Ab initio studies of defect concentrations and diffusion in metal oxides

Kilian Frensch

Condensed Matter Theory Group
Department of Physics
Imperial College London

A thesis submitted for the degree of

Doctor of Philosophy

2011
“In physics, you don’t have to go around making trouble for yourself - nature does it for you.”

- Frank Anthony Wilczek
Declaration of Originality

I certify that, except where appropriately referenced, the work is that of the author alone and has not been previously submitted to qualify for any other academic award.

- Kilian Frensch, July 2011
Abstract

This work presents a methodology for determining the concentrations and diffusion coefficients of point defects in metal oxides using ab initio calculations of defect formation energies and diffusion barriers, and the binding energies of defect-impurity clusters. The methodology is applied to analyse the long-standing mysteries surrounding the mechanism of self-diffusion in α-Al₂O₃.

Al₂O₃ is a prototypical large band gap ceramic with extensive applications, many of which depend on its defect chemistry. In particular, point defect concentrations, that vary with temperature and impurity doping, govern diffusion properties such as creep, sintering, or the oxidation rate of Al-containing alloys. Experimental measurements of the self-diffusion coefficients in bulk alumina reveal three important truths that theory cannot reconcile, collectively termed the ‘corundum conundrum’. First, large experimental activation energies for oxygen and aluminum diffusion and low theoretical formation energies imply unreasonably high diffusion barriers of ~ 5eV. Second, aluminum diffusion is orders of magnitude faster than oxygen diffusion. Third, the oxygen diffusion coefficient is relatively insensitive to aliovalent doping, increasing by a factor of 100 on heavy Mg²⁺-doping, and decreasing by a similar amount on Ti⁴⁺-doping.

We attempt to resolve this conundrum by calculating the formation energies and binding energies of a raft of native point defects and defect-impurity clusters as functions of temperature T and oxygen partial pressure \( p_{O_2} \), and the diffusion barriers of the native defects, using density functional theory. We then use a thermodynamic mass action approach to determine the concentrations of the defects and clusters, and the diffusion coefficients of the defects, as functions of T, \( p_{O_2} \), and the concentrations of aliovalent dopants, [Mg²⁺] and [Ti⁴⁺]. In the process, we discover new ground-state defect structures for the aluminum vacancy and oxygen interstitial, and demonstrate that diffusion of aluminum vacancies and interstitials occurs by extended vacancy and interstitialcy mechanisms, and oxygen interstitials by a dumbbell interstitialcy mechanism, all of which yield much lower migration barriers than previous theory.

Unfortunately, the results do not demonstrate the experimentally-found insensitivity of the oxygen diffusivity to aliovalent doping. This could be an artefact of approximations within density functional theory and our methodology, and we show that modest changes in the calculated binding energies lead to significant defect cluster-
ing. This defect clustering could result in a buffering mechanism that can explain the insensitivity of the diffusivity to aliovalent doping, and may occur in other ionic materials. More accurate calculations, employing hybrid functionals or quantum Monte Carlo methods, may be necessary to elucidate this effect, but are currently computationally intractable for our purposes.
Acknowledgements

This PhD has undeniably been the most challenging episode of my life this far. It would not have been possible without the immense support I received during the time. Above all, I would like to thank Nick, the most devoted and brilliant young scientist I have met, for his neverending help and advice. Much of this work is due to him. Matthew and Mike have been a fountain of knowledge, motivation, and inspiration, and the best supervisors a student could ask for. Arthur contributed crucial elements to this work at a pivotal stage, and was a driving force during the final year of my work. I would like to thank my friends, in particular Ola and Adam, for putting up with much of the frustration that accompanied the work and making the time all the more fun. I could thank Racheal for any one of a number of things over the past years, but each would only describe a small part of what makes her special. Suffice it to say that she is most definitely ‘the better half’ and has contributed enormously to this thesis.

My family has always supported me and I thank them for their guidance and love. My father has been an inspiration to me from a young age, and I dedicate this thesis to him, my mother, and Antonia, who have always been the nucleus of my motivation.
# Table of Contents

1 Introduction 11

2 The many body problem 15
   2.1 Introduction ........................................... 15
   2.2 The Hamiltonian ........................................ 16
      2.2.1 The Born-Oppenheimer approximation .............. 18
   2.3 Density Functional Theory ............................ 19
      2.3.1 The first Hohenberg-Kohn theorem ............... 19
      2.3.2 The second Hohenberg-Kohn theorem ............ 20
      2.3.3 The Kohn-Sham ansatz ............................ 21
      2.3.4 Exchange-correlation functionals ............... 25
      2.3.5 Basis sets ........................................ 27
      2.3.6 Pseudopotentials ................................. 28
   2.4 Quantum Monte Carlo ................................ 30
      2.4.1 Variational Monte Carlo ......................... 30
      2.4.2 Diffusion Monte Carlo ............................ 32
      2.4.3 Trial wave functions ............................. 39
      2.4.4 Wave function optimisation ..................... 41

3 Point defects in solids 43
   3.1 Introduction .......................................... 43
   3.2 Defect species and notation .......................... 45
   3.3 Defect concentrations in the dilute limit ............ 47
      3.3.1 The law of mass action .......................... 47
      3.3.2 Defect concentrations in the canonical ensemble 48
      3.3.3 Intrinsic and extrinsic defects ................ 54
   3.4 Thermodynamics of defect formation energies ........ 57
      3.4.1 Zhang-Northrup formalism ....................... 57
      3.4.2 Chemical potentials ............................. 59
   3.5 Density functional theory calculations of defect formation energies .... 65
      3.5.1 Vibrational effects ............................... 65
      3.5.2 Band gap corrections ............................. 68
      3.5.3 Finite-size effects ............................... 71
### Table of Contents

3.6 Diffusion kinetics of defects ........................................ 78
  3.6.1 Mechanisms of diffusion ..................................... 79
  3.6.2 Activation energies and diffusion barriers ..................... 81
  3.6.3 Fick’s laws .................................................. 83
  3.6.4 Diffusion coefficients from random walk diffusion ............ 86
  3.6.5 Correlation effects .......................................... 88
  3.6.6 Experimental methods ........................................ 93

4 Point defects and diffusion in Al$_2$O$_3$ .......................... 97
  4.1 Introduction ..................................................... 98
  4.2 Experimental background ......................................... 98
    4.2.1 Oxygen diffusion ......................................... 99
    4.2.2 Oxygen diffusion buffering in doped alumina ............... 102
    4.2.3 Aluminium diffusion ................................... 105
    4.2.4 Anisotropy ............................................... 107
    4.2.5 Alternative species ...................................... 108
  4.3 Previous theoretical work ....................................... 109
  4.4 Density functional theory calculations .......................... 115
    4.4.1 Geometry and basis set convergence ......................... 115
    4.4.2 Defect species and complexes ................................ 122
    4.4.3 Defect formation energies and cluster binding energies .. 145
    4.4.4 Chemical potentials ..................................... 152
    4.4.5 Band gap corrections ...................................... 154
    4.4.6 Finite-size scaling ...................................... 156
    4.4.7 Vibrational contributions .................................. 160
    4.4.8 Diffusion barriers ....................................... 162
  4.5 Defect concentrations ........................................... 179
  4.6 Diffusion coefficients .......................................... 190
  4.7 Impurity solubilities ........................................... 194
  4.8 Quantum Monte Carlo calculations ............................... 198
    4.8.1 Trial wave functions ...................................... 199
    4.8.2 Jastrow optimisation ...................................... 200
    4.8.3 QMC band gaps ........................................... 202
    4.8.4 Errors .................................................... 203
    4.8.5 Diffusion barriers ....................................... 206
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5  Conclusion</td>
<td>209</td>
</tr>
<tr>
<td>A  Point defects in solids</td>
<td>213</td>
</tr>
<tr>
<td>A.1 Arrhenius relationships of activation energies of diffusion</td>
<td>213</td>
</tr>
<tr>
<td>A.2 Einstein equation from random walk diffusion</td>
<td>214</td>
</tr>
<tr>
<td>A.3 Diffusion coefficients from random walk diffusion</td>
<td>215</td>
</tr>
<tr>
<td>A.4 Correlation factors for native defect diffusion in alumina</td>
<td>217</td>
</tr>
</tbody>
</table>
Physicists, and scientists in general, enjoy classifying things into distinct categories. It is therefore only natural that they classify themselves, creating the great divide in physics, between ‘experimentalists’ and ‘theorists’. Of course, each group considers itself superior to the other, occasionally even suggesting that the other merely exists to support it. However, physics is impervious to all this. It remains physics, regardless of who studies it in what way. Neither experimentalists nor theorists ‘invent’ physics with their discoveries. In fact, both rely on the other for their own recognition, whether it be a theorist in need of experimental verification of their prediction, or an experimentalist in need of a theoretical explanation of their result. Ultimately, every physicist is actually part both, and if we wish to have a complete understanding of the world around us, experiment and theory must walk hand in hand.

This study was motivated by the gap between experiment and theory in the understanding of point defect diffusion in \( \text{Al}_2\text{O}_3 \), or alumina. For almost half a century, tracer diffusion studies on alumina doped with aliovalent impurities have shown that the concentrations of the native defects in alumina are strongly buffered. Theory has been unable to explain the cause of this buffering, in part due to the theoretical methods employed. This disparity between theory and experiment has been coined the ‘corundum conundrum’ in a recent paper [80]. The current study attempts to resolve this conundrum using a full \textit{ab initio}, meaning from first-principles, approach. The results from this approach suggest that density functional theory calculations employing an LDA functional are unable to rationalise this long-standing conundrum, and that further work using hybrid functionals or quantum Monte Carlo methods is necessary. Although results are presented for alumina, the theoretical
formalism underlying the current study is general, and can be applied to a wide range of other metal oxides.

In Chapter 2, the theoretical methods used in this study are presented: density functional theory (DFT) and quantum Monte Carlo (QMC). These are two very different approaches to solving the many electron problem that permeates all electronic structure calculations. DFT achieves this by recasting the fundamental description of electrons in solids, normally expressed in terms of wave functions, such that the electron density is the fundamental variable. Although in principle an exact method, the implementations of DFT render it approximate, but with the advantage of computational efficiency. QMC is the method of choice for achieving higher accuracy in condensed matter systems. QMC comprises a large body of stochastic methods, two of which, variational Monte Carlo (VMC) and diffusion Monte Carlo (DMC), are presented and used in this work. DMC is gaining recent recognition as the benchmark for accuracy in solid state systems, albeit at considerable computational cost.

In Chapter 3, the theory of point defects in solids and determination of defect properties and diffusion coefficients is discussed. A formalism derived in this study is presented for determining the concentrations of defects in the dilute limit. This formalism relies on accurate calculations of defect formation energies and chemical potentials. For the latter, crucial changes to an existing method are introduced, without which large errors are incurred. For defect formation energies, the necessity of using band gap and finite-size corrections, as well as including vibrational effects is emphasised. Finally, theoretical expressions for diffusion coefficients are derived, with particular application to alumina. The formalism presented in this section is of general applicability.

In Chapter 4, the framework presented in the preceding chapter is applied to alumina. After a brief overview of the previous experimental and theoretical work that motivated this study, the important point defects and clusters in alumina are presented, and their formation and binding energies, calculated using DFT, discussed. The effects of including the corrections to DFT formation energies presented in Chapter 3 are shown. Thereafter, migration barriers are determined for the principal native defects in alumina using DFT. The formation energies, cluster binding energies, and migration barriers are then combined in the formalism from Chapter 3 to determine defect concentrations and diffusion coefficients as a function of impurity doping, temperature, and oxygen partial pressure. The results are unable
to resolve the corundum conundrum, but show that modest changes in the binding energies of defect-dopant clusters can provide an explanation for the conundrum. Further work using methods of higher accuracy can show whether such changes in binding energies could result from errors intrinsic to DFT or approximations in our methodology, but were beyond the scope of this study. Finally, the migration barriers for the four native defects are determined using DMC in order to benchmark the previous DFT results and provide further insight into the capabilities of QMC methods.
“To do good work, one must first have good tools.”
- Chinese proverb

2

The many body problem

Contents

2.1 Introduction ....................................................... 15
2.2 The Hamiltonian .................................................. 16
  2.2.1 The Born-Oppenheimer approximation ...................... 18
2.3 Density Functional Theory ....................................... 19
  2.3.1 The first Hohenberg-Kohn theorem ......................... 19
  2.3.2 The second Hohenberg-Kohn theorem ....................... 20
  2.3.3 The Kohn-Sham ansatz ...................................... 21
  2.3.4 Exchange-correlation functionals ......................... 25
  2.3.5 Basis sets .................................................. 27
  2.3.6 Pseudopotentials ........................................... 28
2.4 Quantum Monte Carlo ............................................ 30
  2.4.1 Variational Monte Carlo .................................. 30
  2.4.2 Diffusion Monte Carlo .................................... 32
  2.4.3 Trial wave functions ...................................... 39
  2.4.4 Wave function optimisation ............................... 41

2.1 Introduction

The capacity of any theory is measured by its ability to accurately describe experimental results and predict previously unknown phenomena. The advent of quantum mechanics in the early 20th century was met with universal controversy, surpassed only by admiration for its success. Precocious explanations of hitherto mysterious experimental data, coupled with the demonstrated ability to predict profound phenomena, have since cemented the value of quantum mechanics as a theory for describing nature. Suddenly, the properties of any system could be found by solving
Chapter 2: The many body problem

Schrödinger’s equation for its Hamiltonian. Unfortunately, Schrödinger’s equation cannot be solved analytically for systems containing more than two interacting particles. This dilemma, coined the many body problem, has prompted the development of diverse approaches to its solution. In some situations, the many body problem can be simplified to neglect a raft of quantum effects. Methods such as classical inter-atomic potentials follow this approach by constructing parametrised potentials usually fit to experimental data, and often suffice to model specific properties of systems. However, many physical phenomena result from quantum effects. \textit{Ab initio} methods aim to accommodate these effects and solve the Schrödinger equation from first-principles, without adopting any empiricism. Two of the most prominent of these methods are density functional theory (DFT) and quantum Monte Carlo (QMC). Recent computational advancements have enabled the simulation of systems containing hundreds, in some cases even thousands, of atoms using DFT and sparked a flurry in research, ranging from examining the properties of proteins in biology to understanding how matter diffuses along dislocations and grain boundaries in polycrystalline materials. In situations where accuracy is paramount, the more demanding QMC techniques are able to provide insight beyond that found in other methods, and thus provide a benchmark for approximate methods such as DFT. At the core of every \textit{ab initio} technique lies the Hamiltonian of the system in question.

2.2 The Hamiltonian

All electronic structure problems can be addressed by solving the time-dependent Schrödinger equation

\[
\hbar i \frac{\partial \Phi}{\partial t} = \hat{H} \Phi, \tag{2.1}
\]

where $\Phi$ is the many body wave function and $\hat{H}$ the many body Hamiltonian. For a system consisting of $N_e$ electrons of spin $\{\sigma_i\}$ at positions $\{r_i\}$ interacting with $N_n$ nuclei of atomic number $\{Z_\alpha\}$ and mass $\{M_\alpha\}$ at positions $\{R_\alpha\}$, $\hat{H}$ takes the form
Section 2.2: The Hamiltonian

\[
\hat{H} = -\frac{1}{2} \sum_{i=1}^{N_e} \nabla_i^2 - \frac{1}{2} \sum_{\alpha=1}^{N_n} \frac{1}{M_\alpha} \nabla_\alpha^2 - \sum_{i=1}^{N_e} \sum_{\alpha=1}^{N_n} \frac{Z_\alpha}{|r_i - R_\alpha|} \\
+ \frac{1}{2} \sum_{i,j=1}^{N_e} \sum_{j \neq i}^{N_e} \frac{1}{|r_i - r_j|} + \frac{1}{2} \sum_{\alpha=1}^{N_n} \sum_{\beta \neq \alpha}^{N_n} \frac{Z_\alpha Z_\beta}{|R_\alpha - R_\beta|},
\]

(2.2)

when using atomic units, where \( \hbar = e = m_e = 4\pi\varepsilon_0 = 1 \). The first and second terms are the kinetic energy of the electrons and nuclei, the third the Coulomb attraction between the electrons and nuclei, the fourth the electron-electron repulsion and the fifth the nuclear repulsion. Relativistic spin-orbit coupling, magnetic and mass-velocity effects are all assumed to be very small and omitted, but can be included in a perturbation theory approach. In the case of Eq. 2.2, the wave function \( \Phi \) is a function of nucleus position, electron position and spin coordinates, and the time \( t \):

\[
\Phi \equiv \Phi (\{R_\alpha\}, \{r_i, \sigma_i\}, t).
\]

(2.3)

As none of the potential energy terms in the Hamiltonian are explicitly time-dependent, we can perform a separation of variables without loss of generality to obtain

\[
\Phi (\{R_\alpha\}, \{r_i, \sigma_i\}, t) = \Psi (\{R_\alpha\}, \{r_i, \sigma_i\}) \exp (-iEt),
\]

(2.4)

and thereby reduce the many body problem to solving the time-independent, non-relativistic Schrödinger equation

\[
\hat{H} \Psi (\{R_\alpha\}, \{r_i, \sigma_i\}) = E \Psi (\{R_\alpha\}, \{r_i, \sigma_i\}),
\]

(2.5)

with \( \hat{H} \) given in Eq. 2.2 and where \( E \) is the total energy of the system.

However, the time-independent Schrödinger equation is still intractable for the current systems of interest, which contain a large number \( N \) of interacting particles, and thus a corresponding Hilbert space whose dimension rises exponentially with \( N \). This is referred to as the many body problem. We will introduce a series of well-understood approximations in order to simplify the problem to a level such that its solutions are computationally accessible.
2.2.1 The Born-Oppenheimer approximation

The first of these approximations is the Born-Oppenheimer approximation [21], which assumes that, during the time-scale of nuclear motion, the electrons relax to the instantaneous ground-state configuration, allowing us to approximate the nuclei as stationary.

On these grounds, we can set the nuclear kinetic energy term in the Hamiltonian in Eq. 2.2 equal to zero, and substitute the interaction energy of the nuclei with a constant value $E_{\alpha\alpha}$. We have thus replaced the intractable many body Hamiltonian with the still intractable many electron Hamiltonian

$$\hat{H} = -\frac{1}{2} \sum_{i=1}^{N_e} \nabla_i^2 - \sum_{i=1}^{N_e} \sum_{\alpha}^{N_n} \frac{Z_\alpha}{|r_i - R_\alpha|} + \frac{1}{2} \sum_{i=1}^{N_e} \sum_{j \neq i}^{N_e} \frac{1}{|r_i - r_j|} + E_{\alpha\alpha},$$

(2.6)

where we have also chosen to view the Coulomb attraction between electrons and nuclei as an ‘external’ potential

$$V_{\text{ext}}(r_i) = \sum_{\alpha=1}^{N_n} \frac{-Z_\alpha}{|r_i - R_\alpha|},$$

(2.7)

due to the nuclei, acting on each electron.

The Born-Oppenheimer electronic Hamiltonian can be rewritten more compactly as

$$\hat{H} = \hat{T} + \hat{V}_{\text{ext}} + \hat{V}_{ee} + E_{\alpha\alpha},$$

(2.8)

where $\hat{T}$ is the kinetic energy of the electrons, $\hat{V}_{\text{ext}}$ the external potential, and $\hat{V}_{ee}$ the electronic Coulomb repulsion. The electron-electron repulsion term in Eq. 2.6 precludes any exact solutions to the corresponding electronic Schrödinger equation for systems of more than two interacting particles. This is known as the many electron problem, and attempts to recast it in a computable form provide the basis for electronic structure theory and the following sections.
2.3 Density Functional Theory

Density functional theory builds on the earlier Hartree-Fock theory [61, 62, 166, 167] by providing an approach to treat correlated many body systems in a practical framework. This framework revolves around an independent-particle basis, which allows for the computationally inexpensive determination of correlation effects, enabling a quantitatively accurate description of many systems. As such, it has become a cornerstone for modern theory and simulation of materials.

Density functional theory (DFT) is an exact theory of many body systems. The fundamental tenet of DFT states that any property of a system of many interacting particles can be derived from a functional of its ground-state density, \( \rho(r) \). This holds true both for ground-state, as well as excited-state properties. From this, we can replace the complicated \( N \)-electron wave function \( \Psi(r_1, r_2, \ldots, r_N) \) with the simple single particle electron density \( \rho(r) \). The proofs that such functionals linking the ground-state density to the properties of a many particle system exist are surprisingly simple and comprised Hohenberg, Kohn and Mermin’s seminal work [83, 124], for which Kohn was awarded a Nobel Prize in 1998.

2.3.1 The first Hohenberg-Kohn theorem

We start with the electronic Hamiltonian for \( N \) non-relativistic, interacting electrons,

\[
\hat{H} = -\frac{1}{2} \sum_{i=1}^{N} \nabla_{i}^2 + \sum_{i=1}^{N} V_{\text{ext}}(r_i) + \frac{1}{2} \sum_{i=1}^{N} \sum_{j \neq i}^{N} \frac{1}{|r_i - r_j|} \]

\[= \hat{T} + \hat{V}_{\text{ext}} + \hat{V}_{\text{ee}}, \tag{2.9}\]

where \( \hat{T} \) is the kinetic energy of the electrons, \( \hat{V}_{\text{ext}} \) an arbitrary external potential, and \( \hat{V}_{\text{ee}} \) the electronic Coulomb repulsion. Note that \( \hat{V}_{\text{ext}} \) need not be restricted to purely Coulombic potentials. We see that, as long as \( \hat{V}_{\text{ee}} \) remains Coulombic, the Hamiltonian is completely defined by the external potential \( \hat{V}_{\text{ext}} \) and the number of electrons \( N \). From the variational principle [155], minimising the energy functional

\[
E[\Psi] = \frac{\langle \Psi | \hat{H} | \Psi \rangle}{\langle \Psi | \Psi \rangle} \tag{2.10}\]
yields the ground-state wave function and energy. Therefore, ˆV_{ext} and N determine all the properties of the ground state.

The first Hohenberg-Kohn theorem [83] states that the external potential ˆV_{ext} is uniquely determined (up to a constant) by the ground-state electron density ˆρ(r). As ˆρ(r) determines the number of electrons from

\[ \int \rho(r) dr = N, \]  

(2.11)

this means that the full, many body ground-state (and all excited states) is a unique functional of ˆρ(r). The proof of this theorem follows by reductio ad absurdum [119]. However, this interesting result has not yet provided a prescription for dealing with systems of many interacting electrons, as all we have proven is that ˆρ(r) uniquely determines V_{ext}(r). We must still solve the many body problem in the presence of V_{ext}(r), which, in the case of electrons, is the Coulomb potential due to the nuclei.

### 2.3.2 The second Hohenberg-Kohn theorem

The second Hohenberg-Kohn theorem shows that it is possible to define an electronic energy functional of the density ˆρ(r) via

\[ E_V[\hat{\rho}(\mathbf{r})] = T[\hat{\rho}(\mathbf{r})] + V_{ee}[\hat{\rho}(\mathbf{r})] + \int V_{ext}(r)\hat{\rho}(r) dr = F_{HK}[\hat{\rho}(\mathbf{r})] + \int V_{ext}(r)\hat{\rho}(r) dr, \]  

(2.12)

where \( T[\hat{\rho}(\mathbf{r})] \) is the kinetic energy, \( V_{ee}[\hat{\rho}(\mathbf{r})] \) the electron-electron repulsion energy, \( F_{HK}[\hat{\rho}(\mathbf{r})] = T[\hat{\rho}(\mathbf{r})] + V_{ee}[\hat{\rho}(\mathbf{r})] \) the Hohenberg-Kohn functional, and we have made the dependence on \( V_{ext}(\mathbf{r}) \) explicit.

For a given \( V_{ext}(\mathbf{r}) \), the exact ground-state energy of the system is the global minimum of \( E_V[\hat{\rho}(\mathbf{r})] \), and the density that minimises \( E_V[\hat{\rho}(\mathbf{r})] \) is the exact ground-state density \( \rho_0(\mathbf{r}) \):

\[ E_0 \leq E_V[\hat{\rho}(\mathbf{r})]. \]  

(2.13)

This means that if we knew the explicit form of \( F_{HK}[\hat{\rho}(\mathbf{r})] \), we would be able to solve the Schrödinger equation exactly. As we do not know this, we are forced to
introduce approximations.

The original proof [119] of the second Hohenberg-Kohn theorem is restricted to ‘V-representable’ densities that are ground-state densities of the Hamiltonian with some external potential $V_{\text{ext}}(r)$. However, it can be extended to all other densities by the method of Levy and Leib [107, 108, 110, 111]. This formulation also reduces the original minimisation of $\langle \tilde{\Psi} | \hat{H} | \tilde{\Psi} \rangle$ with respect to the $3N$-dimensional trial wave function $\tilde{\Psi}$ to a minimisation of $E_{V}[\tilde{\rho}(r)]$ with respect to the 3-dimensional trial density $\tilde{\rho}(r)$. However, this still involves a search over all antisymmetric wave functions that have the given trial density, rendering this minimisation intractable. In order to get closer to an accurate solution that we can calculate in a reasonable time, it is necessary to find a better way to formulate the universal functional.

2.3.3 The Kohn-Sham ansatz

In 1965, Kohn and Sham [96] developed an ansatz that allows the exact calculation of the ground-state energy of a system using single particle methods. In order to achieve this, they replaced the complicated interacting system with a chosen non-interacting reference system that has the same exact ground-state density as the original. The ground-state density of the auxiliary reference system is then found by minimising a known functional that replaces the explicit interactions that we cannot calculate by a local effective potential, $V_{\text{KS}}(r)$. This method then involves independent particles linked by an interacting density. The single particle equations for the non-interacting system are, in principle, exactly solvable, and all of the difficult many body terms are incorporated into the so-called exchange-correlation functional, $E_{\text{xc}}$. Solving these equations then gives the exact ground-state density and energy with an accuracy limited only by the approximations made to $E_{\text{xc}}$.

Various surprisingly simple approximations have been very successful in describing the properties of a wide range of systems, from metals to wide band gap insulators. From the second Hohenberg-Kohn theorem, it is possible to determine the ground-state energy of a many electron system by minimising an energy functional

$$E[\rho(r)] = F_{\text{HK}}[\rho(r)] + \int V_{\text{ext}}(r)\rho(r)dr$$

$$\equiv T[\rho(r)] + V_{\text{xc}}[\rho(r)] + \int V_{\text{ext}}(r)\rho(r)dr \quad (2.14)$$
determined solely by the electron density $\rho(\mathbf{r})$. The Thomas-Fermi model [175, 55] and others approximate the form of $T[\rho(\mathbf{r})]$ and $V_{ee}[\rho(\mathbf{r})]$, but these approximations do not give accurate results for most many electron systems. Kohn and Sham realised that, rather than approximate the full form of the kinetic energy functional $T[\rho(\mathbf{r})]$, a better approach would involve computing as large a portion of the kinetic energy as possible exactly. The small residual correction could then be handled separately with an approximate functional. Using the Kohn-Sham ansatz to replace the interacting system with a non-interacting reference system gives a Hamiltonian

$$\hat{H}_S = -\frac{1}{2} \sum_{i=1}^{N} \nabla_i^2 + \sum_{i=1}^{N} V_{KS}(\mathbf{r}_i), \quad (2.15)$$

where we have omitted the electron-electron repulsion term and introduced an effective local\(^1\) potential $V_{KS}(\mathbf{r}_i)$, the Kohn-Sham potential. For this system, the ground-state electron density is exactly $\rho(\mathbf{r})$. We can then write the ground-state wave function of the reference system using the Slater determinant from Hartree-Fock theory

$$\Psi_S = \frac{1}{\sqrt{N!}} \det [\psi_1 \psi_2 \cdots \psi_N], \quad (2.16)$$

where the Kohn-Sham orbitals $\psi_i(\mathbf{r})$ are the $N$ lowest eigenstates of the one electron Hamiltonian $\hat{h}_S$

$$\hat{h}_S \psi_i(\mathbf{r}) = \left[ -\frac{1}{2} \nabla^2 + V_{KS}(\mathbf{r}) \right] \psi_i(\mathbf{r}) = \varepsilon_i \psi_i(\mathbf{r}), \quad (2.17)$$

and we have omitted the electrons’ spin, which can, however, be readily incorporated. The kinetic energy in the non-interacting system can now be computed exactly subject to the orthonormality of the orbitals, in the same way as in Hartree-Fock theory,

$$T_S = -\frac{1}{2} \sum_{i=1}^{N} \langle \psi_i(\mathbf{r}) | \nabla^2 | \psi_i(\mathbf{r}) \rangle = -\frac{1}{2} \sum_{i=1}^{N} \int \psi_i^*(\mathbf{r}) \nabla^2 \psi_i(\mathbf{r}) d\mathbf{r}, \quad (2.18)$$

and is uniquely defined for any density. Although $T_S[\rho(\mathbf{r})]$ is not equal to the kinetic

\(^1\)Although the local form is not essential, the simplification is commonly implemented in practice.
energy of the corresponding interacting system, it is similar and can be subtracted from the universal functional $F_{\text{HK}} [\rho (r)]$ known to exist by the second Hohenberg-Kohn theorem. The Kohn-Sham ansatz replaces the total energy functional from Eq. 2.14 with

$$E_{\text{KS}} [\rho (r)] = T_S [\rho (r)] + J [\rho (r)] + E_{\text{xc}} [\rho (r)] + \int V_{\text{ext}}(r)\rho (r)dr, \quad (2.19)$$

where

$$J [\rho (r)] = \frac{1}{2} \int \frac{\rho (r)\rho (r')}{|r - r'|} drdr', \quad (2.20)$$

is the classical Coulomb interaction of the electron density $\rho (r)$ interacting with itself (the Hartree energy), $T_S [\rho (r)]$ is the kinetic energy of the non-interacting system, and $E_{\text{xc}} [\rho (r)]$ is the exchange-correlation energy, which encompasses the difference between $T [\rho (r)]$ and $T_S [\rho (r)]$ and between $V_{\text{ee}} [\rho (r)]$ and $J [\rho (r)]$. To exemplify this, we can write $E_{\text{xc}} [\rho (r)]$ as

$$E_{\text{xc}} [\rho (r)] = F_{\text{HK}} [\rho (r)] - (T_S [\rho (r)] + J [\rho (r)]) = (T [\rho (r)] - T_S [\rho (r)]) + (V_{\text{ee}} [\rho (r)] - J [\rho (r)]). \quad (2.21)$$

We can find the ground-state energy of the Kohn-Sham system, and thus also that of the full interacting system, by minimising the total energy functional given in Eq. 2.19 with respect to varying the Kohn-Sham orbitals $\psi_i (r)$. Using the chain rule gives

$$\frac{\delta E_{\text{KS}}}{\delta \psi_i^*(r)} = \frac{\delta T_S}{\delta \psi_i^*(r)} + \left[ \frac{\delta J}{\delta \rho (r)} + \frac{\delta E_{\text{xc}}}{\delta \rho (r)} + V_{\text{ext}} (r) \right] \frac{\delta \rho (r)}{\delta \psi_i^*(r)} = 0. \quad (2.22)$$

From Eq. 2.18 and the density of the reference system

$$\rho (r) = \sum_{i=1}^{N} |\psi_i (r)|^2, \quad (2.23)$$

we have
Using the method of Lagrange multipliers and imposing the normalisation constraint $\langle \psi_i | \psi_i \rangle = 1$, this yields the Kohn-Sham equations:

$$\hat{H}_{KS} \psi_i(r) = \varepsilon_i \psi_i(r),$$

(2.25)

where $\hat{H}_{KS}$ is the effective Kohn-Sham Hamiltonian

$$\hat{H}_{KS}(r) = -\frac{1}{2} \nabla^2 + V_{KS}(r)$$

(2.26)

and it can be shown [119] that

$$V_{KS}(r) = V_{\text{ext}}(r) + \frac{\delta J}{\delta \rho(r)} + \frac{\delta E_{xc}}{\delta \rho(r)}$$

(2.27)

is the Kohn-Sham potential linking the non-interacting and interacting systems. From the Hohenberg-Kohn theorems, there exists a unique $V_{KS}(r)$ for a given interacting system.

These single particle equations are self-consistent and must be solved in an iterative procedure. Starting from a guess for $\rho(r)$, $V_{KS}(r)$ can be constructed from Eq. 2.27, and then a new $\rho(r)$ can be found from Eqs. 2.25 and 2.23. The total energy is then computed from Eq. 2.19. As no approximation has been made, the Kohn-Sham equations would give the exact ground-state density and energy for the interacting system if the exact form of $E_{xc}[\rho(r)]$ were known. However, it can be proven that the exact functional is not analytic [143, 162], and incorporating the short and long-range effects of many body correlation provides a formidable task.

Thus, we see that the Kohn-Sham ansatz brings us two decisive advantages: first, we have reduced the complicated interacting many body problem to a set of tractable single particle equations, and second, we have separated the universal functional into three parts. The independent particle kinetic energy and long-range Hartree energy are straightforward, leaving the exchange-correlation functional as the final hurdle en route to an accurate and computationally efficient determination of the ground-state properties of solids.
2.3.4 Exchange-correlation functionals

In 1972, Richard Feynman referred to the exchange-correlation energy as the ‘stupidity energy’ [56]; however, this quote would prove premature, as one of the reasons for DFT’s great success in recent years has been that simple approximations to the exchange-correlation functional $E_{xc}[\rho(r)]$ can produce very accurate results in a wide range of systems. There are many flavours of DFT that rely on assuming different functional forms for $E_{xc}[\rho(r)]$. Many solids are quite homogeneous, such that the effects of exchange and correlation are local in character. As an initial approximation to $E_{xc}[\rho(r)]$, it therefore makes sense to account for this.

In the Local Spin Density Approximation (LSDA) [96, 144, 184] the exchange-correlation energy density at a given point is assumed to be the same as in a homogeneous electron gas with that density:

$$E_{xc}^{\text{LSDA}}[\rho^\uparrow(\mathbf{r}), \rho^\downarrow(\mathbf{r})] = \int \rho^\uparrow(\mathbf{r})e_{xc}^{\text{hom}}(\rho^\uparrow(\mathbf{r}), \rho^\downarrow(\mathbf{r}))d\mathbf{r},$$

(2.28)

where $e_{xc}^{\text{hom}}(\rho^\uparrow(\mathbf{r}), \rho^\downarrow(\mathbf{r}))$ is the exchange-correlation energy per electron in a uniform electron gas of spin density $(\rho^\uparrow(\mathbf{r}), \rho^\downarrow(\mathbf{r}))$. In a system with no net spin, the Local Density Approximation (LDA) is found by setting $\rho^\uparrow(\mathbf{r}) = \rho^\downarrow(\mathbf{r}) = \rho(\mathbf{r})/2$. The homogeneous electron gas is a neutral system of electrons moving in a uniform positively charged background.

We expect the LSDA to work best for systems with a nearly uniform charge density, such as nearly-free-electron metals, and not as well for strongly inhomogeneous systems such as atoms and molecules.

However, there are well-documented failures of these approximations, even in relatively homogeneous systems, such as the ‘band gap problem’ and the self-interaction error. The first refers to the underestimation of the band gap of insulators and semiconductors by up to 50%. This arises from the inability of simple exchange-correlation functionals to describe the discontinuity in the exact exchange-correlation potential across the band gap, resulting in an underestimation of the energies of excited states. In addition, it is exacerbated by the second issue, the self-interaction error. While, in Hartree-Fock theory, self-interaction exactly cancels, in DFT it does not. Approximations made to $E_{xc}[\rho(\mathbf{r})]$ are independent of $J[\rho(\mathbf{r})]$, so that the cancellation is only partial. This means that the electrostatic term does not vanish, even for a one electron system.
Despite these complications, the LSDA performs very well for both weakly and strongly inhomogeneous solids. In weakly inhomogeneous solids, geometries are typically very accurate, with bond lengths and lattice parameters determined to within 5% of experiment. Calculations of dissociation energies of strongly inhomogeneous systems such as dimers are particularly challenging, with errors of 10-20% common.

The next step building on the LSDA involves including the gradient of the density in the functional in so-called Generalised Gradient Approximations (GGA). The inclusion of first derivatives of the density allows GGA functionals to accommodate systems with slowly varying densities, and GGAs have shown improvements over the LSDA in a wide range of atoms, molecules and solids. In particular, molecules containing weak interactions tend to be better described, as do cohesive energies of solids. At short ranges, the GGA leads to an exchange energy that is smaller than in the LSDA, which corrects for the systematic overbinding of LSDA in this regime.

Hybrid functionals, which have become the functional of choice for many quantum chemists, attempt to overcome traditional DFT problems by mixing a portion of exact exchange from Hartree-Fock theory with explicit exchange-correlation functionals from DFT. One common hybrid functional, B3LYP [16], parametrises a mixture of Hartree-Fock exact exchange (EXX) and approximations to the gradient corrections to the exchange and correlation energies, $\Delta E_{\text{Becke}}^x$ and $\Delta E_{\text{LYP}}^c$, by Becke [15] and Lee, Yang and Parr [101], respectively, in the form

$$E_{\text{B3LYP}}^{\text{xc}} = E_{\text{xc}}^{\text{LDA}} + a_0 \left( E_{\text{x}}^{\text{EXX}} - E_{\text{x}}^{\text{LDA}} \right) + a_x \Delta E_{\text{x}}^{\text{Becke}} + a_c \Delta E_{\text{c}}^{\text{LYP}},$$

(2.29)

where the coefficients are adjusted to give a good fit to experimental data. Although hybrids have been shown to produce reliable geometries and band gaps in semiconductors [13] as well as insulators, their application to untested materials remains uncertain.

The different types of functionals described above all attempt to approximate the unknown exchange-correlation functional $E_{\text{xc}}$, but there remains no systematic method of improving $E_{\text{xc}}$, and functionals can only be assessed by their ability to reproduce a wide range of experimental results to high accuracy.
Section 2.3: Density Functional Theory

2.3.5 Basis sets

When we deal with a bulk-like structure, we perform a calculation on an infinite periodic array of unit cells of this structure, where each unit cell contains a computationally tractable number of electrons. Using Bloch’s theorem, we can then reduce the infinite number of one electron wave functions in the periodic array to the finite number contained in the unit cell (or even half that for a system with zero net spin), at the cost of introducing an additional integration over the Brillouin zone. The wave functions can then be expanded as discrete sums of plane waves

\[ \psi_{ik}(r) = \sum_G c_{ik}(G) \exp(i(k + G)r), \]

whose wave vectors are reciprocal lattice vectors \( G \) of the crystal. A plane wave (PW) basis set offers a number of advantages, such as the flexibility to describe any form of solution. As the electronic wave functions vary smoothly over the first Brillouin zone, they can be sampled at a finite number of special \( k \)-points. The plane waves used to represent the wave function comprise a discrete set at each \( k \), but remain infinite in number. In order to render the problem computationally efficient, the basis set is made finite by only including those plane waves with a cut-off energy \( E_{\text{cut}} \)

\[ E_{\text{cut}} \leq \frac{\hbar^2 |k + G|^2}{2m_e}. \]

Therefore, the error in the total energy of the system caused by truncating the basis set can be reduced by increasing \( E_{\text{cut}} \). As \( E_{\text{cut}} \to \infty \), the basis set approaches completeness. This makes it possible to examine the convergence of the calculation with respect to the basis set by varying the parameter \( E_{\text{cut}} \). In comparison, other basis sets, such as Gaussian-type orbitals, cannot be as easily systematically examined for convergence.

The main disadvantage of PW basis sets lies in the number of basis functions required for an accurate representation of the Kohn-Sham orbitals. This problem may be mitigated by the use of pseudopotentials, as described in the following section.

Gaussian-type orbitals (GTOs) require a much smaller number of basis functions per atom and are traditionally employed in calculations of atoms and molecules, where it is expected that localised basis functions are particularly suitable. In Cartesians, Gaussian-type orbitals take the general form of a sum
of so called ‘contracted’ Gaussian-type basis functions

\[ \phi_\mu(r) = \sum_{j=1}^{L} d_j g_j(r), \]  

where

\[ g_j(r) = N \exp(-\alpha_j r^2) x^l y^m z^n \]  

are individually normalised Gaussian ‘primitives’, \( g_j(r) \), and \( l, m, \) and \( n \) are integral exponents of the Cartesian coordinates. With the proper choice of \( L \), the length of the contraction, \( \alpha_j \), the contraction exponents, and \( d_j \), the contraction coefficients, the linear combination in Eq. 2.33 can assume any reasonable functional form. This enables the GTOs to be chosen to reflect desired properties of the system. Combinations of Gaussian primitives can approximate the nodal features of atomic orbitals, but the primitives themselves do not resemble exact Hartree-Fock atomic orbitals. Rather, their properties are chosen for computational convenience.

Some advantages of Gaussians are their simple separability into Cartesian components, and their local nature, which facilitates the calculation of ‘atomic-like’ properties such as orbital populations and projected densities of states. However, the loss of universality means a new basis set must be created for each element, or even the same element in different environments. In addition, the computation of forces is difficult, and, in systems which break symmetries, such as ones this study deals with, the efficiency of the calculation is strongly undermined.

### 2.3.6 Pseudopotentials

The tools described in the preceding sections are sufficient to apply density functional theory to many atomic and molecular problems. However, when considering larger systems containing hundreds of atoms, treating all of the electrons explicitly becomes computationally prohibitive. Part of this problem resides in representing the core states in a plane wave basis set. Moreover, the wave functions of valence electrons oscillate rapidly in the region occupied by the core electrons bound by the
strong ionic potential, to maintain the orthogonality between the core and valence electrons. Accounting for this also requires a very large basis set, rendering calculations expensive. Fortunately, it is well known that the tightly-bound core electrons on atoms are almost independent of the chemical environment surrounding them, and that the structure and dynamics of valence electron states alone explain most physical and chemical properties. This is true in particular for the formation of chemical bonds, but also for magnetic behaviour and low-energy excitations. The main reason for the inertness of the core electrons in many processes is the spatial separation of the core and valence states. The highly localised core electrons have little overlap with either the wave functions of valence electrons from their own atom, or those from surrounding atoms. They therefore do not interact directly with neighbouring atoms, and only experience their presence indirectly via the deformation of the valence orbitals, which act only weakly on the tightly-bound core orbitals. This allows for the decomposition into core, valence, and excited states.

In the pseudopotential approximation [146, 34] the core electron states of a given atomic species are assumed to be fixed, and the effects of the nucleus and core electrons are replaced with a smooth effective pseudopotential \( \hat{V}_{\text{pp}} \) in which valence electrons move. The Schrödinger equation then contains an effective potential term instead of the Coulombic potential term for core electrons, reducing the number of electrons that are considered explicitly. The valence electrons are represented by pseudo wave functions that correspond to the pseudopotential and do not exhibit the rapid oscillations of the true wave functions in the core region, reducing the size of the basis set required to represent them. In some cases, this can only be made possible by including semi-core or excited states in addition to the usual valence states. For example, in titanium, the most weakly bound \( 4s \)-state is not well-separated from the \( 3d \)-states. As a result, there is a substantial overlap of the \( 3d \)-orbitals with the valence orbitals from neighbouring atoms. In fact, it has been shown that not only the \( 4s \) and \( 3d \), but also, to a lesser degree, the \( 3s \) and \( 3p \) semi-core electrons in titanium participate in bonding [186]. Discarding any of these electrons from explicit calculation, and including them in a pseudopotential, could result in poor transferability, and may not reproduce accurate bonding effects in differing environments. For calculations involving heavy ions, relativistic effects can be incorporated into the pseudopotential, while valence electrons are still treated non-relativistically.

The neglect of the energetic contribution from core electrons renders total energies alone meaningless: only energy differences carry physical meaning. The pseudopo-
The many body problem was first introduced by Hellmann in 1935 [78] and has since become indispensable for large-scale many body calculations.

2.4 Quantum Monte Carlo

The advent of modern parallel computing has elicited the development of novel techniques for solving Schrödinger’s equation to high accuracy. A prominent exponent of this generation is quantum Monte Carlo (QMC), which characterises a group of methods that employ stochastic sampling. There are many different flavours of QMC, such as variational and diffusion, path integral and reptation Monte Carlo, of which only the first two will be presented here in detail. In general, stochastic methods are highly parallelisable and efficient for high-dimensional problems, such as those encountered in many-electron electronic structure methods, and offer unmatched accuracy, rendering them advantageous for calculations of the static properties of systems. In practice, however, they remain computationally challenging and have been little applied to complex systems such as defects in metal oxides.

2.4.1 Variational Monte Carlo

The variational Monte Carlo (VMC) method, the simpler of the two considered here, combines the variational principle with Monte Carlo integration to evaluate the expectation value of any Hermitian operator \( \hat{O} \). In a variational prescription, the ground-state expectation value of \( \hat{O} \) can be approximated from a parametrised trial wave function \( \tilde{\Psi}(\mathbf{R}) \) by

\[
\langle \hat{O} \rangle = \frac{\int \tilde{\Psi}^*(\mathbf{R}) \hat{O} \tilde{\Psi}(\mathbf{R}) d\mathbf{R}}{\int \tilde{\Psi}^*(\mathbf{R}) \tilde{\Psi}(\mathbf{R}) d\mathbf{R}}.
\] (2.35)

In the case of the ground-state energy, obtained from the Born-Oppenheimer Hamiltonian \( \hat{H} \) in Eq.2.6, \( \langle \hat{H} \rangle \) constitutes an upper bound to the exact ground-state value \( \langle \hat{H} \rangle_0 = E_0 \) that is obtained when \( \tilde{\Psi}(\mathbf{R}) \propto \Psi_0(\mathbf{R}) \):

\[
E_{\text{VMC}} = \langle \hat{H} \rangle = \frac{\int \tilde{\Psi}^*(\mathbf{R}) \hat{H} \tilde{\Psi}(\mathbf{R}) d\mathbf{R}}{\int \tilde{\Psi}^*(\mathbf{R}) \tilde{\Psi}(\mathbf{R}) d\mathbf{R}} \geq E_0.
\] (2.36)

In order to evaluate Eq. 2.36 using Monte Carlo integration, we must rewrite it in the form
\[ I = \int f(R)P(R)dR, \quad (2.37) \]
as the product of a normalised probability density \( P(R) \) and a score function \( f(R) \).

The choice of the importance function \( P(R) \) from which the random sampling points are chosen is not arbitrary and governs the efficiency of the sampling procedure. In our case, it is natural (and also efficient) to set \( P(R) \propto \tilde{\Psi}(R)^2 \). Using this substitution for \( P(R) \), we can rewrite Eq. 2.36 as

\[ E_{\text{VMC}} = \frac{\int \tilde{\Psi}^*(R)\tilde{\Psi}(R)[\tilde{\Psi}^{-1}(R)\hat{H}\tilde{\Psi}(R)]dR}{\int \tilde{\Psi}^*(R)\tilde{\Psi}(R)dR} = \int E_L(R) \left( \frac{\tilde{\Psi}(R)^2}{\int \tilde{\Psi}(R')^2dR'} \right) dR, \quad (2.38) \]

where the score function

\[ E_L(R) = \frac{\hat{H}\tilde{\Psi}(R)}{\tilde{\Psi}(R)} \quad (2.39) \]
is called the ‘local energy’ as it depends on the wave function only at a single point \( R \).

Given a probability distribution \( P(R) \) from which the random sampling points are to be chosen, we now need a method of performing the sampling efficiently. Metropolis et al. [125] devised an algorithm that samples \( P(R) \) by performing a random walk through the configuration space of the probability distribution, generating a sequence of points distributed according to \( P(R) \). For our purposes, the distribution we are trying to sample is \( |\Psi(r_1, r_2, \ldots, r_N)|^2 \), where each point along the walk is a 3N-dimensional vector of electron coordinates, \( R = \{r_1, r_2, \ldots, r_N\} \) (and possibly spin coordinates, which will be omitted for clarity). A particular value of \( R \) is called a ‘walker’, or configuration.

Using the Metropolis algorithm, \( E_{\text{VMC}} \) can be estimated by averaging the local energy \( E_L(R) \) over a sufficiently large set of sampled points \( \{R_1, R_2, \ldots, R_N\} \) from \( P(R) \):

\[ E_{\text{VMC}} \approx \frac{1}{N} \sum_{i=1}^{N} E_L(R_i). \quad (2.40) \]

The accumulation in Eq. 2.40 begins after an equilibration stage during which the
initially random positions of the walkers in the Metropolis algorithm are allowed to execute trial moves until they are correctly sampling the distribution \( P(R) \).

The local energy has the property that it is constant for an exact eigenstate of the Hamiltonian. In general, this is not the case for \( \tilde{\Psi}(R) \), offering a means of improving the trial wave function by minimising the spatially averaged variance of the local energy. In practice, the local energy is not calculated after every Monte Carlo move, as the energies at neighbouring points are correlated and its calculation very intensive, but rather once every correlation length \( c \), where \( c \) is the estimated number of Monte Carlo moves needed to find a new independent sample.

Other observables \( \hat{O} \) may be evaluated analogously to the energy, although it is not ensured that they behave strictly variationally, nor that the variance in their estimate tends to zero as \( \tilde{\Psi}(R) \rightarrow \Psi_0(R) \), unless the commutator \( [\hat{H}, \hat{O}] = 0 \).

The VMC method relies heavily on how well the trial wave function \( \tilde{\Psi}(R) \) approximates the exact ground-state solution \( \Psi_0(R) \). From the variational principle, it is evident that, as the accuracy of the trial wave function is improved, so are the estimates of the ground-state observables such as the energy. In fact, from Eq. 2.38, we see that an error \( \Delta \) in \( \tilde{\Psi}(R) \) yields an error of order \( \Delta^2 \) in the energy \( E_{\text{VMC}} \). This indicates the paramount importance of having an accurate trial wave function as a starting point for the Monte Carlo integration.

### 2.4.2 Diffusion Monte Carlo

Although the trial wave function constitutes an important ingredient to a diffusion Monte Carlo (DMC) calculation, in contrast with variational Monte Carlo, it does not limit the final accuracy of the result. In principle, DMC can project out the exact ground-state solution of the Schrödinger equation for any Hamiltonian. In practice, some important approximations hinder the exact solution for fermionic many-body systems; however, the errors associated with these approximations are typically only about 5% of the correlation energy and fairly well understood.

The starting point for DMC is the time-dependent many body Schrödinger equation written in imaginary time,

\[
-\frac{\partial}{\partial \tau} \Psi(R, \tau) = \left( \hat{H} - E_T \right) \Psi(R, \tau),
\]

(2.41)

where \( \tau = it \) and \( E_T \) is an arbitrary energy shift that does not affect any expectation.
values. The wave function $\Psi(\mathbf{R}, \tau)$ can be expanded in a complete set of stationary eigenstates with time dependence $\exp\left[-(E_i - E_T) \tau\right]$:

$$\Psi(\mathbf{R}, \tau) = \sum_i c_i \phi_i(\mathbf{R}) \exp\left[-(E_i - E_T) \tau\right], \quad (2.42)$$

where $\{E_i\}$ and $\phi_i(\mathbf{R})$ are the energy eigenvalues and eigenstates, respectively. By taking the limit $\tau \to \infty$ and setting $E_T$ equal to the ground-state energy $E_0$, we see that evolving the imaginary-time Schrödinger equation projects out its ground-state component:

$$\lim_{\tau \to \infty} \Psi(\mathbf{R}, \tau) = c_0 \phi_0(\mathbf{R}), \quad (2.43)$$

as all states with energy higher than $E_0$ exponentially decay away. This means that if we can simulate the imaginary time evolution of $\Psi(\mathbf{R}, \tau)$, we can project out its ground-state solution $\phi_0(\mathbf{R})$, and evaluate this to find the ground-state energy $E_0$. If we write the Hamiltonian in Eq. 2.41 in terms of the kinetic energy and a potential $V(\mathbf{R})$ that includes interactions between the particles and interactions with an external potential, we can identify the result as a diffusion equation in the real variable $\tau$, plus an additional growth/decay term with a position-dependent rate constant $E_T - V(\mathbf{R})$:

$$-\frac{\partial}{\partial \tau} \Psi(\mathbf{R}, \tau) = \left(-\frac{1}{2} \nabla^2 + V(\mathbf{R}) - E_T\right) \Psi(\mathbf{R}, \tau). \quad (2.44)$$

In this form, $\Psi(\mathbf{R}, \tau)$ represents the density of diffusing walkers in configuration space, analogous to a density of diffusing particles in real space. Each walker is represented by a $3N$-dimensional position vector $\mathbf{R}$ that contains the configuration of all $N$ electrons in the system. When evolved along a random walk, the walkers diffuse through configuration space and can be used to sample the wave function. This dynamics may be clarified by recasting the imaginary-time Schrödinger equation into an integral form. The equivalent integral expression to Eq. 2.41 is

$$\Psi(\mathbf{R}, \tau + \Delta\tau) = \int G(\mathbf{R} \leftarrow \mathbf{R}', \Delta\tau) \Psi(\mathbf{R}', \tau) d\mathbf{R}', \quad (2.45)$$

where $G(\mathbf{R} \leftarrow \mathbf{R}', \Delta\tau)$ is a Green’s function depending only on the time step $\Delta\tau$, and can be interpreted as a weighted transition probability density for the evolution of the walkers during $\Delta\tau$. As
repeatedly applying the Green’s function to a wave function \( \Psi(R, \tau) \) propagates it in imaginary time and, in the limit \( \tau \to \infty \), projects out its ground-state component. This reduces the problem of solving for the ground-state to finding the Green’s function \( G(R \leftarrow R', \Delta \tau) \) that satisfies Eq. 2.45. The exact Green’s function for many interacting particles is not known for most systems, forcing us to introduce approximations. For a small time step \( \Delta \tau \), the Trotter formula for the exponential of a sum of operators allows us to separate the kinetic and potential energy terms in the Green’s function:

\[
G(R \leftarrow R', \Delta \tau) = \left\langle R \left| \exp \left[ -\Delta \tau \left( \hat{H} - E_T \right) \right] \right| R' \rightangle,
\]

The kinetic energy term is the Green’s function of the diffusion equation, which is a Gaussian of variance \( \sqrt{\Delta \tau} \). Substituting this into Eq. 2.47, the Green’s function for the full Hamiltonian (to order \( (\Delta \tau)^3 \)) becomes

\[
G(R \leftarrow R', \Delta \tau) = (2\pi \Delta \tau)^{-3N/2} \exp \left[ -\frac{(R - R')^2}{2\Delta \tau} \right] \times \exp \left[ -\Delta \tau (V(R) + V(R') - 2E_T) / 2 \right].
\]
that solves the diffusion equation resulting from the kinetic energy term, whereas the reweighting of the Green’s function is a result of the potential energy term.

**Branching**

As walkers diffuse through configuration space, they gain weight in regions of low potential energy and lose weight in regions of high potential energy. If a walker has spent time $\Delta \tau$ at position $\mathbf{R}$, its weight is multiplied by the factor $\exp[-\Delta \tau V(\mathbf{R})]$, meaning that a walker situated in a favourable region of configuration space can quickly amass more weight than all other walkers.

To avoid the inefficiency this causes, a *branching* process can be used to vary the number of walkers (typically 10-1000) as they sample different areas of the potential energy landscape. In this algorithm, the reweighting factor $P$ in the Green’s function,

$$P = \exp[-\Delta \tau (V(\mathbf{R}) + V'(\mathbf{R}) - 2E_T)/2]$$  \hspace{1cm} (2.49)

determines the number of walkers that survive to the next time step from the rules

1. If $P < 1$, the walker continues with probability $P$.

2. If $P \geq 1$, the walker continues and spawns, with probability $P - 1$, another walker at the same position.

Enforcing these rules evolves

$$M_{\text{new}} = \text{INT} \left( P + \eta \right)$$  \hspace{1cm} (2.50)

walkers to the next time step at a given position $\mathbf{R}$, where $\eta$ is a random number between 0 and 1 and INT denotes the integer part of a real number. This causes walkers to disappear in regions of high potential energy and proliferate in regions of low potential energy. However, if left alone, this algorithm would result in the number of walkers either shrinking or growing rapidly. The energy offset $E_T$ affects the overall population of walkers, and must be adjusted periodically to keep the walker population roughly constant.
Importance sampling

Using random walks with branching to solve Eq. 2.45 is inefficient, as the walkers will diffuse into regions where the amplitude of the wave function is low, thereby sampling points that contribute only small amounts to the energy. In addition, the branching is controlled by $V(R)$, which diverges when two particles approach each other, causing the walker population to fluctuate strongly and increase the statistical variance. Both of these difficulties can be overcome by using an importance sampling technique similar to that in VMC and multiplying the exact ground-state wave function $\Psi_0(R, \tau)$ with a trial wave function $\tilde{\Psi}(R)$,

$$g(R, \tau) = \Psi_0(R, \tau)\tilde{\Psi}(R). \quad (2.51)$$

If the trial wave function $\tilde{\Psi}(R)$ is a reasonable approximation to the ground-state, it acts as a bias, guiding the walkers towards regions where the amplitude of the wave function is high. If we now express the imaginary-time Schrödinger equation from Eq. 2.44 in terms of the modified probability distribution $g(R, \tau)$, we obtain the drift-diffusion-reweighting equation:

$$-\frac{\partial g(R, \tau)}{\partial \tau} = -\frac{1}{2} \nabla^2 g(R, \tau) + [E_L(R) - E_T] g(R, \tau) + \nabla \cdot [v_D(R)g(R, \tau)], \quad (2.52)$$

where $E_L(R) = \frac{\hat{H}\tilde{\Psi}(R)}{\tilde{\Psi}(R)}$ is the local energy, and

$$v_D(R) = \frac{1}{2} \nabla \ln \left( |\tilde{\Psi}(R)|^2 \right) \quad (2.53)$$

is the $3N$-dimensional drift velocity that steers the density of walkers towards regions of large $|\tilde{\Psi}(R)|^2$.

The integral equation equivalent to Eq. 2.52,

$$g(R, \tau + \Delta \tau) = \int G^*(R \leftarrow R', \Delta \tau)g(R', \tau)dR', \quad (2.54)$$

contains a modified Green’s function $G^*(R \leftarrow R', \Delta \tau) = (\tilde{\Psi}(R)/\tilde{\Psi}(R'))G(R \leftarrow R', \Delta \tau)$. Adopting a short time-step approximation as for the unmodified case yields
\[ G^*(R \leftarrow R', \Delta \tau) = (2\pi \Delta \tau)^{-3N/2} \]
\[ \times \exp \left[ -\frac{(R - R' - \Delta \tau \nabla \ln (|\tilde{\Psi}(R')|^2))^2}{2\Delta \tau} \right] \]
\[ \times \exp \left[ -\Delta \tau (E_L(R) + E_L(R') - 2E_T) / 2 \right]. \] (2.55)

The branching rate is now controlled by the local energy \( E_L(R) \), rather than \( V(R) \), which is smooth and slow-varying for a suitable choice of trial wave function \( \tilde{\Psi}(R) \), stabilising the walker population.

In the drift-diffusion process defined by Eq. 2.52, we are no longer sampling the ground-state wave function but rather the distribution \( g(R, \tau) = \Psi_0(R, \tau)\tilde{\Psi}(R) \). However, it remains possible to estimate expectation values of operators that commute with the Hamiltonian. For example, the mixed estimator of the energy,

\[ E_{DMC} = \int \left( \frac{\hat{H}\tilde{\Psi}(R)}{\tilde{\Psi}(R)} \right) \left( \frac{\tilde{\Psi}(R)\Psi_0(R, \tau)}{\int \tilde{\Psi}(R)\Psi_0(R, \tau)dR} \right) dR, \] (2.56)

obtained by averaging the local energies at the points on the random walk, gives the ground-state energy.

DMC is less sensitive to the quality of the trial wave function than VMC, and also provides more accurate estimates of ground-state energies. The trial wave function is only used as a starting point for the simulation, and the drift-diffusion process allows the walkers to some extent to correct the distribution \( |\tilde{\Psi}(R)|^2 \) where it differs significantly from \( |\Psi_0(R)|^2 \).

The scheme outlined here can be further tweaked to improve efficiency and minimise statistical variances. For example, incorporating the Metropolis algorithm to accept or reject proposed trial moves from \( R \) to \( R' \) reduces the time step error.

**The fixed node approximation**

In DMC, the imaginary-time Schrödinger equation is interpreted as a drift-diffusion process of a population of walkers. The walker density represents a probability density associated with the wave function. However, while probability densities are everywhere positive by definition, a fermionic wave function must be antisymmetric
under exchange of electrons, leading to both positive and negative regions. This is one example of the infamous ‘fermion sign problem’ that permeates fermionic Monte Carlo methods. A wave function evolved in unconstrained DMC will reach the bosonic ground state, which is symmetric under exchange of particles, unlike the antisymmetric fermionic ground state.

The most obvious solution to this problem is to start with an antisymmetric wave function represented by a population of signed walkers, positive in positive regions and negative in negative regions. The imaginary-time evolution preserves the antisymmetry and so should yield the many-fermion ground state, which would be represented by the difference of the positive and negative weighted densities. Unfortunately, in this method, the positive and negative populations collapse individually to the bosonic ground-state, causing the relative difference in their densities to tend to zero, leading to exponentially growing statistical noise.

The ‘fixed node approximation’ [5] is now the most widely used technique to circumvent this problem. The trial wave function $\tilde{\Psi}(\mathbf{R})$, usually given by a Hartree-Fock or DFT calculation, defines a nodal surface on which $\tilde{\Psi}(\mathbf{R})$ is zero and across which it changes sign. Due to the approximate Green’s function in Eq. 2.55, walkers occasionally attempt to cross these nodes. Importance sampling helps to minimise these crossings, as the drift velocity grows when a walker approaches a node, pushing it away, however it cannot eliminate the effect. In the fixed node approximation, walkers are not allowed to cross the nodes of the trial wave function, which constrains the exact wave function $\Psi_0(\mathbf{R}, \tau)$ to have the same nodes as the trial wave function $\tilde{\Psi}(\mathbf{R})$. This can easily be implemented by either deleting any walker that attempts to cross a node, or, when a Metropolis algorithm is used, rejecting any trial moves that cause a walker to cross a node. It can be shown that deleting walkers leads to a bias proportional to the time step $\Delta\tau$ [180], whereas rejecting trial moves only leads to an $O[\Delta\tau^2]$ bias. The nodes then act as infinite potential barriers that divide the configuration space into disjoint nodal ‘pockets’. This confines walkers to the nodal pockets in which they originate, allowing Eq. 2.52 to be solved individually for each pocket. By the ‘tiling theorem’ [32, 149], all nodal pockets of the ground-state of a Hamiltonian are equivalent by permutation symmetry, thereby giving the same energy. Enforcing the fixed node approximation in DMC then projects out the ground-state energy for a particular nodal surface.

As the approximate nodal surface is determined from a variational calculation such as Hartree-Fock or DFT, fixed node DMC energies are variational, as well. The
DMC energy is greater than or equal to the exact ground-state energy of a given nodal surface. If the trial nodal surface were exact, DMC would give the exact ground-state energy. Unlike VMC, no assumptions are made about the functional form of the wave function between the nodes. Approximating the nodes as fixed leads to an error of ca. 5% in the correlation energy in the system.

### 2.4.3 Trial wave functions

Choosing a suitable trial wave function $\tilde{\Psi}(\mathbf{R})$ is the most important ingredient of a VMC or DMC calculation. A major advantage of Monte Carlo integration is that any computable form of wave function can be used, although care must be taken as the choice of form determines the accuracy and efficiency of a QMC simulation. In VMC, averages taken over $|\tilde{\Psi}(\mathbf{R})|^2$ yield expectation values of operators, whereas in DMC the quality of the nodal surface controls the final accuracy. In both cases, the trial wave function controls the variance of the local energy, and therefore the efficiency of a calculation. In addition, evaluating $\tilde{\Psi}(\mathbf{R})$ and its derivatives is extremely computationally demanding. These considerations demonstrate the importance of having an accurate, yet simple form for $\tilde{\Psi}(\mathbf{R})$.

In quantum chemistry, many-body wave functions are often expressed as linear combinations of determinants. These give accurate representations of the symmetries of true wave functions, but such expansions converge slowly, partly due to the difficulty of describing the cusps that arise from coalescent electrons. The computational cost of using the huge linear combinations of determinants required to simulate solids of significant size has thus far impeded their use.

It has been found that simple single-determinant wave functions of the Slater-Jastrow type [89] are very accurate for solids. In this form, a Slater determinant $D(\mathbf{X})$ is multiplied by a symmetric, non-negative Jastrow correlation factor $\exp [J(\mathbf{X})]$:

$$
\Psi(\mathbf{X}) = D(\mathbf{X}) \exp [J(\mathbf{X})],
$$

(2.57)

where $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, and $\mathbf{x}_i = \{\mathbf{r}_i, s_i\}$ denotes the space and spin coordinates of electron $i$. For computational convenience, the spin variables are normally removed from the Slater-Jastrow wave function by replacing the single, large Slater determinant with two smaller determinants.
\[ \Psi(\mathbf{R}) = D(\mathbf{r}_1, \ldots, \mathbf{r}_N)D(\mathbf{r}_{N+1}, \ldots, \mathbf{r}_N) \exp \left[ J(\mathbf{X}) \right]. \tag{2.58} \]

This function is no longer antisymmetric upon exchange of electrons, but it can be shown to give the same expectation value as the original form in Eq. 2.57 for any spin-independent operator \[63\].

In addition, the representation of the single-particle orbitals is important for the efficiency of a calculation. For the sake of computational efficiency in large systems, it is best to expand the orbitals in a basis of functions localised in real space. This makes Gaussians, which can be evaluated rapidly, favourable. However, recent work has shown how to convert a plane-wave basis set into a set of localised ‘blip’ functions \[2\], which is also computationally efficient and aids convergence analysis.

The Jastrow factor introduces short-ranged correlation by reducing the magnitude of the many-electron wave function when charged particles approach one another. Ideally, it would account for \(N\)-body correlations; however, correlations involving more than three particles have little impact on the variational energy \[85, 180\] and are usually discarded.

For the calculations performed in this study, the Jastrow factor retains one- and two-body correlations, such that

\[ J(\mathbf{X}) = \sum_{i=1}^{N} \chi(x_i) - \frac{1}{2} \sum_{i=1}^{N} \sum_{j \neq i}^{N} u(x_i, x_j), \tag{2.59} \]

where the \(u\) terms describe electron-electron correlations, and the \(\chi\) terms electron-nuclear correlation. The correlation between pairs of electrons depends on their separation, relative spin, and the likelihood of finding other electrons between them.

The two-body \(u\) term reduces the amplitude of the wave function when the separation between two electrons becomes small, thereby reducing the probability of finding two electrons close together and decreasing the electron interaction energy. This is necessary because the antisymmetry of the wave function only explicitly keeps parallel spin electrons apart. Antiparallel spin electrons are allowed to approach too near to one another, raising the electron-electron Coulomb repulsion energy. Therefore, the antiparallel-spin \(u\) function is more repulsive than the parallel-spin one. In non-uniform systems, introducing \(u\) terms spreads the electron density from high density (e.g. bonding) to low density (e.g. interstitial) regions. Unfortunately, the charge densities of Slater determinants from Hartree-Fock or DFT tend to be quite
accurate, and so the cumulative effect of incorporating $u$ terms and the resulting change in charge density only marginally lowers the total energy. To counteract this unfavourable change in density, it is necessary to either reoptimise the orbitals (which is computationally costly), or to introduce one-body $\chi$ terms. When optimised variationally together with the $u$ terms, the $\chi$ terms return the density back to its original distribution. Although the $\chi$ terms could be absorbed into the Slater determinant, they are normally retained in the Jastrow factor for conceptual simplicity.

The $u$ terms must obey the cusp conditions [92]:

$$\left. \frac{du}{dr} \right|_{r=0} = \begin{cases} \frac{1}{2} & \text{antiparallel spins} \\ -\frac{1}{4} & \text{parallel spins} \end{cases}$$

(2.60)

that introduce electron-electron cusps into the wave function. These non-analytic cusps arise whenever charged particles approach each other. At these singularities, the potential and kinetic energies diverge with opposite signs. The electron-nuclear cusps are already included in the Hartree-Fock or DFT orbitals, but the electron-electron cusps are not. Their incorporation has been shown to reduce the average energy and its variance. Equally important, violating the cusp conditions leads to unwelcome singularities in the local energy.

The cusps introduce spin dependence into the Jastrow factor. It can then occur that a linear combination of Slater-Jastrow wave functions is not an exact eigenfunction of $\hat{S}^2$ even though the corresponding sum of Slater determinants is. Fortunately, the effect of this ‘spin contamination’ is usually small and can be ignored [84].

Slater-Jastrow wave functions have met widespread success in both atomic and solid-state calculations. The combined optimized terms in the Jastrow function lower the variational energy significantly, routinely recapturing ca. 80% of the correlation energy. When used together with pseudopotentials, this represents a significant amount of the error in the total energy from a Hartree-Fock or DFT calculation that is necessary for understanding bonding and other effects.

### 2.4.4 Wave function optimisation

Given a parametrised trial wave function $\tilde{\Psi}(\mathbf{R})$, it is now necessary to optimise the parameters to make $\tilde{\Psi}(\mathbf{R})$ represent the ground-state wave function $\Psi_0$ as closely as possible. However, it must be emphasised that, no matter how many variational
parameters are contained in $\tilde{\Psi}(R)$, there is still no guarantee that the assumed functional form is accurate.

For Slater-Jastrow trial wave functions, an ideal optimisation would involve both the Slater determinant $D(X)$ and the Jastrow factor $\exp[J(X)]$ simultaneously. However, the orbitals from Hartree-Fock or DFT tend to be quite accurate, and the computational cost for improving these for a large system has so far limited simultaneous optimisation of $D(X)$ and $\exp[J(X)]$ to atoms and molecules. For solid-state systems such as those considered in this study, it is common to optimise only the $u$ and $\chi$ functions of the Jastrow factor.

The $u$ and $\chi$ terms in the Jastrow factor contain a set of parameters $\{\alpha\}$ that can be optimised according to different criteria, each represented by a cost function. The choice of cost function, and thus optimisation procedure, depends on the type of calculation being performed. Minimising the cost function with respect to the parameters yields an optimised trial wave function.

It is most common to minimise the variance (and thus statistical error) of the VMC energy [181, 93]

$$\sigma^2_E(\alpha) = \frac{\int \tilde{\Psi}^2(\alpha)(E_L(\alpha) - E_{VMC}(\alpha))^2 dR}{\int \tilde{\Psi}^2(\alpha) dR},$$

which conveniently has a minimum value of zero, obtained when the trial wave function $\tilde{\Psi}(\alpha)$ is an exact eigenstate of $\hat{H}$. This allows for a simple extension of optimisation to excited states. When evaluating the variance $\sigma^2_E(\alpha)$, the minimisation begins from a set of configurations distributed according to $\tilde{\Psi}^2(\alpha_0)$, where $\{\alpha_0\}$ is an initial guess of the parameters. The set of configurations is then kept fixed while the parameters $\{\alpha\}$ are adjusted until $\sigma^2_E(\alpha)$ is minimised. This procedure, referred to as correlated sampling, becomes inefficient when the difference $|\{\alpha_0\} - \{\alpha\}|$ becomes large. In order to avoid this problem, the set of configurations can be periodically regenerated when a certain tolerance on the variance is satisfied.

Although it has been suggested that the intuitive method of minimising the energy [176] maximises the efficiency of a DMC calculation [31], variance minimisation is far more common than energy minimisation, primarily due to its numerical stability for large systems (see Ref. [63] and references therein). However, recent computational improvements to energy minimisation implementations have narrowed this gap.


“Crystals are like people; it’s the defects in them that make them interesting.”

- Sir Frederick Charles Frank

3

Point defects in solids

Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1 Introduction</td>
<td>43</td>
</tr>
<tr>
<td>3.2 Defect species and notation</td>
<td>45</td>
</tr>
<tr>
<td>3.3 Defect concentrations in the dilute limit</td>
<td>47</td>
</tr>
<tr>
<td>3.3.1 The law of mass action</td>
<td>47</td>
</tr>
<tr>
<td>3.3.2 Defect concentrations in the canonical ensemble</td>
<td>48</td>
</tr>
<tr>
<td>3.3.3 Intrinsic and extrinsic defects</td>
<td>54</td>
</tr>
<tr>
<td>3.4 Thermodynamics of defect formation energies</td>
<td>57</td>
</tr>
<tr>
<td>3.4.1 Zhang-Northrup formalism</td>
<td>57</td>
</tr>
<tr>
<td>3.4.2 Chemical potentials</td>
<td>59</td>
</tr>
<tr>
<td>3.5 Density functional theory calculations of defect formation energies</td>
<td>65</td>
</tr>
<tr>
<td>3.5.1 Vibrational effects</td>
<td>65</td>
</tr>
<tr>
<td>3.5.2 Band gap corrections</td>
<td>68</td>
</tr>
<tr>
<td>3.5.3 Finite-size effects</td>
<td>71</td>
</tr>
<tr>
<td>3.6 Diffusion kinetics of defects</td>
<td>78</td>
</tr>
<tr>
<td>3.6.1 Mechanisms of diffusion</td>
<td>79</td>
</tr>
<tr>
<td>3.6.2 Activation energies and diffusion barriers</td>
<td>81</td>
</tr>
<tr>
<td>3.6.3 Fick’s laws</td>
<td>83</td>
</tr>
<tr>
<td>3.6.4 Diffusion coefficients from random walk diffusion</td>
<td>86</td>
</tr>
<tr>
<td>3.6.5 Correlation effects</td>
<td>88</td>
</tr>
<tr>
<td>3.6.6 Experimental methods</td>
<td>93</td>
</tr>
</tbody>
</table>

3.1 Introduction

For centuries, humans have exploited the advantageous properties that solids acquire when doped with impurities. In the Bronze Age, tools were made predominantly
from bronze, due to its low melting point and high strength. However, the components of bronze, copper and tin, were not readily available, in contrast with iron. The advent of higher temperature smelting enabled iron to be welded, and the ability to sharpen iron by simple grinding rendered it relatively cheap and practical. During the ensuing Iron Age, the development of alloys such as steel, iron doped with carbon, added strength and rigidity to the material. The twentieth century saw the mass production of stainless steel, a steel alloy containing chromium impurities, that does not stain or corrode as easily as ordinary steel. Today, further refinements in high carbon-content steel alloys are still being made to improve strength and durability.

In these alloys, impurities enhance specific properties of the material. Any solid at finite temperature, even a ‘pure’ one, contains an equilibrium concentration of point defects [169]. The controlled addition of certain impurities to the solid influences the concentrations of these and other defects. These changes, as well as the presence of the impurities themselves, alter many physical properties of the solid and allow materials to be tailored for certain industrial applications. For example, the optical properties of a material are affected by the presence of defects that introduce states with optical transitions. Mechanical properties can be heavily altered by the presence of defects; for example, a high carbon content in steel renders the metal harder and stronger, although less ductile and more difficult to weld. Most notably, perhaps, ionic defects have a pronounced effect on transport properties, such as conductivity and diffusion, enabling processes such as plastic deformation, sintering and creep. Attempts to understand how and why these changes take place have led to the creation of a mammoth body of work in many diverse fields, such as chemistry, engineering, materials science and solid state physics.

Despite considerable theoretical work on the energetics and kinetics of point defects in a wide range of solids, many puzzles remain unsolved. Early simulations of solids using classical inter-atomic potentials were very successful in discerning the geometric relaxations and kinetics of defect structures, but could not account for many important electronic structure effects such as correlation or polarisation. Density functional theory is an established method for calculating the electronic properties of bulk solids, and is now frequently used in a wide range of studies beyond the bulk, from examining the chemical behaviour of molecules at surfaces to the segregation of defects to grain boundaries; however, its limitations must be carefully considered. In situations where DFT has difficulties, such as strongly-correlated systems
and transition metal oxides, quantum Monte Carlo can provide a higher level of accuracy, albeit at considerable computational cost. Due to this increased cost and difficulty, QMC studies have been limited to very simple solids with defects [3, 106]. In this section, we will introduce the main concepts used in the study of point defects and introduce a thermodynamic mass action framework for determining their concentrations in the dilute limit. We then establish a reliable formalism for calculating the key ingredient to this approach, the formation energy of a defect. Finally, we examine the diffusion kinetics of point defects in a solid, presenting a methodology for determining diffusion coefficients within this framework.

### 3.2 Defect species and notation

At absolute zero, atoms in a solid move due to zero-point vibrations. As the temperature is raised, thermal vibrations increase and anharmonicity affects lattice spacing so that lattice defects are produced. As we will later see, lattice defects can also be created when a solid is doped with aliovalent impurities. In addition, electronic excitations are generated, which are normally delocalised in semiconductors, but can be localised at defect sites in insulators, for example when there is charge transfer on an atom that can readily change its valency. Eventually, when the temperature is high enough, phase changes can occur.

Point defects are local irregularities in a solid, the simplest being vacancies, interstitials and substitutionals (see Fig. 3.1). Vacancies are normally occupied lattice sites that become unoccupied, interstitials normally vacant sites that become occupied, and substitutionals impurities that replace host atoms in the lattice. The latter two can either be native to their lattice, or foreign impurities. Of course, these are only elementary examples, and diverse clusters of the above are also possible.

The creation of vacancies involves breaking chemical bonds, whereas interstitials form bonds with their neighbours. Changes in the bonding character of neighbouring atoms affects various local properties, such as bond lengths and the electron density surrounding a defect site. Many of these effects can be modelled using electronic structure methods in order to develop a deeper understanding of how the defect chemistry affects the properties of materials.

In labelling defects, there are several properties to be taken into account. A single defect can be of three types, vacancy, interstitial, or substitutional; it can occupy
Chapter 3: Point defects in solids

Figure 3.1: Various species of point defects (a. interstitial, b. substitutional, c. vacancy) in a schematic 2D lattice.

different types of sites within a crystal lattice; and it can carry a charge. The defect charge may consist of electrons or holes left over after removing a neutral atom from an ionic lattice, or may be created by adding or removing electrons at the site where the defect has been created.

Many different methods have been used to label these properties for specific defects [95], some intuitive, some counter-intuitive, and many confusing. We shall adopt the notation proposed by Kröger and Vink in 1953 [97], which has found wide acceptance.

In Kröger-Vink notation, defects are described by three labels, as seen in this example:

\[ \text{Ti}_i^- \]  

(3.1)

1. **The symbol**: the defect species, indicated by the atomic symbol for interstitials and substitutionals, or by \( V \) in the case of a vacancy.

2. **The subscript**: the site which the defect occupies, indicated by the atomic symbol for substitutionals and vacancies, or by \( i \) in the case of an interstitial.

3. **The superscript**: the ‘charge state’, or difference in charge at the defect site relative to the charge at the same site in the perfect crystal. Indicated by a dot (\( \cdot \)) for an extra positive charge or a slash (\( \slash \)) for an extra negative charge. If the charge at the defect site is the same as in the perfect crystal, we use a superscript \( \times \). The symbolic representation of the charge state can also be replaced with numbers, e.g. \( n = -2 \).
For the example in Eq. 3.1, we have a titanium interstitial carrying a +2 charge in comparison with the charge (in this case zero) expected at that site (i.e. 2 electrons have been added at the defect site in addition to a +4 charged titanium nucleus). The charge may be fully localised on the defect site, or less localised over nearest neighbours. The localised electrons/holes can give rise to defects such as $F$ centres (where electrons are trapped in an anion vacancy) or $V$ centres (holes trapped in cation vacancies) [77]. The charge is important for a number of reasons. For example, because the perfect crystal is electrically neutral, a defect will react to an electric field according to its effective defect charge.

### 3.3 Defect concentrations in the dilute limit

Determining the defect chemistry of a solid that contains a raft of impurities and defects involves writing equilibrium reactions for the formation of all sensible defects and the interactions between them. The possible reactions are almost endless: for example, the Coulomb interaction between charged impurities and defects can cause them to bind into clusters; the addition of extra impurities can give rise to charge compensating defects; and raising the temperature can create more intrinsic defects. The law of mass action relates the concentrations of this medley of defects and clusters to one other by a set of equilibrium constants that depend only on the temperature and pressure. The equilibrium constants can be determined by calculating the Gibbs energy of formation for every possible defect in the system.

#### 3.3.1 The law of mass action

We will begin by introducing the law of mass action within the formalism used in this study, and then explain how accurate calculations of Gibbs free energies can be performed using the DFT or QMC methods discussed previously. For the solids considered, we neglect surface effects, determining the defect chemistry in the bulk of the solid, which may differ greatly from that at the surface. We assume that we are working in the dilute limit, such that all defects are independent and there are no interactions between them beyond those explicitly accounted for. For wide band gap insulators this assumption is sensible, as the concentrations of point defects are generally very low. The dilute approximation fails when the concentration of impurities is very high (for example at high temperatures near the melting point of
a crystal), the compound is nonstoichiometric, or under intense irradiation [39].

For a general reaction in dilute solution,

\[ n_1X_1 + n_2X_2 + \ldots + n_iX_i \rightleftharpoons m_1Y_1 + m_2Y_2 + \ldots + m_iY_i \]  \hspace{1cm} (3.2)

of \( n_i \) reactant species \( \{X_i\} \) and \( m_i \) products \( \{Y_i\} \), the law of mass action asserts that, in equilibrium, the concentrations \( \{[X_i]\} \) and \( \{[Y_i]\} \) are related by

\[ \frac{[X_1]^{n_1} [X_2]^{n_2} \ldots [X_i]^{n_i}}{[Y_1]^{m_1} [Y_2]^{m_2} \ldots [Y_i]^{m_i}} = K(T, p). \]  \hspace{1cm} (3.3)

The equilibrium constant \( K(T, p) \) depends only on the temperature \( T \) and pressure \( p \). In a crystal, where many reactions are occurring simultaneously in thermodynamic equilibrium, all mass action expressions and conservation rules must be obeyed simultaneously. We will now discuss this point in more detail in order to derive a formalism that respects these rules and delivers the concentrations of defects as a function of temperature, pressure, and impurity doping only.

### 3.3.2 Defect concentrations in the canonical ensemble

In an ideal crystal at \( T = 0K \) and constant \( p \), the ground state is found by minimising the total enthalpy \( H \) for the particular assembly of atoms. In most cases, the structure is already given by experiment, and finding the minimum enthalpy in DFT or QMC using the experimental structure as an initial guess then gives the ‘theoretical’ geometry according to the specific set of approximations that have been made (i.e. DFT functional, \( k \)-point sampling, basis set, etc.). At finite \( T \), the equilibrium state is not determined by minimising the enthalpy \( H \), but by minimising the Gibbs free energy

\[ G = H - TS, \]  \hspace{1cm} (3.4)

where \( S \) is the entropy. The defects introduced as \( T \) rises strain the lattice of the perfect crystal, moving neighbouring atoms from their initial positions and expanding the crystal; this process has an enthalpic cost, which can be balanced by entropic profits.

The entropy \( S \) comprises three parts: a configurational term \( S_{\text{conf}} \), a vibrational term \( S_{\text{vib}} \), and the entropy due to electronic excitation. The latter is assumed to be
Section 3.3: Defect concentrations in the dilute limit

negligible and ignored. Configurational entropy measures the number of permutations in which the components of a system can be arranged. When comparing the difference between the Gibbs energy of a perfect crystal and one containing a defect, vibrational entropy arises from changes in the vibrational modes surrounding the defect. In general, creating a vacancy involves removing modes, while creating an interstitial adds modes to the system. In addition, the atoms around a vacant lattice site can vibrate over a larger amplitude (called a ‘softening’ of the phonon modes), whereas the converse holds true in the vicinity of interstitials. Which of these two effects will dominate is difficult to anticipate, and the exact determination of the vibrational modes in a system is a cumbersome process in many body methods such as DFT. However, it can be crucial for an accurate description of the defect energetics, contributing 2eV to the formation energy of self-vacancies in silicon [1]. Finally, when comparing the vibrational entropies of crystals with and without defects, the vibrational entropy of the reference state (e.g. O₂ gas) from/to which any additional atoms were taken/given must be considered.

We now present a formalism for determining the equilibrium Gibbs free energy of a crystal containing point defects. Consider a very large crystal comprising fixed numbers $N^\alpha$ of atoms of various elements $\alpha$ distributed over $N$ identical unit cells. Each unit cell is allowed to contain one of $I$ different species of point defect $i$. Each species $i$ can be either a single defect (whereby a divacancy, for example, counts as a single defect), dopant, defect-dopant cluster, or also perfect crystal, denoted by $i = 0$. Then, we define $n^\alpha_i$ as the number of atoms of element $\alpha$ located in a unit cell when it has a defect of type $i$. A charge $q_i$ is also associated with each defect species. For example, in an ionic model of Al₂O₃, native aluminium and oxygen vacancies carry charges of -3 ($V''_{\text{Al}}$) and +2 ($V'_{\text{O}}$), respectively. Consequently, an AlO divacancy would carry a charge of -1 ($V'_{\text{AlO}}$). However, in the current formalism, defects are allowed to carry any of a number of charges, so that a $V'_{\text{AlO}}$ defect might exist as $V''_{\text{AlO}}$, $V'_{\text{AlO}}$, or $V'_{\text{AlO}}$. Each charge state of the AlO divacancy counts as an individual defect species altogether. Also, an aluminium vacancy ($V_{\text{Al}}$) and an oxygen vacancy ($V_{\text{O}}$) that are nearest-neighbours but lie in neighbouring unit cells would count as a two separate defects, rather than one $V_{\text{AlO}}$ divacancy. In the limit of dilute defects and large unit cells, this should not pose a problem. The choice of unit cell must be large enough to contain the atoms and vacancies that constitute the defect, but the elastic and Coulomb fields arising from the formation of the defect may extend far outside. In addition, each of the $n_i$ unit cells occupied by point defect species $i$
Chapter 3: Point defects in solids

has a multiplicity $m_i$ of different sites and orientations which the point defect may assume. For example, an oxygen vacancy in a five-atom unit cell of alumina has a multiplicity of two sites.

The configurational entropy of the crystal is then given from the ideal solution model by

$$S_{\text{conf}} = k_B \ln \frac{N! \prod_{i=0}^{I} m_i^{n_i}}{\prod_{i=0}^{I} n_i!}.$$  
(3.5)

Using Stirling's approximation, this simplifies to

$$\frac{S_{\text{conf}}}{N} = -k_B \sum_{i=0}^{I} c_i \ln \left( \frac{c_i}{m_i} \right),$$  
(3.6)

where $c_i$ is the concentration of defect species $i$, such that $c_i N = n_i$. The term $(c_i/m_i)$ represents the concentration of defect $i$ per possible site, including orientational degeneracy.

In the canonical ensemble, the total Gibbs free energy $G$ of the system at constant pressure and temperature, neglecting thermal expansion, can then be written as

$$G = \sum_{i=0}^{I} g_i n_i - T S_{\text{conf}}$$

$$= N \left\{ \sum_{i=0}^{I} g_i c_i + k_B T \sum_{i=0}^{I} c_i \ln \left( \frac{c_i}{m_i} \right) \right\},$$  
(3.7)

where $g_i$ is the contribution to $G$ from a cell of type $i$, including its long-ranged elastic and Coulomb fields and vibrational effects. By including not only the $g_i$ from different native point defects, but also those from clusters binding these defects to substitutional impurities, we can explicitly account for the interactions between defects and impurities. We now have $I + 1$ unknowns, namely the $\{c_i\}$ and $N$. If there are $E$ different elements $\alpha$ present in the crystal, there are $E + 2$ constraints, namely

$$\sum_{i=0}^{I} c_i n_i^\alpha N = N^\alpha \quad \alpha = 1, 2, \ldots, E,$$  
(3.8)
Section 3.3: Defect concentrations in the dilute limit

\[ \sum_{i=0}^{I} c_i q_i N = Q N = 0, \quad (3.9) \]

\[ \sum_{i=0}^{I} c_i = 1. \quad (3.10) \]

The first constraint describes conservation of atom number and the second charge neutrality. In Eq. 3.9, \( q_i \) is the charge of each defect species (remember that we define different charge states of the same defect species as different defects altogether), and \( Q \) is the total charge of all defects, which we insist must be zero. For a pure crystal at finite temperature containing no dopants, there will be a number of intrinsic defects, whose charges must balance exactly. When we introduce dopants to the crystal, these can exist either as substitutional or interstitial defects. As interstitials it is almost certain, and as substitutionals likely, that they will introduce charges to the system that must also be balanced. This can be achieved by creating charge-compensating defects, which may either be free and mobile, or bound to the dopant atoms in clusters.

Associating \( E + 2 \) Lagrange multipliers \( \{ \lambda^\alpha \} \), \( \lambda^q \), and \( \lambda^c \) with these constraints and minimizing \( G \) with respect to \( N \) and the \( \{ c_i \} \) subject to the constraints gives \( I + 1 \) equations for \( I + 1 \) unknowns:

\[ \sum_{i=0}^{I} \left( g_i - \sum_{\alpha=1}^{E} \lambda^\alpha n_i^\alpha - \lambda^q q_i \right) c_i + k_B T \sum_{i=0}^{I} c_i \ln (c_i / m_i) = 0, \quad (3.11) \]

\[ g_i + k_B T \ln (c_i / m_i) + k_B T - \sum_{\alpha=1}^{E} \lambda^\alpha n_i^\alpha - \lambda^q q_i - \lambda^c = 0. \quad (3.12) \]

If we sum \( c_i \times \text{Eq. 3.12} \) over \( i \) and subtract \( \text{Eq. 3.11} \) from the result, we get

\[ k_B T - \lambda^c = 0. \quad (3.13) \]

Substituting \( \text{Eq. 3.13} \) into \( \text{Eq. 3.12} \) gives

\[ g_i + k_B T \ln (c_i / m_i) - \sum_{\alpha=1}^{E} \lambda^\alpha n_i^\alpha - \lambda^q q_i = 0. \quad (3.14) \]

If we substitute \( \text{Eq. 3.14} \) into \( \text{Eq. 3.7} \), we get
Chapter 3: Point defects in solids

\[ G = N \left\{ \sum_{i=0}^{I} \sum_{\alpha=1}^{E} \lambda^\alpha n_i^\alpha c_i + \sum_{i=0}^{I} \lambda^q q_i c_i \right\}, \]  
\hspace{1cm} (3.15)

Substituting the constraints \( N^\alpha = \sum_{i=0}^{I} c_i n_i^\alpha N \) and \( QN = \sum_{i=0}^{I} c_i q_i N \) into Eq. 3.15 then gives

\[ G = \sum_{\alpha=1}^{E} \lambda^\alpha N^\alpha + \lambda^q QN, \]  
\hspace{1cm} (3.16)

where we have not yet set \( QN = 0 \) so that we can identify the effect of a small change in the total charge (for example upon adding a dopant or creating a defect).

In equilibrium and at constant temperature, pressure and chemical potentials,

\[ \frac{\partial G}{\partial N^\alpha} = \lambda^\alpha = \mu^\alpha, \]  
\hspace{1cm} (3.17)

where \( \mu^\alpha \) is the chemical potential of element \( \alpha \) throughout the crystal.

Under the same conditions, taking the partial derivative with respect to \( QN \) gives

\[ \frac{\partial G}{\partial (QN)} = \lambda^q. \]  
\hspace{1cm} (3.18)

A positive change in \( Q \) will correspond to removing electrons, a negative change in \( Q \) to adding electrons, and therefore, we can identify

\[ \lambda^q = -\mu_e, \]  
\hspace{1cm} (3.19)

where \( \mu_e \) is the electron chemical potential. Defects can exist in more than one charge state, and, in thermal equilibrium, adding dopants can move the electron chemical potential, controlling the charge state principally present. The electron chemical potential is allowed to adjust itself in the same way as in a semiconductor, where doping by donors contributes electrons to the conduction band, raising \( \mu_e \), while doping by acceptors contributes holes to the valence band, lowering \( \mu_e \). By treating defects in different charge states as different defects altogether, and allowing the electron chemical potential to vary freely, this formalism allows defects to readily change their charge state according to the environment they are placed in.

Substituting the results from Eqs. 3.17 and 3.19 back into Eq. 3.14 then gives us an expression for the concentration of defect species \( i \) in equilibrium:
\[ c_i = m_i \exp \left( -\frac{g_i - \sum_{\alpha=1}^{E} \mu_{\alpha} n_{i\alpha} + \mu_e q_i}{k_B T} \right) = m_i \exp \left( -\frac{\Delta g_i^f}{k_B T} \right). \] (3.20)

The crucial \( \Delta g_i^f \) term is known as the ‘formation energy of defect \( i \) in the charge state \( q_i \)’, and its determination will be discussed at length in Sec. 3.4. We now have \( E + 2 \) unknowns, namely the \( \{\mu_{\alpha}\} \), \( \mu_e \) (together which can be used to find the \( \{c_i\} \)), and \( N \), and still \( E + 2 \) constraints from Eqs. 3.8, 3.9, and 3.10.

**Grand canonical ensemble**

In order to find a self-consistent solution, and to make direct contact with experiment, we switch to the grand canonical ensemble in order to choose consistent values for the chemical potentials \( \{\mu_{\alpha}\} \) equivalent to particular formation conditions of the crystal. In our case, these are the oxygen partial pressure \( p_{O_2} \), temperature \( T \), and dopant concentration. The \( \{\mu_{\alpha}\} \) are defined by the reservoirs with which the crystal is in contact and must be consistent with each other in two respects. First, any phases and reservoirs present in the system must be in equilibrium with respect to chemical reactions, and second, the chemical potentials must be referred to the same reference state. In a metal oxide, this can be achieved by assuming the oxide is in equilibrium with an oxygen atmosphere of partial pressure \( p_{O_2} \) and temperature \( T \). A chemical potential for oxygen \( \mu_O \) can then be selected equivalent to these formation conditions, and the chemical potential of the metal can be found directly from the equilibrium condition of the solid. In practice, \( \mu_O \) is determined by invoking a thermodynamic cycle corresponding to the formation of the oxide, thereby accurately combining zero-temperature DFT total energy calculations of the metal oxide and metal with thermodynamic relationships, as detailed in Sec. 3.4.2. The chemical potentials of any dopants present in the system are adjusted iteratively in order to give a specified target concentration of dopant atoms, whether free or contained in clusters, providing a mean-field-like coupling between the defect concentrations. Given a choice for the chemical potentials \( \{\mu_{\alpha}\} \), it is then straightforward to solve for the concentrations \( \{c_i\} \) of all defect species by adjusting \( \mu_e \) iteratively until charge neutrality is obtained to a specified tolerance. This gives us the concentrations of the defects (and thereby the stoichiometry) and the electron chemical potential as a function of \( p_{O_2} \), \( T \), and dopant concentration. Each defect species \( c_i \), which can
Chapter 3: Point defects in solids

be either a single, free-standing defect or a cluster, has a specified charge state \( q_i \).
The mass action relationship from Eq. 3.3 can be recovered from our formalism. Any
reaction of the kind in Eq. 3.2 can be expressed as

\[
\sum_{i=0}^{I} r_i D_i = 0, \tag{3.21}
\]

where \( D_i \) is the name of the defect species participating in the reaction and \( r_i \) its
coefficient. The reaction must satisfy

\[
\sum_{i=0}^{I} r_i q_i = 0 \tag{3.22}
\]
to conserve charge, and, for each element \( \alpha \),

\[
\sum_{i=0}^{I} r_i n_{i}^{\alpha} = 0 \tag{3.23}
\]
to conserve the number of atoms.
Using our expression for the concentration of defect \( i \) per possible site (including
orientational degeneracy), \( c_i/m_i \), where \( c_i \) is defined as in Eq. 3.20, we can recover
the law of mass action from Eq. 3.3:

\[
\prod_{i=0}^{I} (c_i/m_i)^{r_i} = \exp \left( -\sum_{i=0}^{I} r_i \left( g_i - \sum_{\alpha=1}^{E} \mu_{\alpha} n_{i}^{\alpha} + \mu_{e} q_i \right) \right) \frac{k_B T}{k_B T} \]
\[
= \exp \left( -\sum_{i=0}^{I} r_i g_i \right), \tag{3.24}
\]

omitting vibrational effects. The equilibrium constant can thus be determined from
knowledge of the Gibbs free energies of formation for the various defect species, \( \{g_i\} \).

3.3.3 Intrinsic and extrinsic defects

In a crystal, there are two types of defects, intrinsic and extrinsic. Intrinsic defects
exist in ‘pure’ or nominally undoped crystals, whereas extrinsic defects result from
the presence of impurities or deviations from stoichiometry. Both forms of disorder
require conservation of mass, charge and stoichiometry to be obeyed, as in the
self-consistent approach outlined above. While all conceivable types of disorder consistent with the conservation rules are, in principle, allowed, in reality a handful are so strongly energetically favoured over others that they dominate the defect concentrations by many orders of magnitude. This is due to the large differences in formation energy, generally a few eV, compared with the temperatures in question, usually only a few tenths of an eV.

At finite temperature, pure crystals containing no defects do not exist. Intrinsic disorder is part of the equilibrium state of any real crystal, and the concentrations of intrinsic defects are solely determined by their relative formation energies and the temperature. In fact, if the impurity content is low enough and the temperature high enough to allow the diffusion processes needed to reach equilibrium on the experimental timescale, extrinsic defect concentrations will be negligible in comparison with intrinsic concentrations. In this case, we speak of a ‘pure’ crystal, or the ‘intrinsic regime’. The defects considered here belong to the equilibrium state in the bulk of a solid, and may well differ from those found near dislocations, grain boundaries or surfaces.

The two main types of disorder generally found in ‘pure’ crystals are Frenkel and Schottky disorder. Frenkel disorder occurs when vacancies and interstitials of the same species are in equilibrium. Cation Frenkel disorder, frequent in silver halide compounds, can be expressed using Kröger-Vink notation as

\[ \text{nil} \rightleftharpoons \text{Ag}^+_i + V^\prime_{\text{Ag}}, \tag{3.25} \]

where nil indicates perfect crystal, containing no defects. If this is the dominant form of disorder, cation interstitials and vacancies must be created in almost equal numbers. Similarly, anion Frenkel disorder involves the creation of equal numbers of anion vacancies and interstitials.

Schottky disorder comprises charge neutral combinations of cation and anion vacancies, such as

\[ \text{nil} \rightleftharpoons V^\prime_{\text{Na}} + V^\prime_{\text{Cl}}, \tag{3.26} \]

in NaCl. Schematically, the atoms removed to form the vacant formula unit are placed on new lattice sites at a surface of the crystal to preserve stoichiometry. This surface can be an exterior surface of the crystal, or an internal one along a grain boundary or dislocation. Either way, in the actual Schottky process, surface ions
move out onto new sites, and the vacant sites they leave behind diffuse into the bulk of the crystal.

Other forms of intrinsic disorder, such as antisite defects, where species A occupies the lattice site of species B in a compound containing both species, or anti-Schottky defects (quintets of interstitials) exist but are seldom found.

Extrinsic disorder involves defects that are created as a result of the presence of impurities dissolved in the lattice. In most cases, these impurities will substitute atoms in the host lattice, and only occasionally do they occupy interstitial sites [169]. The main effect impurity ions have on the defect chemistry of a system results from their difference in charge relative to the ion they replace. When the valency of the impurity ion differs from that of the ion it is replacing, it is referred to as an aliovalent impurity. The extra charge introduced by an aliovalent impurity is balanced by the creation of charge-compensating defects, which can either be mobile point defects or bound to the impurity in a complex. In contrast, isovalent substitutional impurities have no first-order effect on the defect chemistry of the host material as they require no extrinsic defects for charge compensation. To ensure continuity between the intrinsic and extrinsic regimes, the initial lattice defects necessary for charge compensation of an impurity must be one of those in the preferred form of intrinsic disorder. When a crystal is heavily doped, the defect chemistry becomes more complicated and the dominant defects cannot be easily predicted, although it has been widely asserted that one form of extrinsic defect dominates over all others [169]. The exact solution in the dilute limit can be found by determining the concentrations of all defects in the system from Eq. 3.20 while maintaining charge neutrality for a specific amount of doping.

Complicated defect clusters involving a few atoms can form upon impurity doping, many of which will be considered in this study. Clusters primarily form due to the strong electrostatic attraction between the oppositely charged impurities and their charge-compensating defects. Additionally, elastic interactions can accommodate clustering, for example when an oversized impurity attracts a vacancy to relieve the local stresses. We will distinguish between a combination of one or more impurities with oppositely charged ionic defects, referred to as a defect cluster or complex, and a combination of two intrinsic ionic defects, called a defect associate.

Understanding what types of defects are created in what amounts upon impurity doping, and how they affect the properties of materials, is the quintessence of this study.
3.4 Thermodynamics of defect formation energies

The main ingredients needed to determine the concentrations of defects in the formalism introduced in Sec. 3.3.2 are the Gibbs formation energies of all considered defects and clusters. We define formation energies using the methodology introduced by Zhang and Northrup [192], which has found wide acceptance and seamlessly links our formalism to quantities that can be calculated accurately in DFT or QMC. The Zhang-Northrup method has been employed in numerous studies of defects in semiconductors (see Ref. [41] and references therein), and, of late, has received increased attention in insulators [122, 195].

3.4.1 Zhang-Northrup formalism

We begin by selecting a sample, called a supercell, of the crystal in question. We imagine that this supercell formed in equilibrium with large reservoirs of some\(^1\) of its constituent atomic species \(\{\alpha\}\) and a reservoir of electrons, whose chemical potentials \(\{\mu_\alpha\}\) and \(\mu_e\) are then equal to those in the crystal. The formation energy of a defect is found by comparing the total Gibbs free energies of a supercell before and after creation of the defect to the chemical potentials of the atomic species and electrons added or removed.

A perfect supercell, namely one containing no defects, of bulk solid contains \(\{n_0^\alpha\}\) atoms each of element \(\{\alpha\}\). Since the concentration of perfect unit cells in Eq. 3.20 is almost unity, the approximate sum rule on the chemical potentials \(\{\mu_\alpha\}\)

\[
g_0 \approx \sum_{\alpha=1}^{E} n_0^\alpha \mu_\alpha
\]

(3.27)

from the ideal solution model is valid, and the total Gibbs energy of the perfect supercell \(g_0^{\text{tot}}\) can be written as

\[
g_0^{\text{tot}} = \sum_{\alpha=1}^{E} n_0^\alpha \mu_\alpha.
\]

(3.28)

To create a defect of type \(i\), we add (positive) or remove (negative) \(\{\Delta n_i^\alpha\}\) atoms

\(^1\)The exact number of phases that must be in equilibrium depends on the Gibbs phase rule. For a two-component system such as Al\(_2\)O\(_3\) with two degrees of freedom (pressure \(p\) and temperature \(T\)), two phases must be in equilibrium. In this study, these phases are chosen to be solid Al\(_2\)O\(_3\) and O\(_2\) gas.
of element \( \{ \alpha \} \) and \( \Delta n_i^e \) electrons. The charge associated with the defect upon adding \( \Delta n_i^e \) electrons is then \( q_i = -\Delta n_i^e \). The change in Gibbs energy, neglecting configurational entropy (which was considered explicitly in our formalism), to form defect \( i \) can then be written as

\[
\Delta g_i^f = g_i^{tot} - \sum_{\alpha=1}^{E} n_i^\alpha \mu_\alpha + q_i \mu_e
\]

\[
= g_i^{tot} - \sum_{\alpha=1}^{E} (n_i^0 + \Delta n_i^\alpha) \mu_\alpha + q_i \mu_e
\]

\[
= g_i^{tot} - g_0^{tot} - \sum_{\alpha=1}^{E} \Delta n_i^\alpha \mu_\alpha + q_i \mu_e,
\]

where \( g_i^{tot} \) is the total Gibbs energy of the defect supercell and the electron chemical potential \( \mu_e \) denotes the change in Gibbs energy per added electron of a distant reservoir of electrons.\(^2\) In our mass action formalism, \( \mu_e \) is dynamically adjusted to ensure that charge neutrality is globally obeyed in the crystal, and the chemical potentials of the constituents of the metal oxide are found using the methodology described in Sec. 3.4.2. Since the electron chemical potential \( \mu_e \) lies within the gap, it is convenient to write \( \mu_e = E_{VBM}^V + \epsilon_F \), where \( E_{VBM}^V \) is the energy of the valence band maximum (VBM) in the crystal containing defect \( i \), and \( \epsilon_F \) lies between 0 and the band gap of the perfect crystal, \( E_g^0 \). We refer to \( \epsilon_F \) as the ‘Fermi energy’, even when \( T > 0 \).

The free energy difference \( g_i^{tot} - g_0^{tot} \) in Eq. 3.29 can be approximated by \( G_i^{tot} - G_0^{tot} \), where \( G_i^{tot} \) is the Gibbs free energy of a large computational supercell containing a single defect of type \( i \), and \( G_0^{tot} \) is the free energy of the corresponding supercell of perfect crystal. The largest contributors to \( G_i^{tot} \) and \( G_0^{tot} \) are the ground-state electronic and vibrational energies. These can be determined in the supercell approximation [35], whereby a computational supercell is replicated in an infinite grid.

\(^2\)The reference level of the electrostatic potential is arbitrary and does not affect an uncharged system. However, for a charged defect, the choice of the zero of potential does affect the defect formation energy unless the electron chemical potential is measured with respect to the same reference. To avoid this, the perfect system and defective system must have the same total charge. Taking the perfect system to be charge neutral, we can compensate the charge \( q_i \) of a defective system by some distant reservoir of electrons and empty states, rendering the defect formation energy independent of the zero of potential.
Section 3.4: Thermodynamics of defect formation energies

of periodic images and its energy is evaluated using DFT or QMC methods.

3.4.2 Chemical potentials

The chemical potentials entering Eq. 3.29 are crucial in determining the formation energies, and therefore concentrations, of defects. In a compound without knowledge of the formation conditions or defect densities, they are inherently ambiguous quantities, as it is difficult to apportion the total Gibbs free energy among the compound’s constituent elements. Fortunately, simple thermodynamic considerations restrict the chemical potentials of these constituent elements to a specific range, regardless of the formation conditions of the crystal. With added knowledge of the pressure and temperature conditions at which the compound formed, the values of the chemical potentials of the constituent atoms can be determined to a high level of accuracy.

The methodology advocated here is applicable to a wide range of metal oxides, but we will use alumina, Al₂O₃, as an example. In alumina, we are interested in calculating the atomic chemical potentials μₐl and μₒ, both of which depend on the oxygen partial pressure pₒ₂ and temperature T of the oxygen atmosphere in which the crystal was annealed. Parametrising μₐl(pₒ₂, T) and μₒ(pₒ₂, T) as functions of pₒ₂ and T enables us to determine the concentrations of defects in alumina according to these annealing conditions.

Although the values of the atomic chemical potentials depend strongly on the experimental conditions, it is possible to place lower and upper bounds on them without any reference to these conditions, using only knowledge of the formation energies of the elements and the crystal. The thermodynamic stability of a crystal implies that, under identical conditions, the chemical potentials of its constituent atoms cannot be higher in the crystal than in their elemental states, otherwise the crystal would decompose. In alumina, the elemental states of oxygen and aluminium are gaseous O₂ and face-centred (fcc) cubic aluminium metal, respectively. This allows us to place an upper bound on the chemical potential of oxygen in alumina, μₒ(pₒ₂, T), at

\[
\muₒ(pₒ₂, T) \leq \mu_{\frac{1}{2}O₂(\text{g})}(pₒ₂, T),
\]

(3.30)

where \( \mu_{\frac{1}{2}O₂(\text{g})}(pₒ₂, T) \) is the total free energy per atom of molecular oxygen gas. Similarly, for \( \muₐl(pₒ₂, T) \),
where $\mu_{\text{Al}(s)}$ is the total free energy per atom of fcc aluminium metal and we have momentarily neglected its comparatively weak pressure and temperature dependence. The corresponding lower bounds can be determined with help of the experimental Gibbs energy of formation of $\text{Al}_2\text{O}_3$ (and similarly for any other oxide) under the same conditions. Combining Eqs. 3.30 and 3.31 with the Gibbs free energy balance for the formation energy $\Delta G_f^{\text{Al}_2\text{O}_3}(p_{\text{O}_2}, T)$ per formula unit of alumina at oxygen partial pressure $p_{\text{O}_2}$ and temperature $T$,

$$\Delta G_f^{\text{Al}_2\text{O}_3}(p_{\text{O}_2}, T) = \mu_{\text{Al}_2\text{O}_3} - 2\mu_{\text{Al}(s)} - 3\mu_{\frac{1}{2} \text{O}_2(g)}(p_{\text{O}_2}, T), \quad (3.32)$$

and the equilibrium condition of the solid

$$\mu_{\text{Al}_2\text{O}_3} = 2\mu_{\text{Al}}(p_{\text{O}_2}, T) + 3\mu_{\text{O}}(p_{\text{O}_2}, T), \quad (3.33)$$

we can complement the upper bounds in Eqs. 3.30 and 3.31 with lower bounds, such that

$$\frac{1}{3}\Delta G_f^{\text{Al}_2\text{O}_3}(p_{\text{O}_2}, T) \leq \mu_{\text{O}}(p_{\text{O}_2}, T) - \mu_{\frac{1}{2} \text{O}_2(g)}(p_{\text{O}_2}, T) \leq 0 \quad (3.34)$$

for $\mu_{\text{O}}(p_{\text{O}_2}, T)$ and

$$\frac{1}{2}\Delta G_f^{\text{Al}_2\text{O}_3}(p_{\text{O}_2}, T) \leq \mu_{\text{Al}}(p_{\text{O}_2}, T) - \mu_{\text{Al}(s)} \leq 0 \quad (3.35)$$

for $\mu_{\text{Al}}(p_{\text{O}_2}, T)$. The upper bound in Eq. 3.34 is commonly referred to as the ‘oxidation’, or ‘O-rich’, limit, the lower bound as the ‘reduction’, or ‘Al-rich’, limit.

These thermodynamic considerations specify an allowed range for the atomic chemical potentials. The equilibrium condition in Eq. 3.33 requires $\mu_{\text{O}}(p_{\text{O}_2}, T)$ and $\mu_{\text{Al}}(p_{\text{O}_2}, T)$ to move in harmony: for example, if one takes its maximum allowed value, the other must take its minimum, and vice-versa.

While this approach suffices in cases where the formation energy $\Delta G_f(p_{\text{O}_2}, T)$ is small, such as in III – V semiconductors, it helps little in materials such as alumina, where $\Delta G_f^{\text{Al}_2\text{O}_3}(p_{\text{O}_2}^0, T^0) \approx -17\text{eV}$ at standard pressure $p^0 = 1\text{atm}$ and temperature $T^0 = 273.15\text{K}$, designated by the superscripts $^0$. The formation energies of defect species depend so strongly on the constituent atomic chemical potentials that the
concentrations of defects can vary by more than 10 orders of magnitude over the thermodynamically allowed range of $\mu_O(p_{O_2}, T)$ and $\mu_{Al}(p_{O_2}, T)$. For metal oxides, an alternative, and more fruitful, route is to make direct contact with experiment by assuming that the oxide has been annealed in an oxygen atmosphere at specific $p_{O_2}$ and $T$, such that $\mu_O(p_{O_2}, T)$ in the solid must equal $\mu_{\frac{1}{2}O_2(g)}(p_{O_2}, T)$ for gaseous oxygen. The aluminium (or similarly for any other metal) chemical potential $\mu_{Al}(p_{O_2}, T)$ can then be determined from the equilibrium condition in Eq. 3.33.

The difficulty then lies in determining a reliable value for the chemical potential $\mu_{\frac{1}{2}O_2(g)}(p_{O_2}, T)$ per atom of molecular oxygen gas. Methods such as DFT struggle with the $O_2$ dimer to the extent that combining different functionals and pseudopotentials yields results that span a variation of ca. 1eV, far too large for the level of accuracy needed in these calculations. Fortunately, this problem can be avoided, and $\mu_{\frac{1}{2}O_2(g)}(p_{O_2}, T)$ determined to a very high level of accuracy, by combining thermodynamics with *ab initio* calculations of other quantities that are much better behaved. The method of Finnis et al. [58, 14], originally proposed for surface energy calculations in NiAl, accurately determines $\mu_{\frac{1}{2}O_2(g)}(p_{O_2}, T^0)$ at standard pressure and temperature from Eq. 3.32, using the experimental value of $\Delta G_{Al_2O_3}^f(p_{O_2}, T^0)$ and approximating $\mu_{Al_2O_3}$ and $\mu_{Al(s)}$ by zero-temperature total energies in DFT or QMC. This greatly reduces the sensitivity of $\mu_{\frac{1}{2}O_2(g)}(p_{O_2}, T^0)$ to the nuances of the electronic structure calculations. The chemical potential $\mu_{\frac{1}{2}O_2(g)}(p_{O_2}, T)$ of oxygen gas at the specific $p_{O_2}$ and $T$ corresponding to the experimental formation conditions of a crystal is then obtained from $\mu_{\frac{1}{2}O_2(g)}(p_{O_2}, T^0)$ with ideal gas relations. Comparison to thermodynamic data [90] shows that this is a reliable approach.

The change in $\mu_{\frac{1}{2}O_2(g)}(p_{O_2}, T)$ as the pressure and temperature are changed from $p^0$ and $T^0$ to any chosen $p_{O_2}$ and $T$, respectively, is governed by the ideal gas expression

$$\mu_{\frac{1}{2}O_2(g)}(p_{O_2}, T) = \mu_{\frac{1}{2}O_2(g)}(p_{O_2}^0, T^0) + \Delta \mu_{\frac{1}{2}O_2(g)}(p_{O_2}, T) + \frac{1}{2} k_B T \ln \left( \frac{p_{O_2}}{p^0} \right). \quad (3.36)$$

3 The standard form of DFT yields only zero-temperature total energies, neglecting the entropic and $pV$ contributions to the Gibbs free energy per particle. For gases, these contributions are large and strongly dependent on pressure and temperature, but can be calculated easily and accurately using simple ideal gas physics. For solids, the $pV$ contribution to the Gibbs free energy per unit cell is negligible at reasonable pressures, and the electronic specific heat is so small that the temperature dependence of the electronic contribution can also be neglected.

4 As noted already, $\mu_{Al_2O_3}$ and $\mu_{Al(s)}$ can be reliably calculated using DFT or QMC and have a far smaller pressure and temperature dependence than $\mu_{\frac{1}{2}O_2(g)}(p_{O_2}, T)$.
Although the temperature dependence of $\Delta \mu_{\frac{1}{2}O_2(g)}(p_{O_2}^0, T)$ is best obtained from experimental data [112], a more elegant, and similarly accurate solution [58, 14] neglects experimental data and integrates from $\Delta \mu_{\frac{1}{2}O_2(g)}(p_{O_2}^0, T^0)$ to any temperature $T$ using the formula for an ideal gas of rigid dumbells [90],

$$\Delta \mu_{\frac{1}{2}O_2(g)}(p_{O_2}^0, T) = -\frac{1}{2} \left\{ (S_{O_2}^0 - C_p^0) (T - T^0) + C_p^0 T \log \left( \frac{T}{T^0} \right) \right\}, \quad (3.37)$$

where $S_{O_2}^0$ is the entropy of gaseous oxygen at standard pressure and temperature and is taken from thermodynamic tables, while $C_p^0$, the heat capacity at constant pressure, is approximated as $7k_B/2$. Even when $T$ is near the melting point, for example at 2000K in alumina, this approximation only incurs an error of ca. 0.036eV. This model, combining ideal gas thermodynamics with reliable experimental and ab initio components, is therefore capable of accurately describing the translational, vibrational, and rotational degrees of freedom of the oxygen dimer at elevated temperatures.

Although the formation of any oxide can be used to obtain an estimate of $\mu_{\frac{1}{2}O_2(g)}(p_{O_2}^0, T^0)$ (offering a useful internal consistency check), we will once again focus on the example of alumina. At standard partial oxygen pressure $p_{O_2} = p^0$ and $T = T^0$, the formation of alumina occurs from fcc aluminium metal and molecular oxygen gas by the reaction

$$2\text{Al}(s) + \frac{3}{2}\text{O}_2(g) \rightleftharpoons \text{Al}_2\text{O}_3(s). \quad (3.38)$$

The Gibbs free energy balance for this reaction is

$$\mu_{\text{Al}_2\text{O}_3}(p_{O_2}^0, T^0) = 2\mu_{\text{Al}(s)}(p_{O_2}^0, T^0) + 3\mu_{\frac{1}{2}O_2(g)}(p_{O_2}^0, T^0) + \Delta G_{f\text{Al}_2\text{O}_3}(p_{O_2}^0, T^0), \quad (3.39)$$

where $\Delta G_{f\text{Al}_2\text{O}_3}(p_{O_2}^0, T^0)$ is the standard Gibbs free energy of formation of the oxide at $T^0$ and $\mu_{\text{Al}_2\text{O}_3}(p_{O_2}^0, T^0)$ is the chemical potential per formula unit of alumina. Rearranging for $\mu_{\frac{1}{2}O_2(g)}(p_{O_2}^0, T^0)$ gives

62
\[ \mu_{\frac{1}{2}O_2(g)}(p_{O_2}^0, T^0) = \frac{1}{3} \left\{ \mu_{Al_2O_3}(p_{O_2}^0, T^0) - 2\mu_{Al(s)}(p_{O_2}^0, T^0) \right\} - \Delta G_{Al_2O_3}^f(p_{O_2}^0, T^0) \] (3.40)

This reformulation is advantageous as it avoids explicit calculations of the oxygen dimer, and rather chooses to evaluate the better behaved metal and oxide using \textit{ab initio} methods such as DFT or QMC. Commonly, \( \mu_{Al_2O_3}(p_{O_2}^0, T^0) \) and \( \mu_{Al(s)}(p_{O_2}^0, T^0) \) are approximated by the corresponding zero-temperature total energies, \( \mu_{Al_2O_3} \) and \( \mu_{Al(s)} \), due to their comparatively small \( pV \) and \( TS \) contributions at \( p^0 \) and \( T^0 \). The remaining quantity needed, \( \Delta G_{Al_2O_3}^f(p_{O_2}^0, T^0) \), may be obtained accurately from tables of experimental data. From Eq. 3.40, we see that \( \mu_{\frac{1}{2}O_2(g)}(p_{O_2}^0, T^0) \) depends on the reference states of the components in the \textit{ab initio} calculation. In the pseudopotential approximation, these will be separated ions and nuclei at rest. If we choose to invoke a thermodynamic cycle on another oxide and treat the calculations on the same footing, we should always obtain the same value for \( \mu_{\frac{1}{2}O_2(g)}(p_{O_2}^0, T^0) \).

Substituting \( \mu_{\frac{1}{2}O_2(g)}(p_{O_2}^0, T^0) \) back into Eq. 3.36 leads to the final result:

\[ \mu_{\frac{1}{2}O_2(g)}(p_{O_2}, T) = \frac{1}{3} \left\{ \mu_{Al_2O_3}(p_{O_2}, T^0) - 2\mu_{Al(s)}(p_{O_2}, T^0) \right\} - \Delta G_{Al_2O_3}^f(p_{O_2}, T^0) + \Delta \mu_{\frac{1}{2}O_2(g)}(p_{O_2}, T) + \frac{1}{2} k_B T \ln \left( \frac{p_{O_2}}{p^0} \right). \] (3.41)

In theory, any tabulated oxide can be used to determine \( \mu_{\frac{1}{2}O_2(g)}(p_{O_2}, T) \) from Eq.3.41 and should give identical results; however, in practice, small variations of order 0.2eV at high temperatures \( (T = 1750K) \) occur due to experimental uncertainties in the Gibbs formation energies of the solids, the neglect of thermal contributions in the \textit{ab initio} calculations, and, for DFT, differences in the treatment of exchange and correlation between systems [90, 58]. The neglect of thermal contributions is partially ameliorated in the ideal gas expression in Eq. 3.37, but a more accurate treatment should explicitly account for vibrational contributions to the oxide and metal total energies in Eq.3.41 by employing density functional perturbation theory (DFPT) or a similar method. In the current study, vibrational effects are fully accounted for with DFPT in the harmonic approximation, contributing < 0.1eV to
Chapter 3: Point defects in solids

The total energies.

The chemical potential of the metallic element in the oxide, in this example aluminium, can now be obtained from the equilibrium condition of the solid as

\[
\mu_{\text{Al}}(p_{\text{O}_2}, T) = \frac{1}{2} \left\{ \mu_{\text{Al}_2\text{O}_3(s)}(p_{\text{O}_2}, T) - 3\mu_{\frac{1}{2}\text{O}_2(\text{g})}(p_{\text{O}_2}, T) \right\}.
\] (3.42)

Commonly, \(\mu_{\text{Al}_2\text{O}_3(s)}(p_{\text{O}_2}, T)\) (or similarly for any other oxide) in Eq. 3.42 is approximated by the zero-temperature total energy \(E_{\text{Al}_2\text{O}_3(s)}^{\text{tot}}\) per unit cell of alumina (calculated in DFT or other \textit{ab initio} methods), under the assumption that the vibrational effects at high temperatures roughly cancel in Eq. 3.29 for the defect formation energy. At experimental annealing temperatures, this is inaccurate, and vibrational contributions become significant. It is imperative that \(\mu_{\text{Al}_2\text{O}_3(s)}(p_{\text{O}_2}, T)\) is determined from the Gibbs free energy per unit cell \(G_{\text{Al}_2\text{O}_3(s)}^{\text{tot}}(p_{\text{O}_2}, T)\), using methods such as DFPT, rather than the zero-temperature energy \(E_{\text{Al}_2\text{O}_3(s)}^{\text{tot}}\), as this error otherwise accumulates entirely in the aluminium chemical potential \(\mu_{\text{Al}}(p_{\text{O}_2}, T)\).

Separating \(\mu_{\text{Al}_2\text{O}_3(s)}(p_{\text{O}_2}, T)\) into a zero-temperature contribution \(E_{\text{Al}_2\text{O}_3(s)}^{\text{tot}}\) and the normally neglected vibrational energy \(\Delta G_{\text{Al}_2\text{O}_3}(p_{\text{O}_2}, T)\),

\[
\mu_{\text{Al}_2\text{O}_3(s)}(p_{\text{O}_2}, T) = G_{\text{Al}_2\text{O}_3(s)}^{\text{tot}}(p_{\text{O}_2}, T) = E_{\text{Al}_2\text{O}_3(s)}^{\text{tot}} + \Delta G_{\text{Al}_2\text{O}_3}(p_{\text{O}_2}, T),
\] (3.43)

we can compare the accurate \(\mu_{\text{Al}}(p_{\text{O}_2}, T)\) that includes vibrational contributions to the inaccurate \(\tilde{\mu}_{\text{Al}}(p_{\text{O}_2}, T)\) that is commonly determined from \(E_{\text{Al}_2\text{O}_3(s)}^{\text{tot}}\):

\[
\mu_{\text{Al}}(p_{\text{O}_2}, T) = \frac{1}{2} \left\{ E_{\text{Al}_2\text{O}_3(s)}^{\text{tot}} + \Delta G_{\text{Al}_2\text{O}_3}(p_{\text{O}_2}, T) - 3\mu_{\frac{1}{2}\text{O}_2(\text{g})}(p_{\text{O}_2}, T) \right\}
\]

\[
= \tilde{\mu}_{\text{Al}}(p_{\text{O}_2}, T) + \frac{1}{2}\Delta G_{\text{Al}_2\text{O}_3}(p_{\text{O}_2}, T).
\] (3.44)

At typical experimental temperatures, \(\Delta G_{\text{Al}_2\text{O}_3}(p_{\text{O}_2}, T)\) for metal oxides such as \(\text{Al}_2\text{O}_3\), \(\text{MgO}\), \(\text{TiO}_2\) or \(\text{Cr}_2\text{O}_3\) lies between 1 and 2eV. Neglect of the vibrational contribution to the total energy of the bulk oxide therefore introduces an error of \(O\) [1eV] to the aluminium chemical potential. After including these vibrational contributions, both atomic chemical potentials are known to a high accuracy, and defect formation energies, and thus concentrations, can be calculated as functions
3.5 Density functional theory calculations of defect formation energies

Density functional theory (DFT) is an established tool for evaluating the static properties of solids and has been successfully used to characterise a vast array of different materials. As DFT is generally used in the pseudopotential approximation, total energies have no physical meaning and only energy differences, such as cohesive and formation energies, can be related to experimental quantities. Unfortunately, these are rarely straightforward differences of two quantities, and a number of factors must be carefully considered so that the ultimate result is accurate enough for the intended purposes. In order to determine defect concentrations that can be compared with experiment, the formation energies of the various defects must be calculated to a high level of accuracy at elevated temperatures and pressures. This cannot be done directly in DFT, and the previous section described how the chemical potentials of constituent atoms can be found at specified $T$ and $p_{O_2}$ by combining DFT with thermodynamic relations.

When calculating defect formation energies in DFT, there are also two main limitations that hamper every calculation: the inexactness of the exchange-correlation functional, and the small numbers of atoms which are computationally tractable. The first leads to a drastic underestimation of the band gap (with the exception of hybrid functionals), and the second introduces defect-defect interactions via the periodic boundary conditions. Careful considerations of these effects are necessary in order to obtain defect concentrations comparable to experiment.

3.5.1 Vibrational effects

In Sec.3.3.2, it was shown that defect concentrations are governed by their respective formation energies, which, from the Zhang-Northrup formula in Eq. 3.29, depend on the chemical potentials of constituent atoms and the Gibbs free energies of computational supercells containing single defects. The preceding section demonstrated how the atomic chemical potentials of a system can be determined at a specified oxygen partial pressure $p_{O_2}$ and temperature $T$. This enables the chemical potentials...
to be related to the formation conditions of an experimental sample.

In Sec. 3.4.2, we established that, if the oxygen chemical potential is determined by the method of Finnis et al. and DFT total energies for the metal and metal oxide are used, this leads to a significant error in the chemical potential of the metallic atom. Furthermore, the use of $T = 0K$ total energies can lead to an imbalance between the formation energies of different defects if the vibrational contributions from various defects differ. We correct for the first of these problems by determining the Gibbs free energy, including vibrational contributions, for the metal and metal oxide. For the formation energies of the defects, we approximate the Gibbs free energies $\{g^{tot}\}$ entering Eq. 3.29 by DFT total energies $\{E^{tot}\}$ of computational supercells at $T = 0K$.

Although ab initio studies of the vibrational free energies of semiconductors and simple metals are plentiful [1, 135], computational demands have limited calculations of more complex solids [173]. Several methods can be used to extend $T = 0K$ calculations to finite temperatures. Thermodynamic integration techniques determine the free energy relative to a known reference system [113, 104]. However, such calculations require Monte Carlo or molecular dynamics (MD) simulations at numerous fixed temperatures and knowledge of a reference system, rendering them unsuitable for our current purposes. Recent advances in density functional perturbation theory (DFPT) [191, 11, 12, 69] have made it the tool of choice for determining vibrational frequencies at every point of the Brillouin zone. The vibrational free energy can then be obtained in the harmonic [118] or quasiharmonic approximation. In the quasiharmonic approximation, anharmonic effects are included through an explicit volume dependence of the vibrational frequencies. This method has been used to determine the thermal expansivity for many simple metals (see Ref. [135] and references therein).

In this study, the Gibbs energies $G_{i}^{tot}(p, T)$ of computational supercells are obtained from DFT total energies $E_{i}^{tot}$ at $T = 0K$, and the vibrational contributions are calculated from DFPT calculations of the phonon modes of the supercells in the harmonic approximation. We now establish the relationship between Gibbs free energies and DFT total energies, and the role that the vibrational free energy plays, in detail.

The equilibrium geometry of a crystal containing $N$ atoms at temperature $T$ is obtained by minimising the Gibbs free energy
Section 3.5: Density functional theory calculations of defect formation energies

\[ G^{\text{tot}} = E^{\text{tot}} + F^{\text{vib}} + \Delta E^{\text{el}} - T S^{\text{el}} + pV, \]

(3.45)

where \( E^{\text{tot}} \) is the ground state \((T = 0\text{K})\) total electronic energy of a computational supercell (neglecting zero-point vibrations) calculated in DFT and \( F^{\text{vib}} \) the vibrational free energy. The next two terms, namely the electronic energy due to thermal excitations \( \Delta E^{\text{el}} \) and the electronic entropy \( T S^{\text{el}} \), are very small and assumed to be negligible. It has been suggested that, in transition metals, this may not be true and the electronic contribution may be significant due to the large density of states near the Fermi energy \([135]\). The final \( pV \) term is small at reasonable pressures and can safely be neglected \([154]\). The last contribution needed to determine the Gibbs free energy is the vibrational term \( F^{\text{vib}} \), which includes the zero-point energy. To calculate \( F^{\text{vib}} \) accurately, it is necessary to determine the vibrational frequencies \( \omega \) across the Brillouin zone, which can be done using DFPT. In the harmonic approximation, thermal expansion is ignored and the frequencies \( \omega \) are evaluated at the equilibrium volume of the statically relaxed supercell. Anharmonic thermal effects can be included by rendering the frequencies volume dependent. The equilibrium volume \( V(T) \) is then determined by minimising the total free energy \( F^{\text{vib}} \) with respect to volume under constant pressure. This is referred to as the quasiharmonic approximation. Unfortunately, accounting for these anharmonic effects requires performing DFPT calculations at various lattice parameters, the high computational cost of which has so far hindered its use in complex systems such as those encountered in this study.

Ignoring anharmonic, \( pV \), and electronic effects, the Gibbs free energy (or, strictly, the Helmholtz free energy) in the harmonic approximation can be written as\(^5\)

\[ G^{\text{tot}}(T) \approx E^{\text{tot}}(T) + F^{\text{vib}}(T). \]

(3.46)

The vibrational contribution \( F^{\text{vib}}(T) \) can be derived from the partition function of a vibrating \( N \)-atom system in the harmonic approximation as \([118, 18]\)

\(^5\)Strictly, the first term in Eq. 3.46 is an enthalpy at finite temperature; however, it can be replaced with an internal energy at zero temperature, as demonstrated in Ref. [74].
\[ F^{\text{vib}}(T) = k_B T \int d\omega \ln \left\{ 2 \sinh \left( \frac{\hbar \omega}{2k_B T} \right) \right\} \sigma(\omega), \quad (3.47) \]

where \( \sigma(\omega) \) is the phonon DOS.

In a DFPT calculation, accuracy is paramount. In order to avoid imaginary vibrational frequencies and similar artefacts, the geometry of a system must be very well converged, requiring high tolerances on forces and fine grid scales and \( k \)-point sampling.

At high temperatures, especially when approaching a phase transition, the limits of the harmonic and quasiharmonic approximations become evident. For example, thermal expansivity can cause the harmonic approximation to fail in capturing of order 15\% of the vibrational energy. Even in the quasiharmonic approximation, phonon frequencies at a given volume are independent of temperature, whereas, in a real crystal, they are not: anharmonic effects can also shift phonon frequencies \[117\] at elevated temperatures. In addition, errors intrinsic to DFT, such as the LDA underestimation of lattice parameters, are aggravated at high temperatures.

### 3.5.2 Band gap corrections

In the region surrounding a point defect, the local geometry and electronic structure differ significantly from those in the bulk. The change in geometry can introduce localised electron orbitals near the defect site, called ‘defect states’, the energy levels of which lie in the band gap. Defect states can either derive from former conduction band states that have been lowered in energy by the presence of the defect, or from valence band states raised in energy (or a mixture of the two). In the second case, if the states are not very strongly localised, they are similar in character to the occupied valence band states of the perfect crystal. In DFT, the energy eigenvalues of defect states derived from the valence band are reasonably accurate, bearing in mind the typical limitations of DFT calculations. However, if the localised defect orbital is composed of conduction band states, the eigenvalue of the defect state will suffer from the DFT band gap problem \[162, 163\], whereby energy gaps to conduction band states are severely underestimated. The eigenvalue of the eigenstate may therefore be far too low. This problem is intrinsic to DFT calculations of defects, and leads to a deceptive underestimation of the formation energies of defects with occupied localised states of this type.
There are several mutually incompatible methods to combat this problem, but none is completely correct and all contain empiricism. In general, it is not straightforward to discern which scheme yields the best results for a specific system, and it is necessary to understand the physics of the system before committing to a remedy.

A commonly-used approach in semiconductors to align DFT formation energies with experimental results rigidly shifts the energy eigenvalues and occupied defect levels by the difference between the experimental band gap and the DFT band gap, $\Delta E_{\text{g}}$ [194, 193]. If there are $m$ electrons occupying defect states thought to derive from the conduction band, these are rigidly shifted by adding $m \times \Delta E_{\text{g}}$ to the total energy of the defect supercell. In semiconductors, this remedy is necessary for agreement between calculated defect formation energies and experiment. However, in wide band gap ionic insulators such as $\text{Al}_2\text{O}_3$, this empirical correction is problematic.

In semiconductors, defect states high in the band gap are primarily delocalised and conduction-band-like in nature, so that a rigid energy shift is sensible. In contrast, in strongly ionic insulators, the defect states can be very strongly localised at the defect site, and it can be unclear whether they are associated with the conduction or valence band (see Fig. 3.2). For example, consider a neutral oxygen vacancy in $\text{Al}_2\text{O}_3$, designated as $V_0^\times$ in Kröger-Vink notation. Two localised electrons reside on the vacant oxygen site where, previously, an oxygen O$^{2-}$ anion with filled 2s and 2p shells lay. However, despite the removal of the oxygen nucleus, the defect states need not derive from the unoccupied aluminium 3s orbitals at the bottom of the conduction band, and may well look rather like the now-missing oxygen s-orbitals. In these circumstances, it is unclear whether the state should be shifted at all. In addition, the wide band gap and large band gap error mean that any applied shift would have a large impact on the formation energy.

Another empirically-motivated approach involves determining the DFT energy level for a reference defect that has a well-defined experimental defect level in the band gap [26]. The difference between the DFT and experimental levels can then be used to shift other calculated defect levels. This ‘reference level’ scheme is very empirical and hinges on the choice of reference defect and the experimental accuracy of its level. Castleton et al. [26] explored using the valence band maximum (VBM) and conduction band minimum (CBM) as the reference states for donor and acceptor levels in Si and InP, but found the scheme to be internally inconsistent and perform worse than other remedies.

The use of a physically-motivated ‘scissor operator’ to stretch the DFT states over...
Figure 3.2: Importance of band gap corrections on defect states. The valence band maximum (VBM) and conduction band minimum (CBM) are shown in blue and green, respectively. The calculated band gap of the perfect crystal in DFT, $E_g^0$ (DFT), differs from the experimental band gap, $E_g^0$ (exp.), by an amount $\Delta E_g^0$ (DFT), referred to as the band-gap error. In experiment and in DFT, valence band-like defect states, $\varepsilon_{\text{val}}^i$ (exp.), lie just above the VBM. However, conduction band-like states that lie just below the CBM in experiment, $\varepsilon_{\text{cond}}^i$ (exp.), are lowered by the DFT band-gap error to somewhere below the DFT CBM, $\varepsilon_{\text{cond}}^i$ (DFT). For these states, a band gap correction can be applied that shifts the energy eigenvalue up by an amount related to $\Delta E_g^0$ (DFT). In addition, there are defect states that exhibit neither predominantly valence nor conduction band character, designated by $\varepsilon_i$ (DFT), for which it is more difficult to devise a suitable correction scheme.

The experimental gap has been suggested as a well-behaved alternative to rigid energy shifts. This has been cruelly achieved by placing conduction band-like states (i.e. those that lie just below the calculated DFT CBM) the same distance below the experimental CBM, and analogously for valence band-like states. However, this is no more physical than a rigid energy shift. A better approach involves examining the form and symmetry of the defect states in order to determine the amount of conduction band nature they contain. Motivated by Baraff and Schlüter [8], this can be implemented by determining an effective value of $m$, the number of electrons localised at the defect site, from the nature of the defect states. In this formalism, a Hamiltonian can be constructed that widens the DFT band gap by $\Delta E_g^0$ to obtain the experimental gap by adding a scissor operator of the form $\Delta E_g^0 \hat{P}_{\text{cond}}$ to the DFT Hamiltonian, where $\hat{P}_{\text{cond}}$ is a projection operator onto the conduction bands of the perfect crystal. A first-order estimate of the contribution of this additional term to the electronic energy of a system with defect levels in the gap may be obtained by summing the expectation values of the scissor operator in the occupied defect
Section 3.5: Density functional theory calculations of defect formation energies

states. This calculates the percentage of conduction band nature in the defect states from their overlap with all the conduction band states of the perfect crystal. The effective correction to the total energy then becomes

\[ m_{\text{eff}} \Delta E_0^g = \sum_{i, \text{def}} \sum_{j, \text{cond}} \left| \langle \psi_i^{\text{def}} | \psi_j^{\text{cond}} \rangle \right|^2 \Delta E_0^g \]

\[ = \sum_{i, \text{def}} \left( 1 - \sum_{j, \text{val}} \left| \langle \psi_i^{\text{def}} | \psi_j^{\text{cond}} \rangle \right|^2 \right) \Delta E_0^g, \]  

(3.48)

where the superscripts ‘val’, ‘cond’, and ‘def’ refer to the valence and conduction states of the perfect crystal and the defect states of the defect supercell, respectively. This approach is still somewhat empirical, and depends on the DFT functional used, so that it does not constitute an exact solution to the problem. However, it is fortunate that the band gap error, as we shall later demonstrate, does not affect the formation energies of the most important defects in the systems studied here. As we are dealing with wide band gap insulators, the most stable defects are generally those with full ionic charges, and these do not have any occupied states in the gap. The scissor operator only corrects defects in charge states below their full ionic values, which we will show generally have higher formation energies than their full-charge counterparts over the experimentally accessible range of the electron chemical potential, \( \mu_e \). 

Unfortunately, the band gap problem will always exist for LDA and GGA defect calculations in DFT, and post-hoc corrections will always contain a degree of empiricism. Alternative approaches involve employing hybrid functionals, which are capable of correctly localising electron states and yield the correct gap, or switching from DFT to QMC, which does not suffer from the band gap problem at all. Unfortunately, the computational cost of hybrid exchange calculations in a plane wave basis prohibited their use in the current study, while QMC, which is even more demanding, suffers other difficulties that we will explain later.

3.5.3 Finite-size effects

A central assumption of the method presented here is that the defects of the system in question are created in the dilute limit. However, when defects form in a material, they need not exist in complete isolation. In a real solid, defects will migrate to form
Chapter 3: Point defects in solids

clusters when it is energetically favourable. In the systems considered in this study, defect concentrations are very low, and as we account explicitly for the formation of various clusters in addition to the basic point defects, the assumption that we are working in the dilute limit is reasonable. In order to determine the concentrations of the various possible defects, we require accurate defect formation energies and binding energies of clusters, the Gibbs energies of which are ideally calculated from infinitely large cells containing an isolated defect or cluster. Unfortunately, it is impossible to use infinitely large computational cells, as this would mean determining an infinite number of wave functions for an infinite number of electrons. In practice we are only able to perform calculations for finite, periodic systems, making use of Bloch’s theorem to help calculate the electronic wave functions.

Hallmark implementations in DFT, such as pseudopotentials and plane wave basis sets, allow the simulation of cells containing hundreds of atoms to a high level of accuracy. As we are concerned with defects in the bulk of a material, we perform calculations in the supercell approximation [105], whereby a chosen computational ‘supercell’ is periodically repeated in an infinite array. This renders the cell boundary bulk-like, eliminating surface effects without the need of an intractably large supercell, and enables the use of efficient computational algorithms. However, the disadvantage is that a supercell containing a defect is surrounded by an infinite array of identical supercells containing the same defect. The restriction to finite computational supercells repeated in an infinite array generates a plethora of considerations.

If the supercell is not large enough to contain the full effect of the long-ranged elastic and Coulomb fields of the defect, the periodic repetition introduces spurious interactions with neighbouring replicas of the defect, the magnitudes of which depend strongly on the size and shape of the supercell. The potential created by a defect is felt by all its periodic copies, and it feels the potential of an infinite lattice of copies of itself. These defect interactions must be considered accurately, as they may modify the calculated defect formation energies and affect important phenomena, such as dopant clustering [173]. The situation is exacerbated when the defect carries a strongly localised charge that gives rise to long-range lattice relaxations [103]. In fact, for any charged supercell, the electrostatic Madelung energy of the infinite array of like-charged point charges is infinite. This divergence is traditionally removed by balancing the monopole charge of the defect with a homogeneous uniform background [105], rendering the computational supercell neutral and enabling use of Ewald’s method [54] for evaluating the Madelung energy. The use of this
neutralising ‘jellium’ background introduces another spurious interaction, this time between the defect and the background, that also depends on the linear dimension $L$ of the supercell, defined as the cube root of the supercell volume $L = \sqrt[3]{V}$ [103]. In a real system, positively charged defects would be associated with a corresponding charge-balancing negative defect nearby.

When calculating the defect formation energy $\Delta G_i^f$ of a neutral defect in a finite-sized supercell, the average value of the potential, $v_B$, cancels out and does not contribute to the total energy. For a charged defect, the reference level is also arbitrary and should not affect $\Delta G_i^f$; the cancellation that makes this possible occurs between the charge $q_i$ of the defect, which contributes $q_i v_B$ to the defect formation energy, and the electron chemical potential term in the Zhang-Northrup formation energy (Eq. 3.29), which contributes $-q_i \mu_e$. If the chemical potential $\mu_e$ is measured relative to the same reference level as the defect calculation, this cancellation is perfect. However, there is a difficulty in choosing a value for $\mu_e$, since the long-ranged Coulomb potential around the charged defect shifts the VBM and CBM up or down in the defect supercell and it is difficult to decide where these shifted levels ought to lie relative to the true levels, which can only be established very far from the defect and thus only by using an impractically large simulation cell (see Fig. 3.3). Therefore, the position $E_{VBM}^i$ of the VBM in the defective supercell ought to be different from that in the perfect supercell, $E_{VBM}^0$, but we are not sure by how much.

There are many different methods that try to correct for the effect of spurious interactions from the potentials of nearby periodic copies of defects and thereby correctly align the chemical potential $\mu_e$. The methods can be broadly cast into two classes, one that tries to calibrate the value of $E_{VBM}^i$ used to determine $\mu_e$ in Eq. 3.29 with the value it takes in the perfect supercell, and a second that retrospectively adjusts the total energy of the defect supercell to remove spurious defect-defect contributions. Members of the first class include simple average potential corrections and more advanced techniques that try to cancel the potential from the periodic images of the defect. The second class comprises different methods of scaling total energies acquired from calculations of supercells of different sizes in order to extrapolate to the infinite-size limit. In the past, these two methods have frequently been considered individual and complementary corrections, but a review by Castleton et al. [26] clearly demonstrates that they both attempt to correct for the same problem and should not be used simultaneously, as this overestimates the required correction.
Figure 3.3: The Coulomb potential of the charged defect shifts the CBM and VBM in a defect supercell (right) compared to a perfect supercell (left).

The simplest approach to align $E_{i}^{\text{VBM}}$ with $E_{0}^{\text{VBM}}$, the average potential correction (or potential realignment), attempts to remove the background potential caused by a lattice of defects. This is achieved by replacing $E_{i}^{\text{VBM}}$ with the sum of $E_{0}^{\text{VBM}}$ and an average potential shift $\Delta V$ [147, 123]

$$E_{i}^{\text{VBM}} \rightarrow E_{0}^{\text{VBM}} + \Delta V,$$  \hspace{1cm} (3.49)

in the Zhang-Northrup formation energy from Eq. 3.29. The shift $\Delta V = V_{i}^{\text{av}} - V_{0}^{\text{av}}$ represents the difference in ‘average’ potentials between a defect supercell, $V_{i}^{\text{av}}$, and the perfect supercell, $V_{0}^{\text{av}}$. The averages are calculated over a chosen range as far from the defect as possible within the supercells. Unfortunately, the definition of ‘far’ from the defect is arbitrary, and there is no unique prescription for choosing the range over which to average. In a semiconductor that contains delocalised defect states, this is not a problem, and the average potential correction can reliably align the band structures of the defect and perfect supercells, giving errors in defect formation energies of individual cells of order 0.1eV and scaled energies to the infinite-size limit to within 0.5eV [26]. However, in an insulator with strongly localised defect charge, the approach is less reliable. Small movements of the plane over which the average potential is estimated can change the defect formation energy by 1eV in supercells of common size, hindering extrapolation to the infinite-size limit. This is because no region in a supercell of an ionic crystal is truly far enough from the defect to be considered bulk-like. Moreover, as the average potential correction does not explicitly account for defect-interaction errors, there is no guarantee that it is suitable for correcting defect formation energies. Fig. 2 in Ref. [82] demonstrates the inaccuracy of this method for native point defect formation energies in reasonably-sized supercells of Al₂O₃.
Ironically, the potential realignment scheme is sometimes successful even when the range over which the potential is averaged is not bulk-like at all. The additional shift in the averaging region due to the effects of electrostatic defect-image interactions can occasionally cancel the error in the realignment, producing accurate results.

More promising routes attempt to cancel the potential from the periodic images of the defect using a potential generated from a periodic array of counter-charges. These solutions try to correct the boundary conditions on the potential a priori during the simulation. For example, Schultz’s local moment counter-charge method (LMCC) [160, 161, 159] creates a model charge density from spherical Gaussians to describe a localised defect charge. This model density is constructed to match the moments of the charge density of the defect cell, so that separating the model density from the total charge density leaves a residual, periodic charge distribution void of long-range multipole interactions. The residual density can then be incorporated in the supercell approximation to yield an electrostatic potential that is uncontaminated by moments from periodic images of defects, while the isolated model density can be examined analytically. However, this approach neglects the polarisation response of the surrounding medium to the defect charge, and assumes that the remaining periodic charge after the moments have been removed closely mirrors the true periodic charge of the crystal, of which there is no guarantee. In addition, there is no analytical correction formula for the LMCC method and Wright and Modine [188] demonstrated that LMCC-corrected total energies of unrelaxed silicon defects often converge considerably slower than the uncorrected values as the supercell size is increased.

The other class of correction schemes attempts to perform total energy calculations of supercells of various sizes but identical symmetry that contain the same defect, and then extrapolate the results to the limit of an infinite supercell. The most commonly applied finite-size correction was proposed by Makov and Payne [115]. This relates the total energy of a cubic supercell of linear size $L$ containing a defect to the infinite ($L \to \infty$) limit by

$$ E_i^{\text{tot}} (L) = E_i^{\text{tot}} (L \to \infty) - \frac{q^2 \alpha}{2 \epsilon} \cdot L^{-1} - \frac{2 \pi q Q}{3 \epsilon} \cdot L^{-3} + O [L^{-5}] , \quad (3.50) $$

where $q$ is the charge of the defect cell, $Q$ its quadrupole moment, $\epsilon$ the relative dielectric constant of the medium in which the charges are embedded, and $\alpha$ the
Madelung constant. The Madelung constant is a property of the shape of the supercell: in a simple cubic system \( \alpha = 2.8373 \); in a hexagonal system its value depends on the relative lengths of the lattice parameters \( a \) and \( c \). Since dipole-dipole terms do not necessarily cancel in non-cubic systems, higher-order \( Q^2 \) contributions may also be important.

The Makov-Payne correction, as well as the aforementioned LMCC method, work very well for isolated ions or molecules and for metals with strong screening [103]. In these cases, there is no difficulty in defining the localised defect charge, there are no problems with polarisability of the bulk or boundary conditions, and, for metals, the Coulomb defect-defect interaction between supercells is negligible. However, for insulators, and to a lesser extent semiconductors, these problems become important, and the above corrections are inadequate for accurate calculations of defect formation energies. Castleton et al. [26] performed an exhaustive comparison between applying just the first, or both the first and second terms of the Makov-Payne correction from Eq. 3.50, and the average potential correction from Eq. 3.49 to semiconductors. Neither of the methods proved universally reliable, but the analysis indicated that the forms, although not the magnitudes, of the terms in the Makov-Payne correction were appropriate. Using fitted parameters for the coefficients in the first and second terms improved the results, giving isolated charged defect formation energies in InP to \( \pm 0.05 \) eV. Shim, Lee and Nieminen [165], Wright and Modine [188], and Lento, Mozos and Nieminen [103] reached similar conclusions.

This work applies our recently proposed approach that exploits the form of the Makov-Payne correction but adds flexibility [82]. In Eq. 3.50, the first-order correction term dominates for single highly-localised defects where the monopole-quadrupole and dipole-dipole terms are small. It contains the ratio \( \alpha/L \), referred to as the Madelung potential \( v_M \). Previous applications of the Makov-Payne correction have fixed \( \alpha \) and determined the total energy as a function of \( L \) for fixed lattice geometry, using the results to estimate \( \epsilon \). However, this is unnecessarily limiting, as it is possible to change the shape of the supercell along with its size. This can be done by varying the number of unit cells along different crystal axes, resulting in different values of \( \alpha \) as well as \( L \). Plotting the total energy against \( v_M \), which is a function of both \( L \) and \( \alpha \), for a wide range of cell shapes and sizes, allows an extrapolation to the infinite-cell limit where \( v_M = 0 \). In practice this is accomplished by fitting a straight line,
to the calculated energies. By comparison with Eq. 3.50, the gradient $b$ can then be used to find an effective dielectric constant from $b = \frac{\sigma^2}{2\epsilon}$.

This variation of Castleton’s finite-size scaling approach has three advantages over a vanilla Makov-Payne style correction. The first is that, as it does not rely on a uniform scaling with $L$, it enables a much larger range of cells to be used within the computational limits. The smallest supercells that can account for the geometric relaxations surrounding a defect should contain at least second-nearest neighbours to the defect site. When using the Makov-Payne correction, isotropically doubling the size of such a supercell twice generally results in an unfeasibly large calculation.

The proposed variation on the method, on the other hand, allows the cell size to be changed along different crystallographic axes rather than simply isotropically. This produces many reasonably small defect supercells of different shapes and sizes, allowing the calculation of a large number of data points for the linear fit in Eq. 3.51, rather than the two or three obtained when the cell shape is fixed. The second advantage lies in the extrapolation technique. In the Makov-Payne correction, fixing a value of $\alpha$ and increasing the cell size uniformly means that the extrapolation (made from a very small number of points) has a large error in the infinite-size limit. With this revised method, it is possible to create cells that have negative Madelung constants. This allows (a significant number of) data points to be calculated either side of the infinite cell-size limit, where $v_M = 0$. Rather than making an unreliable extrapolation to a point outside the range of data, it is therefore possible to mathematically interpolate to the infinite-size limit, which is generally more reliable.

Finally, electrostatic monopole terms are not the only ones that scale as $L^{-1}$; the energy stored in the elastic field arising from geometry relaxations (which are driven by the Coulomb forces) scales as $L^{-1}$, also. This is a reason why many previous studies found that the Makov-Payne correction worked far better on unrelaxed, rather than relaxed, total energies. In the scheme advocated here, the free form of the $L^{-1}$ term can accommodate elastic relaxations as well as purely electrostatic effects.

This approach to system-size extrapolation assumes that the finite-size errors depend only on $v_M$ and neglects quadrupole and higher order terms, which scale differently with system size. For example, finite-size errors arising from uncertainties in the
VBM of the defect supercell, $E_{\text{VBM}}$, scale with the number of atoms contained in the supercell [26], which is essentially $L^3$. The Madelung interpolation approach is therefore not a complete solution, but should only be used as a guide in systems where it can provide reliable results. The importance of the neglected terms is indicated by the magnitude of the deviations of data points from the straight line in Eq. 3.51. In wide band gap insulators with highly-localised defect charges, the higher-order $L^{-3}$ and $L^{-5}$ terms are generally small, and the Madelung interpolation offers a useful tool for dealing with defect-defect interactions.

Accounting for defect-defect interactions in finite supercell calculations of strongly ionic metal oxides is a difficult problem for which many remedies have been proposed. Unfortunately, none of these is capable of exactly solving the problem, and the application of different prescriptions can change estimates of defect formation energies by many eV. The path taken in this work provides an alternative route, allowing reliable extrapolations to the infinite-size limit from the results of calculations for cells of various shapes and sizes. This offers a number of advantages and, as shown in Sec. 4.4.6, can yield formation energies that vary little with cell size.

### 3.6 Diffusion kinetics of defects

Calculating diffusion coefficients from first-principles is an audacious undertaking, and much care must be taken if the results are to complement or explain experimental diffusion data. When defects, such as vacancies and interstitials, diffuse through a solid, they give rise to a diffusive flux of atoms that can be harnessed for a number of purposes. Diffusion enables properties such as sintering, creep, and oxide film growth on alloys, and can be used to transport certain ions, such as oxygen, through a host lattice. The diffusion process can be characterised by either a continuum or atomic model. In the former, the diffusing substance and its host lattice are treated as a continuous medium in order to derive thermodynamic relationships. The atomic model relates macroscopic quantities, such as the flux, to atomic quantities, such as atom jump frequencies and distances. This provides a useful means of deriving expressions for the diffusion coefficient in specific crystals. Many of the atomic quantities appearing in the derivation can be found directly from the geometry of the crystal lattice, and others, such as formation energies and diffusion barriers, from electronic structure calculations.
3.6.1 Mechanisms of diffusion

Understanding diffusion in any solid begins with identifying the mechanism by which defects migrate through the host lattice. In an ordered crystal lattice, a diffusing species can only occupy certain equilibrium lattice sites. A macroscopic diffusion process occurs when a defect performs numerous jumps between these sites. The elementary jumps in this series can be of many kinds, but we will restrict this discussion to those typically found in metal oxides, in particular $\text{Al}_2\text{O}_3$.

The simplest and most common method of diffusion is the vacancy mechanism, first elucidated by Kirkendall [168] and commonly found in alkali halides and most oxides with a NaCl structure. In the vacancy mechanism, a vacancy approaches an atom occupying a lattice site and exchanges positions with it, as seen in Fig. 3.4(a). The ion is immobile until the vacancy approaches it and allows it to jump. The flux of diffusing atoms in one direction therefore balances the flux of vacancies in the opposite direction. As all crystals in thermal equilibrium contain populations of vacancies, this provides an easy path for diffusion. When divacancies are present, this mechanism becomes more complicated, as a divacancy has the ability to dissolve into two vacancies, rotate as a non-rigid dumbbell, or translate directly.

Atoms that naturally dissolve in a host lattice by occupying interstitial lattice sites diffuse by an interstitial mechanism (see Fig. 3.4(b)). They migrate through the crystal by jumping directly from one interstitial site to another, until they drop into a vacancy to occupy a normal lattice site. This mechanism is particularly likely for the diffusion of small impurity atoms that can easily fit into the interstices without greatly perturbing the surrounding geometry. The relatively straightforward interstitial diffusion process generally requires little energy ($< 0.5\text{eV}$) to surmount the diffusion barrier separating interstices, and can frequently be found in cubic lattices. Occasionally, radiation damage leads to interstitial diffusion by producing a high concentration of Frenkel pairs. The interstitial atom of a Frenkel pair is termed a ‘self-interstitial’, and while many recombine with their corresponding vacancy, some may diffuse away from their initial site. Upon extreme radiation damage, the concentration of interstitials can become so large that they precipitate in microscopic clusters. Due to their high mobility, interstitial defects can often travel large distances before annihilating at grain boundaries, dislocations or with vacancies.

An indirect interstitial mechanism, called the interstitialcy mechanism, can occur
Figure 3.4: Elementary jumps for various diffusion mechanisms in a schematic 2D lattice.
(a) Vacancy mechanism
(b) Interstitial mechanism
(c) Interstitialcy mechanism: solid arrows indicate collinear, dotted arrows non-collinear, displacements of diffusing atoms
(d) Dumbbell interstitialcy mechanism: solid arrows indicate atomic displacements of diffusing atoms, the dotted arrow the displacement of the interstitialcy defect.
(e) Crowdion mechanism: solid arrows show atomic displacements of diffusing atoms, the dotted arrow the displacement of the crowdion defect.
when an interstitial atom cannot easily move between interstices. Instead, the interstitial moves into a normal lattice site, pushing the atom previously occupying that site onto a neighbouring interstitial site, as in Fig. 3.4(c). The cooperative motion of the two atoms moves the interstitialcy defect from one interstitial site to another. This mechanism occurs, for example, in silver halides such as AgBr, where the Ag\(^+\) ions are relatively small compared to the Br\(^-\) ions. Occasionally, an interstitial atom and a lattice atom may occupy a lattice site together in a ‘dumbbell interstitialcy’ configuration (see Fig. 3.4(d)). Then, the interstitialcy moves into the lattice site, occupying it fully and pushing the lattice atom into a new dumbbell position at a neighbouring lattice site.

A crowdion mechanism (see Fig. 3.4(e)) can arise when a single lattice site is shared by two atoms, for example when there is a ‘split interstitial’ defect. In some situations, the extra atom can cause a larger number of atoms to displace slightly from their equilibrium lattice positions, not just two. The crowdion configuration can readily move along a row, and, as it diffuses through the solid, displaces each atom in the row by one lattice spacing. The mechanism does not result in three-dimensional diffusion, however, as the crowdion can not migrate between rows. Although this configuration is rarely found in thermal equilibrium, it can result from radiation damage or when successive collisions are confined to specific crystallographic directions.

These mechanisms only pertain to diffusion in the bulk of a crystal, not along dislocations, in grain boundaries, or on surfaces. These cases differ substantially from bulk diffusion and require an entirely separate treatment.

### 3.6.2 Activation energies and diffusion barriers

Bulk diffusion by any mechanism has two requirements: diffusing defects must be present, and they must acquire the thermal energy necessary to complete an elementary jump. The rate of diffusion \(D(p, T)\) is proportional to both the defect concentration and the jump rate. In thermal equilibrium, the probability \(P_i^f(p, T)\) of finding defect species \(i\) at any given site is proportional to the Boltzmann factor

\[
P_i^f(p, T) \propto \exp\left(-\frac{\Delta G_i^f(p, T)}{k_B T}\right),
\]

where the formation energy \(\Delta G_i^f(p, T)\) is the Gibbs energy required to create a single defect \(i\) on that lattice site. In order for the defect to perform an elementary
Chapter 3: Point defects in solids

diffusion jump, it must surmount a potential barrier $\Delta G_{\text{mig}}^i$ separating equivalent lattice sites. The diffusion barrier $\Delta G_{\text{mig}}^i$ varies between defects. When there are a number of non-equivalent paths to nearest-neighbour sites, the barrier may differ among the paths, and defects will diffuse along the path with the lowest barrier. The probability of a defect $i$ acquiring the thermal energy $\Delta E_{\text{mig}}^i$ necessary to jump across the diffusion barrier to a nearest-neighbour site at temperature $T$ is proportional to

$$P_{\text{mig}}^i (p, T) \propto \exp \left( -\frac{\Delta G_{\text{mig}}^i (p, T)}{k_B T} \right).$$ (3.53)

During an elementary jump, the change in the $pV$ contribution to the Gibbs energy is negligible, so that we approximate $\Delta G_{\text{mig}}^i (p, T)$ by the zero-temperature DFT total energy $\Delta E_{\text{mig}}^i$. This approximation neglects vibrational contributions to the migration barrier, which are certainly not negligible at high temperatures. Unfortunately, these contributions cannot be easily determined using DFPT due to the fine tolerances on forces required between atoms while the defect is fixed at the top of the migration barrier. When these tolerances cannot be rigorously enforced, the interatomic force matrix used in DFPT calculations is not converged and leads to a slew of negative phonon frequencies. Hopefully, future work can address this weakness using other methods, such as ab initio molecular dynamics simulations, that exceeded current computational resources.

Diffusion barriers can be calculated using DFT or other methods by making step-by-step moves of a defect to a nearest-neighbour equivalent site. The moving defect (or, in the case of a vacancy, the atom that exchanges positions with the vacancy) is allowed to relax in the plane perpendicular to the vector joining its start and end points, while the position of the plane is moved along the vector with every step. All atoms in the computational supercell should be relaxed at each step, as significant lattice relaxations can occur in the vicinity of a migrating defect. The rate of diffusion $D_i$ of a defect of species $i$ along one of $z$ equivalent paths to nearest-neighbour sites is proportional to the product $z P_f^i P_{\text{mig}}^i$, such that

$$D_i (p, T) = D_{i,0} \exp \left( -\frac{\Delta G_i^f (p, T) + \Delta E_{\text{mig}}^i}{k_B T} \right) = D_{i,0} \exp \left( -\frac{Q_{\text{act}}^i (p, T)}{k_B T} \right),$$ (3.54)

where $Q_{\text{act}}^i (p, T)$ is called the ‘activation energy’ for diffusion, and $D_{i,0}$ (units $m^2/s$) is a pre-exponential factor. Over a wide range of temperatures, experimentally mea-
SURED DIFFUSION COEFFICIENTS $D_i(p,T)$ OFTEN FIT THE ARRHENIUS RELATIONSHIP IN EQ. 3.54 AND CAN BE PLOTTED AS $\ln[D_i(p,T)]$ AGAINST $1/T$. AS $Q_{i,act}(p,T)$ DEPENDS ON $T$, THE GRADIENT OF SUCH AN ARRHENIUS PLOT CAN BE SHOWN (SEE APPENDIX A.1) TO YIELD THE ACTIVATION ENTHALPY $-H_{i,act}(p,T)/k_B$.

Both $D_{i,0}$ AND $Q_{i,act}(p,T)$ DEPEND ON THE DEFECT SPECIES $i$ AND ITS DIFFUSION MECHANISM. WE WILL NOW EXAMINE THE FORM OF THE PRE-EXPONENTIAL FACTOR $D_{i,0}$ USING KINETIC THEORY.

3.6.3 Fick’s laws

The previous section described how a macroscopic diffusion process in a crystal can be experimentally measured and obeys an Arrhenius relationship. The activation energy $Q_{i,act}(p,T)$ for this process comprises two microscopic quantities, the Gibbs energy of formation and the diffusion barrier, that can both be determined using $ab$ initio methods. In the following sections, Fick’s laws will be introduced and the experimental pre-exponential diffusion factor $D_{i,0}$ deduced by recasting the macroscopic diffusion problem in terms of an atomistic random walk. Knowledge of the properties of this walk, coupled with the Gibbs energies of formation and migration, can be used to determine the diffusion coefficient measured in experiment.

Point defects can perform a series of jumps through a crystal by any of the mechanisms described in Sec. 3.6.1. The mechanism of choice differs between defect species and depends on a number of factors, such as the local geometry and electrostatics. In general, intuition can eliminate many unlikely possibilities, but the rest must be analysed as described in the previous section. Defects then repeatedly jump in the direction of the lowest diffusion barrier, creating a random walk.

Diffusion processes are governed by Fick’s laws, that can be derived from elementary random walks that correspond to simple, but real systems. For example, diffusion along the axes of a cubic crystal can be described in terms of a one-dimensional random walk with one type of diffusing atomic species and a single jump distance along each axis. Consider two adjacent lattice planes separated by $\lambda$ normal to the axis of diffusion, as in Fig. 3.5. Plane 1 contains $n_1$ atoms of the diffusing species per unit area, plane 2 $n_2$ atoms. Let $\Gamma_{12}$ be the jump frequency for an atom on a given site on plane 1 to a neighbouring site on plane 2. Then the number of atoms per unit area and time jumping from plane 1 to plane 2 is $j_{12} = n_1\Gamma_{12}$. Similarly, the flux $j_{21}$ of atoms jumping the other way, from plane 2 to plane 1, is $j_{21} = n_2\Gamma_{21}$. Diffusion fluxes are generally the result of concentration gradients,
the result of which is atomic intermixing \[60\]. The imbalance between opposing fluxes from neighbouring regions manifests itself as a net diffusion current acting in response to a concentration gradient. In lieu of driving forces\(^6\) such as an electric field, the net atom flux \(J\) between planes 1 and 2 is

\[
J = j_{12} - j_{21} = n_1 \Gamma_{12} - n_2 \Gamma_{21}.
\] (3.55)

As the jump frequency is independent of direction, \(\Gamma_{12} = \Gamma_{21} = \Gamma_0\), it follows that

\[
J = (n_1 - n_2) \Gamma_0.
\] (3.56)

In the continuum limit where \(\lambda \to 0\), the concentration difference \(n_1 - n_2\) can be rewritten in terms of the concentration gradient and the separation between neighbouring planes,

\[
J = -\lambda \frac{\partial n}{\partial x} \Gamma_0.
\] (3.57)

As the concentration per unit area \(n\) cannot generally be related to experimental quantities, we now introduce the concentration per unit volume, \(c = n/\lambda\), giving

\[
J = -\lambda^2 \frac{\partial c}{\partial x} \Gamma_0 = -D \frac{\partial c}{\partial x},
\]

which is Fick’s first law for diffusion in a single-component system\(^7\). The diffusion coefficient \(D\) contains all the physical knowledge about the diffusing species in the system and is well-defined by experiment. This derivation applies for the specific case where the diffusion direction is along \(x\), so that all concentrations and fluxes have the same value at any position on a given plane perpendicular to the \(x\)-axis. In a real crystal, diffusion is a three-dimensional process, and the directions of the fluxes, forces and gradients can depend on position. Fick’s laws can be readily extended to

\(^6\)Driving forces are those that render the jump frequency dependent on the direction of the jump.

\(^7\)Generally, for a multi-component system, the diffusion coefficient will depend on the concentration gradients of the various components. Then, Fick’s first law can be expressed as \(J_i = -\sum_k D_{ik} (\partial c_k / \partial x)\), where \(J_i\), the flux of species \(i\), may depend on the concentration gradient \((\partial c_k / \partial x)\) of every species \(k\) in the crystal. The ‘partial diffusion coefficients’ \(D_{ik}\) act as weights for the contribution of the concentration gradient of species \(k\) to the total flux \(J_i\) of species \(i\).
three dimensions, but most experiments measure diffusion by determining $D$ along a specific diffusion direction, so a study of planar diffusion suffices to bridge theory and experiment. The macroscopic diffusion coefficient $D$ can be determined from the microscopic geometry of the system in question, as we will see in Sec. 3.6.4.

Fick’s second law can be found from the rate of change of the concentration at plane 1. This can be calculated by considering the atoms which jump between plane 1 and its neighbouring planes 0 and 2, as in Fig. 3.5. This local mixing gives rise to diffusion and allows the global defect population in the crystal to equilibrate [60]. Each atom that arrives at plane 1 adds one atom to the concentration on the plane, and each atom that leaves subtracts one. The change in concentration $\Delta n$ during a small time period $\Delta t$ is then

$$\Delta n = \Delta t (n_0 \Gamma_{01} - n_1 \Gamma_{12} - n_1 \Gamma_{10} + n_2 \Gamma_{21}),$$

(3.58)

where the symbols take their usual meanings. Assuming that the jump frequency is independent of direction, such that $\Gamma_0 = \Gamma_{01} = \Gamma_{12} = \Gamma_{10} = \Gamma_{21}$, we find that

$$\frac{\Delta n}{\Delta t} = \Gamma_0 (n_0 - n_1) - \Gamma_0 (n_1 - n_2).$$

(3.59)

Recasting the concentration differences in terms of the concentration gradients and the separation between neighbouring planes $\lambda$ yields

$$\frac{\Delta n}{\Delta t} = \Gamma_0 \lambda \left( \frac{\partial n_{12}}{\partial x} - \frac{\partial n_{01}}{\partial x} \right),$$

(3.60)
where $\partial n_{12}/\partial x$ is the concentration gradient between planes 1 and 2, and similarly for $\partial n_{01}/\partial x$. In the continuum limit and when $\Delta t \to 0$,

$$\frac{\partial n}{\partial t} = \Gamma_0 \lambda^2 \frac{\partial^2 n}{\partial x^2}. \quad (3.61)$$

Finally, substituting $n$ for the concentration per unit volume, $c = n/\lambda$,

$$\frac{\partial c}{\partial t} = \Gamma_0 \lambda^2 \frac{\partial^2 c}{\partial x^2}, \quad (3.62)$$

and substituting $D = \Gamma_0 \lambda^2$ as for Fick’s first law, we have

$$\frac{\partial c}{\partial t} = D \frac{\partial^2 c}{\partial x^2}, \quad (3.63)$$

which is Fick’s second law for a constant diffusion coefficient $D$.

### 3.6.4 Diffusion coefficients from random walk diffusion

Fick’s laws were derived from a simple one-dimensional random walk with a single jump distance. The derivation applies to diffusion along the $\langle 100 \rangle$ direction in a simple cubic crystal, but not necessarily to other structures and crystal orientations. A general expression for the diffusion coefficient in any structure can be found by accommodating jumps that lie at an angle to the diffusion direction. Such jumps have different projected $x$-components and therefore contribute different amounts to the diffusion flux.

For an actual diffusion process in a crystal during a finite time $\tau$, a defect $i$ performs a number of discrete jumps. Fick’s first law then defines the diffusion coefficient for the defect $i$ as

$$D_i = -\lim_{J_i \to 0} \frac{J_i}{\nabla c_i}. \quad (3.64)$$

The diffusion coefficient $D_i$ can be determined from a continuum diffusion model, or, more relevant for our purposes, from kinetic theory. A general kinetic expression for $D_i$ can be derived from a density function $W_i(X, \tau)$ that gives the probability that a defect at position $x$ at time $t$ will be at $x + X$ at time $t + \tau$. In a homogeneous system with no driving forces $^8$, $W_i(X, \tau)$ does not depend on $x$ or $t$ and $D_i$ is given

---

$^8$The derivation can be easily extended to the more general case involving driving forces or inhomogeneous concentration gradients. The interested reader is referred to the more involved
by the Einstein equation [50]

\[ D_i = \lim_{\tau \to \text{small}} \frac{\langle X^2 \rangle}{2\tau}. \] (3.65)

A derivation of Eq. 3.65 is presented in Appendix A.2. The Einstein equation can be evaluated using kinetic theory, as detailed in Appendix A.3, to find

\[ D_i = \frac{1}{2} \sum_{j=1}^{z} \Gamma_{i,j} x_j^2, \] (3.66)

where \( \Gamma_{i,j} \) is the total jump frequency of a defect \( i \) to its \( j \)th (of \( z \) in total) nearest-neighbour, and \( x_j \) is the displacement along the diffusion direction \( x \) for such a jump. The diffusion coefficient in Eq. 3.66, although derived for one-dimensional diffusion with a concentration gradient only along the \( x \)-axis, is generally valid for diffusion in cubic crystals or along principal axes in noncubic crystals, regardless of the concentration gradients present [116]. In a real crystal, such as \( \text{Al}_2\text{O}_3 \), many jumps do not lie perfectly in the diffusion direction \( x \), but rather at an angle. The contribution \( x_j \) to the diffusion flux along the \( x \)-axis per jump can then be written as the projection onto the \( x \)-axis of a jump of length \( \lambda_i \) made at an angle \( \phi_{i,j} \) to the \( x \)-axis. Since all nearest-neighbour jumps over the lowest diffusion barrier have the same jump distance \( \lambda_i \) and jump frequency \( \Gamma_{i,0} \), the diffusion coefficient in Eq. 3.66 reduces to the expression

\[ D_i = \frac{1}{2} \Gamma_{i,0} \lambda_i^2 \sum_{j=1}^{z} \cos^2 \phi_{i,j}, \] (3.67)

which is valid for all crystals along a specific diffusion direction. The total jump frequency \( \Gamma_{i,0} \) contains much of the physics of the actual diffusion process, including the temperature dependence of the diffusion coefficient, and can be written as

\[ \Gamma_{i,0} \approx \nu_i \exp \left(-\frac{Q_i^{\text{act}}}{k_B T}\right), \] (3.68)

where \( \nu_i \) is the attempt frequency and \( Q_i^{\text{act}} \) the activation energy of the defect \( i \). The main contributor to \( \nu_i \) is the frequency at which the thermal vibrational motion of the defect pushes it in the direction of the barrier to be surmounted. The value of \( \nu_i \) can be estimated from the peak of the local phonon density of states of a defect [discussion in Ref. [116]].
species, which can be calculated in the harmonic approximation in DFPT.

### 3.6.5 Correlation effects

While examining tracer diffusion experiments in 1952, Bardeen and Herring [10, 9] realised that the idealised model in which the jump probabilities do not depend on the directions of previous jumps does not hold in real crystals. The defects are said to perform a *correlated walk*. Consequently, the standard diffusion equation does not apply and must be modified by introducing a correlation factor $f_i$.

Consider a system where a defect $i$ first makes a forward jump and then a correlated reverse jump so that the system returns to its initial state. As the two jumps cancel, neither of them is effective in causing random diffusion, and the effective frequency of random jumps $\Gamma_{i,0}^{\text{eff}}$ can then be expressed as

$$\Gamma_{i,0}^{\text{eff}} = f_i \Gamma_{i,0},$$

(3.69)

where $\Gamma_{i,0}$ is the actual jump frequency and $f_i$ the correlation factor. The value of $f_i$, which lies in the range $[0, 1]$, describes the fraction of jumps that are effective in causing random diffusion. A correlation factor near zero corresponds to a large correlation effect, and one near unity to a small correlation effect between successive jumps.

The difference between $\Gamma_{i,0}$ and $\Gamma_{i,0}^{\text{eff}}$ plays a significant role in the interpretation of experimental data, which thus offers a means of measuring the correlation factor. For example, the isotope effect in diffusion provides a useful experiment for measuring $f_i$ [157]. By classical rate theory [66], two isotopes will have different vibrational, and therefore also jump, frequencies due to their differing masses. The difference in jump frequencies is reflected in a change in the correlation factors and diffusion coefficients.

Correlation effects arise when the jump probabilities at a site depend on the directions of previous jumps, which is always the case when a jump requires a lattice imperfection at a neighbouring site, such as in the vacancy and interstitialcy mechanisms. These cases are described schematically in Figs. 3.6 and 3.7.

For diffusion via a vacancy mechanism, we can consider the motion of a distinguishable tracer atom $M$, that is not able to jump until a vacancy arrives at a neighbouring site. When no driving forces$^9$ are present, the direction in which a

---

$^9$A driving force constitutes any influence that renders the jump frequency between two sites direction dependent.
Section 3.6: Diffusion kinetics of defects

Figure 3.6: Vacancy diffusion mechanism in a schematic 2D lattice. A tracer atom (indicated by the highlighted circle) waits for a vacancy to arrive at a neighbouring site (either B, C, D, or E in the top frame). When a vacancy arrives at B (middle frame), the tracer atom and vacancy can jump in opposite directions. After such a jump, the tracer atom is more likely to perform a reverse jump back into its original position than a jump in another direction (either F, G, or H in the bottom frame).

Vacancy hops is random. As the vacancy can approach M from any direction, the initial atomic jump is in a random direction. After M has performed a jump into a specific neighbouring vacant site, however, the configuration of the atom-vacancy pair is no longer random. Immediately following the exchange, M is more likely to perform a reverse jump and return to the previous configuration. Therefore, the second jump of M is no longer in a random direction, but rather correlated due to the change in local symmetry induced by the initial jump.

In an interstitialcy mechanism, atoms migrate by repeatedly jumping from a lattice site to an interstitial site, and then from the interstitial site to another lattice site. If the tracer atom M initially sits on a normal lattice site, an interstitial can arrive randomly at any neighbouring interstitial site so that M performs an initial jump in a random direction. After M jumps into a specific interstitial site $i_1$ (site B in Fig. 3.7) and the lattice site it jumped from is occupied by the migrating interstitial atom, M can jump to any neighbouring, occupied, lattice site $n$ with equal probability due to the local symmetry. When M performs a jump to one of the lattice sites, the atom sitting on this site jumps to a new interstitial site $i_2$ (site J in Fig. 3.7), so that the interstitialcy itself has moved from $i_1$ to $i_2$. After this jump, M has a greater than
random probability of jumping back into the interstitial site it came from, $i_1$, due to the asymmetric environment of the tracer atom. This introduces a correlation between a given lattice-to-interstitial jump and the preceding interstitial-to-lattice jump. As in the vacancy mechanism, the change in local symmetry due to the jump $M$ makes from an interstitial to a lattice site introduces correlation.

In impure crystals, these arguments become more complicated, in particular as another correlation effect arises between successive jumps if the jump frequency depends on the species of the participating lattice atom. However, these effects will not be addressed here.

Atoms that diffuse via an interstitial mechanism in pure crystals exhibit no correlations since the local symmetry does not depend on the direction of the previous

**Figure 3.7:** Interstitialcy diffusion mechanism in a schematic 2D lattice.
jump, but is always identical. This holds true in the dilute limit when the concentration of interstitial defects is small.

Correlation effects for divacancy diffusion are more difficult to determine, and depend crucially on the mechanism by which the divacancy migrates. For example, if it moves as a rigid dumbbell, the correlations can be examined in a similar manner to the interstitialcy mechanism.

We will follow the methodology in Refs. [36, 4, 116] closely. The correlation factor $f_i$ for a defect $i$ is defined by the relation

$$D_{\text{actual}}^i = f_i D_{\text{random}}^i,$$

(3.70)

where $D_{\text{actual}}^i$ is the measured tracer diffusion coefficient where the defect follows a correlated walk, and $D_{\text{random}}^i$ is the tracer diffusion coefficient when the defect performs the same number of jumps per unit time but follows a random walk. In random walk diffusion (neglecting correlations), we have

$$D_{\text{random}}^i = \frac{1}{2} \Gamma_{i,0} \lambda_i^2 \sum_{j=1}^{z} \cos^2 \varphi_{i,j},$$

(3.71)

from Eq. 3.67, where the sum is taken over the $z$ nearest neighbours with the lowest diffusion barrier, $\varphi_{i,j}$ is the angle between the diffusion axis and the vector joining the sites $i$ and $j$ between which the defect jumps, $\lambda_i$ the jump distance, and $\Gamma_{i,0}$ the jump frequency. Incorporating the correlation factor modifies the random walk diffusion coefficient to [116]

$$D_{\text{actual}}^i = \frac{1}{2} f_i \nu_i \exp \left(-\frac{Q_i^{\text{act}}}{k_B T} \right) \lambda_i^2 \sum_{j=1}^{z} \cos^2 \varphi_{i,j},$$

(3.72)

where we have substituted $\Gamma_{i,0}^{\text{eff}} = f_i \nu_i \exp \left(-\frac{Q_i^{\text{act}}}{k_B T} \right)$ from Eqs. 3.69 and 3.68. This is the expression that will be used in conjunction with DFT calculations of formation energies and diffusion barriers to calculate the diffusion coefficient we wish to compare with experiment.

The value of $f_i$ depends on both the diffusion mechanism of the defect and the geometry of the crystal. Expressions for $f_i$ are derived in Appendix A.4 for the native defects in alumina, as presented in Sec. 4.4.2. For vacancy diffusion in alumina, we find that

$$f_{\text{VO,VA}} = f = \frac{1 + \langle \cos \theta_{\text{VO,VA}} \rangle}{1 - \langle \cos \theta_{\text{VO,VA}} \rangle},$$

(3.73)
Chapter 3: Point defects in solids

where $\theta_{V_O,V_{Al}}$ refers to the angle between any two successive defect jumps for either $V_O$ or $V_{Al}$ defects. For interstitials in alumina, that migrate by interstitialcy mechanisms, we have that

$$f_{O_i,Al_i} = 1 + \langle \cos \theta'_{O_i,Al_i} \rangle,$$

(3.74)

where $\theta'_{O_i,Al_i}$ is the angle between the jump vector for an interstitial to lattice jump and the jump vector for the succeeding lattice to interstitial jump.

The angles in Eqs. 3.73 and 3.74 can be found from the average angle between successive defect jumps:

$$\langle \cos \theta \rangle = \sum_j T_j \cos \theta_j,$$

(3.75)

where $T_j$ is the probability that the defect will take successive jumps with an angle $\theta_j$ between the directions of the two jumps, and the sum contains all jump directions.

As shown in Sec. 4.4.8, an oxygen vacancy in alumina diffuses via a simple vacancy mechanism, and the migration of an aluminium vacancy defect can be viewed as a vacancy mechanism accompanied by two lattice relaxations at the beginning and end points of each jump. In contrast, aluminium interstitials migrate via an interstitialcy mechanism, and oxygen interstitials by a dumbbell interstitialcy mechanism.

In order to determine the correlation factors for defect migration in alumina, we examine the geometry surrounding a single vacancy or interstitial defect in the largest simulation cell feasible in a DFT calculation. This analysis considers the effect of nearest-neighbour relaxations, which change the angles between successive defect jumps from those in a perfect unrelaxed geometry. Then, taking averages over all the possible successive jumps along the migration path with the lowest energy barrier (paths explained in detail in Sec. 4.4.8) and using Eqs. 3.73 and 3.74 gives correlation factors of $f_{V_O} \approx 0.47$ for oxygen vacancies, $f_{V_{Al}} \approx 0.36$ for aluminium vacancies, $f_{Al_i} \approx 0.95$ for aluminium interstitials, and $f_{O_i} = 1.00$ for oxygen interstitials. These compare favourably to calculations for the perfect hexagonal close-packed lattice, as well as previous explicit calculations of the correlation factor for aluminium vacancy diffusion in corundum. In an ideal hcp structure, diffusion by a vacancy mechanism yields a correlation factor of 0.78 (both normal and parallel to the $c$-axis) [36, 131, 116, 4]. The deviation from the ideal value found in alumina has two causes. First, only 4, rather than all, of the 12 nearest-neighbour oxygen atoms actually participate in vacancy diffusion, as their migration barrier is lower than that for the other nearest-neighbours able to induce 3D diffusion. Secondly,
the calculations for vacancy diffusion in an ideal hcp lattice did not account for the
distorted geometry surrounding a vacant site. For comparison, the correlation fac-
tors for vacancy diffusion in simple, body-centred and face-centred cubic structures
are 0.65, 0.73, and 0.78, respectively. Compaan and Haven [36] determined $f_{\text{VAl}}$ for
aluminium vacancy diffusion in the basal plane of corundum to be 0.33, where the
difference to our calculated value arises from consideration of the split vacancy con-
figuration (shown in Sec. 4.4.2) and the distorted local geometry around the defect.
Similarly, our calculation for aluminium interstitials migrating by an interstitialcy
mechanism compares well with typical values for interstitialcy mechanisms, which
range from $2/3$ to, more commonly, around unity [37, 116].

For more complex structures, in particular those where diffusion along different
paths is equally likely, or for many cases of impurity diffusion, the angles between
successive jumps by a defect can be found using a number of involved methods. The
most direct approach, first used by Bardeen and Herring [9, 10] involved enumerating
the possible paths for a migrating defect in order to generate and recursively solve
a set of linear equations describing the probabilities of each of these paths occuring.
This problem was then elegantly recast into matrix algebra by Mullen [131], LeClaire
and Lidiard [100], and others. The matrix formulation allowed for a number of
simplifications arising from symmetry considerations to accelerate computation, but,
as with Bardeen and Herring’s original solution, the accuracy is limited by the size
of the matrix used to describe the transition probabilities, corresponding to a limit
on the number of possible transition sites in the lattice. The larger the lattice, the
more accurate the calculation of $f$. Compaan and Haven [36, 37] translated the
crystal lattice into an electrical network, so that the flow of defects between lattice
sites corresponds to the flow of electrical charge in a network of resistors. Symmetry
considerations allowed them to reduce the size of the networks dramatically, again
lowering the computational time and allowing for larger lattices to be considered.
The advent of combinatorial methods, first used in this context by Schoen and
Lowen [158], eventually allowed for the exact calculation of correlation factors in
many simple lattices without restriction on lattice sizes.

3.6.6 Experimental methods

Diffusion coefficients can be experimentally measured using an assortment of meth-
ods, where the method of choice often depends on the material in question. In
general, mass transport in solids is mediated by vacancies: as vacancies diffuse in one direction, atoms diffuse in the opposite direction. Most diffusion experiments aim to determine the diffusion coefficient of vacancy diffusion in a homogeneous crystal by measuring the rate at which atoms travel in the opposite direction.

Radioactive tracers

For materials whose constituent atomic species have a convenient radioactive isotope, diffusion coefficients are most commonly measured from the diffusion of a dilute radiotracer. The net diffusive flow of the tracers is then solely governed by the concentration gradient of the tracer. Tracers can be introduced to an experimental sample in numerous ways, such as electroplating, vapour or liquid deposition, or high-energy proton bombardment. Evaporation of the tracer from the surface of the sample and other problems with boundary conditions can be avoided by placing the surfaces of two samples, one enriched with the radiotracer, one not, in contact to form a diffusion couple. The prepared sample is then heated to annealing temperatures at which the tracer can acquire the thermal energy necessary to perform diffusive jumps through the solid, either in oxygen gas or a vacuum. After the diffusion anneal, the sample is sectioned perpendicular to the principal diffusion axis (given by the tracer concentration gradient), and the residual tracer concentration or activity in each section measured. The diffusion coefficient $D$ can be found from the solution of Fick’s second law

$$\frac{c}{c_0} = \frac{1}{2} \text{erfc} \left[ \frac{x}{2\sqrt{Dt}} \right],$$

where $c$ is the concentration of the tracer at any point after a time $t$, $c_0$ the initial concentration in the radioactive material, and $x$ the distance in the diffusion direction. For a diffusion couple, the diffusion coefficient can be calculated from the tracer concentration that has crossed the interface separating the two samples.

Ion beam microprobes

Secondary ion mass spectrometry (SIMS) and nuclear reaction analysis (NRA) can be used to determine diffusion coefficients by probing the depth profile resulting from a diffusion anneal. SIMS focusses energetic beams of heavy ions to sputter ions from the surface of a sample for mass analysis. This can be used to determine
the isotopic composition of a sample with depth, and therefore, after a prior diffusion anneal with a given isotope, the diffusion profile. The advantage of using SIMS over other methods is that low concentrations of low activity radiotracers, or even non-radioactive isotopes, can be used in the anneal, and that the technique lends itself to studying anisotropy in \( D \) for non-cubic single crystals.

In NRA, an incoming ion beam enters through the surface of a thin film, losing energy through electronic stopping. At a certain depth, the incoming particles undergo a nuclear reaction with atoms of the film, where the energy of the reaction products is a function of the energy of the incoming ions. The depth distribution of the atoms, such as those from a diffusion anneal, that react with the incoming ions can then be deduced from the measured energy distribution of the reaction products.

**Dislocation loops**

When appropriate tracers are hard to come by or have an inconvenient half-life, diffusion coefficients of the rate-controlling (slower) species in a material can be determined from the shrinkage rates of prismatic\(^{10}\) dislocation loops. These small dislocation loops are isolated in the bulk of a material and can be created when the dislocation substructure of a material is deformed at high temperature. For example, loops form in \( \text{Al}_2\text{O}_3 \) from edge-trapped dislocation dipoles created by basal slip deformation. The loops can be analysed using methods such as transmission electron microscopy (TEM) in order to characterise their shape, orientation and size, and are either of vacancy or interstitial nature. Under annealing, transport of vacancies from the former, interstitials from the latter, causes the loops to shrink and, eventually, annihilate. The sizes of loops between subsequent annealings, and thus the shrinkage rate, can be measured using TEM to yield the diffusion coefficient of the rate-controlling species involved in the loop annihilation process. For further information on dislocation loop annealing and its applicability to metal oxides, Ref. [99] offers details of such experiments in alumina.

\(^{10}\)In prismatic dislocation loops, the Burgers vectors do not lie in the plane of the loops.
“Science walks forward on two feet, namely theory and experiment.”

- Robert Andrews Millikan

## Point defects and diffusion in Al₂O₃

### Contents

| 4.1 | Introduction | 98 |
| 4.2 | Experimental background | 98 |
| 4.2.1 | Oxygen diffusion | 99 |
| 4.2.2 | Oxygen diffusion buffering in doped alumina | 102 |
| 4.2.3 | Aluminium diffusion | 105 |
| 4.2.4 | Anisotropy | 107 |
| 4.2.5 | Alternative species | 108 |
| 4.3 | Previous theoretical work | 109 |
| 4.4 | Density functional theory calculations | 115 |
| 4.4.1 | Geometry and basis set convergence | 115 |
| 4.4.2 | Defect species and complexes | 122 |
| 4.4.3 | Defect formation energies and cluster binding energies | 145 |
| 4.4.4 | Chemical potentials | 152 |
| 4.4.5 | Band gap corrections | 154 |
| 4.4.6 | Finite-size scaling | 156 |
| 4.4.7 | Vibrational contributions | 160 |
| 4.4.8 | Diffusion barriers | 162 |
| 4.5 | Defect concentrations | 179 |
| 4.6 | Diffusion coefficients | 190 |
| 4.7 | Impurity solubilities | 194 |
| 4.8 | Quantum Monte Carlo calculations | 198 |
| 4.8.1 | Trial wave functions | 199 |
| 4.8.2 | Jastrow optimisation | 200 |
| 4.8.3 | QMC band gaps | 202 |
| 4.8.4 | Errors | 203 |
| 4.8.5 | Diffusion barriers | 206 |
4.1 Introduction

The defect chemistry of $\text{Al}_2\text{O}_3$, in particular when doped with impurities, has perplexed experimentalists and theorists alike for decades. $\text{Al}_2\text{O}_3$, or alumina, is arguably the most widely used structural ceramic, and diffusion in its bulk, grain boundaries and dislocations is crucial to understanding a host of high temperature properties, such as the plastic deformation of single crystals, grain growth and sintering of polycrystals, and the oxidation of aluminium-containing alloys, that create alumina scales providing good oxidation resistance in oxidising environments. Although the diffusion of oxygen in alumina has been experimentally and theoretically studied for nearly half a century, theory has been unable to explain many of the phenomena experiment has revealed. In particular, the oxygen diffusion coefficient in alumina doped with aliovalent impurities exhibits a remarkable insensitivity to the concentration of impurities. The inability of theory to explain this effect, and its clear underestimation of defect populations in alumina, is referred to as the ‘corundum conundrum’ [80]. We attempt to resolve this conundrum using a complete and careful $ab\ initio$ approach in order to understand the mechanisms that govern diffusion processes in alumina. This $ab\ initio$ approach begins with understanding the ground-state configurations of defects and their migration paths. In Sec. 4.4.2, it will be shown that previous studies of defect formation energies did not identify the most stable configurations of either the oxygen interstitial or the aluminium vacancy. In addition, we find new lowest-energy diffusion mechanisms for all defects but the oxygen vacancy, and show the importance of extrapolating both formation energies and diffusion barriers to the dilute limit. While the present work is unable to quantitatively reconcile experimental diffusion coefficients with theory, it does hypothesise a qualitative explanation of the corundum conundrum that can be examined with further work using more accurate techniques. Although the work presented here focuses on alumina, the theory introduced in Chapter 3 is widely applicable, and can be used to elucidate details of the defect chemistry in a number of other metal oxides.

4.2 Experimental background

The diffusion of aluminium and oxygen in $\text{Al}_2\text{O}_3$ has been continuously explored using various experimental techniques since the 1960s. The most reliable experi-
mental results are for the diffusion of aluminium, oxygen, and impurities in bulk single crystals (sapphire). This study focuses on the diffusion coefficients and activation energies for bulk self-diffusion of aluminium and oxygen only, as these are the most important defects involved in the transport of mass through alumina, and thereby govern mass diffusion properties.

Despite numerous experimental determinations of the bulk oxygen diffusion coefficient $D_O$ using a range of different techniques, and a growing number of measurements of the bulk aluminium diffusion coefficient $D_{Al}$, many experimental results are still ill-understood. In particular, the order of magnitude of $D_{Al}/D_O$ is under debate, the origin of the spread of an order of magnitude between different measurements of $D_O$ is uncertain, and the reason for anisotropic lattice expansion along the $c$-axis in titanium-doped alumina is unknown. Most importantly, the nature of the ‘buffering’ that reduces the change in $D_O$ to only 1-2 orders of magnitude even upon heavy aliovalent doping is not understood.

The following sections summarise experimental results on oxygen and aluminium diffusion in the bulk corundum polymorph, $\alpha$-$\text{Al}_2\text{O}_3$. Diffusion along grain boundaries and dislocations in $\text{Al}_2\text{O}_3$ polycrystals will not be discussed, although this provides many more puzzles and should be the subject of further work.

### 4.2.1 Oxygen diffusion

The low absolute diffusivity of oxygen and effects of surface damage in diffusion samples render experimental determinations of the oxygen diffusion coefficient $D_O$ in bulk $\text{Al}_2\text{O}_3$ difficult. The resulting spread of roughly one order of magnitude in $D_O$ is shown in Fig. 4.1, which is an Arrhenius plot including experimental results from a number of different studies. The mean of the measured activation energies in undoped alumina, tabulated in Table 4.1, is 6.54eV, with a standard deviation of 0.74eV.

Oxygen diffusion in undoped $\text{Al}_2\text{O}_3$ was first measured in 1960 by Oishi and Kingery [139], who introduced an $^{18}$O radiotracer to a sample under gaseous exchange in an $^{18}$O-enriched atmosphere. They annealed Verneuil-grown\(^1\) single crystal spheres, crushed Verneuil rods, and crushed polycrystals which had been sintered at 1900°C. After the $^{18}$O diffusion anneal, the amount of $^{18}$O exchange was determined from

---

\(^1\)A process involving melting powders under an oxyhydrogen flame and crystallising the melted droplets into a boule.
Table 4.1: Activation energies $Q_{i}^{\text{act}}$ for diffusion in undoped (U), Mg-, and Ti-doped alumina. In the study described in Ref. [138], the results from Ref. [139] were corrected, but this did not affect the activation energy.

mass spectrometric measurements of gas samples taken during the exchange, or by reducing the exchanged oxide to CO by heating with graphite to 1700°C, catalytically converting the CO to CO$_2$, and determining the $^{48}$CO$_2$/^{44}$CO$_2$ ratio.

Oishi extended these measurements to two sets of single crystals with other collaborators [137, 138] using the oxide reduction/mass spectroscopy technique, obtaining tolerable agreement with the initial measurements. The first set contained crushed particles and discs cut from a rod, both from a high purity CVD-grown$^2$ material [137]. The second set comprised Verneuil-grown discs that were buff polished with diamond paste [137] or chemically polished and Ar-ion-milled [138]. Polishing and milling reduces surface roughness and damage, and Oishi et al. found much lower values of $D_O$ for these smoothed samples. They deduced that low-temperature measurements, such as those in the initial experiments by Oishi and Kingery at temperatures below 1650°C, are unreliable and likely to exhibit enhanced diffusion along cracks underneath the sample surface. They also corrected their previous results on crushed particles with improved measurements of the dimensions of the particles, finding good agreement with the diffusion coefficients of the smoothed samples.

After the first swathe of experiments, further measurements [153, 30, 148, 133] on

---

$^2$Chemical vapour deposition (CVD) is a process by which a substrate is exposed to volatile precursors that react and/or decompose on the substrate surface to produce a desired deposit.
undoped Al$_2$O$_3$ single crystals using SIMS and NRA have cemented the value of $D_O$. For the SIMS measurements, Prot and Monty [148], Cawley [29], and Cawley, Halloran, and Cooper [30] all introduced the $^{18}$O tracer under gaseous exchange with $^{18}$O-enriched oxygen; and the latter two studies performed the diffusion anneals in a vacuum of $p_{O_2} < 10^{-10}$ atm, rather than in air or oxygen. Reed and Wuensch [153] also annealed in a vacuum ($p_{O_2} < 1.28 \times 10^{-8}$ atm), but used a diffusion couple involving a film of Al$_2$$^{18}$O$_3$ deposited on Al$_2$O$_3$ single crystals rather than gaseous exchange. All found diffusion coefficients close to those seen in previous experiments at standard pressure, indicating only a small $p_{O_2}$ dependence of $D_O$. Cawley, Halloran, and Cooper [30] and Reddy and Cooper [151] both performed NRA measurements using the $^{18}$O(p, $\alpha$)$^{15}$N reaction to determine $^{18}$O depth profiles resulting from gaseous exchange in the same laboratory, finding $D_O$ to within experimental error of each other, and in line with other measurements. Further investigations by Reddy [151, 150] on the exact same samples used by Reed [152, 153] showed that any differences in $D_O$ must have resulted from differences in the methods of forming the profiles [47], with Reed forging a diffusion couple so that the sample surfaces were not in contact with the surrounding atmosphere during annealing, in contrast to Reddy.

As indicated by Oishi et al.’s revised studies [137, 138], the effects of surface damage in samples on $D_O$ are so pronounced that short circuit diffusion along cracks and dislocations can increase $D_O$ by an order of magnitude, an effect visible as ‘tails’ in the $^{18}$O penetration profile. Mechanical polishing, performed on many samples, can create damage in a thin ($\leq 1 \mu$m) layer beneath the surface, although this can be removed by lengthy annealing prior to diffusion measurements. Many of the later studies [151, 30, 148, 133] annealed their samples prior to introducing the tracer at temperatures and lengths comparable to the actual exchange; however, in all the tracer studies, the diffusion depths were in the range $0.1 - 0.6 \mu$m, within the layer that had been damaged by mechanical polishing.

Problems with surface damage do not affect loop annealing studies such as those performed by Lagerlöf, Mitchell, and Heuer [99], as measurements are made from the bulk of a sample at low temperature. These workers inferred $D_O$ from measurements of the shrinkage rates of small prismatic dislocation loops in alumina single crystals that had been previously deformed at 1400°C, and found good agreement with Prot and Monty’s SIMS data [148]. Lagerlöf et al. used TEM to measure the sizes of ten loops between six sequential annealings in the temperature range $1220^\circ$C $< T <$
1470°C of a sample foil. This procedure allowed them to determine the diffusion
coefficient of the rate-controlling species, assumed to be oxygen under all conditions.
The agreement with previous experiments strongly suggests that oxygen diffusion is
slower than aluminium diffusion.

The most recent studies by Nakagawa et al. [133] on high purity Czochralski\textsuperscript{3} and
Verneuil crystals reinforce previous findings. In addition, an ICP-AES\textsuperscript{4} analysis of
the Czochralski crystal revealed very low concentrations of impurities, with 2.8ppmw
Ti present. Other impurities, such as Ca, Mg, Na, K, Fe, Ba, Zr, Mo, W, Mn, and Cd
may have also been present in amounts below the detection limits. In comparison, a
total of 30ppm cation and 16ppm anion impurities were found in the Verneuil crystal.
These impurity concentrations are much lower than those found in earlier studies
(where explicitly stated), and the universal agreement between studies of varying
impurity content indicates that the effect of impurities on the defect chemistry may
not be as pronounced as conventional thought suggests.

Despite the experimental challenges in determining diffusion coefficients in alumina,
this considerable body of work exhibits a remarkable confluence. The very low
absolute magnitude of the oxygen diffusivity, the use of different techniques, and
the diverse origins of the single crystals all contribute to the experimental spread
of roughly one order of magnitude in $D_O$. However, the expectation that very low
concentrations of impurities, especially aliovalent dopants, should alter $D_O$ by many
orders of magnitude does not fit into this picture. The next section elaborates on
this dilemma.

4.2.2 Oxygen diffusion buffering in doped alumina

The experimental samples used for oxygen diffusion measurements in ‘undoped’
alumina actually contain measurable traces of impurities. Simplistically, any devi-
ation from a charge neutral balance between aliovalent dopants creates charged
point defects in compensating numbers. Doping alumina with large concentra-
tions (i.e. hundreds of ppm) of charged impurities would thereby create charge-

\textsuperscript{3}A process by which a seed crystal, mounted on a rod, is dipped into a molten metal or
semiconductor and pulled outwards while being rotated, in order to extract large single crystals
from the melt.

\textsuperscript{4}Inductively coupled plasma atomic emission spectroscopy (ICP-AES) is a highly sensitive type
of emission spectroscopy that uses a plasma to produce excited atoms that emit an indicative
intensity of electro-magnetic radiation at characteristic wavelengths.
Figure 4.1: Arrhenius plots of the diffusion coefficients for oxygen in undoped alumina (top) and oxygen and aluminium in both undoped and aliovalently-doped alumina (bottom). Values for undoped alumina are shown as bold lines; for Mg-doped alumina as dashed lines; and for Ti-doped alumina as double-dashed lines.
compensating defects in concentrations that would dwarf the low concentrations of intrinsic defects already present in pure alumina, changing the self-diffusion coefficients by orders of magnitude.

For example, let us assume that the detection limit of impurities is 1ppm and that an experimental sample of ‘undoped’ alumina actually has just under 1ppm Ti dopants and no Mg impurities. Mg-doping is expected to result in the creation of charge-compensating oxygen vacancies, while Ti-doping creates aluminium vacancies. Through the Schottky equilibrium, the latter leads to a depletion of the oxygen vacancy concentration. Therefore, the hypothetical sample would contain a lower concentration of oxygen vacancies than a pure sample. Intentionally doping this sample with 500ppm Mg would then require approximately 250ppm extra oxygen vacancies to ensure charge neutrality. If the intrinsic oxygen vacancy concentration is very small, say less than 1ppm, then the oxygen vacancy concentration must increase by a factor of over 250. The oxygen diffusion coefficient $D_O$ would increase by the same factor.

In contrast, experimental work on intentionally-doped alumina paints a very different picture, demonstrating much smaller changes in $D_O$ upon aliovalent doping. Instead, the point defect population controlling the oxygen diffusion coefficient appears to be strongly ‘buffered’.

The most reliable experimental determinations of self-diffusion coefficients in doped alumina are found from loop annealing studies, as these do not suffer from evaporation of dopants, such as Mg into MgO, or surface damage. Lagerlöf, Mitchell, and Heuer [99] measured the shrinkage rates of dislocation loops in 250ppm Mg-doped and 600ppm Ti-doped single crystal alumina at three temperatures each. For $1250^\circ C < T < 1350^\circ C$, they found that $D_O$ in Mg-doped alumina is ca. 50 times higher, and the activation energy $Q_{O}^{\text{act}}$ slightly lower, than in undoped alumina. In Ti-doped alumina between $1400^\circ C < T < 1500^\circ C$, they found that $D_O$ is ca. 13 times lower, and the activation energy $Q_{O}^{\text{act}}$ slightly higher, than in undoped alumina. These modest changes in $D_O$ were the largest experimental changes found by any group studying intentionally-doped alumina, but are small in comparison with those one would expect if the dopants were creating charge-compensating native point defects. This limitation of the changes in $D_O$ upon aliovalent doping is referred to as the ‘buffering’ effect [79, 99].

Another analysis of oxygen self-diffusion in aliovalently-doped alumina, Haneda and Monty’s [73] work on tracer diffusion in Mg-doped and Ti-doped Verneuil-grown
single crystals, supports these findings. They introduced an $^{18}$O radiotracer using the gas/solid exchange technique, performed diffusion anneals at standard pressure, and used SIMS to analyse the depth profiles of $^{18}$O in three samples, containing 30ppm Mg, 130ppm Mg, and 1000ppm Ti. The Mg-doped profiles were in good agreement with each other, both exhibiting an increase in $D_O$ of ca. one order of magnitude over previous measurements of $D_O$ in undoped alumina. The Ti-doped profile agreed well with an earlier determination of $D_O$ in 800ppm Ti-doped alumina by Reddy and Cooper [151], which showed a decrease in $D_O$ of ca. one order of magnitude relative to its value in undoped alumina. Another interesting phenomenon occurs in Ti-doped alumina. Philipps, Mitchell and Heuer [145] examined the precipitation of needle-shaped TiO$_2$ in alumina, and found that it is accompanied by the formation of large prismatic dislocation loops of interstitial character. Precision measurements of the lattice parameters in undoped and Ti-doped alumina reveal that the lattice expansion when doped with 1ppt Ti is $\sim$ 10 times greater along the c-axis than along the a-axis. The cause was suggested to be an interstitial defect complex containing both titanium and oxygen interstitials that elongates the c-axis. Regarding the same study, Heuer [79] also suggests that oxygen interstitial diffusion could occur via an interstitialcy mechanism, which will be examined in Sec. 4.4.8.

### 4.2.3 Aluminium diffusion

Aluminium diffusion in Al$_2$O$_3$ is even harder to measure than oxygen diffusion. The major difficulty is that the only available radiotracer, $^{26}$Al, is difficult to produce and has a long half-life of $7.2 \times 10^5$ years [80]. The resulting low activity renders traditional radiotracer counting experiments extremely challenging. Attempts to bombard alumina foils with protons to produce large amounts of $^{26}$Al fail because impurities also become irradiated and give rise to high-activity isotopes that complicate experiments. Also, one of the natural isotopes of magnesium also has a mass of 26a.u.; this rules out the use of SIMS, because magnesium impurities are often contained in samples [102].

In contrast to the extensive work on the oxygen diffusion coefficient $D_O$ in Al$_2$O$_3$, these difficulties have allowed only three measurements of the aluminium diffusion coefficient $D_{Al}$. The first work by Paladino and Kingery [140] was on undoped polycrystals of grain size 100 – 200µm; the second by Le Gall, Lesage and Bernardini
Paladino and Kingery hot-pressed a $^{26}$Al-containing Al$_2$O$_3$ powder against alumina discs to form diffusion couples, performed diffusion anneals in the temperature range 1670 – 1905°C in an oxygen gas furnace, and sectioned the samples parallel to the couple interface to determine the activity of the sections in a low-level radiation facility. They found an activation energy for aluminium diffusion, $Q_{\text{Al}}^{\text{act}}$, of 477kJ/mol, $3/4$ the value of $Q_{\text{O}}^{\text{act}}$, and a diffusivity $D_{\text{Al}}$ roughly three to four orders of magnitude higher than $D_{\text{O}}$ at 1600°C, creating the long-standing notion that $D_{\text{Al}} \gg D_{\text{O}}$.

However, it is possible that diffusion in grain boundaries rather than in the bulk was being measured [47], and, unfortunately, the low specific activity of the $^{26}$Al-enriched Al$_2$O$_3$, $3 \times 10^4$ dpm ($\beta^+$/g, did not allow single crystal measurements to be made.

Le Gall et al. [102] came to a different conclusion, but the validity of their results remains in question. They employed the same single crystals used by Prot and Monty [148] to study oxygen diffusion, and overcame the difficulties in using $^{26}$Al as a tracer by neutralising a diluted high-radioactivity solution of AlCl$_3$ in HCl with NH$_4$OH to fix the tracer as Al(OH)$_3$ onto the surface of an Al$_2$O$_3$ sample.

The diffusion anneals were performed at four temperatures between 1540 – 1697°C and the resulting penetration profile, only a few µm deep, determined by mechanical sectioning and counting of the residual activity in the remaining sample. The profiles of the three lower-temperature measurements contained ‘tails’ of low concentrations that penetrated more deeply into the material. These tails can arise for two opposing reasons. Either the tails result from pipe diffusion in dislocations forming low angle subboundaries, and the shallow part of the profile (near the surface) represents bulk diffusion, or the tails actually show bulk diffusion, and the shallow part represents enhanced diffusion in surface cracks and damage. Whereas Le Gall et al. [102] used the former interpretation to conclude that $D_{\text{Al}} \sim D_{\text{O}}$, Heuer and Lagerlöf [80] believed the latter, finding $D_{\text{Al}} \gg D_{\text{O}}$ and that the value of $D_{\text{Al}}$ agrees well with Paladino and Kingery’s polycrystalline data [140]. The validity of Heuer and Lagerlöf’s interpretation requires very severe damage resulting from polishing and left after annealing, and, as there has been no microstructural characterisation of the samples, this is open to debate. However, significantly, the highest temperature data point at 1697°C did not show a tail and gave a value $D_{\text{Al}} \sim 10^2D_{\text{O}}$. Le Gall et
interpreted this as enhanced subboundary diffusion, but Heuer [79] argued that it is more likely that annealing at higher temperatures eliminates tails resulting from surface damage, and that the measurement at 1697°C is the only valid data point for the experiment. Due to the small number of measurements taken and the uncertainty in their interpretation, these results are not included in the figures.

The newest determination of $D_{\text{Al}}$ by Fielitz et al. [57] reinforces the traditional assertion that $D_{\text{Al}} \gg D_{\text{O}}$ by comparing simultaneous tracer diffusion of $^{18}\text{O}$ and $^{26}\text{Al}$ in Ti-doped alumina single crystals. Czochralski-grown single crystals were cut into thin slices, mechanically polished, and 128ppm Ti impurities were added to the samples to avoid the effects of unwanted background impurities. After pre-annealing at 1750°C for 10h to remove polishing damage and oxidise Ti$^{3+}$ in the crystal to Ti$^{4+}$, the diffusion anneals were carried out at four different temperatures between 1230°C and 1500°C, and the depth distribution of the isotopes measured using SIMS. Although $^{26}\text{Al}$ could be detected up to significant depths of a few microns, no measurable amount of $^{18}\text{O}$ had diffused into the sample, suggesting that $D_{\text{Al}} \gg D_{\text{O}}$ and contradicting suggestions than AlO molecules mediate diffusion (see Sec. 4.2.5). The results were in accordance with Reddy and Cooper’s [151] measurements of reduced $^{18}\text{O}$ diffusivity in Ti-doped single crystal alumina and Paladino and Kingery’s [140] previous measurement of $D_{\text{Al}}$.

### 4.2.4 Anisotropy

Alumina single crystals exhibit anisotropy of elastic modulus, thermal expansion, and refractive index due to the corundum lattice. The differences with crystal orientation are small; for example, the thermal expansion differs by only about 10% parallel and perpendicular to the c-axis. Unfortunately, the variations in measured values of $D_{\text{Al}}$ and $D_{\text{O}}$ preclude deduction of such small anistropy effects. Diffusional anisotropy in alumina has only been successfully measured for sodium diffusion coefficients [47], also parallel and perpendicular to the c-axis, yielding a 35% enhancement along the c-axis at 1568°C. Whereas sodium diffuses via an interstitial mechanism, aluminium and oxygen vacancies do not, and there is no reason to suggest that there could not be a mechanism of fast diffusion along specific crystallographic directions, such as for aluminium vacancies along the c-axis. Such possibilities will be discussed at length in Sec. 4.4.8, where we determine diffusion barriers for native point defects in alumina.
4.2.5 Alternative species

Doremus [47, 46] attempts to reconcile the puzzling features of diffusion in alumina by suggesting that AlO defects mediate diffusion via a diffusion-reaction mechanism. This idea was initially proposed to explain the difference between the measured and predicted rates of oxidation of aluminium-containing alloys, such as NiCrAl and FeCrAl [47]. The rate of oxidation of these alloys is much faster than the rate calculated from the measured diffusion coefficients of oxygen and aluminium in alumina, and the activation energy (2.49eV) less than half that of $D_O$. The enhanced oxidation rate cannot be explained by the traditional Wagner theory [185], whereby the growth of the oxide layer involves transport of ions through the oxide scale.

Doremus was motivated by the oxidation of silicon, which is controlled by molecular diffusion of water and oxygen molecules through the SiO$_2$ scale [45, 42], and Le Gall et al.’s aluminium diffusion experiments that suggested that $D_{Al} \sim D_O$. However, the SiO$_2$ scale formed during oxidation of silicon is quite open, allowing for easier diffusion of molecules than the dense corundum lattice, and the aluminium diffusion data from Le Gall et al. has been disputed in the literature [79, 47].

According to Doremus’ theory, the formation of AlO vacancies can occur by vaporisation of AlO molecules, which diffuse from the bulk of alumina, are adsorbed on the surface, and subsequently vaporised. The fast growth of oxide scales on aluminium-containing alloys can then be explained as an effect of a diffusion-exchange mechanism involving AlO vacancies. The AlO molecules diffuse rapidly and also exchange with constituents of the solid solvent, for example by reactions such as

\[
2\text{Al}^{18}\text{O} + \text{Al}_2\text{O}_3 \rightleftharpoons 2\text{Al}^{16}\text{O} + \text{Al}_2\text{O}_3.
\]  

As the concentration profile of exchanged Al$_2^{18}$O$_3$ is measured experimentally, the apparent diffusion coefficient calculated from the measured profile is much smaller than the actual diffusion coefficient of the dissolved AlO molecules.

However, Heuer [114] notes concerns regarding the viability of AlO as a vapour species above alumina at elevated temperatures, and the heat of vaporisation of AlO from alumina is 8.14eV [24], which is higher than the activation energies measured for oxygen and aluminium diffusion. In addition, it is not even known whether the diffusion barrier of AlO defects, presumed to migrate via a vacancy mechanism, is low enough to render them relevant to diffusion.

If AlO molecules do participate heavily in diffusion, this mechanism could explain the
apparent equivalence of $D_{\text{Al}}$ and $D_{\text{O}}$ in Le Gall et al.’s measurements, the very small pressure dependence of $D_{\text{O}}$, and also the order of magnitude difference between the oxygen diffusion coefficients reported by Reddy and Cooper [151] and Reddy [150]. Reddy’s results showed that the difference between the two data sets arose from differences in how the diffusion profiles were formed during annealing. Reddy and Cooper formed their diffusion profile while the sample was in contact with oxygen gas, so that AlO molecules would be able to vaporise and form large numbers of AlO vacancies. In contrast, Reddy formed a diffusion couple from two alumina samples containing tracers, limiting the possibility of AlO vaporisation, and thereby leading to a lower measurement of $D_{\text{O}}$.

The possibility of AlO molecules dictating diffusion in alumina is examined in Sec. 4.5 and 4.4.8 by calculating the concentrations and diffusion barrier of AlO defects in doped and undoped alumina.

### 4.3 Previous theoretical work

Despite many attempts to describe the defect chemistry of alumina theoretically, none of the work to date can satisfactorily explain the experimental diffusion data. From the early seventies onwards, classical inter-atomic potential methods have been used to study the formation and binding energies of point defects and clusters in alumina. Unfortunately, the results exhibit a wide spread in Schottky and Frenkel energies, and their incorporation into a mass action approach yields defect concentrations that do not reveal any buffering mechanism that could account for experimental results on self-diffusion coefficients. Recent efforts have focussed on describing defect energetics and properties using density functional theory (DFT); however, these studies have not satisfactorily corrected for the many underlying problems that plague calculations of charged defects in solids using DFT, such as finite-size effects and the underestimation of the band gap. The calculated formation energies, whether using classical potentials or DFT, are so high that defects would only be present in minute amounts, too small to account for measured diffusion coefficients, even in undoped alumina. In addition, no previous work has attempted to connect with experiment by determining the chemical potentials that enter Zhang-Northrup defect formation energies in terms of $p_{\text{O}_2}$ and $T$, as described in Sec. 3.4.2. Both classical potentials and DFT have been sparingly used to analyse the diffu-
sion barriers of native defects in alumina, with only marginal success. The diffusion barriers determined so far are all very high in comparison with similar calculations in other solids, and the mechanisms of diffusion remain uncertain. The difficulty in reconciling experimental diffusion data with theory was dubbed the ‘corundum conundrum’ by Heuer and Lagerlöf in their 1999 review [80].

One of the very first determinations of the intrinsic disorder energetics in alumina by Mohapatra and Kroger [129] combined experimental results on optical absorption with an estimated expression for intrinsic electronic excitation. They were able to deduce the Schottky energy as 3.83eV, and the aluminium Frenkel energy as 4.45eV, both normalised per defect. El-Aiat and Kroger [51] later used experimental measurements of the conductivity in Mg-doped crystals to find the Schottky, oxygen Frenkel, and aluminium Frenkel energies as 6.6±1.4eV, 5.5±1.2eV, and 8.2±1.8eV, respectively.

There have been a number of studies [44, 27, 72] of point defects in alumina using polarisable ion models [43, 19, 20], where a massless shell of charge $Y$ is connected to a massive core of charge $X$ by an isotropic spring of force constant $k$. The parameters of these models are usually optimised so that they give good results for the cohesive energy, lattice parameters, and dielectric and elastic constants. They can also be derived from the electron gas model [76, 86], or the results of ab initio calculations [52]. Although classical shell models have been applied extensively to alumina, they consistently fail to predict the ground-state corundum structure, preferring bixbyte\(^5\) instead. This is indicative of their limit as a quantitatively accurate method for describing defect energetics. Classical shell models generally predict that Schottky and oxygen Frenkel energies are lower than the cation Frenkel energy (see Table 4.2), in agreement with early experimental work. However, the published values for the formation energies vary by at least 2eV, and in some cases nearly 5eV. Small changes in the shell model parameters can account for some of these large differences. For example, Dienes \textit{et al.} [44] use a model containing an empirical parameter tuned to give a good fit to the high-frequency dielectric constant, and variation of this parameter by $< 5\%$ can change formation energies by over 1eV. These uncertainties render it impossible to predict whether Schottky or Frenkel disorder dominates in a pure crystal. Either way, the very high energies found for all forms of disorder mean that, even near the melting point, intrinsic crystals will only contain ca. $10^{-15}$

\(^5\)A structure derived from fluorite by removing $1/4$ of the anions to change the stoichiometry from MO\(_2\) to M\(_2\)O\(_3\).
Table 4.2: Early determinations of the intrinsic defect energies in alumina (all in eV). The first two studies deduce values from experiment, the rest using shell models. The superscript 1 indicates a partial charge model; otherwise the ions assume their fully charged ionic values. Shell model parameters are either fit to empirical data (2) or derived from an electron gas model (3).

<table>
<thead>
<tr>
<th>Schottky</th>
<th>Oxygen Frenkel</th>
<th>Aluminium Frenkel</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.83</td>
<td>-</td>
<td>4.45</td>
<td>Mohapatra et al. [129]</td>
</tr>
<tr>
<td>6.6 ± 1.4</td>
<td>5.5 ± 1.2</td>
<td>8.2 ± 1.8</td>
<td>El-Aiat et al. [51]</td>
</tr>
<tr>
<td>5.7</td>
<td>7.0</td>
<td>10.0</td>
<td>Dienes et al. [44]</td>
</tr>
<tr>
<td>4.18</td>
<td>3.79</td>
<td>5.22</td>
<td>Catlow et al. [27]</td>
</tr>
<tr>
<td>5.14</td>
<td>8.27</td>
<td>7.09</td>
<td>Catlow et al. [27]</td>
</tr>
<tr>
<td>5.86</td>
<td>5.79</td>
<td>6.30</td>
<td>Grimes et al. [72]</td>
</tr>
<tr>
<td>5.17</td>
<td>4.87</td>
<td>6.59</td>
<td>Grimes et al. [72]</td>
</tr>
<tr>
<td>5.15</td>
<td>5.54</td>
<td>7.22</td>
<td>Atkinson et al. [7]</td>
</tr>
</tbody>
</table>

point defects per unit cell, leading many to suggest that defect concentrations are almost entirely determined by impurity concentrations. Also, allowing the ions to take partial charges instead of retaining their fully-charged ionic values generally lowers the formation energies [72], indicating that alumina exhibits a certain degree of covalency. The difficulty in discerning which intrinsic disorder dominates already signals that classical inter-atomic potentials do not deliver an accurate description of point defect energetics.

Shell models have also been used to study cluster binding energies [27] and defect migration paths [44, 87]. Dienes et al. [44] were computationally limited to guessing a location for the saddle point of each migration path, at which point they estimated a value for the diffusion barrier. Their results, 2.9eV for oxygen diffusion to a nearest-neighbour in the basal plane, and > 6eV for aluminium vacancy and interstitial migration, are very large, and indicate that the true diffusion paths may not be the shortest paths between given beginning and end points, but rather follow a curve, or involve more complex mechanisms. Jacobs and Kotomin [87] follow a similar approach and determine the oxygen and aluminium vacancy diffusion barriers as 1.9eV and 3.7eV, respectively; they also find that the Schottky and oxygen Frenkel energies are much lower than the aluminium Frenkel energy.

Later, some attention was given to inspecting the properties of $F$-centres in alumina [170, 94]. These form when an oxygen vacancy traps one ($F^+$-centre) or two ($F$-centre) electrons. Studies using semi-empirical and Hartree-Fock methods suggest that about 90% (for $F^+$-centres) or 80% (for $F$-centres) of the electron density
associated with the defect is localised in the vacancy. The remainder of the electron density was found to be spread over the nearest-neighbours of the vacancy.

Very few studies have attempted to use calculated defect formation and cluster binding energies to determine the concentrations of defects. Lagerlöf and Grimes [98] incorporated the formation and binding energies calculated by Grimes [72] into a mass action approach similar to the one used in the current study, in order to predict the concentrations of native point defects and a number of clusters in undoped and aliovalently-doped systems. Their results demonstrate the amphoteric nature of alumina, whereby oxygen and aluminium diffusion is dominated by oxygen vacancies and aluminium interstitials, or oxygen interstitials and aluminium vacancies, depending on which aliovalent dopant is present in greater abundance. They also find that defect clusters form at low temperatures, including \{\text{Ti}_{\text{Al}} : V_{\text{Al}}'}''\} in Ti-doped and \{\text{Mg}_{\text{Al}}' : V_{\text{O}}'\} in Mg-doped crystals. However, they are unable to account for the buffering mechanism revealed in the diffusion data, whereby the defect populations of oxygen vacancies are found to change by only 1-2 orders of magnitude upon heavy Mg or Ti doping. Instead, they predict that the balance between aliovalent impurities alters the isolated defect populations radically. A slight excess of magnesium or titanium changes the concentrations of charge-compensating defects by up to 15 orders of magnitude at experimental temperatures.

Heuer and Lagerlöf demonstrate that migration energies extracted from the same calculations by Grimes, as well as their loop-annealing data, are unusually large at \(\sim 5\,\text{eV}\), regardless of whether oxygen or aluminium diffusion is assumed to be rate limiting. These migration energies are inferred from activation energies for diffusion found from the loop-annealing experiment, and apparent formation energies deduced from Arrhenius plots of the concentrations of the dominant defects in aliovalently-doped crystals as obtained by Lagerlöf and Grimes. Although no calculations of the diffusion barriers using the potentials of Grimes have been performed, similar calculations in other oxides using shell models [27, 87] suggest that these values are also unreasonably high. Solving this conundrum was the motivation for much of the work presented here.

Atkinson et al. [7] used a similar mass-action approach with Born-model [19, 20] calculations, but for a wider range of dopant ions, to predict the solution and binding energies of various defects as functions of ionic radii. They found that isovalent substitutionals in alumina have very large solution energies, so that they will not stay in solution, but rather form a second phase. Divalent and tetravalent substitutionals
Section 4.3: Previous theoretical work

create anion and cation vacancies in large numbers, respectively, and were thus also unable to account for the buffering effect.

Building on Atkinson’s results, Harding et al. [75] propose that defect clustering could be key to understanding diffusion in alumina. They suggest that the excess compensating defects created when impurities are present are tightly bound in clusters with the impurities. The bound clusters do not contribute to diffusion, and the numbers of isolated defects (primarily oxygen and aluminium vacancies for divalent and tetravalent impurities, respectively) are limited, as most are sucked into clusters.

Unfortunately, despite their computational advantages, classical shell models struggle to describe the energetics of charged defects. The neglect of electronic polarisation, kinetic energy, and exchange and correlation hinders accurate determinations of defect formation energies and cluster binding energies.

Xu et al. [190] were the first to perform ab initio calculations of defects in alumina, examining the energetics, lattice relaxations, and defect states of neutral and charged oxygen vacancies using DFT. They obtain the formation energy of the neutral oxygen vacancy, $V_O^\times$ (i.e. a vacancy occupied by two electrons), as $\Delta E_{V_O^\times}^f = 5.83$eV, but give little insight into their procedure for determining it. Their band gap of 6.31eV is predictably smaller than the experimental gap of 9.57eV [64] due to errors intrinsic to the approximate exchange-correlation functionals used in almost all DFT calculations. They also analyse the F-centre defect states of the charged oxygen vacancy and find significant lattice relaxations beyond the first nearest-neighbours, indicating that large supercells must be fully relaxed.

Formation energies calculated using the Zhang-Northrup formalism presented in Sec. 3.4.1 are always referenced to the atomic chemical potentials of the atomic species involved. Tanaka et al. [174] examined the formation energy and structural relaxation of $V_O^\times$ defects using the Zhang-Northrup formalism in DFT with a GGA functional, finding $\Delta E_{V_O^\times}^f = 7.08$eV in an oxygen-rich environment ($\mu_O$ at the oxidation limit), and $\Delta E_{V_O^\times}^f = 1.63$eV in an aluminium-rich environment ($\mu_O$ at the reduction limit). Remarkably, they observed very small lattice relaxations of less than 1.5% for the aluminium and oxygen ions nearest to the defect, in contrast with earlier empirical work on F-centres. However, their calculations were performed at the $\Gamma$ point only, in small supercells containing < 96 atoms, which will be shown in Secs. 4.4.1 and 4.4.6 to be insufficient to treat formation energies accurately.
The most complete study of point defect energetics in alumina is also one of the most recent. Matsunaga et al. [122] performed DFT calculations of the formation energies and accompanying relaxations of all the native defects in alumina in a number of charge states, following the Zhang-Northrup formalism, also with a GGA functional. They determined the formation energies of the defects as explicit functions of the electron and oxygen chemical potentials. Their values for the Schottky, anion Frenkel, and cation Frenkel energies, equal to 4.01eV, 6.52eV, and 4.95eV, respectively, differ markedly from previous work, suggesting that classical potentials may have considerable difficulties describing defect energetics in alumina. The disagreement could, in part, arise from the fact that the parameters of the classical potentials used were obtained by fitting to the reference state of perfect crystal. Using such potentials to calculate defect formation energies can lead to complications as, for example, the bonding characteristics surrounding an interstitial atom differ greatly from those in the perfect crystal. When examining interstitial defects, Matsunaga et al. highlight the significant distortions in the electron density caused by the interactions of the interstitials with neighbouring ions. This important effect demonstrates the importance of using ab initio techniques, rather than classical models, in determining defect properties in alumina. In addition, Matsunaga et al. determine the formation energies of the bound Frenkel pairs, $\Delta E^{f}_{O_{c}:O}$ and $\Delta E^{f}_{Al_{c}:Al}$, as 5.69eV and 4.50eV, respectively, suggesting a significant stabilisation by forming clusters. They later expanded their work, using the same methodology, to Ti-doped alumina supercells [121], and concluded that titanium substitutional defects prefer to exist in the neutral Ti$_{Al}^0$ state, so that titanium assumes the charge state Ti$^{3+}$. However, they did not account for finite-size effects or the DFT underestimation of the band gap, and did not relax the entire supercell, but only atoms within 3.4Å of the defect. The only previous ab initio calculation of a diffusion barrier in alumina comes courtesy of Carrasco et al. [25], who analysed the formation and diffusion of oxygen vacancies using DFT. They found the formation energy of the neutral oxygen vacancy as $\Delta E_{V_{O}^{\infty}}^{f}$ = 10.14eV, much higher than previous estimates, although it is unclear which value they used for the electron chemical potential. In order to minimise finite-size effects, they used a 3 × 3 × 1 supercell, considerably larger than for previous DFT calculations. They also examined the migration of an oxygen vacancy to its nearest-neighbours, and found a lowest diffusion barrier of 3.7eV, which seems unusually high. Carrasco et al. did not specify to which neighbour the vacancy diffuses, nor the details of their calculations.
To summarise, although there is a profuse body of work on defect energetics in alumina, none of it unshrouds the mechanisms that govern the defect chemistry and diffusion of defects. This study will show how uncertainties in defect formation energies and cluster binding energies can be reduced by careful *ab initio* calculations that account for finite-size effects, the DFT underestimation of the band gap, and vibrational contributions to the energy of the metal chemical potential. A theoretical description of defect diffusion in alumina can be found by incorporating these quantities within a mass-action approach. To date, there have been no studies of $\text{Al}_2\text{O}_3$ combining *ab initio* calculations of defect formation energies with a mass-action approach to defect concentrations.

### 4.4 Density functional theory calculations

DFT calculations of defect formation energies, cluster binding energies, and diffusion barriers in alumina form the backbone of this work. This section begins by introducing the corundum structure and examining the convergence of the DFT geometry and total energy for the different basis sets and Brillouin zone sampling methods used. Well-converged properties are paramount for a faithful representation of defect energetics in alumina. Following this introduction, the various defect species and clusters included in the mass-action approach are presented, and their formation or binding energies determined. New ground-state configurations are found for the aluminium vacancy and oxygen interstitial defects, and a host of previously ignored clusters incorporated. Then, the full DFT calculations are dismantled and the effects of using accurate chemical potentials, applying band gap and finite-size corrections, and including vibrational effects, as laid out in Secs. 3.4.2, 3.5.2, 3.5.3, and 3.5.1, are discussed. Finally, the diffusion mechanisms governing native defects in alumina are presented and the accompanying diffusion barriers determined.

#### 4.4.1 Geometry and basis set convergence

$\alpha$-$\text{Al}_2\text{O}_3$ (space group $\text{R}3\text{c} (D_{6h}^5)$) has a complex corundum structure. Although alumina is a trigonal crystal with a rhombohedral primitive cell (see Fig. 4.2), it is commonly described by a larger hexagonal unit cell (see Fig. 4.3). The corundum structure can be constructed from a bipartite hexagonal close-packed (hcp) analogue, created by letting the oxygen ions occupy a hcp lattice, and then inserting
aluminium ions into the empty octahedral interstitial sites, so that the aluminium ions themselves form another hcp lattice. As the resulting 1:1 composition is incorrect, removing one third of the aluminium ions corrects the coordination to 2:3. This causes the oxygen ions to relax towards interstices containing aluminium ions, distorting them from the bipartite hcp positions, and leading to the corundum structure. In corundum, each aluminium ion is bonded to six oxygen ions arranged in a distorted octahedron, whereas each oxygen ion is bonded to four aluminium ions in a distorted tetrahedron. There are two short Al – O bonds of 1.846 Å and four longer bonds of 1.915 Å. Each oxygen ion has twelve second nearest-neighbour oxygen ions at distances between 2.524 and 2.869 Å. When viewing corundum down the a-axis (see Fig. 4.4), the structure comprises a stacking of alternating layers of aluminium and oxygen ions in the ab-plane. In each aluminium layer, half the aluminium ions are slightly above, and half slightly below the plane. The oxygen ions all lie in the ab-plane, and form triangles of side length 1.441 Å and 1.661 Å. The layers can be described by the stacking sequence (aAbBcAaBbAcB ...), where anion layers are denoted by A, B, and cation layers by a, b, c, repeating every 12 layers along the c-axis.

The rhombohedral primitive cell of corundum contains 10 atoms at positions \( r = ut_1 + vt_2 + wt_3 \), where the oxygen ions lie at fractional coordinates

<table>
<thead>
<tr>
<th>( u )</th>
<th>( v )</th>
<th>( w )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 2\delta_O )</td>
<td>( \frac{1}{2} - 2\delta_O )</td>
<td>( \frac{3}{4} )</td>
</tr>
<tr>
<td>( \frac{1}{4} )</td>
<td>( 1 - 2\delta_O )</td>
<td>( \frac{1}{2} + 2\delta_O )</td>
</tr>
<tr>
<td>( \frac{1}{2} - 2\delta_O )</td>
<td>( \frac{3}{4} )</td>
<td>( 2\delta_O )</td>
</tr>
<tr>
<td>( \frac{1}{2} + 2\delta_O )</td>
<td>( \frac{1}{4} )</td>
<td>( 1 - 2\delta_O )</td>
</tr>
<tr>
<td>( \frac{3}{4} )</td>
<td>( 2\delta_O )</td>
<td>( \frac{1}{2} - 2\delta_O )</td>
</tr>
<tr>
<td>( 1 - 2\delta_O )</td>
<td>( \frac{1}{2} + 2\delta_O )</td>
<td>( \frac{1}{4} )</td>
</tr>
</tbody>
</table>

and the aluminium ions at fractional coordinates

<table>
<thead>
<tr>
<th>( u )</th>
<th>( v )</th>
<th>( w )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \frac{1}{6} - \delta_{Al} )</td>
<td>( \frac{1}{6} - \delta_{Al} )</td>
<td>( \frac{1}{6} - \delta_{Al} )</td>
</tr>
<tr>
<td>( \frac{2}{6} + \delta_{Al} )</td>
<td>( \frac{2}{6} + \delta_{Al} )</td>
<td>( \frac{2}{6} + \delta_{Al} )</td>
</tr>
<tr>
<td>( \frac{4}{6} - \delta_{Al} )</td>
<td>( \frac{4}{6} - \delta_{Al} )</td>
<td>( \frac{4}{6} - \delta_{Al} )</td>
</tr>
<tr>
<td>( \frac{5}{6} + \delta_{Al} )</td>
<td>( \frac{5}{6} + \delta_{Al} )</td>
<td>( \frac{5}{6} + \delta_{Al} )</td>
</tr>
</tbody>
</table>

in terms of the rhombohedral lattice vectors \( t_i \).
Section 4.4: Density functional theory calculations

\[
\begin{pmatrix}
  t_{1x} & t_{1y} & t_{1z} \\
  t_{2x} & t_{2y} & t_{2z} \\
  t_{3x} & t_{3y} & t_{3z}
\end{pmatrix}
= \begin{pmatrix}
  s & 0 & r \\
  -s/2 & \sqrt{3}s/2 & r \\
  -s/2 & -\sqrt{3}s/2 & r
\end{pmatrix}.
\]

(4.2)

The values of \(s\) and \(r\) are defined in terms of the lattice constant \(a_{\text{rhomb}}\) and the angle \(\beta\) between lattice vectors as \(s = 2a_{\text{rhomb}}/\sqrt{3}\sin(\beta/2)\) and \(r = \sqrt{a_{\text{rhomb}}^2 - s^2}\). Crystallographic measurements \([40]\) give \(a_{\text{rhomb}} = 5.136\AA, \beta = 55.280^\circ, \delta_O = 0.027\), and \(\delta_{\text{Al}} = 0.018\). The fractional displacements \(\delta_O\) and \(\delta_{\text{Al}}\) represent the deviation from the bipartite hcp structure (when the coordination between cation and anion is 1:1). Thus, bipartite hcp can be obtained by inserting an extra \(1/3\) aluminium ions at \((0, 0, 0)\) and \((3/6, 3/6, 3/6)\) and setting \(\delta_O = 0\) and \(\delta_{\text{Al}} = 0\).

In the current calculations, the hexagonal unit cell is used. This contains 30 atoms at positions \(\mathbf{r} = u\mathbf{t}_1 + v\mathbf{t}_2 + w\mathbf{t}_3\) and corresponding fractional coordinates

<table>
<thead>
<tr>
<th>(u)</th>
<th>(v)</th>
<th>(w)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1/3 - (\delta_O)</td>
<td>0/6</td>
</tr>
<tr>
<td>1/3 - (\delta_O)</td>
<td>0</td>
<td>0/6</td>
</tr>
<tr>
<td>2/3 + (\delta_O)</td>
<td>2/3 + (\delta_O)</td>
<td>0/6</td>
</tr>
<tr>
<td>0/3 + (\delta_O)</td>
<td>2/3</td>
<td>1/6</td>
</tr>
<tr>
<td>1/3</td>
<td>1/3 + (\delta_O)</td>
<td>1/6</td>
</tr>
<tr>
<td>2/3 - (\delta_O)</td>
<td>3/3 - (\delta_O)</td>
<td>1/6</td>
</tr>
<tr>
<td>1/3 + (\delta_O)</td>
<td>9/3 + (\delta_O)</td>
<td>2/6</td>
</tr>
<tr>
<td>2/3</td>
<td>2/3 - (\delta_O)</td>
<td>2/6</td>
</tr>
<tr>
<td>3/3 - (\delta_O)</td>
<td>1/3</td>
<td>2/6</td>
</tr>
<tr>
<td>0</td>
<td>2/3 + (\delta_O)</td>
<td>3/6</td>
</tr>
<tr>
<td>1/3 - (\delta_O)</td>
<td>1/3 - (\delta_O)</td>
<td>3/6</td>
</tr>
<tr>
<td>2/3 + (\delta_O)</td>
<td>0</td>
<td>3/6</td>
</tr>
<tr>
<td>0/3 + (\delta_O)</td>
<td>1/3 + (\delta_O)</td>
<td>4/6</td>
</tr>
<tr>
<td>1/3</td>
<td>3/3 - (\delta_O)</td>
<td>4/6</td>
</tr>
<tr>
<td>2/3 - (\delta_O)</td>
<td>2/3</td>
<td>4/6</td>
</tr>
<tr>
<td>1/3 + (\delta_O)</td>
<td>1/3</td>
<td>5/6</td>
</tr>
<tr>
<td>2/3</td>
<td>0/3 + (\delta_O)</td>
<td>5/6</td>
</tr>
<tr>
<td>3/3 - (\delta_O)</td>
<td>2/3 - (\delta_O)</td>
<td>5/6</td>
</tr>
</tbody>
</table>

defined by the hexagonal lattice vectors \(\mathbf{t}_i\).
The rhombohedral primitive cell of alumina (left), embedded in a $2 \times 2 \times 1$ supercell (right). Aluminium atoms are coloured blue, oxygen atoms red.

The hexagonal primitive cell (left) is repeated to form a $2 \times 2 \times 1$ supercell (right) for the majority of this work.

\[
\begin{pmatrix}
t_{1x} & t_{1y} & t_{1z} \\
t_{2x} & t_{2y} & t_{2z} \\
t_{3x} & t_{3y} & t_{3z}
\end{pmatrix}
= \begin{pmatrix}
\sqrt{3}a_0/2 & a_0/2 & 0 \\
0 & a_0 & 0 \\
0 & 0 & c_0
\end{pmatrix}.
\] (4.3)

The values of $a_0$ and $c_0$ are related to the rhombohedral unit cell parameters by $a_{\text{rhomb}} = \sqrt{(c_0/\sqrt{3})^2 + (a_0/\sqrt{3})^2}$ and $\sin(\beta/2) = a_0/2a_{\text{rhomb}}$, and experimentally found [189] to be 4.7628Å and 13.0032Å, respectively.

The first step in examining alumina using electronic structure methods is to find the optimised geometry for the given set of approximations. In DFT, these approximations are the chosen exchange-correlation functional, the basis set, the pseudopotential (PSP), and various computational parameters. The optimised DFT geometry can be found when certain quantities are said to be well-converged. Convergence occurs when these quantities, such as the total energy (crucial for defect formation energies) or the forces on atoms (crucial for phonon calculations) lie within a certain tolerance range. We will focus on examining the convergence behaviour with respect to the basis set and \textbf{k}-point sampling. The plane wave DFT calculations here and
Figure 4.4: Schematic of $\alpha$-Al$_2$O$_3$ viewed down the $a$-axis. Crosses (×) indicate vacant octahedral interstices. The aluminium and oxygen ions are stacked in alternating layers parallel to the $ab$-plane. All of the oxygen ions lie within their respective layers, but the aluminium ions lie slightly above or below the layers. The oxygen ions form small and large triangles in their planes with side lengths 1.441 Å and 1.661 Å, respectively.
throughout the rest of the study are performed using CASTEP [33], Gaussian DFT calculations using CRystal [48], and QMC calculations using CASINO [134]. All DFT geometry optimisations use the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm [23, 59, 68, 164], known for its efficiency and robustness.

For a plane wave basis set, there are two criteria against which we require the total energy to converge: the plane wave cut-off and the \( k \)-point sampling grid. The convergence of the total energy of a rhombohedral primitive cell of alumina with plane wave cut-off is plotted in Fig. 4.5 for Vanderbilt ultrasoft pseudopotentials (USPs) [182] and various exchange-correlation functionals. The USPs contain three, six, eight, and twelve valence electrons for aluminium, oxygen, magnesium, and titanium, respectively (magnesium and titanium atoms are not considered now, but later when determining the formation energies and binding energies of dopants and dopant-containing clusters). There is little variation with functional, and the total energy is converged to 0.015\( \text{eV/atom} \) at a plane wave cut-off of 550\( \text{eV} \), which turns out to be accurate enough for our purposes. Moreover, in Sec.4.4.2, it will be verified that energy differences between perfect and defect supercells were converged to within 0.02\( \text{eV/atom} \) for \( E_{\text{cut}} = 550\text{eV} \). This cut-off was used in all subsequent calculations involving USPs. Convergence of the total energy with respect to \( k \)-point sampling is rapid as alumina is a good insulator. Using a \( 2 \times 2 \times 2 \) Monkhorst-Pack [130] grid yields energies converged to within 0.02\( \text{eV/atom} \), while using a \( 3 \times 3 \times 3 \) grid yields energies converged to 0.002\( \text{eV/atom} \) at a plane wave cut-off of 550\( \text{eV} \) for both the LDA and GGA (PBE parametrisation [141]). For the USP calculations presented here, a \( 4 \times 4 \times 4 \) grid is used throughout.

Unlike plane wave basis sets, Gaussian basis sets have no single tuneable parameter that defines their quality. This work uses an 85-11G* basis set for aluminium [28], where eight and five Gaussian primitives, respectively, have been used to expand the core \( s \) and \( sp \) orbitals, and the valence orbitals are described by two functions expanded in one Gaussian primitive each, and a polarisation function is added. Meanwhile, oxygen atoms are described using an 8-411 basis set [132], expanding the core \( s \) orbital in eight Gaussian primitives, and the valence \( sp \) orbital in three basis functions, containing four, one, and one Gaussian primitives, each. The convergence of the total energy is examined with respect to \( k \)-point sampling. We find that the total energy is converged to \( 10^{-6}\text{eV/atom} \) for a \( 4 \times 4 \times 4 \) Monkhorst-Pack grid using both the LDA and hybrid B3LYP functionals. All subsequent calculations involving
Section 4.4: Density functional theory calculations

Gaussians therefore use a $4 \times 4 \times 4$ Monkhorst-Pack grid.

The QMC calculations in this work use trial wave functions whose orbitals originate from plane-wave DFT calculations. The quality of a trial wave function\(^6\) determines the variance of the QMC energy and the efficiency of the QMC calculation. As the orbitals in the Slater determinant of a trial wave function are generally not re-optimised due to the computational cost, it is paramount that the DFT orbitals are as accurate as possible. This can be achieved by increasing the plane wave cut-off beyond the point at which the total DFT energy is converged, leading to more accurate orbitals and a lower-variance trial wave function. Such DFT calculations are time-consuming, but the excess time required is considerably less than the time saved in the subsequent QMC calculations.

Unfortunately, USPs are incompatible with current QMC implementations, so Trail-Needs Dirac-Fock norm-conserving PSPs (TNDFPs) [177, 178] are used to generate the DFT orbitals for the trial wave functions. TNDFP pseudopotentials are much harder than USPs, as evidenced by the higher plane-wave cut-off needed to converge the total energy in Fig.4.5. There are ‘hard’ and ‘soft’ TNDFPs, the former intended for atomic calculations (using a smaller core radius for higher accuracy), and the latter for solids (using a larger core radius). The difference in accuracy for bulk geometries is negligible, but the difference in the cut-off required for convergence is marked, indicating that the soft TNDFPs are better suited for our purposes. The TNDFPs contain three and six valence electrons for aluminium and oxygen, respectively. Unfortunately, one limiting factor in using QMC for defect calculations in alumina is the lack of a suitable pseudopotential for titanium. The only available pseudopotentials for titanium do not include semi-core $s$ and $p$ states, which have been shown to contribute heavily to bonding [186]. This prohibits calculations of any cluster binding energies containing titanium using QMC.

Once the convergence criteria for the DFT calculations have been established, the corresponding optimised bulk cell geometry can be found. For the rhombohedral primitive cell, this involves optimising the parameters $a_{\text{rhomb}}$, $\beta$, $\delta_{\text{Al}}$, and $\delta_{\text{O}}$ simultaneously. Similarly, for the hexagonal unit cell, $a_0$, $c_0$, $\delta_{\text{Al}}$, and $\delta_{\text{O}}$ can be optimised, and the results compared with the rhombohedral primitive cell for consistency. We allow the initial configuration, given by the experimental values [40] of these parameters, to relax in DFT until the forces on atoms are converged to $< |10^{-4}|$ eV/Å and

\(^6\) The ‘quality’ describes how well a trial wave function approximates the ground-state wave function of the system.


**Figure 4.5:** Convergence of the total energy of a rhombohedral primitive cell of Al$_2$O$_3$ using (a) Vanderbilt ultrasoft pseudopotentials (USPs), or (b) Trail-Needs Dirac-Fock norm-conserving pseudopotentials (TNDFPs), with Monkhorst-Pack $k$-point grid densities ranging from $1 \times 1 \times 1$ to $4 \times 4 \times 4$, as a function of plane wave cut-off $E_{\text{cut}}$. In (b), the ‘hard’ TNDFPs converge considerably more slowly than their ‘soft’ counterparts.

**Table 4.3:** Optimised DFT lattice parameters and atomic coordinates for different choices of exchange-correlation functional $E_{\text{xc}}$, basis set (plane waves or Gaussians), and pseudopotential (Vanderbilt USPs, TNDFPs).

<table>
<thead>
<tr>
<th></th>
<th>$a_{\text{rhomb}}$ (Å)</th>
<th>$\beta$ (°)</th>
<th>$\delta_{\text{Al}}$</th>
<th>$\delta_{\text{O}}$</th>
<th>$a_{\text{Al-O}}$ (Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiment [40]</td>
<td>5.136</td>
<td>55.280</td>
<td>0.0187</td>
<td>0.0283</td>
<td>1.846</td>
</tr>
<tr>
<td>LDA (USP)</td>
<td>5.042</td>
<td>55.313</td>
<td>0.0187</td>
<td>0.0283</td>
<td>1.824</td>
</tr>
<tr>
<td>GGA (USP)</td>
<td>5.062</td>
<td>55.182</td>
<td>0.0187</td>
<td>0.0283</td>
<td>1.826</td>
</tr>
<tr>
<td>LDA (TNDFP)</td>
<td>5.146</td>
<td>55.994</td>
<td>0.0187</td>
<td>0.0303</td>
<td>1.845</td>
</tr>
<tr>
<td>GGA (TNDFP)</td>
<td>5.167</td>
<td>54.915</td>
<td>0.0187</td>
<td>0.0303</td>
<td>1.848</td>
</tr>
<tr>
<td>LDA (Gaussians)</td>
<td>5.089</td>
<td>55.604</td>
<td>0.0187</td>
<td>0.0283</td>
<td>1.849</td>
</tr>
<tr>
<td>B3LYP (Gaussians)</td>
<td>5.153</td>
<td>55.443</td>
<td>0.0187</td>
<td>0.0283</td>
<td>1.867</td>
</tr>
</tbody>
</table>

The total energy to $< 10^{-7}$eV. The results are shown in Table 4.3 and indicate that the bulk geometry is well described by all choices of exchange-correlation functional, basis set, and pseudopotential.

### 4.4.2 Defect species and complexes

In order to determine the ground-state structures of defects in alumina, an initial guess for the defect structure is required. This can be more complicated than initially thought. For example, interstitials can be located in many different configurations, and even vacancies, as we will later see in alumina, are not always straightforward.
Section 4.4: Density functional theory calculations

New methods, such as metadynamics, genetic algorithms and random structure searches allow a complete analysis of the configuration space of defects, enabling many different defect configurations to be examined. Although these methods were not used in this work, subsequent studies, particularly of defect clusters, could benefit greatly from these developments.

We will examine the native point defects in alumina, namely oxygen and aluminium vacancies and interstitials, as well as the AlO divacancy suggested by Doremus, and a number of clusters combining native point defects with impurities.

Our studies on defect formation energies have shown that it is important to relax the entire computational supercell, rather than just the atoms nearest to the defect, in order to account for long-ranged elastic effects. Failure to include these relaxations in a finite-size scaling procedure can lead to substantial overestimations of defect formation energies [82]. In this study, the properties of defects are primarily determined in supercells constructed by stacking $2 \times 2 \times 1$, $3 \times 3 \times 1$, $4 \times 4 \times 1$, $2 \times 2 \times 2$, and $2 \times 2 \times 3$ copies of the 30-atom hexagonal unit cell. Cells of other shapes and sizes are also used occasionally. Although the relaxations of first and second nearest-neighbour atoms differ little between these supercells, the additional layers of atoms separating a defect from its periodic images in larger supercells enable a better description of relaxations further from the defect site. This can give rise to significant contributions to the defect formation energy.

**Oxygen vacancy**

Oxygen vacancies are the only straightforward point defects in alumina. The ground-state structure can be found by removing a single oxygen atom and relaxing the geometry of a supercell.

The relaxations found in this study using DFT are summarised in Table 4.4 and compare well with previous DFT [122, 82, 190, 25] and semi-empirical [170] calculations. The differences between results obtained using an LDA or GGA functional in a plane-wave basis, or a B3LYP functional in a Gaussian basis, are minimal for all defects. The results show a strong outward displacement of the two nearest-neighbour aluminium atoms for $V_O^-$ and $V_O^-$ defects, but a small inward displacement for $V_O^\times$ defects. The displacements away from the charged oxygen vacancies are caused by the repulsion between the positive charge on the defect site and the Al$^{3+}$ ions. The 12 second nearest-neighbour oxygen atoms relax inwards slightly,
Table 4.4: The atomic relaxations surrounding native defects in alumina, calculated using Vanderbilt USPs in the LDA. For each defect species, the identities of the first, second, and third nearest neighbours and their distances (in Å) from the defect site are shown, followed by the corresponding distances and percentage changes after the geometry has been relaxed. Since the equilibrium geometry of the aluminium vacancy is a split configuration in which an interstitial aluminium atom separates two vacant aluminium lattice sites, the neighbourhood is referenced to the octahedral interstitial site \( i \) that lies at the centre of the split vacancy. Similarly, the oxygen interstitial, the ground state of which is a dumbbell interstitial configuration centred on an oxygen lattice site, is referenced to an oxygen lattice site. Also, for the aluminium split vacancy, both nearest-neighbour aluminium atoms are no longer nearest neighbours (one has been removed and one has migrated to the interstitial site), so their relaxed distances are not shown.
the distance increasing with the defect charge state. These results are very similar to previous DFT calculations on oxygen vacancies by Matsunaga et al. [122], who used a 120 atom $2 \times 2 \times 1$ computational supercell, similar Vanderbilt USPs to those used in the current study, the same $4 \times 4 \times 4$ Monkhorst-Pack $k$-point mesh, and the GGA functional of Perdew and Wang [142], but only relaxed atoms within a radius of 3.4 Å from the defect. The relatively small sizes of the lattice relaxations around the neutral oxygen vacancy $V_O^\times$ are supported by further DFT studies by Tanaka et al. [174], who reported relaxations of less than 1.5% surrounding the $V_O^\times$ defect, also using a GGA functional, a $4 \times 4 \times 4$ $k$-point mesh, and Vanderbilt USPs. However, Tanaka et al.'s calculations were performed on very small computational supercells containing 54 to 96 atoms, which we will later demonstrate are insufficient to accurately describe the long-ranged Coulomb forces.

The density of states (DOS) of a $V_O^\times$ defect, shown in Fig. 4.6(b) for a $2 \times 2 \times 1$ computational supercell, contains a defect state ca. 1.8 eV above the VBM in the band gap, as reported in other studies [122, 174, 190, 170]. The position of this state is strongly affected by the DFT band gap error, with optical absorption data [53] placing it ca. 6 eV above the VBM. The defect level contains the two electrons added to the supercell after creating the oxygen vacancy, causing the Fermi level $\epsilon_F$, indicated by an arrow, to be pinned at this highest-occupied state. A three-dimensional (3D) plot of the DFT isoelectronic surface in Fig. 4.7 shows that the defect state electron density is highly localised, primarily comprising a diffuse $s$ orbital centred on the vacancy, with additional residue in the lobes of nearest-neighbour oxygen $2p$ states, similar to previous findings [170, 190, 122]. This defect is the well-known $F$-centre in alumina, and removing one electron to create a $V_O$ defect gives rise to an $F^+$-centre. The $s$-like defect orbital has considerable conduction-band character and its energy is therefore underestimated due to the DFT band gap error. An accurate treatment of these states within DFT requires a hybrid functional capable of reproducing the experimental band gap and correctly localising the states.

Aluminium (split) vacancy

Most studies of point defects in alumina assume that the ground-state structure of an aluminium vacancy, like an oxygen vacancy, is found by removing a single aluminium atom from a computational supercell and relaxing the structure. However, this study demonstrates the existence of a previously unseen lower-energy defect state
Figure 4.6: DFT density of states (DOS) of bulk Al$_2$O$_3$, a neutral oxygen vacancy ($V^\times_O$), aluminium vacancy ($V^\times_{Al}$), oxygen interstitial ($O^\times_i$), and aluminium interstitial ($Al^\times_i$), in a $2 \times 2 \times 1$ supercell of alumina. The VBM, which is at the upper edge of the oxygen 2$p$ band, is located at 0eV, and the band gap $E^g_0$ (shown by dashed lines) extending to the aluminium 3$s$ and 3$p$ states is underestimated at ca. 6.88eV (experiment: 9.57eV [64]). For the $V^\times_O$ defect, the Fermi level $\epsilon_F$, indicated by an arrow, lies at the energy of the localised defect orbital shown in Fig. 4.7. For $V^\times_{Al}$, there are no occupied states within the gap, and $\epsilon_F$ lies at the top of the three bound hole states, just above the VBM. Adding electrons to this defect contributes weight to neighbouring oxygen 2$p$ lobes (see Fig. 4.12). The oxygen 2$p$ orbitals on an $O^\times_i$ dumbbell create a defect state just above the VBM, displayed in Fig. 4.15 for $O^\times'_i$. $Al^\times_i$ defects fill two defect states below and at the CBM with three electrons, as detailed in Fig. 4.16.
that cannot be found by this simple procedure.

After an aluminium atom is removed, it is energetically favourable (by ca. 1.1 eV) for the next nearest-neighbour to the vacant aluminium site along the c-axis to migrate into the normally vacant octahedral interstitial site, as depicted in the schematic in Figs. 4.8(a)-(c). This leaves two normally-occupied aluminium sites vacant, while the aluminium atom that migrated sits in an interstitial site. Removing an aluminium atom from the perfect lattice and relaxing the geometry in an electronic structure calculation does not yield this energy minimum, as the optimiser cannot escape the local minimum of the initial ‘simple’ aluminium vacancy configuration. The resulting defect is referred to in this work as an ‘aluminium split vacancy’ and is equivalent to removing two aluminium atoms separated by a vacant octahedral site, and placing an interstitial aluminium atom in this octahedral site. In this sense, it can be viewed as an extended defect containing three lattice sites, as indicated by the dotted rectangle in Fig. 4.8(c).

The aluminium split vacancy defect was found while examining the migration barrier for a simple aluminium vacancy. Starting from an initial configuration where a single aluminium atom is removed from the lattice, there are five inequivalent nearest-neighbour aluminium atoms that can migrate into the now vacant site, shown in Figs. 4.9(a) and (b). The migration barriers for each of these atoms were calculated using DFT in the LDA for a $2 \times 2 \times 1$ supercell. The resulting migration barriers are shown in Fig. 4.10 as a function of the fractional distance along the jump vector between the beginning and end points of each migration. At each calculation step, the migrating atom is allowed to relax in a plane perpendicular to its jump vector.
Figure 4.8: Schematic depicting the creation of an aluminium split vacancy defect. Initially (a), an aluminium vacancy is created, indicated by the hollow aluminium atom encircling an ×. Then, the next nearest-neighbour aluminium atom along the c-axis relaxes (indicated by an arrow with subscript r) into the normally vacant octahedral site (b). As the aluminium atom now resides in an interstitial position, it is designated with a square instead of a circle. In (c), the extended split vacancy defect comprising three lattice sites is denoted by a dotted rectangle, and the substantial lattice contraction shown.

(as described in Sec. 3.6.2). All other atoms in the supercell must also be allowed to relax, as the migrating atom can cause significant relaxations of nearby atoms\(^7\). The lowest-energy path (Path 5 in Fig.4.10) has an energy minimum lower than that at the beginning of the path, indicating a lower-energy ground-state configuration. The position of the energy minimum corresponds to the next nearest-neighbour aluminium atom along the c-axis to the aluminium vacancy migrating into the normally vacant octahedral interstitial site to create the aluminium split vacancy. In a 2 × 2 × 1 computational supercell, the split vacancy ground-state configuration is 0.71 ± 0.05eV lower in energy than the simple vacancy configuration, where the 0.05eV uncertainty results from differences between basis sets (plane waves or Gaussians) and functionals (LDA, GGA, B3LYP) in DFT. The good agreement of results obtained using different approximations is encouraging. Diffusion Monte

\(^7\)In fact, relaxing only nearest and next-nearest neighbours to the migrating atom can create regions near the beginning and end points of the path with unphysical negative barriers. This happens because different sets of atoms are allowed to relax, depending on the location of the migrating atom.
Section 4.4: Density functional theory calculations

Figure 4.9: Nearest-neighbour aluminium atoms to a ‘simple’ aluminium vacancy defect in a $2 \times 2 \times 1$ supercell. The vacancy is indicated with an arrow. The $c$-axis points out of the paper in the left-hand figure (a), and up in the right-hand figure (b). The five inequivalent nearest-neighbour aluminium atoms are coloured green and labelled with numbers. Along the $c$-axis, aluminium atom number 4 is a nearest neighbour of the vacancy, whereas atom number 5 is a next-nearest neighbour, separated by the vacant octahedral interstitial site, designated with an $i$.

Carlo (DMC) calculations described in Sec. 4.8.5 using the same supercell reinforce these results, calculating an energy difference between the simple and split vacancies of $0.72 \pm 0.40 \text{eV}$. The energy difference was determined for cells of varying sizes and shapes in order to extrapolate to the infinite-cell limit using the Madelung interpolation method described in Sec. 3.5.3. The results are shown in Fig. 4.11 as a function of the Madelung potential $v_M$, demonstrating a significant finite-size effect due to elastic effects along the $c$-axis. The extrapolated value in the infinite-cell limit (i.e. $\lim_{v_M \to 0}$) is $1.22 \text{eV}$.

Several classical potentials were tested to determine whether they could correctly reproduce the split vacancy configuration as the ground-state structure of the aluminium vacancy. Shell potentials from Catlow et al. [27], Gale et al. [65], Kislov et al. [94], and Lewis and Catlow [109], as well as a Buckingham potential from Sun et al. [172] were used to relax $2 \times 2 \times 1$ and $4 \times 4 \times 2$ supercells of alumina containing a single aluminium vacancy (where the optimised lattice parameters had been previously found from the rhombohedral unit cell). All remained in the simple vacancy configuration. Then, the next nearest-neighbour aluminium atom along the $c$-axis...
Figure 4.10: Energy barriers for the migration of the five inequivalent aluminium atoms from Fig. 4.9 to the vacant aluminium lattice site using Vanderbilt USPs in the LDA in a $2 \times 2 \times 1$ supercell. The migrating distances along each path are shown in the legend. Atom numbers 4 and 5, both migrating along the $c$-axis, are inequivalent due to the coordinated stacking of aluminium layers: number 4 is a nearest neighbour to the vacant site, while number 5 is a next-nearest neighbour, separated by the vacant octahedral interstitial site. The lowest maximum of all paths belongs to path 5, which has a minimum 0.71 eV lower than the initial stable configuration, the ‘simple’ aluminium vacancy. This configuration, called the ‘aluminium split vacancy’, is obtained when the next-nearest neighbour aluminium atom along the $c$-axis migrates into the normally vacant octahedral interstitial site, leaving empty lattice sites either side of the interstitial site, as depicted in Fig. 4.8.

was placed in the octahedral interstitial site, and the entire supercell relaxed while the migrating aluminium atom was only allowed to relax in a plane perpendicular to the jump vector. This was done so that, if there was no energy well for the split vacancy configuration using a given potential, the aluminium atom would not migrate back into its original position. The differences in energy between the split vacancy and the simple vacancy configurations are shown in Table 4.5. Only one potential, from Gale et al., gave a lower energy for the split vacancy configuration, and the parameters for this particular potential were the only ones not chosen empirically, but rather by fitting to the results of ab initio Hartree-Fock calculations. This is not surprising as most potentials (that of Gale et al. being the only exception here) do not even yield corundum as the correct ground-state structure for bulk alumina, but rather bixbyite. The shape of the migration barrier for the migrating aluminium
atom was also analysed for all the potentials. Unsurprisingly, only that of Gale et al. mirrored the behaviour of DFT, whereas all others simply produced a Gaussian-like, non-negative diffusion barrier. The potential of Kislov et al. produced neither of these two results, but rather yielded an extremely low (< 0.1 eV), roughly constant barrier along most of the migration path. This highlights the limitations of simple empirical potentials for describing defects and diffusion in alumina. More advanced potentials, such as those from Streitz and Mintmire [171] using the embedded-atom method, or Jahn et al. [88], using the aspherical ion model, may perform better, but are limited in implementation. Newer ab initio-based classical potentials, with parameters determined from a large ‘learning set’ of DFT calculations [17, 136, 22], are likely to be able to reproduce the split vacancy as the ground-state configuration and account for other effects only seen in ab initio calculations. Unfortunately, the computational cost of the $O[100 - 1000]$ DFT calculations required to generate the potential already eclipses that needed for a full ab initio evaluation of the defects. Such a classical potential would; however, come into its own for studies of defect dynamics.

The relaxations of the nearest neighbours of the aluminium split vacancy are tabulated in Table 4.4. As the centre of the split vacancy lies at the now occupied octa-
Table 4.5: The results of shell model calculations of the split vacancy configuration. The left-hand column lists different shell models. The right-hand column lists the differences between the total energy of a $2\times2\times1$ computational supercell containing an aluminium split vacancy $E_{V_{A1(s)}}^{\text{tot}}$, and one containing a simple aluminium vacancy, $E_{V_{A1}}^{\text{tot}}$. Only one potential, that of Gale et al., finds the split vacancy configuration to be energetically favourable relative to the ‘simple’ vacancy.

<table>
<thead>
<tr>
<th>Potential</th>
<th>$E_{V_{A1(s)}}^{\text{tot}} - E_{V_{A1}}^{\text{tot}}$ (eV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Catlow et al. [27]</td>
<td>+0.26</td>
</tr>
<tr>
<td>Gale et al. [65]</td>
<td>−0.44</td>
</tr>
<tr>
<td>Kislov et al. [94]</td>
<td>+0.07</td>
</tr>
<tr>
<td>Lewis and Catlow [109]</td>
<td>+0.10</td>
</tr>
<tr>
<td>Sun et al. [172]</td>
<td>+0.21</td>
</tr>
</tbody>
</table>

hedral interstitial site, the relaxations are referenced to the interstitial site. The first nearest-neighbours, normally two aluminium atoms in a perfect crystal, no longer exist, as one has been removed to create the vacancy, and the other has migrated into the interstitial site, so these distances are omitted. Beyond this, the six nearest oxygen atoms relax towards the split vacancy, rather than away, as in the case of a simple vacancy. The six next-nearest neighbour aluminium atoms relax outwards. The two next-nearest neighbour aluminium atoms along the $c$-axis strongly relax towards the defect, causing their spacing to reduce from 9.00Å in the perfect crystal to 8.58Å. None of these relaxations depends strongly on the charge state of the split vacancy.

The DOS for a neutral $V_{A1}^s$ defect, shown in Fig. 4.6(c), shows no occupied states within the band gap, and $\epsilon_F$ lies at the bound valence band states that are just above the VBM. When electrons are added to the defect, for example to create an $V_{A1}^{m}$ defect, the highest-occupied states are the oxygen $2p$ orbitals filled by the additional electrons, whose isosurface is shown in Fig. 4.12(a) and (b). These electrons are highly localised on the nearest-neighbour oxygen atoms of the two aluminium atoms flanking the aluminium atom sitting in the octahedral interstitial site along the $c$-axis.

When multiple aluminium split vacancies are created on a common vector along the $c$-axis, the defects prefer not to lie exactly on the $c$-axis, but rather slightly off-centred on the $a$-axis by ca. 0.01Å. This staggered stacking lowers the energy very slightly, for two vacancies in a $2\times2\times3$ supercell by ca. 0.1eV. This effect
Section 4.4: Density functional theory calculations

Figure 4.12: The $|\psi(r)|^2 = 0.1$ isosurface of the highest occupied molecular orbital (HOMO) in a $2 \times 2 \times 1$ supercell of alumina containing a single $V'''_{\text{Al}}$ defect in the split vacancy configuration, projected onto (a) the $ab$-plane and (b) $ac$-plane. The highest occupied state is made primarily of $2p$ orbitals from oxygen atoms surrounding the neighbouring aluminium atoms along the $c$-axis. The aluminium atom at the centre of the split vacancy resides on the octahedral interstitial site, denoted by $i$ in the right-hand figure.

cannot be observed when creating a single vacancy, and requires very large supercells to analyse accurately, beyond current computational capabilities. As we are only concerned with defects in the dilute limit, it suffices for our purposes to consider single aluminium split vacancies that lie on the $c$-axis, and we ignore this effect. Future developments in linear-scaling DFT [67] could provide further insight into this phenomenon.

Oxygen (dumbbell) interstitial

We examine two different oxygen interstitial configurations: a simple interstitial and a dumbbell interstitial. The first, commonly considered in other studies [44, 122, 98, 7], occupies the same site as an aluminium interstitial, the octahedral interstitial site. However, electrostatics suggest that this site is fairly unaccommodating. A simple oxygen interstitial is surrounded by six oxygen nearest-neighbours at a distance of 1.82Å in the form of two tetrahedra whose common vertex is the interstitial atom (akin to Fig. 4.13(a) for aluminium interstitials). Previous calculations demonstrate a very high defect formation energy for this configuration. For example, Dienes et al. [44] find $\Delta G^f_{i(\text{oct})} = 10.5$eV, where the subscript ‘oct’ indicates the octahedral interstitial site. From a physical standpoint, this configuration does not seem likely, especially given the large space available elsewhere in corundum, such as between the stacked aluminium and oxygen layers.
Figure 4.13: Left (a): a $2 \times 2 \times 1$ supercell of $\text{Al}_2\text{O}_3$ containing an $\text{Al}^{3-}$ defect (black) and its six oxygen nearest-neighbours (green), forming two tetrahedra. Right (b): schematic of an aluminium interstitial defect. The aluminium interstitial atom, indicated by a blue square, resides on the normally vacant octahedral interstitial site that separates groups of aluminium atoms (blue circles) along the $c$-axis (pointing up). The lattice expansion along the $c$-axis is indicated by the outward relaxation of the nearest-neighbour aluminium atoms (values represent $\text{Al}^{3-}$ vs. perfect crystal).

In this study, the dumbbell interstitial configuration was found by examining the migration barrier of a simple oxygen interstitial atom from one octahedral site to another. This was done as described in Sec. 3.6.2. When performing this procedure on an oxygen interstitial atom moving from an octahedral interstitial site to any nearest-neighbour equivalent interstitial site, the oxygen atom moved perpendicular to the migration vector to form a dumbbell interstitial configuration with a neighbouring oxygen lattice atom for all 15 trial steps along the vector. In a $2 \times 2 \times 1$ computational supercell, the dumbbell interstitial configuration is 0.58eV lower in energy than the simple interstitial configuration. DMC gives a very similar energy difference of $0.60 \pm 0.10$eV. In addition, displacing a simple oxygen interstitial by more than 0.1Å in the $ab$-plane from the octahedral interstitial site and allowing the geometry to fully relax also causes the oxygen atom to migrate towards the nearest-neighbour oxygen atom (dependent on the exact direction of the displacement) to form a dumbbell, rather than into the interstitial site. The octahedral interstitial site contains a small local minimum which the standard geometry optimiser generally cannot escape, but is certainly not the global minimum.
Figure 4.14: The oxygen dumbbell interstitial defect in a $2 \times 2 \times 1$ computational supercell, projected (a) onto the $ab$- and (b) onto the $ac$-planes. The two oxygen atoms comprising the dumbbell interstitial are coloured green.

When an oxygen dumbbell interstitial is created, an oxygen lattice atom is displaced from its original site and forced out of its stacked oxygen layer into a neighbouring aluminium layer (see Figs. 4.14(a) and (b)). The oxygen interstitial atom resides opposite this atom in the other aluminium layer sandwiching the layer of oxygen atoms from which the lattice oxygen atom originated. The dumbbell has a bond length of 2.13Å, centred on the now empty lattice site of one of the oxygen dumbbell atoms, and is easily rotated around the crystal axes. The energetic cost for rotating up to 15° from its energy minimum around either the $a$-, $b$-, or $c$-axis is less than 0.2eV.

The geometry surrounding the oxygen dumbbell interstitial, centred on an oxygen lattice site, relaxes as shown in Table 4.4. For charged $O_i$ defects, the four nearest-neighbour aluminium atoms relax towards the dumbbell, whereas they relax away from $O_i^x$ defects. The twelve second-nearest neighbour oxygen atoms are pushed away from the dumbbell by increasing amounts for higher charge states, whereas the converse relationship holds for the two next nearest-neighbour aluminium atoms.

None of the classical potentials applied earlier to the aluminium split vacancy defect is able to find the ground-state of the oxygen interstitial defect correctly. In fact, none of the potentials used in this study even finds a local minimum when an oxygen dumbbell interstitial obtained from a DFT calculation is allowed to relax. Instead, the atoms spontaneously return to the ‘simple’ configuration in which the oxygen interstitial atom resides in the octahedral site. Interestingly, Catlow et al. [27], who
Figure 4.15: The $|\psi(r)|^2 = 0.1$ isosurface of the HOMO in a $2 \times 2 \times 1$ supercell of alumina containing a single O''$_i$ dumbbell defect, projected onto (a) the ab- and (b) the ac-planes (with atoms explicitly shown for clarity). Most of the weight associated with the highest-occupied state is concentrated on the $2p$ orbitals of the two oxygen atoms comprising the oxygen interstitial dumbbell defect, with smaller weights on neighbouring oxygen $2p$ lobes.

compared defect energetics for two potentials, one with parameters fit to experimental lattice data and one derived from Hartree-Fock calculations, noted a large difference in formation energy between the two only for simple oxygen interstitial defects. This could be because the potential fitted to experimental data had no knowledge in its ‘training set’ of configurations in which oxygen ions approached each other closely. The resulting O$^{2-}$-O$^{2-}$ potential is therefore too repulsive at short range. More advanced potentials, or those based on a significant training set of DFT calculations, are essential for studying defects in alumina.

The DOS of an O$^x_i$ defect, shown in Fig. 4.6(d), has a defect state just above the VBM, constructed from the $2p$ lobes of both atoms in the dumbbell, and, to a lesser extent, from adjacent oxygen $2p$ orbitals. In the O''$_i$ defect, this defect level is occupied, as demonstrated by the HOMO isosurface shown in Fig. 4.15.

Aluminium interstitial

Aluminium interstitials occupy the normally vacant octahedral site in the lattice, creating a continuous sequence of five aluminium atoms along the c-axis, as depicted in Figs. 4.13(a) and (b). Although the ground-state configuration of the defect is straightforward, its migration, described in Sec. 4.4.8, is not.
4.4: Density functional theory calculations

Figure 4.16: The 3D isosurface corresponding to $|\psi(r)|^2 = 0.1$ in a $2 \times 2 \times 1$ supercell of alumina containing (a) a single Al$_i^\prime$ or (b) an Al$_i^\times$ defect, projected onto the ab-plane. For the Al$_i^\prime$ defect, the highest-occupied state is strongly localised on the next-nearest neighbour oxygen 2$p$ and aluminium 3$s$ orbitals, whereas the additional electron added to create an Al$_i^\times$ defect is substantially more delocalised across the supercell.

The geometrical relaxations surrounding an aluminium interstitial defect in a $2 \times 2 \times 1$ supercell, as calculated using the LDA, are summarised in Table 4.4 and generally show a strong outward relaxation of the two nearest-neighbour aluminium ions, and an inward relaxation of the six second-nearest neighbour oxygen ions. However, for the neutral Al$_i^\times$ defect, the nearest-neighbour oxygen ions relax away from the defect site, breaking this trend. These results compare favourably with previous DFT calculations of the aluminium interstitial [122, 82].

The density of states (DOS) of the Al$_i^\times$ defect, shown in Fig. 4.6(e) for the $2 \times 2 \times 1$ supercell, features a defect state high in the band gap, ca. 4.80eV from the VBM, in good agreement with previous DFT findings [122]. The first two electrons added to an Al$_i^\prime$ defect to create Al$_i^\times$ are localised in this defect state, shown in Fig. 4.16(a). Due to the predominantly aluminium 3$s$ character of the defect state, the defect level sits just below the conduction band edge. The final electron required to create Al$_i^\times$ goes into the highest-occupied state, lying at the CBM and indicated by the arrow in the DOS of Fig. 4.6(e). This state is much more delocalised (see Fig. 4.16(b)).

AlO Vacancy

The complex corundum structure and aluminium split vacancy structure create many possible ground-state configurations for AlO vacancies. In order to enumerate all these possibilities, we create single aluminium vacancies in the simple and
split configurations. For each of these two possibilities, we consider combining the aluminium vacancy with all sensible choices of neighbouring oxygen vacancies. Beginning with the ‘simple’ aluminium vacancy, there are six first-nearest-neighbour oxygen atoms, three at distance 1.82\ Å from the vacancy, and three at distance 1.94\ Å (for simplicity, all distances in this section are as in the perfect crystal). Similarly, there are three second-nearest-neighbour oxygen atoms at a distance of 3.17\ Å, and further neighbours as listed in Table 4.6.

The symmetry of the aluminium split vacancy reduces the number of nearest-neighbour oxygen vacancies to be considered. From the perspective of the aluminium atom that resides in the octahedral interstitial site, there are only three inequivalent shells of nearest-neighbours less than 4\ Å away, rather than five. Oxygen vacancies further than 4\ Å from an aluminium vacancy were not considered.

Creating either a relaxed simple or a split aluminium vacancy, and then removing one of the neighbouring oxygen atoms listed in Table 4.6 and relaxing the structure, leads to \( i = 9 \) different ground-state \( V'_\text{AlO}(i) \) structures. Each structure has a ground-state energy, referenced in Table 4.6 to the energy of the most stable configuration found, which consists of a simple aluminium vacancy with a nearest-neighbour oxygen vacancy. Interestingly, when creating AlO vacancies, it is favourable to have a simple, rather than a split, aluminium vacancy.

The DOS of a \( V''''\text{AlO} \) defect, shown in Fig. 4.17(a), is similar to a superposition of the DOS’s of the \( V'''\text{Al} \) and \( V''\text{O} \) defects. As in the case of \( V''\text{O} \), a state of \( s \) character appears near the middle of the band gap, shown in the isosurface plot in Fig.4.17(b). However, this defect state lies significantly higher, ca. 1.8eV, in the gap than for \( V''\text{O} \). Removing electrons to create a \( V'_\text{AlO} \) defect causes \( \epsilon_F \) to fall to the VBM, as the highest-occupied state then consists almost solely of weight on neighbouring oxygen 2p lobes, as for \( V'''\text{Al} \).

Dopants

In the mass-action approach outlined in Sec. 3.3.2, dopants play a crucial role in linking experimental diffusion data with the calculated formation energies of defects and defect-dopant clusters. For a given target dopant concentration, comprising free dopants and dopants bound into defect-dopant clusters, the dopant chemical potential is fixed. The concentrations of all defects and clusters in the crystal are then determined according to constraints such as charge neutrality. From these
### Section 4.4: Density functional theory calculations

<table>
<thead>
<tr>
<th>Al vacancy type</th>
<th>O neighbourhood</th>
<th>(d_{\text{Al-O}})</th>
<th>(E_{\text{V}<em>{\text{AlO}}}^{\text{tot}}(i) - \min \left{ E</em>{\text{V}_{\text{AlO}}}^{\text{tot}}(i) \right})</th>
</tr>
</thead>
<tbody>
<tr>
<td>simple</td>
<td>1 NN (tot. 3)</td>
<td>1.82 Å</td>
<td>0.00eV</td>
</tr>
<tr>
<td>simple</td>
<td>1 NN (3)</td>
<td>1.94 Å</td>
<td>0.88eV</td>
</tr>
<tr>
<td>simple</td>
<td>2 NN (3)</td>
<td>3.17 Å</td>
<td>1.75eV</td>
</tr>
<tr>
<td>simple</td>
<td>3 NN (3)</td>
<td>3.37 Å</td>
<td>1.97eV</td>
</tr>
<tr>
<td>simple</td>
<td>4 NN (3)</td>
<td>3.50 Å</td>
<td>1.57eV</td>
</tr>
<tr>
<td>simple</td>
<td>5 NN (3)</td>
<td>3.80 Å</td>
<td>1.51eV</td>
</tr>
<tr>
<td>split</td>
<td>1 NN (6)</td>
<td>1.94 Å</td>
<td>0.86eV</td>
</tr>
<tr>
<td>split</td>
<td>2 NN (6)</td>
<td>3.24 Å</td>
<td>1.10eV</td>
</tr>
<tr>
<td>split</td>
<td>3 NN (6)</td>
<td>3.50 Å</td>
<td>0.35eV</td>
</tr>
</tbody>
</table>

**Table 4.6:** Relative ground-state energies of possible AlO vacancy defect configurations in a \(2 \times 2 \times 1\) computational supercell. In all cases, the aluminium vacancy assumes either a simple or split configuration. Simple \(V_{\text{Al}}\) defects have six inequivalent neighbouring oxygen atoms (each three-fold degenerate) less than 4 Å from the vacancy site, at distances \(d_{\text{Al-O}}\), shown in the third column. Similarly, split \(V_{\text{Al}}\) defects have three inequivalent oxygen neighbours, each six-fold degenerate. The final column tabulates the ground-state energy of each particular AlO vacancy configuration, relative to the configuration that has the lowest total energy. The most stable configuration consists of a simple aluminium vacancy and an oxygen vacancy on the nearest available lattice site.

**Figure 4.17:** The DOS (a) and a 3D isosurface of the highest-occupied state (b) for a \(V_{\text{AlO}}^{\prime\prime\prime}\) defect. The calculations were performed in a \(2 \times 2 \times 1\) supercell of alumina, using Vanderbilt USPs in the LDA. \(\epsilon_F\) (shown by an arrow in (a)) is pinned by a defect state ca. 3.6eV above the VBM (which is set to 0eV). The \(|\psi(\mathbf{r})|^2 = 0.1\) isosurface of this defect state, shown in (b), is projected onto the \(ab\)-plane. The lion’s share of weight lies in a diffuse \(s\)-like orbital on the vacant oxygen site, just as for a \(V_{\text{O}}^{\prime}\) defect. Removing two electrons to create \(V_{\text{AlO}}^{\prime}\) yields a highest state similar in character to that of a \(V_{\text{Al}}^{\prime\prime\prime}\) defect, primarily weighted on neighbouring oxygen \(2p\) orbitals.
Table 4.7: The geometric relaxations of nearest-neighbours to interstitial and substitutional dopants in alumina, calculated using Vanderbilt USPs in the LDA. For each defect species, the distances to (in Å) and identities of the first, second, and third nearest-neighbours (1 NN, etc.) in the perfect crystal are shown, followed by the corresponding distances and percentage changes after the geometry has been relaxed. Interstitials impart significantly larger strains on the lattice, causing non-negligible long-range relaxations.

<table>
<thead>
<tr>
<th>Defect species</th>
<th>1 NN Change</th>
<th>2 NN Change</th>
<th>3 NN Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bulk Al</td>
<td>1.88 (6 O)</td>
<td>2.72 (4 Al)</td>
<td>3.17 (3 Al, 3 O)</td>
</tr>
<tr>
<td>Mg(_{\text{Al}})</td>
<td>2.00 +6.38%</td>
<td>2.75 +1.10%</td>
<td>3.15 −0.63%</td>
</tr>
<tr>
<td>Mg(_{\text{Al}})</td>
<td>2.01 +6.91%</td>
<td>2.75 +1.10%</td>
<td>3.14 −0.95%</td>
</tr>
<tr>
<td>Mg(_{\text{Al}})</td>
<td>1.88 ±0.00%</td>
<td>2.74 +0.74%</td>
<td>3.17 ±0.00%</td>
</tr>
<tr>
<td>Interstitial (\text{i})</td>
<td>1.89 (2 Al)</td>
<td>1.94 (6 O)</td>
<td>2.71 (6 Al)</td>
</tr>
<tr>
<td>Mg(_{\text{i}})</td>
<td>2.25 +19.05%</td>
<td>1.94 ±0.00%</td>
<td>2.78 +2.58%</td>
</tr>
<tr>
<td>Mg(_{\text{i}})</td>
<td>2.24 +18.52%</td>
<td>1.93 −0.52%</td>
<td>2.78 +2.58%</td>
</tr>
<tr>
<td>Mg(_{\text{i}})</td>
<td>2.25 +19.05%</td>
<td>1.93 −0.52%</td>
<td>2.78 +2.58%</td>
</tr>
<tr>
<td>Bulk Al</td>
<td>1.88 (6 O)</td>
<td>2.72 (4 Al)</td>
<td>3.17 (3 Al, 3 O)</td>
</tr>
<tr>
<td>Ti(_{\text{Al}})</td>
<td>1.93 +2.66%</td>
<td>2.81 +3.31%</td>
<td>3.19 +0.63%</td>
</tr>
<tr>
<td>Ti(_{\text{Al}})</td>
<td>1.93 +2.66%</td>
<td>2.81 +3.31%</td>
<td>3.20 +0.95%</td>
</tr>
<tr>
<td>Ti(_{\text{Al}})</td>
<td>1.98 +5.32%</td>
<td>2.78 +2.21%</td>
<td>3.18 +0.32%</td>
</tr>
<tr>
<td>Interstitial (\text{i})</td>
<td>1.89 (2 Al)</td>
<td>1.94 (6 O)</td>
<td>2.71 (6 Al)</td>
</tr>
<tr>
<td>Ti(_{\text{i}})</td>
<td>2.12 +12.17%</td>
<td>2.06 +6.19%</td>
<td>2.74 +1.11%</td>
</tr>
<tr>
<td>Ti(_{\text{i}})</td>
<td>2.17 +14.81%</td>
<td>2.03 +4.64%</td>
<td>2.79 +2.95%</td>
</tr>
<tr>
<td>Ti(_{\text{i}})</td>
<td>2.22 +17.46%</td>
<td>1.99 +2.58%</td>
<td>2.81 +3.69%</td>
</tr>
<tr>
<td>Ti(_{\text{i}})</td>
<td>2.29 +21.16%</td>
<td>1.94 ±0.00%</td>
<td>2.83 +4.43%</td>
</tr>
<tr>
<td>Ti(_{\text{i}})</td>
<td>2.35 +24.33%</td>
<td>1.89 −2.58%</td>
<td>2.84 +4.80%</td>
</tr>
</tbody>
</table>

concentrations, diffusion coefficients are then deduced.

The two most common aliovalent impurities found in experimental samples of Al\(_2\)O\(_3\) are magnesium and titanium. During the calculations in Sec. 4.4.3, we find that interstitial magnesium and titantium impurities have significantly higher formation energies than substitutionals, and are therefore present in virtually negligible concentrations under all feasible experimental conditions.

Both defects are very straightforward: substitutionals are found on a normally occupied aluminium lattice site, interstitials on the normally vacant octahedral interstitial site. The atomic relaxations surrounding these impurities are summarised in Table 4.7. Interstitials place much more strain on the lattice than substitutionals, causing their aluminium nearest-neighbours to strongly relax outwards, and exhibiting significant long-range relaxations up to third-nearest neighbours.
The DOS’s of magnesium and titanium substitutionals contain defect states in the band gap (see Fig. 4.18). The highest-occupied state for Mg\textsubscript{Al}, shown in an isosurface plot in Fig. 4.19(a), is largely constructed from nearest-neighbour oxygen 2\textit{p} orbitals. Removing an electron to create a neutral Mg\textsubscript{Al}\textsuperscript{x} defect reveals a similar DOS with slightly less weight on the nearest-neighbour oxygen 2\textit{p} orbitals, leaving a localised hole behind in the 2\textit{p} band, as discussed in previous empirical potential calculations (see Ref. [27] and references therein and [127]) and reported from experimental electron spin resonance (ESR) [38] and optical absorption evidence [187], as well as thermally stimulated polarisation and depolarisation experiments [183]. We find that the Mulliken charges only change significantly between the Mg\textsubscript{Al} and Mg\textsubscript{Al}\textsuperscript{x} defects for the six nearest-neighbour oxygen atoms of the magnesium substitutional. The hole is bound to the substitutional as it costs Coulomb energy to pull the positively-charged hole away from the negatively-charged magnesium substitutional that would be left behind in its absence.

Titanium substitutionals feature two defect states close together, ca. 2.2eV below the conduction band edge. For a Ti\textsubscript{Al}\textsuperscript{x} defect, the first of these states is occupied and resembles a titanium 3\textit{d} orbital localised on the defect site (see the isosurface in Fig. 4.19(b)). A previous DFT study by Matsunaga \textit{et al.} [121] on titanium substitutionals and interstitials also found defect levels in the band gap mainly composed from titanium 3\textit{d} orbitals strongly localised on the defect site, and Mohapatra \textit{et al.} confirm these findings from optical data [128]. For titanium interstitials, both the present work and Matsunaga’s study find defect states with a similar nature.

\textbf{Defect-Dopant Clusters}

Defect-dopant clusters bind charge-compensating defects to impurities. Harding \textit{et al.} [75] have suggested that this binding causes the buffering effect experimentally observed for diffusion in doped alumina. We examine this hypothesis using DFT studies in Sec. 4.5. Although a raft of different clusters is possible and 23 have been specifically considered in this study, only those that we find to exist in significant concentrations and selected other examples will be highlighted here.

Under magnesium doping, only two prevalent clusters form: trapped vacancies, \{Mg\textsubscript{Al} : V\textsubscript{O}\}, which are considered for the charge states (0, +1, +2), and trapped interstitials, \{Mg\textsubscript{Al} : Al\textsubscript{i}\}, considered for charge states (+1, +2, +3), both shown schematically in Fig. 4.20. Other charge states, as well as other magnesium-based
Figure 4.18: The densities of states of $2 \times 2 \times 1$ supercells containing (a) magnesium and (b) titanium substitutionals, from DFT calculations using Vanderbilt USPs in the LDA. The Fermi level $\epsilon_F$ (indicated by an arrow) is set by defect states in the band gap for both substitutionals. For Ti$_{Al}$, the state, shown in an isosurface plot in Fig. 4.19(b), lies 2.2eV below the conduction band and comprises titanium 3$d$ states. For Mg$_{Al}$, the state lies ca. 0.5eV above the VBM and comprises primarily nearest-neighbour oxygen 2$p$ orbitals, as seen in Fig. 4.19(a). Removing an electron to create Mg$_{Al}$ causes $\epsilon_F$ to fall to the top of the VBM, leaving a localised hole above the oxygen 2$p$ band.

Figure 4.19: 3D isosurfaces describing the highest-occupied state for Mg$_{Al}$ ((a), defect coloured gold) and Ti$_{Al}$ ((b), defect coloured pink) substitutionals, corresponding to $|\psi(r)|^2 = 0.1$, projected onto the ab-plane. Both states are localised around the defect site, consisting of fully-occupied nearest-neighbour oxygen 2$p$ orbitals in the case of Mg$_{Al}$, and partially-occupied titanium 3$d$ orbitals in the case of Ti$_{Al}$. 
Section 4.4: Density functional theory calculations

Figure 4.20: A schematic of the dominant clusters in Mg-doped alumina: \{\text{Mg}\text{Al} : V_O\} and \{\text{Mg}\text{Al} : \text{Al}_i\}. For \{\text{Mg}\text{Al} : V_O\}, the oxygen vacancy is indicated by a circle containing an \times, and the magnesium substitutional by a brown triangle. For \{\text{Mg}\text{Al} : \text{Al}_i\}, the aluminium interstitial is symbolised by a blue square in an interstitial lattice site neighbouring the Mg\text{Al} substitutional.

clusters, exist, but are more than 20 orders of magnitude lower in concentration. For both dominant clusters, the native point defect and the impurity are nearest neighbours and it is found that moving the defects further apart incurs an enthalpic penalty. These and other clusters clusters resemble superpositions of native defects and dopants; their DOSes and isosurfaces can be inferred from their constituent defects and so are not presented here.

For titanium-doped sapphire, a number of clusters are significant to the defect chemistry. The most abundant are \{\text{Ti}\text{Al} : V_{\text{Al}}\}, found in charge states \((-3, -2, -1)\), and \{2\text{Ti}\text{Al} : V_{\text{Al}}\}, in charge states \((-2, -1, 0)\). Both of these clusters contain an aluminium split vacancy, with Ti\text{Al} substitutionals in various positions.

For \{\text{Ti}\text{Al} : V_{\text{Al}}\}, the three possibilities with relatively low ground-state energies are shown in Fig. 4.21(a). The titanium substitutional can either occupy an aluminium lattice site neighbouring the split vacancy perpendicular to the c-axis, a nearest-neighbour aluminium site along the c-axis, or the interstitial lattice site in the centre of the split vacancy. Although the last configuration is in fact a combination of one Ti\text{i} and two V\text{Al} defects, we will group it as a \{\text{Ti}\text{Al} : V_{\text{Al}}\} cluster for simplicity, due to its identical charge states and similar structure.

For \{2\text{Ti}\text{Al} : V_{\text{Al}}\} clusters, there are also three possible combinations with comparatively low ground-state energies, shown in Fig. 4.21(b). The Ti\text{Al} defect can occupy
Figure 4.21: Schematics of the dominant clusters in Ti-doped alumina: (a) \{Ti_{Al} : V_{Al}\}; and (b) \{2Ti_{Al} : V_{Al}\}. Aluminium vacancies are symbolised by small circles with an $\times$, aluminium atoms residing in the interstitial site at the centre of the split vacancy by blue squares, titanium atoms in this site by green squares, and titanium substitutionals by green triangles. For \{Ti_{Al} : V_{Al}\}, the three relevant clusters differ in the position of the titanium atom, which can either reside in the interstitial site at the centre of the split vacancy (1), on a nearest-neighbour aluminium site along the c-axis (2), or on a nearest-neighbour perpendicular to the c-axis (3). Similarly, for \{2Ti_{Al} : V_{Al}\}, the titanium atoms can lie on the aluminium sites flanking the split vacancy (1), one of these sites and the interstitial site (2), or the two nearest neighbours to the split vacancy (3). All other configurations have comparatively high energies.

The two aluminium sites neighbouring the split vacancy perpendicular to the c-axis, the two nearest-neighbour aluminium sites flanking the split vacancy along the c-axis, or one of the nearest-neighbours along the c-axis and the interstitial site in the centre of the split vacancy. The latter cluster is really a combination of one Ti$_i$, one Ti$_{Al}$, and one V$_{Al}$ defect, but is grouped as a \{2Ti$_{Al}$ : V$_{Al}$\} cluster for simplicity.

The relative energetics of these configurations of \{Ti$_{Al}$ : V$_{Al}$\} and \{2Ti$_{Al}$ : V$_{Al}$\} clusters are listed in Table 4.8, which shows that in both cases, the titanium substitutionals prefer to sit on the neighbouring aluminium sites along the c-axis. Only these configurations are included in the mass action framework as the ground states of both clusters. Combinations comprising Ti$_{Al}$ defects neighbouring the split vacancy along and perpendicular to the c-axis are found to have relatively high ground-state energies and are also excluded from the framework.
Section 4.4: Density functional theory calculations

<table>
<thead>
<tr>
<th>Configuration in Fig. 4.21(a)</th>
<th>$E_{\text{tot}}^{\text{Ti}\text{Al}:\text{VAl}}''(i) - \min \left{ E_{\text{tot}}^{\text{Ti}\text{Al}:\text{VAl}}''(i) \right}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.74 eV</td>
</tr>
<tr>
<td>2</td>
<td>0.00 eV</td>
</tr>
<tr>
<td>3</td>
<td>0.53 eV</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Configuration in Fig. 4.21(b)</th>
<th>$E_{\text{tot}}^{\text{2Ti}\text{Al}:\text{VAl}}'(i) - \min \left{ E_{\text{tot}}^{\text{2Ti}\text{Al}:\text{VAl}}'(i) \right}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.00 eV</td>
</tr>
<tr>
<td>2</td>
<td>0.71 eV</td>
</tr>
<tr>
<td>3</td>
<td>0.48 eV</td>
</tr>
</tbody>
</table>

Table 4.8: Relative ground-state energies of the different $\{\text{Ti}\text{Al}:\text{VAl}\}$ and $\{\text{2Ti}\text{Al}:\text{VAl}\}$ cluster configurations shown schematically in Figs. 4.21(a) and (b) for a $2 \times 2 \times 1$ computational supercell. For each of the three configurations for both clusters (denoted by their number in the schematics), the total energy is shown in relation to the lowest energy of all configurations. In both clusters, the titanium substitutionals prefer to reside on nearest-neighbour aluminium sites along the $c$-axis to the split vacancy defect.

Alternatively, Ti$_{\text{Al}}$ dopants can bind O$_\text{i}$ defects to form $\{\text{Ti}\text{Al}:\text{O}_\text{i}\}$ clusters, with dominant charge states $(-2, -1, 0)$. In this situation, the oxygen interstitial can either reside in the octahedral interstitial site, or form a nearest-neighbour oxygen dumbbell interstitial to the dopant atom. The cluster comprising an oxygen dumbbell is found to have a ground-state energy 1.00 eV lower than the simple oxygen interstitial configuration in a $2 \times 2 \times 1$ supercell, so this configuration was used for all subsequent calculations of formation energies and defect concentrations.

The clusters shown in Fig. 4.22 are examples of $\{3\text{Ti}\text{Al}:\text{VAl}\}$ clusters created by removing an aluminium atom to create a split vacancy, and then substituting titanium for three aluminium atoms, either associated with the vacancy or nearest neighbours to it. $\{3\text{Ti}\text{Al}:\text{VAl}\}$ clusters are found to form in moderate concentrations (for specifics see Sec. 4.5) in a neutral charge state, but do not significantly affect the defect chemistry of doped alumina. Many more clusters were considered in this study, but proved unnecessary to explain the behaviour of doped alumina.

4.4.3 Defect formation energies and cluster binding energies

Using the methodology presented in Sec. 3.4, defect formation energies and cluster binding energies can be calculated in alumina. In the Zhang-Northrup formalism, the formation energy $\Delta G^f_i$ of a defect $i$ is given by Eq. 3.29:
Figure 4.22: A schematic of $\{3\text{Ti}_1\text{Al}_3 : V_{\text{Al}}\}$ clusters formed by creating an aluminium split vacancy and substituting three aluminium atoms for titanium dopants. The symbols take their usual meanings. On the left, all atoms associated with the aluminium split vacancy have been substituted for titanium, on the right, one example permutation is shown of the four possible when placing three titanium dopants on the five (1, 2, 3, NN1, NN2) indicated aluminium sites (discounting the configuration on the left). All permutations were considered in the mass action approach.

$$\Delta G^f_i = G^t_i - G^t_0 - \sum_{\alpha=1}^{E} \Delta n^\alpha_i \mu_\alpha + q_i \mu_e,$$

where $G^t_i$ and $G^t_0$ are the Gibbs total energies of the defective and perfect computational supercells and are approximated by the $T = 0\text{K}$ total energies $E_i^{\text{tot}}$ and $E_0^{\text{tot}}$, $\Delta n^\alpha_i$ the number and $\mu_\alpha$ the chemical potential of the atoms of type $\alpha$ added (positive $\Delta n^\alpha_i$) or removed (negative $\Delta n^\alpha_i$) to form the defect, $q_i$ the defect charge state, and $\mu_e$ the electron chemical potential. We define the Fermi level $\epsilon_F$ via $\mu_e = E_i^{VBM} + \epsilon_F$, where $E_i^{VBM}$ is the energy of the VBM. The formation energy $\Delta G^f_i$ depends on $\epsilon_F$ unless $q_i = 0$. Therefore, for any particular choice of $\epsilon_F$, different charge states of a defect have different formation energies, and the charge state with the lowest formation energy will have the highest concentration, dominating at low temperature. In all plots of defect formation energies, only the lowest-energy charge state at each $\epsilon_F$ is shown, with points marking the crossovers between different charge states.

The formation energies of the native defects in alumina, including the AlO vacancy, are plotted in Fig. 4.23 against $\epsilon_F$ at typical experimental conditions of $T = 1750\text{K}$. 

146
Table 4.9: The Schottky and Frenkel energies in alumina at $T = 1750\text{K}$, $p_{O_2}/p^0 = 0.2$. Previous studies have indicated a preference for Schottky disorder; however, many shell models predict that the anion Frenkel energy is lower than the cation Frenkel energy, a finding that this study, and other DFT studies [122], does not support.

<table>
<thead>
<tr>
<th>Disorder type</th>
<th>Energy (eV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anion Frenkel</td>
<td>6.04</td>
</tr>
<tr>
<td>Cation Frenkel</td>
<td>4.08</td>
</tr>
<tr>
<td>Schottky</td>
<td>3.12</td>
</tr>
</tbody>
</table>

and $p_{O_2}/p^0 = 0.2$. These calculations use Vanderbilt USPs in the LDA, include all corrections discussed in Sec. 3.5. Although $\epsilon_F$ can theoretically take values between 0 (at the VBM) and the DFT band gap, $E^g_0$, self-consistent determinations of defect concentrations using the mass-action approach introduced in Sec. 3.3.2 show that $\epsilon_F$ is restricted to a range whose size depends on the existence and strength of any buffering effect on the defect concentrations. For $T = 1750\text{K}$ and $p_{O_2}/p^0 = 0.2$, this range is indicated by a shaded area in Fig. 4.23. In an intrinsic crystal, $\epsilon_F$ would sit near the intersection between the $V_{\text{O}}$ and $V''_{\text{Al}}$ lines to ensure charge neutrality, reflecting the fact that Schottky defects are the prominent form of intrinsic disorder, as evidenced in Table 4.9, that shows the calculated intrinsic disorder energies. The cation Frenkel energy per defect is higher, and the anion Frenkel far higher, than the Schottky energy, due to the relatively high formation energy of oxygen interstitials. These results compare well with previous DFT findings by Matsunaga et al. [122], and contrast with the results of shell model calculations, which predict a much lower anion Frenkel energy, in some cases even lower than the Schottky energy.

Upon aliovalent doping, $\epsilon_F$ can either fall (Mg-doping) or rise (Ti-doping). Fig. 4.23 shows that within the range found to be accessible for $\epsilon_F$ at $T = 1750\text{K}$, $p_{O_2}/p^0 = 0.2$, $V_{\text{O}}$ and $V''_{\text{Al}}$ defects are the most stable, depending on the level of $\epsilon_F$. $V'_{\text{AlO}}$ defects, which have only a small dependence on $p_{O_2}$ and $T$ due to cancellation between changes in $\mu_{\text{O}}$ and $\mu_{\text{Al}}$, are slightly higher in energy, and could therefore play an interesting role in the defect chemistry. Interstitial defects, in contrast, are significantly higher in energy, in part due to their strong lattice relaxations. Interestingly, all defects assume charge states corresponding to formally charged ions except for oxygen interstitials, which are neutrally charged over the allowed range.

---

8Using the DFT band gap rather than the experimental value prevents spurious high-charge states from appearing stable within DFT.
Table 4.10: Defect formation energies of the native defects in alumina at \( T = 1750K \), \( p_{O_2}/p^0 = 0.2 \), and \( \epsilon_F = 1.6eV \), using Vanderbilt USPs with a plane wave cut-off of 550eV.

<table>
<thead>
<tr>
<th>Defect</th>
<th>( \Delta G_f^T ) (eV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( V_{\dot{O}} )</td>
<td>3.13</td>
</tr>
<tr>
<td>( Al_{\dot{\cdot}} )</td>
<td>5.07</td>
</tr>
<tr>
<td>( V_{Al}^{'''n} )</td>
<td>3.10</td>
</tr>
<tr>
<td>( O_{1\times}/O_{1}/O_{1}'' )</td>
<td>5.50/7.63/8.95</td>
</tr>
</tbody>
</table>

The formation energies of the four native defects in alumina were also examined using the formalism presented in Sec. 3.4, but with a hybrid functional, B3LYP, in a Gaussian basis set. B3LYP was able to reproduce the band gap of alumina to a higher accuracy, giving a value of 8.32eV, so that no band gap corrections need to be applied. This provides a gauge for the accuracy of LDA calculations of native point defects in order to highlight the errors arising from band gap corrections. The formation energies, plotted in Fig. 4.24 for \( T = 1750K \) and \( p_{O_2}/p^0 = 0.2 \), show the same qualitative trends those obtained using using the LDA in a plane wave basis, but exhibit slight quantitative differences. The Fermi level is still set low in the gap for an intrinsic crystal, and \( V_{\dot{O}} \) and \( V_{Al}^{'''n} \) defects dominate in this regime. This is mirrored in the intrinsic disorder energetics, where the Schottky energy (2.70eV) is much smaller than the cation Frenkel (4.24eV), and anion Frenkel (5.46eV) formation energies, as also observed in the LDA calculations. Both interstitial defects are raised in energy with respect to the vacancies when compared with the LDA results. In addition, B3LYP stabilises intermediate charge states for oxygen defects over a wider range of \( \epsilon_F \), rendering \( O_{1}^{i} \) defects more stable than \( O_{1}^{\times} \) defects in intrinsic crystals, and creating a larger range of stability for \( V_{\dot{O}} \). The existence of \( F^+ \)-centres in alumina verifies that the B3LYP functional is an improvement over the LDA. It is worth noting that a large portion of the difference in formation energies between
Figure 4.23: Defect formation energies of the native defects in alumina at typical experimental annealing conditions of $T = 1750\text{K}$ and $p_{O_2}/p_0 = 0.2$, as functions of the Fermi level $\epsilon_F$, where $0 < \epsilon_F < E_g^{0}\text{(DFT)}$, calculated using Vanderbilt USPs in the LDA with a plane wave cut-off of 550eV. Selected charge states of defects are indicated along the plot, but some are omitted for clarity. The position of $\epsilon_F$ varies throughout the gap, depending on the types of dopants present and their concentrations. At these formation conditions, $\epsilon_F$ is found to vary within the shaded area. In this regime, the most stable defects are $V_O^-$ and $V_{Al}'''$.

defects calculated in the LDA and in B3LYP is due to the neglect of finite-size corrections for calculations when using the latter functional. These corrections, that were computationally intractable for the chosen Gaussian basis set, raise the formation energies of charged defects significantly, which would bring the vacancy defect formation energies roughly in line with each other. In addition, many of the B3LYP calculations involving intermediate charge states were difficult to converge, and AlO defects were entirely unfeasible due to convergence issues and the large computational resources required. In summary, the formation energies from B3LYP, to a large degree, support the LDA results.

The findings presented here are supported by previous DFT calculations by Tanaka et al. [174] and Xu et al. [190]. These found relatively high formation energies in line with those reported here, and indicated the necessity for relaxing large supercells
Figure 4.24: Defect formation energies of the native defects in alumina at $T = 1750\text{K}$ and $p_{O_2}/p^0 = 0.2$ using a hybrid functional, B3LYP, as functions of the Fermi level $\epsilon_F$, where $0 < \epsilon_F < E_g^0$ (DFT). The results are qualitatively similar to those obtained using the LDA (Fig. 4.23), but show slight differences. Vacancy formation energies are lower than in the LDA by $< 1\text{eV}$, interstitial formation energies higher, and there are larger regions of stability for the $+1$ and $-1$ charged $V_O$ and $O_i$ defects.

to deal with finite-size errors. However, these studies did not account for finite-size errors via a correction scheme, nor corrected for the DFT underestimation of the band gap, rendering direct comparison with the current results difficult. Matsunaga et al. [122] applied average potential corrections to remove defect-defect interactions and a full $m \times \Delta E_g^0$ band gap correction, both of which we have shown in the past to be unreliable [82].

Cluster binding energies

Defects can either exist as isolated, free defects, or they can aggregate together, or with impurities, to form defect clusters. The binding energy

$$\Delta G_{\text{bind}}^i = \sum_{i=1}^{N_{\text{defects}}} \Delta G_{i}^{f} - \Delta G_{\text{cluster}}^{f}$$

(4.5)
Section 4.4: Density functional theory calculations

Cluster species | $\Delta G_{i}^{\text{bind}}$ (eV) \\
--- | ---
{MgAl : V\text{O}} & 1.21 \\
{MgAl : Al\text{i}} & 1.65 \\
{T\text{iAl} : V\text{Al}} & 1.13 \\
{2Ti\text{Al} : V\text{Al}} & 2.37 \\
{3Ti\text{Al} : V\text{Al}} & 2.51 \\
{T\text{iAl} : O\text{i}} & 1.42 \\
{2Ti\text{Al} : Al\text{i}} & 1.54 \\
{3Ti\text{Al} : Al\text{i}} & 1.70 \\

Table 4.11: Cluster binding energies of the most significant clusters at $T = 1750K$, $p_{O_{2}}/p^{0} = 0.2$. 

The binding energies of the most common defect-dopant clusters in alumina are tabulated in Table 4.11. Many of the clusters, such as {MgAl : V\text{O}}$^\cdot$, {MgAl : Al\text{i}}$^-$, {T\text{iAl} : V\text{Al}}$''$, and {2Ti\text{Al} : V\text{Al}}$'$, have high binding energies and should therefore feature prominently in the defect chemistry of alumina. Other clusters, including {T\text{iAl} : O\text{i}}$^x$, have relatively low binding energies in comparison with the sum of the formation energies of their constituent defects, and are presumably very low in concentration.

The only previous determinations of cluster binding energies in alumina use empirical shell models. Catlow et al. [27] calculated the binding energies of neutral clusters and found $\Delta E_{\text{bind}}^{\text{bind}} (\{2\text{Ti}_{\text{Al}} : O\text{i}^{\cdot}\}) = 3.05\text{eV}$, $\Delta E_{\text{bind}}^{\text{bind}} (\{3\text{Ti}_{\text{Al}} : V\text{Al}^{''}\}) = 5.23\text{eV}$, $\Delta E_{\text{bind}}^{\text{bind}} (\{2\text{Mg}_{\text{Al}} : V\text{O}^{\cdot}\}) = 2.56\text{eV}$, and $\Delta E_{\text{bind}}^{\text{bind}} (\{2\text{Mg}_{\text{Al}}^{'} : V\text{Al}^{''}\}) = 2.43\text{eV}$. Similarly, Grimes [72] used a full-charge, polarisable shell model and calculated the binding energies of a raft of clusters, finding values larger than those of Catlow et al. by ca. 0.3eV.
4.4.4 Chemical potentials

In Secs. 3.4 and 3.5, the application of various corrections to out-of-the-box DFT defect formation enthalpies was introduced in order to obtain accurate defect concentrations to be compared with experimental data. We now strip out each of these corrections separately, to demonstrate their impact on defect formation energies.

Defect formation energies depend on the chemical potentials of any atoms removed or added to the perfect crystal in order to create defects, as seen in Eq. 4.4. For alumina, these atoms are aluminium and oxygen. Conventionally, the oxygen chemical potential $\mu_O$ is calculated as half of the DFT total energy of the oxygen dimer, combined with $pV$ and entropic contributions, and the aluminium chemical potential $\mu_{Al}$ found from the equilibrium condition of the solid (Eq. 3.33). Alternatively, $\mu_O$ can be calculated from highly-accurate experimental data combined with DFT total energies of bulk alumina and aluminium metal, using the method first proposed by Batyrev, Alavi, and Finnis (BAF) [58, 14], by invoking a thermodynamic cycle corresponding to formation of alumina (see Sec. 3.4.2 for details).

To illustrate the difference between these two approaches, the defect formation energy of an oxygen vacancy, $\Delta G_{V_O}^f$, is shown in Figs. 4.25(a) and (b) as calculated using both methods and a range of pseudopotentials and exchange-correlation functionals.
Other defects yield very similar results, and are therefore omitted. In Fig. 4.25(a), $\mu_O$ is calculated using the conventional method, using pseudopotentials constructed with the same $E_{xc}$ functionals used in the defect calculations (e.g. Vanderbilt ultrasoft pseudopotentials within the LDA). All other corrections to the defect formation energies are still included. In Fig. 4.25(b), $\mu_O$ is found from Eq. 3.41 using the BAF method, replacing the DFT calculation of the oxygen dimer by experimental data and DFT calculations of bulk alumina and aluminium metal. The approach in Fig. 4.25(a) produces defect formation energies that depend significantly on the choice of pseudopotential and $E_{xc}$, giving a spread of ca. 1eV between different implementations. This is caused by the difficulty DFT has in describing the rapidly varying density of the oxygen dimer and the spin-polarised oxygen atom. The BAF method, on the other hand, avoids this challenging calculation, yielding formation energies that differ by less than 0.2eV for charged, and 0.5eV for neutral, oxygen vacancies. This suggests that a substantial difference between calculated defect formation energies in DFT using various pseudopotentials and exchange-correlation functionals can be alleviated by avoiding DFT calculations of molecular oxygen.

Although the BAF approach provides a reliable link to experimental conditions, it is not entirely error-free. Three of these errors are difficult to quantify. First, there are small experimental uncertainties in the value of $\Delta G^f_{Al_2O_3}(p_{O_2}^0, T^0)$ entering $\mu_O$ in Eq. 3.41. Second, the DFT calculations of the bulk oxide and metal contain errors intrinsic to the LDA or GGA. These can perhaps be partially ameliorated by employing a hybrid functional, which has been seen to correct certain errors, such as the overbinding of solids [13]. Third, the lack of transferability of the metal pseudopotential between the metal and metal oxide may affect the DFT calculations. Fortunately, as evidenced by recent quantum Monte Carlo (QMC) studies on native defect formation energies in alumina [81], which show reasonable agreement between DFT and QMC, these errors appear to be very small. The pseudopotential errors can be rigorously benchmarked by performing all-electron calculations of both the metal and metal oxide and comparing these to the pseudopotential results. The B3LYP calculations in this study use an all-electron Gaussian basis set, and the agreement between the formation energies obtained using the full-core B3LYP method and the pseudopotential LDA method is encouraging. Finally, although vibrational contributions at room temperature are included in calculations of the Gibbs free energies of the metal and metal oxide needed to compute $\mu_O$, anhar-
Table 4.12: Effective factors $m_{\text{eff}}$ describing the percentage of conduction band nature in defect states containing electrons for the native defects and dopants in alumina. When a charge state is not shown, it is implied that $m_{\text{eff}} = 0$. $m_{\text{eff}}$ is calculated from Eq. 3.48 and then used in the band gap correction given in Eq. 4.6 to shift the total energies of defect supercells. For example, for an oxygen vacancy, the applied corrections are: $2 \times 0.37 \times \Delta E_g^0$ for $V_{O}^\times$, $1 \times 0.37 \times \Delta E_g^0$ for $V_{O}$ and $0 \times 0.37 \times \Delta E_g^0$ for $V_{O}$.

4.4.5 Band gap corrections

In Sec. 3.5.2, several schemes for dealing with the DFT band-gap error were outlined. In DFT, energy gaps to conduction band states are underestimated, leading to a smaller band gap than found experimentally. A recent study infers the experimental band gap in alumina from optical absorption data, measuring $E_g^0 (\text{exp.}) = 9.57 \text{eV}$ [64]. The DFT gap found using Vanderbilt USPs in the LDA is $E_g^0 (\text{DFT}) = 6.88 \text{eV}$, which is $\Delta E_g^0 = 2.69 \text{eV}$ smaller than the experimental gap. This result compares to previous DFT results of 5.82eV [122] and 6.31eV [190]. We observe differences in the DFT band gap of $\pm 0.3 \text{eV}$ between different pseudopotentials and functionals, except for B3LYP. Using B3LYP with Gaussians in the LDA produces a more accurate band gap of 8.32eV, as expected from a hybrid functional.

The work presented here applies a modification, discussed in Sec. 3.5.2, of the approach proposed by Baraff and Schlüter [8] to correct for the band gap error. The total energy of a defect containing $m$ electrons in a localised state thought to derive from the conduction band (e.g. for $\text{Al}_{i}^{m}$, $m = 0$, for $\text{Al}_{i}^{+}$, $m = 1$, for $\text{Al}_{i}^{+}$, $m = 2$, and for $\text{Al}_{i}^{\times}$, $m = 3$) is rigidly shifted upwards by a term
Section 4.4: Density functional theory calculations

Figure 4.26: The defect formation energy of an oxygen vacancy, $\Delta G_{V_O}^{f}(T = 1750K, p_{O_2}/p^0 = 0.2)$, when applying different band gap corrections. The spread between applying no band gap correction (i.e $m_{eff} = 0$ in Eq. 4.6) and a full correction ($m_{eff} = 1$) is 5.38eV for the highest charge state, $V_O^{\times}$, when $\epsilon_F$ is at the top of the DFT band gap. The more physically motivated overlap method, that calculates the amount of conduction band nature in defect states to generate $m_{eff} = 0.37$ using Eq. 3.48, yields a value in between these two extremes.

$$ m \times m_{eff} \times \Delta E_{0g}^g, \quad (4.6) $$

where $m_{eff}$ is an effective scaling factor that describes the amount of conduction band nature in the defect states. The inclusion of this effective factor prevents defect states that are not very conduction band-like (for example electrons added to an O_i interstitial, that derive entirely from oxygen 2p orbitals) from being shifted by the large values of $\Delta E_{0g}^g$ encountered in wide band gap insulators. $m_{eff}$ is calculated using Eq.3.48 from the overlap between the wave functions of perfect and defect supercells, necessarily using norm-conserving (TNDFP) pseudopotentials. The values found for native defects and dopants in alumina are listed in Table 4.12, while those for defect-dopant clusters are omitted, as they can easily be inferred from their constituent defects.

Fig. 4.26 demonstrates the effect of applying a band gap correction to $V_O$ defects for different values of $m_{eff}$. The most common approaches are to apply no band gap correction ($m_{eff} = 0$) or a ‘full’ correction ($m_{eff} = 1$) that rigidly shifts up all electrons in defect states by $\Delta E_{0g}^g$. In alumina, $\Delta E_{0g}^g$ is very large and varying the value of $m_{eff}$ between 0 and 1 creates an enormous spread of over $5.38eV = 2 \times \Delta E_{0g}^g$ for
the formation energy of a neutral oxygen vacancy, $V_O^\times$. Calculating the conduction band nature in the defect states using the approach described in Sec. 3.5.2 yields $m_{\text{eff}} = 0.37$ for $V_O$ and $V_O^\times$ defects, and thereby gives formation energies that straddle the two extremes, but are more physically meaningful. From Fig. 4.7, it is evident that none of the visible weight of the defect state is distributed on neighbouring aluminium sites, suggesting only minimal conduction band-like character. In addition, this value of $m_{\text{eff}}$ closely resembles the position of the localised defect state for $V_O$ (relative to the VBM) divided by the DFT band gap, $(\varepsilon_{\text{def}} - E_{\text{VBM}}^0)/E_g^0$, suggesting that this may provide a first approximation for $m_{\text{eff}}$. In contrast, when Matsunaga et al. [122] apply a full band gap shift, they raise the Al$_i^\cdot$, Al$_i^\bullet$, and Al$_i^\times$ defects so high in energy that they only appear as the lowest-energy defects $> 2\text{eV}$ above the top of the DFT band gap. This unphysical result is rectified using the overlap band gap correction advocated here.

At high temperatures, additional changes in the band gap arise, but are assumed to be negligible. Due to thermal expansivity, the lattice constant in alumina at 1750K is 1-2% larger than at 0K, producing a slight change in the band gap. These effects are small and presumed to cancel between supercells entering defect formation energies. Although the band gap error is by far the most severe and uncontrolled error in the present work, it only affects defects in high charge states that contain conduction band-like defect states, such as Al$_i^\times$, which are shown in Sec. 4.5 to have relatively insignificant concentrations in real crystals, as the Fermi level is confined to a specific range in the band gap. Thus, although the errors are substantial, they are not expected to influence any important results.

4.4.6 Finite-size scaling

Many of the defects and clusters presented in Sec. 4.4.2 carry a strongly localised charge that gives rise to significant defect-defect interactions and renders finite-size scaling challenging. In addition, the anistropic dielectric tensor and elasticity in alumina hamper analytic corrections of these finite-size effects.

The present work uses the Madelung interpolation scheme presented in Sec. 3.5.3 to extrapolate defect formation energies to the infinite-cell limit from calculations of formation energies in supercells of different shapes and sizes by performing a linear fit.
Section 4.4: Density functional theory calculations

\[ E(v_M) = E_\infty + bv_M, \quad (4.7) \]

where the dependent variable, the Madelung potential \( v_M \), is a function of the cell geometry. The gradient \( b = q^2/2\epsilon \) can then be used to find an effective dielectric constant \( \epsilon_{\text{eff}} \), and the extrapolated formation energy in the infinite-cell limit is found by adding the correction

\[ \Delta E = -\frac{q^2v_M}{2\epsilon_{\text{eff}}} \quad (4.8) \]

to the formation energy corresponding to one of the supercells. In practice, the \( 2 \times 2 \times 1, 2 \times 2 \times 2, 2 \times 2 \times 3, 3 \times 3 \times 1, \) and \( 4 \times 4 \times 1 \) supercells are used with USPs at a plane wave cut-off of \( E_{\text{cut}} = 400\text{eV} \) in the LDA, and the correction added to the formation energy of the \( 2 \times 2 \times 1 \) cell. In order to accelerate geometry optimisations in larger cells, the ground-state geometry of smaller cells are embedded into the larger structures as an improved starting point for the optimiser. Suitable \( \mathbf{k} \)-point grid densities were applied to the larger supercells to keep the grid density constant between supercells, although the dependence of formation energies on \( \mathbf{k} \)-point sampling was minimal.

Fig. 4.27 shows the corrected and uncorrected formation energies of the native defects, dopants, and important defect-dopant clusters in alumina as a function of \( v_M \). All results are determined for charge states \( q_i \) corresponding to formally-charged ions, as these tend to be the most important charge states and suffer the largest finite-size errors. In contrast, neutral defects such as \( \text{O}_i^{\times} \) contain no first-order Madelung/Makov-Payne correction term. As \( q_i \) for a given defect \( i \) is increased, the Madelung interpolation (or, similarly, the Makov-Payne correction) becomes more reliable, since the charge density surrounding the defect becomes increasingly bulk-like. For example, removing an aluminium atom from the lattice to create the neutral \( V^{\text{Ne}_i}_{\text{Al}} \) defect depletes the defect site of the three electrons previously donated to oxygen neighbours. This causes significant changes in the surrounding charge density. Adding these electrons to create a \( V^{\text{m10}}_{\text{Al}} \) defect returns the density to a bulk-similar state. As the charge at the defect site is highly localised, the Makov-Payne correction and its derivatives work very well. In contrast, those situations where defects change the surrounding charge density significantly and add polarisation effects are more difficult to cope with. This agrees with previous research into the screening of defect-localised charges in diamond [165].
Figure 4.27: Finite-size scaling graphs for the native defects, dopants, and important defect-dopant clusters in alumina, all in charge states corresponding to formally-charged ions, calculated using Vanderbilt USPs in the LDA with a plane wave cut-off of 400eV. The Madelung correction method works very well for native defects containing strongly localised charges, leaving an error of $<0.2$eV. Larger errors of up to 0.4eV arise in clusters due to the strong lattice relaxations surrounding the defects.
Section 4.4: Density functional theory calculations

The corrected energies for the native defects and substitutional dopants, that contain strongly localised charges on the defect sites, have uncertainties of less than 0.2 eV. The small residual errors comprise dipole-dipole and monopole-quadrupole terms, which cannot be expressed in terms of $v_M$, but decay more rapidly with $L$. In contrast, the results for clusters, including $V'_{AlO}$, are less impressive, but still adequate. Some clusters, such as $\{Mg_{Al} : V_O\}$ and $\{2Ti_{Al} : V_{Al}\}'$ perform very well, although their uncorrected formation energies already show little variance with supercell size. Other clusters, such as $V'_{AlO}$, $\{Mg_{Al} : Al_i\}$, and $\{Ti_{Al} : V_{Al}\}''$ have uncertainties of up to 0.4 eV. This error is measured as the largest difference between the extrapolated infinite-cell energies from different sized supercells. The increased uncertainty is due to the significant lattice relaxations surrounding the defects, in particular from strains along the c-axis. Small supercells, such as $2 \times 2 \times 1$, $3 \times 3 \times 1$, and $4 \times 4 \times 1$, are simply inadequate when dealing with clusters that perturb the lattice significantly. As we will see in Sec. 4.5, $\{Ti_{Al} : V_{Al}\}''$ defects are present in significant concentrations in Ti-doped alumina. This renders the large uncertainties in the finite-size corrections for this cluster one of the principal errors of the present work, and we will later explore the effect that this error could have on the defect chemistry of doped alumina. Although other clusters, such as $\{Ti_{Al} : O_i\}'$ and $\{Mg_{Al} : Al_i\}'$ are also difficult to encompass in our finite-size correction method, we will show that these are present in lower concentrations, so that the errors in these cases are less critical.

The errors in the finite-size extrapolations performed in this study are still acceptable for our purposes. For example, changing the formation energy of a single defect by 0.4 eV, the largest error found for any defect, while keeping all other defect formation energies constant can change the concentrations of defects by up to one order of magnitude. Therefore, the finite-size corrections used in this study generate defect formation energies accurate enough to resolve defect concentrations to an accuracy of roughly one order of magnitude.

The values of the effective ‘permittivity’ $\epsilon_{eff}$ found from linear fits of the uncorrected formation energies range from 9.7 to 14.4, comparable in order of magnitude to the components of the dielectric tensor in alumina. This tensor is anisotropic, which, in part, explains the variation found in $\epsilon_{eff}$. However, the large differences primarily arise from the band gap error that yields incorrect dielectric properties, and the extremely high concentration of defects due to periodic replicas of defect supercells, whose dielectric properties may differ from the bulk.
4.4.7 Vibrational contributions

Vibrational effects enter defect formation energies in two ways. First, Zhang-Northrup defect formation energies, expressed as

\[ \Delta G_i^f = G_i^{\text{tot}} - G_0^{\text{tot}} - \sum_{\alpha=1}^{E} \Delta n_{i\alpha} \mu_{\alpha} + q_i \mu_e \]  

(4.9)

contain the difference between the Gibbs free energy of a defect-containing supercell, \( G_i^{\text{tot}} \), and a perfect supercell, \( G_0^{\text{tot}} \), of alumina. Each of these comprises a ground-state \( T = 0 \text{K} \) total energy \( E^{\text{tot}} \) (omitting zero-point effects) and a vibrational free energy \( F^{\text{vib}} \), as described in Sec. 3.5.1. Although usually ignored, as in this study, there is no reason to assume that the latter will not vary among supercells containing different defects. Second, as described in Sec. 3.4.2, using the BAF method to determine the oxygen chemical potential \( \mu_O \) and neglecting the vibrational free energy \( F^{\text{vib}}_0 \) for bulk Al\(_2\)O\(_3\) causes an error to accumulate in the metal chemical potential, \( \mu_{\text{Al}} \). This error can be corrected by adding half the change in the vibrational free energy of bulk Al\(_2\)O\(_3\) from \( T = 0 \text{K} \) to a given \( T \), \( \Delta G_{\text{Al}_2\text{O}_3} \), to the commonly-used aluminium chemical potential omitting vibrational effects, \( \bar{\mu}_{\text{Al}} \):

\[ \mu_{\text{Al}}(p_{\text{O}_2}, T) = \bar{\mu}_{\text{Al}}(p_{\text{O}_2}, T) + \frac{1}{2} \Delta G_{\text{Al}_2\text{O}_3}(T). \]  

(4.10)

In this study, the vibrational free energy of bulk alumina is calculated in the harmonic approximation using density functional perturbation theory (DFPT) in order to correct for the second of these problems. For this, norm-conserving (TNDFP) pseudopotentials are used.

When performing DFPT calculations, it is imperative that the geometry is very well-converged. The most important factors for this are accurate forces and \( \mathbf{k} \)-point sampling. For the calculations presented here, forces are converged to \( 5 \times 10^{-4} \text{eV/Å} \), and a \( 4 \times 4 \times 4 \) Monkhorst-Pack grid is used. Higher force tolerances could not converge with the BFGS optimiser. Using lower grid densities was found to produce a slew of phonon modes with small imaginary frequencies. At the tolerances used in the current study, the difference in the change in Gibbs free energy from \( T = 0 \text{K} \) to 1750K from that found from the experimental heat capacity of perfect alumina is \( 0.15\text{eV} \). The dynamical matrix was determined in DFPT on a \( 4 \times 4 \times 4 \) Monkhorst-
Figure 4.28: The difference in Gibbs free energies from 0K to T of the uncorrected and vibrationally-corrected aluminium chemical potentials $\Delta \mu_{Al} = \mu_{Al}(T) - \mu_{Al}(T = 0K)$, as a function of temperature T. $\mu_{Al}$ and $\mu_{O}$ are determined using the BAF method. Neglecting the vibrational contributions to bulk alumina, $F^{vib}_0$, accumulates an error of $O[1-2eV]$ into $\mu_{Al}$ at experimental temperatures of 1500-2000K.

Pack grid shifted to contain the $\Gamma$ point, and the vibrational free energy determined as

$$F^{vib}(T) = k_B T \sum_{\mathbf{k}} w_{\mathbf{k}} \sum_{i=1}^{3N} \ln \left\{ 2 \sinh \left( \frac{\hbar \omega_i(\mathbf{k})}{2k_B T} \right) \right\}.$$ (4.11)

In Fig. 4.28, the effect of correcting $\mu_{Al}$ when using the BAF method by accounting for the vibrational free energy of bulk alumina is shown. Even at low temperatures, this correction is crucial, and at $T = 2000K$, it grows to over 2eV. Neglecting this contribution leads to a very accurate determination of $\mu_{O}$ and a very poor determination of $\mu_{Al}$ for given $T$ and $p_{O_2}$. As $\mu_{Al}$ enters the formation energies of all aluminium-containing defects (e.g. $V_{Al}$, $\{Mg_{Al} : Al_i\}$, etc.), these will then either be over- or under-bound by $O[1eV]$ at experimental conditions.

If this weren’t enough, the neglect of anharmonicity at high temperatures can exacerbate these errors. The harmonic approximation does not account for the volume and temperature dependence of vibrational frequencies. The former can be quantified within DFPT by determining the phonon spectra at different cell volumes. Although this quasiharmonic approximation (QHA) exceeds current computational capabilities, previous studies on MgO and MgSiO$_3$ [179, 91] show that the volume changes at high temperatures are fairly small (1-2%) and lead to changes of $< 10\%$ in the vibrational contributions to the free energy at high temperatures. Recent $ab$
initio studies on aluminium show that, cumulatively, anharmonicity accounts for changes $< 1\%$ in the vibrational free energy and $< 3\%$ in the heat capacity at temperatures near the melting point [70]. Although these changes are not insignificant, they do not make qualitative changes in the current results. Current computational limitations prevent accurate determinations of anharmonic effects, but future work should ascertain the errors due to their neglect in alumina.

4.4.8 Diffusion barriers

In order to get from theoretical defect concentrations to theoretical diffusion coefficients that can be compared with experiment, we need to know the diffusion barriers of isolated defects that contribute to diffusion. The complex ground-state structures of some of the native point defects in alumina, as demonstrated in this section, lead to exotic diffusion mechanisms, making alumina an interesting case study for diffusion.

Previous studies of diffusion barriers in alumina have been very limited, primarily due to computational resources. Dienes et al. [44] performed a systematic study using a polarisable ion model. For aluminium vacancies, they found the diffusion barrier to a nearest-neighbour along the c-axis to be 3.8eV, and, surprisingly, that the split-vacancy configuration is 6.6eV higher in energy than the normal vacancy. They also found a barrier of 3.8eV for the aluminium vacancy to move in the basal plane. For the oxygen vacancy, interestingly, they found a barrier of 5.2eV to a nearest-neighbour in the basal plane, and a lower barrier of 2.9eV to a nearest-neighbour outside the basal plane. For simple aluminium interstitials, they found a barrier of 4.8eV. Unfortunately, Dienes et al. were computationally restricted to guessing the atomic positions for these configurations, rather than testing a number of different positions, but the high value of the split vacancy configuration, as also discussed in Sec. 4.4.2, already suggests that shell models may be inadequate for the calculation of migration barriers in alumina. In the only previous ab initio study of diffusion barriers in alumina, Carrasco et al. [25] find an oxygen vacancy migration barrier of 3.7eV, which, as mentioned in Sec. 4.3, seems very high. Unfortunately, there are too few details of the method used to find this barrier in this paper for us to draw any more conclusions.

The diffusion barriers calculated in this section are found using the moving-plane method described in Sec. 3.6.2. All calculations, unless specifically stated, are per-
formed in a $2 \times 2 \times 1$ supercell, except for the Madelung interpolations used for finite-size scaling of the migration barriers, which are performed in the $2 \times 2 \times 1$, $3 \times 3 \times 2$, $3 \times 2 \times 2$, $2 \times 2 \times 2$, $3 \times 3 \times 1$, $4 \times 4 \times 1$ and $2 \times 2 \times 3$ supercells.

**Oxygen vacancy**

Oxygen vacancies in alumina diffuse by a simple mechanism. In perfect alumina, every oxygen atom has twelve nearest-neighbour oxygen atoms, of which there are just four that are inequivalent by symmetry. This means that there are four different paths along which an oxygen vacancy can migrate to a nearest-neighbour oxygen lattice site. In addition, due to the open space in the structure, it seems plausible that an oxygen vacancy could migrate along the $c$-axis to a nearest-neighbour site, albeit 4.26Å away. Equivalently, these processes can be viewed from the perspective of the migrating oxygen atom.

Each of these five possible migration paths are examined and the corresponding migration barriers plotted in Figs. 4.29(a) and (b). The barrier along the $c$-axis is significantly higher than to the first set of oxygen nearest-neighbours, at 10.84eV. For the remaining four, the highest barrier (2.78eV) corresponds to the nearest-neighbour furthest, the lowest barrier (1.01eV) to the neighbour nearest, to the vacancy. The path corresponding to the lowest barrier does not allow for migration throughout the crystal, but merely along the sides of the small triangles connecting three oxygen atoms, as can be seen in Figs.4.30(a) and (b). The second-lowest barrier (1.79eV) is surmounted along a path to the third nearest-neighbour to the defect, a distance of 2.68Å from the vacancy in the perfect crystal, and is the lowest barrier allowing for full diffusion throughout the crystal. Finally, the third-lowest barrier, only slightly higher than the preceding one (1.88eV), corresponds to migration to the second nearest-neighbour oxygen atom.

Migration along the lowest-barrier path allowing for 3D diffusion throughout the crystal, Path 2 in Figs. 4.29(a) and (b), is schematically depicted in Figs. 4.30(a) and (b). Surmounting the associated barrier allows for movement between stacked oxygen layers, as well as in the plane of the layer.

The migration barrier corresponding to the lowest-energy path of the $V_O$ defect was examined using a classical potential from Gale *et al.* [65] to gauge the potential’s performance for such defect calculations. The migration barrier along Path 2 in Figs. 4.29(a) and (b) was found to be 0.77eV with Gale *et al.*’s potential, ca. 1eV.
Figure 4.29: The migration barriers along paths to the four distinct nearest-neighbour oxygen atoms to a vacancy (Paths 1-4), as well as along the c-axis (Path 5) (the latter is omitted in (b)), measured against the fractional distance along each migration path. The distances of each path in the legend in (a) relate to the positions of the neighbours to the defect site in the perfect crystal, as the distances are different for the defect crystals, depending on the charge state of the defect. Differences between the migration barriers of defects in different charge states were examined by comparing the maxima along each path, and found to be < 0.05 eV. The lowest barrier (1.01 eV), along Path 3, does not allow for 3D diffusion through the crystal, so the barrier along Path 2 (1.79 eV), is used for comparison to experiment.

lower than the value found in DFT.

Aluminium split vacancy

The migration of an aluminium split vacancy is a complicated process, either occurring by a crowdion mechanism along the c-axis, or a cooperative movement of atoms in the basal plane. The split vacancy has a three-fold axis of symmetry, with three inequivalent basal aluminium nearest-neighbours and one along the c-axis, as shown in Fig. 4.31(a). Any of these four neighbours can migrate into one (or two) of the vacancies comprising the split vacancy defect. Other neighbouring aluminium atoms were either considered to be too far from the defect site, or likely to cause significant lattice distortions if migrated into the split vacancy. These possibilities were discounted and constitute a possible point for further research.
Section 4.4: Density functional theory calculations

Figure 4.30: A schematic of the lowest-energy migration step of a single oxygen vacancy in alumina. On the left (a), a vacancy is created, symbolised by the empty oxygen atom encircling a ×, and the movement of the nearest-neighbour oxygen atom with the lowest migration barrier is indicated by an arrow. On the right (b), the migrating oxygen atom is shown to move into the vacant site, while the vacancy moves in the opposite direction.

In this study, five migration paths were considered for the aluminum split vacancy. Shown in Fig. 4.31(a), the four possibilities in the basal plane are: \( \text{Al}_5 \rightarrow V_2 \), \( \text{Al}_4 \rightarrow V_1 \), \( \text{Al}_3 \rightarrow V_1 \), and \( \text{Al}_4 \rightarrow V_2 \). Alternatively, the split vacancy can migrate along the c-axis, where both \( \text{Al}_1 \rightarrow V_1 \) and \( \text{Al}_2 \rightarrow V_2 \) simultaneously, similar to a crowdion mechanism (see Sec. 3.6.1 for a description). For the basal migration paths, the split vacancy jumps from one vector parallel to the c-axis to another, so that the perfect lattice structure is restored on the vector with the original split vacancy. The migration barriers corresponding to these paths are shown in Fig. 4.31(b). Diffusion along the c-axis, with a barrier of 1.00eV, is much easier than into the basal plane. However, as most experiments measure 3D diffusion, the vacancy must diffuse into the basal plane, for which the limiting migration barrier, 2.24eV, corresponds to the shortest path, \( \text{Al}_5 \rightarrow V_2 \).

This basal migration path \( \text{Al}_5 \rightarrow V_2 \) is shown schematically in Figs. 4.32(a) and (b). The basal aluminium nearest-neighbour to one of the vacancies comprising the split vacancy migrates towards the vacancy, while the aluminium atom residing in the interstitial site relaxes towards the second vacancy in the split vacancy. Simultaneously, the next nearest-neighbour aluminium atom along the c-axis to the migrating atom relaxes into the vacant octahedral site, creating an aluminium split vacancy.
Figure 4.31: One of the three equivalent sets of nearest-neighbours to an aluminium split vacancy (a) and the corresponding energy barriers for the considered migration paths (b). The split vacancy has three distinct basal nearest-neighbours (Al$_3$, Al$_4$, Al$_5$) and one along the c-axis (Al$_2$). If one of the first three migrates into one of the vacant sites in the split vacancy, it restores the perfect crystal structure, while leaving behind it a split vacancy. In the final path the split vacancy moves along the c-axis, such that Al$_1$ $\rightarrow$ V$_1$ and Al$_2$ $\rightarrow$ V$_2$ migrate simultaneously. This process has a low energy barrier of 1.00eV, while the lowest basal migration path, Al$_5$ $\rightarrow$ V$_2$, has a barrier of 2.24eV.

along a neighbouring parallel vector along the c-axis.

Locating the saddle point for this migration is more complex than for a simple vacancy, as the lowest energy path involves a migration of three aluminium atoms (Al$_5$, Al$_1$, and Al$_6$ in Fig. 4.31(a)) that turns out to be neither simultaneous nor sequential, but cooperative. If the migration were simultaneous, all three atoms would move together from their initial to their final positions. If the migration were sequential, some of the aluminium atoms would fully migrate to their final positions before the others began leaving their initial positions. We find that a mixture of these two possibilities occurs, that we call cooperative motion.

When calculating the diffusion barrier in Fig. 4.31(b), only the migrating atom in the basal plane (i.e. Al$_5$) was constrained to relax in a plane perpendicular to its jump vector. All other atoms were allowed to relax freely, including those taking
Figure 4.32: A schematic of aluminium vacancy diffusion in alumina. A basal nearest-neighbour aluminium atom to one of the vacancies comprising the split vacancy (Al$_5$ in Fig. 4.31(a)) migrates towards this vacancy (V$_2$), causing the aluminium atom in the interstitial site (Al$_i^1$) to relax into the other vacancy (V$_1$), and a split vacancy to form where the migrating atom initially resided (Al$_6$ relaxes into the empty octahedral interstitial site). Alternatively, the migration can be viewed as a simple vacancy mechanism involving an *extended* split vacancy defect, as indicated by the dotted rectangles.

part in the cooperative migration. This ‘lack’ of constraints during the Al$_5$ → V$_2$ migration leads to the atom initially sitting on the interstitial site (Al$_i^1$) relaxing into the neighbouring vacancy before the migrating atom (Al$_5$) has performed a quarter of its jump. Similarly, by this point, the second migrating atom (Al$_6$) has already relaxed into its neighbouring empty interstitial site to create a new split vacancy. Thus, the two relaxing atoms move before the migrating atom has completed a quarter of its jump into the basal plane. This is referred to as ‘cooperative’ motion between the three aluminium atoms.

To demonstrate that this is the most likely scenario, the energy maximum at the half-way point along the Al$_5$ → V$_2$ migration was analysed in detail. Al$_5$ was constrained to move in a plane perpendicular to its jump vector at this half-way point, while the relaxing atoms (Al$_i^1$ and Al$_6$) were similarly constrained at intervals between the beginning and end points of their relaxation paths. Plotting the total energies of these configurations with respect to the ground-state energy of the split vacancy defect produces an energy landscape, shown in Fig. 4.33(a). From this, we
can infer that the relaxing atoms prefer to either stay in their initial positions, or, even more so, fully relax into their final positions, as occurred during the ‘unconstrained’ migration. This landscape gives great credence to the proposed migration mechanism by severely limiting the alternatives possible. The maximum of the landscape corresponds to the maximum of the migration barrier when all three atoms are moved simultaneously, as plotted in Fig. 4.33(b).

Given the complexity of the corundum structure and aluminium vacancy migration path found in this study, a more complicated diffusion mechanism cannot be excluded as a possibility. A proper analysis of this would require another method, as numerous DFT calculations are too taxing. A good candidate for this would involve a thorough sampling of the energy space using metadynamics, potentially in conjunction with a novel classical potential fitted to DFT data containing the correct ground-state defect structures.

The classical potential from Gale et al. [65] was also used to examine the $V_{Al}$ migration barrier, as it was the only potential found in Sec. 4.4.2 that could reproduce the split vacancy as the ground-state aluminium vacancy structure. For the $Al_5 \rightarrow V_2$ migration, it yielded a comparable migration barrier of 2.80eV in a $2 \times 2 \times 1$ supercell, and for the crowdion mechanism along the c-axis 0.75eV in a $3 \times 2 \times 2$ supercell. Starting from the split vacancy configuration in a $4 \times 4 \times 2$ supercell, a molecular dynamics (MD) simulation in the NVT ensemble at a temperature of 2000K (just below the melting point) over 100ps was run using this potential. The split vacancy wandered up the c-axis and performed one basal migration along the $Al_5 \rightarrow V_2$ path during this time frame. The observed trajectory for the basal migration was a cooperative movement along approximately the same path as that deduced from the DFT calculations. Using the same supercell over a longer run of 500ps showed the split vacancy perform two hops into the basal plane, indicating that the potential from Gale et al. has the ability to describe diffusion in alumina adequately.

Oxygen dumbbell interstitial

Oxygen dumbbell interstitials diffuse by a dumbbell interstitialcy mechanism, as described in Sec. 3.6.1, to a nearest-neighbour oxygen lattice atom. In a perfect crystal, oxygen lattice atoms have twelve oxygen nearest-neighbours, of which four are inequivalent. Creating an oxygen dumbbell interstitial, centred on an oxygen lattice site, breaks this symmetry, so that each constituent of the dumbbell has
Figure 4.33: An energy landscape depicting the effect of different configurations of the two relaxing atoms partaking in the split vacancy migration mechanism (a), and the migration barrier when all three atoms move simultaneously (b). For the energy landscape, the migrating atom, Al\textsubscript{5}, is held at the half-way point along its migration path and allowed to relax in a plane perpendicular to the jump vector, while the two relaxing atoms, Al\textsubscript{i} and Al\textsubscript{6}, are placed at fractional distances along their migration paths and similarly relaxed. The landscape shows that the relaxing atoms prefer to fully move into their final positions, supporting evidence from an unconstrained Al\textsubscript{5} \rightarrow V\textsubscript{2} migration. The most unfavourable position is for all three atoms to lie at their half-way points simultaneously, as would occur at the peak of the migration barrier when they move simultaneously (b).
twelve inequivalent nearest-neighbour oxygen atoms. The migration barrier for the dumbbell can be found by migrating one of its oxygen atoms towards each of its nearest-neighbour oxygen atoms, so that it forms a new dumbbell with these, while the other atom from the original dumbbell relaxes back into an oxygen lattice site. The separation of the either oxygen atom in the dumbbell to its neighbours ranges from 2.15 to 4.07 Å (for the -2 charged dumbbell). After fully examining the six shortest migration paths, the remaining six were probed halfway along the migration path. These samples revealed significantly higher migration barriers, so that these paths could be safely ignored. In addition, a seventh path was considered for migration along the c-axis, due to the open space available in the structure. The migration barriers along the six shortest paths (all under 3.00 Å), as well as along the c-axis (3.30 Å), are shown in Fig. 4.34 for a $2 \times 2 \times 1$ computational supercell. The two paths with the significantly lowest barriers (both 0.54 eV) belong to the two closest neighbours to the migrating oxygen atom, at distances of 2.15 and 2.24 Å. Although the barrier corresponding to the latter path (Path 5 in Fig. 4.34) is slightly lower, they are identical within the accuracy of the calculations. Examining the maxima of these two paths in larger supercells, such as $4 \times 4 \times 1$ and $2 \times 2 \times 3$, reveals that the latter path does indeed have a lower migration barrier, by ca. 0.1 eV. Together with the relatively low migration barrier for oxygen vacancy diffusion, this contrasts with the traditionally-held view of large, immobile oxygen ions.

The migration path corresponding to the lowest barrier (and all other possible paths for that matter) is difficult to depict for the complex corundum structure, and therefore cannot be displayed in a simple schematic as for the other native defects. Instead, it is projected onto the ab-, ac-, and bc-planes in Figs. 4.35(a)-(c). One oxygen atom from the dumbbell interstitial (designated in grey) migrates towards the corresponding nearest-neighbour oxygen atom, displacing this from its original lattice site and forming the new dumbbell interstitial (green), centred on a now vacant lattice site. Simultaneously, the other oxygen atom from the original dumbbell relaxes back into the originally vacant oxygen lattice site upon which the dumbbell was centred. This mechanism is also depicted for a simpler structure in Sec. 3.6.1.

**Aluminium interstitial**

As simple interstitial defects, the migration of aluminium interstitials is generally considered to occur by a simple interstitial mechanism. We show that this is not
Figure 4.34: The migration barriers along paths to the six closest nearest-neighbour oxygen atoms to an oxygen dumbbell interstitial (Paths 1-6), as well as along the \( c \)-axis (Path 7), calculated using DFT in a \( 2 \times 2 \times 1 \) supercell. The two lowest barriers correspond to the two shortest paths (3 and 5), and are identical to within errors (0.54 eV). Calculations of the maxima along these paths in the much larger \( 4 \times 4 \times 1 \) and \( 2 \times 2 \times 3 \) supercells show that the barrier along Path 5 is ca. 0.1 eV lower than that for Path 3.

In the case, as a similar interstitialcy mechanism has a significantly lower migration barrier.

From the schematic in Fig.4.36(a), we see that, for the aluminium interstitial to migrate to a nearest-neighbour empty octahedral interstitial site, it must pass through a layer of oxygen atoms, along a path of length 3.48 Å. The migration barrier for this path is plotted in Fig. 4.37, leading to a high barrier of 4.35 eV in a \( 2 \times 2 \times 1 \) supercell. However, there is a much shorter path available, that also allows for diffusion throughout the crystal. Either of the two aluminium atoms neighbouring the interstitial along the \( c \)-axis can migrate straight into an interstitial site, only 2.75 Å away, without having to pass through the oxygen layer. When one of these atoms is placed on points along this jump vector to calculate the migration barrier, the interstitial atom simultaneously relaxes back into the lattice site that the migrating atom has just vacated, creating a concerted motion between the two aluminium atoms known as an interstitialcy mechanism. This mechanism is described in Sec. 3.6.1 and shown
Figure 4.35: The lowest-energy migration path for an oxygen dumbbell interstitial defect, projected onto the ab- (a), ac- (b), and bc-planes (c) from DFT calculations in a $2 \times 2 \times 1$ supercell. Three oxygen atoms take part in each migration, the two forming the original dumbbell and a third that forms a new dumbbell with one of the first two. In this case, the atoms of the original dumbbell are shown in grey, and those of the final dumbbell in green. In (c), the oxygen atom on the right side of the original dumbbell relaxes back into an oxygen lattice site, becoming a red oxygen lattice atom. Its relaxation steps are shown in pink. The other half of the dumbbell migrates towards a nearest-neighbour oxygen atom 2.24Å away, forming half of the new dumbbell (green). The steps along its path are coloured in light green. Finally, the lattice oxygen that this interstitial oxygen atom approaches, on the far left in (c), relaxes to form the other half of the new dumbbell (green). Its relaxation steps are shown in pink.

schematically for alumina in Fig. 4.36(b). The migration barrier along this path, that includes the migration of one of the neighbours of the interstitial to a new interstitial site, as well as the relaxation of the original interstitial into a lattice site, is only 3.13eV. The entire process can alternatively be viewed as the migration of an extended interstitial complex comprising the interstitial atom and its two neighbours along the c-axis, indicated by a dotted box in Fig. 4.36(b).

Gale et al.’s classical potential [65] was also used to analyse the Al$_i$ migration barrier, but gave mixed results. Although a barrier of 3.37eV coincided decently with the DFT value for simple interstitial diffusion, the barrier for interstitialcy diffusion was significantly higher, at 5.01eV. Given the migration path, this is odd, but perhaps points to the difficulties that classical potentials have in describing aluminium interstitials, which, in the past, has led to very high aluminium interstitial defect
Figure 4.36: A schematic of aluminium interstitial diffusion in alumina. On the left (a), the movement of the migrating atom, either of two nearest-neighbours flanking the aluminium interstitial along the c-axis, is indicated by an arrow. On the right (b), this atom migrates into a neighbouring interstitial site, while the original interstitial simultaneously relaxes (indicated with an arrow containing the subscript ‘r’) into the now vacant lattice position. The concerted movement, called an interstitialcy mechanism, can also be seen as the migration of an extended aluminium interstitial defect (denoted by a dotted box), comprising the interstitial atom and its two neighbours along the c-axis, from one interstitial site to another.

Figure 4.37: The migration barriers for the two considered diffusion mechanisms of an aluminium interstitial, extended interstitial diffusion (Path 1) and simple interstitial diffusion (Path 2), calculated in a $2 \times 2 \times 1$ supercell. The former has a far lower barrier of 3.13eV, compared with 4.35eV, and involves concerted movements between an aluminium interstitial and one of its nearest-neighbours along the c-axis, as depicted in Fig. 4.36.
Figure 4.38: The migration barriers for individual migration of the aluminium (a) and oxygen (b) vacancies contained in a $V_{\text{AlO}}$ defect to its aluminium or oxygen nearest-neighbours, respectively. The distances in the legends refer to the relaxed structure of the $V_{\text{AlO}}$ defect, as strong lattice relaxations cause the nearest-neighbour separations to change dramatically. The end points on the plot are non-zero as they correspond to a different structure than the start points, as the AlO vacancy has been partly filled by either an aluminium or oxygen atom, separating the divacancy into two non-nearest-neighbour vacancies.

formation energies (see Table 4.2).

AlO vacancy

The diffusion of AlO vacancies in alumina can occur by many different mechanisms, of which two are considered in detail in this study. There are four nearest-neighbour aluminium atoms to the aluminium vacancy contained in a $V_{\text{AlO}}$ defect, at distances between 2.40 and 2.80Å. Six oxygen atoms that are bonded to these four aluminium atoms are also nearest-neighbours to the oxygen vacancy contained in the $V_{\text{AlO}}$ complex. Either one of the aluminium atoms can migrate into the aluminium vacancy in the $V_{\text{AlO}}$ defect, and then one of its previously bonded oxygen atoms can follow (or vice-versa), or the two can migrate together as a dumbbell. For the first scenario, the migration barriers are found by migrating each aluminium and oxygen atom into the aluminium or oxygen vacancy individually. The results, plotted in Figs. 4.38(a) and (b), show that the three aluminium atoms nearest to the aluminium vacancy all have similar migration barriers between 2.10 and 2.41eV,
Figure 4.39: The migration barriers of the AlO vacancy as a mixed dumbbell to nearest-neighbour AlO pairs. The AlO pairs are migrated into the AlO vacancy while allowing their bond to stretch perpendicular to the migration path. The path numbers correspond to the oxygen nearest-neighbours from Fig. 4.38(b), who are partnered with the aluminium atoms they share a bond with from Fig. 4.38(a). The resulting barriers are much higher than for separate migration of the constituent vacancies.

whereas the furthest aluminium neighbour has a barrier of 3.17eV, while the nearest oxygen neighbour has a migration barrier over 1eV lower than all other oxygen neighbours, at 1.12eV.

Alternatively, these six oxygen atoms and their four aluminium partners are able to migrate into the AlO vacancy as mixed dumbbells, where the centre of mass of the dumbbell is constrained to move along a jump vector to nearest-neighbour AlO pairs. This allows the AlO bond to stretch or rotate along the migration path. The corresponding dumbbell migration barriers are shown in Fig. 4.39. The migration barriers, all > 3eV, are significantly higher than for the separate migration of the constituent vacancies, suggesting that simple dumbbell diffusion is unlikely. Selected other paths were examined that included rotations of oxygen atoms around aluminium nearest-neighbours to the aluminium vacancy, but gave rise to barriers in excess of 8.50eV.

However, the aluminium and oxygen defects comprising the migrating AlO dumbbell need not move together, they can also move in lock-step, with one atom taking a
Figure 4.40: The migration barrier corresponding to the aluminium atom that shares the oxygen atom with the lowest migration barrier in Fig. 4.38(b) (Path 2) as a nearest-neighbour with the AlO vacancy, after the oxygen atom has fully migrated into the oxygen vacancy contained in the \( V_{\text{AlO}} \) defect. The barrier is referenced to the ground-state energy of a \( V_{\text{AlO}} \) defect, and, as in Figs. 4.38(a) and (b), the start and end points correspond to different defect structures. At the start, the oxygen atom has already fully migrated into the oxygen vacancy, so that the total energy of the supercell is higher than for the ground-state AlO vacancy. At the end point, the AlO vacancy has been restored on the nearest-neighbour sites. The energy maximum along the migration path is 2.30eV.

step forward, and the second following. Alternatively, one of the two can migrate into the AlO vacancy alone, for the other to follow. Given the relatively high migration barriers of aluminium part-vacancy in Fig. 4.38(a), it is likely that an oxygen atom will part-fill the AlO vacancy, allowing for an aluminium atom to follow. This requires two processes: first, the oxygen neighbour with the lowest migration barrier in Fig. 4.38(b) (1.12eV) fills the oxygen vacancy, and second, the aluminium atom that shares the migrating oxygen atom with the aluminium vacancy fills the aluminium vacancy. For the second step in this process, we find that the migration barrier for the aluminium vacancy is lowered to 2.30eV, as seen in Fig. 4.40. This was the value subsequently used for the total migration barrier of an AlO vacancy, as it represents the minimum energy barrier for a full migration of both constituent atoms in the defect. Alternatively, the two atoms can move in lock-step, such that they are separated by either 0.1, 0.2, \ldots, 0.9 fractions of the migration path, as a more or less-stretched dumbbell. This was not found to lower the diffusion barrier, and the minimum energy path was found to be that where the oxygen atom had fully
Density functional theory calculations

Section 4.4: Density functional theory calculations

migrated into the oxygen vacancy before any significant movement of the aluminium atom.

Finite-size effects

Just as defect formation energies must be corrected to account for the small size of computational supercells, so must migration barriers. Such calculations have not been performed in the past for this system, but this study suggests that, in particular for complex migration paths, they are necessary. Using the same method as for defect formation energies, described in Sec. 3.5.3, we examine the migration barriers for each of the native point defects in alumina. Beginning with the relaxed ground-state geometry of each defect, the migrating atom(s) are placed at the saddle point of the lowest-energy path. The migrating atom(s) are then allowed to relax in a plane perpendicular to their jump vector in each respective supercell, while the rest of the atoms in the supercell are fully relaxed. The difference in energy of the relaxed configuration at the top of the migration barrier and the original relaxed geometry is then taken as the migration barrier for a given supercell. For these calculations, the $2 \times 2 \times 1$, $3 \times 3 \times 2$, $3 \times 2 \times 2$, $2 \times 2 \times 2$, and $2 \times 2 \times 3$ supercells were used with Vanderbilt USPs and a plane wave cut-off of 400eV in the LDA. Little dependence ($< 0.05$eV) of the migration barrier on the plane wave cut-off was found beyond 400eV, allowing for more efficient, lower cut-off calculations.

Fig. 4.41 shows the uncorrected and corrected migration barriers as a function of the Madelung potential $v_M$. The uncorrected values show a clear supercell size-dependence, with spreads in the values obtained for different supercells ranging from 0.07 to 1.71eV. For all defects except oxygen interstitials, the migration barrier decreases with decreasing $v_M$. For the corrected values, the Madelung interpolation from Sec. 3.5.3 was used, revealing much smaller spreads, ranging from 0.02 to 0.60eV. These dependencies emphasise that the long-ranged elastic and Coulomb fields cannot be ignored when determining diffusion barriers in oxides. The finite-size errors are larger for defects with more complex diffusion mechanisms, such as aluminium vacancies and interstitials. In both of these cases, an accurate description of the stress along the c-axis is crucial for determining an accurate migration barrier. For example, in the $2 \times 2 \times 1$ supercell, an aluminium split vacancy has only one set of nearest-neighbour aluminium atoms along the c-axis. In periodic
boundary conditions, this means that there is only one aluminium atom separating periodic images of the split vacancy from each other. Without any ‘bulk’ aluminium atoms separating the defects, lattice relaxations along the c-axis cannot be accurately described. In supercells of greater dimension along the c-axis, there are more aluminium atoms that can relax to accommodate these strains, so that the cell boundary is significantly more bulk-like. Defects that migrate by simple mechanisms, such as the oxygen vacancy, and already have a number of nearest-neighbour shells within the supercell that can relax to lower the energy significantly, exhibit much smaller finite-size errors. This is to be expected and shows that the finite-size corrections here are, to a large degree, dealing with elastic effects. The migration
Table 4.13: The migration barriers $\Delta E_{i}^{\text{mig}}$ of the four native point defects in alumina, extrapolated to the infinite-cell limit using the Madelung interpolation method.

<table>
<thead>
<tr>
<th>Defect species $i$</th>
<th>$\Delta E_{i}^{\text{mig}}$ (eV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V_O$</td>
<td>1.73</td>
</tr>
<tr>
<td>$V_{Al}^{''}$</td>
<td>1.27</td>
</tr>
<tr>
<td>$O_{Al}^{''}$</td>
<td>0.66</td>
</tr>
<tr>
<td>$Al_{i}^{'''}$</td>
<td>1.31</td>
</tr>
</tbody>
</table>

4.5 Defect concentrations

Following the methodology presented in Sec. 3.3.2, the DFT formation and binding energies calculated in Sec. 4.4.3 are combined in a mass-action approach to determine defect and cluster concentrations as a function of dopant concentrations, temperature $T$ and oxygen partial pressure $pO_2$. The concentration $c_i$ of defect $i$ is given by

$$c_i = m_i \exp \left( -\frac{\Delta G_i^f}{k_B T} \right),$$  \hspace{1cm} (4.12)

where $\Delta G_i^f$ is the defect formation energy (or analogously for cluster binding energies) and $m_i$ the multiplicity of defect $i$.

The Brouwer-type diagram in Fig. 4.42 shows the self-consistent concentrations of native defects and the most prominent defect-dopant clusters at typical annealing conditions of $T = 1750K$, $pO_2 = 0.2p^0$ as a function of aliovalent dopant concentration. The right-hand side of the diagram corresponds to magnesium doping, the left-hand side to titanium doping. The centre of the plot coincides with a doping level of $10^{-12} = 10^{-6}$ppm, beyond modern detection techniques. In theory, the diagram can be expanded around its centre to show lower levels of doping, but
this is unnecessary to explain the defect chemistry of alumina at the conditions presented here. The highest levels of doping, at the left and right extremes along the \(x\)-axis, correspond to 50000 ppm, which suffices for current purposes and may also approach the level where the dilute ideal solution approximation loses validity. In Fig. 4.42, the blue, red, green, and grey lines show \([V''_{\text{Al}}]\), \([V'_{\text{O}}]\), \([\text{Al}^+_{\text{r}}]\), and \([V'_{\text{AlO}}]\) respectively, where \([i]\) represents the per site concentration of defect species \(i\). These are the most numerous charge states of these native defect species under any experimentally-attainable conditions and doping levels. Oxygen interstitials, which form as neutral \([O\hat{x}_i]\) defects, are many orders of magnitude lower in concentration than the other native defects due to their relatively high defect formation energy, and are not explicitly shown. The pink line shows \([\text{Ti}_{\text{Al}}]\), and the olive line \([\text{Mg}^\prime_{\text{Al}}]\). Finally, the purple, brown, and cyan lines show \([\{\text{Ti}_{\text{Al}} : \text{V}_{\text{Al}}\}'\], \([\{\text{Mg}_{\text{Al}} : \text{V}_{\text{O}}\}'\], and \([\{2\text{Ti}_{\text{Al}} : \text{V}_{\text{Al}}\}'\], respectively. For Fig. 4.42, only the concentrations of the dominant clusters are shown explicitly, although many more were considered, some of which are shown in Fig. 4.43(a) and described hereafter. At the bottom of Fig. 4.42, the red line shows how the Fermi level \(\epsilon_F\) varies (right-hand scale) with doping.

The trends revealed in Fig. 4.42 explain many aspects of the point defect phenomenology of alumina. The concentrations of defect species clearly depend on both the type and level of doping. The centre of the graph shows the undoped, or intrinsic, regime, where defect concentrations are very low: \([V''_{\text{Al}}]\) \(\approx\) \([V'_{\text{O}}]\) \(\approx\) \(10^{-9}\), \([V'_{\text{AlO}}]\) \(\approx\) \(10^{-12}\), \([\text{Al}^+_{\text{r}}]\) \(\approx\) \(10^{-15}\). Charge neutrality in this regime is maintained by the Schottky mechanism, \(2[V''_{\text{Al}}] \approx 3[V'_{\text{O}}]\). At \(T = 1750K\) and \(p_{O_2} = 0.2p_0\), this equilibrium occurs when \(\epsilon_F = 1.60eV\), just below the position in the band gap where oxygen vacancies introduce localised defect states.

### Mg-doped alumina

Focusing on magnesium-doped alumina first, the vast majority of magnesium substitutes as \(\text{Mg}^\prime_{\text{Al}}\) up to very high levels of Mg-doping. Departing from ‘pure’ alumina to higher levels of magnesium doping (i.e. moving from the centre of Fig. 4.42 along the \(x\)-axis towards the right), several changes in the concentrations of native defects are noteworthy. Focusing on the oxygen vacancy concentration, \([V'_{\text{O}}]\) remains flat until \([\text{Mg}]\) rises to ca. \(10^{-9} = 10^{-3}\) ppm, as it is the dominant charged defect and easily absorbs any change in the relatively minute concentration of \(\text{Mg}^\prime_{\text{Al}}\). When \([\text{Mg}^\prime_{\text{Al}}]\) \(\approx\) \([V'_{\text{O}}]\), the charge introduced by additional magnesium impurities requires
Figure 4.42: A Brouwer-type diagram of defect and cluster concentrations as a function of magnesium concentration (right-hand side; defined as ‘positive’ concentration) and titanium concentration (left-hand side; defined as ‘negative’ concentration) at $T = 1750\text{K}$, $p_{O_2} = 0.2p^0$. Only the most populous native defects and clusters are shown, in their dominant charge states. At the bottom, the variation of the Fermi level $\epsilon_F$ with impurity doping is shown (right-hand scale).

$[V'O]$ to rise substantially to ensure charge neutrality. Simultaneously, $[V''''\text{Al}]$ must fall because the Schottky equilibrium requires the product $[V'O]^3 [V''''\text{Al}]^2$ to remain constant, as

$$[V'O]^3 [V''''\text{Al}]^2 = k_S \exp \left( \frac{\Delta G^f_S}{k_B T} \right),$$

(4.13)

where $k_S$ is a proportionality constant, and $\Delta G^f_S$ the normalised Schottky energy. As $[\text{Mg}]$ is increased to extremely high levels (the right hand side edge of Fig. 4.42 corresponds to $[\text{Mg}] = 50000\text{ppm}$, which is 2-3 orders of magnitude higher than the impurity levels seen in experimentally-doped samples), $[V'O]$ increases slower than for lower levels of doping. This is because charge neutrality in this very heavily doped regime is largely ensured by $\{\text{Mg}\text{Al} : V'O\}$ clusters. As $[\text{Mg}]$ is increased from zero,
the concentration \([\{\text{Mg}_{\text{Al}} : V_{\text{O}}\}']\) rises in tandem with \([\text{Mg}_{\text{Al}}']\) until \([V_{\text{O}}] \approx [\text{Mg}_{\text{Al}}']\). From this point onwards, \([V_{\text{O}}]\) rises in line with \([\text{Mg}_{\text{Al}}']\), and \([\{\text{Mg}_{\text{Al}} : V_{\text{O}}\}']\) rises by the product \([V_{\text{O}}] \times [\text{Mg}_{\text{Al}}']\). At even higher levels of Mg-doping, when we reach the point where \([\{\text{Mg}_{\text{Al}} : V_{\text{O}}\}'] \approx [\text{Mg}_{\text{Al}}]\), \([V_{\text{O}}]\) begins to level off, while the \([\{\text{Mg}_{\text{Al}} : V_{\text{O}}\}']\) clusters become the dominant charge-compensating defect to subsequent Mg-doping due to their significant binding energy. This occurs because, at all levels of doping, the relationships between these three defects must follow the equation

\[
\text{Mg}_{\text{Al}}' + V_{\text{O}}^- \rightleftharpoons \{\text{Mg}_{\text{Al}} : V_{\text{O}}\}', \quad (4.14)
\]

which is governed by the mass action expression

\[
\frac{[\text{Mg}_{\text{Al}}'] [V_{\text{O}}]}{[\{\text{Mg}_{\text{Al}} : V_{\text{O}}\}']} = k_0 \exp\left(-\frac{\Delta G_{\text{bind}}^{\text{Mg}_{\text{Al}} : V_{\text{O}}}}{k_B T}\right), \quad (4.15)
\]

where \(k_0\) is a proportionality constant and \(\Delta G_{\text{bind}}^{\text{Mg}_{\text{Al}} : V_{\text{O}}}\) the binding energy of the \([\{\text{Mg}_{\text{Al}} : V_{\text{O}}\}']\) cluster, tabulated in Table 4.11.

Equivalently, the defect chemistry can be described by the behaviour of the Fermi level. In the intrinsic regime, \(\epsilon_F\) is constant, as any additional charged dopants are easily compensated by minute changes in the dominant defect concentration, \([V_{\text{O}}^-]\). Beyond this regime, \(\epsilon_F\) must fall to create charge-compensating defects, until \([\text{Mg}_{\text{Al}}'] \approx [\{\text{Mg}_{\text{Al}} : V_{\text{O}}\}']\). This change in \(\epsilon_F\) lowers the formation energy of oxygen vacancies, increasing their population to ensure charge neutrality. When \([\text{Mg}_{\text{Al}}'] \approx [\{\text{Mg}_{\text{Al}} : V_{\text{O}}\}']\), \(\epsilon_F\) begins to level off, as additional [Mg] introduces little net charge, since increases in \([\text{Mg}_{\text{Al}}']\) are primarily balanced by increases in \([\{\text{Mg}_{\text{Al}} : V_{\text{O}}\}']\).

The concentrations of other native defects can be explained analogously to \(V_{\text{O}}^-\). As \(\epsilon_F\) falls, \(\Delta G_{V_{\text{Al}}''}^f\) rises, and the concentration \([V_{\text{Al}}'']\) falls. \(V_{\text{Al}}''\) defects behave similarly but less pronouncedly, due to their lower charge state. The concentration of aluminium interstitials, that also act as charge-compensating defects under magnesium doping, mirrors that of \(V_{\text{O}}^-\), albeit at lower absolute concentrations, as \(\Delta G_{\text{Al}^+}^f > \Delta G_{V_{\text{O}}^-}^f\). Other defect-dopant clusters, such as \([\text{Mg}_{\text{Al}} : \text{Al}^+]'\), behave similarly to \([\text{Mg}_{\text{Al}} : V_{\text{O}}]'\), but are much lower in concentration. Meanwhile, the concentration of neutrally-charged oxygen interstitials (not shown due to their very low concentrations) remains constant upon doping, as \(\epsilon_F\) does not affect \(\Delta G_{O_{\text{i}}}^f\), or, they do not contribute any charge to the crystal.
Section 4.5: Defect concentrations

Ti-doped alumina

Titanium-doped alumina exhibits similar behaviour, primarily governed by the reaction $\text{Ti}_\text{Al} + V''''_\text{Al} \rightleftharpoons \{\text{Ti}_\text{Al} : V_\text{Al}\}''''$ and mass action relation

$$\frac{[\text{Ti}_\text{Al}] [V''''_\text{Al}]}{[\{\text{Ti}_\text{Al} : V_\text{Al}\}'''']} = k_1 \exp \left( -\frac{\Delta G^\text{bind}_{\{\text{Ti}_\text{Al} : V_\text{Al}\}''''}}{k_B T} \right).$$  \hspace{1cm} (4.16)

When small amounts of titanium are added to a pure crystal, it overwhelmingly forms $\text{Ti}_\text{Al}$ substitutionals, and the charge is balanced by $V''''_\text{Al}$ defects in the intrinsic regime. This contrasts with previous assertions made by Matsunaga et al. [121], who proposed that $[\text{Ti}_\text{Al}] \approx [\text{Ti}_\text{Al}]$ when comparing the formation energies of $\text{Ti}_\text{Al}$ and a number of neutral defect clusters that bind $\text{Ti}_\text{Al}$ for different values of $\mu_\text{Al}$ and $\mu_\text{O}$. There is no reason to assume that $\text{Ti}_\text{Al}$ cannot exist freely, or bound in charged clusters, and Matsunaga et al.’s inadequate treatment of band gap errors and neglect of finite-size effects casts doubt on this assertion.

We see in Fig. 4.42 that over the intrinsic regime, where the impurity concentrations are less than $10^{-8}$, both the dominant defects $[V''''_\text{Al}]$ and $[V''''_\text{O}]$ are almost constant on a logarithmic scale, as determined by the Schottky equilibrium. This is because only small fractional adjustments in $[V''''_\text{Al}]$ and $[V''''_\text{O}]$ are required to compensate the charges introduced by aliovalent impurities, whether they are titanium or magnesium. This causes $\epsilon_F$ to remain roughly constant. Eq. 4.16 limits the amount of titanium that can exist in clusters. In the intrinsic regime, the concentration of trapped vacancies, $[\{\text{Ti}_\text{Al} : V_\text{Al}\}'''']$ remains very low, but grows linearly in proportion to $[\text{Ti}_\text{Al}]$. Once $[\text{Ti}]$ is increased such that $[\text{Ti}_\text{Al}] \approx [V''''_\text{Al}]$, significantly more $V''''_\text{Al}$ defects must be produced to balance the positive charge from $\text{Ti}_\text{Al}$ dopants, causing $\epsilon_F$ to rise. Simultaneously, the Schottky equilibrium suppresses $[V''''_\text{O}]$. All the while, $[\{\text{Ti}_\text{Al} : V_\text{Al}\}'''']$ rises as $[\text{Ti}_\text{Al}]^2$, since by Eq. 4.16 both $[V''''_\text{Al}]$ and $[\text{Ti}_\text{Al}]$ are growing linearly, until we reach very high levels of doping, where $[\text{Ti}] \approx 5000$ppm and $[\{\text{Ti}_\text{Al} : V_\text{Al}\}''''] \approx \frac{1}{2} [\text{Ti}_\text{Al}]$, after which $\{\text{Ti}_\text{Al} : V_\text{Al}\}''''$ clusters compensate further increases in $[\text{Ti}]$. At even more extreme levels of doping, $\{2\text{Ti}_\text{Al} : V_\text{Al}\}'$ clusters, governed by the reaction $2\text{Ti}_\text{Al} + V''''_\text{Al} \rightleftharpoons \{2\text{Ti}_\text{Al} : V_\text{Al}\}'$ and mass action relation

$$\frac{[\text{Ti}_\text{Al}]^2 [V''''_\text{Al}]}{[\{2\text{Ti}_\text{Al} : V_\text{Al}\}']} = k_1 \exp \left( -\frac{\Delta G^\text{bind}_{\{2\text{Ti}_\text{Al} : V_\text{Al}\}'}}{k_B T} \right) \hspace{1cm} (4.17)$$

surpass $\{\text{Ti}_\text{Al} : V_\text{Al}\}''''$ clusters as the primary charge-compensating defect to further
increase in [Ti]. When this occurs, $[V''''_{Al}]$ decreases by Eq. 4.17, and $\epsilon_F$ consequently falls again. However, these high levels of doping approach the limits of the dilute ideal solution approximation and the solubility of TiO$_2$ ($\sim 10000$ ppm [156]), and so these results must be treated with great care.

Although the defect-dopant clusters $\{\text{Mg}_{Al} : V_O\}$, $\{\text{Ti}_{Al} : V_{Al}\}''$, and $\{2\text{Ti}_{Al} : V_{Al}\}'$ are the most numerous in alumina, other clusters are also present in lower concentrations. Of the 23 clusters included in this study, those whose per site concentrations are higher than $10^{-30}$ in a heavily-doped system at $T = 1750K$, $p_{O_2} = 0.2p^0$ are shown in Fig. 4.43(a). For magnesium-doped alumina, the concentration of $\{\text{Mg}_{Al} : \text{Al}_i\}^-$ clusters is 2-5 orders of magnitude lower than the dominant cluster, $\{\text{Mg}_{Al} : V_O\}^-$. In titanium-doped alumina, the next most populous clusters behind $\{\text{Ti}_{Al} : V_{Al}\}''$ and $\{2\text{Ti}_{Al} : V_{Al}\}'$, $\{3\text{Ti}_{Al} : V_{Al}\}^\times$ and $\{\text{Ti}_{Al} : \text{O}_i\}^\times$, are 15 orders of magnitude lower in concentration (at heavy doping). Oxygen interstitials, whether free-standing or bound into $\{\text{Ti}_{Al} : \text{O}_i\}$ clusters, play no role in the defect chemistry of alumina due to their very low concentrations.

Under most experimental conditions, a single charge state $q_i$ for each defect tends to dominate all others. However, defects can still exist in other charge states, albeit in much lower concentrations. Fig. 4.43(b) shows the concentrations of $V''''_{Al}$, $V''_{Al}$, $V'_{Al}$, and $V^\times_{Al}$ defects at $T = 1750K$, $p_{O_2} = 0.2p^0$. The most dominant state is generally $V''''_{Al}$, with the exception of heavily Mg-doped samples (where the concentration of all $V_{Al}$ defects is very low). Following this, $V''_{Al}$ defects are 2-3 orders of magnitude lower in concentration in Ti-doped samples; the $V'_{Al}$ concentration is four orders of magnitude lower than the $V''_{Al}$ concentration; and the $V^\times_{Al}$ concentration is five orders of magnitude lower than the $V'_{Al}$ concentration. When $\epsilon_F$ moves significantly between the undoped and heavily doped regimes, the changes in concentration of each of these charge states differ due to the $q_i\mu_e$ contribution to the formation energy.

When examining the oxygen or aluminium diffusion coefficients, it is important to sum over all the charge states of free defects, as any of these can participate in diffusion.

Fig. 4.44 shows the variation in defect concentrations for various different temperatures and oxygen partial pressures. Fig. 4.44(a) corresponds to $T = 1517K$, $p_{O_2} = 0.2p^0$, (b) to $T = 1982K$, $p_{O_2} = 0.2p^0$, and (c) to $T = 1750K$, $p_{O_2} = 10^{-5}p^0$. Changes in experimental conditions feed into defect concentrations via changes in the oxygen chemical potential $\mu_O$ (and thereby also $\mu_{Al}$), that alter defect formation
Figure 4.43: Selected secondary cluster concentrations (a) and the concentrations of various charge states of a $V_{Al}$ defect (b), as a function of impurity doping at $T = 1750$K, $p_{O_2} = 0.2p^0$. Included in (a) are the dominant defect clusters for Mg- and Ti-doping, \{Mg$_{Al}$ : $V_O$\}, \{Ti$_{Al}$ : $V_{Al}$\}$''$, and \{2Ti$_{Al}$ : $V_{Al}$\}$'$, for reference. Apart from \{Mg$_{Al}$ : Al$_{i}$\}$^{-}$, all other cluster concentrations are far lower than the dominant species.
Figure 4.44: The concentrations of the most populous native defects and clusters in alumina at $T = 1517\,\text{K}$, $p_{\text{O}_2} = 0.2p^0$ (a), $T = 1982\,\text{K}$, $p_{\text{O}_2} = 0.2p^0$ (b), and $T = 1750\,\text{K}$, $p_{\text{O}_2} = 10^{-5}p^0$ (c).
energies. For example, an increase in $T$ or a decrease in $p_{O_2}$ corresponds to lowering $\mu_O$ and raising $\mu_{Al}$ (by $3/2$ the amount of $\mu_O$). This, in turn, lowers the formation energies of $V_O$ and $Al_i$ defects, and raises those of $V_{Al}$ and $O_i$. In addition, these changes in $\Delta G_f^i$ shift the Fermi level at which charge neutrality is obtained in the intrinsic regime by the Schottky mechanism, leading to further changes in the formation energies. Solving for the defect concentrations self-consistently accounts for all of these factors simultaneously.

Several global themes are evident, regardless of doping and formation conditions. First, the concentrations of native interstitials are generally lower than those of vacancies, so that the Schottky mechanism always dominates intrinsic crystals. Second, the AlO divacancy and the oxygen interstitial exist as $V'_{AlO}$ and $O_i^\times$, not as $V_{AlO}^\times$ and $O_i^\prime$ as previously assumed. For the latter, the dumbbell interstitial configuration, that causes the two oxygen ions forming the dumbbell to be significantly more covalently bonded than bulk oxygen ions, leads to this reduced charge state. This contrasts strongly with the traditional ceramists’ view of ionicity in alumina. Third, the dominant defect-dopant clusters are always $\{Mg_{Al}: V_O\}$, $\{Ti_{Al}: V_{Al}\}''$, and $\{2Ti_{Al}: V_{Al}\}'$. Fourth, no other noteworthy clusters or charge states are seen in the sampled range of $T$ and $p_{O_2}$.

In contrast, many effects caused specifically by changes in $T$ and $p_{O_2}$, and thereby in $\epsilon_F$, can be inferred from Figs. 4.44(a)-(c). For example, raising the temperature increases defect concentrations across the board, pushing the impurity content at which charge-compensating defects are formed in significant amounts relative to their intrinsic concentrations to higher levels of doping. The Fermi level $\epsilon_F$ rises as $T$ increases or $p_{O_2}$ decreases. Among other things, an increase in $\epsilon_F$ causes $\Delta G_{V_{AlO}}^{IF}$ to decrease relative to $\Delta G_{Al_i}^{IF}$, and thereby $[V'_{AlO}]$ to increase relative to $[Al_i^{-}]$, which can easily be identified in the intrinsic regime of Figs. 4.44(a)-(c)). Also, in 1977, Mohapatra et al. [128] attributed the behaviour of electronic and ionic conductivity with decreasing $p_{O_2}$ in Ti-doped alumina to the formation of $Ti_i^{\times}$ defects. This study refutes this assertion, as the proportion of $Ti_i^{\times}$ defects is roughly constant, regardless of $p_{O_2}$. A number of other inferences can be made from these plots, but will not be explicitly examined here.
In summary, these results do not demonstrate the buffering effect seen experimentally, whereby the oxygen diffusion coefficient $D_O$ is relatively insensitive to impurity doping. For buffering to occur, $D_O$, and thereby $[V_O^-]$, must be limited to rise by a factor of roughly 10 on heavy Mg-doping, and fall by a similar factor on heavy Ti-doping. In contrast, in our results $[V_O^-]$ rises by 3-6 orders of magnitude upon Mg-doping and falls by 2-4 orders of magnitude upon Ti-doping. However, the results do suggest that at extreme levels of doping, $[V_O^-]$ levels off as the dominant defect-dopant clusters rise in concentration enough such that they become the primary charge-compensating defects.

We suggest that defect-dopant clustering could potentially still be the physical cause for the buffering phenomenon seen experimentally, and that two reasons could have prevented us from demonstrating this. First, this study could have simply omitted the necessary charge-compensating clusters that buffer the oxygen vacancy population. Although 23 different clusters were considered here, there is no reason to believe that other clusters, or clusters examined here but in other configurations, cannot act as the primary charge-compensating defects in heavily-doped alumina. Using a method such as metadynamics with an accurate classical potential fitted to a DFT dataset that contains different defect ground states in alumina could indicate whether such stable clusters exist. Second, the limits of the LDA and the correction schemes applied in this study (finite-size corrections in particular) could have prohibited the current study from calculating accurate cluster binding energies. For example, the unphysically small LDA band gap leads to an unphysically large dielectric constant, which means that highly-charged defects are over-screened and their ground-state energies thereby abnormally lowered. This stabilises separated highly-charged defects, such as the charge-compensating isolated defects $V_O^-$ and $V_{Al}^{'''}$, relative to bound clusters. Further work involving hybrid functionals, that can reproduce an accurate band gap in alumina, can examine the validity of this assertion in alumina. Unfortunately, current implementations of hybrid functionals are computationally too taxing for calculating cluster binding energies in bulk alumina.

As a thought experiment, if we arbitrarily assume that the LDA underbinds the dominant clusters in alumina by 1eV and raise the binding energies of these clusters by this amount, we obtain the Brouwer-type diagram in Fig. 4.45.

For Ti-doped alumina in this thought experiment, we see that $\{Ti_{Al} : V_{Al}\}''$ de-
Section 4.5: Defect concentrations

Figure 4.45: Thought experiment: the concentrations of native defects and important clusters in alumina at $T = 1750\text{K}$, $p_{O_2} = 0.2p^0$ when raising the binding energies of $\{\text{Mg}_{Al} : V_O\}$, $\{\text{Ti}_{Al} : V_{Al}\}''$, and $\{2\text{Ti}_{Al} : V_{Al}\}'$ by 1eV each.

Defects begin to act as the primary charge-compensating defect at much lower levels of doping due to their increased binding energy. The crossover point where $\{\text{Ti}_{Al} : V_{Al}\}'' \approx \frac{1}{2} [\text{Ti}_{Al}]$ has moved from $[\text{Ti}] \approx 5000\text{ppm}$ in the original system, to $[\text{Ti}] \approx 1\text{ppm}$ in the hypothetical system. From this point onwards, $\{\text{Ti}_{Al} : V_{Al}\}''$ clusters compensate further increases in $[\text{Ti}]$, causing $[V_{Al}']$, and, by the Schottky equilibrium, $[V_O]$ to level off. This is an example of a buffering mechanism that limits $[V_O]$ to change by only a factor of 30 from no titanium doping to the hundreds of ppm used in heavily-doped experiments. Assuming that those defects hoovered up and bound in clusters are immobile, this limits the number of free native defects available for diffusion upon doping. This mechanism could cause the buffering effect experimentally observed for $D_O$. As before, at extremely high levels of Ti-doping, $\{2\text{Ti}_{Al} : V_{Al}\}'$ clusters become the dominant charge-compensating defect, causing $[V_{Al}']$ to decrease, and $[V_O]$ to increase.

Similarly, for Mg-doping, the point at which $[\{\text{Mg}_{Al} : V_O\}] \approx [\text{Mg}_{Al}]$ has moved from $[\text{Mg}] \approx 10000\text{ppm}$ to $[\text{Mg}] \approx 100\text{ppm}$. From this point on, $[V_O]$ stabilises to
a constant value, where any additional rise in [Mg] does not affect its population, as increases in [Mg] are effectively buffered by \( \{\text{Mg}_{\text{Al}} : \text{V}_O\} \) defects. As a corollary, when \( [\text{Mg}'_{\text{Al}}] \approx [\{\text{Mg}_{\text{Al}} : \text{V}_O\}] \), \( \epsilon_F \) levels off, as additional [Mg] introduces no net charge, since increases in \( [\text{Mg}'_{\text{Al}}] \) are balanced by increases in \( [\{\text{Mg}_{\text{Al}} : \text{V}_O\}] \). When \( \epsilon_F \) is constant, it has no effect on the formation energies and, thus, concentrations of charged defects such as \( \text{V} \cdot \cdot \cdot \text{O} \). As for Ti-doped alumina in the thought experiment, \( [\text{V} \cdot \cdot \cdot \text{O}] \) only changes by a factor of 300 from no magnesium doping to the levels used in heavily-doped experiments. Whether or not the LDA causes errors of \( O[1eV] \) in the binding energies of the dominant defect-dopant clusters in alumina can be examined by higher-level methods such as DFT using hybrid functionals or even QMC. Unfortunately, the necessary calculations to examine this are currently computationally intractable.

### 4.6 Diffusion coefficients

In this section, we compare calculated diffusion coefficients from the DFT defect concentrations in Sec. 4.5 and migration barriers in Sec. 4.4.8 to experimental data to highlight how the current study is unable to reproduce the buffering effect that limits changes in the oxygen diffusion coefficient upon significant aliovalent doping. In Sec. 3.6, it was shown that the diffusion coefficient for a defect species \( i \), such as \( \text{V}_O \), can be expressed by the Arrhenius equation

\[
D_i = \frac{1}{2} f_i \nu_i \exp \left( -\frac{Q^\text{act}_i}{k_B T} \right) \lambda_i^2 \sum_{j=1}^{z} \cos^2 \varphi_{i,j} = D_{i,0} \exp \left( -\frac{Q^\text{act}_i}{k_B T} \right),
\]

(4.18)

where \( D_{i,0} \) is termed the diffusion ‘pre-exponential’. The correlation factors \( f_i \) were determined in Sec. 3.6.5 as \( f_{\text{V}_O} \approx 0.47, f_{\text{V}_{\text{Al}}} \approx 0.36, f_{\text{Al}} \approx 0.95, \) and \( f_{\text{O}} = 1.00 \), and the activation energies are approximated as sums of the formation energies and migration barriers found in the preceding sections, \( Q^\text{act}_i \approx \Delta G^f_i + \Delta E^\text{mig}_i \). The jump distances \( \lambda_i \) are taken from the lowest-energy diffusion barriers calculated in Sec. 4.4.8, and the vibrational frequency \( \nu_i \) estimated as \( \approx 1.5 \times 10^{13} \) from the peak of the phonon spectra of the defects. This leaves only the angles \( \varphi_{i,j} \) spawning the jump vectors of the defect \( i \) to nearest-neighbour site \( j \) and the diffusion axis for determination. These can be very simply evaluated by choosing a diffusion axis. In
order to draw comparison with experiment, and as we have established in Sec. 4.4.8 that fast aluminium vacancy diffusion along the \(c\)-axis can occur, we evaluate the \(\varphi_{i,j}\) for \textit{basal} diffusion to find:

\[
\sum_{j=1}^{z} \cos^2 \varphi_{V_{O,j}} \approx 2.27
\]

\[
\sum_{j=1}^{z} \cos^2 \varphi_{V_{Al,j}} \approx 2.95
\]

\[
\sum_{j=1}^{z} \cos^2 \varphi_{Al_{i,j}} \approx 3.00
\]

\[
\sum_{j=1}^{z} \cos^2 \varphi_{O_{i,j}} \approx 1.90
\]

using the relaxed defect geometries. For the aluminium split vacancy configuration, either of the two vacancies can, in theory, diffuse to one of three equivalent nearest-neighbours sporting the lowest diffusion path. However, only one of the two vacancies will actually do so, while the other will subsequently relax towards the aluminium atom sitting in the interstitial site in order to recreate the bulk structure. Therefore, only one of the two vacancies contributes to the diffusion flux, so that the sum \(\sum_{j=1}^{z}\) only runs over the nearest-neighbours of one of the two identical vacancies. Similarly, for the extended aluminium interstitial defect, only one of the two neighbours flanking the interstitial aluminium atom along the \(c\)-axis will migrate to one of the three equivalent nearest-neighbours with a lowest-energy diffusion path, so that the sum also only runs over the neighbours of one of the two identical aluminium atoms that have the ability to migrate. Analogously, only the two nearest-neighbours of one of the two oxygen atoms comprising the oxygen dumbbell interstitial are considered.

Predicted diffusion coefficients at typical experimental doping levels, calculated from Eq. 4.18, are compared with the ‘best available’ experimental data in Fig. 4.46. For \(D_{O}\), the most reliable diffusion data comes from recent dislocation loop annealing measurements performed by Lagerlöf \textit{et al.} [99] on undoped, Mg-, and Ti-doped single crystal samples. The only other measurement of \(D_{O}\) in Mg-doped alumina, by Haneda and Monty [73], differs from the results obtained by Lagerlöf \textit{et al.} by three orders of magnitude (as shown in Fig. 4.1), highlighting the significant experimental
Figure 4.46: Calculated diffusion coefficients for (a) oxygen vacancies (red) and (b) aluminium vacancies (blue) and interstitials (green) in undoped, Mg-doped, and Ti-doped alumina (solid, dashed, double-dashed lines, respectively). Experimental loop annealing data from Lagerlöf et al. [99] (a) and radio-tracer data from Fielitz et al. [57] (b) (both black) are shown for comparison.

errors attached to measurements of $D_O$ in doped alumina. These errors arise in particular from existing background impurities in the sample, damaged surfaces in the case of radiotracer data, and small datasets. The difficulty in experimentally determining $D_{Al}$ is reflected in a paucity of data. We use the most recent study comparing $^{18}$O and $^{26}$Al tracer diffusion in Ti-doped alumina single crystals, by Fielitz et al. [57] for comparison, but note that there is no true experimental consensus on $D_{Al}$.

In Fig. 4.46(a), calculated values of $D_{Vo}$ are compared with experimental determinations of $D_O$ for undoped, Mg-, and Ti-doped crystals. Due to the comparatively high formation energy of oxygen interstitials, oxygen transport is mediated by oxygen vacancies.

The calculated oxygen diffusion coefficient from DFT for undoped alumina agrees well with experimental data, lying within one order of magnitude. However, for doped alumina, this is not the case. Experimentally, the buffering effect causes $D_O$ to rise by a factor of roughly 10 on Mg doping, and fall by a similar factor on
4.6: Diffusion coefficients

Table 4.14: Activation energies $Q_{\text{act}}^i$ (all in eV) for oxygen and aluminium diffusion in undoped, Mg-, and Ti-doped crystals, inferred from the Arrhenius plots in Fig. 4.46 and compared with available loop annealing (\(^1\)) and tracer diffusion (\(^2\)) experiments from Refs. [99] and [57].

<table>
<thead>
<tr>
<th></th>
<th>$Q_{\text{act}}^{\text{VO}}$</th>
<th>$Q_{\text{act}}^{\text{VO}}$ (exp.)(^1)</th>
<th>$Q_{\text{act}}^{\text{VAI}}$</th>
<th>$Q_{\text{act}}^{\text{Al}}$</th>
<th>$Q_{\text{act}}^{\text{Al}}$ (exp.)(^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>undoped</td>
<td>6.1</td>
<td>6.1 ± 0.2</td>
<td>5.6</td>
<td>5.6</td>
<td>-</td>
</tr>
<tr>
<td>Mg-doped</td>
<td>2.2</td>
<td>5.8 ± 0.4</td>
<td>8.7</td>
<td>0.2</td>
<td>-</td>
</tr>
<tr>
<td>Ti-doped</td>
<td>8.6</td>
<td>7.0 ± 0.4</td>
<td>1.6</td>
<td>9.2</td>
<td>3.9 ± 0.3</td>
</tr>
</tbody>
</table>

Ti doping. The calculated diffusion coefficient from DFT, in contrast, falls by 2-4 orders of magnitude upon Ti doping, and rises by 3-6 orders of magnitude upon Mg doping. The large changes in the calculated oxygen diffusion coefficient upon doping occur because, as discussed in Sec. 4.5, the oxygen vacancy concentration changes by many orders of magnitude when the crystal is doped with a few hundred ppm of impurities. As suggested in Sec. 4.5, defect clustering could still be the mechanism that limits changes in defect populations upon doping and leads to the buffering effect, but inadequacies in the current study have prevented us from demonstrating this. Interestingly, the results for Mg-doped alumina disagree with experiment more so than those for Ti-doped alumina, despite the most significant error, the finite-size corrections applied, being greater for the titanium-containing clusters. This suggests that it is possible that another, more stable, magnesium-containing cluster that was not considered in the present framework may be the culprit.

Calculated activation energies are inferred from Fig. 4.46 and compared with experiment in Table 4.14. Only the value for $Q_{\text{act}}^{\text{VO}}$ in undoped alumina, 6.1eV, agrees well with the experimental determination, $Q_{\text{O}}^{\text{act}} = 6.1 ± 0.2$eV. Experimentally, $Q_{\text{O}}^{\text{act}}$ falls by $\sim 0.3$eV on Mg-doping, and rises by $\sim 0.9$eV on Ti-doping. The calculated values for $Q_{\text{act}}^{\text{VO}}$ replicate these trends, but overestimate the changes by several eV.

Fig. 4.46(b) compares calculated values of $D_{\text{VAI}}$ and $D_{\text{Al}}$ in undoped, Mg-, and Ti-doped crystals with Fielitz et al.’s experimental tracer study on Ti-doped samples. The lack of available experimental data renders comparison difficult. The calculated diffusion coefficient of the fastest diffusing species in Ti-doped alumina, $D_{\text{VAI}}$, lies within two orders of magnitude to the corresponding experimental data, but their activation energies, 1.6 and 3.9 ± 0.3eV, respectively, do not agree. The activation energies for other defects and doping environments can be found in Table 4.14. With the exception of Mg-doping, for almost all other doping levels and experimentally-attainable conditions, the diffusivity of aluminium exceeds that of oxygen, upholding
the long-held notion that $D_{\text{Al}} > D_{\text{O}}$, although the difference is not as great as previously believed. In addition, the calculated diffusion coefficients suggest the existence of a heretofore unconsidered phenomenon. As might have been predicted from the calculated defect concentrations in Fig. 4.42, aluminium transport is mediated by $V_A'^{\prime\prime\prime}$ defects in undoped and Ti-doped, and $\text{Al}_i'$ defects in Mg-doped crystals. This transition occurs because Mg-doping increases $[\text{Al}_i']$ to the point that it exceeds $[V_A'^{\prime\prime\prime}]$. Since $\text{Al}_i$ and $V_{\text{Al}}$ have similar migration barriers, $\text{Al}_i'$ defects become the dominant diffusing species in these situations. For undoped and Ti-doped crystals, $D_{\text{Al}_i}$ is significantly lower than $D_{V_{\text{Al}}}$.

Comparatively, across all doping regimes and temperatures, $V_{\text{AlO}}'$ divacancies contribute negligibly to diffusion due to their relatively high migration barriers and low concentrations. This evidence renders the notion that AlO complexes could mediate oxygen diffusion in alumina [47, 46] unlikely.

Errors in defect formation energies and migration barriers due to intrinsic DFT errors, the pseudopotential approximation, finite-size effects, and anharmonicity could accumulate to $O(1\text{eV})$, which is 3 orders of magnitude at 1750K. Errors in the prefactor of the diffusion coefficient, $D_{i,0}$, are roughly one order of magnitude (as $0 < f_i \leq 1$, the entire phonon spectrum only spans a factor of four in $\nu$, $\lambda_i$ is very well-defined, and the angles $\varphi_{i,j}$ also only contain small errors). These errors are significant, and, as seen in Sec. 4.5, prevent the current study from calculating a defect chemistry in alumina where the oxygen vacancy concentration is buffered upon alovalent doping. This insufficiency manifests itself in a DFT oxygen diffusion coefficient that changes by many orders of magnitude upon heavy doping, rather than the experimentally-determined one order of magnitude. In order to obtain accurate theoretical diffusion coefficients that can be reliably compared with experimental data, more work is recommended using higher level methods such as hybrid functionals or QMC. These may be able to determine cluster binding energies more accurately and thereby correctly account for the buffering phenomenon in alumina.

4.7 Impurity solubilities

MgO doping has a myriad of effects on alumina, such as reducing grain boundary mobility, which prevents discontinuous grain growth and allows pores to reside on the boundaries for extended periods [71]. Many experimental attempts to charac-
terise the mechanisms underlying these processes are performed at relatively low temperatures (≤ 1800°C) and high MgO doping levels (≥ 200ppm) that have been suggested to lie beyond the solubility limit of MgO \[71\]. Knowledge of solubility limits in alumina can aid experimentalists in finding adequate conditions to investigate the effects of doping on grain boundaries. Solubilities can be determined within the framework presented here, by comparing self-consistent calculations of dopant chemical potentials, such as $\mu_{\text{Mg}}$, in heavily-doped alumina with their values at the same $p_{\text{O}_2}$, $T$ in their respective oxides (i.e. MgO).

We determine the chemical potential of oxygen gas $\mu^{\text{Al}_2\text{O}_3}_{\frac{1}{2}\text{O}_2(g)}(p_{\text{O}_2}, T)$ at $p_{\text{O}_2}$, $T$ using the method of Finnis et al.\[58, 14\] by invoking a thermodynamic cycle corresponding to the formation of $\text{Al}_2\text{O}_3$ and applying ideal gas relations as

$$
\mu^{\text{Al}_2\text{O}_3}_{\frac{1}{2}\text{O}_2(g)}(p_{\text{O}_2}, T) = \frac{1}{3} \left\{ \mu_{\text{Al}_2\text{O}_3}(p_{\text{O}_2}, T^0) - 2\mu_{\text{Al}(s)}(p_{\text{O}_2}, T^0) \right. \\
- \Delta G^f_{\text{Al}_2\text{O}_3}(p_{\text{O}_2}, T^0) \right\} + \Delta \mu_{\frac{1}{2}\text{O}_2(g)}(p_{\text{O}_2}, T) + \frac{1}{2}k_B T \ln \left( \frac{p_{\text{O}_2}}{p^0} \right). 
$$

(4.19)

In Eq. 4.19, the chemical potentials of solid $\text{Al}_2\text{O}_3$ and fcc aluminium at standard conditions, $\mu_{\text{Al}_2\text{O}_3}(p_{\text{O}_2}^0, T^0)$ and $\mu_{\text{Al}(s)}(p_{\text{O}_2}^0, T^0)$, are approximated by DFT total energies at zero pressure and ideal stoichiometry, including vibrational contributions in the harmonic approximation. The standard Gibbs free energy of formation of alumina, $\Delta G^f_{\text{Al}_2\text{O}_3}(p_{\text{O}_2}^0, T^0)$, and the entropy of gaseous oxygen, $S_{\text{O}_2}^0$, are found from experiment, and the heat capacity at constant pressure is $C_p^0 = \frac{7}{2}k_B/2$. The superscript ‘$\text{Al}_2\text{O}_3$’ indicates that alumina is used for the thermodynamic cycle, but, in principle, the calculated value of the oxygen chemical potential at given $p_{\text{O}_2}$, $T$ will be the same when another oxide is used. In practice, small errors $O [0.2eV]$ arise from experimental uncertainties in $\Delta G^f$, the use of pseudopotentials, inadequate k-point sampling of the phonon density of states, anharmonicity, and the approximations within DFT.

The chemical potential of aluminium in $\text{Al}_2\text{O}_3$ is then found from the equilibrium condition of the solid as

$$
\mu_{\text{Al}}(p_{\text{O}_2}, T) = \frac{1}{2} \left\{ E_{\text{Al}_2\text{O}_3(s)}^{\text{tot}} + \Delta G_{\text{Al}_2\text{O}_3}(T) - 3\mu_{\frac{1}{2}\text{O}_2(g)}(p_{\text{O}_2}, T) \right\}, 
$$

(4.20)
neglecting the small \( pV \) contribution to the oxide and explicitly separating the DFT
zero-temperature total energy \( E_{\text{Al}_2\text{O}_3(s)}^{\text{tot}} \) and the vibrational contribution \( \Delta G_{\text{Al}_2\text{O}_3}(T) \)
in the harmonic approximation) to the Gibbs energy of solid \( \text{Al}_2\text{O}_3 \).

This method can be applied to the thermodynamic cycle of the formation of any
oxide, yielding the same result for \( \mu_{\frac{1}{2}\text{O}_2(g)}(p_{\text{O}_2}, T) \) if the same methods and approximations are made. Therefore, the oxygen chemical potential found from the
formation of \( \text{MgO} \) can be written as

\[
\mu_{\frac{1}{2}\text{O}_2(g)}(p_{\text{O}_2}, T) = \left\{ \mu_{\text{MgO}}(p_{\text{O}_2}^0, T^0) - \mu_{\text{Mg(s)}}(p_{\text{O}_2}^0, T^0) \\
- \Delta G_{\text{MgO}}^f(p_{\text{O}_2}^0, T^0) \right\} + \Delta \mu_{\frac{1}{2}\text{O}_2(g)}(p_{\text{O}_2}^0, T) + \frac{1}{2} k_B T \ln \left( \frac{p_{\text{O}_2}}{p_{\text{O}_2}^0} \right),
\]

where the symbols take their corresponding meanings, and, apart from the
small errors mentioned above, we find that

\[
\mu_{\frac{1}{2}\text{O}_2(g)}(p_{\text{O}_2}, T) \equiv \mu_{\text{Al}_2\text{O}_3}(p_{\text{O}_2}, T) \approx \mu_{\text{MgO}}(p_{\text{O}_2}, T).
\]

This gives the chemical potential of magnesium in \( \text{MgO} \) at \( p_{\text{O}_2}, T \) as

\[
\mu_{\text{Mg}}(p_{\text{O}_2}, T) = \left\{ E_{\text{MgO(s)}}^{\text{tot}} + \Delta G_{\text{MgO}}(T) - \mu_{\frac{1}{2}\text{O}_2(g)}(p_{\text{O}_2}, T) \right\},
\]

where the large influence of the vibrational contributions, \( \Delta G_{\text{MgO}}(1750K) \approx -1.20 \text{eV} \),
is noteworthy. The vibrational free energy for bulk \( \text{MgO} \) is determined using DFPT
in the harmonic approximation with TNDFPs in the LDA and a \( 4 \times 4 \times 4 \) Monkhorst-
Pack grid shifted to contain the \( \Gamma \) point, for the \( 4 \times 4 \times 4 \) (MgO) supercell. The
neglect of anharmonicity is likely to be non-negligible at these temperatures.

Considering the equilibrium between an oxygen gas atmosphere, solid \( \text{Al}_2\text{O}_3 \), and
solid \( \text{MgO} \), all at the same \( p_{\text{O}_2} \) and \( T \), it is possible to determine the solubility of
\( \text{MgO} \) in \( \text{Al}_2\text{O}_3 \). By choosing specific formation conditions \( p_{\text{O}_2} \) and \( T \), the oxygen
gas chemical potential \( \mu_{\frac{1}{2}\text{O}_2(g)}(p_{\text{O}_2}, T) \) and the chemical potentials of magnesium,
\( \mu_{\text{Mg}}(p_{\text{O}_2}, T) \), and, of course, aluminium, \( \mu_{\text{Al}}(p_{\text{O}_2}, T) \), in their respective oxides can
be calculated using Eqs. 4.19, 4.23, and 4.20. Given the chemical potentials of oxy-
gen and aluminium, the self-consistent calculations of the defect chemistry of \( \text{Al}_2\text{O}_3 \)
yield the concentrations of the defects and clusters considered as a function of in-
**Figure 4.47:** Predicted solubility of MgO (red solid line) compared with experimental data using scanning electron microscopy (Greskovich and Brewer [71]), wave-length-dispersive spectroscopy (Miller et al. [126]) and an electron microprobe (Ando and Momoda [6]). Common methods of measuring solubilities, such as x-ray diffractometry, are inconvenient for the low solubility of MgO.

incremental impurity concentration [Mg]. At each of these increments, the chemical potential of magnesium in alumina, $\mu_{\text{MgO}}(p_{\text{O}_2}, T)$, is adjusted until the total concentration of magnesium, whether free or bound to native defects, is equal to the specified concentration at that increment. Determining the full defect chemistry of doped Al$_2$O$_3$ thereby yields the chemical potential of Mg as a function of its concentration. The solubility of MgO at specified $p_{\text{O}_2}$ and $T$ in Al$_2$O$_3$ is then given by the concentration at which the Mg chemical potential in alumina matches the value in solid MgO, given from Eq. 4.23.

The calculated solubility is compared to sparse experimental data in Fig. 4.47. The solubility of MgO is overestimated by ca. 3 orders of magnitude, and it is difficult to compare trends between calculated and experimental data due to a paucity of experimental samples. To complicate matters, Greskovich and Brewer [71] performed their experiments on MgO solubilities in a hydrogen atmosphere, rendering direct comparison difficult.

Unfortunately, numerous errors plague these calculations, the largest of which arise from the self-consistent calculations of the concentrations of defects in alumina. This governs the relationship between the Mg chemical potential in alumina and the MgO solubility, and depends on the formation and binding energies of all defects and clusters included in the framework. These suffer, above all, DFT and finite-size errors. The latter we have already quantified as 0.4eV for some of the ill-
behaveled clusters in the study. For the former, a paper [106] comparing the defect formation energies of self interstitials in silicon in DFT to DMC found errors of 1.1-1.6eV, depending on the DFT functional chosen. Although, as will be shown in the following section, we find much smaller errors between DFT and DMC for the diffusion barriers of native defects in alumina, there is no reason to believe that errors for cluster formation energies in alumina will not be larger, due to the mixed bonding found in some of these structures. In addition, there are smaller errors in determining the chemical potentials in their respective oxides, as mentioned above. Cumulatively, we estimate that these errors amount to $O(1\text{eV})$, which is equivalent to roughly 3 orders of magnitude at 1750K, and can account for differences with experimental data. However, the results still suggest that determining solubility limits from self-consistent DFT calculations of defect energetics has little predictive power and can, at best, complement experiment.

4.8 Quantum Monte Carlo calculations

The term Quantum Monte Carlo (QMC) covers several highly accurate methods for determining the electronic structure of materials, two of which, variational Monte Carlo (VMC) and diffusion Monte Carlo (DMC), were described in Sec. 2.4. DMC can, in principle, determine the exact ground-state energy of a system, subject only to the fixed-node approximation. In this study, DMC calculations of diffusion barriers in alumina are used to benchmark the accuracy of the preceding DFT study. Previous comparisons between DFT and QMC have suggested that DFT overbinds solids [106, 13], but a hybrid functional has been able to reproduce DMC results [13] for defects in silicon. When dealing with defects, DFT suffers from the band gap error and, for defects carrying a highly localised charge, introduces self-interaction errors. The only previous study of defect formation energies using QMC in alumina agreed better than expected with DFT results [81]. This somewhat surprising result bodes well for the reliability of DFT formation energies and thus defect concentrations. Although the existing QMC calculations did not account for the aluminium split vacancy and oxygen dumbbell interstitial structures, this should not affect the outcome drastically. Although it would be optimal to calculate all \textit{ab initio} quantities in this study using DMC, computational demands render this impractical. In addition, the lack of a
suitable pseudopotential containing the semi-core $3s$ and $3p$ states that participate significantly in bonding [186], for titanium prohibits calculations of cluster binding energies containing titanium. In the future, perhaps, these problems will be overcome and thorough QMC benchmarks of current DFT work will be obtained, but in the meantime, it is still possible to gauge the capability of DFT to calculate migration barriers, as this involves only energy differences between two different states of the same defect.

### 4.8.1 Trial wave functions

In order to balance accuracy and efficiency, the trial wave functions generated for the QMC calculations in this study derive from highly-converged plane wave DFT calculations.

Unfortunately, ultrasoft pseudopotentials are incompatible with current QMC implementations, necessitating the use of considerably harder, norm-conserving pseudopotentials. This study uses the Trail-Needs pseudopotentials presented in Sec. 4.4.1, as these are specifically generated for accurate QMC calculations. In addition, the herculean computational cost of DMC renders defect calculations in large supercells unfeasible. Fortunately, the $2 \times 2 \times 1$ supercell containing 576 electrons (for perfect alumina) used for the majority of the DFT studies presented here, with the exception of calculations for finite-size scaling, is tractable in QMC. The DFT ground-state structures of the native point defects and the geometries at the energy maxima along their migration paths are determined in the $2 \times 2 \times 1$ supercell at a plane wave cut-off of 3000eV in the LDA. This yields a well-converged basis of one electron orbitals from which the determinantal parts of the QMC trial functions are constructed. These DFT structures are then used for the subsequent VMC and DMC calculations.

As a plane wave basis set is unnecessarily inefficient for QMC calculations, the plane wave DFT orbitals are converted into a ‘blip’ representation [2]. Blip functions are localised cubic splines centred at points on a regular grid. Orbitals are represented as linear combinations of blip functions at each grid point, so that the cost to evaluate an orbital at any single point is fixed, regardless of system size. In contrast, the cost of evaluating the plane-wave representation of an orbital at a given point scales as $L^3$, where $L$ is the linear dimension of the simulation cell.
4.8.2 Jastrow optimisation

In the current QMC calculations, electron correlation is introduced into the trial wave function via a Jastrow factor that multiplies the single-particle orbitals obtained from DFT. Not only can a ‘good’ Jastrow factor recover a significant portion of the correlation energy, it can also help a trial wave function mimic the true ground-state wave function, and thereby lower the variance and increase the efficiency of a DMC calculation. Although the presence of a Jastrow factor changes the VMC energy, it does not affect the final DMC energy as the nodes of the trial wave function remain fixed.

When choosing which terms to include in a Jastrow factor, it is crucial to strike a balance between accuracy and computational tractability. The more suited the form is to describing the system in question, by offering sufficient variational freedom, the more effectively the energy can be minimised. However, as more terms and parameters are included, it becomes more arduous to optimise the Jastrow factor. This study employs a Jastrow factor comprising one-, two-, and three-body $\chi$, $u$, and $f$ terms, as described in Sec. 2.4.3. Each of these terms is represented by a power series expansion containing optimisable coefficients, the form of which guarantees that the function approaches zero smoothly at an optimisable cut-off. The polynomial representation of the $u$ function is also constrained to ensure that the cusp conditions discussed in Sec. 2.4.3 are obeyed. The free parameters were optimised using the accelerated variance minimisation scheme of Ref. [49]. The $u$ terms are functions of electron-electron separation $r_{ij}$, expanded up to $r_{ij}^8$, with separate parameters for $u_{\uparrow\uparrow}$ and $u_{\uparrow\downarrow}$. This gives a total of eight optimisable variables, namely seven free parameters and one cut-off length for each of the $u_{\uparrow\uparrow}$ and $u_{\uparrow\downarrow}$. The $\chi$ terms are functions of electron-ion distance $r_{iI}$, and are expanded similarly to the $u$ terms. Naturally, different $\chi$ terms are chosen for the two atomic species, aluminium and oxygen. In addition, as the electronic neighbourhood surrounding a defect can be very inhomogeneous, separate $\chi$ terms are used for all aluminium and oxygen nearest neighbours to a defect site, both at its equilibrium position and along its migration path. Atoms comprising defects and migrating atoms are also treated separately. This freedom allows the nature of the electron correlation to change in the vicinity of a defect. Also, for the oxygen vacancy, a $\chi$ term centred on the vacant site is included, allowing optimisation of the free charge surrounding the site. Due to computational demands, the same could not done for the aluminium split...
vacancy, as this contains two separate vacancy sites and quite a large number of ‘special’ atoms that also require their own $\chi$ terms. The $f$ terms are functions of $r_{ij}$, $r_{ii}$, and $r_{jj}$, expanded to second order in each separation. These terms contain eight free parameters, in addition to the cut-off length, for optimisation.

The optimisation of the $\chi$ terms for the aluminium split vacancy defect is very unstable. Separate $\chi$ terms are used for the split vacancy aluminium atom that sits in the interstitial site, the nearest-neighbour aluminium atom that migrates into one of the split vacancies, the 2NN oxygen atoms surrounding the defect site, and the 3NN aluminium atoms (as both 1NN aluminium atoms have vacated their sites). The $\chi$ terms at the split vacancy and migrating aluminium atom sites are negative for the ground-state structure of the defect. At the height of the migration barrier, the $\chi$ term on the migrating aluminium atom changes sign, suggesting that the electron density differs strongly from that in the ground state. This renders the optimisation of the $\chi$ very difficult. Changes on the other sites, although smaller, are still significant, such as a decrease in the value of $\chi$ on the aluminium 3NN during the migration.

The goal of the optimised Jastrow factor is to provide an accurate description of electron correlation. The fraction of the correlation energy recovered by the Jastrow factor can be estimated as

$$E_{i}^{\text{corr}} = \frac{E_{i}^{\text{VMC}} - E_{i}^{\text{HF}}}{E_{i}^{\text{DMC}} - E_{i}^{\text{HF}}},$$

where $E_{i}^{\text{VMC}}$ and $E_{i}^{\text{DMC}}$ are the VMC and DMC total energies found when the Jastrow factor is included in the calculation, and $E_{i}^{\text{HF}}$ is found from a VMC calculation omitting the Jastrow factor and using only the Slater determinant derived from the Kohn-Sham DFT orbitals. The latter is actually the DFT exact-exchange energy since the orbitals are not calculated from Hartree-Fock theory. As this is slightly higher than the true Hartree-Fock energy, the correlation energy as defined here is slightly larger (more negative) than that defined relative to the true Hartree-Fock energy. The major assumption underlying Eq. 4.24 is that $E_{i}^{\text{DMC}}$ is the correct ground-state energy, and thereby recovers $\approx 100\%$ of the correlation energy.

The correlation energies recovered by the Slater-Jastrow trial wave functions for supercells containing bulk alumina and various defects are listed in Table 4.15. The Jastrow factors perform worse in defect calculations, due to the inhomogeneous electron density surrounding the defect. In order to recover a larger portion of the
Chapter 4: Point defects and diffusion in Al$_2$O$_3$

<table>
<thead>
<tr>
<th>Species</th>
<th>$E_{i}^{\text{corr}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>bulk</td>
<td>91%</td>
</tr>
<tr>
<td>$V_{\text{O}}$</td>
<td>87%</td>
</tr>
<tr>
<td>$V_{\text{Al}}'''$</td>
<td>76%</td>
</tr>
<tr>
<td>$O_{i}''$</td>
<td>79%</td>
</tr>
<tr>
<td>Al$_{i}^-$</td>
<td>86%</td>
</tr>
</tbody>
</table>

Table 4.15: The percentage of correlation energy retrieved by the inclusion of a Jastrow factor in the trial wave function, calculated from Eq. 4.24. The Jastrow factors are less successful for defect calculations due to the inhomogeneous electron density surrounding the defects and the limited variational freedom.

correlation energy, a number of separate $\chi$ terms for neighbours surrounding the defect are necessary, and possibly also a spatially inhomogeneous $u$ term. However, these inclusions would render the variance minimisation intractable, suggesting that defect calculations using simple VMC may be a dead-end in terms of computational cost versus achievable accuracy. As the accuracy of a DMC simulation is only limited by the nodes of the trial wave function, DMC seems the more obvious candidate for benchmarking DFT calculations involving defects.

4.8.3 QMC band gaps

One of the major advantages of QMC methods over DFT is their ability to accurately describe the energies of excited states. This enables DMC to reproduce the experimental band gap in alumina, which, as described in Ref. [81], can have an impact on the formation energies of defects in high charge states.

There are two primary ways to calculate the band gap using QMC methods. First, as in DFT, the gap can be determined as

$$ E_0^{\text{gap}} = E_0^{\text{tot}} (N - 1) + E_0^{\text{tot}} (N + 1) - 2E_0^{\text{tot}} (N), $$

(4.25)

where $E_0^{\text{tot}} (N)$ is the total energy of a supercell of perfect material containing $N$ electrons, and $E_0^{\text{tot}} (N \pm 1)$ the corresponding total energies of supercells containing $N + 1$ and $N - 1$ electrons. This method contains significant finite-size errors due to the added or removed electron and the need to include a neutralising background charge. In QMC calculations, where the electron positions are fixed explicitly at each time step, the interaction with the background incurs an error of $\frac{1}{2}q^2v_M$ [81].
### Section 4.8: Quantum Monte Carlo calculations

**Charged supercells**

<table>
<thead>
<tr>
<th>Method</th>
<th>Charged supercells</th>
<th>Excitation</th>
</tr>
</thead>
<tbody>
<tr>
<td>DFT (LDA-USP)</td>
<td>6.83</td>
<td>6.88</td>
</tr>
<tr>
<td>DMC</td>
<td>10.6 ± 0.5</td>
<td>9.4 ± 0.4</td>
</tr>
<tr>
<td>Experiment [64]</td>
<td>9.57</td>
<td>-</td>
</tr>
</tbody>
</table>

**Table 4.16:** The band gap of alumina (in eV) calculated using the two methods described in Sec. 4.8.3 in DFT and DMC, compared with a recent experimental determination. DFT suffers from the band gap error discussed in Sec. 3.5.2, while the QMC values are considerably more accurate. The errors on the DMC values refer to statistical uncertainty.

This must be corrected when calculating defect formation energies [81] or band gaps using Eq. 4.25 in QMC, but cancels in calculations of migration barriers.

The second method estimates the band gap by calculating the energy of an exciton created by promoting an electron from the highest-occupied state to the lowest unoccupied state in the Slater determinant,

\[
E_{0}^{\text{gap}} \simeq E_{0}^{\text{tot}}(\text{ex}) - E_{0}^{\text{tot}}(\text{gs}),
\]

where \( E_{0}^{\text{tot}}(\text{ex}) \) and \( E_{0}^{\text{tot}}(\text{gs}) \) are the total energies of supercells of perfect material in its first excited state and ground-state, respectively. Although Eq. 4.26 neglects the contribution from the exciton binding energy (experimentally found to be 0.13eV [64]), this omission is smaller than the Coulomb errors arising in the first method.

The QMC band gaps calculated using both methods are compared with their DFT counterparts and the best available experimental data in Table 4.16. For the DFT values of excitation gaps, differences between the eigenvalues at the bottom of the conduction band and the top of the valence band are used. The DFT values all suffer from the band gap error discussed in Sec. 3.5.2, underestimating the gap by almost 3eV. The DMC results are much better, differing by less than half an eV when using the excitation method, and just over 1eV when using differences between charged supercells.

#### 4.8.4 Errors

Although DMC is, in principle, an exact method, certain approximations are made to render its application computationally practical. The most important of these are time step errors, the pseudopotential localisation error, and the fixed node approximation. In addition, calculating DMC migration barriers involves comparing DMC...
total energies of supercells whose geometries derive from DFT, implicitly assuming that the DFT geometries are good.

**Transferability**

In order to examine the validity of the latter assumption, it is necessary to verify that the optimised geometry does not differ greatly between DFT and DMC. As described in Sec. 4.4.1, the corundum structure has four independent geometric parameters that can be optimised. Unfortunately, there are no methods available to accurately determine forces in large, periodic systems in order to fully optimise the geometry using QMC methods. Even if there were, such calculations would not be feasible for alumina due to the large number of parameters and optimisation steps necessary to find the energy minimum. In order to quantify the error associated with transferring geometries from DFT to DMC, we can examine the effect of varying a single geometric parameter on the total energy. Fig. 4.48(a) shows the change in total energy of a formula unit of alumina (taken from the $2 \times 2 \times 2$ rhombohedral supercell) in DFT and DMC when the parameter $a_{\text{rhom}}$ is varied to change the cell size at constant shape. In DFT, small changes in $a_{\text{rhom}}$ are found to have a larger effect on the total energy, and a smaller effect on other lattice parameters, than the other geometric parameters. For each fixed value of $a_{\text{rhom}}$, the rest of the geometry is reoptimised in DFT. For the DMC simulations, the Jastrow factor is reoptimised for each of these geometries. The $x$-axis in Fig. 4.48(a) expresses the change in the volume of the primitive cell for the chosen values of $a_{\text{rhom}}$. The minimum for the DFT curve, where $V/V_0^{\text{DFT}} = 1$, corresponds to the optimised geometry in Sec. 4.4.1. The minimum for the DMC curve corresponds to $V/V_0^{\text{DFT}} = 0.96$, suggesting that the DFT equilibrium volume is slightly too large. This result agrees well with experimental lattice parameters, that yield an experimental equilibrium volume $V^{\text{EXP}}$ such that $V^{\text{EXP}}/V_0^{\text{DFT}} = 0.94$. Although the agreement is not perfect, the results shown in Fig. 4.48(a) demonstrate that the equilibrium lattice parameters do not differ significantly between DFT and DMC.

The error in transferability is expected to be more pronounced for defect supercells, especially when the atomic positions correspond to the maximum of the migration barrier. Unfortunately, computational limitations prohibit a detailed analysis of this scenario, as the supercells required for these calculations are very large (at least the $2 \times 2 \times 1$ hexagonal cell), and VMC and DMC simulations of such large systems are
extremely costly.

Additionally, our previous calculations [81] comparing DFT, DMC, and experimental cohesive and atomisation energies of bulk alumina, its atomic constituents in their standard states, and the AlO dimer, found that DMC vastly outperforms DFT. The largest discrepancies were, unsurprisingly, found for the oxygen dimer, underscoring the importance of avoiding its explicit calculation in DFT, as alluded to in Sec.3.4.2.

**Time step bias**

As DMC is only exact in the limit of an infinitesimally small time step $\Delta \tau$, using finite values for $\Delta \tau$ incurs a time step error. This error can be systematically examined and corrected for. However, if the chosen time step is already short enough that its associated error is small compared with the statistical error of the simulation, time-step errors can be safely neglected. Note that simulations using a shorter time step require more steps to reach the next statistically-independent sampling point; this increase in the correlation time manifests itself in added computational time. Both the number of steps necessary for decorrelation and the computational time required for a certain error bar scale as $1/\Delta \tau$, so that a balance needs to be found between accuracy and efficiency.
Chapter 4: Point defects and diffusion in Al$_2$O$_3$

Fig. 4.48(b) shows the total energy per formula unit of the $2 \times 2 \times 2$ rhombohedral supercell of perfect Al$_2$O$_3$ as a function of $\Delta \tau$. The time step error, given by 
$$
\lim_{\Delta \tau \to 0} [E_{\text{tot}}^{\text{tot}}(\Delta \tau)] - E_{\text{tot}}^{\text{tot}}(\Delta \tau),
$$
converges very quickly for small time steps. The error for $\Delta \tau = 0.005 \text{a.u.}$, the time step used to calculate migration barriers, is less than 0.005eV/atom, which is much smaller than the statistical error achieved during the long DMC runs in this study (ca. 0.05eV/atom). As there is no reason to believe that time step errors differ greatly between the stable and migration-barrier configurations of defect supercells, they can safely be neglected.

4.8.5 Diffusion barriers

In the DFT studies of diffusion barriers in Sec. 4.4.8, the total energy of the system as the migrating atoms of a defect were moved incrementally along a migration vector was compared with the ground-state total energy of the defect to create an energy profile describing the migration. Analogous calculations are not computationally feasible in DMC. However, the energy profile is unlikely to differ significantly between DFT and DMC, and calculations of the heights of migration barriers are feasible.

To determine a migration barrier, two QMC trial wave functions were generated as described in Sec. 4.8.1 from the DFT orbitals of 3000eV TNDFP LDA calculations. One of the two trial wave functions was for the ground-state of the defect and the other for the geometry at the top of the DFT migration barrier. It is reassuring that the results of the TNDFP DFT calculations differed very little from those of the Vanderbilt USP DFT calculations presented in Sec. 4.4.8, with the DFT migration barriers agreeing to within 0.05eV for all defects. Electron correlation was introduced by multiplying the Slater determinant of single-particle orbitals with a Jastrow factor constructed and optimised as described in Sec. 4.8.2. In optimising the Jastrow factor, VMC studies were performed. The VMC migration barriers were several eV too large, highlighting the difficulties in describing electron correlation using VMC. Therefore, VMC was only used to generate the trial wave functions for the ensuing DMC simulations.

The migration barriers calculated using DMC and DFT are compared in Table 4.8.5. The DFT values stem from TNDFP LDA calculations in the $2 \times 2 \times 1$ supercell, and so are not directly comparable with the values presented in Sec. 4.4.8, which were extrapolated to the infinite-cell limit. Similar finite-size corrections could not be
The results show agreement between DFT and DMC to within the statistical error bars of the simulations. This demonstrates the ability of DFT to correctly describe the mixed bonding environments encountered when defects migrate through $\text{Al}_2\text{O}_3$. Generally, DMC produces slightly lower migration barriers for vacancies, and slightly higher barriers for interstitials, in line with the expectation that DFT tends to overbind the solid [81]. However, the differences in barrier energies are so small, and those in computational time so great, that DMC will be restricted to a supporting role when studying defect migration in systems such as alumina.
Conclusion

This thesis presents a formalism for determining the concentrations and diffusion coefficients of defects in metal oxides, applied to the archetypal metal oxide, alumina. Despite a plethora of studies, both experimental and theoretical, many aspects of diffusion in alumina remain unexplained. In particular, experimental results from a wide range of studies show that the oxygen diffusion coefficient is relatively insensitive to aliovalent doping. This insensitivity has been suggested to derive from strongly buffered defect populations that change little even when a crystal is heavily doped. By employing ab initio calculations of defect formation energies, cluster binding energies, and migration barriers from density functional theory (DFT), this work examines this hypothesis.

In order to correctly describe defects and diffusion in any system, it is imperative that the ground-state structures of the defects and their diffusion mechanisms are known. This work demonstrates that many preconceptions of the native point defects in alumina and their migration are incorrect by unearthing new ground-state structures and diffusion mechanisms for several defects. Aluminium vacancies are found to form a split vacancy defect where a next nearest-neighbour aluminium atom along the c-axis divides the vacancy into two parts by relaxing into the normally-vacant interstitial site, while oxygen interstitials form a dumbbell interstitial together with a lattice oxygen atom, centred on a normal oxygen lattice site. These two ground-state configurations are significantly more stable than previously used structures. Novel diffusion mechanisms are found for both aluminium vacancies and interstitials, as well as oxygen interstitials. The first two migrate by concerted vacancy and interstitialcy mechanisms, respectively, while oxygen dumbbell interstitials migrate by a dumbbell interstitialcy mechanism. In addition, the diffusion barriers for all
four defects are found to be significantly lower than previously assumed. These results are demonstrated to be unobtainable using simple classical potential models, requiring knowledge of the electronic structure of the system. To compound this, it is shown to be imperative that the defect formation energies, cluster binding energies, and migration barriers are determined to a high level of accuracy. In particular, chemical potentials determined using the Finnis, Alavi, and Lozovoi method [58] must be corrected to include vibrational effects, as the neglect of these contributions lumps entirely in the metal chemical potential, and does not cancel out as previously assumed. Finite-size corrections must be applied both to defect formation energies and to migration barriers. Their incorporation is shown to drastically reduce the migration barriers of defects that place large strains on specific crystal axes in insufficiently-sized supercells. Finally, the accuracy of the migration barriers calculated in DFT is benchmarked with diffusion Monte Carlo (DMC), the most accurate method available for describing solids. The two methods are shown to agree to within the statistical error bar of DMC.

By incorporating defect formation energies and cluster binding energies calculated in DFT within a mass action framework, the concentrations of defects and defect-dopant clusters as a function of doping are examined. Combining these defect concentrations with calculated migration barriers of the four native point defects in alumina yields theoretical predictions of diffusion coefficients in alumina. The results from this study are unable to demonstrate the insensitivity of the oxygen diffusivity to aliovalent doping, but suggest that this could be caused by inadequacies in the current methodology, which can be overcome using more accurate approaches. In this study, it is found that the concentrations of charge-compensating defects, primarily aluminium and oxygen vacancies in alumina, change by many orders of magnitude upon heavy aliovalent doping. This causes the oxygen diffusion coefficient to equally change by many orders of magnitude, which is inconsistent with experimental findings. Agreement with experiment can be found by arbitrarily raising the binding energies of the dominant defect-dopant clusters by 1eV. Then, the buffering of defect concentrations in alumina can be shown to arise from the formation of tightly-bound defect-dopant clusters. Unfortunately, without this arbitrary change, the concentrations of these clusters calculated in the LDA are too low to create this phenomenon. These results suggest that two further pieces of research are needed to determine whether or not defect-dopant clustering is the cause of the buffering effect in alumina. First, it is necessary to examine whether other
clusters not considered in the current study, or those considered here but in different ground-state structures, exist in high concentrations. This can be done using a technique such as metadynamics with an accurate classical potential derived from DFT. Second, calculations of defect formation energies and cluster binding energies using more accurate methods, such as hybrid functionals or quantum Monte Carlo, can show whether errors intrinsic to the LDA have prevented the current study from determining accurate defect concentrations.

As the formalism presented in Chapter 3 is widely applicable to a raft of metal oxides, further work applying the same approach to other technologically ubiquitous materials, such as TiO$_2$, ZrO$_2$, or CeO$_2$, is encouraged. For alumina, future developments in complex classical potentials derived from ab initio studies can be measured against results found in this study. Modern potentials should be able to reproduce the correct ground-state structures of the native defects in alumina, not just the crystal structure. Once such potentials have been generated, they can be used to determine the lowest-energy migration paths in alumina to a higher level of accuracy by sampling the energy landscape using metadynamics, or similar methods. Furthermore, grain boundary diffusion in alumina unveils a raft of questions that theory has yet to answer.

One leitmotif that permeates the results of this thesis is the importance of interdisciplinary work. Not only must experimental and theoretical physicists collaborate, but so should scientists and engineers from different backgrounds and fields of research. This collaboration brings together vital expertise that would otherwise be dearly missed, and symmetrically shares the wealth of knowledge that we have built. In the end, only if we can mirror nature’s pluralism can we begin to understand and marvel at its beauty.
A.1 Arrhenius relationships of activation energies of diffusion

The activation enthalpy for diffusion measured in experiments is obtained from the slope of an Arrhenius plot of \( \ln [D_i (p, T)] \) against \( 1/T \), where we define the diffusion coefficient \( D_i (p, T) \) as

\[
D_i (p, T) = D_{i,0} \exp \left( -\frac{Q_{\text{act}}^i (p, T)}{k_B T} \right).
\] (A.1)

Strictly, the activation energy \( Q_{\text{act}}^i (p, T) \) in Eq. A.1 is the sum of the Gibbs energies of formation \( \Delta G_f^i (p, T) \) and migration \( \Delta G_{\text{mig}}^i (p, T) \): \( Q_{\text{act}}^i (p, T) = \Delta G_f^i (p, T) + \Delta G_{\text{mig}}^i (p, T) \).  

To show that the slope of the Arrhenius plot of \( \ln [D_i (p, T)] \) against \( 1/T \) yields an activation enthalpy, rather than a Gibbs energy, we begin by recognising that \( Q_{\text{act}}^i (p, T) \) is not independent of temperature, so that the slope

\[
\frac{d \ln [D_i (p, T)]}{d \left( \frac{1}{T} \right)}
\] (A.2)

is not simply \( Q_{\text{act}}^i (p, T) \). The gradient in Eq. A.2 can be expanded (simplifying the notation for clarity) as

\[1\text{We approximate } \Delta G_{\text{mig}}^i (p, T) \text{ as } \Delta E_{\text{mig}}^i \text{ by omitting vibrational contributions, but this does not affect the current derivation.}\]
\[ \frac{d \ln [D_i]}{d \left( \frac{1}{T} \right)} = - \frac{1}{k_B} \frac{d \left( \frac{Q^{\text{act}}_i}{T} \right)}{d \left( \frac{1}{T} \right)} = \frac{T^2}{k_B} \frac{d \left( \frac{Q^{\text{act}}_i}{T} \right)}{dT}. \]  

(A.3)

Noting that \( Q^{\text{act}}_i = H^{\text{act}}_i - T S^{\text{act}}_i \), we have that

\[ d \left( \frac{Q^{\text{act}}_i}{T} \right) = d \left( \frac{H^{\text{act}}_i - T S^{\text{act}}_i}{T} \right) = \frac{d H^{\text{act}}_i}{T} + H^{\text{act}}_i d \left( \frac{1}{T} \right) - d S^{\text{act}}_i. \]  

(A.4)

Also, as \( d H^{\text{act}}_i = T d S^{\text{act}}_i + V d P \) and we are working at constant pressure, \( \frac{d H^{\text{act}}_i}{T} = d S^{\text{act}}_i \), so that Eq. A.4 becomes

\[ d \left( \frac{Q^{\text{act}}_i}{T} \right) = H^{\text{act}}_i d \left( \frac{1}{T} \right) = - \frac{H^{\text{act}}_i}{T^2} dT. \]  

(A.5)

Substituting the result from Eq. A.5 into the original equation for the slope of the Arrhenius plot in Eq. A.3 gives

\[ \frac{d \ln [D_i]}{d \left( \frac{1}{T} \right)} = - \frac{H^{\text{act}}_i}{k_B}, \]  

(A.6)

so that the measured slope of the Arrhenius plot of \( \ln [D_i (p, T)] \) against \( 1/T \) yields the activation enthalpy \( H^{\text{act}}_i \), rather than the Gibbs energy of activation \( Q^{\text{act}}_i \).

**A.2 Einstein equation from random walk diffusion**

A general kinetic expression for \( D_i \), called an Einstein equation, can be derived from a density function \( W_i (X, \tau) \) that gives the probability that a defect at position \( x \) at time \( t \) will be at \( x + X \) at time \( t + \tau \). In a homogeneous system with no driving forces, at time \( \tau \), each diffusing species \( i \) must be at some position \( X \), such that

\[ \sum_{\text{all } X} W_i (X, \tau) = 1 \]  

(A.7)

with moments

\[ \sum_{\text{all } X} X^m W_i (X, \tau) = \langle X^m \rangle , \]  

(A.8)
Section A.3: Diffusion coefficients from random walk diffusion

where \( \langle \ldots \rangle \) denotes an average of a large number of diffusing species. We can write the concentration \( c_i(x, t + \tau) \) of the diffusing species as a function of \( x \) at time \( t + \tau \) as

\[
c_i(x, t + \tau) = \sum_{all X} c_i(x - X, t) W_i(X, \tau), \tag{A.9}
\]

where \( c_i(x - X, t) \) is the concentration at an earlier time \( t \) and the sum runs over all possible values of \( X \). The rate at which the concentration is changing can be found by expanding \( c_i(x, t + \tau) \) and \( c_i(x - X, t) \) in a Taylor series around \( X = 0 \) and \( \tau = 0 \):

\[
c_i(x, t) + \tau \frac{\partial c_i}{\partial t} = \sum_{all X} \left\{ c_i(x, t) - X \frac{\partial c_i}{\partial x} + \frac{X^2}{2} \frac{\partial^2 c_i}{\partial x^2} \right\} W_i(X, \tau),
\]

where the higher-order terms in \( \tau \) and \( X \) have been truncated on the assumption that \( W_i(X, \tau) \) becomes increasingly localised near \( X = 0 \) as \( \tau \to 0 \). Simplifying with Eq. A.8 leaves

\[
\frac{\partial c_i}{\partial t} = -\frac{\langle X \rangle}{\tau} \frac{\partial c_i}{\partial x} + \frac{\langle X^2 \rangle}{2\tau} \frac{\partial^2 c_i}{\partial x^2}, \tag{A.10}
\]

and hence, since \( \langle X \rangle = 0 \) for a simple random walk,

\[
\frac{\partial c_i}{\partial t} = \frac{\langle X^2 \rangle}{2\tau} \frac{\partial^2 c_i}{\partial x^2}, \tag{A.11}
\]

which has the same form as Fick’s second law in Eq. 3.63. The diffusion coefficient \( D_i \) is then given by the Einstein equation [50]

\[
D_i = \lim_{\tau \to 0} \frac{\langle X^2 \rangle}{2\tau}. \tag{A.12}
\]

A.3 Diffusion coefficients from random walk diffusion

The Einstein equation in Eq. A.12 can be evaluated using kinetic theory to give an expression for the diffusion coefficient \( D_i \) for one-dimensional diffusion with a concentration gradient only along the \( x \)-axis. The total displacement \( X \) along a
APPENDIX A: Point defects in solids

given diffusion direction (here taken to be the $x$-axis) during a finite time $\tau$ from a series of $n$ diffusion jumps each of length $x_i$ each is

\[ X = x_1 + x_2 + \ldots + x_n, \]  

(A.13)
such that the squared displacement $X^2$ can be written as

\[
X^2 = (x_1 + x_2 + \ldots + x_n)^2 \\
= \sum_{\gamma=1}^{n} x_{\gamma}^2 + 2 \sum_{\gamma=1}^{n-1} \sum_{\gamma+1}^{n} x_\gamma x_\delta.
\]  

(A.14)

The mean squared displacement $\langle X^2 \rangle$ is found by averaging $X^2$ over a large number of jump series $X$. Each single jump $x_i$ within a series is independent of all others, so that it is equally likely to be in the positive or negative diffusion direction. This means that the $x_\gamma x_\delta$ cross-terms in Eq. A.14 cancel in the limit of a large number of jumps, leaving

\[
\langle X^2 \rangle = \sum_{\gamma=1}^{n} \langle x_{\gamma}^2 \rangle.
\]  

(A.15)

Assuming that all jumps to the $z$ nearest-neighbours with the lowest diffusion barrier are equally likely, and jumps to other sites very unlikely, Eq. A.15 simplifies to

\[
\langle X^2 \rangle = \left( \sum_{j=1}^{z} n_j x_j^2 \right) = \sum_{j=1}^{z} x_j^2 \langle n_j \rangle,
\]  

(A.16)

where $n_j$ is the total number of jumps in the direction of the $j$th nearest-neighbour and $x_j$ is the displacement along the diffusion direction $x$ for each jump. The total jump frequency of the defect $i$ to the $j$th nearest-neighbour, $\Gamma_{i,j}$, is then given by

\[
\Gamma_{i,j} = \frac{\langle n_j \rangle}{\tau},
\]  

(A.17)

so that

\[
\langle X^2 \rangle = \tau \sum_{j=1}^{z} \Gamma_{i,j} x_j^2.
\]  

(A.18)

Substituting the result from Eq. A.18 into the Einstein relation in Eq. A.12, we find
that

\[ D_i = \frac{1}{2} \sum_{j=1}^{z} \Gamma_{i,j} x_j^2. \]  

(A.19)

### A.4 Correlation factors for native defect diffusion in alumina

In general, both the diffusion coefficient \( D_i^{\text{actual}} \) and correlation factor \( f_i \) for a defect \( i \) are tensor quantities. In cubic crystals, however, they are isotropic; and, in hexagonal and tetragonal crystals, they are the same for every direction in the basal plane [116]. In the dilute limit, the correlation factor is given by

\[
 f = \lim_{\tau \rightarrow \text{small}} \frac{\langle X^2 \rangle_{\text{actual}}}{\langle X^2 \rangle_{\text{random}}},
\]  

(A.20)

where \( \langle X^2 \rangle \) is the mean squared displacement from a series of \( n \) jumps, taking place over time \( \tau \) and we have omitted the defect subscripts \( i \) for clarity. The restriction on \( \tau \) holds true whether or not driving forces are present and is described in depth in Ref. [116]. The displacement \( X \) is then comprised of individual jumps

\[
 X = x_1 + x_2 + \ldots + x_n,
\]  

(A.21)

and it follows that

\[
 X^2 = \sum_{\gamma=1}^{n} x_{\gamma}^2 + 2 \sum_{\gamma=1}^{n-1} \sum_{\delta=\gamma+1}^{n} x_{\gamma} x_{\delta}.
\]  

(A.22)

When averaging \( X^2 \) over many series of jumps, the diagonal and cross-terms in Eq. A.22 are independent of each other, so that

\[
 \langle X^2 \rangle = \sum_{\gamma=1}^{n} \langle x_{\gamma}^2 \rangle + 2 \sum_{\gamma=1}^{n-1} \sum_{\delta=\gamma+1}^{n} \langle x_{\gamma} x_{\delta} \rangle.
\]  

(A.23)

For a truly random walk, the contribution from the summations over \( \langle x_{\gamma} x_{\delta} \rangle \) is zero as all jumps are independent and have equal probability of being in the positive or negative direction. When correlations are present, we obtain

\[
 f = 1 + \lim_{n \rightarrow \infty} \left( 2 \sum_{\gamma=1}^{n-1} \sum_{\delta=\gamma+1}^{n} \langle x_{\gamma} x_{\delta} \rangle / \sum_{\gamma=1}^{n} \langle x_{\gamma}^2 \rangle \right),
\]  

(A.24)

where the limit \( n \rightarrow \infty \) accounts for all correlations, including those between widely separated jumps where \( \delta \gg \gamma \).
Appendix A: Point defects in solids

The value of $f$ depends on the diffusion mechanism and the geometry of the structure, and the following derivations are only valid in the dilute limit, when the diffusion is not near impurity atoms.

For vacancy diffusion in alumina, all jumps corresponding to the lowest migration barrier have the same length, allowing us to set $\langle x_\gamma x_\delta \rangle = \langle x_\gamma^2 \cos \theta_{\gamma \delta} \rangle$, where $\theta_{\gamma \delta}$ is the angle between the $\gamma$th and $\delta$th jumps caused by the migration of a single vacancy. As $\sum_{\gamma=1}^n \langle x_\gamma^2 \rangle = n \langle x_\gamma^2 \rangle$, Eq. A.24 becomes

$$f = 1 + \lim_{n \to \infty} \frac{2}{n} \sum_{\gamma=1}^{n-1} \sum_{\delta=\gamma+1}^n \langle \cos \theta_{\gamma \delta} \rangle. \quad (A.25)$$

Expanding, Eq. A.25 becomes

$$f = 1 + \lim_{n \to \infty} \frac{2}{n} \sum_{\gamma=1}^{n-1} \langle \cos \theta_{\gamma,\gamma+1} + \cos \theta_{\gamma+1,\gamma+2} + \ldots + \cos \theta_{\gamma,n} \rangle, \quad (A.26)$$

and using the spherical cosines rule, we have that

$$\cos \theta_{\gamma,\gamma+2} = \cos \theta_{\gamma,\gamma+1} \cos \theta_{\gamma+1,\gamma+2} + \sin \theta_{\gamma,\gamma+1} \sin \theta_{\gamma+1,\gamma+2} \cos \Theta, \quad (A.27)$$

where $\Theta$ is the angle between plane $A$, spanned by the vectors corresponding to jumps $\gamma$ and $\gamma+1$, and plane $B$, spanned by the vectors of jumps $\gamma+1$ and $\gamma+2$. The average of the second term on the RHS of Eq. A.27 is proportional to $\sum_{\Theta} P_\Theta \cos \Theta$, where $P_\Theta$ is the jump probability for a certain $\Theta$, and equals zero for crystals with at least two-fold rotational symmetry around the jump vectors\footnote{This does not hold strictly for the case of oxygen diffusion in the distorted hexagonal close packed (hcp) lattice; however the error will be similar to the deviation from the bond lengths in a perfect hcp structure (ca. 5%).}, giving

$$\langle \cos \theta_{\gamma,\gamma+2} \rangle = \langle \cos \theta_{\gamma,\gamma+1} \cos \theta_{\gamma+1,\gamma+2} \rangle. \quad (A.28)$$

As the correlation between the $\gamma$th and $\gamma+1$th jump does not differ from that between the $\gamma+1$th and $\gamma+2$th jump in the absence of driving forces,

$$\langle \cos \theta_{\gamma,\gamma+1} \rangle = \langle \cos \theta_{\gamma+1,\gamma+2} \rangle. \quad (A.29)$$

Also, correlations are independent, so that

$$\langle \cos \theta_{\gamma,\gamma+1} \cos \theta_{\gamma+1,\gamma+2} \rangle = \langle \cos \theta_{\gamma,\gamma+1} \rangle \langle \cos \theta_{\gamma+1,\gamma+2} \rangle. \quad (A.30)$$

Combining Eqs. A.28, A.29, and A.30, we have

$$\langle \cos \theta_{\gamma,\gamma+2} \rangle = (\langle \cos \theta_{\gamma,\gamma+1} \rangle)^2, \quad (A.31)$$
and, by iteration,
\[ \langle \cos \theta_{\gamma,\gamma+m} \rangle = \langle \cos \theta \rangle^m, \]  
(A.32)

where \( \theta \) refers to the angle between any two successive defect jumps. Eq. A.26 then becomes
\[ f = 1 + \lim_{n \to \infty} \frac{2}{n} \left[ \langle \cos \theta \rangle + \langle \cos \theta \rangle^2 + \ldots + \langle \cos \theta \rangle^n \right] (n - 1), \]  
(A.33)

and, simplifying the geometric series, the correlation factor for vacancy diffusion in alumina becomes
\[ f_{V_O,V_{Al}} \equiv f = \frac{1 + \langle \cos \theta_{V_O,V_{Al}} \rangle}{1 - \langle \cos \theta_{V_O,V_{Al}} \rangle}, \]  
(A.34)

When interstitials migrate in alumina, this occurs via an interstitialcy mechanism (described in depth in Sec. 4.4.8). Then the above derivation applies, but only every second defect jump is correlated to the previous jump, so that \( \cos \theta_{\gamma,\gamma+m} = 0 \) if \( m > 1 \) and \( \cos \theta_{\gamma,\gamma+1} = 0 \) half of the time, reducing Eq. A.26 to
\[ f_{O_i,Al_i} = 1 + \langle \cos \theta'_{O_i,Al_i} \rangle, \]  
(A.35)

where \( \theta'_{O_i,Al_i} \) is the angle between the jump vector for an interstitial to lattice jump and the jump vector for the succeeding lattice to interstitial jump.


References


References


References


References


References


References


References


