds to a sinusoidal density. The only way this can be determined is via a search, which may prove expensive. However, such calculations would only ever have to be performed once, so the computational cost may be justifiable in the long run.

As well as obtaining exact results for the interacting system, we would also need to know the Kohn-Sham orbitals corresponding to a given sinusoidal density. This could be achieved quite easily via an OEP style search, in which we minimise the deviation of the density from the sinusoidal target.

7.3 Variational Correlation Holes

7.3.1 Introduction

Earlier, we discussed the physical meaning of the correlation hole and the correlation kinetic and potential energies. This understanding of the physics of correlation can be used to justify a possible method of treating correlation in DFT calculations that involves using the correlation hole itself as a variational object.

7.3.2 Definition of the Functional

Supposing we have a given density, $\rho(\mathbf{r})$, and a corresponding set of Kohn-Sham orbitals $\{\phi_i(\mathbf{r})\}$, then we can directly evaluate every term contributing to the energy

in the Kohn-Sham framework except for the correlation energy. The reason we cannot evaluate the correlation energy exactly is that it is a many-electron property, and therefore not directly related to the non-interacting Kohn-Sham system defined by the orbitals.

However, if the correlation hole $h_C(\mathbf{r}'|\mathbf{r})$ is known, then the correlation potential energy, V_C, can be directly evaluated simply by calculating the Coulomb interaction between the hole and an electron at $\mathbf{r}$. Also, while there is no direct means of evaluating the correlation kinetic energy, T_C, we know that it must be related to the shape of the correlation hole.

If we could fix the shape of the correlation hole, as well as the orbitals and the density, then the only energy term free to vary would be T_C. The many-body wavefunction would therefore be the function that minimised T_C subject to the necessary constraints. For example, if we fixed the correlation hole to be zero everywhere, then the lowest energy wavefunction would equal that of the non-interacting system and T_C would equal zero. Any non-zero shape for correlation hole forces the wavefunction to distort away from this non-interacting case, resulting in a positive T_C. Loosely speaking, the larger the deviation of the correlation hole from zero, the more the wavefunction has to distort, and the larger the value of T_C.

The shape of the correlation hole that will naturally be adopted by a system will be the shape that minimises the total correlation energy, subject to constraints. The constraints on the correlation hole ensure that it corresponds to an exchange anti-symmetric many-electron wavefunction of density $\rho(\mathbf{r})$. One specific constraint is that the correlation hole around any point must integrate to zero. By becoming negative at short-range, and positive at long range, the correlation potential energy is lowered, while the correlation kinetic energy is increased. Minimisation of the correlation energy can therefore be viewed as a balance between reduced potential energy and increased kinetic energy.

T_C and V_C can be viewed as a functionals of the correlation hole shape. Also, the corresponding energies per electron, $t_C(\mathbf{r})$, and $v_C(\mathbf{r})$, can be viewed as functionals of the correlation hole surrounding the point $\mathbf{r}$. While $v_C(\mathbf{r})$ is readily evaluated

from the shape of the hole via

$$v_C(\mathbf{r}) = \frac{1}{2}\int d\mathbf{r}' \frac{h_C(\mathbf{r}'|\mathbf{r})}{|\mathbf{r}-\mathbf{r}'|}, \tag{7.4}$$

there is no such explicit expression for $t_C(\mathbf{r})$. Nevertheless, if a good approximate expression could be found for $t_C[h_C]$, then we would have a method of evaluating the correlation energy involving minimisation with respect to the correlation hole. That is, our correlation functional would be of the form:

$$E_C = \int d\mathbf{r}\,(t_C(\mathbf{r}) + v_C(\mathbf{r}))\,\rho(\mathbf{r}), \tag{7.5}$$

where

$$t_C(\mathbf{r}) = \min_{h_C(\mathbf{r}'|\mathbf{r})} t_C[h_C]. \tag{7.6}$$

A possible means of approximating $t_C[h_C]$ would be to assume that it can be obtained by integrating some local function of the hole, τ_C i.e.

$$t_C(\mathbf{r})[h_C] = \int d\mathbf{r}' \tau_C\left(h_C(\mathbf{r}'|\mathbf{r})\right). \tag{7.7}$$

It makes more sense however to allow this function, τ_C, to depend on other quantities as well, such as the distance, $|\mathbf{r}-\mathbf{r}'|$, and the density at $\mathbf{r}'$, or perhaps the exchange hole, $h_X(\mathbf{r}'|\mathbf{r})$, or derivatives of h_C, so we may have a functional of the form

$$t_C(\mathbf{r})[h_C] = \int d\mathbf{r}' \tau_C\left(h_C(\mathbf{r}'|\mathbf{r}), |\mathbf{r}-\mathbf{r}'|, \rho(\mathbf{r}'), \cdots\right). \tag{7.8}$$

In order to parameterise such a functional we could make use of QMC studies of the homogeneous electron gas. While data on the shape of the correlation hole is available in the literature [104], more detailed information about the wavefunction is harder to come across. A quantity of particular interest would be a two-particle correlation kinetic energy density, $t_C(\mathbf{r},\mathbf{r}')$, as this could be directly related to τ_C. Even more useful in fact would be QMC studies in which the correlation hole was fixed to be some shape other than its natural one. This would provide very detailed information on the relationship between $h_C(\mathbf{r}'|\mathbf{r})$ and T_C.

7.3.3 Spin Dependence

An important consideration when developing functionals such as this is the spin-dependence of the correlation hole. This is relevant even for systems that are not

spin-polarised. Because the exchange interaction only causes particles to avoid each other if they have like spins, the exchange hole is always spin-polarised, irrespective of the average polarisation of the system. This means that if we are thinking about the shape of the correlation hole, we must bear in mind that particles of like-spin are *already* likely to be well separated due to the exchange interaction. The correlation interaction should therefore affect unlike spins much more than it affects like spins, and the correlation holes for like and unlike spins may be very different. For this reason it may make more sense to formulate the variational hole method in terms of spin-dependent quantities.

7.3.4 Minimisation Procedures

We envisage that minimisation of the correlation energy with respect to the shape of the correlation hole could proceed via standard numerical methods such as conjugate gradients [38]. Of course this would require us to be able to differentiate whatever form we have chosen for $t_C[h_C]$, which may complicate matters. We also have to be careful to impose certain constraints on $h_C(\mathbf{r}'|\mathbf{r})$. For example, the sum rule implies that the integral of the correlation hole must equal zero. Also, exchange anti-symmetry means that the exchange-correlation hole must satisfy the following constraint,

$$\frac{h_{XC}(\mathbf{r}'|\mathbf{r})}{h_{XC}(\mathbf{r}|\mathbf{r}')} = \frac{\rho(\mathbf{r}')}{\rho(\mathbf{r})}, \tag{7.9}$$

which, if enforced, would mean that the minimisations of the holes surrounding each electron coordinate, $\mathbf{r}$, would have to proceed simultaneously. Other restrictions must also exist that may be more difficult to enforce efficiently, one such restriction being that the correlation hole actually corresponds to some many-electron wavefunction. It may be the case, however, that some restrictions could be relaxed without altering the results severely.

7.3.5 Potential Drawbacks

While potentially very promising as an idea, the variational hole method is admittedly a long way from practical realisation. Attempts to develop of the method

towards a computationally feasible implementation are likely to encounter a number of problems, and we ought to mention some of the potential problems that we already foresee.

Firstly, the computational cost is likely to be very high indeed. We would be manipulating objects that are at least $\sim N_p^2$ in size. Further, for each step in the self-consistent DFT calculation, we would have to do a separate optimisation of the correlation hole for every point on the grid. Having said that, it may be possible to combine the minimisations in some way to speed the process up.

It is possible that sampling of the correlation hole would require a finer grid that in a standard calculation, which would increase computational demand. On the other hand, for large systems, it may be possible to set a cut-off radius beyond which the hole can be assumed to be zero, which would ultimately lead to linear scaling. This would, however, depend on the nature of the particular system under study.

7.4 Final Remarks

This work has demonstrated some of the capabilities of DFT-based techniques in predicting properties of condensed matter from first principles. At present the major challenges for future development of DFT techniques are improving accuracy and improving efficiency, particularly in relation to scaling with system size. We have shown in this work some of the present methods that exist in relation to improving accuracy, mostly concerning the calculation of electronic band structures. We have also, in this final chapter, discussed some possible directions for greater improvements in accuracy in the future, through the further development of advanced exchange-correlation functionals. Much work is underway in at present in developing implementations of DFT that scale linearly with system size [105, 106]. This should, with the continuing advances in computer power, extend the scope of first principles calculations towards systems of much greater size and complexity than are currently possible. However, there are currently problems in applying these methods to metallic systems. Most of the present work however is based around local

functionals, which, as we have discussed in this work, have limited accuracy. In our view, therefore, the ultimate goal of future DFT development should be to combine linear scaling techniques with advanced exchange-correlation functionals.

Appendix A

Commonly Used Symbols and Abbreviations

In this appendix we list some symbols and abbreviations that are commonly used throughout this work, with a brief description of their meaning.

A.1 Variables

A.1.1 Integer Variables

N	number of electrons
N_I	number of ions or nuclei
Z	atomic number
N_k	number of **k**-points
N_b	number of occupied bands
N_p	number of plane waves in basis
n, m, i, j	electron or orbital index
ν	valence orbital index
c	conduction orbital index

A.1.2 Real Scalar Variables

t	time
Ω	unit cell volume
E_{TOT}	total energy
V_{I-I}	Ewald energy
E	total electronic energy
T	kinetic energy
V	potential energy
V_{ext}	potential energy of electrons due to external potential
V_{int}	potential energy from electron-electron repulsion
F	internal electronic energy
F_S	non-interacting electronic energy
T_S	non-interacting kinetic energy
$V_{int}^{(S)}$	non-interacting internal potential energy
V_H	Hartree energy
E_{XC}	exchange-correlation energy
V_X	exchange energy
E_C	correlation energy
T_C	correlation kinetic energy
V_C	correlation potential energy
λ	electron-electron coupling constant
E_{cut}	plane-wave cut-off energy
f	occupancy of a band
k_s	Thomas-Fermi screening constant

A.1.3 Vector Variables

$\mathbf{r}$	position vector (may include spin)
$\mathbf{R}_I$	position vector of ion or nucleus (may include spin)
$\mathbf{R}$	real lattice vector
$\mathbf{a}$	first unit cell vector
$\mathbf{b}$	second unit cell vector
$\mathbf{c}$	third unit cell vector
$\mathbf{k}$,$\mathbf{q}$	Bloch wave vector
$\mathbf{G}$	reciprocal lattice vector
$\mathbf{K}$	general reciprocal space vector
$\mathbf{a}^*$	first reciprocal unit cell vector
$\mathbf{b}^*$	second reciprocal unit cell vector
$\mathbf{c}^*$	third reciprocal unit cell vector

A.1.4 Tensor Variables

$\epsilon_{\alpha\beta}$	unsymmetrised strain tensor
$\sigma_{\alpha\beta}$	stress tensor

A.1.5 Other Variables

σ	spin coordinate

A.2 Fields

A.2.1 Real 3D Scalar Fields

ρ	electron density
v_{ext}	external potential
ε	energy per electron
μ_{KS}	Kohn-Sham potential
t	kinetic energy per electron
v	potential energy per electron
t_s	non-interacting kinetic energy per electron
v_H	Hartree energy per electron
μ_H	Hartree potential
ε_{XC}	exchange-correlation energy per electron
μ_{XC}	exchange-correlation potential
v_X	exchange energy per electron
μ_X	exchange potential
ε_C	correlation energy per electron
μ_C	correlation potential
v_C	correlation potential energy per electron
t_C	correlation kinetic energy per electron

A.2.2 Complex 3D Scalar Fields

ψ	Kohn-Sham orbital, including spin
ϕ	Kohn-Sham orbital, spatial part only
c	spatial-only Kohn-Sham orbital in reciprocal space

A.2.3 2-Particle Objects

$\rho(\mathbf{r}, \mathbf{r}')$	electron pair density
$\rho_S(\mathbf{r}, \mathbf{r}')$	pair density of non-interacting system
$\rho(\mathbf{r}\|\mathbf{r}')$	conditional electron density
$\rho_S(\mathbf{r}\|\mathbf{r}')$	conditional density of non-interacting system
$h_{XC}(\mathbf{r}\|\mathbf{r}')$	exchange-correlation hole
$h_X(\mathbf{r}\|\mathbf{r}')$	exchange hole
$h_C(\mathbf{r}\|\mathbf{r}')$	correlation hole
$\bar{h}_{XC}(\mathbf{r}\|\mathbf{r}')$	coupling constant averaged exchange-correlation hole
$\bar{h}_C(\mathbf{r}\|\mathbf{r}')$	coupling constant averaged correlation hole

A.2.4 Many-Particle Objects

Ψ_{MB}	many-body wavefunction of electrons and nuclei
Ψ	many-electron wavefunction
Ψ_S	wavefunction of non-interacting system

A.3 Abbreviations

DFT	density functional theory
EXX	exact exchange
FT	Fourier transform
FFT	fast Fourier transform
GGA	generalised gradient approximation
HF	Hartree-Fock
LDA	local density approximation
LSDA	local spin density approximation
MGGA	meta-generalised gradient approximation
MP	Monkhorst-Pack
PPD	parallelepiped
QMC	quantum Monte-Carlo
sX-LDA	screened exchange
WZ	wurtzite
ZB	zinc blende

Appendix B

Units and Physical Constants

Due to the small scale of the systems to which electronic structure calculations are applied, the use of S.I. units is somewhat inconvenient. For this reason, throughout most of this work, we generally use units that are more appropriate for these calculations. All equations, expressions, and formulas should be taken to be in *atomic units* unless otherwise stated. Masses and charges are usually quoted in atomic units (m_e and e respectively), lengths are be quoted in either atomic units (bohr) or Ångstroms (Å), and energy will be quoted in either atomic units (Ha) or electron-volts (eV).

To enable the reader easily to convert the units used in this work into S.I. units, if so desired, in this appendix we include a table of relevant physical constants, and a table of conversion factors in terms of those constants. Data on fundamental constants is from NIST [107].

B.1 Physical Constants

Name	Symbol	Value (S.I. units)
Planck Constant/2π	$\hbar$	1.054 571 68(18) $\times 10^{-34}$ Js
Speed of Light	c	2.99 792 458(exact) $\times 10^{8}$ ms^{-1}
Fine Structure Constant	α	7.297 352 568(24) $\times 10^{-3}$
Elementary Charge	e	1.602 176 53(14) $\times 10^{-19}$ C
Electron Mass	m_e	9.109 3826(16) $\times 10^{-31}$ kg

B.2 Atomic - S.I. Conversion Factors

Quantity	Atomic Unit		Conversion Atomic $\rightarrow$ S.I.	
Length	1 Bohr	=	$\frac{\hbar}{m_e c \alpha}$	= 5.291 772 108(18) $\times 10^{-11}$ m
Mass	1 m_e	=	m_e	= 9.109 3826(16) $\times 10^{-31}$ kg
Time	1 aut	=	$\frac{\hbar}{c^2 \alpha^2 m_e}$	= 2.418 884 326 505(16) $\times 10^{-17}$ s
Charge	1 e	=	e	= 1.602 176 53(14) $\times 10^{-19}$ C
Energy	1 Ha	=	$\alpha^2 m_e c^2$	= 4.359 744 17(75) $\times 10^{-18}$ J

B.3 Other Units

Unit	Value in S.I. Units	Value in Atomic Units
Å	10^{-10} m	1.889 726 1249(64) Bohr
eV	1.602 176 53(14) $\times 10^{-19}$ J	3.674 932 45(31) $\times 10^{-2}$ Ha
GPa	10^{9} Pa	3.398 9135(58) $\times 10^{-5}$ Ha Bohr^{-3}

Appendix C

Implicit Mathematical Elements

For convenience and tidiness, within many equations and formulas throughout this work, there may be extra elements that are not explicitly written down but should be taken to be present unless clearly stated otherwise. In this appendix, we list all of the situations in which these implicit mathematical elements are present.

C.1 Extra Variables

C.1.1 Spin Degrees of Freedom

Often we write the coordinates of an electron simply as $\mathbf{r}$, when in fact there should also be a spin coordinate, σ, included. In such cases, the spin coordinate should be considered to be included in $\mathbf{r}$ even though it is not explicitly written down, i.e.

$$\mathbf{r} \equiv (\mathbf{r}, \sigma). \tag{C.1}$$

Integrals over electronic coordinates should also be considered to include a summation over spin coordinates, i.e.

$$\int d\mathbf{r} f(\mathbf{r}) \equiv \sum_{\sigma} \int d\mathbf{r} f(\mathbf{r}, \sigma). \tag{C.2}$$

Of course, $\mathbf{r}$ often means just the spatial coordinates, so the above formulas should not be taken to apply in general. It should always be clear from the context whether or not spin is being included or not.

The information in this section also applies to the spin degrees of freedom of nuclei, which should be considered to be included in the coordinates, $\mathbf{R}_I$.

C.2 Implicit Factors in Summations

C.2.1 Summing over Kohn-Sham Orbitals

When summing over Kohn-Sham orbitals, using the spatial-part only functions, $\phi_{i\mathbf{k}}(\mathbf{r})$, or, in reciprocal space, $c_{i\mathbf{k}}(\mathbf{G})$, there should be a factor of 2 before each term to account for the 2 spin states. The exception to this is when N is odd, in which case there is no implicit factor of 2 in the final term.

C.2.2 Summing over k-points

When summing over $\mathbf{k}$-points, there should always be a factor of $1/N_k$ before the summation sign. This is usually omitted from formulas, but should still be taken to be present, so we have

$$\sum_{\mathbf{k}} \equiv \frac{1}{N_k} \sum_{\mathbf{k}}. \tag{C.3}$$

Appendix D

Derivation of Non-Interacting Quantities

The derivations of equations for quantities related to the Kohn-Sham Slater determinant are rather cumbersome and are therefore included here as an appendix rather than in the main text.

D.1 The Particle Density

The particle density, $\rho(\mathbf{r})$, is the probability density for finding *any* particle at position $\mathbf{r}$. Since all the particles are identical, this must be N times the probability density for finding a *particular* particle at $\mathbf{r}$, i.e.

$$\rho(\mathbf{r}) = N \sum_{\sigma} \int d\mathbf{r}_2 \cdots \int d\mathbf{r}_N \Psi_S^*(\mathbf{r}, \mathbf{r}_2, \cdots \mathbf{r}_N) \Psi_S(\mathbf{r}, \mathbf{r}_2, \cdots \mathbf{r}_N). \tag{D.1}$$

Now, substituting the expression for Ψ_S of Eq. (1.33), we have

$$\begin{aligned} \rho(\mathbf{r}_1) &= N \frac{1}{N!} \sum_{\sigma_1} \int d\mathbf{r}_2 \cdots \int d\mathbf{r}_N \left(\sum_{n=1}^{N!} (-1)^{n-1} \prod_{i=1}^{N} \psi_i^*(\mathbf{r}_{P_n(i)}) \right) \\ &\quad \times \left(\sum_{m=1}^{N!} (-1)^{m-1} \prod_{j=1}^{N} \psi_j(\mathbf{r}_{P_m(j)}) \right). \end{aligned} \tag{D.2}$$

Any terms in which the first permutation, $\{P_n(i)\}$, differs from the second permutation, $\{P_m(i)\}$, must equal zero, since they will involve at least one integration of

a product of orthogonal orbitals. We thus have

$$\rho(\mathbf{r}_1) = \frac{1}{(N-1)!}\sum_{\sigma_1}\int d\mathbf{r}_2 \cdots \int d\mathbf{r}_N \sum_{n=1}^{N!}\prod_{i=1}^{N}\psi_i^*(\mathbf{r}_{P_n(i)})\psi_i(\mathbf{r}_{P_n(i)}). \tag{D.3}$$

Integrals over all coordinates, $\mathbf{r}_j$, are effectively just integrals of products of orbitals, $\psi_i^*(\mathbf{r}_j)\psi_i(\mathbf{r}_j)$, which simply return factors of 1. We therefore have

$$\rho(\mathbf{r}_1) = \frac{1}{(N-1)!}\sum_{\sigma_1}\sum_{n=1}^{N!}\psi^*_{P_n^{-1}(1)}(\mathbf{r}_1)\psi_{P_n^{-1}(1)}(\mathbf{r}_1), \tag{D.4}$$

where $P_n^{-1}(i)$ is the inverse permutation of $P_n(i)$. For each of the N possible subscripts $P_n^{-1}(1)$ there are $(N-1)!$ permutations, hence we can write this as

$$\begin{aligned}\rho(\mathbf{r}) &= \frac{1}{(N-1)!}(N-1)!\sum_{\sigma}\sum_{i}^{N}\psi_i^*(\mathbf{r})\psi_i(\mathbf{r}) \\ &= \sum_{i}^{N}\psi_i^*(\mathbf{r})\psi_i(\mathbf{r}),\end{aligned} \tag{D.5}$$

which is the equation for the density in terms of the Kohn-Sham orbitals. This looks the essentially the same in terms of the spatial-only orbitals, $\phi_i(\mathbf{r})$, except that the spin summation is no longer necessary:

$$\rho(\mathbf{r}) = \sum_{i}^{N}\phi_i^*(\mathbf{r})\phi_i(\mathbf{r}). \tag{D.6}$$

D.2 The Pair-Density and Related Quantities

In analogy with Equation (1.18), the pair-density of the non-interacting system is defined as

$$\rho_S(\mathbf{r},\mathbf{r}') = N(N-1)\sum_{\sigma\sigma'}\int d\mathbf{r}_3\cdots\int d\mathbf{r}_N \Psi_S^*(\mathbf{r},\mathbf{r}',\mathbf{r}_3,\cdots\mathbf{r}_N)\Psi_S(\mathbf{r},\mathbf{r}',\mathbf{r}_3,\cdots\mathbf{r}_N). \tag{D.7}$$

Now substituting the expression for Ψ_S of Equation (1.33), we have

$$\begin{aligned}\rho_S(\mathbf{r}_1,\mathbf{r}_2) &= \frac{1}{(N-2)!}\sum_{\sigma_1\sigma_2}\int d\mathbf{r}_3\cdots\int d\mathbf{r}_N\left(\sum_{n=1}^{N!}(-1)^{n-1}\prod_{i=1}^{N}\psi_i^*(\mathbf{r}_{P_n(i)})\right) \\ &\times\left(\sum_{m=1}^{N!}(-1)^{m-1}\prod_{j=1}^{N}\psi_j(\mathbf{r}_{P_m(j)})\right).\end{aligned} \tag{D.8}$$

Most terms in which the first permutation, $\{P_n(i)\}$, differs from the second permutation, $\{P_m(j)\}$, will equal zero, since they will involve integration of a product of orthogonal orbitals. The only terms for which this is not the case are those in which the second permutation can be generated from the first by simply swapping the positions of indices 1 and 2, since these are not integrated over. This means we can group the terms into those in which the permutations are the same, with a positive prefactor, and those in which the permutations differ by exchange of these indices, with a negative prefactor, as follows:

$$\rho_S(\mathbf{r}_1,\mathbf{r}_2) = \frac{1}{(N-2)!}\sum_{\sigma_1\sigma_2}\int d\mathbf{r}_3\cdots\int d\mathbf{r}_N\left(\sum_{n=1}^{N!}\prod_{i=1}^{N}\psi_i^*(\mathbf{r}_{P_n(i)})\psi_i(\mathbf{r}_{P_n(i)})\right)$$

$$-\frac{1}{(N-2)!}\sum_{\sigma_1\sigma_2}\int d\mathbf{r}_3\cdots\int d\mathbf{r}_N\left(\sum_{n=1}^{N!}\prod_{i=1}^{N}\psi_i^*(\mathbf{r}_{P_n(i)})\psi_i(\mathbf{r}_{R_{12}(P_n(i))})\right), \qquad \text{(D.9)}$$

where R_{12} is a permutation operator that swaps the positions of the numbers 1 and 2.

Each integral is now effectively some prefactor times an integral over products of orbitals, which simply return factors of 1. We can thus integrate out all coordinates except for $\mathbf{r}_1$ and $\mathbf{r}_2$:

$$\rho_S(\mathbf{r}_1,\mathbf{r}_2) = \frac{1}{(N-2)!}\sum_{\sigma_1\sigma_2}\left(\sum_{n=1}^{N!}\psi^*_{P_n^{-1}(1)}(\mathbf{r}_1)\psi_{P_n^{-1}(1)}(\mathbf{r}_1)\psi^*_{P_n^{-1}(2)}(\mathbf{r}_2)\psi_{P_n^{-1}(2)}(\mathbf{r}_2)\right)$$

$$-\frac{1}{(N-2)!}\sum_{\sigma_1\sigma_2}\left(\sum_{n=1}^{N!}\psi^*_{P_n^{-1}(1)}(\mathbf{r}_1)\psi_{P_n^{-1}(1)}(\mathbf{r}_2)\psi^*_{P_n^{-1}(2)}(\mathbf{r}_2)\psi_{P_n^{-1}(2)}(\mathbf{r}_1)\right), \qquad \text{(D.10)}$$

where $P_n^{-1}(i)$ is the inverse permutation of $P_n(i)$.

For each of the $N(N-1)$ possible pairs of subscripts, $P_n^{-1}(1)$, $P_n^{-1}(2)$, there are $(N-2)!$ permutations, hence we can write this as

$$\begin{aligned}\rho_S(\mathbf{r},\mathbf{r}') &= \sum_{\sigma\sigma'}\sum_{ij}\psi_i^*(\mathbf{r})\psi_i(\mathbf{r})\psi_j^*(\mathbf{r}')\psi_j(\mathbf{r}') \\ &\quad -\sum_{\sigma\sigma'}\sum_{ij}\psi_i^*(\mathbf{r})\psi_i(\mathbf{r}')\psi_j^*(\mathbf{r}')\psi_j(\mathbf{r}) \\ &= \rho(\mathbf{r})\rho(\mathbf{r}') - \sum_{\sigma\sigma'}\left|\sum_i\psi_i^*(\mathbf{r})\psi_i(\mathbf{r}')\right|^2, \qquad \text{(D.11)}\end{aligned}$$

which is the equation for the pair density of the non-interacting system in terms of the Kohn-Sham orbitals. This looks the essentially the same in terms of the spatial-only orbitals, $\phi_i(\mathbf{r})$, except that the spin summation is no longer necessary:

$$\rho_S(\mathbf{r},\mathbf{r}') = \rho(\mathbf{r})\rho(\mathbf{r}') - \left|\sum_i \phi_i^*(\mathbf{r})\phi_i(\mathbf{r}')\right|^2 . \tag{D.12}$$

This is related to the conditional density of the non-interacting system, and also to the exchange hole, which are given respectively by:

$$\begin{aligned} \rho_S(\mathbf{r}'|\mathbf{r}) &= \frac{\rho_S(\mathbf{r},\mathbf{r}')}{\rho(\mathbf{r})} \\ &= \rho(\mathbf{r}') - \frac{|\sum_i \phi_i^*(\mathbf{r})\phi_i(\mathbf{r}')|^2}{\rho(\mathbf{r})}, \end{aligned} \tag{D.13}$$

and

$$\begin{aligned} h_X(\mathbf{r}'|\mathbf{r}) &= \rho_S(\mathbf{r}'|\mathbf{r}) - \rho(\mathbf{r}') \\ &= -\frac{|\sum_i \phi_i^*(\mathbf{r})\phi_i(\mathbf{r}')|^2}{\rho(\mathbf{r})}. \end{aligned} \tag{D.14}$$

D.3 The Non-Interacting Kinetic Energy

We begin with the definition of T_S as the expectation value of the kinetic energy operator, $\hat{T}$, in the Kohn-Sham Slater determinant:

$$\begin{aligned} T_S &= \langle \Psi_S|\hat{T}|\Psi_S\rangle \\ &= -\frac{1}{2}\int d\mathbf{r}_1 \cdots \int d\mathbf{r}_N \Psi_S^* \sum_i \nabla_{\mathbf{r}_i}^2 \Psi_S. \end{aligned} \tag{D.15}$$

Now substituting the expression for Ψ_S of Equation (1.33), we have

$$\begin{aligned} T_S = -\frac{1}{2N!}\int d\mathbf{r}_1 \cdots \int d\mathbf{r}_N \left(\sum_{n=1}^{N!}(-1)^{n-1}\prod_{i=1}^{N}\psi_i^*(\mathbf{r}_{P_n(i)})\right) \\ \times \sum_j \nabla_{\mathbf{r}_j}^2 \left(\sum_{m=1}^{N!}(-1)^{m-1}\prod_{k=1}^{N}\psi_k(\mathbf{r}_{P_m(k)})\right). \end{aligned} \tag{D.16}$$

Any terms in which the first permutation, $\{P_n(i)\}$, differs from the second permutation, $\{P_m(i)\}$, must equal zero, since they will involve at least one integration of a product of orthogonal orbitals. We thus have

$$T_S = -\frac{1}{2N!}\sum_j \int d\mathbf{r}_1 \cdots \int d\mathbf{r}_N \sum_{n=1}^{N!}\prod_{i=1}^{N} \psi_i^*(\mathbf{r}_{P_n(i)})\nabla^2_{\mathbf{r}_j}\psi_i(\mathbf{r}_{P_n(i)}). \tag{D.17}$$

The operator, $\nabla^2_{\mathbf{r}_j}$, only operates on orbitals for which $P_n(i) = j$. Therefore integrals over all coordinates, $\mathbf{r}_k$, where $k \neq j$ are effectively just integrals of products of orbitals, $\psi_i^*(\mathbf{r}_k)\psi_i(\mathbf{r}_k)$, which simply return factors of 1. So for each term we can integrate out all the coordinates except $\mathbf{r}_j$, which is associated with the $\nabla^2_{\mathbf{r}_j}$ operator:

$$T_S = -\frac{1}{2N!}\sum_j \int d\mathbf{r}_j \sum_{n=1}^{N!} \psi^*_{P_n^{-1}(j)}(\mathbf{r}_j)\nabla^2_{\mathbf{r}_j}\psi_{P_n^{-1}(j)}(\mathbf{r}_j), \tag{D.18}$$

where $P_n^{-1}(i)$ is the inverse permutation of $P_n(i)$. Now, from here we can see that the subscript on the $\mathbf{r}_j$ is no longer meaningful and that the summation can be reordered to read

$$\begin{aligned} T_S &= -\frac{1}{2N!}\sum_i \int d\mathbf{r} \sum_{n=1}^{N!} \psi_i^*(\mathbf{r})\nabla^2\psi_i(\mathbf{r}) \\ &= -\frac{1}{2}\sum_i \int d\mathbf{r}\psi_i^*(\mathbf{r})\nabla^2\psi_i(\mathbf{r}), \end{aligned} \tag{D.19}$$

which is the standard equation for the non-interacting kinetic energy. This equation looks the same in terms of the spatial-only orbitals, $\phi_i(\mathbf{r})$:

$$\begin{aligned} T_S &= -\frac{1}{2N!}\sum_i \int d\mathbf{r} \sum_{n=1}^{N!} \phi_i^*(\mathbf{r})\nabla^2\phi_i(\mathbf{r}) \\ &= -\frac{1}{2}\sum_i \int d\mathbf{r}\phi_i^*(\mathbf{r})\nabla^2\phi_i(\mathbf{r}), \end{aligned} \tag{D.20}$$

where there is an implicit factor of 2 where appropriate (see Appendix C).

D.4 The Non-Interacting Internal Potential Energy

We begin with the definition of $V_{int}^{(S)}$ as the expectation value of the internal potential energy operator, $\hat{V}_{int}$, in the Kohn-Sham Slater determinant:

$$\begin{aligned} V_{int}^{(S)} &= \langle \Psi_S | V_{int}^{(S)} | \Psi_S \rangle \\ &= \frac{1}{2} \int d\mathbf{r}_1 \cdots \int d\mathbf{r}_N \Psi_S^* \sum_{i \neq j} \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|} \Psi_S. \end{aligned} \tag{D.21}$$

Now substituting the expression for Ψ_S of Equation (1.33), we have

$$\begin{aligned} V_{int}^{(S)} &= \frac{1}{2N!} \int d\mathbf{r}_1 \cdots \int d\mathbf{r}_N \left(\sum_{n=1}^{N!} (-1)^{n-1} \prod_{i=1}^{N} \psi_i^*(\mathbf{r}_{P_n(i)}) \right) \\ &\quad \times \sum_{j \neq k} \frac{1}{|\mathbf{r}_j - \mathbf{r}_k|} \left(\sum_{m=1}^{N!} (-1)^{m-1} \prod_{l=1}^{N} \psi_l(\mathbf{r}_{P_m(l)}) \right). \end{aligned} \tag{D.22}$$

Most terms in which the first permutation, $\{P_n(i)\}$, differs from the second permutation, $\{P_m(l)\}$, will equal zero, since they will involve integration of a product of orthogonal orbitals. The only terms for which this is not the case are those in which the second permutation can be generated from the first by simply swapping the positions of indices j and k. This means we can group the terms into those in which the permutations are the same, with a positive prefactor, and those in which the permutations differ by exchange of these indices, with a negative prefactor, as follows:

$$\begin{aligned} V_{int}^{(S)} &= \frac{1}{2N!} \sum_{j \neq k} \int d\mathbf{r}_1 \cdots \int d\mathbf{r}_N \sum_{n=1}^{N!} \prod_{i=1}^{N} \frac{\psi_i^*(\mathbf{r}_{P_n(i)}) \psi_i(\mathbf{r}_{P_n(i)})}{|\mathbf{r}_j - \mathbf{r}_k|} \\ &\quad - \frac{1}{2N!} \sum_{j \neq k} \int d\mathbf{r}_1 \cdots \int d\mathbf{r}_N \sum_{n=1}^{N!} \prod_{i=1}^{N} \frac{\psi_i^*(\mathbf{r}_{P_n(i)}) \psi_i(\mathbf{r}_{R_{jk}(P_n(i))})}{|\mathbf{r}_j - \mathbf{r}_k|}, \end{aligned} \tag{D.23}$$

where $R_{jk}(i)$ is a permutation operator that swaps the positions of the numbers j and k.

The Coulomb term, $1/|\mathbf{r}_j - \mathbf{r}_k|$, can be considered to be a multiplicative operator that operates only on orbitals for which the subscript on $\mathbf{r}$ is equal to either j or

k. Integrals over all coordinates, $\mathbf{r}_l$, where $l \neq j,k$ are effectively just integrals of products of orbitals, $\psi_i^*(\mathbf{r}_l)\psi_i(\mathbf{r}_l)$, which simply return factors of 1. So for each term we can integrate out all the coordinates except $\mathbf{r}_j$ and $\mathbf{r}_k$, which are associated with the $1/|\mathbf{r}_j - \mathbf{r}_k|$ factor:

$$V_{int}^{(S)} = \frac{1}{2N!}\sum_{j\neq k}\int d\mathbf{r}_j \int d\mathbf{r}_k \sum_n^{N!} \frac{\psi^*_{P_n^{-1}(j)}(\mathbf{r}_j)\psi_{P_n^{-1}(j)}(\mathbf{r}_j)\psi^*_{P_n^{-1}(k)}(\mathbf{r}_k)\psi_{P_n^{-1}(k)}(\mathbf{r}_k)}{|\mathbf{r}_j - \mathbf{r}_k|}$$

$$-\frac{1}{2N!}\sum_{j\neq k}\int d\mathbf{r}_j \int d\mathbf{r}_k \sum_n^{N!} \frac{\psi^*_{P_n^{-1}(j)}(\mathbf{r}_j)\psi_{P_n^{-1}(j)}(\mathbf{r}_k)\psi^*_{P_n^{-1}(k)}(\mathbf{r}_k)\psi_{P_n^{-1}(k)}(\mathbf{r}_j)}{|\mathbf{r}_j - \mathbf{r}_k|}, \quad \text{(D.24)}$$

where $P_n^{-1}(i)$ is the inverse permutation of $P_n(i)$. Now, from here we can see that the subscripts on the $\mathbf{r}_j$ and $\mathbf{r}_k$ are no longer meaningful, other than in distinguishing one from the other, and that the summation can be reordered to read

$$\begin{aligned} V_{int}^{(S)} &= \frac{1}{2N!}\sum_{ij}\int d\mathbf{r}\int d\mathbf{r}' \sum_n^{N!} \frac{\psi_i^*(\mathbf{r})\psi_i(\mathbf{r})\psi_j^*(\mathbf{r}')\psi_j(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} \\ &\quad -\frac{1}{2N!}\sum_{ij}\int d\mathbf{r}\int d\mathbf{r}' \sum_n^{N!} \frac{\psi_i^*(\mathbf{r})\psi_i(\mathbf{r}')\psi_j^*(\mathbf{r}')\psi_j(\mathbf{r})}{|\mathbf{r}-\mathbf{r}'|} \\ &= \frac{1}{2}\sum_{ij}\int d\mathbf{r}\int d\mathbf{r}' \frac{\psi_i^*(\mathbf{r})\psi_i(\mathbf{r})\psi_j^*(\mathbf{r}')\psi_j(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} \\ &\quad -\frac{1}{2}\sum_{ij}\int d\mathbf{r}\int d\mathbf{r}' \frac{\psi_i^*(\mathbf{r})\psi_i(\mathbf{r}')\psi_j^*(\mathbf{r}')\psi_j(\mathbf{r})}{|\mathbf{r}-\mathbf{r}'|}. \end{aligned} \quad \text{(D.25)}$$

Noting that each integral contains an implicit sum over spin states, we see that this equation looks the same in terms of the spatial-only orbitals, $\phi_i(\mathbf{r})$:

$$\begin{aligned} V_{int}^{(S)} &= \frac{1}{2}\sum_{ij}\int d\mathbf{r}\int d\mathbf{r}' \frac{\phi_i^*(\mathbf{r})\phi_i(\mathbf{r})\phi_j^*(\mathbf{r}')\phi_j(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} \\ &\quad -\frac{1}{2}\sum_{ij}\int d\mathbf{r}\int d\mathbf{r}' \frac{\phi_i^*(\mathbf{r})\phi_i(\mathbf{r}')\phi_j^*(\mathbf{r}')\phi_j(\mathbf{r})}{|\mathbf{r}-\mathbf{r}'|}, \end{aligned} \quad \text{(D.26)}$$

where there is an implicit factor of 2 where appropriate (see Appendix C). Now, with reference to Equation (D.5) we see that the part with positive prefactor can be re-written in terms of the density so that we have

$$V_{int}^{(S)} = \frac{1}{2}\int d\mathbf{r}\int d\mathbf{r}' \frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} - \frac{1}{2}\sum_{ij}\int d\mathbf{r}\int d\mathbf{r}' \frac{\phi_i^*(\mathbf{r})\phi_i(\mathbf{r}')\phi_j^*(\mathbf{r}')\phi_j(\mathbf{r})}{|\mathbf{r}-\mathbf{r}'|}. \quad \text{(D.27)}$$

This internal potential energy is usually separated into the Hartree energy, V_H, and exchange energy, V_X, so that

$$V_{int}^{(S)} = V_H + V_X, \tag{D.28}$$

where

$$V_H = \frac{1}{2}\int d\mathbf{r}\int d\mathbf{r}'\frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}, \tag{D.29}$$

and

$$V_X = -\frac{1}{2}\sum_{ij}\int d\mathbf{r}\int d\mathbf{r}'\frac{\phi_i^*(\mathbf{r})\phi_i(\mathbf{r}')\phi_j^*(\mathbf{r}')\phi_j(\mathbf{r})}{|\mathbf{r}-\mathbf{r}'|}. \tag{D.30}$$

Bibliography

[1] R. O. Jones and O. Gunnarson, Rev. Mod. Phys **61**, 689 (1989).

[2] M. C. Payne, M. P. Teter, D. C. Allan, T. A. Arias, and J. D. Joannopoulos, Rev. Mod. Phys. **64**, 1045 (1992).

[3] M. D. Segall, P. J. D. Lindan, M. J. Probert, C. J. Pickard, P. J. Hasnip, S. J. Clark, and M. C. Payne, J. Phys. Condens. Matter **14**, 2717 (2002).

[4] R. M. Martin, *Electronic Structure: Basic Theory and Methods* (Cambridge University Press, 2004), iSBN: 0521782856.

[5] M. S. Hybertsen and S. G. Louie, Phys. Rev. B **34**, 2920 (1986).

[6] G. Theurich and N. A. Hill, Phys. Rev. B **64**, 73106 (2001).

[7] M. Born and J. R. Oppenheimer, Ann. Physik **389**, 457 (1927).

[8] R. M. Martin, *Electronic Structure: Basic Theory and Methods* (Cambridge University Press, 2004), iSBN: 0521782856.

[9] R. C. Grimm, J. Comp. Comm. **7**, 134 (1971).

[10] J. B. Anderson, J. Chem. Phys. **63**, 1499 (1975).

[11] W. M. C. Foulkes, L. Mitas, R. J. Needs, and G. Rajagopal, Rev. Mod. Phys. **73**, 33 (2001).

[12] R. J. Bartlett and J. F. Stanton, Rev. Comp. Chem. **5**, 65 (1995).

[13] P. Hohenberg and W. Kohn, Phys. Rev. **136**, B864 (1964).

[14] T. L. Gilbert, Phys. Rev. B **12**, 2111 (1975).

[15] W. Kohn and L. J. Sham, Phys. Rev. **140**, A1133 (1965).

[16] J. C. Slater, Phys. Rev. **81**, 385 (1951).

[17] D. R. Hartree, Proc. R. Soc. London **A113**, p. 621 (1928).

[18] P. Stevens, J. F. Devlin, C. F. Chabolowski, and M. J. Frisch, J. Phys. Chem **98**, 11623 (1994).

[19] P. P. Rushton, S. J. Clark, and D. J. Tozer, Phys. Rev. B **63**, 115206 (2001).

[20] D. M. Ceperley and B. J. Alder, Phys. Rev. Lett. **45**, 566 (1980).

[21] J. P. Perdew and Y. Wang, Phys. Rev. B **33**, 8800 (1986).

[22] J. P. Perdew and Y. Wang, Phys. Rev. B **45**, 13244 (1992).

[23] J. P. Perdew, Phys. Rev. Lett. **55**, 1665 (1985).

[24] J. P. Perdew, K. Burke, and Y. Wang, Phys. Rev. B **54**, 16533 (1996).

[25] J. P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996).

[26] P. P. Rushton, Ph.D. thesis, University of Durham (2002).

[27] J. P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. **78**, 1396 (1997).

[28] R. Orlando, R. Dovesi, C. Roetti, and V. R. Saunders, J. Phys. Condens. Matter **2**, 7769 (1990).

[29] V. R. Saunders, R. Dovesi, C. Roetti, M. Causà, N. M. Harrison, R. Orlando, and C. M. Zicovich-Wilson, *CRYSTAL'98 User's Manual (University of Torino, Torino)* (2003), see http://www.theochem.unito.it/.

[30] J. M. Soler, E. Artacho, J. D. Gale, A. Garcia, J. Junquera, P. Ordejon, and D. Sanchez-Portal, J. Phys. Cond. Matt. **14**, 2745 (2002).

[31] N. W. Ashcroft and N. D. Mermin, *Solid State Physics* (Holt Saunders, 1976).

[32] H. J. Monkhorst and J. D. Pack, Phys. Rev. B **13**, 5188 (1976).

[33] J. C. Phillips, Phys. Rev. **112**, 685 (1958).

[34] J. C. Phillips and L. Kleinman, Phys. Rev. **116**, 287 (1959).

[35] M. L. Cohen and V. Heine, Solid State Physics **24**, 37 (1970).

[36] D. Vanderbilt, Phys. Rev. B **41**, 7892 (1990).

[37] L. Kleinman and D. M. Bylander, Phys. Rev. Lett. **48**, 1425 (1982).

[38] W. H. Press and S. A. Teukolsky, *Numerical Recipes* (Cambridge, 1992).

[39] S. J. Clark, M. D. Segall, C. J. Pickard, P. J. Hasnip, M. J. Probert, K. Refson, and M. C. Payne, Zeit. Kryst. **220**, 567 (2005).

[40] H. Hellmann, *Einfuhrung in die Quantenchemie* (Deuticke, Leipzig, 1937).

[41] R. P. Feynman, Phys. Rev. **56**, 340 (1939).

[42] S. N. Mohammad and H. Morkoc, Prog. Quant. Electr. **20**, 361 (1996).

[43] J. Orton and C. T. Foxon, Rep. Prog. Phys. **61**, 1 (1998).

[44] S. Jain, M. Willander, J. Narayan, and R. Van Overstaeten, J. Appl. Phys. **87**, 965 (2000).

[45] B. Monemar and G. Pozina, Prog. Quant. Electr. **24**, 239 (2000).

[46] J. Neugebauer, T. Zywietz, and M. Scheffler, Phys. Rev. Lett. **80**, 3097 (1998).

[47] J. Fritsch, O. F. Sankey, K. E. Schmidt, and J. B. Page, Phys. Rev. B **57**, 15360 (1998).

[48] T. Zywietz, J. Neugebauer, and M. Scheffler, Appl. Phys. Lett **73**, 487 (1998).

[49] T. Zywietz, J. Neugebauer, M. Scheffler, J. Northrup, and C. G. Van de Walle, MRS Internet J. Nitride Semicond. Res. **3**, 26 (1998).

[50] J. E. Northrup, J. Neugebauer, R. M. Feenstra, and A. R. Smith, Phys. Rev. B **61**, 9932 (2000).

[51] F. Wang, P. Kruger, and J. Pollmann, Phys. Rev. B **64**, 035305 (2001).

[52] C. G. Van de Walle and J. Neugebauer, Phys. Rev. Lett. **6**, 066103 (2002).

[53] R. M. Feenstra, J. E. Northrup, and J. Neugebauer, MRS Internet J. Nitride Semicond. Res. **7**, 3 (2002).

[54] O. Madelung (Ed.), *Semiconductors - Basic Data* (Springer, 1996).

[55] M. H. Lee, Ph.D. thesis, University of Cambridge (1994).

[56] M. H. Lee, *Kinetic energy tuning for optimising pseudopotentials and projector reduction*, ψ_k Newsletter 67, page 145 (http://psi-k.dl.ac.uk/newsletters/News_67/newsletter_67.pdf).

[57] J. Moreno and J. M. Soler, Phys. Rev. B **45**, 13891 (1992).

[58] M. I. J. Probert and M. C. Payne, Phys. Rev. B **67**, 075204 (2003).

[59] W. A. Harisson, *Electronic Structure and the Properties of Solids* (Dover, New York, 1989).

[60] P. L. Taylor and O. Heinonen, *A Quantum Approach to Condensed Matter Physics* (Cambridge University Press, 2002).

[61] D. R. Lide, *CRC Handbook of Chemistry and Physics, 73rd ed.* (CRC Press, Boca Raton, Florida, 1992).

[62] M. R. Ranade, F. Tessier, A. Navrotsky, V. J. Leppert, S. H. Risbud, F. J. DiSalvo, and C. M. Balkas, J. Phys. Chem. B **104**, 4060 (2000).

[63] M. Fuchs, J. L. F. Da Silva, C. Stampfl, J. Neugebauer, and M. Scheffler, Phys. Rev. B **65**, 245212 (2002).

[64] I. Vurgaftman, J. R. Meyer, and L. R. Ram-Mohan, J. App. Phys. **89**, 5815 (2001).

[65] C. Kittel, *Introduction to Solid State Physics* (Wiley, New York, 1996).

[66] A. Munkholm *et. al.*, Phys. Rev. Lett. **83**, 741 (1999).

[67] P. L. Taylor and O. Heinonen, *A Quantum Approach to Condensed Matter Physics* (Cambridge University Press, 2002).

[68] F. Gygi and A. Baldereschi, Phys. Rev. B **34**, 4405 (1986).

[69] A. Seidl, A. Görling, P. Vogl, J. A. Majewski, and M. Levy, Phys. Rev. B **53**, 3764 (1996).

[70] J. Lento, Master's thesis, Helsinki University of Technology (1997).

[71] S. Chawla and A. G. Voth, J. Chem. Phys. **108**, 4697 (1998).

[72] V. S. Fock, Z. Phys **61**, 126 (1930).

[73] J. E. Robinson, F. Bassani, R. S. Knox, and J. R. Schrieffer, Phys. Rev. Lett. **9**, 215 (1962).

[74] L. H. Thomas, Proc. Cambridge Philos. Soc **23**, 542 (1927).

[75] E. C. Fermi, Z. Phys. **48**, 73 (1928).

[76] D. M. Bylander and L. Kleinman, Phys. Rev. B **41**, 7868 (1990).

[77] A. Görling, Phys. Rev. B **53**, 7024 (1996).

[78] M. Städele, J. A. Majewski, P. Vogl, and A. Görling, Phys. Rev. Lett. **79**, 2089 (1997).

[79] M. Städele, M. Moukara, J. A. Majewski, P. Vogl, and A. Görling, Phys. Rev. B **59**, 10031 (1999).

[80] J. D. Talman and W. F. Shadwick, Phys. Rev. A **14**, 36 (1976).

[81] J. B. Krieger, Y. Li, and G. J. Iafrate, Phys. Rev. A **45**, 101 (1992).

[82] O. Gunnarson, M. Jonson, and B. I. Lundquist, Phys. Rev. B. **59A**, 177 (1976).

[83] O. Gunnarson, M. Jonson, and B. I. Lundquist, Solid State Commun. **54**, 765 (1977).

[84] J. A. Alonso and L. A. Girifalco, Phys. Rev. B **17**, 3735 (1978).

[85] P. P. Rushton, D. J. Tozer, and S. J. Clark, Phys. Rev. B **65**, 235203 (2002).

[86] S. K. Ghosh and R. G. Parr, Phys. Rev. A **34**, 785 (1986).

[87] A. D. Becke and M. R. Roussel, Phys. Rev. A **39**, 3761 (1989).

[88] J. Tau, J. P. Perdew, V. N. Staroverov, and E. Scuseria, Phys. Rev. Lett. **91**, 146401 (2003).

[89] R. M. Martin, *Electronic Structure: Basic Theory and Methods* (Cambridge University Press, 2004), iSBN: 0521782856.

[90] *Accelrys Software Inc.*

[91] J. Wu, W. Walukiewicz, K. M. Yu, J. W. Ager III, E. E. Haller, H. Lu, W. J. Schaff, Y. Saito, and Y. Nanishi, Applied Physics Letters **80**, 3967 (2002).

[92] R. Asahi, W. Mannstadt, and A. J. Freeman, Phys. Rev. B **59**, 7486 (1999).

[93] C. B. Geller, W. Wolf, S. Picozzi, A. Continenza, R. Asahi, W. Mannstadt, J. Freeman, and E. Wimmer, App. Phys. Lett. **79**, 368 (2001).

[94] L. Sham and M. Schlüter, Phys. Rev. Lett. **51**, 1888 (1983).

[95] L. Sham and M. Schlüter, Phys. Rev. B **32**, 3883 (1985).

[96] J. P. Perdew and A. Zunger, Phys. Rev. B **23**, 5048 (1981).

[97] M. C. Gibson, S. J. Clark, S. Brand, and R. A. Abram, AIP Conf. Proc. **772**, 1125 (2005).

[98] K. Xiong, J. Robertson, M. C. Gibson, and S. J. Clark, App. Phys. Lett. **87**, 18 (2005).

[99] M. C. Gibson, S. Brand, and S. J. Clark, Phys. Rev. B **73**, 125120 (2006).

[100] R. M. Martin, *Electronic Structure: Basic Theory and Methods* (Cambridge University Press, 2004), iSBN: 0521782856.

[101] O. H. Nielsen and R. M. Martin, Phys. Rev. B **32**, 3780 (1985).

[102] O. H. Nielsen and R. M. Martin, Phys. Rev. B **32**, 3792 (1985).

[103] M. Nekovee, W. M. C. Foulkes, and R. J. Needs, Phys. Rev. Lett. **87**, 36401 (2001).

[104] J. P. Perdew and Y. Wang, Phys. Rev. B **46**, 12947 (1992).

[105] D. R. Bowler, T. Miyazaki, and M. J. Gillan, J. Phys. Cond. Matt. **14**, 2781 (2002).

[106] C. K. Skylaris, P. D. Haynes, A. A. Mostofi, and M. C. Payne, J. Chem. Phys. **122**, 084119 (2005).

[107] http://physics.nist.gov/cuu/Constants/index.html.