A DFT investigation of Al-based atomically precise epitaxy

Richard Smith

Thesis submitted for the degree of PhD

Department of Physics and Astronomy

University College London

December 2020

Declaration

I, Richard Smith confirm that this thesis is a presentation of my original research work. Where information has been derived from the literature, this has been indicated in the references.

Acknowledgements

I would like to thank my supervisors Dr David Bowler and Dr Michael Gillan for their guidance and support over the entire period of my study.

Abstract

This thesis is about the growth and placement of dopants in silicon semiconductor devices and specifically acceptor dopants as device dimensions enter the nanoscale. Single-atom donor dopant devices have already been demonstrated in the laboratory. Using density functional theory (DFT) and the aluminium atom we now show how acceptor sites might be fabricated and characterize their electronic behaviour.

The thesis opens with a review of the physical basis of statistical doping and the operation of the silicon CMOS transistor which is the most widespread microfabricated device by a wide margin. We show how downscaling requires ever-increasing accuracy in dopant placement and illustrate using some current process techniques. Next, we describe some prototype single-dopant devices and the chapter concludes with a description of a phosphorus nuclear spin qubit and its application.

Chapter 2 outlines the theoretical basis of the DFT nanostructure models found in later chapters and chapter 3 presents some elementary calculations intended to validate the local DFT environment. Chapters 4, 5 and 6 are based on published papers produced during the course of this work and listed on page 11. In chapter 4 we introduce patterned atomic layer epitaxy (PALE), an experimental fabrication technique for Si nanostructures. Chapters 5 and 6 describe how PALE could be applied to locate Al dopant atoms in an Si substrate. The final chapter offers some calculations indicating the electronic behaviour of this dopant when embedded in Si nanostructures of various kinds.

Impact Statement

The work in this thesis is clearly theoretical, and so any anticipated impact is likely to be academic. Accordingly, most of the material has already been published in the *Journal of Physics and Condensed Matter*. The knowledge shared in this way builds on existing results in the precision doping field, which in due course will become established concepts in the semiconductor and nanoelectronics disciplines. To maximize the immediate impact, we selected our topics in liaison with *Zyvex Labs Inc.,* a commercial research organization currently developing atomically precise manufacturing (APM) technology. Therefore, we are also confident of the practical relevance of our work. Moreover, recent results showing that boron is ill-suited to precision doping (Škereň et al., 2020) should stimulate interest in aluminium, as we propose here.

More broadly, the context of this work is the demise of Moore's law and the ending of a 50 year period of exponential growth in computing power brought about by downscaling and large scale integration. Over this period key actors (e.g. Metcalfe, 1995; Ballmer, 2007) have made predictions which now seem hopelessly mistaken. However, an earlier forecast (Feynman, 1960) appears prescient. He calculated that a facsimile of the *Encyclopedia Britannica* could be written on the head of a pin with room to spare, as this area would provide at least 1000 surface atoms to encode each 'dot' of graphical data in the Encyclopedia. Feynman's example implies a bit density of about 10^{10} bits/mm^2. In 1970 Intel announced the first integrated circuit memory, a kilobit DRAM having a density of 10^2 bits/mm^2. Since then, sustained development has produced gigabit chips having the anticipated pinhead densities, a huge technical accomplishment and principal driver of the revolution in computer usage and the emergence of 'Big Tech'.

Earlier disruptive technologies such as printing and nuclear fission caused turbulence, but society adjusted and eventually found a new equilibrium. As computer and information technologies have matured and become pervasive, their harmful consequences are clearly visible. These include reduced opportunity in labour markets, income inequality and corruption of the political process by social media. Equilibrium has not returned and predicting the future remains a risky business.

Contents

List of figures

List of tables

List of publications

[1] **Smith, R**; Bowler, D R (2018). *Reaction paths of alane dissociation on the Si(001) surface*. J. Phys.: Condens. Matter 29 105002

[2] **Smith, R**; Bowler, D R (2017). *Alane adsorption and dissociation on the Si(001) surface*. J. Phys.: Condens. Matter 29 395001

[3] **Smith, R**; Brázdová, V; Bowler, D R (2014). *Hydrogen adsorption and diffusion around Si(001)/Si(110) corners in nanostructures*. J. Phys.: Condens. Matter 26 295301

Chapter 1

Introduction

This thesis is about the growth and placement of dopants in silicon semiconductor devices as the scale of those devices is reduced. In the limit of downscaling we can anticipate devices containing individually addressable dopant atoms. Silicon-based devices based on donor dopants have already been demonstrated in the laboratory. Using density functional theory (DFT) we now show how acceptor sites might be fabricated and characterize their electronic behaviour.

In this chapter we review the physical basis of silicon (Si) doping and the operation of the CMOS transistor which is the most voluminous microfabricated device by a wide margin. We show how downscaling requires ever-increasing accuracy in dopant placement and illustrate using some current process techniques. Next, we describe some prototype single-dopant devices and conclude with a description of the rather more speculative nuclear spin qubit and its application.

1.1 Crystalline silicon

Although DFT analysis can in principle reveal all the electronic properties of Si that are of interest, it is helpful to start from a qualitative viewpoint. The structure of Si in the solid state (MP/°C ≈ 1420) is determined by the nature of its inter-atomic bonding which in turn depends on its (Ne)$3s^2 3p^2$ electronic configuration. One of the 3s orbitals is notionally 'promoted' to the 3p level yielding a (Ne)$3s^{\uparrow} 3p^{x} 3p^{y} 3p^{z}$ configuration. The 3s and 3p orbitals now hybridize to form four *sp3* orbitals, each containing one valence electron and identical except for their orientation in space. These hybrid orbitals overlap with those of neighbouring atoms to form covalent bonds, so that each atom has four neighbours arranged in a tetrahedral coordination. Overall stability is increased, more than offsetting the initial hybridization cost. When large numbers of Si atoms bond in this way they form a homogeneous, periodically ordered structure or, in other words, a crystal. However, perfect ordering is rarely achieved throughout a sample: it will usually have a 'polycrystalline' state consisting of many microscopic zones of perfect ordering separated by defects and impurities.

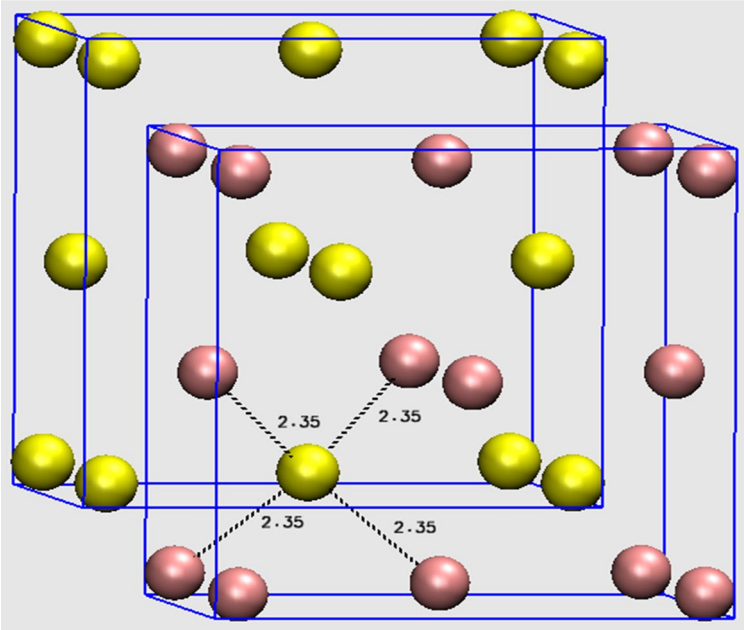


Figure 1.1 Intersecting face-centred cubic lattices (yellow, pink; a = 5.43 Å) give rise to tetrahedral coordination of the Si atoms with a bond length of 2.35 Å.

Crystalline silicon has the same structure as carbon atoms in a diamond crystal, consisting of two interleaved face-centred cubic lattices, with their origins at $(0,0,0)$ and $(¼, ¼, ¼)$. This organization (figure 1.1) yields a packing density of 34%, with 8 atoms in the unit cell and a lattice dimension a $\approx$ 5.43 Å. The relatively open structure encourages the fast diffusion of impurities (dopants) in silicon.

1.2 Band gap formation

In solid state physics the band gap in a crystal (a prerequisite of semiconductor behaviour) is attributed to Bragg reflection of the electronic wavefunctions by the ionic cores (Kittel, 1995). The reflections cause standing waves and accumulations of electrons having differing potential energies. This energy difference is the band gap and its width depends on the strength of the periodic ionic potential. However, we can continue to view the silicon crystal as a large molecule and borrow from the molecular orbital (MO) theory of chemistry to account for the band gap and basic semiconductor behaviour.

In the MO model each pair of overlapping hybridized orbitals are linearly combined to form a bonding and an anti-bonding orbital. The bonding orbital corresponds to constructive (in-phase) overlap and the anti-bonding orbital to a destructive (out-of-phase) overlap. In aggregate, all the bonding orbitals lie in a group of closely spaced energy values or valence band. The anti-bonding orbitals behave in a similar way to form the conduction band. If the valence and conduction bands do not overlap, an energy or band gap is said to exist. The tetrahedral bonding accounts for all four of silicon's valence electrons and means that each 3s and 3p orbital is half-filled. A fully filled valence band and a band gap are the essential prerequisites for a semiconductor. At very low temperatures there will be no free electrons

available to transport charge and the material behaves like an insulator (figure 1.2(a)). At higher temperatures, electrons may be sufficiently excited to cross the band gap to one of the many unoccupied states in the conduction band. When this happens a vacancy or hole is left in the valence band, forming an electron-hole pair (EHP) or *intrinsic* charge carrier. The width of the band gap depends on the strength of the periodic ionic potential, and in silicon is found experimentally to be $\approx$ 1.12 eV. EHP concentration is $\approx 10^{10}$ cm^{-3} at room temperature and the resulting resistivity ($\approx 2\text{x}10^{5}$ Ωcm) is high. But it is possible to create additional charge carriers (and greater levels of conductivity) by purposely introducing impurities into the crystal – a process called doping.

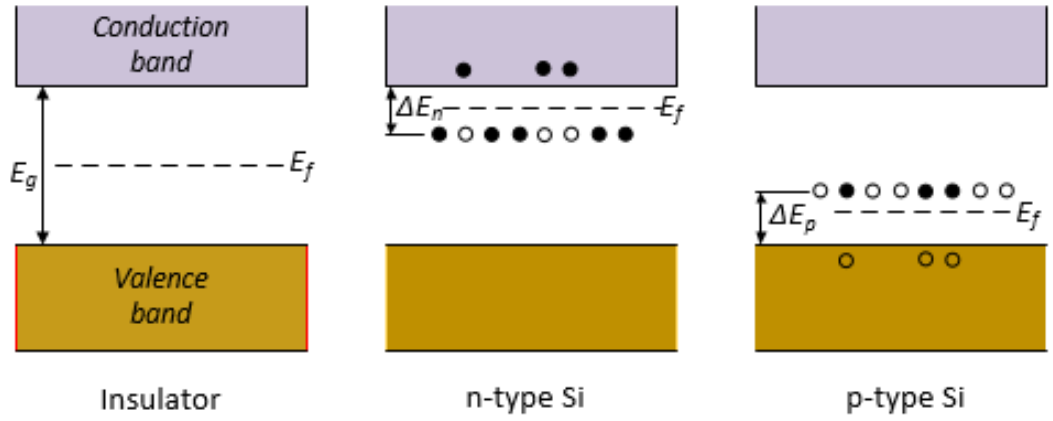


Figure 1.2 (Left) band gap E_g in ultrapure crystalline Si at low temperature, showing occupied valence band states (shaded) and unoccupied conduction band states. Fermi level E_f is the energy at which the probability of occupation is ½. (Centre) n-type Si doped with a group V donor element. Donated electrons occupy states ΔE_n beneath the conduction band. (Right) p-type silicon doped with a group III acceptor dopant with holes ΔE_p above the valence band (adapted from Greenwood; Earnshaw (1985) page 333).

1.3 Doping

Silicon is in column IV of the periodic table. If an atom from column V (e.g. phosphorus or arsenic) occurs in the silicon crystal as a defect it would make four covalent bonds with neighbouring atoms and also introduce an additional electron not involved in bonding, and to some extent delocalized. In silicon, the surplus electrons due to a column V impurity occupy energy levels about $0.03 - 0.05$ eV below the edge of the conduction band at $T = 0$ and at ≈ 100 K all the donor atoms are ionized, with the electrons promoted to levels within the conduction band. When silicon is doped with arsenic at a concentration of 10^{16} atoms cm^{-3}, the electron population in the conduction band increases by about five orders of magnitude compared with undoped material at room temperature. Its resistivity falls to about 5 Ωcm over a wide temperature range. This is characterized as *n-type* behaviour, and the electrons as *extrinsic* charge carriers (figure 1.2 (centre)).

An equivalent scenario unfolds when a column III impurity (e.g. boron) is used as dopant. These impurities have only three valence electrons, so some bonding orbitals are unfilled and MO theory cannot readily account for the resulting current flows. But one can observe that, in a fixed volume of space, the same current flows into specific unoccupied states as

would flow from all the remaining states if they were occupied by particles of charge $+e$ (opposite to the electronic charge) i.e. holes. So, whenever convenient, one can consider current to be carried entirely by hole carriers filling states unoccupied by electrons. In this case (figure 1.2 (right)) the silicon valence levels are not fully filled, and conductivity controlled by holes apparently having positive charge. At room temperature, hole concentration in the valence band far exceeds free electron concentration in the conduction band and holes comprise the majority carrier, otherwise characterized as *p-type* behaviour (Streetman; Banerjee, 2015).

1.4 Junctions

Real semiconductor devices possess at least one junction between p-type and n-type material. When the materials are joined a single crystal is formed with holes diffusing from the p side to the n, and electrons from the n to the p. This diffusion current causes an electrostatic field to build across the junction, causing an opposing drift current that reduces the net current flow to zero. At equilibrium (figure 1.3) the potential gradient is maintained in the near vicinity of the junction, resulting in a region depleted of charge carriers and known as the *space-charge* or *depletion* region. Its presence means the conduction band does not reflect the doping profile exactly, a phenomenon known as band-bending. The width of the depletion region depends on the relative carrier concentrations in the p and n regions. With moderate doping (p and n concentrations $\approx 10^{16}\ \mathrm{cm}^{-3}$), the gradient across the depletion region in Si would be about $0.6\ \mathrm{V}$, and the width about $2\ \mu\mathrm{m}$.

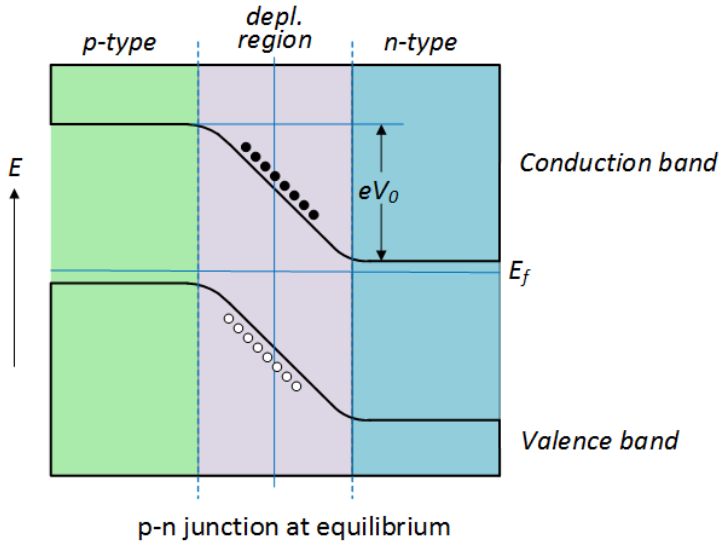


Figure 1.3 p-n junction having common Fermi level showing the contact potential gradient V_0. The valence and conduction bands are eV_0 higher on the p side of the junction than on the n side (attribution as fig 1.2).

If an external voltage is applied across the junction the voltage drop appears across the depletion region. If the p side is made positive the depletion region narrows, and charge carriers can flow across. When the external polarity is reversed, the region widens and no current flows. This is the rectifier action characteristic of a two-terminal p-n junction device or diode. The bipolar junction transistor (BJT) is a three-terminal device with two junctions,

e.g. p-n-p. In this case an external current flow into the central (base) layer can modulate the flow between the outer layers (emitter and collector), giving a switching or amplification function.

1.5 Fabrication basics

A brief description of silicon device fabrication is provided so that recent developments in the field can be visualized. The essential point is that fabrication is a planar process with devices constructed in layers in a sequence of low-level operations. The starting material is an optically flat disk of crystalline silicon or wafer typically 300 mm in diameter and capable of accommodating several hundred devices, each potentially containing many millions of junctions. The devices are later separated by sawing. The operations are oxidation, implantation, lithography, thin-film deposition and epitaxy (Franssila, 2010).

1.5.1 Oxidation

A silicon surface contains unsatisfied bonds and is highly reactive. A thin oxide film (SiO_2) forms on exposure to the atmosphere but layers of controllable thickness are produced by heating the wafers in an O_2 flow at $1100 - 1300$ K , yielding an oxide layer with thickness in the range $0.01 - 1$ μm. The oxide creates a non-conducting barrier between layers and can also accept a coating of photoresist in preparation for lithography. The widespread use of silicon as a semiconductor material is due in part to the ease of oxidation and insulating effect.

1.5.2 Lithography and etching

In this operation the oxide surface is spin-coated with an ultraviolet-light sensitive photoresist material. The photoresist is a high molecular weight polymer dissolved in organic solvent, characterized by its radius of gyration or tendency to coat evenly. The resist is exposed through a mask defining the extent of each junction. The photomask is a chrome-coated glass plate carrying a pattern created in a process analogous to laser printing. With positive resist, the exposed areas are polymerized. When the wafer is placed in a developer the polymerized resist softens and can be washed away. The oxide cover on the unmasked areas can then be selectively etched away and the area doped to create semiconductor junctions. In this 'wet-etch' scenario SiO_2 is etched by dilute hydrofluoric acid leaving a H-passivated Si surface, a key benefit. In an alternative process, often referred to as reactive ion etching (RIE) the etchant is a plasma of gas such as carbon tetrafluoride CF_4. Here Si-O bonds are broken and replaced by thermodynamically favoured Si-F bonds. RIE is capable of finer resolution than wet-etching and is preferred as device dimensions are reduced. However, ultimate resolution is set by the combined limitations of the optical, resist, developing and etching processes. Advances in all these areas and some ingenuity make it possible to etch 7 nm features with an extreme (sub-100 nm) UV light source (Samsung, 2019).

1.5.3 Ion Implantation

Ion implantation is how dopant impurities are introduced into the silicon lattice. It works by accelerating impurity ions to high energies (between 10 keV and 200 keV) shooting them into the semiconductor. This causes disorder in the lattice, which is repaired by subsequent heat treatment. Implantation is blocked by the oxide layer or by a resist or photomask, provided it is thick enough. The final dopant concentration has a Gaussian profile peaking at a depth determined by the beam energy and exposure time and damage to the lattice outside this region is not significant. Implantation has supplanted gas diffusion as device dimensions have reduced, because it is easier to produce shallow localized regions having relatively high dopant concentrations. However, some limited diffusion always occurs naturally, following an implantation.

1.5.4 Thin-film deposition

Metallic thin-film deposition provides the means to connect junctions to the outside world. Sputtering is a form of physical vapour deposition (PVD) in which a metal (e.g. aluminium) is ionized by bombarding a sample with highly charged argon ions in a vacuum chamber. The metal ions land on a masked wafer, to form conductive tracks. The tracks make ohmic (low resistance, non-directional) interconnections between devices or to pads that will make the external connections. Later, the pads are joined to terminals on the device package by welding.

In chemical vapour deposition (CVD) the film is created from gas phase components, catalysed by the silicon surface. An important CVD process involves gaseous tungsten fluoride WF_6 and silane SiH_4 precursors. This reaction deposits metallic tungsten that can fill holes in the substrate, allowing the interconnection of metallization layers and an increase in circuit density.

1.5.5 Epitaxy

Epitaxy is a special form of thin-layer deposition. Whereas CVD and PVD generally form an amorphous or polycrystalline layer, epitaxy produces a film that is an extension of the substrate that preserves its lattice structure. In molecular beam epitaxy (MBE) each constituent to be deposited is heated in in a cylindrical cell with an aperture, and the cells and the substrate are located in a vacuum chamber. Narrow beams of atoms flow out of the cells and impinge on the substrate forming a molecular layer.

1.6 Field effect transistors

Although the junction transistor was prevalent circa 1960 by the end of the decade it had been superceded by the field effect transistor (FET). An n-channel metal-oxide-silicon (NMOS) FET is formed by implanting two heavily doped n regions (the source and drain) into a lightly doped p substrate. The substrate is covered in a thin insulation layer and a third layer of metal (the gate) bridges the channel between the n regions beneath. The drain electrode is maintained at a positive potential (relative to the substrate which is connected to the source) but the p-n junction potentials prevent current flow across the channel. When

a positive gate potential is applied the potential barrier is lowered and an electron current can flow from the source to the drain. In the complementary p-channel form p and n regions are interchanged, and the PMOS devices can be fabricated on the same substrate as the NMOS as shown in figure 1.4. The principal advantage of the FET over the BJT is that it is a voltage-controlled device where the gate presents a capacitive load, so gate current flows only when the device changes state. This simplifies interconnection and reduces power consumption.

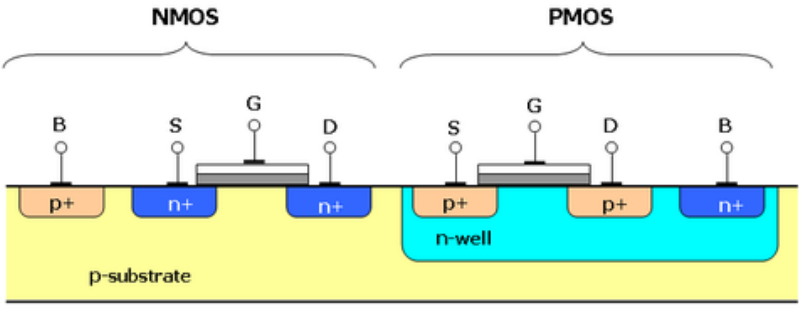


Figure 1.4 Sketch of CMOS doping profile showing PMOS transistor formed in a n-type well on a p-type Si substrate. The B connections are to the device bodies and are strapped to the respective source electrodes in most circuit configurations (adapted from an image by R Mirhosseini / CC by 2.0).

Designers can combine complementary MOSFETs (CMOS) to realize Boolean logic functions. A simple example is the inverter or NOT function shown at figure 1.5, from which any other Boolean function can be realized. Two of these, suitably coupled, can form a single-bit static memory cell (SRAM) that will retain its contents while power is applied. An integrated circuit or IC (e.g. a microprocessor or memory array) is formed when large numbers of gates and memory elements are fabricated on a single silicon surface and interconnected by metallization layers. CMOS ICs (complementary MOS) devices are the core product of the modern semiconductor industry.

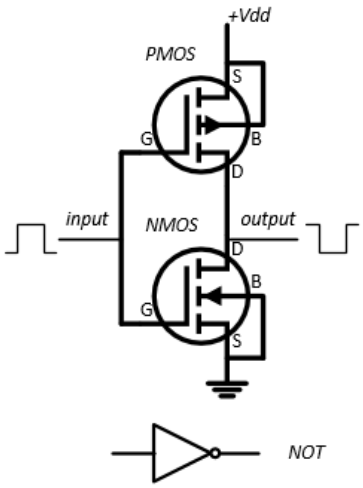


Figure 1.5 CMOS inverter circuit with series connected PMOS and NMOS transistors with interconnected drains, and the device body (substrate) connected to the respective source electrodes. When the input is grounded (logic 0), the positive supply +VDD (logic 1) appears at the output, and vice versa. The equivalent Boolean NOT symbol is also shown (author's MS VISIO® image).

1.7 Large scale integration and downscaling

The driver of integration is economic: wafer processes are capital-intensive, but each wafer contains hundreds of dies (chips) and the wafers are processed in batches, so unit costs are kept low. But the nature of lithography means that per-chip costs do not rise in proportion to the number of transistors, providing the incentive to maximize the functionality per chip.

Advances in process technology stimulated by economic forces have led to a huge rise in the number of devices per chip, and an equivalent reduction in the size of each transistor. The metric for this is feature size, i.e. the size of smallest etched element e.g. the pitch of metallized lines in the interconnect layers. Figure 1.6 shows microprocessor transistor count from 1990 to the present day. This straight-line plot with semi-log vertical axis is the exponential relationship known as Moore's Law (after G. Moore co-founder of the Intel Corporation). Of course, it is not a physical law but an empirical prediction of the interaction of semiconductor technology with market economics.

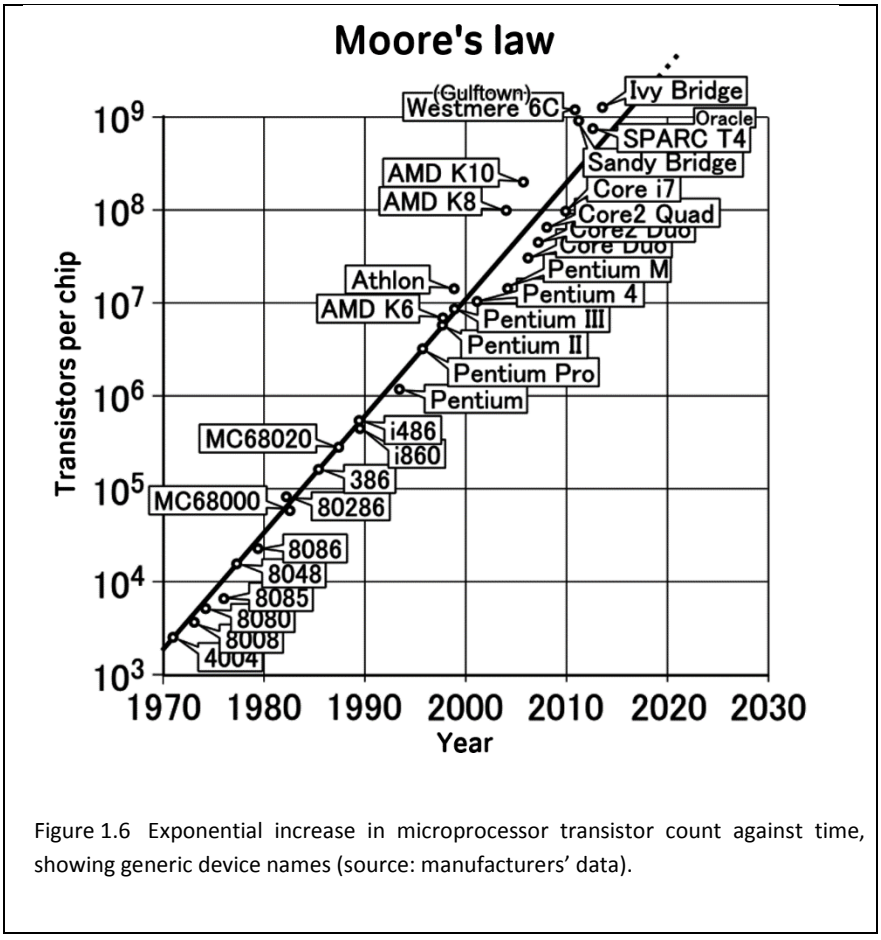


Figure 1.6 Exponential increase in microprocessor transistor count against time, showing generic device names (source: manufacturers' data).

Reduced areas allow more chips per wafer and lower overall fabrication costs. MOSFET capacitances are also lowered, which improves switching speed and reduces power switching dissipation. However, power dissipation density ($\mathrm{W/cm^2}$) will rise excessively unless supply voltages are lowered in line with device dimensions. In the constant field scaling model, device dimensions and all terminal potentials are reduced by a factor S ($S >$

1) and dopant concentration increased by factor S to preserve electric field gradients within the device, and power dissipation density is unaltered. However lowered supply voltages mean the FET threshold voltage (the gate voltage required to bias the device into conducting mode) must also be reduced or switching speed is impaired. But this degrades switch performance (the current ratio i_{on}/i_{off}) due to current leakage between the electrodes. In the constant voltage model voltages are unaltered, but power density rises by S^3 and dopant concentration by S^2. In practice voltages have been scaled conservatively (from 5 V to 1 V or $S \approx 5$), while transistor dimensions have fallen by a factor $S \approx 10^6$ But the resulting increased field strengths and dopant concentrations can have adverse side-effects, such as breakdown of the gate insulation, channel surface scattering and lowered long-term reliability.

Generally, increased circuit density has been achieved incrementally through physical trade-offs and concurrent improvements in process technology including multi-layered structures with internal metallized interconnects, strained semiconductor materials, and new gate insulators amongst others. However, aggressive increases in density have not been matched by an equivalent rise in device efficiency, so overall heat dissipation has tended to increase.

1.8 Non-evolutionary change

Moore's law is a good description of an empirical, non-fundamental relationship but cannot go on for ever. Extrapolated for another 20 years, the current 10 nm feature size would become sub-atomic so the devices of that era could not be the CMOS integrated circuits of today. In any case the device scaling methodology described above does not consider many other chip performance and reliability issues, e.g., interconnect performance, device isolation and external connection. In 2016 the authoritative *International Technology Roadmap for Semiconductors*, produced biannually by expert industry groups, announced the cessation of publication and signalled the end of Moore's Law as a useful metric. A vast amount of research is now directed to finding replacement materials and processes that would allow scaling to continue, or perhaps presage some new 'beyond-MOS' technology. Several proposals for FETs having exotic channel materials have been made, including (amongst others) single electron devices and Si nanowires. At the time of writing (2020) none can demonstrate a decisive advantage and it seems likely that the regime of incremental improvement will continue for a few years, terminating as feature sizes approach 5 nm.

1.8.1 Single electron devices

A single electron device (SED) consists of conducting islands separated by tunnel barriers. By applying bias voltages to the islands (usually referred to as quantum dots or QDs) it is possible control the flow of charge between them on an electron-by-electron basis. The tunnel barrier and the QD can be modelled as 'tunnel' and 'normal' capacitors in series, and the charge on the QD is quantized in a Coulomb blockade staircase or CB, see figure 1.7. To be useful as a CMOS substitute the QD is connected to two tunnel capacitors to form a three terminal device or single electron transistor (SET) reminiscent of a MOSFET, but with

different characteristics. Nevertheless, it has been shown (Takahashi et al., 2002) that SETs can be combined to form inverters and basic logic elements leading in theory to the possibility of large scale integration. Interestingly, the QD charging energy increases as the island size is reduced, so it could swamp variations in background energies due to ambient conditions. However, the island size corresponding to a charging energy of $250\,\mathrm{meV} \approx 10\mathrm{k_B T}$ at room temperatures) is $\approx 1\,\mathrm{nm}$, far too small for any current lithographic process. A dependency on low-temperature operation seems to preclude adoption as a CMOS replacement.

More fundamentally, the so-called 'offset' problem may prove insurmountable. This arises because the SET has no natural ground to which signal levels can be referred. The presence of background charge tends to shift the SET transfer characteristic away from its optimum point, making reliable switching impossible. If the device were implemented with P donor dots atoms on a silicon substrate then some form of isolation (e.g. through the introduction of Al acceptor atoms, the subject of this work) would be desirable.

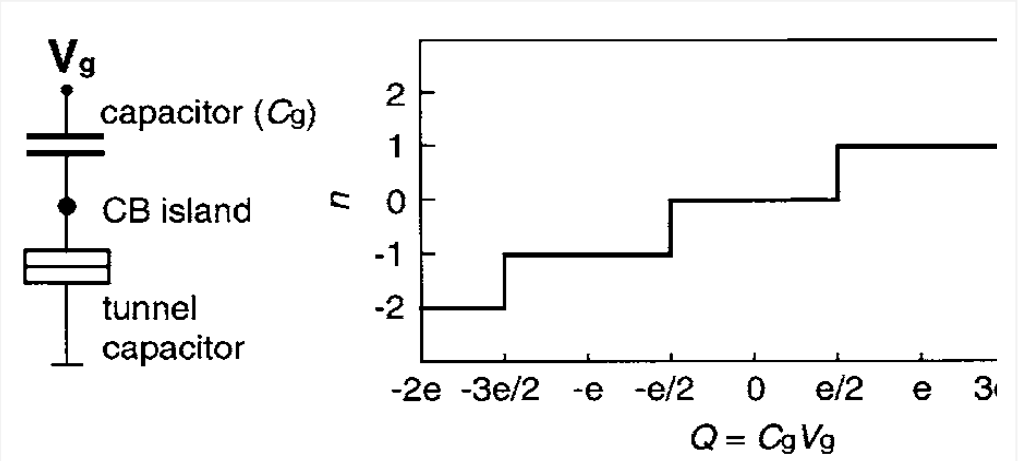


Figure 1.7 Equivalent circuit and stability diagram of a single-electron box. In the diagram, n is the electron number in the island, and Q and V_g are the external charge and voltage to the normal capacitor with a capacitance C_g (Takahashi et al., 2002).

1.8.2 Si nanowire FET

This has been regarded as the most promising candidate for mainstream CMOS devices (Iwai et al., 2012) because the fabrication processes are expected to be compatible with existing CMOS techniques. The MOSFET channel (currently a fin or thin layer formed from strained SOI growth) is replaced by a stack of Si nanowires each of which is jacketed by gate material. Gate control is thereby enhanced, reducing current leakage along the wires to low levels (i.e. obviating the short-channel effect). Low leakage reduces the chip's off-state power dissipation, permitting a higher packing density. The wires themselves would be $5 - 10\,\mathrm{nm}$ in diameter and fabricated by etching, finishing with a hydrogen anneal to provide the gate insulation (figure 1.8). At these dimensions conductivity is dominated by quantum effects. Modelling (ibid, 2012) suggests that a thin nanowire restricts the freedom for carrier scattering, giving quasi one-dimensional conduction, good electron mobility and high drive (on-state) currents. However, some issues remain unresolved including the effect of surface

states on conductivity and the fabrication of effective contacts with the source and drain electrodes and the early promise of this technology remains unrealized.

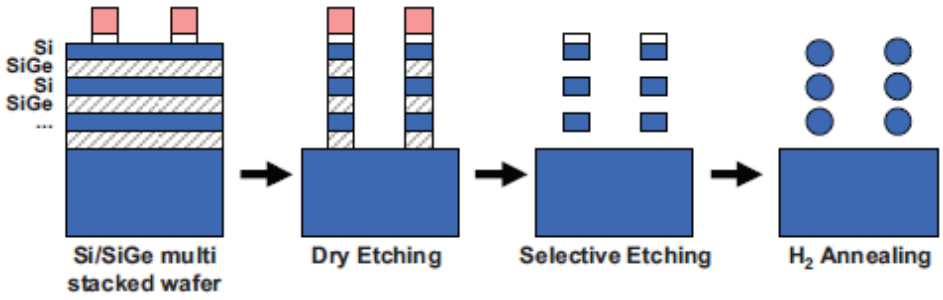


Figure 1.8 Si nanowire FET fabrication process. The 'gate stack' is several Si and sacrificial SiGe layers formed by epitaxy. A masked plasma-etch step creates a fin structure and selective chemical etch based on HF is used to remove the SiGe layers, leaving freestanding Si wires. The H_2 annealing step rounds the nanowire corners and gives a smooth crystalline surface. Subsequently (not shown) wrap-around gates are formed by oxidation and tungsten deposition. (Iwai, 2012)

1.9 Quantum computing

1.9.1 Computable problems

Even if a successor to CMOS exploiting quantum mechanical effects is discovered, its objective would remain limited to the implementation of the classical stored-program architecture dating back to the 1950s. A stored-program computer (SPC) has a central processing unit, random access memory capable of holding both data and instructions and a logic unit holding the current instruction and its memory address. The SPC is itself a realization of the theoretical Turing machine (after its inventor A. M. Turing). The value of the Turing machine is that it can determine what kind of problems a stored-program machine can solve. If a computation cannot be performed on a theoretical Turing machine, then it cannot be performed on any classical machine, no matter how powerful. If a Turing machine can perform the computation, then it is *computable*. Turing also showed the existence of *uncomputable* problems, but they must be distinguished from those which might be computable but only over an unreasonably long period of time.

1.9.2 Qubits and qubit registers

A quantum computer (QC) has a radically different structure and operating principle, offering the possibility of attacking uncomputable or otherwise intractable problems. The memory of a quantum computer consists of two-state quantum systems called qubits. A qubit can be measured in either of its two basis states, but when it is not measured exists in continuum of states that are linear combinations (superpositions) of the basis states. An n-bit qubit register with n qubits can exist in 2^n states, each described by a pair of complex numbers. On measurement it will collapse into just one state described by a single real eigenvalue, with a probability determined by the incidence of that state in the totality of superposed states.

1.9.3 Gates and networks

The QC possesses gate circuitry to perform primitive qubit operations, e.g. rotating a base state into a superposition of both. The gates are combined into networks to execute some desired transformation on a register and the networks are analogous to the stored program in a classical computer. The power of the QC arises from its ability to transform *all* the superposed states of a register in a single operation, creating a superposition of the transformed states in another register, suggesting an exponential increase in performance. To realize this gain, new quantum algorithms that can exploit parallelism and able to interpret the probabilistic nature of the results must be devised. Some have already emerged, notably Shor's algorithm (Shor, 1995) for finding prime factors in polynomial rather than exponential time. It envisages a sequence of classical and quantum procedures, the latter calling for gate networks wired for modular exponentiation and the discrete Fourier transform.

1.9.4 Decoherence

Any real QC will suffer from decoherence, the unavoidable loss of quantum state information through interaction with the surrounding environment. The qubit coherence time (persistence) sets a limit on the number of gate operations available to run an algorithm. Present-day QC's (e.g. the IBM Q environment) utilize superconductive or ion-entrapment phenomena and achieve coherence times of about 100 microseconds. Quantum transitions occur on a femtosecond timescale, but housekeeping overheads will make the effective clock rate (gates per second) much slower. Consequently, attention has focused on qubit error correction protocols allowing longer-running algorithms. These operate analogously to the Hamming code scheme for classical bits which can recover single-bit errors by encoding additional protection bits. However, it has been suggested (Preskill, 1996) that a QC capable of factorizing a 130-digit integer would require of the order 10^6 qubits to achieve a similar level of protection, far larger than any current QC.

1.9.5 Solid state qubits

Several ideas for solid state qubit devices have been advanced, motivated by the prospect of exploiting existing Si fabrication expertise to produce a mass qubit memory. Kane (1996) suggests that the two spin states of the ^{31}P atomic nucleus could form a qubit basis. This is attractive because P is an effective Si dopant and its spin coherence time is long at cryogenic temperatures. The nuclear environment is naturally protected and a pure ^{28}Si host crystal would have no native spin to interfere with the donor spin states. But current mask-based dopant deposition cannot provide the precision needed to accurately locate a single dopant nor the metallic control gates envisaged by Kane, and his solid-state qubits have yet to reach the prototype stage (figure 1.9).

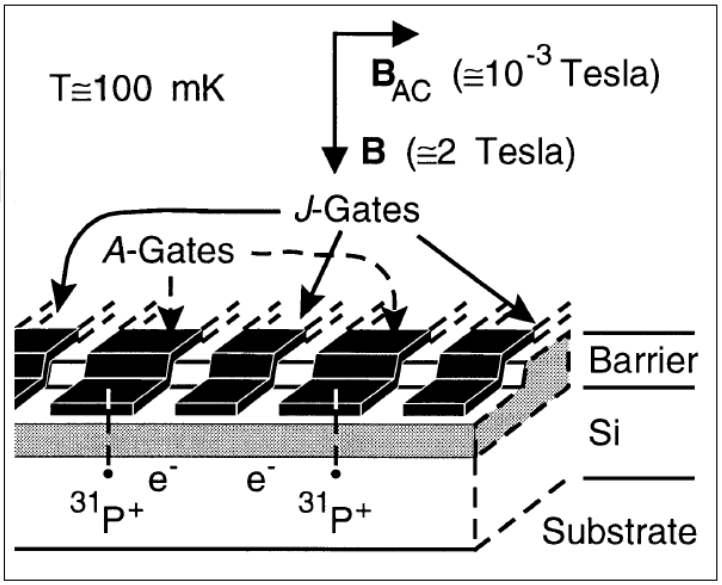


Figure 1.9 Illustration of two cells in a one-dimensional array containing ^{31}P donors and electrons in a Si host, separated by a barrier from metal gates on the surface. The degeneracy of the nuclear spin states is split by the external magnetic field B, and the spin state is flipped by the alternating field B_{AC}. 'A gates' control the resonance frequency of the nuclear spin qubits; 'J gates' control the electron-mediated coupling between adjacent nuclear spins (Kane, 1996).

1.10 Summary

This chapter has provided the technological context in which the development of new nanoscale circuit devices will occur, as existing semiconductor processes approach their physical limits. As the latter can already produce sub 10 nm devices it may seem that the scope for new processes has diminished but nanoscale research is still in its infancy, perhaps resembling Shockley's point-contact transistor of 1947.

DFT modelling of P dopant deposition in phosphine CVD (Warschkow, O et al., 2005) and subsequent incorporation (Warschkow, O et al., 2016) has shown that precision doping of this impurity is theoretically feasible , and some prototype P-based devices have been demonstrated (Fuechsle, M et al., 2012; He et al., 2019). It is the anticipated requirement for complementary acceptor devices that motivates this work, which supplies an analogous DFT modelling for the Al dopant and alane gas precursor. DFT and its application to small silicon nanostructures are discussed in chapters 2 and 3. Chapter 4 describes hydrogen-passivated PALE (Patterned Atomic Layer Epitaxy), an atomically precise nanostructure fabrication technique and also provides some modelling of the passivation process. Al deposition and incorporation are covered in chapters 5 and 6 and the thesis concludes with an examination of the electronic behaviour of the Al dopant when embedded in nanostructures of various kinds.

Chapter 2

Theoretical Background

2.1 Historical note

In the following chapters, we simulate the processes of silicon nanostructure fabrication at the atomic level. The models imply a knowledge of the total energy E of a system of N electrons in the presence of ions located at R_I. The potential energy surface $E(R_I)$ or PES generally has vast numbers of maxima and minima at unknown locations, but the lowest energy corresponds to the ground state structure, and paths between its minima determine the feasibility of processes such as adsorption, diffusion and incorporation.

The present calculations are performed using the Density Functional Theory (DFT) formalism. This is a relatively recent innovation and allows realistic modelling of structures containing hundreds of atoms, which were previously inaccessible. Nevertheless, DFT emerges from a stream of work dating from 1926, when the Schrodinger equation (SE) was first put forward and the search for approximate and practical methods of application began. This chapter will outline DFT as it applies to a crystalline solid such as silicon, but first we look briefly at the earlier and complementary *Hartree-Fock* (HF) approach as it embodies ideas that reappear in the DFT setting and also introduces some necessary terminology.

Both HF and DFT make the *Born-Oppenheimer* (BO) approximation under which electronic and ionic motions can be considered separately and the R_I enter the electronic energy calculation only parametrically. Pathways on PES can be explored by making ground state calculations at a sequence of carefully chosen locations $\{R_I\}$, which would ideally include the critical saddle points.

After the BO approximation the Hamiltonian for the N-electron system takes the form

$$\hat{H} = -\sum_{i=1}^{N} \frac{\hbar^2}{2m_e} \nabla_{\vec{r}_i}^2 + \sum_{i=1}^{N} v(\vec{r}_i) + \sum_{i=1}^{N} \sum_{j<i}^{N} u(\vec{r}_i, \vec{r}_j) \qquad (1)$$

In which the implicit dependence on R_I has been suppressed. The three terms on the right are (in order) the kinetic energy of each electron, the interaction energy of each electron with the ions and the interaction energy between different electrons. The SE

$$\hat{H}\Psi = \hat{H}\Psi(\vec{r}_1, \dots \vec{r}_N) = E\Psi \qquad (2)$$

Is intractable because the electronic states are correlated. Nevertheless, the first (*circa* 1930) results were achieved by neglecting all inter-electronic interaction and assuming $\hat{H}$ could be approximated by the sum of N independent one-electron Hamiltonians $\hat{h}$, so that

$$\hat{H} = \sum_{i=1}^{N} \hat{h}_i \tag{3}$$

and

$$\hat{h}_i = -\frac{\hbar}{2m_e}\nabla^2_{\vec{r}_i} - \upsilon(\vec{r}_i) \tag{4}$$

Each $\hat{h}_i$ satisfies the one-electron SE

$$\hat{h}\psi = E\psi \tag{5}$$

The eigenfunctions defined by this equation are the spin orbitals of MO theory and can be denoted by $\chi_j(\vec{x}_i), (j = 1,2,..)$ where $\vec{x}_i$ is a vector of coordinates giving the position of electron i and its spin state, with corresponding eigenenergies E_j. It is convenient to label the spin orbitals so that the orbital with $j = 1$ has the lowest energy and $j = 2$ is the next lowest, etc. Then, the eigenfunctions of $\hat{H}$ are the products of the one-electron spin orbitals

$$\Psi_{HP}\,(\vec{x}_1, \vec{x}_2, \dots, x_N\,) = \chi_{j1}(\vec{x}_1)\chi_{j2}(\vec{x}_2)\dots\chi_{jN}(\vec{x}_N) \tag{6}$$

where Ψ_{HP} is called a *Hartree product wavefunction* whose energy is $E_{j1} + \cdots + E_{jN}$. However, the Hamiltonians of (3) and (5) do not include interelectronic repulsion, which is difficult to calculate because it depends not on one electron but all possible simultaneous pairwise interactions. Hartree found that $\langle \Psi_{HP} \,|\hat{H}|\, \Psi_{HP} \rangle$ would be minimized if the one-electron operators $\hat{h}_i$ were defined not as (4) above but by

$$\hat{h}_i = \frac{\hbar}{2m_e}\nabla^2_{\vec{r}_i} - \upsilon(\vec{r}_i) + V_i\{j\} \tag{7}$$

where the final term represents an interaction potential that may be computed as

$$V_i\{j\} = \sum_{i \neq j} \int \frac{\rho_j}{\hat{r}_{ij}}\, d\hat{r} \tag{8}$$

In which ρ_j is the charge (probability) density associated with electron j. The interaction potential (also called the *Hartree potential*) is analogous to the attractive second term except that the nuclei are treated as point charges whereas the electrons are treated as delocalized charge densities that must be integrated over all space. This means each electron feels the effect of others in an average, rather than an instantaneous way.

The Hartree product is an uncorrelated or independent electron wavefunction because

$$|\Psi_{HP}\,(\vec{x}_1, \vec{x}_2, \ldots, x_N\,)|^2 d\vec{x}_1, d\vec{x}_2, \ldots, d\vec{x}_N \tag{9}$$

which is the probability of finding electron-one in a volume element $d\vec{x}_1$ centred at $\vec{x}_1$, electron-two at $d\vec{x}_2$ is equal to the product of the probabilities

$$\left|\chi_{j1}(\vec{x}_1)\right|^2 d\vec{x}_1 \left|\chi_{j2}(\vec{x}_2)\right|^2 d\vec{x}_2 \ldots \left|\chi_{jN}(\vec{x}_N)\right|^2 d\vec{x}_N \tag{10}$$

i.e. the probability of finding an electron at one point in space is unaffected by the positions of all the other electrons. But electrons are fermions and the Hartree product should therefore change sign when two electrons are interchanged, which is not the case with the simple product formulation. If the one-electron orbitals are combined in a determinantal form (the Slater determinant) the wavefunction is both antisymmetric and independent of electron labelling. This can be seen for two electrons with spin orbitals χ_j and χ_k occupied:

$$\Psi_{SD}(\vec{x}_1, \vec{x}_2) = 2^{-1/2} \begin{vmatrix} \chi_j(\vec{x}_1) & \chi_k(\vec{x}_1) \\ \chi_j(\vec{x}_2) & \chi_k(\vec{x}_2) \end{vmatrix} \tag{11}$$

where the $2^{-1/2}$ is a normalization factor. When the rows (electrons) are interchanged, the wavefunction changes sign in accordance with the antisymmetry principle. If the spin orbitals are expressed as products of spatial and spin eigenfunctions and the determinant expanded, Ψ_{SD} is shown to be uncorrelated for electrons of opposite spin but correlated for electrons of the same spin. The resulting *exchange correlation* energy causes a small increase in the energy of the N-electron system.

However, the presence of the Hartree interaction potential makes the set of one-electron equations non-linear and not directly soluble for the χ_i. To find solutions, trial wavefunctions are approximated as linear combinations of a finite set of functions $\{\phi_i\}$, $(i = 1, .., K)$ so that

$$\chi_J(\vec{x}) = \sum_{i=1}^{K} \alpha_{j,i}\, \phi_i(\vec{x}) \tag{12}$$

The functions $\{\phi_i\}$ are called a *basis set* and usually formed from hydrogenic atomic orbitals or Gaussians. The $\alpha_{j,i}$ are then treated as parameters in an iterative (referred to as a *self-consistent field* or *SCF*) procedure that relies on the *variational principle*, that the trial wavefunctions will always yield a total energy greater than the exact ground state energy E_g:

$$E_{SCF} = \frac{\int \Psi_{HP}^* H \Psi_{HP}\, d\tau}{\int \Psi_{HP}^*\, \Psi_{HP}\, d\tau} \geq E_g \tag{13}$$

The variational method involves choosing the $\alpha_{j,i}$ so that E_{SCF} is minimized:

$$\frac{\delta E_{SCF}}{\delta \alpha_{j,i}} = 0 \tag{14}$$

producing a new Ψ_{HP} and E_{SCF}. The procedure is then repeated until E_{SCF} converges, i.e. approaches a limiting value to within some preset tolerance, e.g. 10^{-5} eV. Such convergence is not guaranteed, and it may be necessary to use a different basis set that better spans the wave function space (Szabo; Ostlund, 1996) or delivers shorter calculation times. When the converged spin orbitals are combined in a Slater determinant the resulting

wavefunction is the best variational ground state approximation available at this level of theory. This is the essence of the Hartree-Fock method that remains a staple of computational chemistry, with a huge volume of ensuing development aimed at improving the many-body wavefunction to take fuller account of electron correlation. One approach is to form excited determinants that include one or more *unoccupied* spin orbitals and use these as basis functions to approximate the N-electron wavefunction. Correlation energy is *defined* as the difference between the true ground-state energy and the best possible HF approximation known as the *Hartree-Fock limit.* Unsurprisingly, the correlation problem arises again in DFT, which includes an explicit (albeit approximate) correlation energy contribution in the Hamiltonian (see sections 2.4 and 2.5 below).

A general drawback of these 'post-HF' developments is the complexity of the N-electron HF wavefunction. This creates a so-called 'exponential wall' of calculation (Kohn, 1998), becoming insuperable in configurations of more than 10 or 20 chemically active atoms. Dirac (1930) suggested that the atomic state was completely determined by the 3-dimensional electronic density, foreseeing a great reduction in the burden of calculation. However, it was not until the 1960s that a rigorous proof of this was advanced, and the accompanying formalism employed several ideas from HF theory, including the one-electron assumption, the Hartree energy and self-consistent calculations.

2.2 Density functional theory

DFT is another approach to the solution of the SE for a system of atoms where the electronic Hamiltonian can be written in the form of (1) above, i.e. where the Born-Oppenheimer approximation has been applied. DFT provides universal prescriptions for the kinetic and electrostatic interaction components of the Hamiltonian:

$$\hat{T} = -\frac{\hbar}{2m_e}\sum_{i=1}^{N}\nabla_i^2, \quad \hat{U} = \frac{e^2}{8\pi\epsilon_0}\sum_{i\neq j}\frac{1}{|\vec{r}_i - \vec{r}_j|} \tag{15}$$

where $\hat{T}$ is the kinetic energy of the electrons and $\hat{U}$ the sum of all the Coulomb interactions with each other. The notation

$$\hat{H}_0 = \hat{T} + \hat{U} \tag{16}$$

refers to the Hamiltonian of the electronic system by itself. When $\hat{H}_0$ is treated as system-independent all Coulombic systems having the same number of electrons differ only in their *external potential*

$$\hat{V} = \sum_{i=1}^{N} v(\vec{r}_i) \tag{17}$$

which is the sum of all Coulomb interactions of the electrons with the nuclei. The SE is then:

$$\left(\hat{H}_0 + \hat{V}\right)\Psi(\vec{r}_1, \vec{r}_2 \ldots, \vec{r}_n) = E\Psi(\vec{r}_1, \vec{r}_2 \ldots, \vec{r}_n) \tag{18}$$

and we want the ground-state energy E_g for any potential $v(r)$. This is essentially the same intractable SE that arises in the HF theory when the Hartree and exchange energies were included. There some electronic correlation effects were neglected and Ψ approximated as a product of one-electron wavefunctions. DFT provides a way of including some correlation effects within a tractable computational method, which is usually easier to apply than advanced HF methods. This is done by promoting the electronic density $\rho(\vec{r})$ from just one of many observables to become the key variable, on which all other observables can depend. The assumption that T and V could be treated as functionals of electronic density, computed with no reference to a wavefunction was first advanced by Thomas and Fermi in 1927, without a rigorous justification. This was eventually provided in two fundamental theorems dating from 1964 known as the Hohenberg-Kohn theorems, stated in the following section with proofs shown in Appendix A.1.

2.3 The Hohenberg-Kohn theorems

The aim is to calculate the ground-state energy E_g of the system and the electronic density distribution $\rho_g(r)$ in the ground state. (This energy is not the *total* energy since it excludes the contribution from Coulombic inter-nuclear interactions, but this is constant in the BO approximation and can be added back later.) E_g is the expectation value of $\hat{H}_0 + \hat{V}$ computed with the many-electron ground-state wavefunction Ψ:

$$E_g = \langle \Psi^* | \hat{H}_0 + \hat{V} | \Psi \rangle \tag{19}$$

assuming Ψ to be normalized so that $\langle \Psi^* | \Psi \rangle = 1$. The electron density is also an expectation value:

$$\rho(\vec{r}) = \langle \Psi^* | \hat{n}(\vec{r}) | \Psi \rangle \tag{20}$$

where $\hat{n}(r)$ is the operator defined as:

$$\hat{n}(\vec{r}) = \sum_{i=1}^{N} \delta(\vec{r} - \vec{r}_i) \tag{21}$$

If the (non-degenerate) ground state energy is known, then in principle Ψ and hence $\rho(\vec{r})$ are uniquely determined by $V(\vec{r})$. although Ψ is far too complicated to calculate exactly. DFT assumes the converse – that specifying the electronic density $\rho(\vec{r})$ is enough to uniquely determine the ground state properties, including the ground state energy E_g. The first Hohenberg-Kohn shows that this assumption is justified.

- *Theorem 1*: *It is impossible that two external potentials $v(\vec{r})$ and $v'(\vec{r})$ whose difference $v(\vec{r}) - v'(\vec{r})$ is not a constant give rise to the same ground-state density distribution $\rho_g(\vec{r})$.*

The corollary is that the (non-degenerate) ground-state density $\rho(\vec{r})$ uniquely determines $v(\vec{r})$ up to an additive constant, and hence Ψ and E_g. The proviso about the additive constant simply reflects the fact that if the external potential is changed by adding a constant

to it (equivalent to shifting the zero of energy) then this does not change the ground state Ψ or the ground state $\rho_g(\vec{r})$.

The theorem implies that the ground-state energy can be expressed as:

$$E_g = \int d\vec{r}\rho(\vec{r})v(\vec{r}) + F[\rho(\vec{r})] \tag{22}$$

where $F[\rho(\vec{r})]$ is some universal functional of the density $\rho(\vec{r})$ representing the expectation value of $\hat{H}_0 = \hat{T} + \hat{U}$ (the total kinetic energy plus the total electron-electron interaction energy) when the ground-state density is $\rho(\vec{r})$.

Because of theorem 1, specifying the ground state density uniquely determines the value of F and hence the total ground state energy. However, the theorem does not indicate how the functional or density can be calculated. But Theorem 2 asserts that the ground state energy will be found by a minimization procedure:

- *Theorem 2: The ground state energy for a given external potential $v(\vec{r})$ is correctly obtained by minimizing the functional $E_g = \int d\vec{r}\rho(\vec{r})v(\vec{r}) + F[\rho(\vec{r})]$ with respect to $\rho(\vec{r})$ subject to a fixed number of electrons N, and the resulting $\rho(\vec{r})$ gives the correct density distribution of the ground-state.*

In order to turn these theorems into a useful procedure for calculating ground-state energies, it is necessary to reformulate the total energy expression. This is the subject of the next section.

2.4 Terms in the total energy

Now we want to express $F[\rho(\vec{r})]$ as a sum of kinetic energy, Hartree energy, exchange energy and correlation energy. The Hartree energy is the same electrostatic interaction energy seen in the HF theory. It is usual to lump the exchange and correlation energies together into a term called 'exchange-correlation energy'. This term therefore includes the the energy arising in quantum-mechanical effects acting between parallel and antiparallel spin electrons omitted in the HF formulation. The ground-state energy can then be expressed as functionals of $\rho(\vec{r})$:

$$E_g = \int d\vec{r}\rho(\vec{r})v(\vec{r}) + E_{kin}[\rho(\vec{r})] + E_{Har}[\rho(\vec{r})] + E_{xc}[\rho(\tilde{r})] \tag{23}$$

with the RHS terms defined as follows. The first term is the exact expression for the interaction of the electrons with the external potential as before. The second term is defined to be the kinetic energy of *non-interacting* electrons having the density distribution $\rho(\vec{r})$, i.e. it is not the expectation value $\langle\hat{T}\rangle$ in the Hamiltonian at equation (15) above, which applies to a system of interacting electrons. In practice the KE is obtained from notional single electron wavefunctions rather than an explicit functional, as will be seen later. The third term is the Hartree energy which appears as

$$\frac{1}{2}e^2 \int d\vec{r}d\vec{r}' \frac{\rho(\vec{r})\rho(\vec{r}')}{|\vec{r}-\vec{r}'|} \tag{24}$$

when written in atomic units in terms of electronic density. The factor $1/2$ prevents double-counting, but the density integral will include an erroneous element of self-interaction excluded from the Hamiltonian summation. The final term $E_{xc}[\rho(\vec{r})]$ is the exchange-correlation energy and defined to be that part of the energy not accounted for in the other terms $E_V = \int d\vec{r}\rho(\vec{r})v(\vec{r}), E_{kin}$ and E_{Har} Under this definition, E_{xc} includes components to compensate for the kinetic energy error mentioned above. Unfortunately, there is no known formula for calculating E_{xc} exactly so any offset is only partial. But it is possible to make good approximations to E_{xc} which we describe in the next section.

2.5 Exchange-correlation functionals

Although $E_{xc}[\rho(\vec{r})]$ has 'difficult' constituents the first HK theorem means that it is a unique and universal functional of the electronic density. This suggests a strategy for the construction of XC functionals: start with an approximate theoretical model and enhance it progressively with new levels of theory or even empirical data. An approximation for $E_{xc}[\rho(\vec{r})]$ that works surprisingly well is the Local Density Approximation (LDA) based on the exchange-correlation energy per electron $\epsilon_{xc}^0(\rho)$ of the uniform electron gas (UEG) and proposed by Kohn and Sham (1965) although their paper acknowledges earlier work by Thomas, Fermi and Slater. It was later extended to spin-polarized systems (Von Barth and Hedin, 1972). In the UEG model a system of N electrons with finite, slowly varying density moves against a background of positive charge, so that the whole system is neutral. Ignoring spin polarization, the exchange energy contribution per electron ϵ_x^0 can be written as:

$$\epsilon_x^0(\rho) = -C\rho^{\frac{1}{3}}(\vec{r}) \tag{25}$$

where C is a constant and the minus sign indicating that the exchange contribution lowers the total energy. This can be rationalized as the presence of an *exchange hole* surrounding electrons with parallel spin and a consequence of the Pauli principle. This absence of charge increases the inter-electron distance and hence lowers their interaction energy.

Unfortunately, no simple analytic form for the correlation energy $\epsilon_c^0\big(\rho(\vec{r})\big)$ has been discovered. However, its value can be accurately calculated point-by-point for a wide range of values of r_s in quantum Monte-Carlo (QMC) simulations (Ceperley and Alder 1980) and a functional form obtained by curve fitting. The following expression is an example, valid for an unpolarized gas in the high-density range $\big(0 < r_s < 1,\ r_s = [3/4\pi\rho(\vec{r})]^{1/3}\big)$ (Perdew and Zunger, 1981):

$$\varepsilon_c^0[\rho(\vec{r})] = C_1 + C_2 \ln r_s + r_s(C_3 + C_4 \ln r_s) \tag{26}$$

where ε_c^0 is the per-electron correlation energy contribution and all the C_i are constants. Another formula applies when $r_s \geq 1$. This contribution is also negative and generally significantly smaller than the exchange contribution. It can be characterized as the interaction between antiparythgallel electrons leading to a charge displacement referred to

as a *correlation hole*. The XC energy of the LDA can be calculated by multiplying $\varepsilon_{xc}^{0}[\rho(\vec{r})]$ by the local electron density and integrating over space

$$E_{xc}[\rho(\vec{r})] = \int \rho(\vec{r})\, \epsilon_{xc}^{0}[\rho(\vec{r})] d^3\vec{r} = \int \rho(\vec{r}) \left[\epsilon_{x}^{0}[\rho(\vec{r})] + \epsilon_{c}^{0}[\rho(\vec{r})]\right] d^3\vec{r} \tag{27}$$

The LDA is generally successful in predicting structures and macroscopic properties but has shortcomings e.g. a tendency to overbind, overestimating cohesive energies and underestimating lattice parameters. In a real system electron density does not vary slowly as assumed in the LDA model. Unfortunately, simply adding a partial dependency on density gradients (the *gradient expansion model* or GEA) did not yield a systematic improvement when tried. One of the reasons for this failure was the violation of the *sum rule* for the exchange and correlation holes[1]. G*eneralized gradient approximations* (GGAs) are constructions based on the LDA but include a functional dependence on the density gradient:

$$E_{xc}^{GGA}[\rho(\vec{r}), s] = \int \epsilon_{xc}^{LDA}[\rho(\vec{r})]\rho(\vec{r})F(s)d^3\vec{r} \tag{28}$$

where $F(s)$ is an enhancement factor fitted to enforce various physical constraints. Here s depends on the both the electron density and its gradient:

$$s = C\frac{|\nabla\rho(\vec{r})|}{\rho^{4/3}(\vec{r})} \tag{29}$$

GGAs are typically (but not always) more accurate than the LDA, for example reducing the error in bond dissociation energy calculations. But unlike the LDA, there is no single universal form and many GGAs have evolved, each with its own version of $F(s)$. The *Perdew-Wang 91 functional* (PW91) (Perdew, 1991; Perdew, Wang 1992) is constructed so that it satisfies the sum rule for the exchange hole, ensures that the exchange part of the density is always be negative and satisfies the Lieb-Oxford bound[2] condition (Lieb, Oxford 1981). Although PW91 depends on statistical UEG data it is nonempirical and has become a standard functional in the field of in solid state physics.

The PBE (Perdew, Burke, Ernzerhof 1996) functional employed here is a simplified and improved version of PW91 yet satisfying most of its constraints. It has proven to be accurate and efficient and is supported by the PAW pseudopotentials found in VASP. In the PBE form the exchange enhancement factor is given by:

$$F_x^{PBE}(s) = 1 + \kappa - \kappa/(1 + \mu s^2/\kappa) \tag{30}$$

where κ and μ are fitted constants. The form for correlation is expressed as the local correlation plus an additive term H:

[1] The hole is the displaced charge that forms around a point test charge. For the exchange hole, the sum of the displaced charge should the negative of the test charge. For the correlation hole, it should be zero.

[2] A lower bound on the exchange energy $E_x \geq C\int \rho^{4/3}\, d^3r$ where C lies between -1.44 and -1.68.

$$E_c^{PBE}[\rho_\uparrow(\vec{r}), \rho_\downarrow(\vec{r})] = \int \rho(\vec{r})[\epsilon_c^0(r_s, \xi) + H(t, r_s, \xi)]\, d^3\vec{r}. \tag{31}$$

Here $\xi = (\rho_\uparrow - \rho_\downarrow)/(\rho_\uparrow + \rho_\downarrow)$ is the relative spin polarization and t a scaled density gradient $t = |\nabla\rho(\vec{r})|/2\rho(\vec{r})gk_s$ where $g = \left[(1+\xi)^{2/3} + (1-\xi)^{2/3}\right]/2$ and $k_s = (4k_F/\pi)^{1/2}$ is the local screening wave vector, k_F being the Fermi wavevector. Unlike exchange energy, correlation energy is dependent on relative spin polarization and cannot be separated into spin-up and spin-down parts, complicating the formulation. An expression for H and its derivation are given in (Perdew, 1991) and (Perdew, Burke, Ernzerhof 1996).

There are other classes of XC functional, including *meta-GGA* (containing a dependency on second-order density gradients) and *hybrid* (a fraction of the exchange energy is derived from an orbital-dependent HF calculation) both of which consume more (for hybrids, typically tenfold) processing power than GGAs.

However, all functionals are approximations and none is accurate in all properties of interest. No matter what functional is invented, someone will always find a case where it fails. A notable example is the underestimation of the band gap in crystalline Si, observed in all LDA and GGA functionals. This is returned to in the next chapter.

2.6 The Kohn-Sham equation

Following the publication of the Hohenberg-Kohn theorems Kohn and Sham provided a practical means to exploit them (Kohn, Sham 1965). The Kohn-Sham ansatz is that the exact ground state density can be written as the ground state density of a fictitious system of noninteracting particles. The ground state is then determined by minimising the total energy with respect to $\rho(\vec{r})$ while holding the total number electrons constant. This is done by taking the ground state density $\rho_g(\vec{r})$ and making a small change so that $\rho(\vec{r}) = \rho_g(\vec{r}) + \delta\rho(\vec{r})$, and requiring the resulting change δE is zero be zero to the first order in $\delta\rho(\vec{r})$. Using the total energy expression (23) we can write the total energy as:

$$E = \int d^3\vec{r}\rho(\vec{r})v(\vec{r}) + T[\rho(\vec{r})] + G[\rho(\vec{r})] \tag{32}$$

where $T = E_{kin}[\rho(\vec{r})]$ and $G = E_{Har}[\rho(\vec{r})] + E_{xc}[\rho(\vec{r})]$. Then:

$$\delta E = 0 = \int d^3\vec{r}\left[v(\vec{r}) + \frac{\delta T}{\delta\rho(\vec{r})} + \frac{\delta G}{\delta\rho(\vec{r})}\right]\delta\rho(\vec{r}) \tag{33}$$

subject to the constraint:

$$\int d^3\vec{r}\delta\rho(\vec{r}) = 0. \tag{34}$$

This notation employs functional derivatives[3], e.g. $\delta G/\delta\rho(\vec{r})$. For (33) to be true for any arbitrary $\delta\rho(\vec{r})$ satisfying (34) the ground-state condition on $\rho(r)$ is:

[3] See Appendix A.3

$$\mu = \frac{\delta T}{\delta \rho(\vec{r})} + \frac{\delta G}{\delta \rho(\vec{r})} + v(\vec{r}) \tag{35}$$

where μ is a Lagrange undetermined multiplier[4]. The DFT 'trick' is to write:

$$\frac{\delta G}{\delta \rho(\vec{r})} + v(\vec{r}) = v_{eff}(\vec{r}) \tag{36}$$

so that the ground state condition becomes:

$$\frac{\delta T}{\delta \rho(\vec{r})} + v_{eff}(\vec{r}) = \mu \tag{37}$$

and to recall that T was defined as the kinetic energy of non-interacting electron systems where $G = 0$, since G is the sum of the Hartree and exchange-correlation energies and both of these vanish for non-interacting electrons. So, we have from (36) $v(\vec{r}) = v_{eff}(\vec{r})$ and

$$\frac{\delta T}{\delta \rho(\vec{r})} + v(\vec{r}) = \mu \tag{38}$$

and we see that the ground-state density of the interacting system in the electrostatic field $v(\rho)$ is identical to the ground-state density of the non-interacting system in the field $v_{eff}(\rho)$. Because the kinetic energy is now easy, we can solve the SE:

$$\left[-\frac{\hbar^2}{2m}\nabla^2 + v_{eff}(\vec{r})\right]\psi_n(\vec{r}) = \epsilon_n \psi_n(\vec{r}), \tag{39}$$

where the eigenvalues ϵ_n appear in place of μ. Then the lowest $N/2$ states $\psi_n(\vec{r})$ can be filled with a spin-up and spin-down electron and the density obtained as:

$$\rho(\vec{r}) = 2\sum_{i}^{occ} |\psi_i(\vec{r})|^2 \tag{40}$$

The equation (39) is a set of coupled one-electron orbital SEs in the same form seen in HF theory. As in HF theory the nuclei are fixed and electronic energy is minimized in a SCF scheme, resulting in a new electronic density distribution. A geometrical optimization may follow, in which the nuclei are moved classically to lower-energy configurations. The entire process is repeated until the inter-atomic forces acting on each atom are acceptably close to zero and the position on the PES is a stationary point. Some further detail is provided at page 48 below.

In the DFT context (39) is usually called the *Kohn-Sham equation* and the effective potential v_{eff} the *Kohn-Sham potential*, denoted by v_{KS}. However, apart from the highest energy level (corresponding to ionization energy) the Kohn-Sham eigenstates $\psi_i(\vec{r})$ and eigenvalues ϵ_i do not have a physical meaning. The electron density $\rho(\vec{r})$ is the only variable with a physical reality.

[4] See Appendix A.4, where the object function f replaces E and constraint ϕ replaces ρ.

2.7 The Kohn-Sham potential

To calculate the KS potential appearing in (39) the functional derivative $\delta G/\delta\rho(\vec{r})$ must be known. Since $G = E_{Har}[\rho(\vec{r})] + E_{xc}[\rho(\vec{r})]$:

$$\frac{\delta G}{\delta\rho(\vec{r})} = \frac{\delta E_{Har}}{\delta\rho(\vec{r})} + \frac{\delta E_{xc}}{\delta\rho(\vec{r})}. \tag{41}$$

But $\delta E_{Har}/\delta\rho(\vec{r})$ is the Hartree potential:

$$\frac{\delta E_{Har}}{\delta\rho(\vec{r})} = \int dr' \frac{\rho(\vec{r}')}{|\vec{r} - \vec{r}'|} = v_{Har}(\vec{r}) \tag{42}$$

and similarly, $\delta E_{xc}/\delta\rho(r)$ is an exchange-correlation potential. If the LDA were adopted, then:

$$\frac{\delta E_{xc}}{\delta\rho(\vec{r})} = \frac{d}{d\rho(\vec{r})}\left[\rho(\vec{r})\epsilon_{xc}^{0}\left(\rho(\vec{r})\right)\right] = v_{xc}(\vec{r}) \tag{43}$$

using whatever formula has been adopted for ϵ_{xc}^{0}.

The KS potential is therefore:

$$v_{KS}(\vec{r}) = v(\vec{r}) + v_{Har}(\vec{r}) + v_{xc}(\vec{r}) \tag{44}$$

i.e. the sum of the external (ionic) potential $v(\vec{r})$, the Hartree potential $v_{Har}(\vec{r}\,)$ and the exchange-correlation potential $v_{xc}(\vec{r})$. Although the KS equation appears to treat the system using one-electron orbitals (as though correlation did not exist) correlation is being included via the exchange-correlation potential $v_{xc}(\vec{r})$.

2.8 Periodic supercells and Bloch's theorem

The preceding sections have shown how the ground-state energy of an N-electron system can be found by mapping it into an equivalent observable in a single-particle system. But for crystalline silicon we still need to represent wavefunctions extending over the entire solid for an infinitely large number of electrons. This is done by recognizing the periodic nature of the solid and this section introduces some terminology used to discuss periodic systems (Ashcroft; Mermin, 1976).

If $\hat{a}_{1,}\hat{a}_2$ and $\hat{a}_3$ are three linearly independent vectors then a *Bravais lattice* consists of all the points with position vector $\vec{R}$, a linear combination of the *lattice vectors* $\hat{a}_{1,}\hat{a}_2$ and $\hat{a}_3$ such that $\vec{R} = \sum_{i=1}^{3} L_i\hat{a}_i$ and L_i are integers in the range minus infinity to infinity.

A *primitive cell* is a volume that exactly fills the entire space when translated through all the vectors of the Bravais lattice. If the cell exactly fills the space when translated through a subset of the lattice points it is called a *unit cell* or *supercell*.

The *reciprocal lattice* of a Bravais lattice is the set of wave vectors $\{\vec{G}\}$ that have the same periodicity as the original lattice i.e.

$$e^{i\vec{G}.(\vec{r}+\vec{R})} = e^{i\vec{G}.\vec{r}} \tag{45}$$

or

$$e^{i\vec{G}.\vec{R}} = 1 \tag{46}$$

The reciprocal lattice is itself a Bravais lattice. The reciprocal lattice of the reciprocal lattice is the original (sometimes called *direct*) lattice.

The *first Brillouin zone* (BZ) is the smallest volume entirely enclosed by planes that are perpendicular bisectors of the reciprocal lattice vectors drawn from the origin. This definition (Kittel, 2008) is analogous to that of the Wigner-Seitz cell in real space. It is a primitive cell of the reciprocal lattice having its translational symmetry and can be taken to be the lattice cell containing the point $\vec{G} = 0$.

We can assume that the KS potential $v_{KS}(r)$ will also be periodic with the periodicity of the supercell:

$$v_{KS}(\vec{r}) = v_{KS}(\vec{r} + \vec{R}). \tag{47}$$

Bloch's theorem states that in a periodic potential each electronic wavefunction can be written as the product of a cell-periodic part and a wavelike part:

$$\psi_{\vec{k}}(\vec{r}) = e^{\mathrm{i}\vec{k}.\vec{r}}\, u_{\vec{k}}(\vec{r}) \tag{48}$$

where $\vec{k}$ is a wavevector and $u_k(r)$ has the same periodicity as the supercell. The implication of this is that the eigenfunctions fall into classes, each class having a particular wavevector $\vec{k}$ and that it is possible to solve the SE for each value of $\vec{k}$ independently. This is an important result in solid state physics, which surprised Bloch when he discovered it in Vienna in 1927. However, the physicist Werner Heisenberg was on hand and confirmed Bloch's calculation (Hoddeson et al., 2007).

Since $u_{\vec{k}}(\vec{r})$ is periodic it can be expressed as a Fourier series:

$$u_{\vec{k}}(\vec{r}) = \sum_{\vec{G}} c_{\vec{k}+\vec{G}} e^{\mathrm{i}(\vec{k}+\vec{G}).\vec{r}} \tag{49}$$

The periodic boundary conditions

$$\psi(\vec{r} + L_i\vec{a}_i) = \psi(\vec{r}) \tag{50}$$

where L_i is an integer number, together with Bloch's theorem (48) imply that electronic states are allowed only at some of the $\vec{k}$ vectors in the periodic system:

$$\psi_k(\vec{r} + L_i\hat{a}_i) = e^{iL_i\vec{k}.\hat{a}_i}\psi(\vec{r}), \qquad i = 1,2,3. \tag{51}$$

This requires

$$e^{iL_i\vec{k}.\hat{a}_i} = 1. \tag{52}$$

If $\vec{k}$ is expressed as a linear combination of the reciprocal lattice vectors

$$\vec{k} = \sum_i x_i \vec{b}_i \tag{53}$$

where $\hat{b}.\hat{a} = 2\pi\delta_{ij}$, and this expression substituted in (52) which becomes:

$$e^{2\pi i L_i x_i} = 1 \tag{54}$$

and so, the only allowed $\vec{k}$ vectors are of the form

$$\vec{k} = \sum_{i=1}^{3} \frac{m_i}{L_i}\hat{b}_i \tag{55}$$

where m_i must be an integer.

However, since $\vec{R}$ is infinitely large there will still be an infinitely large number of $\vec{k}$ points in the first BZ. But the electronic wavefunctions at $\vec{k}$ points which are close together will be almost identical and the electronic wavefunctions over a region of $\vec{k}$ space can be replaced by the wavefunction at a single point. Moreover, at each $\vec{k}$ point we need consider only the energies of occupied KS orbitals corresponding to the Lagrange multipliers of equation (34) above, as these define the ground state energy. However, a practical calculation will usually include a small number of unoccupied orbitals. In any event, we can anticipate a feasible calculation involving a finite number of electronic states and finite number of $\vec{k}$ points.

A frequently used procedure for acquiring the location of $\vec{k}$ points is the Monkhorst-Pack (MP) mesh (Monkhorst and Pack, 1976). According to the MP prescription, the set of $\vec{k}$ points is obtained from the formula:

$$k_{ijk} = w_i\hat{b}_1 + w_j\hat{b}_2 + w_k\hat{b}_3, \qquad w_r = \frac{(2r - q_r - 1)}{2q_r}, r = 1,2,3,\dots,q_r \tag{56}$$

and q_r is an integer that determines the number of points in the set in the r-direction of the reciprocal axes. With this construction the set contains $(q_1 \times q_2 \times q_3)\,\vec{k}$ points uniformly spaced in the Brioullin zone. If the $q_i$are all odd, the set contains the reciprocal image of the real-space origin, i.e. the Γ point. The $\vec{k}$ point mesh is sufficiently dense for a calculation if the total energy has converged with respect to the number of $\vec{k}$ points. A denser mesh is needed for metals than other materials, to overcome discontinuities in the Fermi surface. However, this is a numerical integration characteristic rather than a material property.

2.9 Plane wave basis sets

Although the wavefunction integrations in reciprocal space are now (in principle) tractable, the wavefunction expansion (49) still contains an infinite number of plane wave terms and is not a basis for practical calculations. However, these terms have a simple interpretation in terms of Schrodinger's equation: they have kinetic energy E

$$E = \frac{\hbar^2}{2m_e}\left|\vec{k} + \vec{G}\right|^2. \tag{57}$$

The infinite sums can be truncated by removing all terms with kinetic energy greater than some *cut-off* value:

$$E_{cut-off} = \frac{\hbar^2}{2m_e}\left|\vec{G}_{cut-off}\right|^2 \tag{58}$$

with the justification that terms with lower energies will be more important physically than those with very high energies. The higher energy terms are plane waves that are oscillating on short length scales in real space. These are associated with the tightly bound core electrons that have little effect on the chemical environment, which is determined by the valence electrons. The infinite sum then reduces to:

$$\psi_{\vec{k}}(\vec{r}) = \sum_{\vec{G}<\vec{G}_{cut-off}} c_{\vec{k}+\vec{G}} e^{i(\vec{k}+\vec{G}).\vec{r}} \tag{59}$$

and these plane wave expansions can be substituted into the KS equation (39) and eventually determine the ground-state energy for the system of atoms. The cut-off energy $E_{cut-off}$ appears as an external, user-defined parameter of the calculation. The more $E_{cut-off}$ is increased, the greater the variational freedom over the orbitals in search of the ground state. Therefore, an increase in $E_{cut-off}$ must yield a decrease in the calculated ground state energy. In practice the ground state energy is found to converge, so that increasing values of $E_{cut-off}$ do not enhance the overall accuracy of the calculation. This relative ease of systematic improvement contrasts with the hierarchy of Gaussian basis sets found in HF calculations, each based on a distinct level of theory.

But the plane wave basis set is generally large. Under a Fourier transform a length scale δ in real space becomes $q = 2\pi/\delta$ in reciprocal space. In a solid the number of plane waves N_G will be:

$$N_G \sim \frac{4}{3}\pi q^3 \times \frac{1}{\Omega_{BZ}} \tag{60}$$

i.e. the volume of a sphere of radius q divided by $\Omega_{BZ} = (2\pi)^3/\Omega$ is the volume of the Brillouin zone and Ω the equivalent volume in the real-space lattice. An estimate can be made from fig 2.1, which shows the spatial extent of the electronic wavefunctions of the H atom in atomic units (i.e. Bohr radius ~ 0.5 Å). These decay exponentially with nuclear distance, but remain significant at ~ 4 a.u. Furthermore, the $1s$ wavefunction should be sampled on an interval $\delta \sim 0.1$ a.u. if curvature near the nucleus is to be represented accurately. Scaling to obtain an estimate of N_G for the silicon lattice (where $a_0 \sim 5$ a.u.) and substituting in (60) gives:

$$N_G \sim \frac{4\pi}{3} \times \left(\frac{2\pi}{0.1}\right)^3 \times \left(\frac{5}{2\pi}\right)^3 \sim 10^5 \tag{61}$$

A basis set of this size is impractical, so the potentials and wavefunctions must be transformed to remove the influence of core states. Without these rapidly varying components, the smoother valence wavefunctions can be represented with a larger sampling interval, leading to a smaller basis set. This is the motivation for the *pseudopotential method*, the subject of the next section.

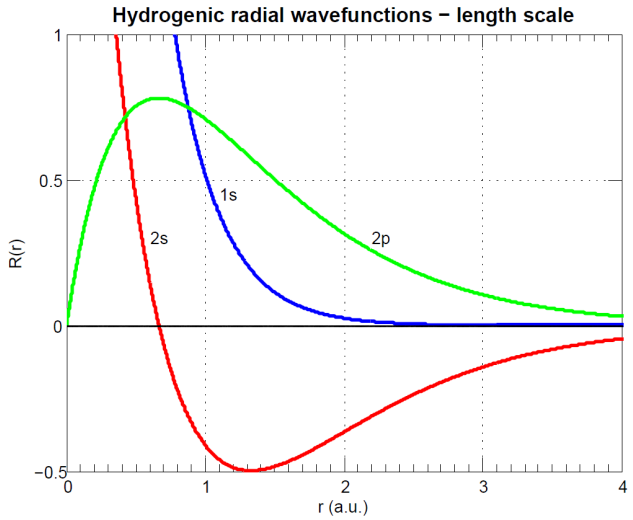


Fig 2.1. Hydrogenic 1s, 2s and 2p wavefunctions in atomic units (1 a.u. ≈ 0.5 Å), computed from Legendre polynomials. Their gradients and spatial extent determine the number of plane waves needed for an accurate representation in reciprocal space (author's MATLAB® image).

2.10 Pseudopotential method

In order to achieve convergence with a manageable size of plane wave basis set, the strong electron-ion interaction must be replaced with a sufficiently weak simulated potential, or *pseudopotential*. The basic idea of the pseudopotential method is to project the SE for the valence electrons onto the subspace orthogonal to the core orbitals, eliminating the nodal structure of the valence orbitals close to the core but without modifying them in the interstitial region where chemical bonding occurs. Simulated potentials that retain the core electrons have been devised e.g. FLAPW (Blügel and Bihlmayer, 2006) but these are complex and demand considerable expertise in their usage. However, these all-electron methods can deliver the highest accuracy and set the standard against which others are judged.

Several pseudopotential schemes have been proposed, to satisfy the conflicting requirements of accuracy, transferability and computational efficiency. A description of the so-called *norm-conserving* pseudopotential method follows, which is known to result in 'hard' (requiring a very large basis set) pseudopotentials when applied to the first-row elements and systems with *d* or *f* electrons. The *projector augmented-wave* (PAW) scheme, which combines plane waves in the interstitial region with spherical waves around the core, is also described. PAWs take the core electrons into account but avoid the need for large basis sets. It is the preferred scheme in the VASP package.

2.10.1 Removal of core electrons

This is based on the observation that core orbitals have a limited spatial extent and much lower energy than the valence orbitals. They are therefore relatively unaffected by the chemical environment of the valence orbitals and can be held fixed (*frozen*) during energy minimization. Freezing the core electrons reduces the calculation size but they cannot be discarded because without them the valence orbitals would collapse towards the nucleus as minimization proceeded, since the requirement that they remain orthogonal to the core orbitals would no longer exist. It turns out that when the depth of the nuclear potential well is artificially reduced (forming a pseudopotential) the lowest bound states of the valence electrons can have energies identical to those found in the true potential field when the lowest bound states are occupied by core electrons, as shown in fig 2.2. This can be rationalized through the removal of the nuclear screening afforded by the core electrons. However, the adjustment in well depth is different for valence *s*, *p* and *d* states. The dependence on angular momentum is called *non-locality.* Nevertheless, the shape of the valence orbitals beyond the core radius is unaltered by the change in potential, so the DFT variational procedure still converges even in the absence of the core electrons, and the energetics of the whole system are correctly reproduced.

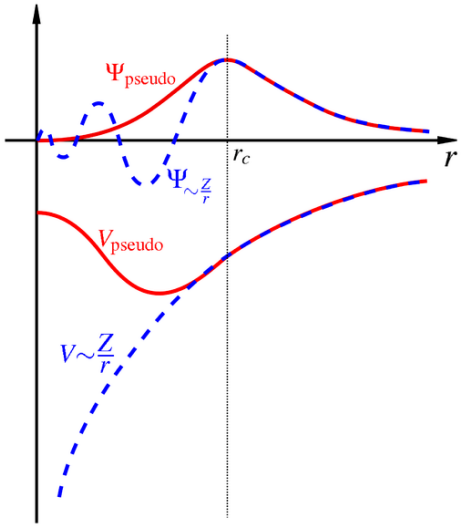


Fig 2.2. Comparison of a wavefunction in the Coulomb potential of the nucleus (blue) to the one in the pseudopotential (red). The real and the pseudo wavefunction and potentials match above a certain cut-off radius r_c called the core radius (image by W Quester / CC by 2.0)).

2.10.2 Transferability and norm conservation

In addition to correctly reproducing valence eigenvalues and long-range atomic wavefunctions, a 'good' pseudopotential should be usable in a variety of chemical environments without change. It has been found that *transferability* improves when two further conditions are imposed on the pseudopotentials and pseudoorbitals.

The first, *norm conservation* requires that the integrals from 0 to r of the real and pseudo charge densities agree for $r = r_c$, for each valence state (r_c is the core radius). This guarantees (by Gauss's law) that the potential field produced outside r_c is the same for the

real and pseudo charge distributions. The second condition is that the logarithmic derivatives of the real and pseudo wavefunctions and their first energy derivatives agree for $r = r_c$. This ensures the quantum scattering properties of the ionic potential wells are reproduced with minimum error as bonding or banding shifts eigenenergies away from the atomic levels.

Fortunately, the two conditions are equivalent so that if a pseudopotential is norm-conserving it will automatically meet the other condition. Norm conservation can be expressed as:

$$\int_0^{r_c} dr r^2 \psi^2(r) = \int_0^{r_c} dr r^2 \chi^2(r) \tag{62}$$

where $\psi(r)$ is the valence wavefunction given by an all-electron calculation in a real potential field and $\chi(r)$ the corresponding valence pseudo wavefunction. As noted above $\psi(r)$, $\chi(r)$ and r_c are evaluated independently for each angular momentum quantum number. These integrals are related to the logarithmic derivatives by the identity (Hamann, 1979):

$$2\pi \left[(r\psi)^2 \frac{d}{d\varepsilon} \frac{d}{dr} \ln(\psi) \right]_{r_c} = 4\pi \int_0^{r_c} \psi^2 r^2 dr \tag{63}$$

For the best accuracy r_c should be as small as possible but for some elements (e.g. third-row transition metals) this would mean the inclusion of fluctuating semicore orbitals in the pseudo wavefunction. The resulting pseudopotential is termed *hard* and requires large basis sets and high cut-off energy energies, making calculations relatively expensive. The norm-conserving condition can be relaxed, leading to *ultrasoft* pseudopotentials having a smoother, more economical wavefunction. These were first derived by Vanderbilt (1990). The correction for the missing charge is handled by placing an augmentation charge density inside the core region.

2.10.3 Projector augmented-wave method

The projector augmented-wave scheme (PAW) was first proposed by Blöchl (1994) and later shown to be related to the ultrasoft scheme by Kresse and Joubert (1999). The PAW pseudopotentials are implemented in VASP and preferred by its authors over the original ultrasoft types. It is motivated by the desire to reproduce the oscillatory behaviour of valence electron wavefunctions (near the ionic cores) without the need for a fine sampling grid and large basis set.

In the PAW method the required wavefunction $|\psi\rangle$ (i.e. the true single-particle all-electron KS wavefunction) is mapped by a linear transformation from a fictitious *auxiliary wavefunction* $|\tilde{\psi}\rangle$, which is smooth and hence converges quickly when expanded in a plane wave basis. Physical properties are evaluated by reconstructing the true wavefunctions. The operator for a back transformation from the auxiliary to the true wavefunctions is defined as:

$$\hat{T} = 1 + \sum_R \hat{S}_R, \qquad |\psi\rangle = \hat{T}|\tilde{\psi}\rangle \tag{64}$$

where the summation runs over atomic sites. The operator $\hat{S}$ acts only within a spherical augmentation region Ω_R surrounding site R, so $|\tilde{\psi}\rangle$ is modified only within the augmentation region and both wavefunctions are equal outside the region. The augmentation region corresponds to the core region of other pseudopotential schemes. Inside Ω_R the true valence wavefunctions can be represented in a basis of *partial waves* ϕ_i^R, typically solutions of the KS Schrodinger equation for an isolated atom, i.e. the product of a radial wave function and a spherical harmonic. The index i is needed to index the angular momentum quantum numbers. For each of these partial waves we define a corresponding smooth *auxiliary partial wave* $\tilde{\phi}_i^R$, such that:

$$|\phi_i^R\rangle = (1 + \hat{S}_R)|\tilde{\phi}_i^R\rangle \Leftrightarrow \hat{S}_R|\tilde{\phi}_i^R\rangle = |\phi_i^R\rangle - |\tilde{\phi}_i^R\rangle \tag{65}$$

for all i, R. This completely defines $\hat{T}$, given ϕ and $\tilde{\phi}$.

Outside Ω_R $\hat{S}_R$ should do nothing so (65) implies that the auxiliary partial waves are equal to the partial waves. Inside, they can be any smooth continuation, e.g. a linear combination of polynomials or Bessel functions. Because the operator $\hat{T}$ is linear, the expansion coefficients c_i can be written as an inner product with a set of so-called *projector functions* $|\tilde{p}_i\rangle$:

$$c_i = \langle \tilde{p}_i^R | \tilde{\psi} \rangle. \tag{66}$$

It can then be shown that (Rostgaard, 2009):

$$\hat{T} = 1 + \sum_R \sum_i \left(|\phi_i^R\rangle - |\tilde{\phi}_i^R\rangle\right) \langle \tilde{p}_i| \tag{67}$$

so that the all electron KS wavefunction $|\psi\rangle$ can be obtained from the transformation:

$$|\psi\rangle = |\tilde{\psi}\rangle + \sum_R \sum_i \left(|\phi_i^R\rangle - |\tilde{\phi}_i^R\rangle\right) \langle \tilde{p}_i^R | \tilde{\psi} \rangle \tag{68}$$

This decomposition separates the original wavefunctions into auxiliary wavefunctions which are smooth everywhere and an oscillating contribution confined to the augmentation spheres. These are treated separately, with the localized part (indicated by the superscript R) represented on atom-centred radial grids. Smooth functions are indicated by a tilde $\sim$. The delocalized parts (no superscript R) are all smooth and so can be represented on coarse Fourier or real-space grids.

The PAW method retains the frozen core approximation and the core electrons are decomposed in a similar way. However, no projector functions are necessary in this case.

2.11 Calculations in reciprocal space

In previous sections it was shown that the key problem of DFT calculations is the solution of the KS equations (39) and (44) i.e.

$$\widehat{H}_{KS}(r)\psi_i(\vec{r}) = \left[-\frac{\hbar^2}{2m}\nabla^2 + v_{KS}[\rho](\vec{r})\right]\psi_i(\vec{r}) = \epsilon_i\psi_i(\vec{r}),$$

$$v_{KS} = v_{PS}(\vec{r}) + v_{Har}[\rho](\vec{r}) + v_{XC}[\rho](\vec{r})$$

where v_{PS} now denotes the external potential as derived from a pseudopotential approximation as discussed in the preceding section. Solutions are obtained by expanding the KS wavefunctions ψ_i in a basis set and applying a variational procedure to determine the basis set coefficients. The following sections show how these equations are converted into matrix form and solved by numerical optimization methods.

2.11.1 The KS Hamiltonian matrix

Each wavefunction can be expanded in a plane wave basis set (59)

$$\psi_{i,\vec{k}}(\vec{r}) = \sum_{\vec{k}+\vec{G}<\vec{G}_{cut}} c_{\,i,\vec{k}+\vec{G}}\, e^{\mathrm{i}(\vec{k}+\vec{G}).\vec{r}}$$

where the coefficients c are refined at each iteration. Substituting this expansion into the KS equation results in a matrix eigenvalue equation in reciprocal space:

$$\sum_{\vec{G}'} H_{\vec{k}+\vec{G},\vec{k}+\vec{G}'}c_{i,\vec{k}+\vec{G}'} = \varepsilon_i c_{i.\vec{k}+\vec{G}} \tag{69}$$

where $H_{\vec{k}+\vec{G},\vec{k}+\vec{G}'}$ is the matrix element of the Hamiltonian $\widehat{H}_{KS}$ between states $\vec{k}+\vec{G}, \vec{k}+\vec{G}'$. $c_{\mathrm{i},\vec{k}+\vec{G}}$ and ε_i are the eigenvectors and eigenvalues for the discrete set of solutions of the matrix equations labelled $i = 1,2,\ldots$ for a given $\vec{k}$. The matrix elements can be shown to be:

$$H_{\vec{G},\vec{G}'}(\vec{k}) = \langle\vec{k}+\vec{G}|\widehat{H}_{KS}|\vec{k}+\vec{G}'\rangle$$

$$= \frac{1}{2}\left|\vec{k}+\vec{G}\right|^2\delta_{\vec{G},\vec{G}'} + v_{PS}(\vec{G}-\vec{G}') + v_{Har}\,(\vec{G}-\vec{G}') + v_{XC}(\vec{G}-\vec{G}'). \tag{70}$$

(69) and (70) are the basic Schrodinger equations in a periodic crystal and support the practical calculations to be described in the following sections. The equations show that the SE must be solved for each $\vec{k}$ separately and that the eigenvalues and eigenvectors are independent unless they differ by a reciprocal lattice vector. At each $\vec{k}$ point the eigenstates (again labelled $i = 1,2,\ldots$) may (in principle) be found by diagonalizing the Hamiltonian matrix (70) in the basis of the Fourier components $\vec{k}+\vec{G}$. In practice, the matrix is too large to diagonalize by conventional methods and an approximate, partial diagonalization is achieved using iterative numerical optimization.

2.11.2 Fast Fourier Transforms

In reciprocal space the kinetic energy term in (70) is diagonal and easy to obtain, but the exchange correlation and Hartree potentials v_{XC} and v_{Har} are defined in terms of the real space electronic density $\rho(r)$. This density is given by (40) above:

$$\rho(\vec{r}) = 2\sum_{occ}|\psi_n(\vec{r})|^2 \tag{71}$$

or

$$\rho(\vec{r}) = \frac{2}{\Omega}\sum_{occ}\sum_{\vec{G},\vec{G}'} c^*_{\vec{G}}c_{\vec{G}'}\, e^{i(\vec{G}-\vec{G}').\vec{r}} \tag{72}$$

and

$$\rho(\vec{G}) = \frac{2}{\Omega}\sum_{occ}\sum_{\vec{G},\vec{G}'} c^*_{\vec{G}}c_{\vec{G}'} \tag{73}$$

in terms of the plane wave coefficients, where $\Omega = \hat{a}_1.(\hat{a}_2 \times \hat{a}_3)$ is a normalising factor equal to the volume of the originating real space cell. Unfortunately, to obtain the electronic density using (73) would need $N^2_{\vec{G}}$ evaluations of the basis coefficients c, a relatively large computational overhead. However, it is possible to apply a fast Fourier transform (FFT) to get the wavefunctions in real space and form a simple point product to get the electronic density. Then the inverse transform can be applied to return to reciprocal space. Crucially, the computational cost of FFT scales as $N_G \log N_G$ and is small compared to matrix diagonalization, although some efficiency is lost in parallel computing architectures.

Because the supercell is periodic the reciprocal vectors $\vec{G}$ are discrete with constant spacing. Imposition of a kinetic energy cut-off imposes a maximum value on $|\vec{G}|$ resulting in a limited number of $\vec{G}$, and these define a grid in reciprocal space. The grid should enclose the sphere of $\vec{G}$ vectors so it can reproduce any function that can be expressed as a linear combination of the G vectors. Any excess will be orthogonal to those inside the $\vec{G}$ vector sphere and represent energies above the cut-off value.

Although the electronic density ρ is a real quantity (73) shows that it arises from the inner products of complex vectors with moduli in the range $(0 \rightarrow |\vec{G}_{cut}|)$. It will therefore be a function containing vectors with moduli in the range $(0 \rightarrow 2|\vec{G}_{cut}|)$. If the density is to be represented without loss of accuracy, a grid spacing at least twice as dense as that required to represent the plane wave expansion of the KS wavefunctions is needed. For the transformation to be invertible the same mesh density is required in real space, although the spacing is different.

The role of FFT in the calculation of the Hartree energy is shown schematically in fig. 2.3 below. The same idea can be applied to the other energy components, but the detail will depend on the pseudopotential approximation and XC functional adopted for the calculation.

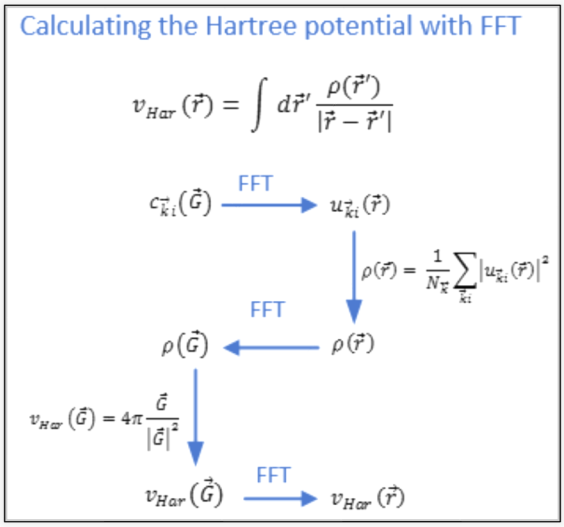


Fig 2.3. The Hartree energy is difficult to calculate in real space, because the integrand contains the singular factor $1/(\vec{r} - \vec{r}')$. Schematic shows the FFT action on the reciprocal space wavefunction coefficients $c_{\vec{k}i}$ producing the equivalent real space coefficients $u_{\vec{k}i}$. The electronic density observable is then calculated in real space and projected back to reciprocal space where the Hartree integral takes on a simple form, via Poisson's equation. The result is then returned to real space for inclusion in the KS potential. The FFT operations are inexpensive to compute, imposing an overhead scaling as $O(N \log N)$, where N is the number of grid points.

2.12 Electronic optimization

In the preceding sections it has been shown that solution of KS equation (39) requires that the inverse of the Hamiltonian matrix H in equation (69) be successively approximated within an SCF procedure. The objective is to find the electronic density distribution that minimizes the total energy, which will then approximate the ground-state energy. Arriving at this *self-consistent* density is the core problem of DFT and is often referred to as *electronic optimization* or *static relaxation*, as the atoms remain stationary throughout. The matrix coefficients H_{ij} are derived from the basis set coefficients, and for matrices of reasonable size $(\text{say } N \leq 500)$ diagonalization could be done exactly e.g. by the Cholesky-Householder (CH) procedure. However, the computation time scales as N^3 irrespective of the number of eigenvalues returned. In the DFT setting this is a real difficulty, since the basis set consists of many $(N \approx 10^{4-5})$ plane waves and a CH calculation would be impractical. Moreover, it would yield N eigenvalues when a much smaller number (roughly one half the number of occupied orbitals plus a few unoccupied) are sought. Consequently, plane wave energy calculations adopt iterative optimizing techniques that avoid treating the KS matrix in its entirety.

There are a several of these each with its own claims e.g. *blocked Davidson* and *residual minimization/direct inversion in the iterative subspace (RMM/DIIS)* (for a comparative review see Woods et al., 2019). They vary in their efficiency, but all yield the same result: the ground state total energy. Studies (Science, March 2016) confirm that calculations in solids yield

energies differing by as little as ~ 1 meV per atom irrespective of the choice of algorithm and pseudopotential. Their significance arises because static relaxation is the rate-limiting factor in a DFT calculation and ultimately determines the effectiveness of the entire method. Some schemes (e.g. RMM-DIIS) show superior scaling on multiple-core machines. We describe a generic Quasi-Newton (QN) algorithm included in the VASP package (Kresse et al., 2009) which is reasonably efficient and simpler to present than some other schemes (Kresse, Furthmüller, 1996).

Most iterative algorithms build an expansion set of trial vectors $\{|b_i\rangle, i = 1, \dots, N_a\}$ from which the best approximation to the exact eigenvectors and eigenvalues is calculated. This set is much smaller than the number of plane waves $N_a \ll N_g$ and usually slightly exceeds the number of occupied bands N_b. All algorithms generate *correction vectors* which are combined with the trial vectors in various ways. A *sequential* algorithm optimizes one band at a time whereas a *blocked* algorithm can process several bands at each step. In this sequential scheme, the expansion set is primed with a guess at $|b_1\rangle$.

A key quantity is the *residual vector* $\vec{R}$

$$|R(\phi_n)\rangle = \left(H - \epsilon_{app}\right)|\phi_n\rangle \tag{74}$$

where ϵ_{app} and ϕ_n are an approximate eigenpair solution of the SE $H\phi = \epsilon\phi$ that minimizes the *Rayleigh quotient*

$$\epsilon_{app} = \frac{\langle\phi_n|H|\phi_n\rangle}{\langle\phi_n|\phi_n\rangle} \tag{75}$$

which is stationary at the exact solution. The norm $\langle R|R\rangle$ can be taken as a measure of the error in the residual vector. The strategy adopted by iterative methods is find a vector increment $|\delta\phi_n\rangle$ which, if added to $|\phi_n\rangle$ yields a very small (ideally zero) residual vector, i.e.

$$|R(\phi_n + \delta\phi_n)\rangle = |R(\phi_n)\rangle + \left(H - \epsilon_{app}\right)|\delta\phi_n\rangle = 0. \tag{76}$$

and $\left|\bar{\phi}_n\right\rangle = |\phi_n\rangle + |\delta\phi_n\rangle$ results in a minimum residual vector and satisfies

$$\left(H - \epsilon_{app}\right)\left|\bar{\phi}_n\right\rangle = 0 \tag{77}$$

i.e. $\left|\bar{\phi}_n\right\rangle$ is an eigenvector of H. Unfortunately, the formal solution

$$|\delta\phi_n\rangle = -\left(H - \epsilon_{app}\right)^{-1}|\phi_n\rangle = 0 \tag{78}$$

is no easier to solve than the diagonalization of H as it requires matrix inversion. So, we seek a matrix K (referred to as the *preconditioning matrix*) which, when multiplied with the residual vector, produces an approximate correction vector:

$$|\delta\phi_n\rangle = K|R\rangle. \tag{79}$$

K might be derived by simply assuming that $\left(H - \epsilon_{app}\right)$ is diagonal so that

$$K = -\sum_{\vec{q}} \frac{|\vec{q}\rangle\langle\vec{q}|}{\langle\vec{q}|H - \epsilon_{app}|\vec{q}\rangle} \tag{80}$$

where $\vec{q}$ runs over all the plane waves in the basis set. This diagonal approximation to $|\delta\phi_n\rangle$ could be a basis for iteration: $|\phi_n\rangle$ would be replaced by $|\phi_n^{new}\rangle = |\delta\phi_n\rangle + |\phi_n\rangle$ and ϵ_{app} by $\epsilon_{app}^{new} = \langle\phi_n^{new}|H|\phi_n^{new}\rangle / \langle\phi_n^{new}|\phi_n^{new}\rangle$, which is called a *Newton step* after its resemblance to the Newton-Raphson iterative procedure for finding the zeros of function. However, the process is not guaranteed to converge for an arbitrary Hamiltonian H and some further refinement is needed.

In a sequential algorithm it is usual to make a new correction vector orthogonal to the current expansion set, by removing the projections of the existing vectors along the new vector:

$$|\delta\phi_n^{\perp}\rangle = \left(|\delta\phi_n\rangle - \sum_m |\phi_m\rangle\langle\phi_m|\,\delta\phi_n\rangle\right) \tag{81}$$

where the expression in brackets is the *Gram-Schmidt* orthogonalization formalism[5] with m ranging over the expansion set. Now the preconditioning matrix is $(1 - \sum_m |\phi_m\rangle\langle\phi_m|) \times K$ and iteration sequence

$$|g_n\rangle = \left(1 - \sum_m |\phi_m\rangle\langle\phi_m|\right) \times KR \tag{82}$$

should generate a preconditioned residual vector $|g_n\rangle$ orthogonal to all its predecessors at each step. In the QN iteration, $|g_n\rangle$ is used to define a small auxiliary eigenvalue problem

$$\langle b_i|H - \varepsilon|b_j\rangle = 0 \tag{83}$$

with basis set (at step N)

$$|b_{i,i=1,N-1}\rangle \quad = \left\{|\phi_n\rangle, |g_n^1\rangle, |g_n^2\rangle, |g_n^3\rangle, \ldots, |g_n^{(N-1)}\rangle\right\} \tag{84}$$

The auxiliary eigenvectors are linear combinations of $|\phi_n\rangle$ and the preceding residual vectors. The lowest can form input to the next step, while the current residual vector is added to the expansion set. The process terminates when the energy is converged or norm of the residual falls below some predetermined value, resuming with the next band.

In this scheme (and most other optimization methods) H is neither stored nor fully evaluated. Partial inversion of H is achieved row-by-row in a restricted subspace, namely the expansion set, using matrix-vector and vector-vector multiplications.

[5] in matrix notation, $\sum_m \phi_m \otimes \phi_m(\delta\phi_n) = \sum_m \phi_m \phi_m^T (\delta\phi_n) = \sum_m \phi_m(\phi_m . \delta\phi_n)$.

2.13 Charge mixing

In the preceding section it was implied that each successive approximation $\rho_{out}(\vec{r})$ to the electronic density generated within the SCF cycle is returned as $\rho_{in}(\vec{r})$ to the next. In practice this usually leads to an instability known as *charge sloshing*, preventing convergence. Sloshing is more likely to occur when the Fermi surface is poorly defined, and a small change in charge distribution can lead to a relatively large change in the Hartree potential. An empirical remedy is charge mixing, whereby the output of several prior iterations is combined to form the input to the next. In functional terms the objective is to minimize the norm of the residual vector

$$R[\rho_{in}] = \rho_{out}[\rho_{in}] - \rho_{in} \tag{85}$$

so that self-consistency ($\rho_{in} \sim \rho_{out}$) is achieved in as few iterations as possible. The simplest prescription is linear mixing:

$$\rho_{in}^{n+1} = \alpha\rho_{out}^{n} + (1-\alpha)\rho_{in}^{n} \tag{86}$$

and the VASP package sets $\alpha = 0.4$ when linear mixing is requested. Although charge mixing can be viewed as a discrete SCF step applied after electronic minimization charge-sloshing can be prevented by modifying the *preconditioning* step described in the preceding section. The role of the preconditioner is to reduce the clustering of eigenvalues and to compress the eigenvalue spectrum. This reduces ill-conditioning usually present in the Hamiltonian matrix and accelerates convergence towards a minimum, which remains unaltered by the presence of the preconditioner. The Kerker preconditioner (Kerker, 1980) is an example. This scales the Hartree and exchange correlation contributions to the KS potential (44) by confining them to a sphere surrounding each space point. The spherical radius is related to the Thomas-Fermi *screening length* in the UEG, the characteristic length over which electron correlation occurs. The overall effect is to attenuate the long-wavelength components of changing electronic density, which cause ill-conditioning. As the iteration proceeds the radius is increased, and the ultimate self-consistency is unaffected.

2.14 Ionic movement

DFT calculations have been described in the context of the Born-Oppenheimer approximation with static ionic nuclei in a varying electronic density producing a ground state energy value for a single ionic configuration. A DFT code usually contains a relaxation scheme for iteratively moving each ion to a nearby PES location having lower energy and then recalculating the total energy and forces. The iteration ends when the force acting on each ion approaches zero (e.g. ≤ 0.01 eV/Å) in a stable or *optimized* configuration which may have changed in size or shape but represents a local PES minimum. This is a reasonable stopping criterion since it implies a very small change in total energy: if any individual atom moves by 0.1 Å then $\Delta E \approx 0.001$ eV. If three-dimensional freedom is not required, individual ions can be fixed in space or constrained to move in fewer dimensions.

Atomic forces arise from electronic attraction and repulsion by nearby nuclei. They can be calculated using the Hellman-Feynman theorem (Feynman, 1939; Hellman, 1937). After the

BO approximation the ground state energy and wavefunctions depend only parametrically on $\{\vec{R}_I\}$ (1) and the theorem states the forces acting on ions are given by the expectation value of the gradient of the electronic Hamiltonian in the ground state:

$$-\vec{F}_I = \nabla_I E_0(\vec{R}) = \frac{\partial}{\partial \vec{R}_I}\langle \Psi_0(\vec{R})|H(\vec{R})|\Psi_0(\vec{R})\rangle \tag{87}$$

$$= \langle \nabla_I \Psi_0|H|\Psi_0\rangle + \langle \Psi_0|\nabla_I H|\Psi_0\rangle + \langle \Psi_0|H|\nabla_I \Psi_0\rangle$$

$$= \langle \Psi_0(\vec{R})|\nabla_I H(\vec{R})|\Psi_0(\vec{R})\rangle$$

where because energy is at a variational minimum, derivative terms involving $\nabla_I \Psi_0$ vanish. The only terms in the Hamiltonian that depend explicitly on nuclear position are the interaction E_{II} and the external pseudopotential. After electronic optimization accurate density wavefunctions for Ψ_0 are available so the $\vec{F}_I$ can be obtained easily by differentiation of these terms. If the basis set depends on atomic position an additional term called the *Pulay force* would arise, but the plane-wave basis set is non-local.

Several schemes exist for determining optimum ionic movement. The simplest is simply to move each ion downwards in the direction of the force. VASP supports a QN-based force minimization in which the ions, rather than the electrons move and the step size is scaled from the force by an external parameter. A variation allows the step size and direction to be determined classically in a molecular dynamics calculation conditioned by a damping parameter. The chosen method will depend on how close to optimum the initial configuration is, and how aggressively the ions can be moved. For example, if two chemically-bonded atoms are configured with a bond length that is much shorter than the equilibrium length, repulsive forces may drive the atoms apart preventing any subsequent convergence.

2.15 DFT algorithm

With sufficient theoretical and computational detail available, it is possible to present a generic algorithm for solving the KS equation (fig 2.4). The apparent simplicity hides a great deal of lower-level complexity, caused by the many processing options offered in a real software package and the highly specialized implementations of the numerical methods employed.

For clarity, electron spin has been largely ignored in the preceding discussion. This is reasonable in systems having an even number of electrons in covalent bonds but otherwise (e.g. Al-Si bonding) spin would be explicitly considered. A spin-sensitive calculation can be thought of as two (spin up/down) calculations proceeding in parallel, with distinct values for v_{eff}^{σ} and ρ^{σ} maintained throughout. The calculations are coupled because the exchange-correlation contribution to the KS potential depends on the total electronic density, which must be conserved throughout the calculation. Assuming that self-consistency can be achieved the two spin densities are summed and a total energy calculated. In either case the algorithm must be primed with some initial guess at the potential and densities. This

can be done by placing the atomic electron density at each ionic coordinate and taking random complex numbers as the basis coefficients of the density wavefunctions. After a few iterations at constant density (allowing the wavefunction to stabilize) the calculation would then proceed.

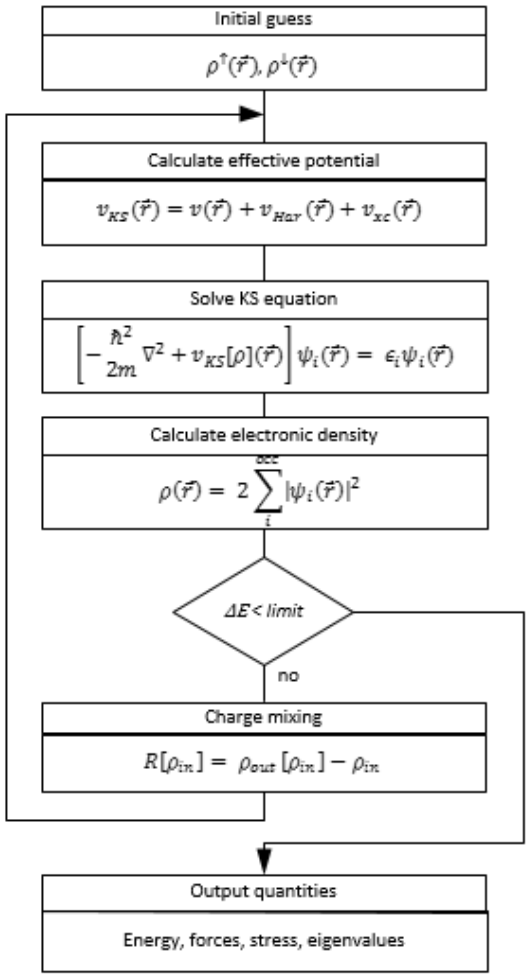


Fig 2.4. Flowchart of the self-consistency loop for the solution of the non-linear Kohn-Sham equation. In a spin-sensitive calculation, two loops would be iterated in parallel producing electronic densities for both spins (after R M Martin, 2004)

With ρ_{in}, the current input electronic density to hand, the effective potential v_{ks} can be calculated. The electronic density is the independent variable of the calculation. In principle, the effective potential can be calculated in either real or reciprocal space and a procedure for the Hartree potential was given earlier. The next step is the solution of the KS equation. This involves calculating the elements of the Hamiltonian matrix and iteratively inverting it to find the eigenvalues and wavefunctions. This latter aspect of the calculation can be regarded as a 'black box' hiding a numerical procedure that solves the KS matrix equation to the required level of accuracy, without treating the matrix in its entirety. In general, the procedure with the lowest convergence time will be preferred.

Finally, a new output density ρ_{out} is calculated and a new ground state energy. If the change ΔE in energy is less than a pre-set stopping criterion (normally $10^{-4} - 10^{-5}$ eV) the algorithm terminates, and the current density distribution assumed to be the unique self-consistent solution of the KS equation. Otherwise, the new density is subject to any desired charge mixing and the new density distribution taken as input to the next iteration.

2.16 Transition State Theory and NEB methods

Nudged Elastic Band (NEB) methods are force-based forms of ionic optimization yielding discrete representations of the *minimum energy path* (MEP) between the initial and final states of a chemical transition of interest, such as surface diffusion (Jonsson et al., 1998). At a typical activation energy (≈ 0.6 eV) diffusion events occur many times per second at room temperature. But there are $\approx 10^{10}$ vibrational periods between such events and no classical dynamics simulation could possibly keep track of them. However, a statistical method, namely *Transition State Theory* (TST) (Eyring, 1935) gives good results over longer timescales. In its *harmonic* approximation form (hTST) it relates the energy difference ΔE between the initial configuration and a transition state † on the MEP with the reaction rate k by the equation:

$$k_{init\to\dagger} = \frac{\nu_\dagger^1 \times \nu_\dagger^2 \times \dots \nu_\dagger^{3N}}{\nu_{init}^1 \times \dots \times \nu_{init}^{3N-1}} e^{\frac{-\Delta E}{k_B T}} \tag{88}$$

where $\nu_\dagger^i$ and ν_{init}^j are the normal mode vibrational frequencies at the transition state and local minimum respectively, N the number of atoms in the simulation, k_B the Boltzmann constant and T the prevailing temperature. At a transition state exactly one of the $\nu_\dagger^i$ will be imaginary, and the remainder real, whereas at a local minimum all the frequencies are real. This formulation assumes the reaction rate is sufficiently slow for a Boltzmann energy distribution to be established in the reactants and neglects quantum effects such as zero-point energy and tunnelling.

The MEP may have one or more minima between the endpoints representing stable intermediate configurations, giving rise to two or more maxima which are transition states or saddle points. The overall rate is determined by the highest energy saddle point. Locating the highest saddle point from the total energy and its first derivative (i.e. ionic force) is the motivation of NEB methods. The resulting energy can be incorporated into (88) along with a notional value of ν (e.g. $\nu = 10^{12} - 10^{14}\ s^{-1}$) to get an estimate of the reaction rate.

An initial approximation to the MEP may arise intuitively but can otherwise be obtained by the linear interpolation between the initial and final states, with the images connected with springs to form a band. A band with $N + 1$ images can be denoted by $[\hat{R}_0, \hat{R}_1, \hat{R}_2, \dots, \hat{R}_N]$ where $\hat{R}_0$ and $\hat{R}_N$, which are fixed, correspond to the initial and final states which have been found in earlier ionic optimizations as described above. A path is a MEP only if the force corresponding to the energy gradient on each image is tangential to the path. Then the $N - 1$ intermediate images are concurrently adjusted using a projection scheme, characteristic of NEB methods. The force acting on the image is taken

to be the sum of the projections of the true (potential-derived) force perpendicular to the local tangent and the spring force along the local tangent:

$$\hat{F}_i = \hat{F}_i^s\big|_{\parallel} - \nabla E(\hat{R}_i)\big|_{\perp} \tag{89}$$

where the perpendicular component of the true force is given by subtracting out the tangential component

$$\nabla E(\hat{R}_i)\big|_{\perp} = \nabla E(\hat{R}_i) - \big(\big|\nabla E(\hat{R}_i)\big|.\hat{\tau}_i\big)\hat{\tau}_i \tag{90}$$

and E is the energy of the system and $\hat{\tau}_i$ the normalized local tangent at image i. The spring force is:

$$\hat{F}_i^s\big|_{\parallel} = k\big(\big|\hat{R}_{i+1} - \hat{R}_i\big| - \big|\hat{R}_i - \hat{R}_{i-1}\big|\big)\hat{\tau}_i. \tag{91}$$

The local tangent at image i is estimated from the adjacent images $\hat{R}_{i-1}$ and $\hat{R}_{i+1}$ (e.g. by normalizing the line segment between the two) while the true force $-\nabla E(\hat{R}_i)$ is obtained from a concurrent DFT electronic structure calculation. The images are all instantaneously moved along the force vectors $\hat{F}_i$ (Jonnson et al., 1998). The images converge on the MEP ($F_i \to 0$) but will not necessarily coincide with any saddle points. The actual saddle point energy must be obtained by interpolation. This cycle repeats until all the forces fall below a convergence criterion.

In the climbing-image (Cl-NEB) variation employed later (Henkelman, 2000) one image, having the highest energy after a few iterations, is selected and constrained by a new force projection to move uphill along the elastic band to a saddle point. The projection on this one image is now given by:

$$\hat{F}_{i_{max}} = -\nabla E(\hat{R}_{i_{max}}) + 2\nabla E(\hat{R}_i)\big|_{\parallel} \tag{92}$$

$$= -\nabla E(\hat{R}_{i_{max}}) + 2\big(\big|\nabla E(\hat{R}_{i_{max}})\big|.\hat{\tau}_i\big)\hat{\tau}_i. \tag{93}$$

while the others continue with the original projection. The climbing-image force is the true force with the component along the local path inverted, and the image does not experience any spring forces. Under this regime the image is pushed up the potential gradient, and when $\hat{F}_{i_{max}}$ converges will be located on the highest saddle point of the MEP.

2.17 Conclusion

This chapter has presented some basic DFT theory and the plane-wave pseudopotential concept as a mechanism for atomistic modelling. Its implementation in the VASP package (Kresse et al., 2009) supports the calculations described in the following chapters. The purpose of this section is to identify some limitations of the DFT approach and comment on the accuracy of the calculations.

Although the Kohn-Sham equation is derived from a many-electron SE it cannot deliver exact solutions. An intrinsic uncertainty exists because the functional assumed by the Hohenberg-

Kohn theorem (which leads to the unique ground state energy) is never known exactly. This uncertainty is generally much greater than any due to SCF convergence or other source and can only be quantified by careful comparison with experimental measurements. However, the non-empirical PBE GGA functional chosen for the present calculations is known to perform well in solid-state scenarios.

There are some situations in which DFT cannot be expected to approach physical accuracy. One is the calculation of excited states. This is simply because the variational principle on which the second HK theorem depends does not apply to excited states. Another well-known inaccuracy is the underestimation of the band gap in semiconductor or insulating materials. A DFT calculation for crystalline Si typically gives a value ~ 0.5 eV, but experimental values are ~ 1.0 eV. This was originally thought to be a shortcoming of the LDA or GGA exchange-correlation functionals, but it has been shown (Rinke at al, 2008) that even a formally exact functional would suffer from the same underlying problem.

Finally, a fundamental limitation of DFT is the computational cost of solving the mathematical problems it presents. Even a very large computer with thousands of cores will take days to optimize a silicon nanostructure consisting of just a thousand atoms. But a single droplet of water 1 micron in radius contains $\sim 10^{11}$ atoms. No conceivable advance in current computer technology would allow DFT examination of a system of this size. In this connection the application of DFT even to small clusters of water molecules has proven arduous – a review of recent progress appears in (Gillan et al., 2016).

Chapter 3

Some preliminary calculations

3.1 Introduction

For a given choice of exchange-correlation functional and pseudopotential, accurate and efficient DFT calculations require judicious specification of the supercell and accompanying external parameters, of which the most critical are the *k* mesh density and the plane wave cut-off energy. In practice the supercell is usually determined empirically, and parameters adjusted incrementally until calculated energies are essentially independent of their exact values. In this chapter we document this procedure using simple, generic Si supercells representing bulk and surface structures, comparing the results with experimental values and similar calculations performed elsewhere. This serves two purposes: to validate our VASP environment and to show how more complex supercells, as featured in subsequent chapters were created. These calculations used the PBE GGA functional (page 32) and PAW pseudopotentials (page 41).

3.2 Reciprocal space and *k* points

In chapter 2 It was shown that the KS effective potential is evaluated iteratively during the SCF procedure, each iteration involving many Brillouin zone integrations of the form:

$$\int_{BZ} g(k)dk \qquad (1)$$

where the $g(k)$ are components of the KS effective potential. On page 35 It was suggested that these integrals could be approximated by summing the functions $g(k)$ over a mesh of *k* points given by the Monkhorst-Pack prescription. This is a numerical integration technique requiring the specification of the number of *k* points in each reciprocal space direction. As noted above, these numbers must be determined empirically for each supercell and species configuration and before any other calculation is undertaken.

Table 3.1 shows the results of total energy calculations in a cubic supercell containing 8 Si atoms with an experimental lattice constant value[6] of 5.431 Å and with the Monkhorst pack *k* point mesh shown in the first column. The calculated per-atom total energies (in the second column) show that a $(5 \times 5 \times 5)$ mesh is needed to achieve convergence to ± 1 meV, and that a denser mesh yields no benefit.

[6] Greenwood; Earnshaw (1985), page 373.

Mesh $m_x \times m_y \times m_z$	***E/atom (eV)***	***No. of k Points in IBZ***	$\tau_{mesh}/\tau_{1\times1\times1}$
1x1x1	-4.1752	1	1.0
2x2x2	-5.3447	4	1.2
3x3x3	-5.4102	14	2.0
4x4x4	-5.4235	32	2.7
5x5x5	-5.4243	63	5.7
6x6x6	-5.4247	108	9.8
7x7x7	-5.4248	172	16.6
8x8x8	-5.4248	256	26.2
9x9x9	-5.4248	365	32.0
10x10x10	-5.4248	500	46.5
11x11x11	-5.4248	666	68.3
12x12x12	-5.4248	864	72.2
13x13x13	-5.4248	1099	112.3
14x14x14	-5.4248	1372	131.4
15x15x15	-5.4248	1688	192.6
16x16x16	-5.4248	2048	210.0

Table 3.1 Total energy calculations for bulk Si with k points given by the Monkhorst pack method. The supercell contains 8 atoms. The processing time τ_{mesh} has been normalized to τ_1. Supercell size is $5.43 \times 5.43 \times 5.43$ Å, as shown in fig 1.1 on page 13.

The third column shows the number of *k* points in the *irreducible* Brillouin zone (IBZ), which is the zone that remains after removal of those which are symmetrically equivalent. The IBZ energies are weighted so that the final total is correct. The VASP software finds a 2-fold symmetry in this supercell, because the number of *k* points in the IBZ is not $M = m_x \times m_y \times m_z$ but either $M/2$ or $(M/2) + 1$. When M is odd the additional point is the Γ point which lies on the IBZ boundary. Unless the Γ point energy is specifically needed, it is more efficient to choose *k* points lying in the interior of the IBZ. The last column shows the processor time taken for the energy calculation, normalized to the value from the single *k* point calculation. These times show that achieving ±1 meV accuracy is relatively costly on this measure. A tolerance of ±5 meV, giving ΔE values to within ±0.01 eV is usually acceptable in the DFT context.

Mesh	***E/atom (eV)***	***No. of k Points in IBZ***	$\tau_{mesh}/\tau_{1\times1\times1}$
1x1x2	-5.0934	2	1.8
2x2x2	-5.4557	4	3.5
2x2x3	-5.4528	6	3.6
3x3x3	-5.4198	14	8.4
3x3x4	-5.4211	18	7.2
3x3x5	-5.4208	23	8.8
4x4x4	-5.4256	32	12.2
4x4x5	-5.4245	40	15.5
5x5x5	-5.4246	63	30.9

Table 3.2 As table 3.1, for a supercell with 16 atoms. τ_M values are comparable with table 3.1. Supercell size is $7.68\times 7.68 \times 5.43$ Å.

Table 3.2 shows equivalent results for a larger supercell containing 16 Si atoms. Because

$$V_{BZ} = \frac{(2\pi)^3}{V_{cell}} \tag{2}$$

the reciprocal cell is smaller, the volume per *k* point reduces and integration accuracy improves. In this case the cell is not uniform, and one can assign fewer k points on the longer axes without sacrificing accuracy. So, a given accuracy is achieved with fewer *k* points. This can be seen from the table, with a $4 \times 4 \times 5$ mesh yielding the same accuracy as a $5 \times 5 \times 5$ in the smaller supercell. The processing times do not scale with the cube of system size ($2^3 = 8$ in this case) because some parts of the VASP code scale by smaller factors and the prefactor for the cubic part is relatively small.

The cohesive energy E_{COH} of a crystal can be defined as the energy that must added to separate it into free atoms at rest and at infinite separation, at 0 K. To compare the energies of tables 3.1 and 3.2 with published values of E_{COH} for Si the energy of an isolated Si atom must be subtracted:

$$E_{COH} = E_{cell} - E_{atom} \tag{3}$$

A separate VASP calculation returned -0.8042 eV for E_{atom}. Substituting in (3) gives $E_{COH} = -4.62$ eV, in good agreement with a published value of -4.63 eV (Kittel, 2008).

3.3 The cut-off energy

As noted earlier (page 38) the infinite sums of the Fourier representation of the KS electronic wavefunction (49) must be truncated for use in practical energy calculations. This is done by retaining terms having kinetic energy less than a *cut-off* value, so that the expansion consists of plane waves having kinetic energy $E < E_{cut-off} = (\hbar^2/2m_e)\left|\vec{G}\right|^2$. The cut-off value then appears as an external parameter to the calculation. This definition implies that the number of plane waves changes discontinuously with the cut-off value and will be different at each k point, but that effect is reduced with a denser k point mesh. The dependency on the $\vec{G}$ vectors means that the basis set is modified whenever the supercell alters in size or shape.

The actual limit value depends on the pseudopotential species and pseudopotential files will contain a default setting that may well have been determined empirically. If a supercell configuration contains mixed species the largest cut-off value must be taken. For the VASP Si PAW pseudopotential the default value is 250 eV, said to produce energies accurate to within a few millivolts (Kresse et al., 2009). Nevertheless, it is worthwhile performing some bulk calculations to verify that the recommended value is correct. For this pseudopotential the results of table 3.3 show that total energy is converged with the advertised accuracy with the recommended cut-off. They further show that the total energy value could be lowered by up to 0.01 eV using larger cut-off energies. At 250 eV the calculation generates ≈ 45000 plane waves, half the number needed at a 390 eV setting, indicating a large saving in processing time with only a moderate loss of accuracy.

Cut-off energy (eV)	*No. of plane waves*	*E/atom (eV)*	*ΔE*
90	10368	-5.2865	
95	14112	-5.3093	-0.0228
100	14112	-5.3286	-0.0193
110	15680	-5.3480	-0.0194
130	18000	-5.3672	-0.0192
150	24576	-5.3785	-0.0113
170	31104	-5.3898	-0.0113
190	36288	-5.3996	-0.0098
210	44800	-5.4070	-0.0074
230	44800	-5.4115	-0.0045
250	44800	-5.4160	-0.0045
270	52920	-5.4195	-0.0035
290	69120	-5.4219	-0.0024
310	73728	-5.4233	-0.0014
330	73728	-5.4241	-0.0008
350	82944	-5.4247	-0.0006
370	90000	-5.4252	-0.0005
390	90000	-5.4256	-0.0004

Table 3.3 Total energy per atom E against cut-off energy and size of plane wave basis set for the VASP Si PAW pseudopotential, in the 16-atom supercell. *k* point setting is $M = 4$. The recommended cut-off energy is 250 eV.

3.4 The bulk Si lattice constant

The diamond structure of crystalline silicon described in chapter 1 is well known and simple to parametrize for input to a DFT calculation. But this structure is only one of several that exist at various pressures, each with its own geometry (McMahon and Nelmes, 1993). Ranking the stabilities of all possible structures with DFT would be a large undertaking, but

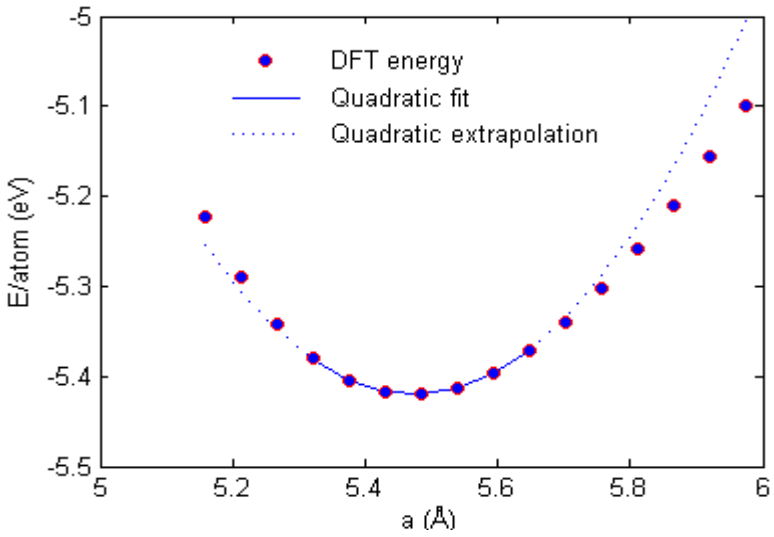


Figure 3.1 Calculated energies for 16-atom supercell for various values of lattice parameter *a*, with $(4x4x4)$ *k* point sampling. Quadratic fit is given by $y = 1.6682x^2 - 18.2755x + 44.6172$.

we can verify that the diamond structure and experimental lattice parameter value do represent a minimum energy configuration and then compare a by-product of the calculation, the *equilibrium bulk modulus* with experimental values.

A series of total energy calculations for the 16-atom supercell of the preceding section with M=4, and the lattice parameter values varying between -10% and $+10\%$ of the experimental value $(5.431\ \text{Å})$ is shown graphically in fig 3.1 above. This supercell and sampling mesh are known to yield well converged energy values (table 3.2).

The graph shows that the energy is minimized near this value, say a_0. This is reassuring as it indicates that ionic optimizations will tend to converge in this vicinity and not at some spurious minimum elsewhere. The energy can be approximated by a truncated Taylor expansion around a_0:

$$E(a) = E(a_0) + \alpha(a - a_0) + \beta(a - a_0)^2 \tag{4}$$

with $\alpha = dE/da|_{a0}$ and $\beta = 1/2\, d^2E/da^2\,|_{a0}$. At $a = a_0$

$$E(a) \cong E(a_0) + \beta(a - a_0)^2 \tag{5}$$

and we try quadratic curve fitting to estimate $E(a_0)$, a_0 and β. This gives $E(a_0) = -5.4254$ eV, in agreement with value in table 3.2 for this k point mesh. The calculated value of $a_0 = 5.47$ Å is greater than the experimental value, showing that the PBE-GGA functional has reversed the overbinding tendency of the LGA functional. The value $\beta = 1.6682$ can be used to estimate the equilibrium bulk modulus B_0 using the definition:

$$B_0 = V\frac{d^2E}{dV^2} \tag{6}$$

where V is the notional crystal volume occupied by a single atom and the derivative is taken at the equilibrium lattice parameter. Changing the variable in (6) from V to a gives

$$B_0 = \frac{16}{9} \times \frac{1}{a_0} \times \beta \cong 0.54\ \text{eV/Å}^3 \cong 86\ \text{GPa} \tag{7}$$

This compares to a published experimental value $B_0 \cong 99$ GPa (Moll et al., 1995). Away from the lattice equilibrium the energy cannot be represented by a quadratic function, as shown by the extrapolation in fig 3.1. A better approximation is the Birch-Murnaghan equation of state, a cubic in a (Birch, 1947). If this were fitted to the energy data of fig 3.1 slightly different values of a_0 and B_0 would be obtained. However, the quadratic result suffices for the present purpose: to show that the lattice model and other external DFT parameter settings yield results consistent with experiment. This approach is similar to that adopted by (Sholl; Stecker, 2009) when examining the crystal structure of Cu.

3.5 Supercells with surfaces and structures

The infinitely repeated supercell model can support structures which are seemingly non-periodic, or periodic in only one or two dimensions, by the introduction of vacuum spaces.

The original idea was published by Evastorev at al, (1975) who were able to impose 3-dimensional translational symmetry on molecular models of solids by regarding them as cyclic structures and applying mathematical group theory. However, translational symmetry is implicit in the periodic supercell scheme of VASP DFT and vacuum spaces can be freely deployed without code change, as shown in Fig 3.2.

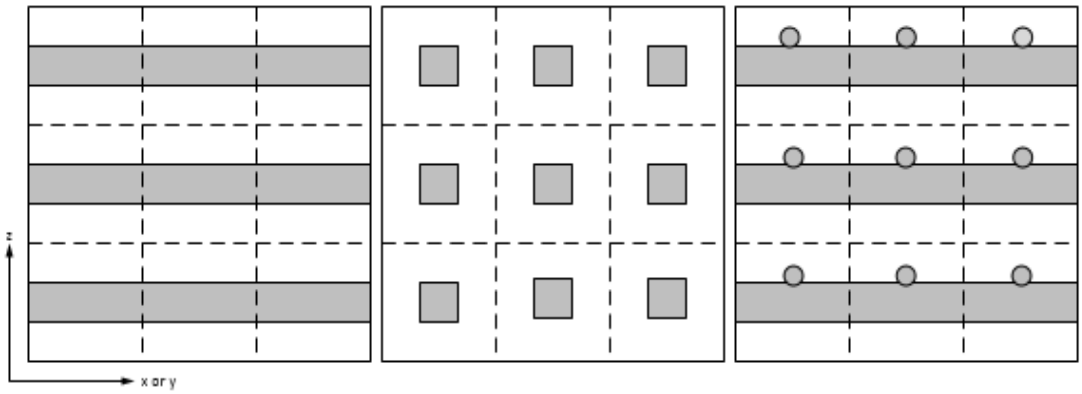


Figure 3.2 2-dimensional sketch of 9 supercell replicas of systems with less than 3-dimensional periodicity. The dashed lines denote supercell boundaries; shaded areas contain atoms, white depict vacuum. Left: surface slab (2-dimensional periodicity); centre: wire nanostructure (1-dimensional periodicity); right: surface slab with adsorbate (2-dimensional periodicity).

When a supercell contains vacuum space there will usually be a 'long' axis passing through the vacuum. The corresponding reciprocal space dimension will therefore contain regions of zero electronic density. If the vacuum interval is sufficiently large there will be no interaction between replica systems and consequently no band dispersion to sample. A single reciprocal space coordinate coordinate is therefore enough, obtained by setting is $M_{long} = 1$ in the Monkhorst-Pack prescription. The mesh settings in the other axes are determined through convergence testing as already described.

When a bulk surface is cut it leaves *undercoordinated* atoms on the surface, with fewer bonds than they had in the crystal. In Si these are valence bonds and the electrons remain localized, giving rise to *dangling bonds*. An energy-lowering reorganization occurs naturally, minimizing the number of such bonds. On the Si(100) surface this results in dimerization, with a dangling bond at each end of the dimer bond. A further relaxation can occur in which the dimer bonds are pairwise tilted *(buckled)* out of the horizontal plane. Here surface dimerization and buckling are introduced by construction, and the equilibrium bond lengths determined by ionic optimization (page 48).

It is also important to ensure there is no interaction between surfaces or their periodic images. The structure of fig 3.3 models the buckled Si(100) upper surface with a vacuum space of 13 Å above, and the lower passivated with H atoms. These surfaces together with intervening bulk-like Si atoms are the slab of figure 3.2 (left) above, the two vacuum spaces having been merged. The chosen height of the vacuum region is typical in this setting, and sufficient for the electronic density to decline to zero before encountering the adjacent periodic surface.

The reorganization of the upper surface distorts the bulk structure beneath, which is compromised when the cell depth is too small. Table 3.4 shows the change ΔE in calculated energies as the cell depth is increased. Each depth increment introduces 16 Si atoms into the cell. Convergence to within 0.01 eV (with respect to cell depth) occurs at ≈ 40 Å, although some quantities (e.g. band width at the valence band edge) are not completely converged at much greater depths (Sagisaka et al., 2017). When surface phenomena are compared (e.g. adsorption energies) the calculations should be made in cells of the same size.

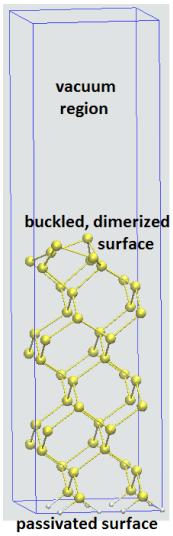


Figure 3.3 Optimized supercell with 48 Si atoms (yellow) with dimerized and buckled surface in the (100) plane, a 13 Å vacuum region above and 8 passivating H atoms (grey) below. Blue lines represent periodic boundaries. H positions and the lowest Si layer were fixed during optimization. Cell dimensions are $(7.68 \times 7.68 \times 29.40)$ Å.

Cell depth (Å)	***Number of atoms***	***Total energy (eV)***	ΔE	τ
23.97	40	-194.03964		1.0
29.40	56	-280.83089	-86.79125	1.9
34.83	72	-367.60086	-86.76996	3.7
40.26	88	-454.36843	-86.76757	6.1
45.69	104	-541.13602	-86.76759	9.2
51.12	120	-627.90439	-86.76836	13.6
56.55	136	-714.67235	-86.76796	20.4
61.98	152	-801.44166	-86.76931	29.5

Table 3.4 Total energies for surface structures as Fig 3.4, of varying depths. The cell depth includes a vacuum space of 13 Å. Each depth increment incorporates a further 16 Si atoms into the structure. The Monkhorst-Pack *k* point mesh for all structures was $(4 \times 4 \times 1)$. Ionic optimization was terminated when forces fell below 0.02 eV/Å. τ represents the processing time taken to optimize relative to that of the 40 atom cell.

3.6 Electronic Density of States

The density of states (DOS) is a measure of the number of states (eigenstates) existing at a given energy E or within an energy window:

$$g(E) = \sum_i \delta(E - \epsilon_i) \tag{8}$$

where ϵ_i denotes the energy of an individual electron. The electronic DOS condenses the electronic properties at all possible locations in reciprocal space into a simple form. A calculated DOS helps analyze the electronic structure of a system and supports experimental techniques such as scanning tunnel spectroscopy, where the surface DOS can be related to the measured tunnelling current.

A DFT DOS calculation is performed in two steps. First, the ground state electronic density is obtained from a static SCF calculation using enough *k* points to ensure a well-converged charge density. A second, non-SCF calculation involving a much finer *k* point mesh is then performed against the ground state density distribution obtained from the first. This technique delivers the eigenstates at many points in reciprocal space in a single electronic minimization step, effectively probing the entire Brillouin zone. The energy window is then divided into bins and each eigenstate cast into the appropriate bin, taking account of any *k* point weighting.

Since the total number of electrons is finite a plot of the eigenstate count against bin number (i.e. $g(E)$ against E) would be a comb-like set of vertical lines, correctly depicting the DOS at 0 K but at odds with experimental spectra which are broadened by thermal fluctuation. So, it is usual to apply Lorentzian broadening to the peaks, significantly altering their appearance (see Figs 3.4 ($\sigma = 0.02$ eV) and 3.5 ($\sigma = 0.15$ eV), where σ is the peak width).

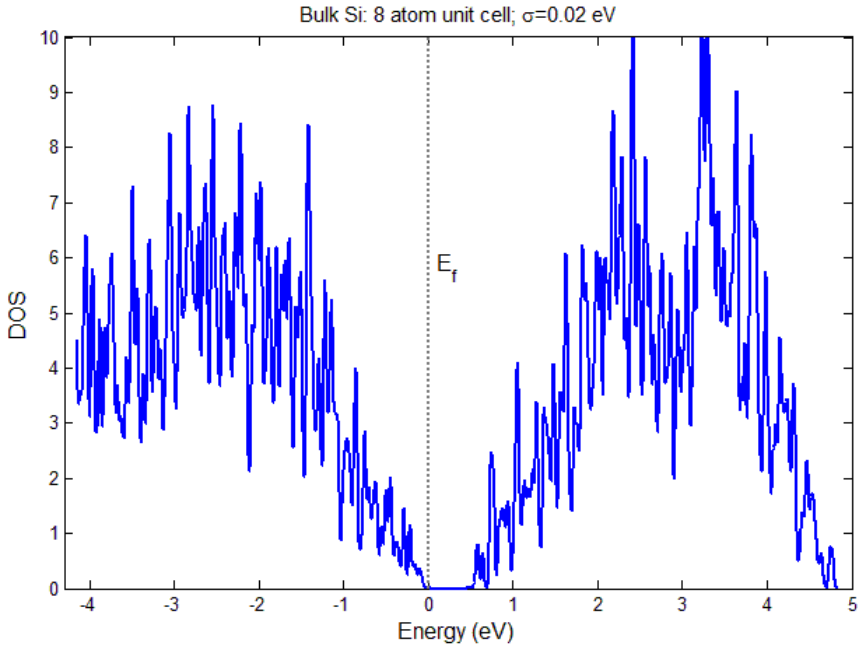


Figure 3.4 Calculated electronic DOS for bulk Si, using the 8-atom supercell and a $(16x16x16)$ Monkhorst-Pack mesh giving 2024 irreducible k points. The energy window is set to 9 eV and the Fermi level is at 0 eV. Gaussian smoothing is employed ($\sigma = 0.02$ eV). The predicted band gap (≈ 0.6 eV) is in poor agreement with experimental values *circa* 1.1 eV but in good agreement with other DFT calculations (see page 112). Author's MATLAB rendering of VASP DOSCAR data.

The step change in electronic density at the Fermi level implies the functions being integrated in reciprocal space change discontinuously (from non-zero values to zero), and many k points would be needed to yield a converged result. This complication is avoided by 'smearing' the wavefunctions, i.e. enforcing continuity by inserting artificial intermediate values and which cause partial band occupancies to arise after integration. The smearing can be done in several ways but here the implicit step function is replaced with a smooth Gauss-like function:

$$f\left(\frac{\varepsilon-\mu}{\sigma}\right)=\frac{1}{2}\left(1-\mathrm{erf}\left(\frac{\varepsilon-\mu}{\sigma}\right)\right) \tag{9}$$

$$\mathrm{erf}(x)=\frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}\,\mathrm{dt} \tag{10}$$

where μ is the Fermi level and σ again controls the degree of smoothing. A small and fictitious ($\approx \sigma$) electronic temperature is introduced, creating partial band occupancies around the Fermi level and facilitating numeric integration. It is reassuring that both plots show the characteristic 'narrow' band gap of bulk Si and that this property is evidently insensitive to the value of σ. This reflects a relatively simple electronic structure at the Fermi level. For metals with a complex DOS (e.g. Rhodium or Vanadium) a different approach might be adopted e.g. the tetrahedron method. Here the k points define a set of tetradhedra that fill the IBZ and an integration performed over each tetrahedron, using linearly interpolated coordinates.

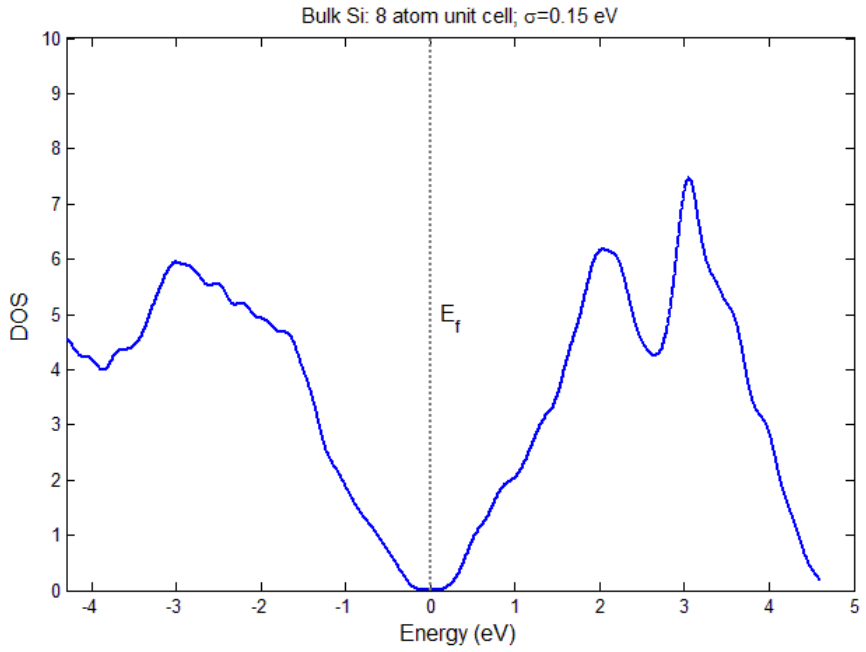


Figure 3.5 As Fig 3.4, broadened with $\sigma = 0.15$ eV. The area under the DOS plot must correspond to the total number of states which is unchanged, so broadening will lower the peak values. However, discontinuities in the slope of the DOS can persist. These are mathematical artifacts known as van Hove singularities.

The presence of dopant impurities within bulk Si or a nanostructure can be modelled by substituting a single dopant atom into a large supercell, while holding the lattice constant fixed. The impurity may give rise to new electronic states near the Fermi level and visible in the calculated DOS spectrum. The manipulation of these states by external electric fields opens the prospect of engineered nanoscale devices. Some DOS calculations of this kind are presented in chapter 7.

3.7 Band structure

The electronic DOS condenses the properties of the electronic states at all possible positions in the IBZ into a simple form. An alternative view of the same data can be obtained by resolving the states into bands. The band structure represents the energy of available states along a series of lines in reciprocal space, usually forming a closed loop beginning and ending at the origin (Γ point) and passing through other points of high symmetry (see table 3.5). This presentation can help identify possible electronic and optical transitions. The appropriate route through the Brillouin zone depends on the supercell, e.g. for a nanowire structure one would take the path corresponding to the axial direction. For the face-centred cubic (fcc) structure of Si (where band dispersion is independent of supercell orientation) it is usual to adopt the segments $L-\Gamma-X$ or $W-L-\Gamma-X-W$. Here the lattice is described by its rhombohedral primitive cell whose translation vectors connect the lattice point at the origin with those at the face centres. In this scheme the Si diamond (consisting of two intersecting fcc cells) is obtained with just two Si atoms, one at the origin and the other displaced by $(a/4\,,a/4\,,a/4)$, where a is the length of the conventional cubic lattice.

Like the DOS calculation, band structure requires two consecutive runs: a static SCF run and another taking the computed charge density and giving the band energies along the directions of interest. A small number (10 or 20) *k* points, derived by linear interpolation between the reciprocal coordinates of table 3.5 is enough to create smooth band structure plots.

Symmetry point	***Reciprocal coordinates (units of $\hat{b}_1, \hat{b}_2, \hat{b}_3$)***	***Cartesian coordinates (units of $2\pi/\hat{a}$)***
Γ	$(0,0,0)$	$(0,0,0)$
X	$(1/2\,,0,1/2)$	$(0,1,0)$
W	$(1/2,1/4\,,3/4)$	$(1/2\,,1,0)$
L	$(1/2\,,1/2\,,1/2)$	$(1/2\,,1/2\,,1/2)$
Δ	$(1/4\,,0,1/4)$	$(0,1/2\,,0)$
Λ	$(1/4\,,1/4\,,1/4)$	$(1/4\,,1/4\,,1/4)$

Table 3.5 A list of high symmetry points in the first Brillouin zone of the face-centred cubic lattice.

Fig 3.6 shows the structure of the first six bands of bulk Si calculated with the two-atom supercell, along the path $L-\Gamma-X$. There are four valence bands (occupied at $T=0$) of which two are degenerate, two conduction bands (unoccupied at $T=0$) and an intervening band gap (≈ 0.6 eV) as in the earlier DOS plots. The valence band maximum and conduction

band minimum are not vertically aligned in k space, indicating an indirect semiconductor. The appearance of the plots indicates that the KS energies and electronic density wavefunctions vary smoothly in *k* space, confirming that *k* point sampling is a reasonable scheme for evaluating the varying quantities in the IBZ.

By way of comparison, fig 3.7 shows the results of another Si band structure calculation based on the '$k.p$' method. In this (non-DFT) approach expressions for the energy in the vicinity of a high symmetry *k* point are obtained in terms of parameters whose values are experimentally determined (Cardona, Pollack 1966). Fig 3.6 shows good agreement with the $k.p$ calculation apart from the band gap, where the latter returned a value ≈ 1.20 eV which is quite close to the experimental value ≈ 1.1 eV and more than twice the DFT result.

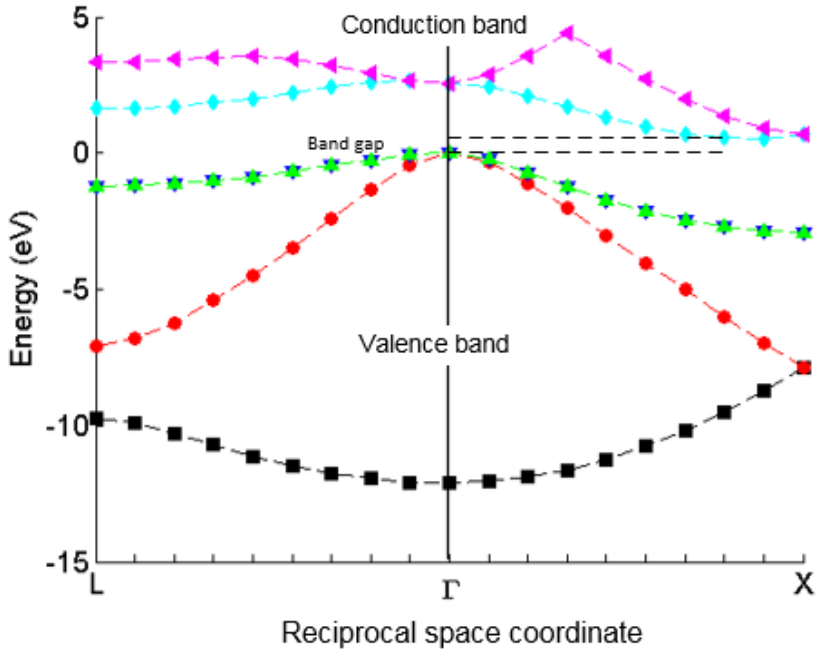


Figure 3.6 Bulk Si band structure calculated with a 2-atom supercell. The horizontal axis is divided into 19 equal increments on the $L - \Gamma - X$ segment of reciprocal space. Coloured markers represent the lowest 6 eigenstates available at each coordinate on the segment. (Author's MATLAB rendering of VASP EIGENVAL output)

That DFT is inaccurate in this respect is to be expected since it fixes the number of electrons (n) and places them all in ground-state levels in the valence band. But the true band gap is the removal/addition energy $E_g = \varepsilon_{n+1}(n+1) - \varepsilon_n(n)$ due to an *incremental* electron entering or leaving the conduction band. The KS version of the band gap is instead $E_g^{KS} = \varepsilon_{n+1}(n) - \varepsilon_n(n)$, with respect to the ground-state electrons. Another source of inaccuracy is partial removal of the self-interaction energy, due the approximate nature of the XC energy (page 31). However, the band gap issue has not prevented the widespread application of DFT/GGA to semiconductor problems.

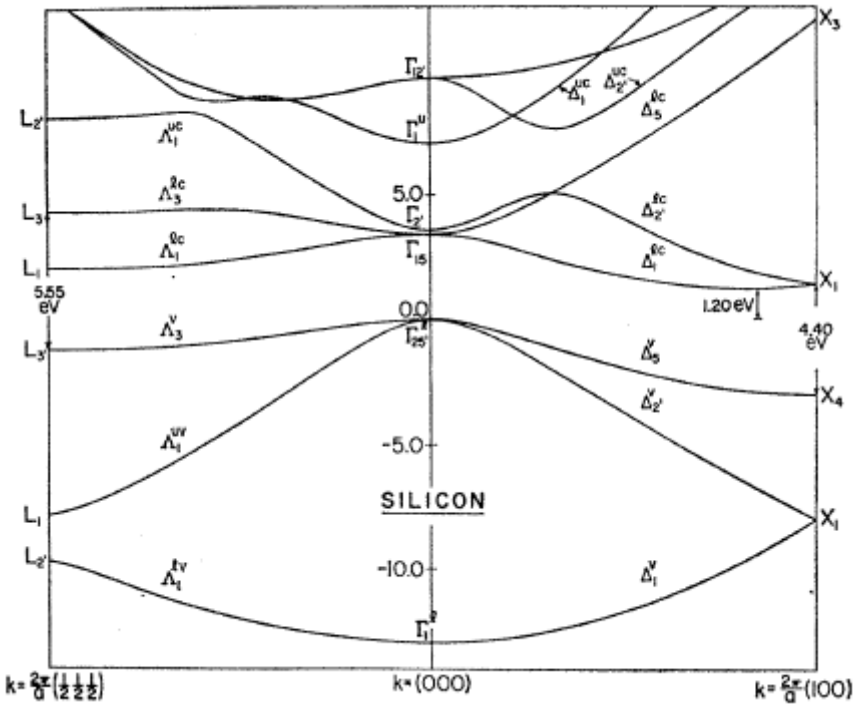


Figure 3.7 Bulk Si band structure with further conduction bands, calculated by the $k.p$ method and showing a band gap of 1.20 eV (Cardona, Pollack 1966).

Chapter 4

H corner diffusion

4.1 Summary of earlier work

In a preliminary study DFT was used to study H adatoms on the top and sidewalls of a notional Si nanostructure, a pillar defined by Patterned Atomic Layer Epitaxy (PALE) and having a (100) growth surface and (110) sidewalls. A chevron-shaped supercell containing the (001) and (110) faces was built and structurally optimized and several H adsorption sites were identified. The availability of those data motivated this study, i.e. the identification of possible diffusion pathways that might remove top surface H during the PALE process and thereby compromise it. The results were subsequently published (Smith; Brázdová; Bowler, 2014), and that paper forms the basis of this chapter.

4.2 PALE

Historically, bulk silicon chip fabrication has been a planar process based on repetitive optical lithography. The progressive reduction in device surface area has impaired switching characteristics, leading to the introduction of three-dimensional structures (e.g. finFETs) that can be realized in planar processes. This trend has prompted interest in other ways of engineering three-dimensional silicon growth, particularly PALE (Martin, 2014; Walsh, Hersham 2009; Lyding et al., 1994; Shen et al., 1995) is a thin-film fabrication technique capable (in principle) of atomic precision, and here refers to the selective growth of silicon structures on the Si(100) surface using monohydride atoms as a mask. The mask is patterned by an adapted scanning tunnelling microscope (STM) and new silicon deposited from a disilane (Si_2H_6) precursor in a CVD (page 17) reaction. Both the lithography and deposition are performed in UHV conditions. The deposition process is complex and results in an incomplete monolayer that is itself hydrogen terminated. This is caused by an inherent nucleation defect referred to as an *antiphase boundary* or APB and discussed below. Consequently, the lithography-deposition cycle must be repeated several times to create a monolayer. The process is depicted in fig 4.1 and the following sections provide additional detail.

4.2.1 Passivation

Hydrogen passivation of the Si(100) surface is convenient for several reasons. First, it retains the order of the underlying Si$(100)2 \times 1$ reconstruction, in which the surface periodicity changes to a pair-wise dimerization. Second, the passivated substrate is quite robust and retains its integrity even after exposure to ambient conditions. Finally, two

desorption regimes are available which provide useful control of precision in the lithography process.

The passivated surface is prepared by heating a Si(100) sample to 1250 K to clean it and then dosing with gaseous H_2 at ≈ 650 K (Lyding et al., ibid). The H_2 molecules are atomized by a filament heated to 1500 K situated close to the sample. These operations are conducted in an ion-pumped UHV system (5×10^{-11} Torr). The exposure required is 1200 L (1 Langmuir (L) = 10^{-6} Torr seconds). Spontaneous desorption (depassivation) occurs when the surface temperature exceeds 750 K, with the hydrogen atoms becoming mobile at ≈ 650 K. This implies substantially lower growth temperatures, typically 500 – 550 K.

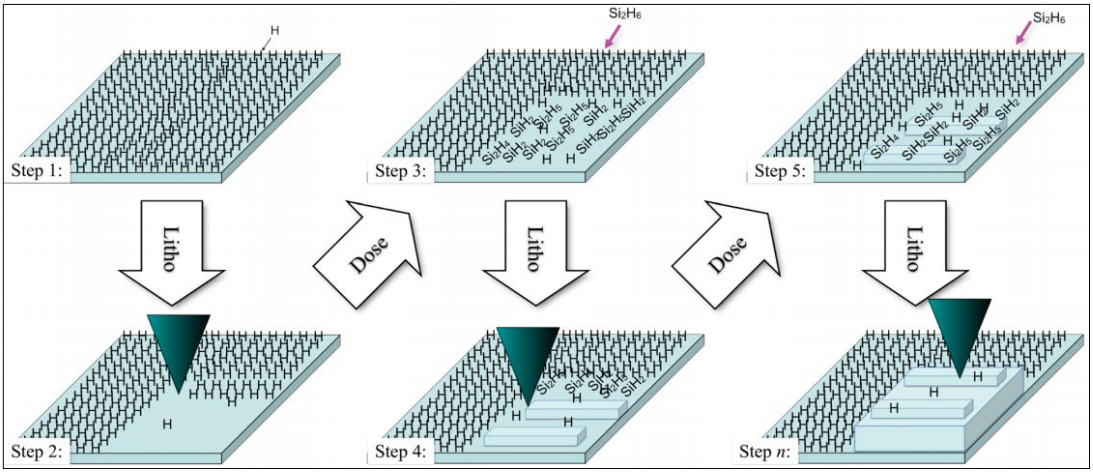


Figure 4.1 Schematic of the PALE process. Step 1: A H-passivated Si(100) surface is stabilised at growth temperature 500 – 550 K under UHV. Step 2: A pattern is etched by STM lithography. Step 3: The tip is retracted, and disilane is introduced to saturate the pattern. Step 4: Lithography is repeated to remove the H from the adsorbed disilane fragments. Step 5: The disilane dosing step is repeated. Step n: After many cycles of Steps 4 and 5, multiple monolayers of Si are built up. (Image courtesy Zyvex Labs LLC)

4.2.2 STM Lithography

The optimum values of STM bias voltage and current, together with the electron dosage necessary were determined experimentally by (Lyding et al., ibid). Two desorption modes were discovered, determined by the bias voltage. In the high-bias mode (> 6.5 V) the Si-H bond is directly excited from the σ bonding to the σ* antibonding state by injecting electronic charge at a rate ≈ 0.1 mC/cm. This is the electronic dose needed to remove more than 90% of the H atoms along a line and results a line width of about 5 nm. In the low-bias (< 6.5 V) mode the desorption yield falls by several orders of magnitude and is dependent on the bias voltage and electron dosage. No incoming electron has enough energy to break the Si-H bond but a vibrational mode can be excited which eventually leads to dissociation of the H atom through a competitive heating and cooling mechanism. Electronic dosage in this mode can vary from 4 to 20 mC/cm.

Resolution in the high-voltage mode is limited to ≈ 50 Å, which corresponds to about 20 atomic rows. The low-voltage mode was re-examined by (Tong; Wolkow 2006) who found that H atoms could be made to desorb in pairs by adjusting the tip tunnelling current. Selective pairwise desorption could occur via intradimer, interdimer and interrow pathways

while still yielding atomic precision. Further study (Ballard et al., 2014) showed the low-voltage mode could excite spurious desorption away from the patterned area, and unwanted repassivation of previously exposed areas. An idealized representation of H desorption leaving the desired a pattern of *dangling bonds* (DBs) is shown in fig 4.2 below.

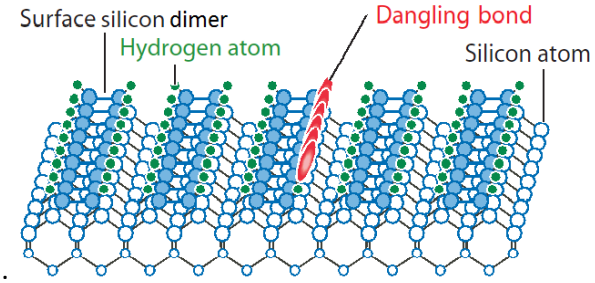


Figure 4.2 Idealized STM lithography of single row of the dimerized Si(100)-2x1:H surface. Dangling bonds shown in red, monohydride passivation in green (Hashizumi et al., 1996)

4.2.3 Growth using disilane precursor

This proceeds in cycles as indicated in fig 4.1. The lithography and dosage steps must be done consecutively at constant temperature, since temperature cycling would cause drift and difficulty in reacquisition of the STM image. Disilane (rather than silane) is used as precursor because it dissociates at a lower temperature that does not disturb the passivation of the unpatterned areas. Growth temperatures lie in the range $500 - 550$ K.

The adsorption process is complex (Owen et al., 1997) but can be summarized as follows. After disilane disassociation the SiH_3 groups lose a hydrogen atom and the active adsorbate consists of SiH_2 ions. These diffuse across the patterned surface and migrate to locations in the trenches between dimer rows, with each Si bonding with a DB from opposing rows. Adjacent adsorbate groups react to form Si_2H_2 groups, evolving an H_2 molecule. There follows a dehydrogenation step that creates a new Si dimer in the next molecular layer. This is the start of a *dimer island* (DI) which can expand by nucleation along the line of the trench below. Most of the nucleation required to form a monolayer occurs in the first litho-dosage cycle, but further complete coverage is approached after 3-4 cycles (i.e. steps 4/5 in fig 4.1). The successive litho/deposition steps are necessary to treat unreacted disilane fragments and growth anomalies caused by APBs (Owen et al, 2011; Chadi, 1987).

An APB arises when neighbouring DIs grow out of phase – a gap or kink will appear at the interface. APBs can occur either parallel (A-type) or perpendicular (B-type) to dimer rows, which mutually interfere to prevent DI nucleation (see fig 4.3). A defect may become embedded in the growth, not filling in in the manner of conventional MBE.

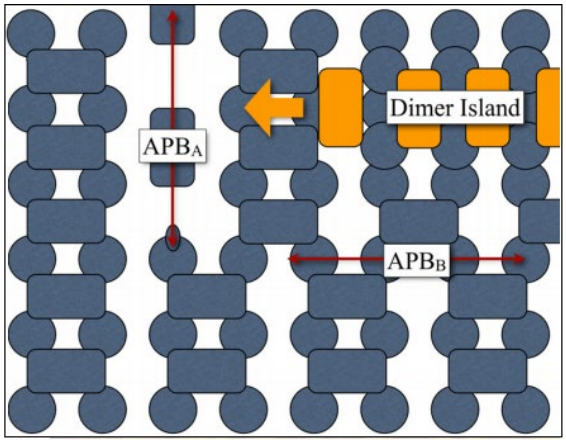


Figure 4.3 The Si(100) growth surface, showing vertical dimer rows following disilane dosing. Dimer islands (Dis) can form both vertically and horizontally (shown in yellow). When the rows are kinked a B type APB_B occurs, creating favourable sites for further nucleation. However, the vertical A type APB_A can act as a barrier to expansion of the horizontal DIs and prevent fill in and island growth. (Image courtesy Zyvex Labs LLC)

4.2.4 Outlook for PALE

As described, Si PALE is a laboratory procedure and great deal of development would be required before it could be exploited commercially. The principal problem is that the growth in the second and subsequent monolayers is critically dependent on the removal of the last few percent of hydrogen atoms from the surface. These residual H atoms inhibit diffusion of the SiH_2 groups and DI formation and are themselves prone to diffuse along trenches in the (100) surface with a predicted energy barrier of the order 1.7 eV (Bowler et al., 1998), (Durr; Hofer 2013) in approximate agreement with experiment (Owen et al., 1996). Other theoretical calculations on the (110) surface predict somewhat smaller barriers, i.e. 1 eV (Brázdová; Bowler 2011). Surface diffusion effects would be compounded if energetic H atoms could diffuse from the walls of a structure (having the (110) orientation) onto the (100) growth surface. This possibility is determined by the PES surrounding the corner regions and is examined in the remainder of this chapter.

A related *pattern transfer* process avoiding the residual H problem has been described (Ballard et al., 2014). Here a 2D spatial pattern is created by hydrogen depassivation as in PALE and CVD used to create a thin (≈ 2 nm), hard mask of titanium dioxide over the exposed areas. The bulk silicon beneath the unmasked areas is removed by conventional reactive ion etching (page 16). Pillar features up to 17 nm tall with lateral dimensions down to 6 nm have been demonstrated, while retaining the intrinsic atomic placement accuracy of the PALE process.

In another PALE application *quantum-dot cellular automata* (QCA) have been implemented as arrays of tunnel-coupled DBs on the H-passivated Si(100) surface (Wolkow et al., 2013). The QCA is characterized as bistable and edge-driven i.e. input, output and power are delivered to the array edge and no communication with its internals is necessary. Any required computer logic function can be realised as a QCA network (Lent et al., 1993) and a functional QCA could form the basis of a post-CMOS electronics architecture. Wolkow's

devices are of interest due to their low power consumption and ability to operate under ambient conditions.

4.3 Methods

4.3.1 Computational details

Energy calculations on the chevron supercell used DFT as implemented in VASP version 4.6.34, together with the PBE GGA exchange-correlation functional (page 32). The VASP PAW pseudopotentials (page 41) were used and the POTCAR file for the silicon atom was dated 5th January 2001 and that for hydrogen 15th June 2001. The energy cut-off was set to 200 eV. The Brillouin zone was sampled with a $(6 \times 2 \times 1)$ Monkhorst-Pack mesh. Gaussian smearing was applied to fractional occupancies with a width of 0.1 eV. The convergence criterion for forces on atoms was 0.01 eV/Å and for total energy 10^{-4} eV. These parameters yield relative energies and energy barriers reliable to within ± 0.01 eV. Transition state search was performed using the climbing image nudged elastic band (NEB) method as implemented in the VASP Transition State Tools (VTST) code (Jonsson, 1998, Shepherd, 2008). The climbing image variation of NEB converges rigorously onto the highest saddle point using just a single intermediate point and yields accurate barrier energy values (Klimes et al., 2011).

4.3.2 The supercell

Calculations were based on a chevron-shaped supercell developed in earlier work, as mentioned above. The H-terminated structure contained 350 Si atoms with a vertical axis of 50Å in the (112) direction. This orientation exposes approximately equal (100) and (110) surface areas at the apex. A smaller supercell with 160 Si atoms was tried initially but did not yield converged energies with respect to its depth or width. It was enlarged and convergence achieved when the overall dimensions were 8 Å $\times$ 28 Å $\times$ 50 Å including a vacuum region of 12 Å (fig 4.4).

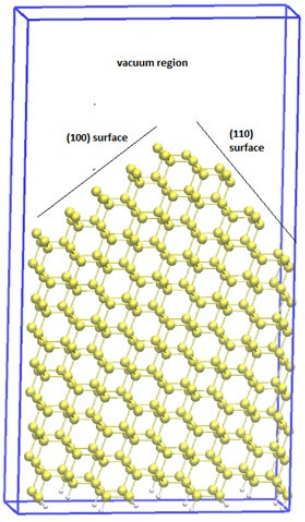


Figure 4.4 Final unoptimized chevron slab supercell with 8 Å $\times$ 28 Å cross-section and 350 Si atoms and 24 H atoms. The Si atoms are coloured yellow and the passivating H atoms (at the base of the cell) are outlined in grey. The blue lines represent the computational unit cell.

4.3.3 Finding H adsorption sites and diffusion pathways

The H adsorption sites are likely to be 3-coordinate atoms on the (110) surface or dimers on the (001) surface. In both cases a single dangling bond is exposed and its replacement by a covalent Si-H bond causes a reduction in the total energy. Single H atoms were placed 1.5 Å above each potential adsorption site on the reconstructed surface (in the direction of the dangling bond) and the structure re-optimized. Total structure energy was typically reduced by approximately 4 eV when compared to the clean structure. Binding energies of the adsorbed hydrogen atom were calculated with respect to the energy of a gaseous H_2 molecule:

$$E_{bind} = E_{adsorb} - E_{clean} - \tfrac{1}{2}E_{H_2} \quad (1)$$

where E_{adsorb}, E_{clean} and E_{H_2} represent the total energy of the optimised structure with the adsorbed H atom, the optimised energy of the clean Si reconstruction and the total energy of gas-phase hydrogen molecule, respectively. A negative binding energy indicates an energy gain on adsorption; a positive binding energy indicates an energy loss. A diffusion pathway will be formed between two adjacent adsorption sites provided that the energy required to surmount any intervening energy barriers is comparable with the thermal energy acquired by the mobile atom. One selects the lowest energy sites close to the step edge and uses the NEB climbing-image optimizer to calculate the diffusion barriers. The NEB technique is discussed again on page 95.

4.4 Results

4.4.1 Reconstruction and characterization

As the (001) surface normal swings through a right-angle to the (110) direction it passes through a number of intermediate surface planes e.g. $[114]$, $[113]$, $[111]$ and $[331]$ (Battaglia, 2009). Each of these surfaces has its own reconstruction strategy but all contain the prototypical 3-coordinate surface atom found on the bulk-truncated (111) surface. Consequently, we can expect the reconstruction of the apex region to include the hexagonal pattern seen on the (111) unreconstructed surface. This can be seen in fig 4.5 (left), while the relaxed structure can be seen in fig 4.5 (right) which also shows the extended bond lengths expected in the presence of delocalized electrons. Away from the apex, the figure shows characteristic dimerization and buckling on the (100) surface and out-of-plane buckling of the zig-zag rows of the (110) surface. Distortion in the bulk structure is greatest in the region beneath the apex and extends to a depth of 20 Å. These observations suggest that the reconstruction is plausible, offering a sensible basis for the calculation of a model potential energy surface. An exhaustive characterization of the step edge reconstruction would require examination of larger structures in multiple orientations and would be unlikely to have a large effect on the diffusion barriers.

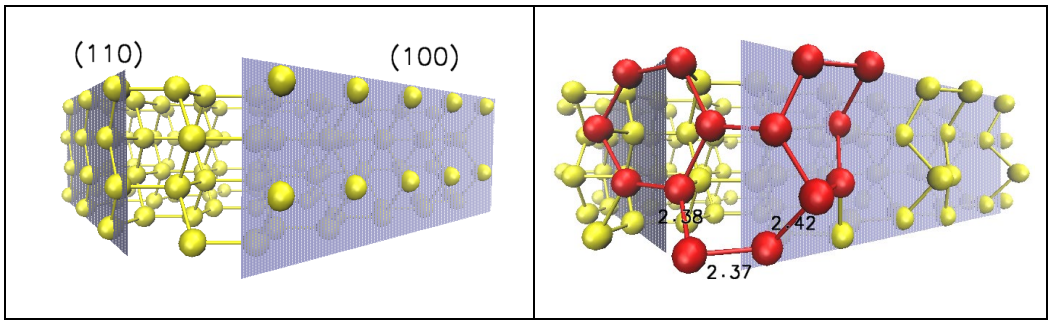


Figure 4.5 (Left) The apex of the 350-atom unoptimized chevron slab viewed from the (112) direction, showing the (100) and (110) surfaces. (Right) Optimized structure, showing (red) buckled (111)-like hexagonal patterns with extended bond lengths (Å). The bond length in bulk Si is 2.35 Å.

4.4.2 The potential energy surface

16 possible adsorption sites on the reconstructed surfaces were assessed. Six of these gave total energies falling within a 0.3 eV range while the remainder was at least 1 eV higher and not considered further (an increment of 1 eV reduces the adsorption probability by a factor of $\approx 10^4$). These values are shown in table 4.1 and the sites are depicted in fig 4.6.

Site	Relative energy	Binding energy
1	0.29	-1.66
2	0.06	-1.88
3	0.00	-1.95
4	0.02	-1.92
5	0.29	-1.65
6	0.13	-1.81

Table 4.1 Stability of H adsorption sites on the chevron surface. All values are in eV. Site labels are those shown in fig 4.6 (left).

Sites 3, 4 and 5 are situated on the chevron apex so a diffusion path consisting of the two hops $3-4$ and $4-5$ was considered. Each is analogous to the kinetics of a chemical reaction with a single transition state corresponding to the highest saddle point in the potential energy landscape lying between the end points. As noted above, the climbing image variation of NEB returns the energy at the highest saddle point and so a single-image NEB calculation per hop suffices in this case. Additional images can provide further points on the reaction path corresponding to the route taken by diffusing atoms, although the computational cost is considerable. Since only relative barrier heights are of interest, single-image calculations were performed.

The results of the NEB calculations are shown in table 4.2 and represented graphically in fig 4.7. Fig 4.6 (right) gives a rough indication of the actual diffusion path. The effective barrier for the path is the greater of the hop barriers and is asymmetrical, due to the differing starting energies. Diffusion from (100) to (110) has a barrier of 1.72 eV, while the reverse process has a barrier of 1.99 eV. These can be compared with published values of 1.66 eV $\pm$

$0.15\ \mathrm{eV}$ (Bowler et al., 1998) and $1.75 \pm 0.02\ \mathrm{eV}$ (Owen et al., 1996) for intrarow H diffusion on the (001) surface and $1.17\ \mathrm{eV}$ (intrarow) and $1.49\ \mathrm{eV}$ (interrow) on the $\mathrm{Si}(110)$ surface (Brázdová; Bowler 2011). The diffusion from the top of the pillar (the (001) surface) to the side of the pillar (the (110) surface) has a comparable barrier to diffusion on the (001) surface, while diffusion in the opposite direction has a significantly larger barrier.

End point	Energy	$\Delta E_{(011 \rightarrow 100)}$	$\Delta E_{(100 \rightarrow 110)}$
3	-1947.01		
barrier	-1945.44	1.57	1.55
4	-1946.98		
barrier	-1945.00	1.99	1.72
5	-1946.71		

Table 4.2 End point and NEB barrier energies for a two-hop diffusion pathway around the chevron apex. All values are in eV. Site labels are those shown in fig 4.6 (left). This data is shown graphically in fig 4.4.

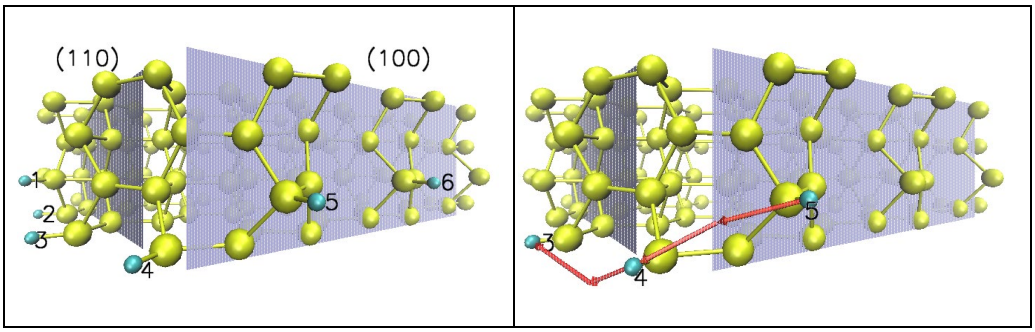


Figure 4.6 (Left) A composite representation of the 6 most stable absorption sites from 16 surveyed. Si atoms and H atoms yellow and blue, respectively. (Right) Approximate path followed by diffusing H atom (red) in NEB climbing image simulation.

4.5 Discussion and conclusions

The stability of hydrogen adatoms at positions near the intersection of the Si (100) and (110) surfaces, as might be occur during Si nanopillar growth by patterned ALE, has been investigated. A diffusion pathway around the corner was found and the barrier energies on it calculated. In this configuration, the hydrogen is more stable on the (110) face than the (100) face.

In the PALE context one is concerned with whether hydrogen will leave the top of the nanopillar for the sidewall or vice versa. The growth process involves a disilane gas source and leaves hydrogen on the sidewalls. Hydrogen migration onto the growth surface would interfere with subsequent STM lithography and possibly compromise the entire process. From the barriers it is clear that, at least for this configuration, diffusion from the growth surface of a nanopillar will only be activated once diffusion on the substrate is activated and

diffusion back onto the top will occur only at higher temperatures. The difference in barrier energy ($\approx 0.3\,\mathrm{eV}$) is significant in DFT terms and equivalent to a process temperature increase of $\approx 100\,\mathrm{K}$.

The actual diffusion rate ν (s^{-1}) can be estimated from the energy barrier by the Arrhenius equation:

$$\nu = \nu_{hop} \times e^{-\frac{\Delta E}{k_B T}} \tag{2}$$

where ν_{hop} is the attempt frequency, ΔE the energy barrier, k_B the Boltzmann constant and T the ambient temperature. ν_{hop} is generally found to lie in the range $10^{12} - 10^{13}\ \mathrm{s}^{-1}$ and since the rate expression is dominated the negative exponential energy term we can take $\nu_{hop} = 10^{13}\ \mathrm{s}^{-1}$ to get an upper bound. If the PALE process temperature of $550\,\mathrm{K}$ is assumed and adsorption sites are assumed to be occupied with a probability of one, then we could estimate a rate of $1.7 \times 10^{-3}\ \mathrm{s}^{-1}$ or approximately one diffusion event every 10 minutes *off* the nanopillar. The reverse process would be three hundred times less frequent at $550\,\mathrm{K}$.

We can extrapolate from this single event to an entire nanopillar, under the conditions used in PALE. A pillar with a side of $5\,\mathrm{nm}$ would have 100 edge sites available, generating a diffusion event off the pillar every 5 or 6 seconds, assuming that there was an empty site to reach. The reverse process produces a diffusion event every half hour. These results indicate that there would be net hydrogen migration off the pillar, which is a desirable outcome in PALE terms. There is little reason to be concerned about higher temperatures, as the process temperature of $550\,\mathrm{K}$ is chosen to avoid desorption of the hydrogen resist (it begins to show mobility at temperatures exceeding $600\,\mathrm{K}$). Although sidewall hydrogen diffusion may not be a critical issue in a PALE manufacturing process other obstacles remain. Chief amongst these is the formation of anti-phase boundaries (APBs) in the growth surface due to the collision of islands with different registry. These APBs can trap hydrogen beneath the surface, leading to surface roughening and a reduction in the depassivation yield from subsequent STM lithography. The reduction in yield can be mitigated by adjusting the STM parameters prior to each depassivation step, but the remanent hydrogen causes cumulative surface damage which halts epitaxial growth after 2 or 3 monolayers. The remediation of APBs and improvement of the quality of the silicon is beyond the scope of this thesis.

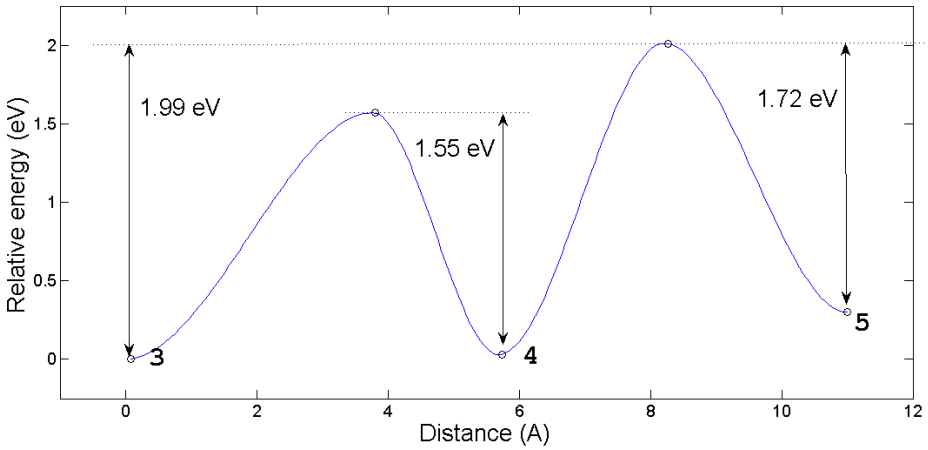


Figure 4.7 Diffusion barriers for a hydrogen atom traversing the PES in the apex region of the chevron slab. The adsorption sites are labelled as in fig 4.6 and the barrier energies derived from two NEB climbing-image calculations. The curve is a spline fit to the five data points. Zero energy at adsorption site 3 corresponds to a calculated value of -1947.01 eV. Distance travelled increases in the $(110) \rightarrow (100)$ direction.

Chapter 5

Alane adsorption and dissociation on the Si(100) surface

5.1 Background

Ever since the transistor was first developed in 1948, dopants have been used to control the characteristics of semiconductor devices. Although a relatively low dopant concentration (≈ 10^{13} atoms cm^{-3}) is sufficient to materially change substrate conductivity each successive reduction of device dimensions has required a corresponding increase in dopant concentration (Dennard, 1974). But concentration is ultimately limited by the overlap of the ground state wavefunctions of neighbouring impurity atoms and onset of the metallic phase (*metal-insulator transition*) at ≈ 10^{19} ions cm^{-3}.

To ensure reliable operation a device requires a statistically significant number (100s or 1000s) of dopant ions in its active region such as the MOSFET channel. If there are too few charge carriers unacceptable performance variations will arise. For example, a channel with dimensions $50 \times 50 \times 10\ \mathrm{nm}^3$, comparable with present-day devices, might contain as few as 100 carriers when strongly doped to a concentration of $5 \times 10^{18}\ \mathrm{cm}^{-3}$.

If dopant atoms can be confined to a 2-dimensional sheet with local concentration $N^{2D}\ \mathrm{cm}^{-2}$ then an equivalent bulk concentration $N^{3D} = (N^{2\mathrm{D}})^{3/2}\ \mathrm{cm}^{-3}$ is attained. This process (delta doping) requires accurate placement of the dopant atoms, achievable by interrupting substrate growth during MBE or CVD or by ion implantation and annealing. Recently, a complete PALE-based strategy for the fabrication of atomic scale silicon devices with donor (P) dopants has emerged (Simmons; Fuechsle, 2013). It includes:

- A full theoretical understanding of the deposition and incorporation processes beginning with the gaseous phosphine precursor;
- Electronic activation of the dopants by low-temperature Si epitatxial overgrowth, while minimizing diffusion and segregation of the dopant structure;
- Realization of a variety of functional donor doped Si devices including nanowires, quantum dots and a single-atom transistor.

Less progress has been seen with acceptor dopants. The variety of devices which can be fabricated with both p and n-type dopants is much greater than when only p-type is available. These might include p-n junction devices such as the tunnel FET or improved n-type devices that would benefit from an increased barrier potential around active elements e.g. qubit devices.

Historically boron has been used as an acceptor dopant, introduced either by ion implantation or through direct addition to molten silicon. However, it is unsuited to precision PALE on Si because its diborane precursor can deposit boron atoms in pairs and promotes H desorption from the surface (Wang, Hamers 1996; Wang et al., 1996). Its small size would cause a delta-doped layer to be strained (Sarrubi, 2010; Škereň et al., 2020), causing relatively fast diffusion within bulk Si and tending to smear out atomically precise dopant profiles. Aluminium, adjacent to silicon in the periodic table, may be a better choice. Unfortunately, the phosphine analogue alane (AlH_3) is not a useful precursor, existing in a solid crystalline form at room temperature and decomposing at higher temperatures. However, it can be synthesized by evaporating metallic aluminium in a molecular hydrogen stream at low pressures (Breissacher, Siegel 1964). Alternatively, the amine alanes are donor-acceptor complexes known to be viable precursors in thin film deposition of Al (Jones; Hitchman, 2009). The trimethylamine complex decomposes in the gaseous phase giving alane and the tertiary amines (Gladfelter et al., 1989), and it is plausible that this reaction would also be effective in the PALE setting. This work is motivated by the expectation that Al will emerge as a viable acceptor dopant for Si in the PALE fabrication process. This will complement P donor doping, increasing the range and functionality of molecular devices. The initial goal will be creation of Si structures with embedded delta-doped Al layers.

A survey of all possible adsorption and subsequent dissociation modes of the alane molecule on the Si(100) surface is attempted. Although this might seem to imply many configuration possibilities, the actual number (≈ 60) remains manageable because the H atoms are required to stay near the initial adsorption site at each dissociation. This approach is similar to that of (Warschkow, 2005) for phosphine adsorption and reflects the highly selective nature of PALE deposition. The survey reveals the relative stability of each intermediate configuration and the dissociation pathways that are energetically favoured. The results have been published (Smith; Bowler, 2017) and are the basis of this chapter.

5.2 Methods

5.2.1 Structural survey

We use DFT to survey all feasible AlH_n structures on Si(100). We show a progressive increase in stability as dissociation proceeds and characterize the more stable surface configurations using simulated STM and electron localization plots. A full kinetic analysis is presented in chapter 6. As noted above, the initial adsorption assumes that any required precursor reaction has already occurred and that free alane molecules are available within bonding distance of the substrate. There are four possible initial absorption sites as shown in fig 5.1.

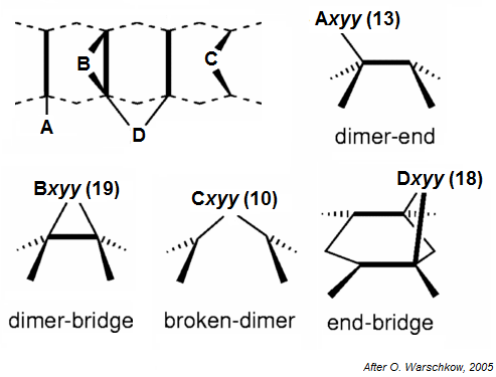


Figure 5.1: Perspective views of adsorption sites of AlH_3 on the Si(100) surface. Adatom A binds at a *dimer-end* position of a surface dimer; B binds to two Si atoms on the same dimer in the *dimer-bridge* position, leaving the dimer intact; *broken-dimer* position C is similar to B, but breaks the dimer and D binds to Si atoms on two adjacent dimers in the *end-bridge* position. Dissociation is modelled by removing an H from the adatom and placing it nearby. This creates a new surface configuration identified by appending a number *xyy* where x indicates the number of H atoms remaining bonded to Al, i.e. *x=3* represents the initial adsorption, *x=0* indicates a fully dehydrogenated Al atom. yy is an enumerator. The respective number of identified structures appears in parentheses.

5.2.2 Computational details

All calculations used density functional theory as implemented in VASP with the PBE GGA functional (page 32). The VASP PAW potentials for aluminium, silicon and hydrogen were used (page 41). These potentials describe both core and valence electrons and the files (POTCAR) were dated 4/5th January 2001 and 15th June 2001, respectively. A 400 eV energy cut-off was used. This value is required for proper operation of the aluminium PAW pseudopotential.

The convergence criterion for atom forces was set to 0.02 eV/Å and that for total energy to 10^{-6} eV. These parameters yield relative energies reliable to within 0.02 eV when the Brillouin zone sampling mesh is set appropriately. For the supercell employed here, energy values were found to converge with a $(3 \times 3 \times 1)$ Monkhorst-Pack mesh. These calculations used a quasi-Newton algorithm for ionic relaxation.

5.2.3 Supercell

The Si(100) surface was modelled on a slab of eight Si layers with a (4×2) surface cell reconstruction, separated by a 12 Å vacuum gap. This surface dimension $(15.36$ Å $\times$ 15.36 Å$)$ has been adopted in other studies of this kind (Brázdová; Bowler, 2011) and accommodates two dimer rows of buckled dimers (four in each row) at approximately $18°$ to the surface plane. The relatively large surface supports adsorption configurations spanning adjacent dimer rows. There is less agreement over optimum cell depth (Sagisake et al., 2107), and the chosen value is a compromise that achieves reasonable convergence and acceptable processing times. The experimental bulk Si lattice parameter $(5.431$ Å$)$ was

used and is within 1% of the PBE lattice constant. The bottom layer of Si atoms was left in bulk-like positions, terminated with pairs of hydrogen atoms, and fixed.

During optimization a single $\mathrm{AlH}_x + (3-x)\mathrm{H}$ ensemble is adsorbed on the surface while the deepest Si and H termination layers are constrained in fixed positions. Dissociation is modelled by progressively detaching atoms from the Al centre and placing them nearby on the surface. The energy change $\Delta E_{Al\,H_x}$ at each stage is calculated by:

$$\Delta E_{AlH_x} = E_{AlH_x+(3-x)H} - E_{Si(100)} - E_{AlH_3} \tag{3}$$

where $E_{Si(100)}$ is the energy of the clean optimized supercell, E_{AlH_3} the energy of an optimized alane molecule *in vacuo* and $E_{AlH_x+(3-x)H}$ the optimized energy of the supercell including the adsorbed AlH_x and dissociated $(3-x)\mathrm{H}$ species.

5.2.4 Electron localization function (ELF)

The ELF (Becke; Edgecombe, 1990) is a function of electronic density which is large in regions where electron pair density is high such as covalent bonds and lower in regions of delocalized electronic density. It provides a useful quantitative representation of the chemical bond in molecules and crystals (Savin et al., 1997), and is employed here to depict $\{\mathrm{H}_{0-3Al}\} \leftrightarrow \mathrm{Si}(100)$ interactions. The ELF function $\eta(\boldsymbol{r})$ can be computed from the orbitals as the definition is:

$$\eta(\boldsymbol{r}) = \frac{1}{1+(D/D_h)^2}, \tag{4}$$

$$D = \frac{1}{2}\sum_{1=1}^{N}|\nabla\psi_i|^2 - \frac{1}{8}\frac{|\nabla\rho|^2}{\rho}, \tag{5}$$

$$D_h = \frac{3}{10}(3\pi^2)^{2/3}\rho^{5/3}. \tag{6}$$

with the electronic density $\rho(r)$ given by:

$$\rho = \sum_{i=1}^{N}|\psi_I|^2, \tag{7}$$

and the sums are over the singly-occupied Kohn-Sham (or Hartree-Fock) orbitals $\psi_i(\boldsymbol{r})$. $D(\boldsymbol{r})$ is the probability of finding an electron near a reference electron of the same spin, and $D_h\big(\rho(\boldsymbol{r})\big)$ is the value of $D(\boldsymbol{r})$ for a homogeneous electron gas. It is interesting to note the same dependency on kinetic energy density (the Laplacian of the orbitals) that occurs in 'meta-GGA' functionals e.g. TPSS (Tao et al., 2003). The ELF formulation inverts $D(\boldsymbol{r})$ and rescales it with respect to the HEG. A low probability, leading to a high ELF, implies a localized electron and vice versa. A perfectly localized orbital, such as the H_2 bonding orbital, would have an ELF of 1. High ELF in an interatomic region can be interpreted as

covalent bonding, with any asymmetries attributed to bond polarity. The HEG represents a fully delocalized state with an ELF of 0.5. Values lower than 0.5 are more difficult to interpret but the ELF generally passes through zero between local maxima, termed *attractors (*Fuster et al., 2000). An isovalue of 0.8 has proven to be a useful bonding indicator in classical valence compounds.

For the high stability configurations, we show the ELF as contour plots in sections through the supercell, rendered by the author's MATLAB programs. For dimer-end, dimer-bridge and broken-dimer configurations the section is the vertical plane containing the Al atom and the dimer, unless otherwise noted. For the other configurations, the plane is usually horizontal or parallel to the dimer row. The chosen isovalues are separated by an interval of 0.2, with an additional contour in the high ELF region.

A complete set of ELF plots is available on figshare (Smith; Bowler, 2017).

5.2.5 Simulated STM images

Simulated STM images can show that a theoretical adsorption configuration has an electronic structure compatible with experimental appearance. Conversely, they can aid the identification of experimental images. Therefore, we provide topographical (constant current) images for the high stability configurations discovered in our survey. These have been prepared using the Tersoff-Hamann approximation (Tersoff; Hamann, 1985) as implemented by the bSKAN 3.3 program (Hofer, 2003) with graphics produced by the author's MATLAB programs. Under this approximation the tunnelling current is proportional to the local density of surface states at the centre of the STM tip, whose own electronic structure is not explicitly modelled.

We show representative simulated surface images for both positive (1.5V) and negative (-2.0V) bias voltages. The positive value indicates current flow into unoccupied surface states (electrons move from tip to surface) and the negative a flow from occupied surface states (electrons move from surface to tip).

A complete set of STM images is available on figshare (Smith; Bowler, 2017).

5.3 Results and discussion

5.3.1 Overview of the entire decomposition pathway

Some 60 configurations were evaluated, showing progressive increase in stability as dissociation proceeds. Fig 5.2 shows the calculated energies as columns of bars versus the dissociation stage horizontally. Stability of a configuration depends on the nature of the Al-Si bonding, and the local disposition of the adsorbed H atoms. Eight incorporation configurations are shown to demonstrate feasibility of these structures.

In each group, there are a few structures notably more stable than any other in the same group, and a thermodynamically favoured dissociation pathway is likely to involve these configurations. We have characterized these *high stability* structures using ELF and simulated STM plots and show their relative energies and bond lengths in Table 5.1. Of

course, some structures may be rendered inaccessible by kinetic considerations, and an analysis based on DFT NEB (nudged elastic band) calculations is the subject of chapter 6.

	Config.	ΔE (eV)	Al-Si (Å)	Si-Si (Å)
H3Al: initial adsorption	**D301**	-0.84	2.58/*(4.41)*	2.37/2.36
	A301	-0.78	2.61	2.39
	B301	+0.01	*(4.01)/(4.15)*	2.36
H2Al: first dissociation	**D205**	-2.31	2.55/2.62	2.35/2.40
	B201	-2.11	2.45/2.81	2.43
	B206	-2.08	2.48/2.54	2.54
	A201	-2.08	2.49	2.44
HAl: second dissociation	**D106**	-3.27	2.49/2.49	2.42/2.43
	D103	-3.22	2.46/2.51	2.38/2.42
	D104	-3.02	2.44/2.58	2.37/2.50
	B101	-2.93	2.40/2.43	2.48
	A103	-2.49	2.58/2.63	2.48/2.52
	C106	-2.24	2.43/2.46	*(3.90)*
Al: third dissociation	**D004**	-3.85	2.48/2.49	2.42/2.46
	D002	-3.67	2.47/2.47	2.41/2.42
	B008	-3.67	2.42/2.60/2.63	2.40/2.53
	D001	-3.60	2.46/2.47	2.42/2.42
	D005	-3.57	2.48/2.48	2.38/2.39
	A001	-3.28	2.61	2.41
	C004	-3.07	2.37/2.38	*(4.75)*
Si: incorporation	**D059**	-3.84	2.42/2.42/2.44	
	D058	-3.74	2.39/2.39/2.44	
	D057	-3.71	2.39/2.39/2.45	
	D056	-3.69	2.40/2.40/2.47	
	C050	-3.29	2.40/2.41/2.45	

Table 5.1: Calculated relative energies and bond lengths for structures identified in fig 5.2 and discussed in the text. In the initial adsorption and dissociation cases, the Al-Si column gives the length of the surface bond(s) with the adsorbed Al. For the incorporation cases the lengths of the two subsurface bonds are given, followed by the length of the Al-Si heterodimer. Column Si-Si gives the length of the adsorbing dimer(s). For comparison, dimer length on the reconstructed bare Si(100) surface is ≈ 2.36 Å in this supercell. Values in parentheses indicate inter-atomic distances, i.e. the absence of bonding.

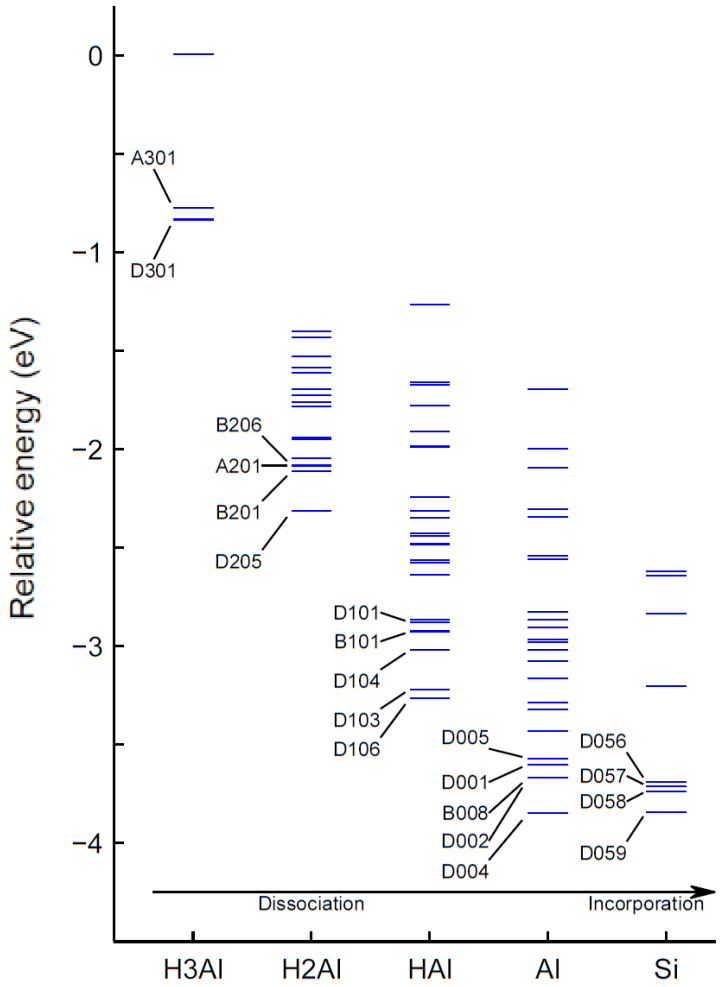


Figure 5.2: Overview of relative energies for alane dissociation and incorporation configurations considered in the survey. Configurations are grouped on the horizontal axis by the degree of dissociation. High stability (low energy) configurations are labelled using the numbering scheme outlined at fig 5.1. Configuration energies are relative to the sum of bare surface and free alane energies. A full listing of the structures and relative energies is available on figshare (Smith; Bowler, 2017).

5.3.2 The Si(100) surface

Fig 5.3 shows ELF and simulated STM output for the bare, reconstructed Si(100) surface. The alternately buckled dimers are $0.2\ \mathrm{eV}$ more stable than when parallel to the surface plane and are the most stable reconstruction possible. At ambient temperatures, the dimers 'flip' at a rate greater than the STM can accommodate, and so the STM images shown may not be observed. However, the presence of an adsorbing Al or H atom will be sufficient to 'pin' the dimer in the buckled configuration, justifying use of the reconstruction.

The STM filled state plot shows that reconstruction eliminates one dangling bond and concentrates electronic density at the 'up' dimer end, and dimer length is found to be $\approx 2.36\ \mathrm{Å}$ in this supercell. The filled state STM plot shows the DOS centred on the surface atoms. As the ELF is determined over occupied states it might be expected to correspond with the filled-state STM image, although no theoretical basis has been established for this.

However, the plot reveals large attractor regions above the 'up' dimer ends with ELF values exceeding 0.9, characteristic of a non-bonding (lone) electron pair.

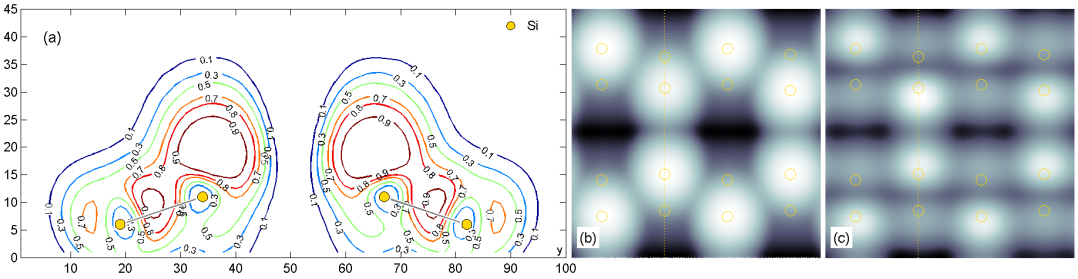


Figure 5 3: ELF plot and simulated STM images for the bare reconstructed Si(100) surface. The ELF plot (a) is perpendicular to the surface in a plane containing a pair of dimers (indicated). The simulated STM images (b) and (c) correspond to tip bias voltages of $-2.0\ V$ and $+1.5\ V$ respectively, and the yellow dotted lines mark the position of the contour plane. The superimposed yellow circles indicate Si dimer atoms. The horizontal scale is $100 \approx 15.36$ Å.

5.3.3 Initial adsorption: H3Al↔Si(100)

Stable configurations were discovered at dimer-end, dimer-bridge and end-dimer sites. No stable broken-dimer configuration was found, with an H atom tending to detach and migrate to the adjacent dimer row or adopt a central position 'buried' beneath the dimer. The dimer-bridge configuration (B301) showed a slight surface repulsion and was not considered further. Fig 5.4 shows the two remaining structures and their relative energies and fig 5.5 shows the ELF plot and simulated STM images for the dimer-end configuration A301 where a bond with stability -0.78 eV was found with the up-atom and the Al atom in pyramidal coordination. No bonding was possible with the down-atom.

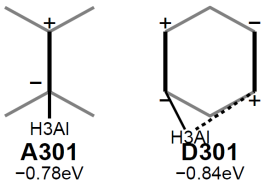


Figure 5.4: Schematic representation of initial adsorption configurations after structural optimization, showing relative energies. Dimers are represented by heavy vertical lines; the '-/+' signs indicate the 'up/down' ends, respectively. All configuration files are available on figshare (Smith; Bowler, 2017).

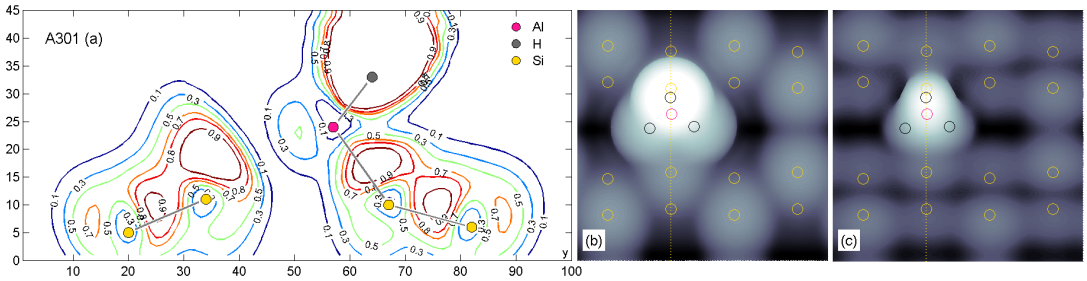


Figure 5.5: ELF plot and simulated STM images for the initial dimer=end configuration A301. The ELF plot (a) indicates the Al-Si, Al-H and Si-Si bonds. The high-ELF region surrounding the H ligand shows the polar nature of the Al-H bond. In the STM images (b) (c), the Al, Si and H atom locations are superimposed and yellow dotted lines mark the position of the contour plane. Images (b) and (c) correspond to tip bias voltages of -2.0 V and +1.5V respectively.

The electronic structure of the Al atom is $[\mathrm{Ne}]3s^2 3p^1$. The alane molecule valence shell contains six electrons and a further two would complete the octet. These are provided by the excess electronic density at the surface dimer 'up' end and form a dative bond with alane acting as a Lewis acid and the substrate as Lewis base. The Al-Si bond length of $2.61\ \text{Å}$ obtained here can be compared with $\approx 2.08\ \text{Å}$ calculated for the dative Al-N bond in ammonia alane (Marsh, 1995). Although both have *sp* hybridization the latter has greater *s* character due to the H ligands of the ammonia. This, together with the greater electronegativity of the nitrogen atom, account for the shorter Al-N bond. However, the ELF plots of figs 5.3 and 5.4 are consistent with the Lewis adduct model.

The adduct model implies that the end-bridge configuration possessing two surface bonds is unfeasible. This was confirmed by our optimization of configuration D301 which resulted in an asymmetrical configuration with only one bond substrate bond (see Table 1) and the ELF plot (not shown) confirmed the presence of just a single bond. The increased stability (0.06 eV) compared with the dimer-end configuration is due to a relative rotation of the alane molecule which was not explored during the optimization of the dimer-end configuration.

These results show the initial alane adsorption modes are analogous to those of phosphine, which bonds into the 'down' atom at dimer-end sites, but is unstable in the dimer-bridge, broken-dimer and end-bridge configurations (Warschkow, 2005).

5.3.4 First dissociation: {H2Al+H}↔Si(100)

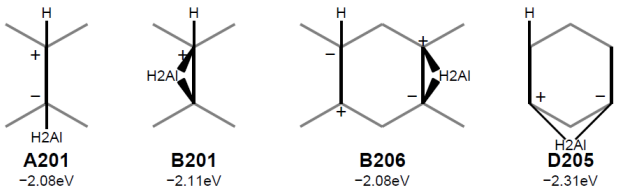


Figure 5.6: Schematic representation of high stability configurations after the first dissociation, showing and relative energies

At this stage the high stability configurations (fig 5.6) show energies decreasing by $1.3 - 1.5$ eV below the initial adsorption, with end-bridge configuration D205 the most stable. In the absence of kinetic barriers, these large margins suggest the initial configurations will be relatively short-lived on the surface. The end-dimer A201 and end-bridge D205 configurations were the most stable of their kind by margins of 0.3 and 0.5 eV respectively but two dimer-bridge configurations B201 and B206 had similar energies, differing only in the placement of the migrating H atom. A broken-dimer configuration appeared with the H atom placed beneath the dimer level in an apparently three-centred bond, but it was at least 0.5 eV less stable than the high stability group and not considered further. Fig 5.7 shows the ELF plots and simulated STM images for configurations A201, B206 and D205.

The end-dimer configuration A201 has the Al atom in trigonal planar coordination and an Al-Si bond of length of 2.49 Å, noticeably shorter than that of the A301 configuration. We surmise that the ligands, now having predominantly sp^2 hybridization, provide improved overlap with the surface orbitals. This can be seen in the ELF plot as an enlarged inter-nuclear region with value 0.9 or greater. The effect of the adsorbed H on the down-dimer atom is to level the dimer, with both atoms making 2-centre, 2-electron bonds.

In the dimer-bridge and end-bridge configurations Al adopts a tetrahedral configuration, although the bond angles are far from ideal. The end bridge configuration is the more stable by a margin of 0.2 eV. In the dimer-bridge cases B201 and B206 the Si-Si dimer bond lengths are 2.43 Å and 2.54 Å respectively, with both surface atoms making 2-centre, 2-electron bonds with the metal. In the end-bridge configuration D205 the dimer bond lengths are 2.35 Å and 2.40 Å, closer to the bare surface value and indicating that the stability gain occurs through sharing the adsorption stress across surface dimers.

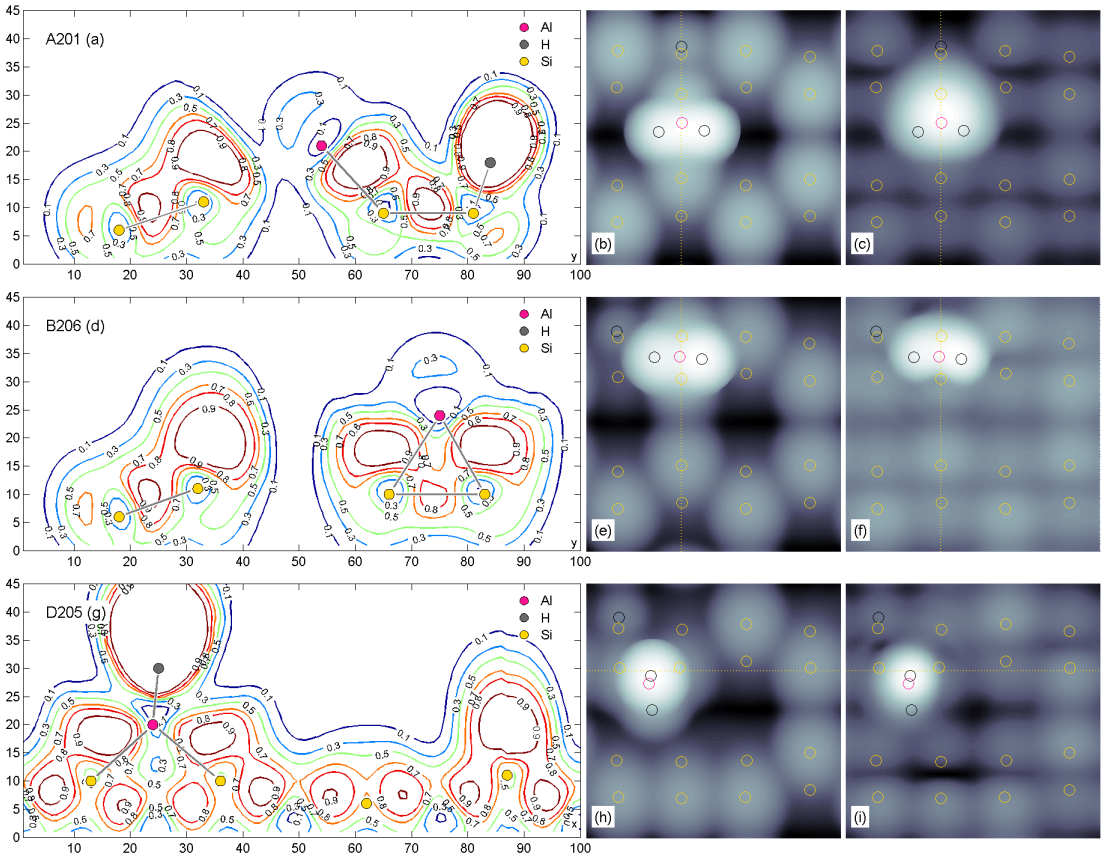


Figure 5.7: ELF plots and simulated STM images for first dissociation configurations A201, B206 and D205. In the ELF plots (a) and (d) the contour map plane passes through the surface dimers perpendicular to the dimer row, and in (g) the plane is parallel to the dimer row. In the STM images, the Al, Si and H atom locations are superimposed and yellow dotted lines mark the position of the ELF contour plane. Images (b, e, h) and (c, f, i) correspond to tip bias voltages of -2.0 V and +1.5V respectively.

5.3.5 Second dissociation: {HAl+2H}↔Si(100)

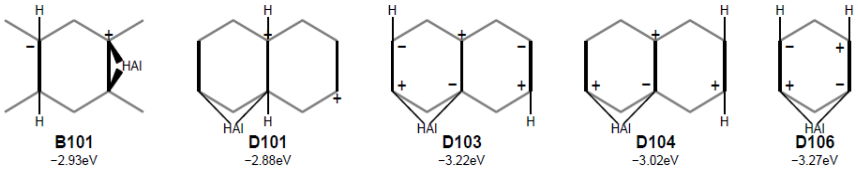


Figure 5.8: Schematic representations of high stability configurations after the second dissociation, showing relative energies. These are the 5 most stable of the 28 configurations examined at this stage. The 10 most stable configurations were all bridged.

The loss of a further H ligand increased stability by up to 0.9 eV, depending on the surface configuration. In the absence of significant kinetic barriers, this energy loss would prompt dissociation in the PALE environment. The high stability configurations are all end or dimer-bridge (see fig 5.8); these are the configurations likely to appear on a pathway towards complete dissociation and incorporation. We take D103 as representative of the end-bridge configurations and show ELF plots and simulated STM images for B101 and D103 in fig 5.9.

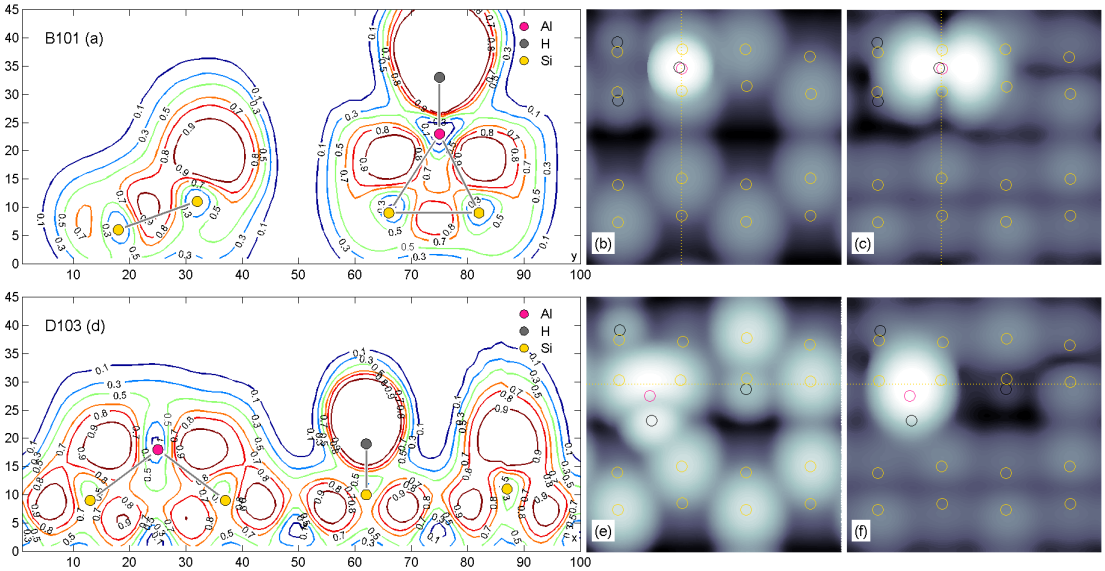


Figure 5.9: ELF and simulated STM plots for second dissociation configurations B101 and D103. For B101 the ELF contour map (a) plane passes vertically through the surface dimers. For D103 (d) the plane is parallel to the dimer row. In the STM images, the Al, Si and H atom locations are superimposed and yellow dotted lines mark the position of the ELF contour plane. Images (b, e) and (c, f) correspond to tip bias voltages of -2.0 V and +1.5V respectively. These configurations are ≈ 0.6 - 0.9 eV more stable than at the previous stage.

The ELF plots are like those of the bridged configurations of the previous stage (see fig 5.7) with the Al atom now adopting a trigonal planar, rather than a tetrahedral coordination. In the dimer-bridge case B101 the adsorbate bonds shorten to 2.40 Å and 2.43 Å compared to 2.48 Å and 2.54 Å in B206, allowing the dimer bond to shorten to 2.48 Å from 2.54 Å. This improved bonding can be attributed to the increased *s* character of the adsorbate bonds feeding into the dimer bond. As before, the sharing of surface stress in the end-bridge configuration D103 is responsible for its additional (≈ 0.3 eV) stability.

The most stable end and broken-dimer configurations are almost 0.5 eV less stable and are depicted in fig 5.10. Although their relative stabilities indicate they are unlikely to participate in a dissociation pathway they are of interest because they show the HAl fragment preserving a trigonal planar coordination with the substrate surface.

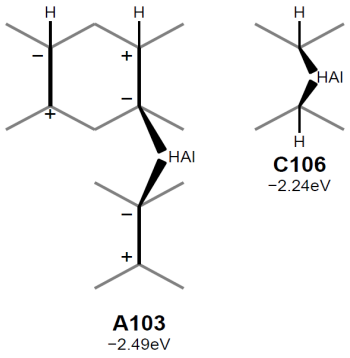


Figure 5.10: Schematic representation of the most stable end and broken-dimer configurations after second dissociation. Structural optimization of dimer-end configuration A103 has moved the HAl fragment to a position bridging dimer rows.

The corresponding ELF plots are shown in fig 5.11. Configuration A103 in fig 5.11 (a) shows the Al atom located in a trigonal planar coordination between dimer rows, bridging to an up-dimer atom in each. Although the Si-Al-Si bond angle is a near perfect $119°$ the lack of stability is due to the elongated adsorbate and Si-Si dimer bonds of this configuration (Table 5.1). Similar results were seen in several other bridged-row configurations in the survey. In the broken-dimer configuration C106 (b) the Al-Si bonds are shorter, but loss of the Si-Si dimer bond outweighs any gain in stability.

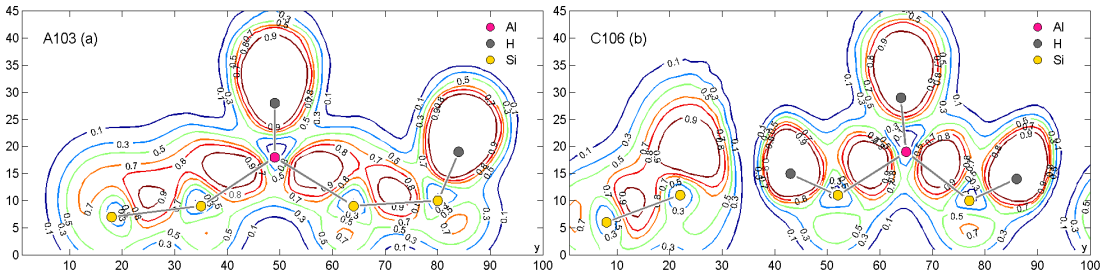


Figure 5.11: ELF plots of end-dimer (a) and broken-dimer (b) configurations after the second dissociation. Both show the HAl fragment in trigonal planar coordination with the Si(100) surface, bridging a single dimer row (C106) or adjacent tows (A103). These are the most stable configurations of their type, but are ≈ 0.5 eV less stable than any bridged and end-dimer configuration at this stage.

5.3.6 Third dissociation: {Al+3H}↔Si(100)

27 configurations were examined; all types were represented but the eight most stable were all the end or bridged-dimer variety. The most stable configuration D004 gains approximately 0.6 eV stability over its counterpart at the previous stage; a smaller energy loss than was seen in the first and second dissociations. Several configurations in the survey had increased energies, reflecting the reduced coordination possibilities available at this

stage. The five most stable span configurations span an energy range of less than 0.2 eV and are depicted in fig 5.12.

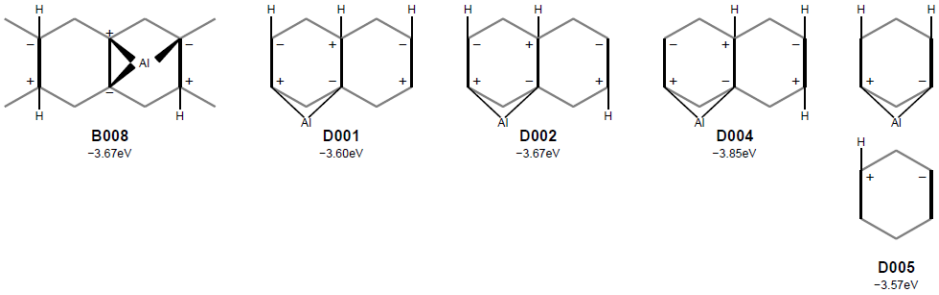


Figure 5.12: Schematic representations of high stability configurations after the third dissociation. In configuration B008 structural optimization has moved the Al atom from its starting dimer-bridge position to a mid-dimer location making three surface bonds. In configuration D005 a H atom has been placed on the adjacent dimer row, but the Al does not bridge the rows.

After optimization, the dimer-bridge configuration B008 had the Al adatom located between adjacent dimers, adopting a trigonal pyramidal coordination with three surface bonds. To illustrate this the ELF plot fig 5.13 (a) is taken in the horizontal plane containing the Al atom, above the dimers and at roughly the same elevation as nearby H atoms. Attempts to induce a square planar Al configuration, with four surface bonds and no surface H were unsuccessful. The bright STM images in the unfilled state images Figs 13 (c) and (f) reflect the adsorbate's vacant *p* orbital in this coordination.

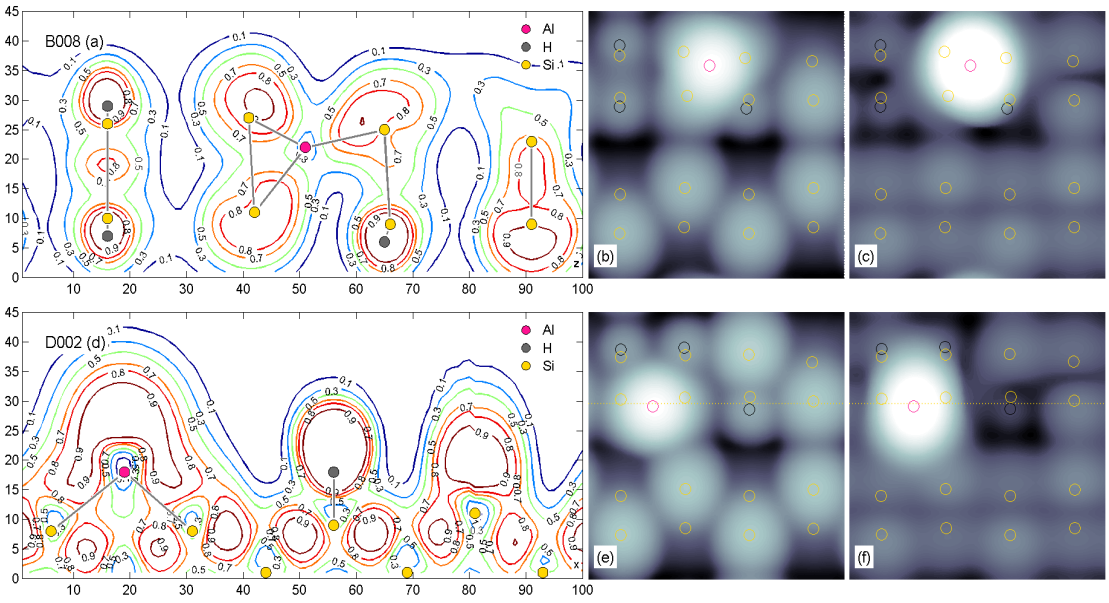


Figure 5.13: ELF plots and STM images of bridged and end-dimer configurations after the third dissociation. The dimer-bridge ELF plot B008 (a) is taken in a horizontal plane (parallel to the surface) and shows the Al atom with three surface bonds. The end-bridge plot D002 (d) is taken in a vertical (perpendicular to the surface) plane. In the STM images, the Al, Si and H atom locations are superimposed and yellow dotted lines mark the position of the ELF contour plane. Images (b, e) and (c, f) correspond to tip bias voltages of -2.0 V and +1.5V respectively. These configurations are ≈ 0.3 - 0.6 eV more stable than at the previous stage.

The four end-bridge configurations D001, D002, D004 and D005 are similar in character, differing in H placement only, and we take D002 as representative. ELF and simulated STM images for this configuration are shown at fig 5.13 (d), (e) and (f). The ELF plot fig 5.13 (d) is taken perpendicular to the surface and shows the Al adatom in trigonal planar coordination with two surface bonds and a large hybridized lone-pair region above.

The most stable end (A001) and broken-dimer (C004) configurations are shown schematically at fig 5.14. Configuration A001 has Al and three H atoms adsorbed on adjacent dimers, saturating them. The corresponding ELF plot at fig 5.15 (a) shows the Al veering along the trench between the dimer rows, but not bridging them as was seen for configuration A103 (fig 5.11 (a) above). Here the Al-Si bond length of 2.61 Å is identical to that found in the initial adsorption case A301 with similar lengths in the respective Si-Si dimers (Table 1). This suggests the same dative covalent character for the surface bond, with the unpaired Al valence electrons arranging themselves to maximize mutual repulsion. However, the single surface bond means that the configuration ≈ 0.3 eV less stable than any of the bridged modes. Several other mid-trench configurations were tried, but none proved particularly stable.

The broken-dimer configuration C004 at fig 5.15 (b) has a perfectly linear Al coordination with predominantly *sp* hybridization with Al-Si bond lengths of ≈ 2.37 Å, the shortest in the survey. It is interesting that this is a minimum energy configuration even though the Al-Si bonds do not pass through the regions of highest ELF. However, elimination of the surface dimer prevents any gain in overall stability, yielding a configuration ≈ 0.5 eV less stable than any bridged mode at this stage.

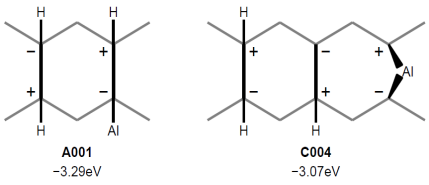


Figure 5.14: Schematic representation of most stable end and broken-dimer configurations after the third dissociation.

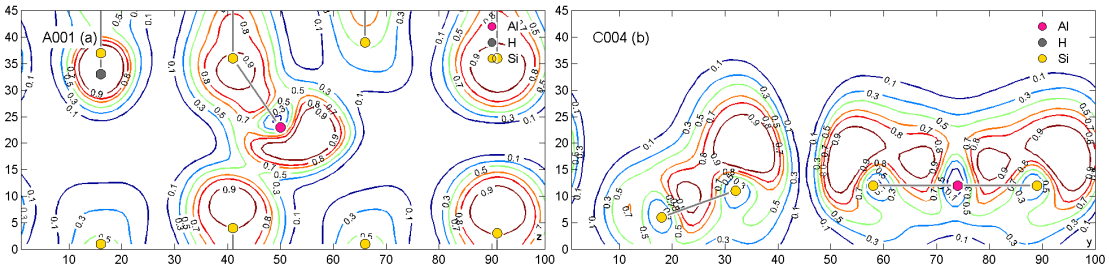


Figure 5.15: ELF plots of end-dimer (a) and broken-dimer (b) configurations after the third and final dissociation. The plot for the end-dimer configuration (A001) is taken parallel to the surface through the midpoint of the Al-Si bond. The plot for the broken-dimer configuration C004 is taken in a vertical plane containing the dimer atoms. These are the most stable configurations of their type but are respectively $\approx 0.3\ eV$ and $0.5\ eV$ less stable than bridged and end-dimer configurations at this stage.

5.4 Incorporation

In the PALE process the surface reaction terminates when all unpassivated bonding sites become occupied, either by precursor fragments or hydrogen adatoms. The dopant atoms must then be incorporated into the surface as Si-Al heterodimers, prior to the deposition of further Si layers. The replacement of an Si-Si dimer by the heterodimer involves the breaking of surface bonds and requires elevated temperatures. Successful incorporation would result in the appearance of ejected Si atoms as surface adatoms and could be confirmed by STM examination. After ejection from the surface the Si adatom could reside in any one of the three bridged sites B, C or D and a systematic survey of all heterodimer structures having three adsorbed H, an incorporated Al and an Si adatom is beyond the present scope. Instead we have optimized a small number of configurations of each type to illustrate the energetics of Al incorporation.

We examined eight incorporation configurations. Each has an ejected Si adatom with two surface bonds and an incorporated Al forming a Si-Al heterodimer. Their relative energies appear in Table 5.1 and are represented graphically in the rightmost column of fig 5.2. The four configurations with the greatest stability were of the end-bridge variety and are shown schematically at fig 5.16. They differ only in the placement of H atoms and fall within a 0.15 eV energy span. The most stable (D059, -3.84 eV) has almost the same stability ($\Delta E = 0.004$ eV) as configuration D004 at the final stage of dissociation. This margin is less than DFT accuracy and would result in a theoretical 50% incorporation assuming both states were equally stable and both kinetically accessible.

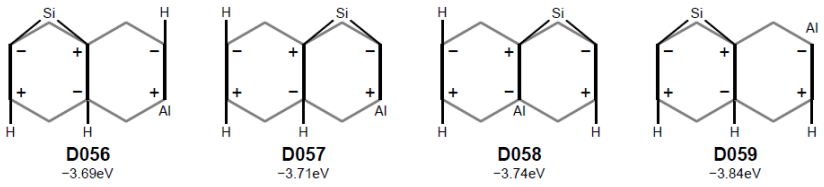


Figure 5.16: Schematic representation of high stability configurations after incorporation.

We take configuration D058 as representative and show ELF and simulated STM plots at fig 5.17. Although the Al atom replaced an 'up' Si it becomes the 'down' atom after incorporation. It has a pyramidal coordination with two subsurface bonds of length 2.39 Å and the heterodimer of 2.44 Å. The adjacent dimers are levelled. The ELF plot fig 5.17 (a) confirms the covalent character of these bonds. The filled-state STM image fig 5.17 (b) shows the absence of a dangling bond.

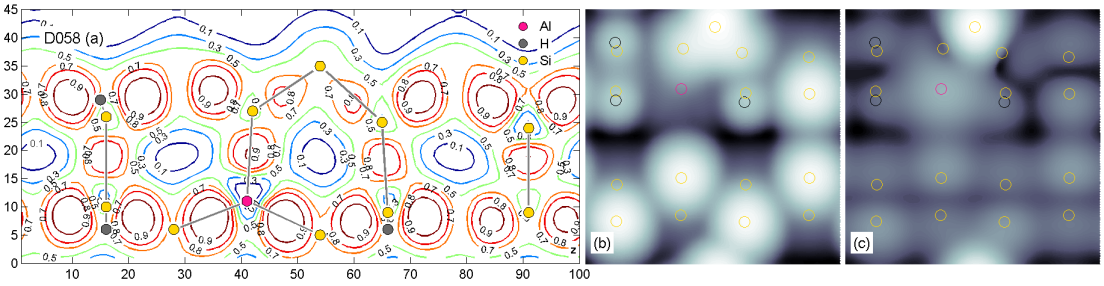


Figure 5.17: ELF and simulated STM images for incorporation configuration D058. The ELF plot is taken in the horizontal plane containing the Al atom. The STM images have the locations of the Al, Si and H atom locations superimposed. Images (b) and (c) correspond to tip bias voltages of -2.0 V and +1.5V respectively.

5.5 Conclusion

DFT has been employed to study the structure and energetics of the AlH_x species which come from the adsorption and dissociation of AlH_3 on the Si(100) surface, also considering several incorporation scenarios. A progressive, though declining, gain in stability is found as the dissociation and incorporation proceeds. The initial surface bond is dative and tetrahedral with the adsorbate fragment adopting trigonal geometries as dissociation proceeds. At each stage, high stability structures are likely to occur on any dissociation pathway and dimer bridging dominates. Structure have been characterized using ELF plots and simulated STM images to aid experiment.

As noted above, the methodology of this chapter follows that of (Warschkow et al., 2005) who offered a rigorous DFT study of phosphine adsorption and dissociation on a confined Si(001) surface area. The conclusion, that adsorbed PH_3 undergoes a progressive deprotonation through PH_2+H, PH+2H and P+3H remains unchallenged, and the intermediate forms seen in the accompanying STM observations were successfully assigned to the adsorption structures of fig 5.1 above. Warschkow found PH_3 was likely to adsorb intact and calculated a stability gain of ~ 0.6 eV, confirming an earlier finding of (Hamers; Wang, 1996) that the molecule bonds datively with the electron-poor 'down' atom of the buckled substrate dimer whereas we find AlH_3 bonding datively with the 'up' atom, with a gain of ~ 0.8 eV. The thermodynamics of AlH_3 decomposition, i.e. an overall calculated stability gain of ~ 3.0 eV are broadly similar to PH_3 which gains an equivalent ~ 2.4 eV. However, the adsorbed P atom stabilizes by a further ~ 0.5 eV on incorporation while the respective Al configurations are likely to be metastable. Both dopants are then in end-bridge configurations, D2 in Warschkow's notation and D004 in ours. P incorporation was already known to occur at an annealing temperature of ~ 350 C (Curson; Schofield et al., 2004) and

the process was not seen as problematic. Moreover, the lithographic H atoms of PALE could be sacrificed in the construction of a planar device, where incorporation is followed by Si overgrowth. This technique was employed in the precision doping of a P atom transistor (Fueschle et al., 2012), a significant milestone en route to the Australian group's ultimate objective, a silicon-based qubit memory.

The ability to incorporate acceptor dopants as well as donors in Si(001) with atomic precision should significantly advance the capabilities of patterned ALE. It opens the possibility of p-n junctions fabricated with atomic precision, as well as local control of the electrostatic potential using both positive and negative dopant ions. We keenly anticipate experimental measurements of these structures as a first realisation of this.

Chapter 6

Reaction paths of alane dissociation on the Si(100) surface

6.1 Background

In the previous chapter, alane (aluminium hydride AlH_3) was proposed as a suitable precursor for Al deposition as an acceptor dopant. Although difficult to synthesize, an energetic analysis implies that it will adsorb and dissociate on the Si(100) surface, yielding Al bonded in dimer-bridging modes. It was assumed that the H ligands of alane would detach sequentially, re-adsorbing in the immediate vicinity as should occur in a PALE process. Some surface configurations having an incorporated Al atom were also investigated. Many were less stable than the bridging configurations available as starting points (this relative stability is also seen with P incorporation (Warschkow, 2007)). In the absence of kinetic barriers, one might expect reversal of the incorporation reaction, unless another forward pathway leading to a configuration with better stability is available.

This chapter now resolves these results by providing a kinetic analysis of the available dissociation and incorporation pathways of alane on the Si(100) surface. DFT calculations yield the activation energies and hence the expected reaction rates of each step under PALE process conditions. The maximum energy (transition state) configurations are shown in diagrammatic form and an incorporation procedure based on removal of dissociated H and Si surface diffusion, prior to incorporation.

These findings have been published (Smith; Bowler 2018) and together with the earlier results provide a testable route to Al incorporation in the Si(100) surface to support experimental work in this area.

6.2 Methods

6.2.1 Terminology

We continue with the configuration naming scheme introduced earlier and summarized at fig 1. A pathway segment is defined by a starting and ending configuration e.g. A301-B202 is a path between the initial dimer-end adsorption configuration A301 and the bridged-dimer first-dissociation configuration B202. Fully dissociated configurations (having a bare Al adatom) have zero as the second character of the name e.g. B002, D004. Incorporation configurations are categorized as end-bridge (because they give rise to an Si adatom in that configuration) and are numbered sequentially from 50 upwards, i.e. D050, D051 and so on.

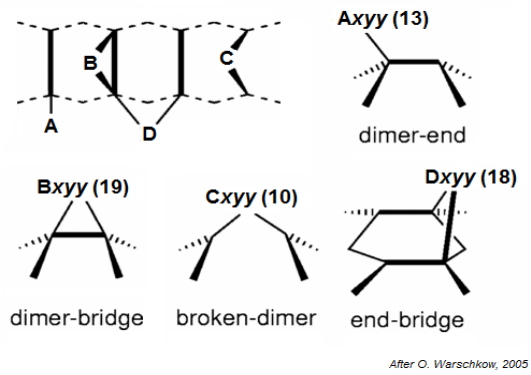


Figure 6.1: Perspective views of adsorption sites of AlH_3 on the Si(100) surface, reproduced from [7]. Adatom A binds at a *dimer-end* position of a surface dimer; B binds to two Si atoms on the same dimer in the *dimer-bridge* position, leaving the dimer intact; *broken-dimer* position C is like B, but breaks the dimer and D binds to Si atoms on two adjacent dimers in the *end-bridge* position. Dissociation is modelled by removing an H from the adatom and placing it nearby. This creates a new surface configuration identified by appending a number *xyy* where x indicates the number of H atoms remaining bonded to Al, i.e. *x=3* represents the initial adsorption, *x=0* indicates a fully dehydrogenated Al atom. yy is an enumerator. The respective number of identified structures appears in parentheses.

6.2.2 Computational details

The supercell and VASP parameterization introduced in chapter 5 were retained for these calculations.

Transition state search was performed using the climbing image nudged elastic band (CL-NEB) method (Henkelman et al., 2000) as implemented in the VASP Transition State Tools (VTST). The theoretical basis of the method was provided at page 51 and it is known to yield accurate barrier energy values (Klimes, 2010). The calculation was initialized using by six intermediate images on each path segment, obtained by linear interpolation from the minimum-energy end points. The convergence criterion for atom forces was set to a maximum of $0.05\ \mathrm{eV/\AA}$, and the maximum number of image optimizations was set to 200 (see the following section). We used the FIRE algorithm (Bitzek, et al., 2006) to optimize the intermediate images. This is one of several force-based optimizers that could have been used and is known to perform well in the VTST environment.

6.2.3 NEB convergence considerations

In some path segments, images constructed by linear interpolation can result in atomic trajectories rather far from the MEP. When inter-atomic distances become small, large repulsive forces are generated, and the NEB algorithm can sample high-energy regions of the potential energy surface without ever satisfying the convergence conditions. When this situation arose the trajectory was adjusted manually, while the force criterion was progressively increased from its initial setting of $0.02\ \mathrm{eV/\AA}$ to a maximum of $0.05\ \mathrm{eV/\AA}$. If convergence could not be achieved, the segment was discarded.

Difficulties also arise when the dissociating H in an interpolated image approaches that in another known configuration. Since the latter is located at a local energy minimum subsequent movement will be confined. The resulting MEP must then traverse steep energy gradients as the spurious Si-H bond is broken, and the calculation may again terminate without result. In most such cases the intermediate configuration appeared as an end-point on other MEPs, so no further action was needed. In others, the intermediate configuration showed a small gain in stability, attributed to the shallow valleys in the energy surface associated with rotational movement of the H ligands. In these cases, the intermediate configuration was re-optimized and replaced an original survey point.

6.2.4 Activation energy and reaction rate

The energy difference between the initial configuration and the highest saddle point on an MEP is the activation energy E_A which is related to the reaction rate k by the Arrhenius equation:

$$k = \nu e^{-E_A/k_B T} \tag{8}$$

where ν is the attempt frequency, k_B the Boltzmann constant and T the prevailing temperature. This simple formulation arises from harmonic transition state theory (Keeler, Wothers 2003) and assumes the reaction rate is sufficiently slow for a Boltzmann energy distribution to be established in the reactants and neglects quantum effects such as zero-point energy and tunnelling. Since the rate is dominated by the exponent term it is common to use an estimate of ν (e.g. $\nu = 10^{12} - 10^{14}\ s^{-1}$). An expression for ν in the harmonic approximation is given on page 51.

In table 6.1 below we show the activation energies yielding a single transition attempt over various timescales for range of process temperatures. These are estimates based on the Arrhenius rate using the boundary values of ν given above and may assist the reader in determining the feasibility of the reactions to be presented later.

PALE process temperature (K)	Activation energy (eV) causing 1 attempt in interval:		
	1 second	1 minute	1 hour
150	0.36-0.42	0.41-0.47	0.46-0.55
200	0.48-0.56	0.55-0.63	0.62-0.70
250	0.60-0.69	0.68-0.78	0.77-0.87
300	0.71-0.83	0.82-0.94	0.93-1.05
350	0.83-0.97	0.96-1.10	1.08-1.22
400	0.95-1.11	1.09-1.25	1.23-1.39

450	1.07-1.25	1.23-1.41	1.39-1.57
500	1.19-1.39	1.37-1.57	1.54-1.74
550	1.31-1.53	1.50-1.72	1.70-1.92

Table 6.1 Calculated activation energies (from the hTST/Arrhenius equation with $\nu = 10^{12} - 10^{14}\ s^{-1}$) to give rates of 1 per second, 1 per minute and 1 per hour (e.g. a reaction with $E_A = 1.00$ eV would be activated within 1 hour at 300 K). At temperatures greater than 550 K the H passivation layer, essential to the PALE process, becomes mobile and ultimately desorbs from the substrate.

6.3 Results and discussion

6.3.1 Reaction pathways

In chapter 5, approximately 70 surface configurations at various stages of dissociation were evaluated. Although one expects any minimum energy pathway (MEP) to traverse the more stable configurations at each stage, many sterically feasible segments involving end configurations of lower stability were nevertheless examined. A segment was considered feasible if it possessed an unobstructed H migration route but did not call for re-diffusion of previously dissociated H and the high energy levels needed to break an Si-H bond. Under these constraints approximately 110 feasible segments were available, of which 40 yielded converged results in the MEP calculation. A calculation was considered non-converging when it exceeded the nudging algorithm's iteration limit, as discussed above. This limit (and the number of intermediate images) were chosen pragmatically bearing in mind the relatively large number of computationally expensive calculations undertaken. Some non-converging calculations were repeated with altered intermediate locations and relaxed convergence conditions, but without additional result.

The 40 MEP segments can be combined to give 14 pathways from initial adsorption to full dissociation, with an Al adatom and three surface H. These fall into two groups of seven, differing only in their initial configurations, i.e. the A301 dimer-end configuration and the D301 end-bridge configuration. Earlier, it was concluded that these configurations were similar in character, differing only by a correlated rotation of the three H ligands, and this is consistent with the present NEB findings. Therefore, the discussion is limited to the pathways that start with the A301 configuration, focussing on those that play out on adjacent dimers in a single row.

Fig 6.2 compares the stabilities and activation energies of the seven pathways, which include the final incorporation MEPs. The data points are summarized at table 6.2. For comparison, a similar figure in the preceding chapter (fig 5.2 on page 82) shows D301 and other configurations not lying on any pathway. The figshare repository (Smith; Bowler 2017) includes a complete set of VASP configuration files.

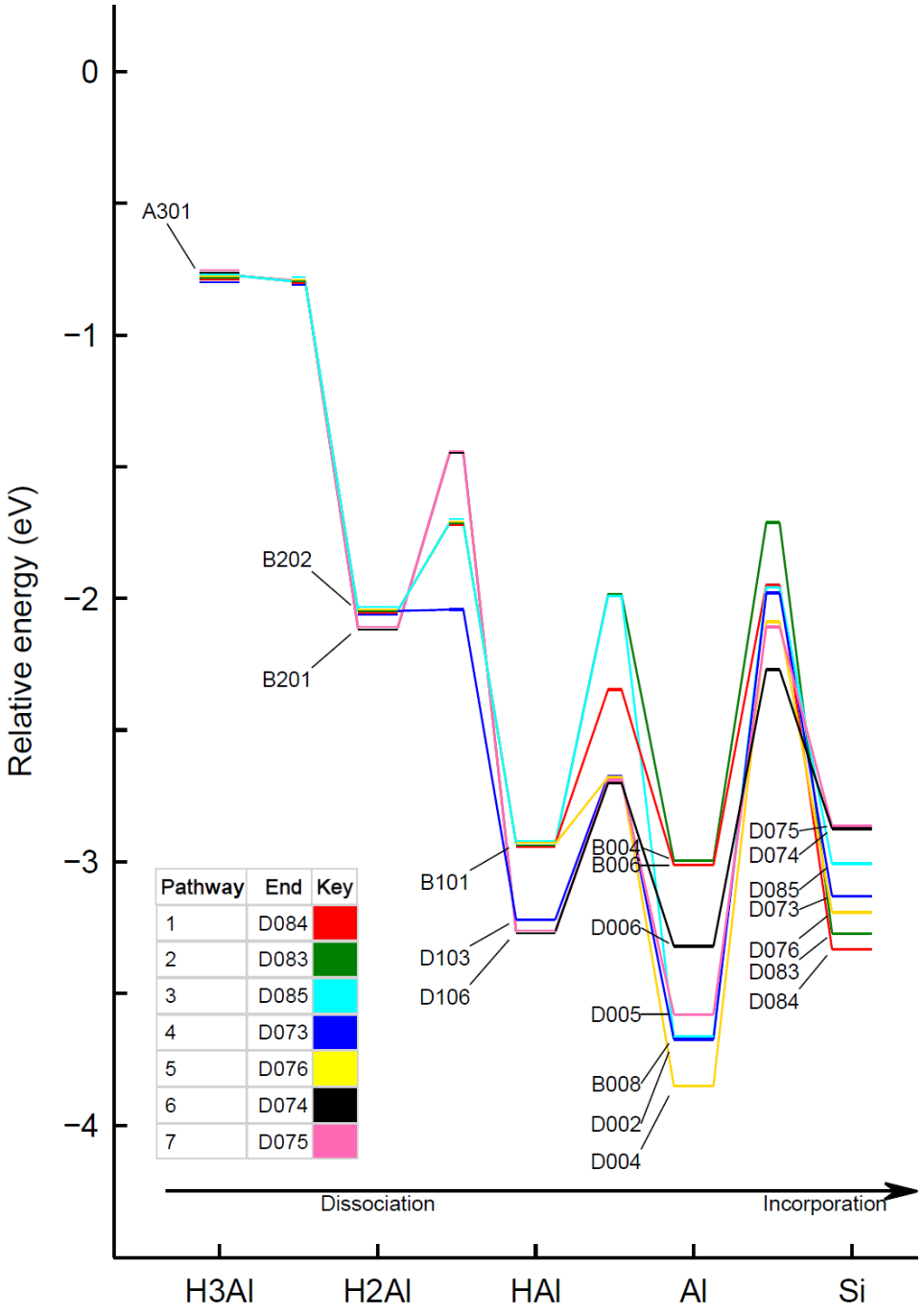


Figure 6.2 Seven continuous dissociation and incorporation pathways of alane, on Si(100). Vertical columns of bars show relative configuration energies (eV) at each stage. Energies are relative to the sum of bare surface and free alane energies. A pathway is indicated by a succession of bars of the same colour, e.g. the pathway 5 through configurations *A301-B202-B101-D004-D076* is coloured yellow. The intervening columns of shorter bars indicate the calculated relative transition state energies, i.e. the highest saddle-point energy along the MEP between adjacent configurations. The data points appear in table 2 below.

Pathway	*MEPs* $E_A/\Delta E_{cumulative}$ (eV)			
1	*A301-B202*	*B202-B101*	*B101-B006*	*B006-D084*
	0.00/-2.05	-0.01/-2.93	0.58/-3.01	1.06/-3.33
2	*A301-B202*	*B202-B101*	*B101-B004*	*B004-D083*
	0.00/-2.05	0.01/-2.93	0.95/-2.97	1.26/-3.27
3	*A301-B202*	*B202-B101*	*B101-B008*	*B008-D085*
	0.00/-2.05	0.01/-2.93	0.24/-3.67	1.40/-3.01
4	*A301-B202*	*B202-D103*	*D103-D002*	*D002-D073*
	0.00/-2.05	0.34/-2.29	0.37/-3.67	1.56/-3.13
5	*A301-B202*	*B202-B101*	*B101-D004*	*D004-D076*
	0.00/-2.05	0.01/-2.93	0.94/-3.85	1.87/-3.20
6	*A301-B201*	*B201-D106*	*D106-D006*	*D006-D074*
	0.00/-2.11	0.67/-3.33	0.59/-3.32	1.23/-2.87
7	*A301-B201*	*B201-D106*	*D106-D005*	*D005-D075*
	0.00/-2.11	0.67/-3.33	0.60/-3.78	1.63/-2.86

Table 6.2 Pathway number, per-MEP activation energy E_A and relative cumulative energy change $\Delta E_{cumulative}$ for the seven dissociation and incorporation pathways shown at fig 6.2. Activation energies are derived from a six-image VASP Cl-NEB calculation. A lowering of relative energy (i.e. a negative energy change) indicates a gain in stability, and vice versa.

An ideal pathway would pass through successive low-energy configurations, gain stability at each stage, and terminate in incorporation. As seen in chapter 5, the energetics favour end-bridge configurations as dissociation proceeds with some intermediate configurations of this kind acquiring greater stabilities than succeeding fully dissociated or incorporation scenarios which are our real interest. Any kinetic barrier should be surmountable at temperatures not impairing PALE passivation through H diffusion, say $450 - 500\,\mathrm{K}$. One can also expect incorporation to be hindered kinetically by the breaking of surface bonds. If a pathway fails to yield a progressive gain in stability, then (in the absence any other possibility) its reactions will tend to reverse, and the sequence terminate on the lowest energy configuration which is both stable and kinetically accessible at the prevailing temperature.

Table 6.2 shows the energy changes and activation energies for each dissociation and incorporation MEP, for each pathway. The remainder of our discussion is structured as follows: in sections 6.3.2 and 6.3.3 the first and second dissociations, where the relative absence of kinetic barriers characterizes all pathways, are discussed. In section 6.3.4 we describe two pathways (1 and 2 in table 6.2) that, while not involving the lowest-energy configurations, are nevertheless kinetically and thermodynamically feasible and appear to terminate in stable incorporation. In section 6.3.5 we discuss pathways 3, 4 and 5 where low-energy fully-dissociated configurations are sampled but do not lead to stable incorporation. Pathways 6 and 7 in table 6.2 involve configurations occupying two surface dimer rows and are of lesser interest and not discussed, but the calculation results are available in the figshare repository. Finally, in section 6.3.6 we describe an incorporation scenario involving surface migration of the ejected Si adatom.

6.3.2 First dissociation: $AlH_3 \rightarrow AlH_2$+H

There are two MEPs involved in the first dissociation, resolving to the dimer-bridge configurations B201 ($-2.11\,\mathrm{eV}$, fig 6.3(a)) and B202 ($-2.05\,\mathrm{eV}$, fig 6.3(b)) respectively, differing only in the ultimate location of the dissociated H ligand. No MEP was found to the

other low-energy configurations B206 (-2.08 eV) and D205 (2.32 eV), which are assumed to be sterically inaccessible. An MEP to configuration A201 (-2.08 eV) was discovered but gave rise to a relatively unstable configuration at the second dissociation (D109, -2.63 eV) from which no onward MEP was found.

The MEPs show similar stability gains (1.34 eV and 1.27 eV respectively) and are characterized by the absence of kinetic barriers, indicating that dissocation should occur immediately after the initial adsorption and proceed independently of the ambient temperature. The diagrams suggest a shallow PES basin in the vicinity of the Al-Si bond.

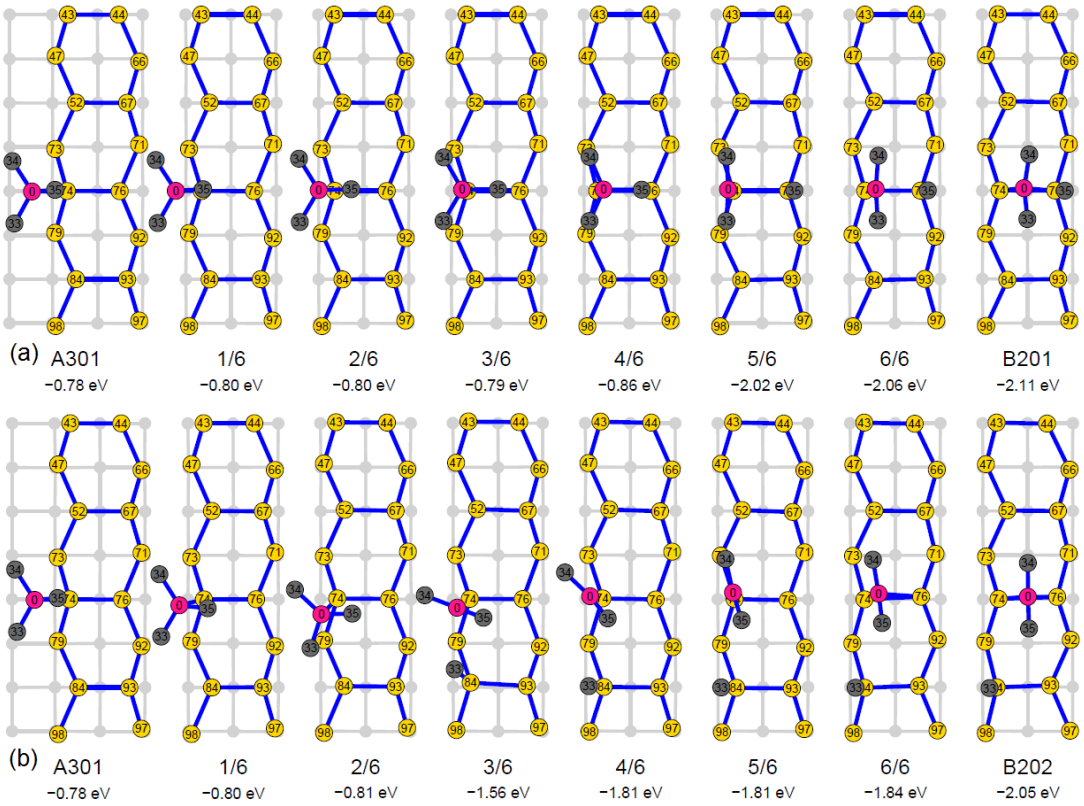


Figure 6.3 Representations of two MEPs of the first dissociation of alane adsorbed on Si(100), corresponding to segments A301-B201(a) and A301-B202(b). The reaction proceeds from left to right through intermediate points 1/6, 2/6 etc. In both cases the drop in relative energies indicates the absence of kinetic barriers. Images are derived from a VASP Cl-NEB calculation and each represents a single row of alternately buckled Si dimers, viewed from above, with Si, Al and H atoms coloured yellow, pink, and grey respectively. The atom numbers are zero-based indices into the originating VASP coordinate files.

6.3.3 Second dissociation: $AlH_2+H \rightarrow AlH+2H$

The second dissociation can proceed through three MEPs, terminating in the dimer-bridge configuration B101 (-2.93 eV, fig 6.4(a)) and the end-bridge configurations D103 and D106 (-3.07 eV, -3.33 eV figs 6.4(b), (c) respectively). However, whereas the transition to configuration B101 proceeds without significant kinetic barrier, those to the end-bridge

configurations encounter barriers of 0.34 eV and 0.67 eV respectively. The incomplete pathway to configuration D109 (-2.63 eV) mentioned above also encounters a barrier of 0.60 eV. The TST/Arrhenius relation (equation (1) above) indicates that even the lowest of these barriers would slow the second dissociation rate by a factor of 10^6 compared to the barrier-free rate. Given that the adsorbed fragment samples all directions equally one may assume reaction kinetics will dominate to ensure the onward pathways from B101 are sampled ahead of the others, possessing slightly increased stabilities.

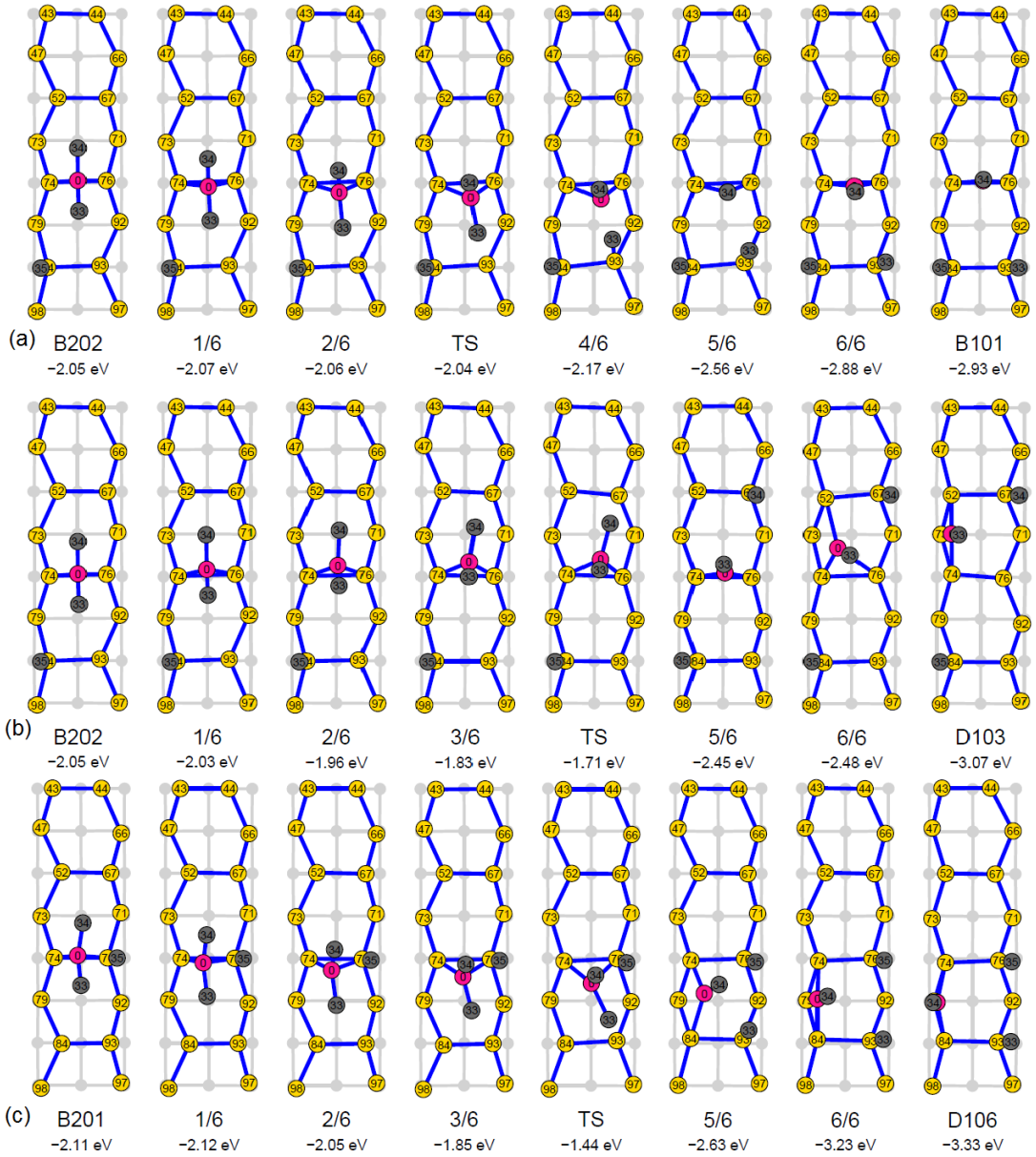


Figure 6.4 Representations of three MEPs of the second dissociation of alane adsorbed on Si(100), corresponding to segments B202-B101 (a), B202-D103 (b) and B201-D106 (c). MEP (a) shows an insignificant forward barrier of 0.02 eV at image 3, whereas MEPs (b) and (c) have barriers of 0.34 eV and 0.67 eV both at image 4. Image numbering and colouring convention as for fig 6.3. In MEP (a), Al atom 0 is obscured by H ligand 34 in the images labelled '5/6' and 'B101'.

6.3.4 Pathways 1 & 2: stabilization on incorporation

These pathways are characterized by incorporation MEPs that show increased stability (see figs 6.5(a), (b)). The lowering of energy (by 0.32 eV and 0.30 eV respectively) stabilizes the incorporation through a corresponding increase in the energy barrier to reversal. The forward energy barriers (1.06 eV and 1.26 eV) are surmountable within the constraints of the PALE process and so we might consider conditioning the PALE environment to favour these pathways while frustrating others involving configuration B101. This would be feasible using an automated STM, where complex operational sequences can occur under program control. For example, on pathway 2 the dissociation plays out on just two adjacent dimers, i.e. atoms (74, 76) and (84, 93) in the figures. If these were initially de-passivated an ambient temperature of ≈ 350 K would allow the exclusive evolution of the fully-dissociated configuration B004. Then, after de-passivating a third dimer and adjusting the temperature to ≈ 450 K, incorporation in configuration D083 should follow.

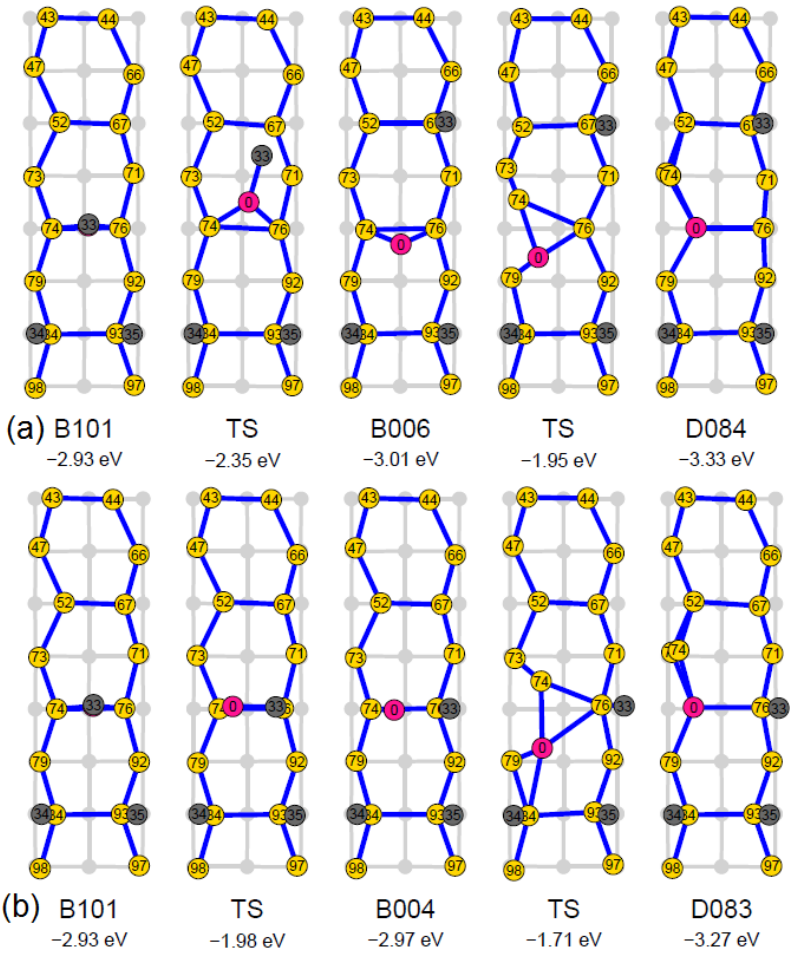


Figure 6.5 MEP representations of pathways 1 (a) & 2 (b) of the third dissociation of alane adsorbed on Si(100), followed by stability-enhancing incorporation steps. On pathway 1, incorporation (B006-D084) imposes an activation energy barrier of 1.06 eV and results in a stabilization of 0.32 eV. For pathway 2, the barrier to incorporation (B004-D083) is 1.26 eV and the stabilization is 0.30 eV. Even so, these pathways fail to yield stable incorporation (see text). Image derivation and colouring convention as for fig 6.3. In the images of configuration B101, Al atom 0 is obscured by H ligand 33.

Unfortunately, this scenario is unrealistic since the Al atom in configuration B004 will migrate to the lower energy end-bridge configuration D004 (seen on pathway 5, fig 6.6) as soon as the third dimer is cleared. This Al migration MEP B004-D004 (not shown) does not exhibit a significant activation energy and will be sampled before incorporation into D083, which presents a relatively large energy barrier. Incorporation on pathway 1 is frustrated similarly.

6.3.5 Pathways 3, 4 & 5: metastable incorporation

Fig 6.6 shows the third dissociation and incorporation stages of pathways characterized by destabilization on incorporation. Such destabilization ($> 0.50\ \mathrm{eV}$) imply that the end configurations (respectively D085, D073 and D076) are metastable and that the incorporation reactions would rapidly reverse, unless onward routes yielding sufficiently lowered energies were available. These reactions all require elevated temperatures and those on pathways 4 and 5 are not feasible under the temperature constraint of the PALE process.

At room temperatures, the kinetic barriers to incorporation are not surmountable and the fully dissociated configurations of pathways 3 (B008) and 4 (D002) should be visible by STM inspection. The effective (considering adsorption and dissociation only) activation energies are respectively $0.24\ \mathrm{eV}$ and $0.37\ \mathrm{eV}$. Both configurations have the same stability of $-3.37\ \mathrm{eV}$, but B008 is kinetically favoured and more likely to occur by a factor $\approx 10^2$. Although the fully dissociated configuration D004 on pathway 5 is the most stable encountered, its activation energy of $0.94\ \mathrm{eV}$ means it is even less likely to be observed at room temperature (by a factor $\approx 10^{12}$).

If the expected evolution to configuration B008 is found to occur in practice, one could then consider some selective de-passivation that would enable stable incorporation by allowing migration of the ejected Si adatom to a new location having suitably low energy. A possible scenario is described in the next section.

6.3.6 Post-incorporation Si migration

In the structural survey of chapter 5, a group of incorporation configurations (fig 6.7, D056-8) with stabilities comparable with the low-energy fully-dissociated configurations (fig 5(a) B008, (b) D002, (c) D004) emerged. However, all these incorporation configurations possess a 3-coordinated surface Al atom with an ejected Si adatom bridging a pair of Si-Si dimers. This organization will not evolve naturally during a low-temperature incorporation, which necessarily results in a 4-coordinated Al with the Si adatom bridging the Al-Si heterodimer and adjacent Si-Si dimer. The 4-coordinated configurations are seen to be less stable than those having 3-coordinated Al, which (in the PALE environment) could only be attained by exposing additional dimers so that the Si adatom could migrate away from the Al end of the heterodimer.

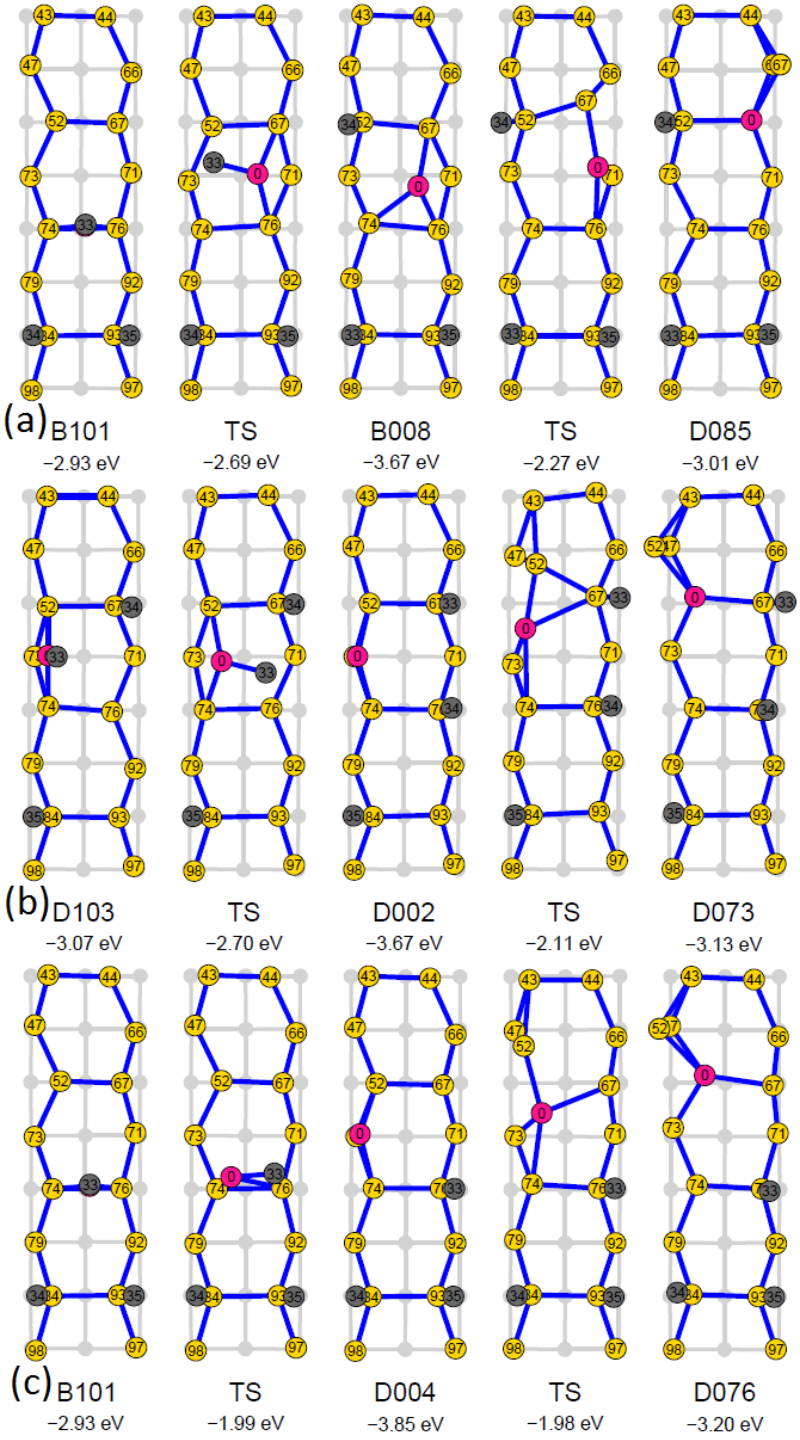


Figure 6.6 MEP representations of pathways 3 (a), 4 (b) & 5 (c) for the third dissociation of alane adsorbed on Si(100), followed by incorporation. A stability loss of at least 0.50 eV in the incorporation segments B008-D085, D002-D073 and D004-D076 indicate the end configurations are metastable. The activation energies for incorporation on pathways 4 and 5 (1.56 eV, 1.87 eV) are unachievable at PALE process temperatures. At room temperatures, fully-dissociated configuration B008 on pathway 3 is likely to predominate (see text). Image derivation and colouring convention as for fig 6.3.

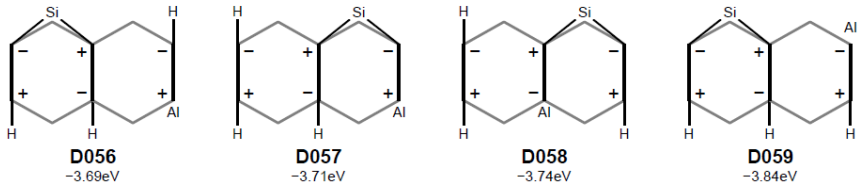


Figure 6.7 High stability incorporation configurations discovered in the structural survey of chapter 5. These are more stable than the fully dissociated configurations on pathways 3 and 4 while D059 has a stability comparable with that of D004 on pathway 5. This is due to the more favourable 3-coordination of the Al atom seen here.

Earlier, it was suggested that migration of the ejected Si adatom might lower the energies of our metastable incorporation configurations, thereby yielding an unconditionally stable configuration. Focussing on migration away from configuration D085 on pathway 3, configurations with relative energies significantly less than its predecessor B008 (-3.67 eV) were sought but without success. Some configurations did exhibit lowered energies, but none sufficiently low to stabilize the incorporation.

Equivalently, configuration B008 was destabilized by removing its adsorbed H atoms (H34, H35 and H36, see fig 6.6(a)). The new configuration, B021, was structurally optimized and showed Al-Si bond lengths increased by $\approx 1\%$ compared to B008, with the adsorbate coordination remaining unchanged. A relative energy of -2.96 eV was obtained by subtracting the bare surface energy. Additional configurations simulating a feasible onward migration and incorporation pathway, were also produced, bearing in mind the restricted size of the simulation cell.

Fig 6.8 shows the MEPs for these reactions. Incorporation (B021-D095) remains metastable but the stability loss is reduced to 0.14 eV from 0.66 eV obtained in the presence of adsorbate H. The first post-incorporation migration step (D095-D096) yields no stability gain but the second (D096-D097) indicates a gain of 0.25 eV, now sufficient to overcome the loss on incorporation. The energy barrier to incorporation is 1.17 eV, corresponding to a PALE process temperature of 400 K for reasonable activation within 60 seconds. The migration steps have lower barriers (respectively 0.65 eV and 0.80 eV) which would be surmountable at this temperature. The effective barrier to migration is lower than that presented by the reversal of incorporation, and so migration is favoured on both kinetic and thermodynamic grounds.

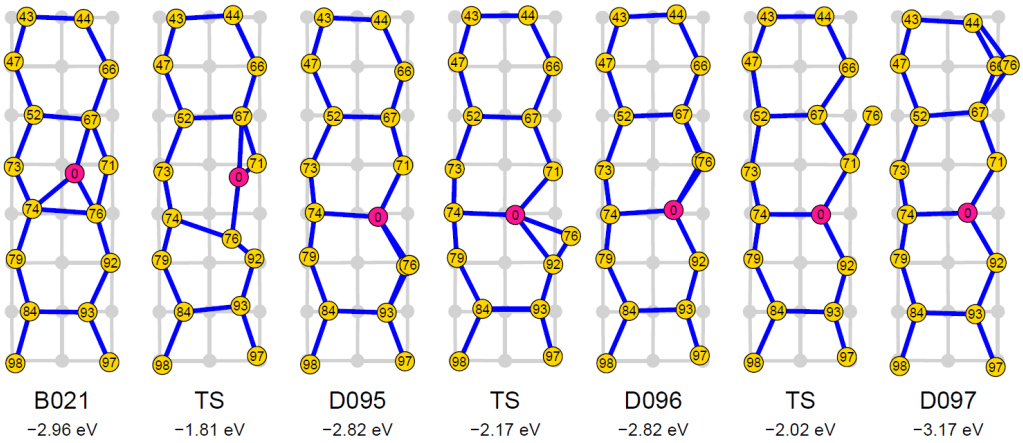


Figure 6.8 MEP representation of an Al incorporation and Si migration pathway from the fully-dissociated configuration B021 to a low-energy configuration D097 via metastable configurations D095 and D096. Configuration B021 was derived from B008 by removing its dissociated H. Configurations D095 and D096 are isomeric and a consequence of the restricted cell size. The migrating Si atom is 76. Ultimately, configuration D097 is kinetically and thermodynamically favoured (see text). Energies are relative to the bare surface energy. Image derivation and colouring convention as for fig 6.3.

6.4 Conclusion

Building on the results of an earlier structural survey, climbing-image NEB DFT calculations and transition state theory have been used to analyse the decomposition and incorporation of alane (AlH_3) on the H-passivated Si(100) surface. Decomposition, resulting in a bridged Al adatom and surface H, proceeds without significant kinetic barrier but, as with P, the energetics of the incorporated Al are close to those of the Al adatom. Furthermore, the constraints of the PALE scenario, where reactions are confined to an area spanning three or four adjacent dimers, prevent the natural evolution of a stable incorporation configuration. Rather, one would expect Al adatoms to become 'trapped' on the surface at room temperature and incorporate reversibly at elevated temperatures.

However, in an alternative scenario in which a bridged Al adatom is destabilized by the removal of dissociated H after decomposition, incorporation is eventually stabilized by surface migration of the ejected Si adatom. The migration pathway is determined by the extent of the surface exposure after decomposition and it is likely that other routes to stable incorporation could be found.

As already mentioned (page 92), the methodology of this and the preceding chapter follows that of Warschkow and others in their studies of PH_3 behaviour on the Si(100) surface. Their earlier DFT work was supported by STM observations showing decomposition pathways which could not be rationalized as simple temperature-dependent transitions between low energy intermediate configurations. Schofield; Curson et al., (2006) proposed a diffusion mechanism whereby the intermediate PH_2 fragment could move along a dimer row, or jump to an adjacent row prior to adopting a stable PH+2H configuration. This work did not cover

the transition to P+3H but the entire dissociation sequence was later dealt with exhaustively by (Warschkow; Curson et al., 2016). The latter paper concluded that dissociation to P+3H would require an activation energy of ~ 1 eV, consistent with experiment and comparable with the barriers to Al+3H found here. Intra-row diffusion of the P adatom during incorporation was reported by (Bennett; Warschkow, 2009); this paper also covered surface diffusion of ejected Si atoms and predicted incorporation at 530 C, lower than the experimental value of 650 C, but still too high for PALE compatibility. This is especially relevant to our findings for AlH_3, even in the absence of experimental data, as we see that some surface diffusion of the adatom is needed to achieve incorporation. But although PH_3 decomposition on the Si(100) surface has been studied extensively, the goal of incorporation at room temperature remains elusive.

However, the ability to incorporate acceptor dopants as well as donors in $\mathrm{Si}(100)$ with atomic precision will significantly advance the capabilities of patterned ALE. It opens the possibility of p-n junctions fabricated with atomic precision, as well as local control of the electrostatic potential using both positive and negative dopant ions. We keenly anticipate experimental measurements of these structures as a first realisation of this.

Chapter 7

Al doped Si nanostructures

7.1 Introduction

In chapter 1 it was shown how doping can determine the electronic properties of bulk Si. Later chapters described the fabrication of Si nanostructures and the atomically precise incorporation of Al dopants using PALE. These structures would be combined as building blocks with donor-doped equivalents to form basic circuit elements such as diodes, transistors and inverters, and eventual integration into larger-scale devices (Cui, Lieber 2001). These devices would be junction-based and created by p and n co-doping of a single structure (Colinge et al., 2010; Ng, Tong 2011). We now examine some primitive Al doped structures which might be built with PALE to see how their properties depend on shape, size, surface treatment and dopant location.

Nanowires (NWs) are the simplest one-dimensional nanostructure. They are of particular interest since they be can fabricated as both interconnects and as building blocks from from which computational circuit elements could be assembled (Huang et al., 2001). Existing literature describes fabrication and doping of Si NWs using non-lithographic means such as the vapour-liquid-solid (VLS) mechanism (Peng; Lee 2011, Wagner; Ellis 1964) but only relatively small examples (diameter ≤ 10 nm) can be investigated by atomistic methods. Even with a very large computer (as here) VASP calculation times can become excessive when structure sizes exceed 1000-1500 atoms. This problem does not arise with $\mathrm{O}(N)$ codes (e.g. CONQUEST) that calculate interactions on a per-atom basis within a local part of space rather than for the totality of atoms in the system. Computational effort then scales with the local volume and ultimate performance depends on efficient implementation of the local basis set (Bowler, Miyazaki 2010).

The VLS process produces NWs whose diameter follows that of catalytic gold nanoclusters. Growth is conditioned by the surface energetics of the Au/Si LS interface, and for smaller NWs occurs in the $[110]$ direction with hexagonal cross section (Wu al., 2004). DFT models can readily employ an H terminated hexagonal slab of variable depth as supercell (Kumarasinghe, Bowler 2020; Ng et al., 2011). On the PALE Si(100) surface atoms are situated on a 3.84 Å square grid and both octagonal and square-section models are geometrically convenient, although other configurations are possible. We select the square section as we wish to investigate NWs in lateral contact with a Si substrate. Further, we model small cuboid (cell, pillar) structures, accessible to VASP. The required calculations were described in chapter 3 and should reveal the electronic behaviour near the Fermi level (which characterizes a functional dopant) and how it is affected by surfaces, dopant locations and concentration. We model NWs of increasing size (the largest equivalent to ~ 5 nm

diameter) and establish qualitative trends in behaviour. Then we choose smaller, but representative dimensions for the 3D structures so that calculation times remain manageable.

Aluminium is a p-type dopant that can enter the silicon lattice substitutionally as shown in the ELF plot fig 7.1 (a). This suggests the Al nucleus abstracts electronic density equivalent to one electron from the lattice to form four covalent bonds (also see page 84). The Al-Si bonds each have length ~ 2.40 Å compared with ~ 2.35 Å seen for bulk Si-Si bonds. Once incorporated, it is essential that the dopant remain static and not diffuse within the lattice. Fig 7.1(b) shows that two adjacent Al have defective lattice bonding, compromising any p-type character of the neighbourhood. The calculations of this chapter assume that dopant atoms are isolated, immobile and distant from each other.

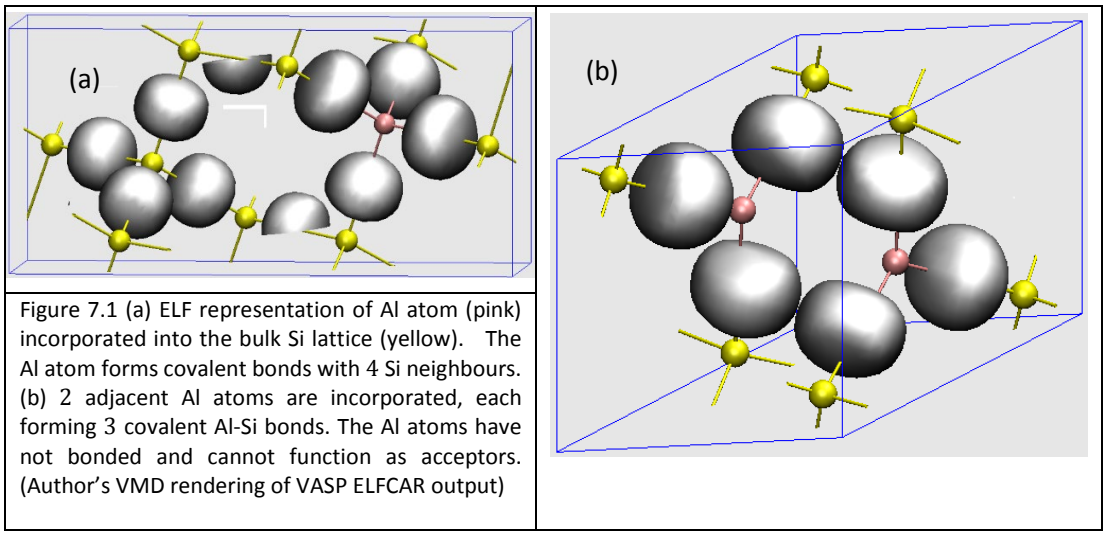


Figure 7.1 (a) ELF representation of Al atom (pink) incorporated into the bulk Si lattice (yellow). The Al atom forms covalent bonds with 4 Si neighbours. (b) 2 adjacent Al atoms are incorporated, each forming 3 covalent Al-Si bonds. The Al atoms have not bonded and cannot function as acceptors. (Author's VMD rendering of VASP ELFCAR output)

7.2 Methods

All the calculations of this chapter are based on DFT as implemented in the VASP package. The GGA PBE functional and PAW pseudopotentials are discussed at pages 32 and 39 respectively, and the techniques for state density and band structure calculations are described at pages 60 and 63 respectively. DFT total energies were converged to within 10^{-5} eV with a 400 eV energy cut-off. For the NW structural relaxations used a Monkhorst-Pack mesh with a density of $(1\text{x}1\text{x}2)$ where the NW is taken to lie along the z direction. The NW was constrained axially but free to expand radially. The NWs were placed in supercells with 11 Å vacuum spacing and relaxed with an atomic force criterion of 0.05 eV/Å. The ridge, cell and pillar nanostructures were handled similarly, i.e. allowed to expand along surface normals but otherwise constrained.

The graphical images were created in MATLAB from VASP text output files, using programs written by the author and briefly described in Appendix B.

7.3 Results and discussion

7.3.1 Pure Si NWs

Reported studies usually describe catalysed vapor-liquid-solid (VLS) growth in the [110] and [111] directions, creating NWs with hexagonal cross-section (Ma et al., 2003; Wu et al., 2004). STM-based PALE naturally creates structures having square or rectangular cross-section where growth occurs in the [001] direction. Fig 7.2 shows schematic diagrams of the smallest (a) and largest (c) square section NWs considered, in terms of side length, which varies from 1.92 nm to 6.56 nm. The supercells containing these structures allow for an

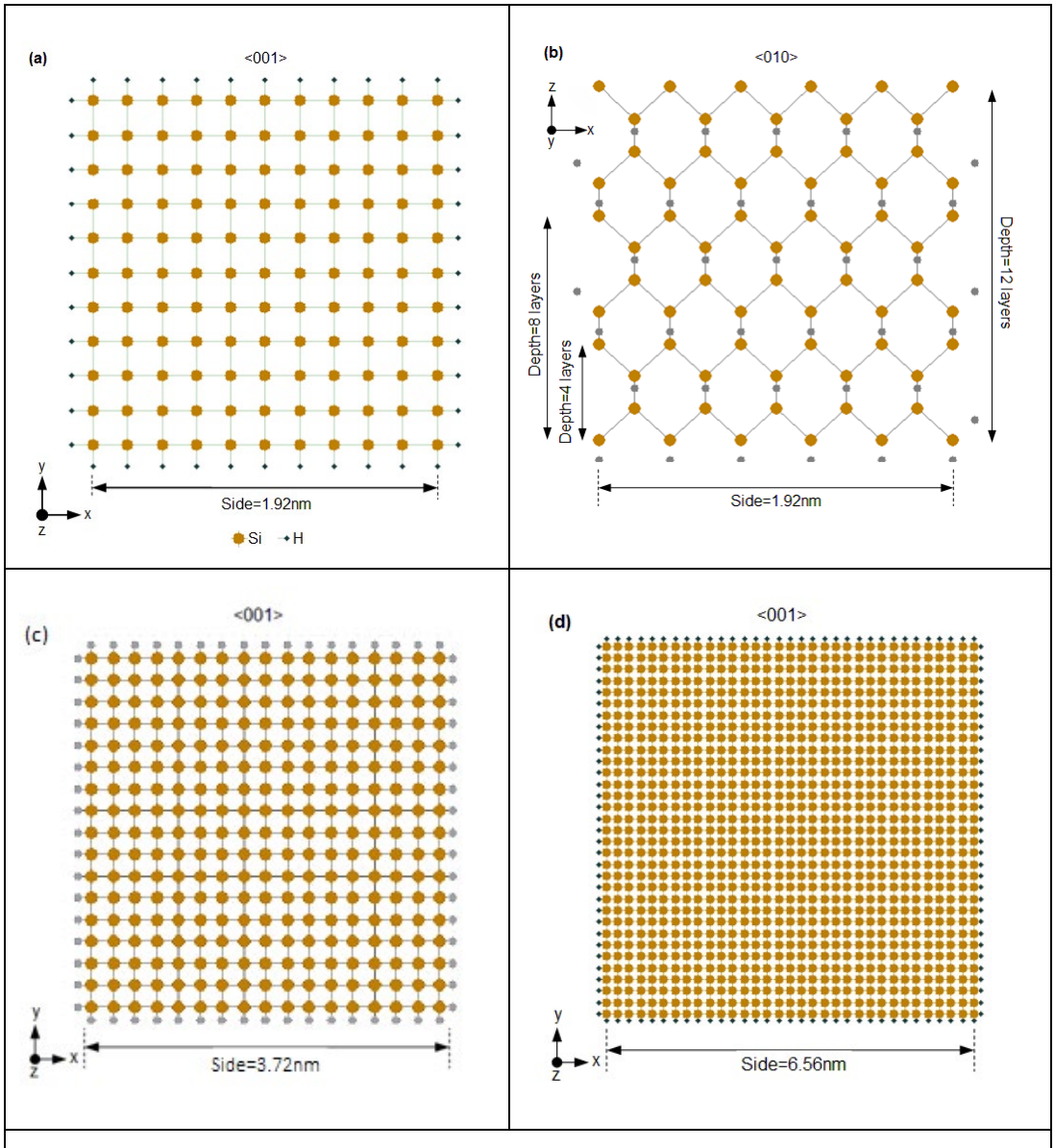


Figure 7.2 (a), (c) and (d): Cross-sectional view of [001] oriented H passivated undoped square Si NWs of side length 1.92 nm, 3.72 nm and 6.56 nm, respectively. (b) shows the side elevation [010] of the smallest NW and the different simulation cell lengths employed – 4, 8 and 12 layers. These are representations of freestanding NWs, and do not show the surrounding vacuum region of the supercell.

11 nm vacuum gap in both the x and y directions and extend indefinitely in the z (axial) direction. The side view (b) shows the layers included in the simulation cell (4, 8 or 12 layers) along with H passivation of the [010] surface.

We start by considering the effect of side length on electronic structure, using a repeat length of 4 layers. Fig 7.3 (a) shows the band gap increasing as the side length is reduced (a quantum confinement effect) and declining towards the calculated bulk value as the side length is increased. The largest NW considered here has a side length of 6.56 nm or 36 Si layers (fig 7.2 (d)). This is large enough to approximate bulk behaviour in the central region of the NW. Fig 7.3 (a) also shows that experimental band gaps (in approximately circular cross-section NWs) converge to the room temperature bulk value (~1.1 eV) with increasing diameter. The calculated results for these square section NWs show a similar convergence

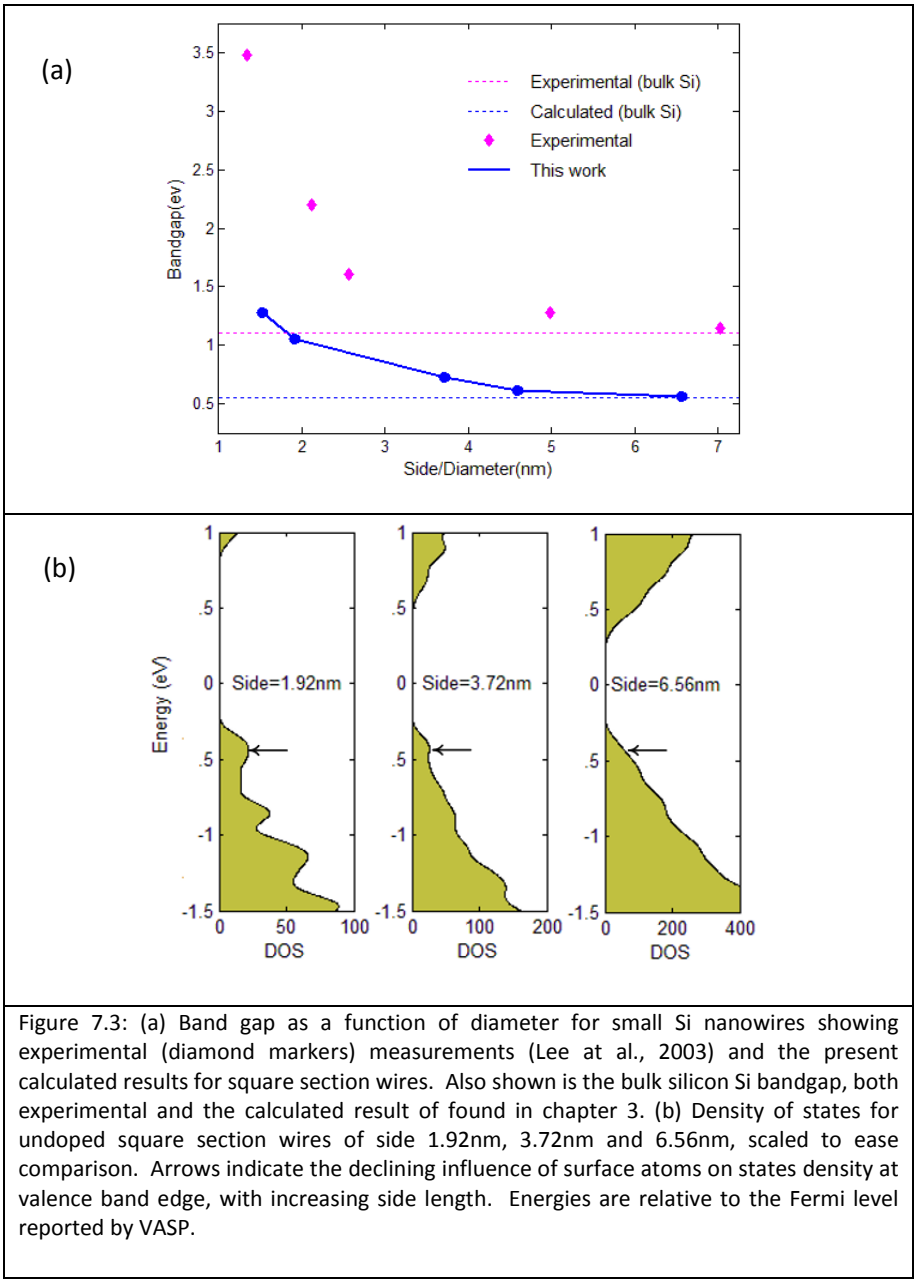


Figure 7.3: (a) Band gap as a function of diameter for small Si nanowires showing experimental (diamond markers) measurements (Lee at al., 2003) and the present calculated results for square section wires. Also shown is the bulk silicon Si bandgap, both experimental and the calculated result of found in chapter 3. (b) Density of states for undoped square section wires of side 1.92nm, 3.72nm and 6.56nm, scaled to ease comparison. Arrows indicate the declining influence of surface atoms on states density at valence band edge, with increasing side length. Energies are relative to the Fermi level reported by VASP.

to the calculated bulk band gap which differs from the experimental value, a well-known DFT discrepancy arising from the use of a GGA functional. This could be avoided (at increased processing cost) with a hybrid functional, but it is usually ignored in theoretical work.

Fig 7.3 (b) shows DOS for the full NW strongly affected by side length, and the smoothing effect caused by the presence of more atoms in the simulation. The appearance of DOS and PDOS plots also depends on the level of artificial smoothing applied in the calculation, as discussed at page 60.

All the state density plots of this chapter employ the Gaussian smoothing technique implemented in VASP, with the smoothing parameter value σ set to 0.08 eV. This value was chosen empirically to avoid excess detail and ease qualitative comparisons. The action of a dopant hinges on its effect on electronic states near the edges of the conduction and valence bands, which dominate charge transport properties. In a confined nanostructure it is important to know which atoms contribute most to these states. Fig 7.4 shows partial densities of states (PDOS) projected onto individual atoms situated at varying distances from the NW centre. At the valence band edge (VBE), locations nearer the surface contribute least irrespective of NW size. Conversely, centrally located atoms always make a greater contribution. In these models, relaxation leaves Si-Si bond lengths near the surface unchanged from the bulk and the effects on the PDOS must be due to passivation, with the Si-H bonds contributing relatively fewer states. Similar radial PDOS behaviour has been noted in other NW studies (Ng, Tong, 2012). However, in the smaller NWs and at lower energies (e.g. where indicated by an arrow on the energy axes) the PDOS oscillates and is greater at some asymmetric dopant positions. This effect is not seen in H-passivated hexagonal Si NWs grown in the $[110]$ direction (Kumarasinghe, Bowler 2020) and can be attributed to the sharper corners present here. In the presence of an acceptor dopant the Fermi level moves closer to these states and performance might be improved, as proper p-type behaviour requires a plentiful supply of states in this region. In the largest NW (c) surface effects are diluted and the PDOS resembles that seen in hexagonal NWs (ibid.).

7.3.2 Al doped Si NWs

We now examine the effect of doping in the two smaller NWs shown in fig 7.2, with the aim of showing how the Al dopant and its radial positioning affects the electronic band structure. As before, dopant locations are chosen to reveal the influence (if any) of the NW surfaces on the band structure. The large NW of fig 7.2 (side length 6.56 nm (d)) has been excluded, allowing examination of larger axial dimensions within manageable calculation times. Fig 7.3 (a) shows that the band gap of the undoped NW tends towards the bulk value as the side length is increases from 4 to 5 nm, and we assume that study of larger NWs would not reveal any new qualitative behaviour.

In these calculations the open shell of the trivalent Al atom gives rise to a non-zero relative spin polarization:

$$\zeta = \frac{\rho_\uparrow - \rho_\downarrow}{\rho_\uparrow + \rho_\downarrow} \qquad (1)$$

where $\rho_\uparrow$, $\rho_\downarrow$ are the densities of 'spin-up' and 'spin-down' electrons, respectively. With just a single dopant atom spin the effect of ζ might be expected to be small and spread over the entire lattice, but this can be confirmed by performing spin-sensitive calculations as described in chapter 2. Spin sensitivity is exposed in the exchange correlation functional. Exchange energy is independent of spin polarization but in the LDA and GGA functionals correlation energy is represented an elaborate function of ζ fitted to quantum Monte Carlo data from the homogeneous electronic gas (Perdew at al, 1996).

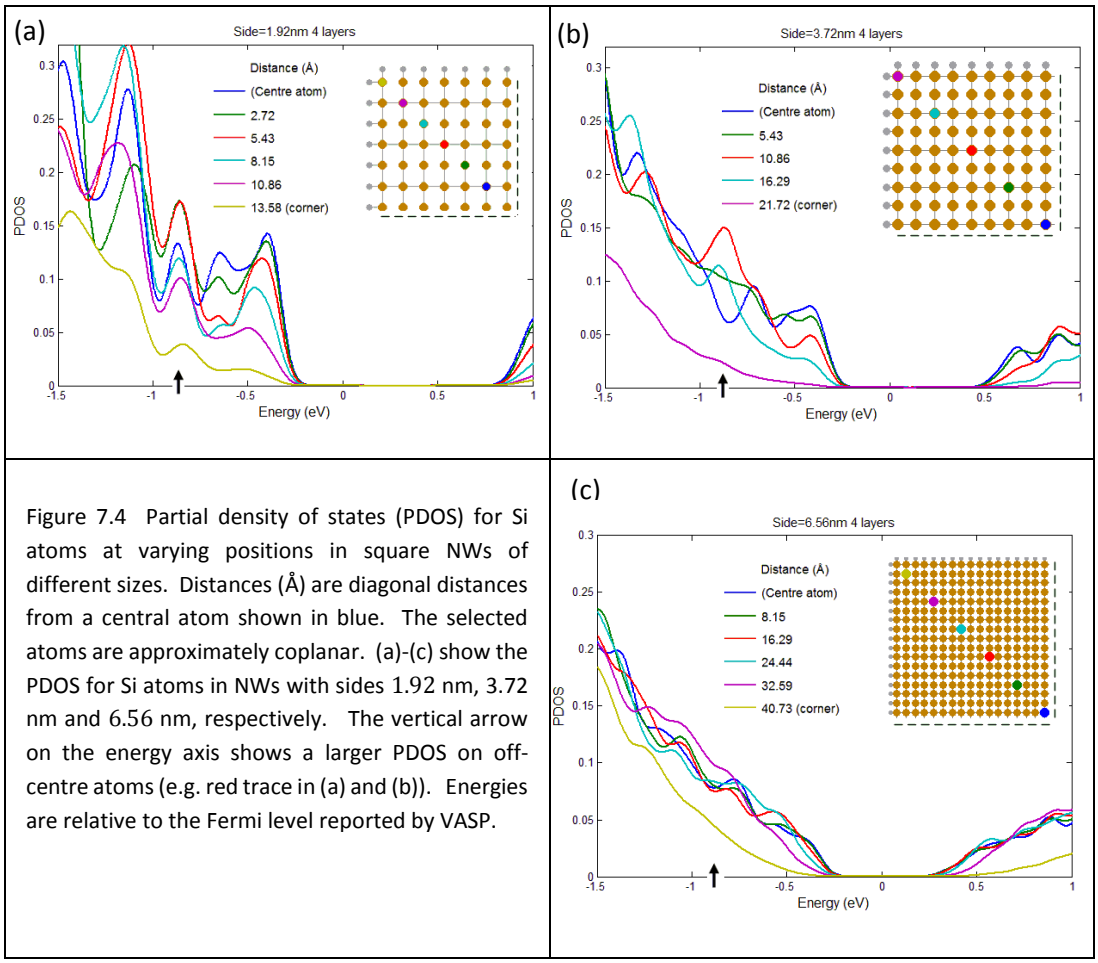


Figure 7.4 Partial density of states (PDOS) for Si atoms at varying positions in square NWs of different sizes. Distances (Å) are diagonal distances from a central atom shown in blue. The selected atoms are approximately coplanar. (a)-(c) show the PDOS for Si atoms in NWs with sides 1.92 nm, 3.72 nm and 6.56 nm, respectively. The vertical arrow on the energy axis shows a larger PDOS on off-centre atoms (e.g. red trace in (a) and (b)). Energies are relative to the Fermi level reported by VASP.

The spin-sensitive calculation produces two energy eigenvalues for each band. The lower of these is conventionally taken to correspond to a 'spin-up' orientation and the higher to 'spin-down'. This terminology is somewhat misleading because the spin direction is arbitrary, although it is assumed to be same for all electrons. DFT allocates a spin-up and spin-down electron to each band in ascending energy order, representing the electronic ground state configuration at $T = 0$. With a single dopant atom all bands are doubly occupied except the highest, which defines the VBE and has its spin-up state occupied and its spin-down state unoccupied (alternatively, occupied by a hole state). The Fermi level can be taken to lie at the VBE, where the electronic density should ideally experience a step change to zero. DFT packages generally smear the step edge to ease the numerical integration of the density wavefunctions in k space, causing non-integer occupancies to arise (also see chapter 3 pages

60-63). The band structure diagrams of this chapter have been produced with minimal Gaussian smoothing ($\sigma = 0.01$ eV) and non-integer occupancies occur at only a few k points per band with values close to 0 or 1. These artificial values must be distinguished from *partial occupancy* where a band may be occupied over a range of adjacent k points but unoccupied over a neighbouring range. This can be seen in some of the following band structure diagrams, where edge occupancies are indicated by circular markers. At each k point a filled marker indicates the highest energy occupied band (the HOMO of Molecular Orbital theory) and an unfilled the lowest unoccupied (LUMO) and partial occupancy is indicated by connected markers traversing the band gap. Prior to making this determination, any smearing occupancy is eliminated by rounding.

The effect of increasing the axial length of a doped NW can now be tested, using the NW of fig 7.2 (a) of side length 1.92 nm and depths of 0.55 nm (4 layers), 1.1 nm (8 layers) and 1.6 nm (12 layers) as shown in fig 7.2 (b). Each structure contains a single Al dopant atom, so these depths are also the dopant spacings and yield concentrations (Al/Si ratio) of $1/121$, $1/241$ and $1/362$ respectively. Fig 7.5 compares the band structures of these NWs in the axial direction which is the only symmetric path in the Brillouin zone. The diagrams show band curvature varying with the number of supercell layers. This is a dispersive effect caused by the interaction of adjacent periodic images (i.e. in the limit of dopants interacting on a solely nearest neighbour basis, the band would become parabolic). In the 4-layer model (a) the dopant spacing is only 0.55 nm, leading to some electronic delocalization and partial occupancy appearing at the VBE. The 8-layer (c) and 12-layer (d) models show a flattening of the bands as the dopant spacing increases, indicating negligible interaction at a spacing of 1.1 nm and beyond and an absence of partial occupancy. The spin-up 4-layer band structure (b) contains no unoccupied band at the VBE and resembles that of the same sized undoped NW (e). This indicates spin-up states are not involved in the acceptor doping mechanism and spin-up band structures have not been calculated for the other nanostructures considered later in this chapter.

The HOMO-LUMO energy interval contributes to the width of the depletion region (page 15) formed when p and n-type regions come into contact and is typically ~ 0.7 V in a conventionally doped silicon diode or transistor. An ideal p-type nanostructure would show a uniform transition over an interval of a few tenths of a volt, but confinement and concentration effects may result in a non-uniform or non-existing transition as seen in the 4-layer model. When the axial length of the NW is extended to 8 and 12 layers, a sharp transition (< 0.1 eV) is sustained throughout the NW.

Figs 7.6 (a), (b) compare the total DOS of the NWs of fig 7.5, for both spin orientations. An increased state density near the VBE and a shifted Fermi level compared with the undoped is visible. These effects are greater in the 8 and 12-layer models but the spin-up configuration does not introduce an extra hole band level as noted above.

Figs 7.6 (c), (d) show the DOS projected onto the dopant atom, again for both spin orientations. These show that the hole state is localized on the impurity, as would be expected at $T = 0$. There is little difference in the PDOS for the 8 and 12-layer cells, but both are markedly greater than that of 4-layer cell. This confirms at least 8 layers are needed

if interaction between the dopant images is to be avoided. Accordingly, subsequent NW calculations employ a repeat length of 2 (8 layers), allowing the modelling of larger wires with negligible dopant interaction.

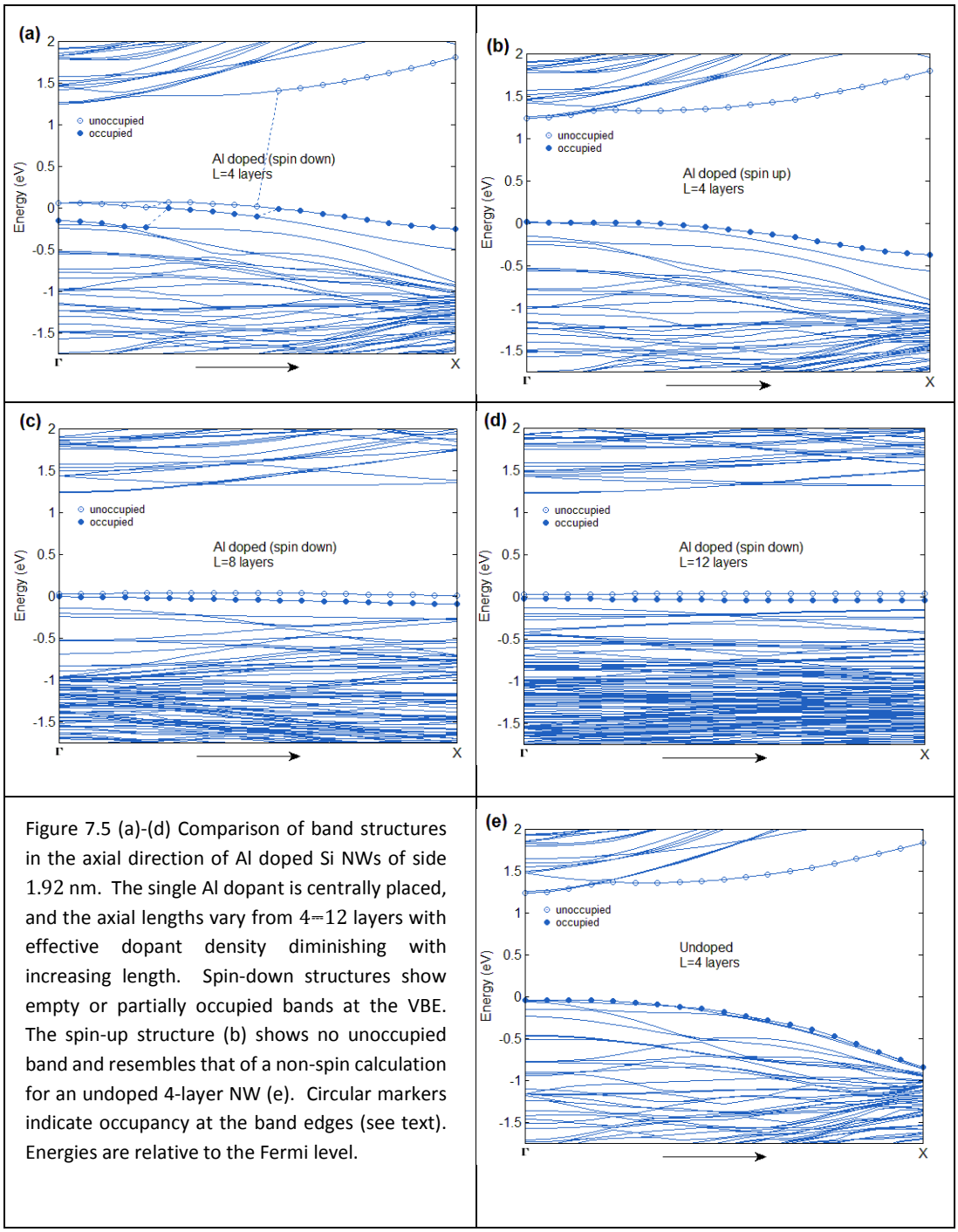


Figure 7.5 (a)-(d) Comparison of band structures in the axial direction of Al doped Si NWs of side 1.92 nm. The single Al dopant is centrally placed, and the axial lengths vary from 4–12 layers with effective dopant density diminishing with increasing length. Spin-down structures show empty or partially occupied bands at the VBE. The spin-up structure (b) shows no unoccupied band and resembles that of a non-spin calculation for an undoped 4-layer NW (e). Circular markers indicate occupancy at the band edges (see text). Energies are relative to the Fermi level.

To investigate the effect of dopant positioning on the DOS, PDOS and band structure, square section H-passivated NWs of sides 1.92 nm and 3.72 nm were modelled. The results for the smaller NW in an 8-layer supercell with 241 Si atoms are shown in fig 7.7. An Al atom substitutes for an Si atom near the centre (position 1, (b)), in a corner (position 3, (d)) and approximately mid-way between (position 2, (c)). To aid interpretation, band structure,

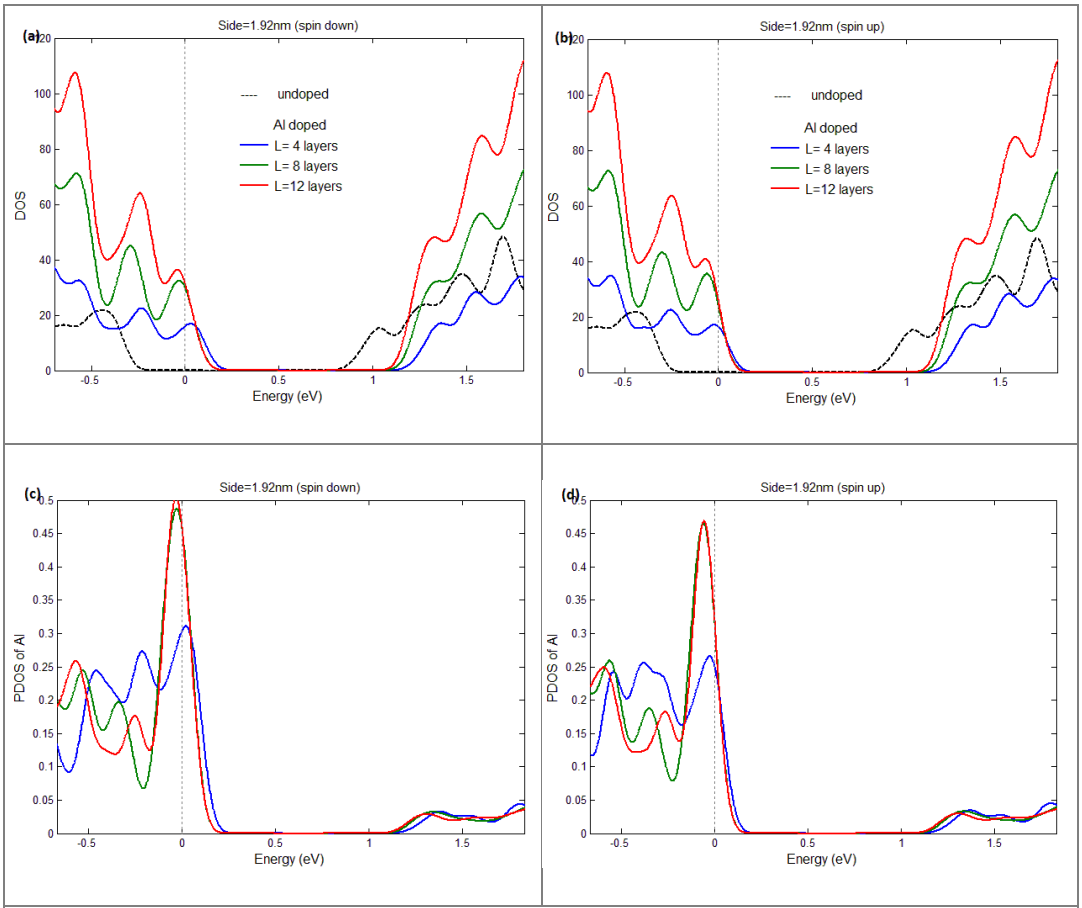


Figure 7.6 (a), (b) comparison of DOS of doped (solid lines) NWs of side 1.92 nm. Cell lengths of 4,8 and 12 layers are compared for wires containing a single, centrally placed Al dopant for the spin-down and spin-up orientations of the Al electron, respectively. The DOS for an undoped (dashed line) 4-layer NW is shown for comparison. (c), (d) comparison of the PDOS experienced by the Al atom in the same three doped NWs, for both spin orientations. In each case, energies are shown relative to the Fermi level, indicated by vertical dotted lines.

DOS and PDOS plots are shown with a common energy axis. Comparing the PDOS and band structure plots it is apparent that the LUMO band corresponds to the state projection on the Al atom, irrespective of its position. The LUMO is flat, leading to a localized peak in the DOS which can also be seen. The state density projection onto the dopant changes with its location. At the centre the Al hole state and a relatively small number of surrounding Si valence states predominate, i.e. the surface induced PDOS peaks around -0.4 eV are relatively small. As the dopant moves towards the surface these increase in size, reflecting the change in dielectric from NW to vacuum. The declining permittivity increases the attractive Coulombic force exerted by the nuclei, lowering electronic energy and accentuating the PDOS below the Fermi level.

Fig 7.8 shows corresponding results when the NW side length is increased to 3.72 nm. This supercell has 8 layers and contains 577 Si atoms. The lowered confinement immediately reduces the band gap width, as was seen in fig 7.3 (a), but this effect is not dopant related.

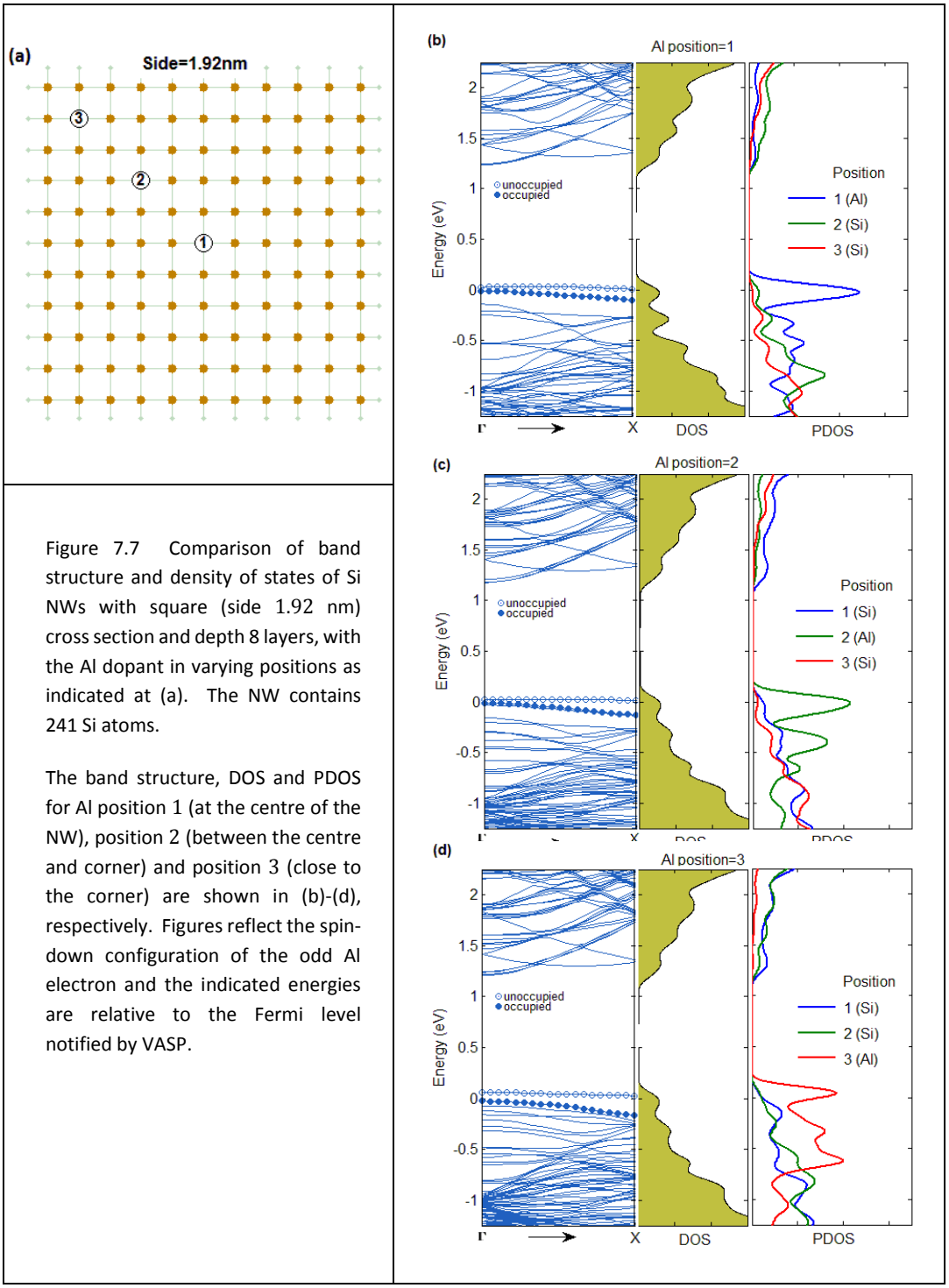
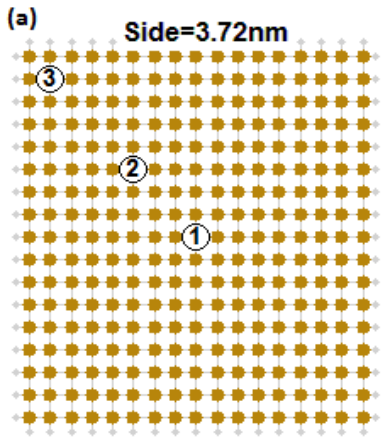


Figure 7.7 Comparison of band structure and density of states of Si NWs with square (side 1.92 nm) cross section and depth 8 layers, with the Al dopant in varying positions as indicated at (a). The NW contains 241 Si atoms.

The band structure, DOS and PDOS for Al position 1 (at the centre of the NW), position 2 (between the centre and corner) and position 3 (close to the corner) are shown in (b)-(d), respectively. Figures reflect the spin-down configuration of the odd Al electron and the indicated energies are relative to the Fermi level notified by VASP.

Atoms in the central region are relatively isolated from surface effects, compared to the smaller NW. Although the PDOS plots show dopant-induced states at the Fermi level the band structure plots show holes occupying distinct, closely spaced bands near the VBE and an eventual reversion to intrinsic behaviour, dependent on dopant location. These effects were not seen in the smaller NW of fig 7.7, suggesting that the perturbing effect of the dopant was sufficient to lift any band degeneracy in that case. While the breakdown of p-type behaviour might be due to reduced dopant concentration its onset



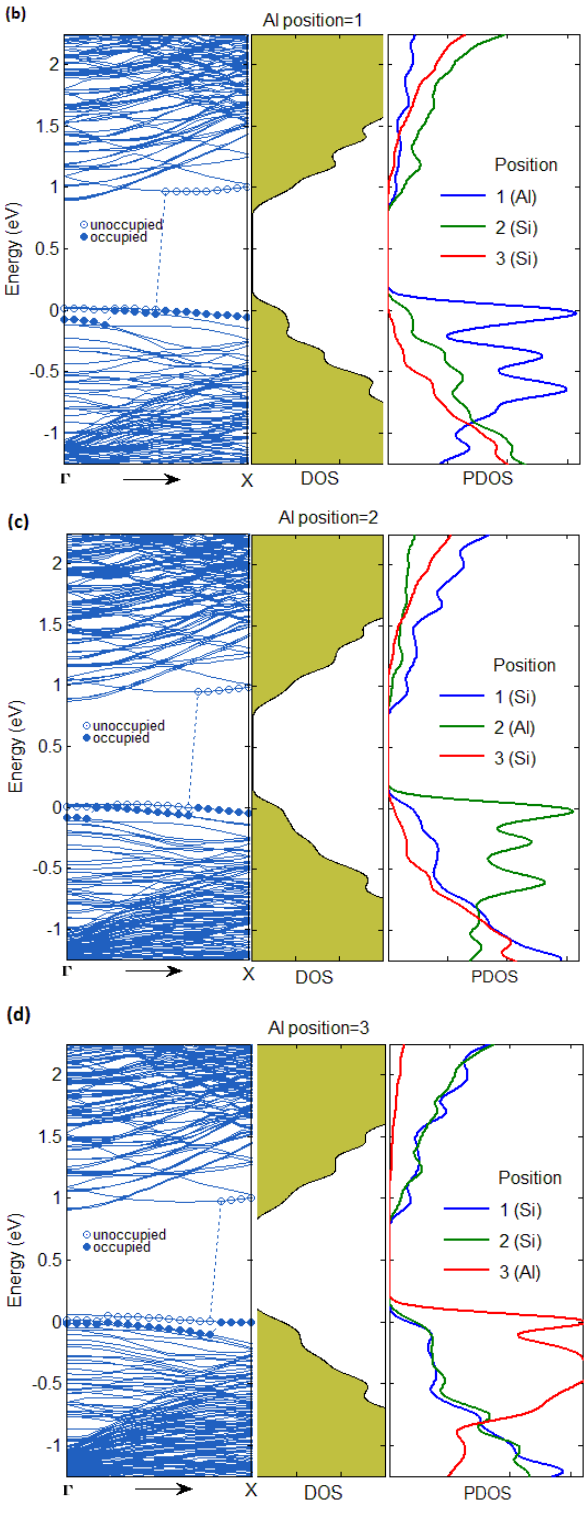


Figure 7.8 Comparison of band structure and density of states of Al doped Si NWs with square (side 3.72 nm) cross section and depth 8 layers, with the single Al dopant in varying positions as indicated in (a). The NW contains 577 Si atoms.

The band structure, DOS and PDOS are shown for Al at position 1 (the centre of the NW), position 2 (between the centre and corner) and position 3 (close to the corner) in (b)-(d), respectively. Figures reflect the spin-down configuration of the odd Al electron and the indicated energies are relative to the Fermi level notified by VASP.

also depends on dopant location and so surface confinement must play a role. It seems that surface proximity accentuates the state density at lower energies (e.g. fig 7.9 (d), red plot) and contributes to consistent p-type behaviour throughout the IBZ. To clarify this point, we

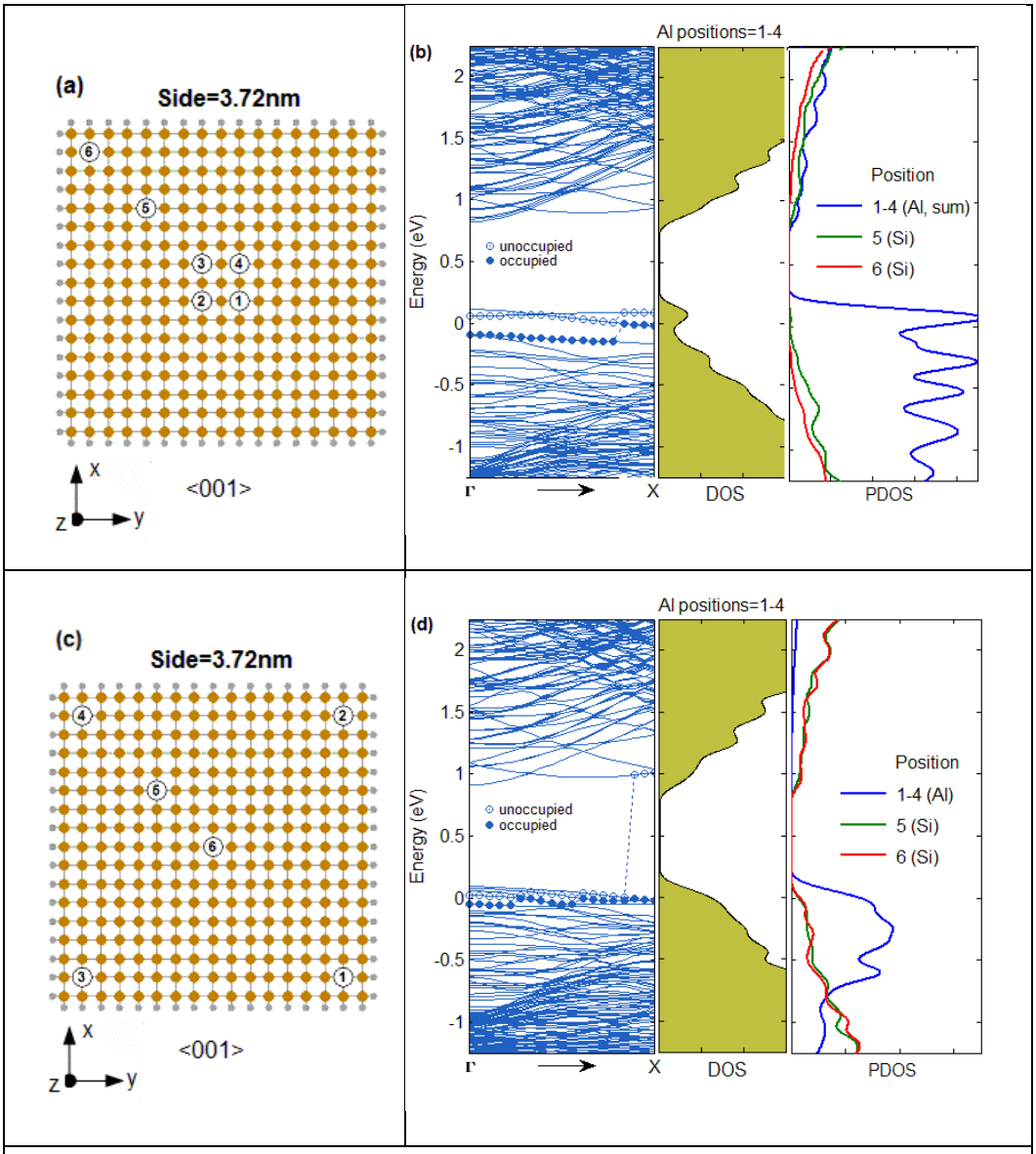


Figure 7.9 The NW of fig 7.8, but containing 4 Al and 574 Si atoms. In configuration (a) 4 Al dopants are centrally placed, while at (c) they occupy corner locations. In (b) the blue PDOS plot is the sum over the Al atoms at positions $1-4$. In (d) it is the single Al in position 1 but the projections at positions $2-4$ are similar.

make two further calculations for this NW at an increased dopant density. In fig 7.9 (a) there are 4 centrally placed Al atoms while in fig 7.9 (c) a dopant atom is located at each corner. These configurations are similar to those employed by (Ng and Tong, 2012) in a study of B and P co-doped NWs. The increased concentration $(1/144)$ configurations may be compared with dopant positions 1 and 3 in fig 7.8 (a) where the concentration is $(1/577)$. After structural optimization, the Si-Al bond lengths of the closely spaced dopants of fig 7.9 (a) were all found to lie in the range 2.39 to 2.40 Å and so proper covalent bonding can be assumed. The band structure shown at fig 7.9 (b) shows hole

states at the VBE throughout the IBZ, but again they do not appear uniformly as a single band. This indicates that the breakdown of the band structure seen in the larger NW of fig 7.8 has been suppressed (but not eliminated) by increasing the dopant concentration, and full coverage of the IBZ has been restored. The PDOS plot shown in fig 7.9 (b) is the summation of the projection onto the 4 dopant atoms. With the 4 dopants situated in the corners (fig 7.9 (c)) the Si-Al bond lengths lie in the range 2.42 to 2.45 Å reflecting the increased stress of those locations. Fig 7.9(d) shows the coverage of unoccupied states at the VBE level remains incomplete and resembles that of the low concentration configuration 3 of fig 7.8 (d), where the dopant also occupies a corner location. This is due to the occluded character of the corner locations when compared to the centre. In some applications (e.g. co-doped p-i-n devices) non-uniform coverage and the presence of an intrinsic Si region might be desirable and therefore sought by design.

7.3.3 Ridge nanostructure

The ridge is a NW structure placed horizontally on a substrate of Si atoms. It might function as an interconnect between p-doped regions at the same potential, with the substrate boundary acting as passivation. The supercell consists of the 1.92 nm square structure of fig 7.7 (a) stacked sideways on an 8-layer substrate as shown at fig 7.10 (a) below. The combined structure has a repeat length of 8 layers in the x direction and H-passivation is applied to the ridge surfaces, the substrate surfaces and the substrate base. The supercell has an 11 Å vacuum spacing and contains 619 Si atoms of which 288 occupy the substrate. Four dopant positions are modelled: top corner (1), near centre (2), near base (3) and just inside the substrate (4) and the results summarized in figs 7.10 (b)-(e). These calculations were repeated with substrate depths of 4 and 12 layers for the same ridge dimensions, with similar results.

In this structure the ridge is electronically confined in the y direction, but the substrate is not. Separated, the NW would have a wider band gap than the more bulk-like substrate, as was seen in fig 7.3 (b) above. When combined, this disparity causes dispersal at both band edges (fig 7.7 (b)-(d)) which moderates when the dopant is moved into the substrate region (fig 7.7 (e)). However, inside the ridge incomplete p-type behaviour is apparent irrespective of dopant location.

As the ridge dopant concentration $(1/619)$ is comparable with that of the large NW considered above $(1/577)$, it is worthwhile repeating this calculation at the increased concentration used in that case. The results for 4 centrally placed Al dopants are shown in fig 7.11 below, where the PDOS plot (b) is again the summation over the dopant locations. The band structure shows some irregularity in the HOMO levels (presumably caused by the band gap disparity already noted) but a fully delocalized LUMO at the Fermi level now appears. The result shows that with proportionate doping the p-type ridge could function as an active element in a complex nanostructure configuration.

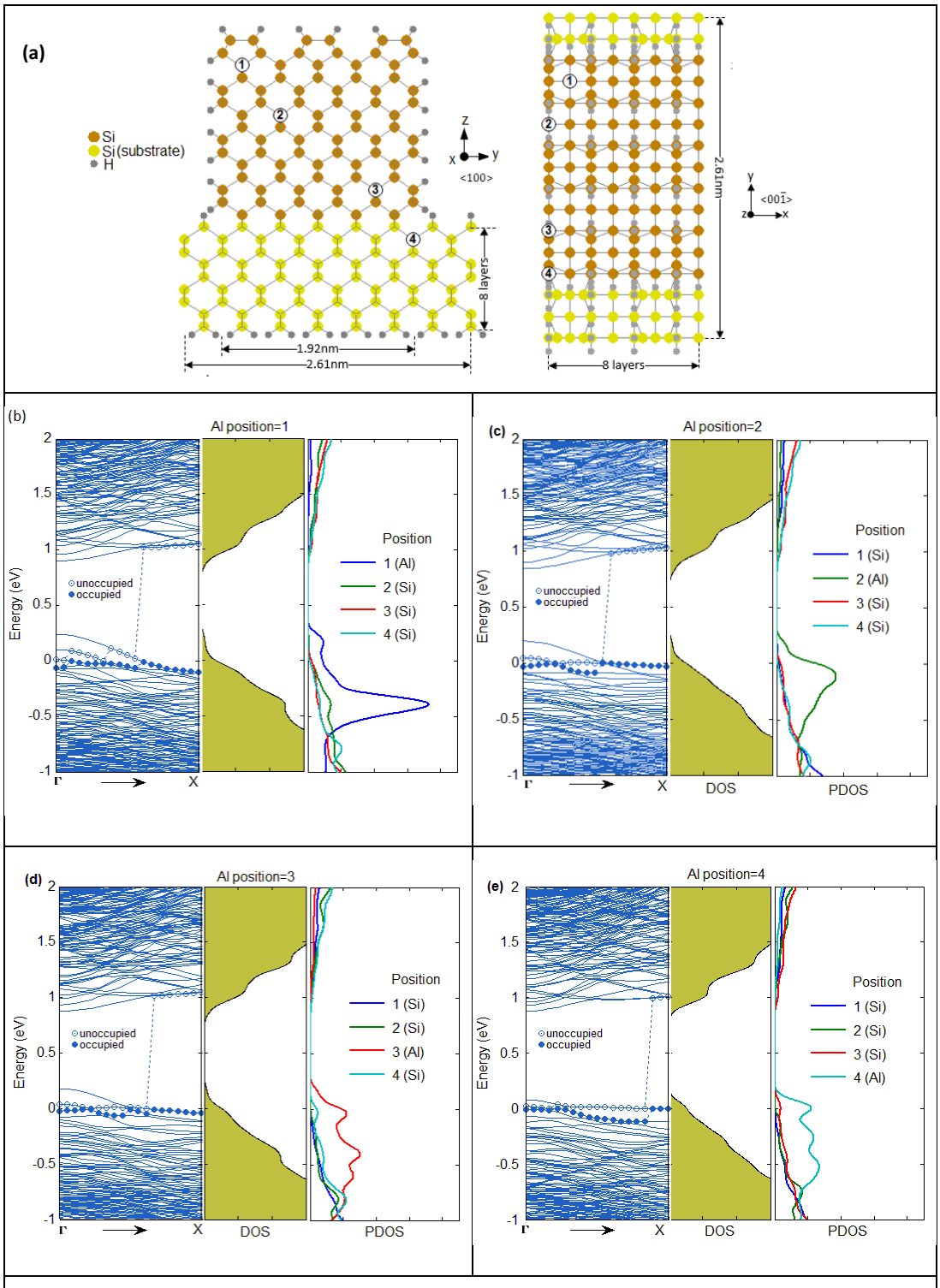
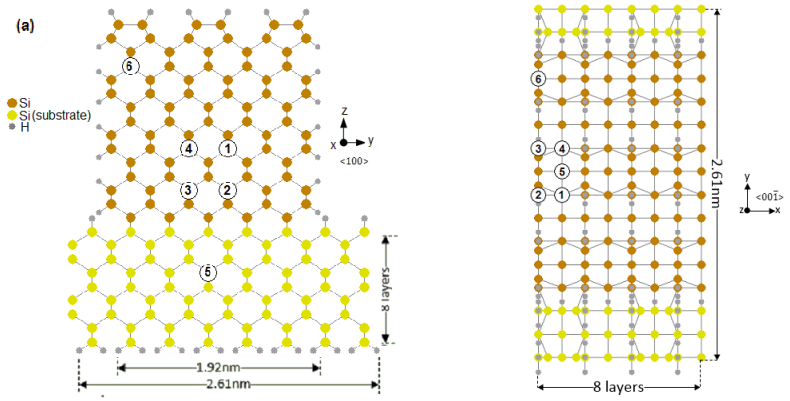


Figure 7.10 Comparison of band structure and density of states of an Al doped ridge structure consisting of a square (side 1.92 nm) section Si NW grown on an 8-layer Si substrate. The combined structure contains 619 Si atoms. (a) shows a cross sectional view perpendicular to the axis of the ridge with 4 dopant positions labelled 1–4, and an overhead view along the z-axis. The ridge atoms are coloured dark yellow, and the substrate atoms light yellow. Passivating H atoms are coloured grey. The simulation cell depth is 8 layers, in the x direction. The band structure, DOS and PDOS are shown for Al at positions 1–4 in (b)-(e), respectively.



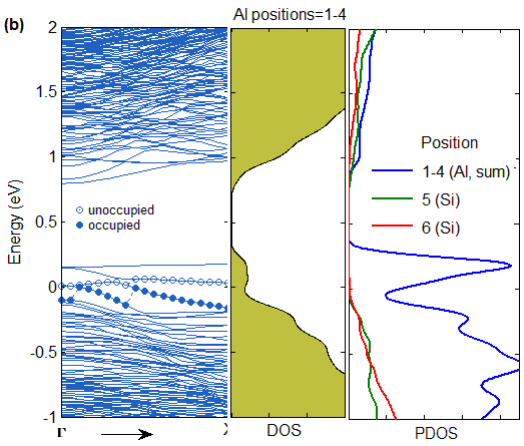


Figure 7.11 The ridge structure of fig 7.10, but containing 4 centrally placed Al and 616 Si atoms, as shown at (a). In (b) the blue PDOS plot is the sum over the projections onto the Al atoms at positions $1-4$.

7.3.4 Cell nanostructure

The cell is an isolated cubic structure passivated on all surfaces. When doped, it could be useful in shielding applications. The shielding effect arises from the superposition of an external field with the induced field within the cell. The Fermi level adjusts to make the net internal field zero, and an opposing charge appears on the exterior of the cell. There is no charge transfer to or from the cell (Griffiths, 1999). One might envisage a network of interconnected cells protecting some sensitive circuit element.

Here the supercell contains $11\text{x}11\text{x}12$ layers with 399 Si and 204 H atoms and an 11 Å vacuum spacing, as shown in fig 7.12 (a) below. Since the cell does not possess periodic boundaries a single k point suffices, taken to be $k = 0$. This makes the Bloch phase factor unity and the charge distribution independent of k (page 35). Consequently there is no band dispersion or structure in the usual sense, only a set of eigenenergies remaining constant throughout the IBZ as shown in figs 7.12 (b) and (c) corresponding to the dopant positions labelled 1 and 2. This lack of band structure would be true in any isolated molecule.

In (b) an unoccupied energy level appears at the VBE, consistent with p-type semiconductor behaviour. Although appearing to coincide with the HOMO this is a purely graphical

artefact. The calculation returns over 1000 distinct eigenenergies most of which are unresolved by this presentation. In (c) the LUMO is elevated, due to the increased PDOS seen beneath the Fermi level at locations near the surface (compare figs 7.7 (c) and 7.8 (c) on pages 117 and 118). This preserves charge neutrality in the larger structure. Both cell configurations show a wide (~ 1.5 eV) band gap, slightly greater than that calculated for the square NW of comparable side length (page 111) and consistent with increased confinement. Of course, this underestimates the actual band gap due to the GGA basis. However, the absence of partial occupation suggests that the p-doped cell could function as active nanocircuit element, if satisfactory ohmic connections could be made.

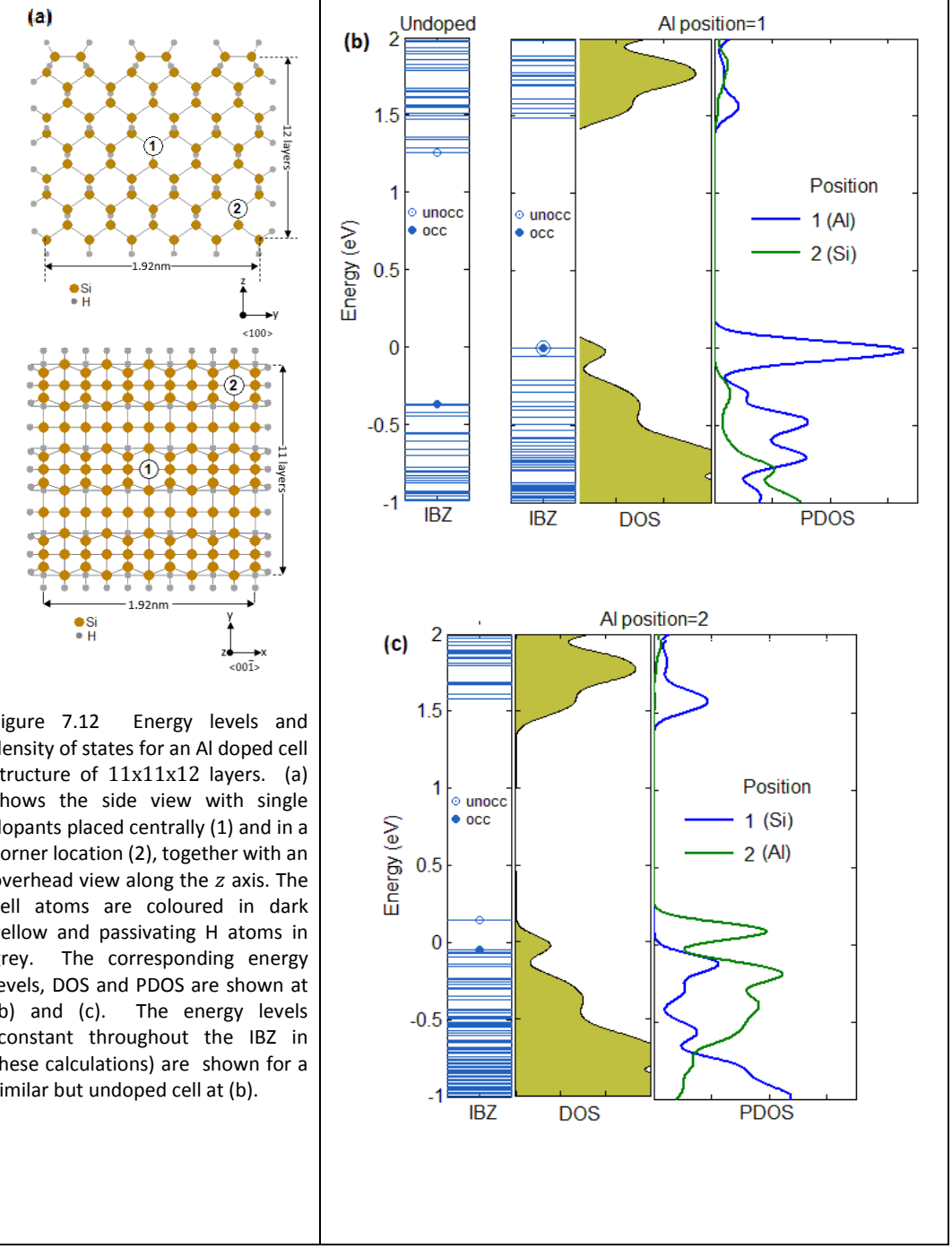


Figure 7.12 Energy levels and density of states for an Al doped cell structure of 11x11x12 layers. (a) shows the side view with single dopants placed centrally (1) and in a corner location (2), together with an overhead view along the *z* axis. The cell atoms are coloured in dark yellow and passivating H atoms in grey. The corresponding energy levels, DOS and PDOS are shown at (b) and (c). The energy levels (constant throughout the IBZ in these calculations) are shown for a similar but undoped cell at (b).

7.3.5 Pillar nanostructure

Finally, we consider the pillar nanostructure, a square section Si cell grown on an 8-layer Si substrate. The supercell repeats indefinitely in the x and y directions, with a vacuum gap of 11 Å and H termination is applied on all the surfaces, as shown at fig 7.13 (a) below. The structure contains 744 Si and 276 H atoms, and of the Si atoms 348 reside in the pillar and the remainder (396) in the substrate. VASP run times for structures of this size can be excessive, even when large scale computing resources can be deployed. To limit run times the cell dimension is reduced from 11x11 to 9x9 layers, which tends to accentuate confinement effects when compared with the NW and ridge structures having a larger cross-sectional area.

The band structure, DOS and PDOS plots corresponding to 4 dopant locations are shown in 7.13 (b)-(e). The IBZ path chosen for the band structure plots corresponds to the x axis. When the path taken corresponds to the z (vertical) axis the structure would become aperiodic and yield flat, cell-like energy levels. The number of k points on the x axis path is 10 (reduced from 20 elsewhere) again to manage run-time requirements.

As might be expected, the cell-like character of the pillar reduces band dispersion compared with the ridge of fig 7.10, with the substrate exerting less influence on the overall electronic structure. This results in relatively flat LUMO bands without partial occupation, even when the dopant is placed inside the substrate region, as in seen in (e). When the dopant occupies a corner location two distinct peaks appear in the state projection (b, blue plot). The peak at the Fermi level is due to the dopant atom, whereas the larger one immediately beneath is due to surface confinement. This can be compared with the ridge structure result at fig 7.10 (b), where the dopant is also in a corner location, but where only a single peak beneath the Fermi level is seen. This is an interesting result as it indicates that uniform p-doping is facilitated by confinement, permitting lowered concentration levels. This was also seen in the large NW (fig 7.8 (d) above) where asymmetric dopant placement yielded better p-type coverage in the low concentration setting. Additionally, the pillar format would assist in the design of complex 3-dimensional circuit configurations. However, here the pillar size is small and probably difficult to fabricate. In a sufficiently large pillar, it is likely that partial LUMO occupation would occur unless concentration were increased. This calculation should therefore be repeated for larger pillars, perhaps using the CONQUEST package already mentioned, although the maximum structure size is limited to ≈ 3000 atoms when running in its DFT (exact diagonalization) mode.

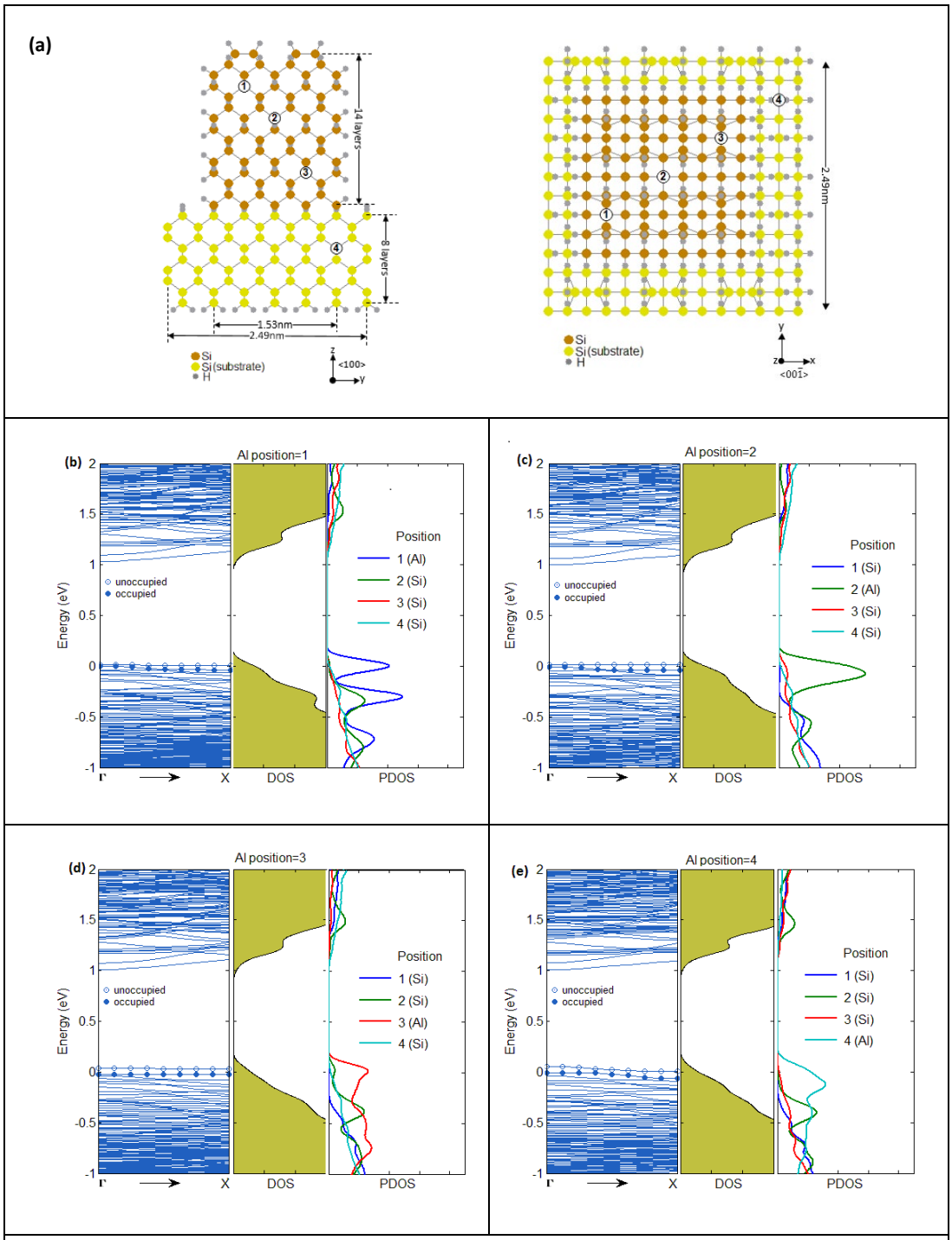


Figure 7.13 Comparison of band structure and density of states in an Al doped pillar structure consisting of a square section Si cell (side 1.66 nm) grown on an 8-layer Si substrate. (a) shows a cross sectional view through the cell perpendicular to the x axis and an overhead view along the z axis, with 4 Al dopant positions labelled $1-4$. The cell atoms are coloured in dark yellow, and the substrate atoms light yellow. Passivating H atoms are coloured grey. The band structure, DOS and PDOS are shown for Al at positions $1-4$ in (b)-(e), respectively.

7.4 Conclusion

The electronic structure of square section Al doped and undoped Si NWs with edge lengths ranging from ~ 2 to ~ 6 nm has been studied. Related ridge, cell, and pillar nanostructures containing up to 1000 atoms have also been investigated. For the undoped NWs, the band gap declines with increasing side length and converges to the theoretical bulk valued for Si. Atoms near the surfaces contribute less to states at the band edges, due to extended Si-Si bond lengths, when compared to centrally located atoms. However, in smaller NWs confinement can cause wide variations in the projected state density on atoms in asymmetric locations.

In the doped structures a uniform p-type behaviour is desirable for participation in junction-based circuit configurations. This is evidenced by the appearance of a flat LUMO band at the VBE, and the absence of partial hole occupation. This has been demonstrated in the smaller wires and the more confined cell and pillar structures but is hindered by the lowered dopant concentration and the more bulk-like nature of the ridge and larger NW structures where the onset of non-uniform behaviour occurs as dopant density falls below $\sim 1/500$. On the other hand, uniform p-type behaviour is enhanced by the increased confinement of NW corner locations and the cell and pillar structures. We find that uniform behaviour is restored when dopant concentration is increased using small clusters of dopant atoms.

Since P is a widely used donor dopant, it is unsurprising that references to its use crop up throughout the literature. In the context of this chapter (Reuβ et al., 2006) described narrow, heavily P doped planar NWs, exhibiting low resistivity and of interest in interconnect applications. The electronic structure of δ-doped Si-P was investigated by (Carter et al., 2009) using large slab supercells containing a single P layer of varying dopant density. These showed the same qualitative behaviour seen in Al i.e. a shifted Fermi level and conduction bands pulled into the band gap by the dopant potential. In the study by (Ng; Tong, 2012) the authors found the onset of bulk-like behaviour in H passivated P doped NWs occurred at a diameter of ~ 5 nm; at smaller diameters the band gap widens, as seen for B and Al. Finally, (Watson et al., 2017) reported on a qubit memory based on a double quantum dot made from P atoms by PALE, a direct descendant of the P atom transistor mentioned at the close of chapter 5.

The scope of these calculations was constrained by processing times, which can become excessive when large structures are optimized with VASP. Our results show that structures should be modelled at actual size, because performance parameters are size dependent. The 3-dimensional structures are relatively small and should be enlarged and the calculations repeated in a suitable DFT environment, as indicated in the text.

Chapter 8

Summary

This work presents a study into the use of Al as an atomically precise acceptor dopant in silicon semiconductor nanostructures. It is motivated by recent progress with P as donor (page 76) and more generally by the demise of Moore's law, which has stimulated a widespread search for alternatives to current CMOS technology. After addressing an area of concern in the proposed fabrication process we model the adsorption of alane (AlH_3) on the Si(100) surface and the subsequent migration and incorporation of the dopant atom. Finally, we investigate the electronic behaviour of Al doped Si nanowires and other cuboid structures. All the calculations were made on optimized structures within density functional theory using the gradient corrected functional of Perdew, Burke and Ernzerhof (PBE) as implemented in the VASP code, embodying periodic boundary conditions with plane wave basis sets (chapter 2). The results delivered by this approximation (i.e. the PBE functional and the associated projector-augmented wave pseudopotentials) were checked and found to agree with experimental data for crystalline Si (chapter 3).

Patterned Atomic Layer Epitaxy (PALE) is a layer-by-layer growth process for nanoscale structures under a mask of of passivating atoms. The original patent application (Randall et al., 2008) gives a general description without identifying the atomic species involved, but here the term refers to CVD growth of an Si nanostructure (e.g. a pillar) on an exposed area created by STM lithography of the H-passivated Si(100) surface. The surface energetics of the walls differ from those on the growth surface and chapter 4 applies transition state theory (TST) to calculate the probability of H diffusion from the walls back onto the growth surface, concluding that such contamination is unlikely. Unfortunately, the dimer-led growth process is frustrated by the appearance of antiphase boundaries (page 68), impeding epitaxial coverage and reducing the effectiveness of successive CVD cycles. This problem (and the inherently slow nature of PALE) represent real obstacles to future progress, but for now it is assumed that appropriate technological solutions will emerge in due course.

Chapter 5 investigates accurate placement of an Al dopant within a Si nanostructure. In the proposed scenario PALE growth cycles using the disilane precursor are interrupted, the surface repassivated and a small target area exposed by fresh lithography. A PALE cycle using an alane (ALH_3) precursor is then inserted. The amine alanes might serve in this role, but new experimental work is needed to confirm their effectiveness in the PALE setting. Some 60 possible alane decomposition pathways are subjected to TST analysis, predicting an initial adsorption of the Al atom, facile dissociation of the H ligands with the adatom finally making either 2 or 3 surface bonds. Incorporation (to a subsurface location) is less straightforward, as no significant gain in stability accrues; an analogous situation arises

during P doping with phosphine. However, in chapter 6 we show that incorporation is achieved after some selective depassivation around the adsorption site which allows surface migration, incorporation and activation at a temperature $\sim 400℃$, well within PALE process limits.

With a route to fabrication established, chapter 7 turns to the electronic performance of the doped nanostructures. We choose square nanowire and cuboid ridge, cell and pillar formations with edge lengths between ~ 2 and ~ 6 nm, in the expectation that these will emerge naturally from PALE growth. P type semiconductor behaviour is characterized by the retention of the intrinsic band gap, the location of the Fermi level at the valence band edge (VBE) and appearance of unoccupied states nearby. With a single Al atom present this behaviour is evident in each formation, fading towards intrinsic behaviour at concentrations $< 1/500$ in the relatively low confinement of the ridge and larger nanowires. In these cases, uniform behaviour is easily restored by an increase in dosage. On the other hand, the more confined cell and pillar formations permit concentrations as low as $\sim 1/800$ while maintaining uniform p type behaviour.

These results are encouraging and suggest that a theoretical study of P and Al co-doped junction diodes and transistors could be undertaken. These devices, together with a capacitive element could form a memory cell and the basis of future large scale integration. The footprint of such a cell would indicate the ultimate physical packing density achievable with this technology, ideally comparable with the pinhead densities referred to in the Impact Statement at the start of this thesis. However, the viability of these devices also depends on their electrical characteristics, which are not revealed in static electronic structure calculations but require an electron transport model and a non-equilibrium DFT, such as density functional perturbation theory (DFPT). These are challenging calculations and the capabilities of VASP and CONQUEST in this area should be assessed.

Appendix A

The Hohenberg-Kohn theorems

This appendix provides proofs for the Hohenberg-Kohn theorems quoted in chapter 2. It also introduces two supporting topics: the functional derivative and Lagrange's method undetermined multipliers. The latter material draws on notes of lectures delivered by Dr Michael Gillan.

A.1 The first Hohenberg-Kohn theorem

- *It is impossible that two external potentials $v(\tilde{r})$ and $v'(\vec{r})$ whose difference $v(\vec{r}) - v'(\vec{r})$ is not a constant give rise to the same ground-state density distribution $\rho_g(\vec{r})$.*

The proof is a *reductio ad absurdum* argument. Let $\hat{H} = \hat{H}_0 + \hat{V}$ and $\hat{H}' = \hat{H}_0 + \hat{V}'$ be the Hamiltonians associated with the two potentials $v(r)$ and $v'(r)$ and their ground state many-electron wavefunctions be called Ψ and Ψ'. These wavefunctions are necessarily different. Then by the variational principle of QM:

$$E' = \langle\Psi'|H'|\Psi'\rangle < \langle\Psi|H'|\Psi\rangle = \langle\Psi|H_0 + V + (V' - V)|\Psi\rangle \tag{9}$$

Hence:

$$E' < E + \langle\Psi|V' - V|\Psi\rangle = E + \int dr\rho(\vec{r})(v'(\vec{r}) - v(\vec{r})) \tag{10}$$

and similarly (after interchanging primed and unprimed quantities):

$$E < E' + \int d\vec{r}\rho'(\vec{r})(v(\vec{r}) - v'(\vec{r})). \tag{11}$$

Adding the two equations, we get:

$$E' + E < E + E' - \int dr(\rho'(\vec{r}) - \rho(\vec{r})(v'(\vec{r}) - v(\vec{r})) \tag{12}$$

so that $\rho(\vec{r})$ cannot equal $\rho'(\vec{r})$, otherwise the inequality would not hold.

A.2 The second Hohenberg-Kohn theorem

- *The ground state energy for a given external potential $v(\vec{r})$ is correctly obtained by minimizing the functional $E_g = \int d\vec{r}\rho(\vec{r})v(\vec{r}) + F[\rho(\vec{r})]$ with respect to $\rho(\vec{r})$ subject to a fixed number of electrons N, and the resulting $\rho(\vec{r})$ gives the correct density distribution of the ground-state.*

It is necessary to show that, if E_g is the ground-state energy associated with external potential $v(\vec{r})$ then:

$$E_g \leq \int d\vec{r} v(\vec{r})\rho'(\vec{r}) + F[\rho'(\vec{r})], \tag{13}$$

where $\rho'(\vec{r})$ is the density associated with any arbitrary potential $v'(\vec{r})$. The equality holds only if $\rho'(\vec{r}) = \rho(\vec{r})$ when (by the first theorem) $v'(\vec{r})$ and $v(\vec{r})$ differ by at most an additive constant.

The theorem follows directly from the variational principle. Let Ψ and Ψ' be the many-electron ground-state wavefunctions associated with different external potentials $v(\vec{r})$ and $v'(\vec{r})$. As the ground states are assumed to be not degenerate Ψ and Ψ' are uniquely defined. Then:

$$E_g < \langle \Psi' | H_0 + V | \Psi' \rangle = \int d\vec{r} v(\vec{r})\rho'(\vec{r}) + F[\rho'(\vec{r})] \tag{14}$$

proving the theorem.

A.3 Functionals

In DFT, the total energy E_{tot} depends on the electronic density distribution $\rho(\vec{r})$ and we consider the variation of when E_{tot} when the function ρ changes. So, we need notation to discuss quantities that depend not on a discrete set of variables, but on a function. Such a quantity is called a *functional*.

Consider a quantity E that depends on a function $f(x)$, with f depending on a single variable x. To indicate that E depends on $f(x)$, it is usual to write $E = E[f(x)]$ and express E in the form of an indefinite integral e.g.:

$$E[f(x)] = \int p(x) f(x) dx \tag{15}$$

$$E[f(x)] = \int f(x)^2 dx \tag{16}$$

where the function $p(x)$ is a fixed function and the integrals are taken over some chosen interval, e.g. range $(-\infty, +\infty)$. In each case, for any chosen function $f(x)$, E has some numerical value.

A.3.1 The functional derivative

The rate of change of a functional with variation of the function on which it depends is called a *functional derivative*. Instead of being a number (or a set of numbers, in the case of partial derivatives), it is a function. When the function $f(x)$ is changed to $f(x) + \delta f(x)$ in equation (15) above, the corresponding change in E is:

$$\delta E = \int p(x) \delta f(x) dx \tag{17}$$

So δE can be expressed as an integral involving $\delta f(x)$. This is true even when $f(x)$ is multiplied by a function which is not constant, as in (16):

$$\delta E = \int 2 f(x) \delta f(x) dx \tag{18}$$

Generally, when the $f(x)$ on which E depends is changed infinitesimally, the change δE can be expressed as an integral of some other function times the change $\delta f(x)$. The function appearing in the integral is called the functional derivative and written $\delta E/\delta f(x)$. The functional derivative is therefore defined by:

$$\delta E = \int \frac{\delta E}{\delta f(x)} \delta f(x) dx \tag{19}$$

in the limit of $\delta f(x)$ becoming infinitesimally small. In the examples given above, the functional derivatives are:

$$\frac{\delta E}{\delta f(x)} = p(x) \tag{20}$$

$$\frac{\delta E}{\delta f(x)} = 2x \tag{21}$$

In the DFT setting $f(x)$ is approximated by a linear combination of fixed basis functions. Let these be $\phi_i(x)$ ($i = 1, 2, \ldots P$), which are not allowed to change. Then we consider all functions $f(x)$ that can be expressed as:

$$f(x) = \sum_{i=1}^{P} c_i \phi_i(x) \tag{22}$$

where c_i are coefficients. Then $E[f(x)]$ depends only on the c_i and the only allowed changes $\delta f(x)$ have the form:

$$\delta f(x) = \sum_{i=1}^{P} \delta c_i \phi_i(x) \tag{23}$$

so that the resulting change of E is:

$$\delta E = \int \frac{\delta E}{\delta f(x)} \delta f(x) dx = \sum_{i=1}^{P} \delta c_i \int \frac{\delta E}{\delta f(x)} \phi_i(x) dx \tag{24}$$

From this, the partial derivative of E with respect to c_i is expressed in terms of the functional derivative as:

$$\frac{\partial E}{\partial c_i} = \int \frac{\delta E}{\delta f(x)} \phi_i(x) dx \tag{25}$$

A.3.2 Stationary point of a functional

In DFT, we must minimize the total energy with respect to the electronic density distribution. The general functional $E[f(x)]$ is stationary with respect to variation of $f(x)$ when its functional derivative is zero: $\delta E/\delta f(x) = 0$ for all x. This means that the variation of E to linear order in $\delta f(x)$ is zero:

$$\delta E = \int \frac{\delta E}{\delta f(x)} \delta f(x) dx = 0 \tag{26}$$

If E is approximated by basis functions as in (22) this condition is satisfied when $\partial E/\partial c_i = 0$ for all i and the problem is equivalent to minimizing E over P variables. To invoke Lagrange's method of undetermined multipliers (see A.4 below) we impose an additional constraint on E, chosen to be:

$$\sum_{i=1}^{P} c_i \int \phi_i(x)dx = N \tag{27}$$

where N is some constant. Then, the Lagrange condition (37) is that the quantity $E - \mu N$ be minimized, where μ is the undetermined multiplier. The condition for the minimum is

$$\frac{\partial E}{\partial c_i} = \mu \frac{\partial N}{\partial c_i} \tag{28}$$

From (19) and (27) this is equivalent to

$$\int \frac{\delta E}{\delta f(x)} \delta f(x)dx = \mu \int \phi_i(x)dx \tag{29}$$

for all i. In the limit of the ϕ_i forming a complete basis as $P \to \infty$, this requires

$$\frac{\delta E}{\delta f(x)} = \mu \tag{30}$$

and this is the general condition for a stationary point of E, subject to the constraint in the form of (27) above.

A.3.3 Functions of more than one variable

In DFT the total energy E_{tot} depends on $\rho(\vec{r})$, itself a function of the positional vector $\vec{r}$ consisting of three variables (x, y, z). The definition (19) still applies, so

$$\delta E_{tot} = \int \frac{\delta E_{tot}}{\delta \rho(\vec{r})} \delta\rho(\vec{r})d\vec{r} \tag{31}$$

where the integral is now over the volume of the system. The condition for a stationary point of E_{tot} is:

$$\frac{\delta E_{tot}}{\delta \rho(\vec{r})} = 0 \tag{32}$$

when there is no constraint. If a constraint of the form:

$$\int \rho(\vec{r}) = N \tag{33}$$

Is applied, the stationary point is:

$$\frac{\delta E_{tot}}{\delta \rho(\vec{r})} = \mu \tag{34}$$

A.4 Lagrange's method of undetermined multipliers

Suppose we have a function $f(x_1, x_2, \dots, x_P)$ depending on P variables x_i, $i = \{1,2 \dots, P\}$. The condition that f has a stationary point (i.e. a minimum, a maximum, or a saddle point) for given values of the variables x_i is:

$$\frac{\partial f}{\partial x_i} = 0 \tag{35}$$

for all x_i. Let $\vec{x} \equiv \{x_1, x_2, \dots, x_P\}$ denote the P-dimensional vector whose components are x_i and $\vec{x}^*$ be a stationary point. Then at a nearby point $\vec{x} = \vec{x}^* + \delta\vec{x}$, a Taylor series expansion shows that

$$\delta f = f(\vec{x}^* + \delta\vec{x}) - f(\vec{x}^*) = \sum_{i=1}^{P} \frac{\partial f}{\partial x_i} \delta x_i + O\left(\delta x_i^2\right) \tag{36}$$

The term linear in δx_i vanishes, because $\partial f / \partial x_i = 0$ for all i and so $\delta f = 0$ as expected.

Now suppose that the x_i are constrained by the condition that another function $\phi(x_1, x_2, \dots, x_P)$ has the fixed value c. In the P-dimensional space of vectors $\vec{x}$, there is some surface on which $\phi(x_1, x_2, \dots, x_P) = c$. We seek the stationary points of f subject to this constraint. Because of the constraint, it is no longer true that $\partial f / \partial x_i = 0$ for all i at a stationary point. Instead, we can show that the constrained stationary point is an unconstrained stationary point of the function

$$f - \mu\phi \tag{37}$$

for an appropriate value of μ. The parameter μ is called the *Lagrange undetermined multiplier*, because it is initially unknown. The value chosen for μ depends on the constant c, and we search for the value of μ that gives the desired value of c.

To verify this, let $\vec{x}^*$ be an unconstrained stationary point of the function F. At this stationary point, we have from (35):

$$\frac{\partial F}{\partial x_i} = \frac{\partial f}{\partial x_i} - \mu \frac{\partial \phi}{\partial x_i} = 0 \tag{38}$$

for all x_i. Then from (36) the change δf at a nearby point $\vec{x} = \vec{x}^* + \delta\vec{x}$ is

$$\delta f = \sum_{i=1}^{P} \frac{\partial F}{\partial x_i} \delta x_i \tag{39}$$

But if $\partial f / \partial x_i = \mu\, \partial\phi / \partial x_i$ for all i, then the change of f is:

$$\delta f = \sum_{i=1}^{P} \frac{\partial f}{\partial x_i} \delta x_i = \mu \sum_{i=1}^{P} \frac{\partial \phi}{\partial x_i} \delta x_i \tag{40}$$

Now for all displacements satisfying the constraint, there is no change of ϕ to linear order in $\delta\vec{x}$, so that $\sum_{i=1}^{P} (\partial\phi / \partial x_i) \delta x_i = 0$, which implies that:

$$\sum_{i=1}^{P} \frac{\partial f}{\partial x_i} \delta x_i = 0 \tag{41}$$

and that $\vec{x}^*$ is a constrained stationary point, as required.

Appendix B

MATLAB programs

This appendix provides descriptions of programs written by the author of this work for processing VASP input and output.

B.1 vasp_menu

Functions collectively providing job submission and related functionality in the UCL HPC (High Performance Computing) environment.

UCL researchers enjoy access to several HPC sites including SALVIATI, LEGION, GRACE, THOMAS and ARCHER. These are large multi-core systems running Linux nodes and a Sun-derived job scheduler (batch queueing) system. However queueing, accounting and login policies originate locally, so job submission procedures are to some extent site-specific.

In this context a job equates to a single invocation of the VASP executable. Input consists of several VASP parameter files and an ionic structure file. Job submission requires assembly of these files (along with an appropriate control script) in an HPC directory which is copied to a run-time directory accessible by the processor nodes. Then the job is enqueued by calling the scheduler with the control script. On successful completion, the run-time directory contains VASP output files which the user can copy back to the workstation for analysis. A project containing hundreds of jobs over multiple HPCs will proliferate thousands of files and provide ample scope for error. It is therefore worthwhile automating the job procedures provided this can be done at reasonable cost and this is the motivation for **vasp_menu.**

vasp_menu is a menu-driven program containing job profiles for static optimization, structural relaxation, NEB and DOS calculations and providing easy access to editors and viewers of the relevant VASP data files. HPC-specific data transfer and job submission scripts are generated dynamically and submitted as WINSCP batch requests. POSCAR structure files are validated and appropriate pseudopotential files automatically assembled. OUTCAR files are checked for normal completion and convergence and energy values extracted. It is impossible to incorporate output files from a failing or unconverged job into input to a subsequent job.

B.2 gen_struct.m

A function to create Si nanostructure configuration files in VASP POSCAR format. The structures are oriented with an Si(100) growth surface in the z direction and present Si(110) surfaces in the x and y directions. The VMD (Humphrey et al., 1996) molecular viewer is called to render a 3-dimensional image of the structure. A user-written VMD extension allows insertion of dopants by double-clicking on selected Si atoms, which are

replaced. The output POSCAR file is placed in a nominated configuration directory compatible with the **vasp_menu** job submission system.

The function can be called from another function or a MATLAB script and accepts keyword parameters as shown in the table below:

keyword	***allowed values***	***note***
'cfg'	'wire'|'ridge'|'cell'|'pillar'|'surface'|'bulk'	Structure type
'side'	numeric	# Si layers
'width'	numeric	# Si layers
'depth'	numeric	# Si layers
'base'	numeric	# Si layers
'base_side'	numeric	# Si layers
'base_depth'	numeric	# Si layers
'side_vac'	numeric	Vacuum interval above (100) surface (Å)
'h'	'y'|'n'	H-terminate surface
'dimerize'	'y'|'n'	dimerize, buckle and H-terminate (100) surface

Table B.1 Keyword parameters to the MATLAB gen_struct function.

Bibliography

Ashcroft, N W; Mermin, N D (1975). *Solid State Physics*. Brooks/Cole

Ballard J B et al. (2014). *Spurious dangling bond formation during atomically precise hydrogen depassivation lithography on Si (100): The role of liberated hydrogen*. J. Vac. Sci. Technol. B32(2) 021805

Ballard J B et al. (2014). *Pattern transfer of hydrogen depassivation lithography patterns into silicon with atomically traceable placement and size control.* J. Vac. Sci. Technol. B32 041804

Ballmer, S A (2007). *'There's no chance that the iPhone is going to get any significant market share. No chance'.* Ballmer was Microsoft CEO from 2000 until 2014. Comment taken from an interview with David Letterman of *USA Today*.

Battaglia, C et al. (2009). *Elementary structural building blocks encountered in silicon surface reconstruction*. J. Phys.: Condens. Matter 21 013001

Becke, A D; Edgecombe, K E (1990). *A simple measure of electron localization in atomic and molecular systems.* J. Chem. Phys. 92 5397

Birch, F (1947). *Finite Elastic Strain of Cubic Crystals*. Physical Review. **71** (11): 809–824

Bitzek, E et al. (2006). *Structural Relaxation Made Simple*. Phys. Rev. Lett. 97, 170201

Blöchl, P E (1994). *Projector augmented-wave method*. Phys. Rev. B, 50:17953-17979

Blügel, S; Bihlmayer G (2006). *Full-Potential Linearized Augmented Planewave Method*. John von Neumann Institute for Computing, Jülich, Germany. NIC Series, Vol. 31, ISBN 3-00-017350-1, pp. 85-129

Bowler, D R et al. (1998). *Hydrogen diffusion on Si(001) studied with the local density approximation and tight binding*. J. Phys.: Condens. Matter 10 3719

Bowler, D R*;* Miyazaki, T (2010). *Calculations for millions of atoms with density functional theory: linear scaling shows its potential.* J. Phys.: Condens. Matter 22 074207

Brázdová, V; Bowler D R (2011). *H atom adsorption and diffusion on Si(110)-(1×1) and (2×1) surfaces*. Phys. Chem. Chem. Phys. 13 11367

Breisacher, P; Siegel B (1964). *Gaseous Alane and Dialane*. J. Am. Chem. Soc. 86 5053

Cardona, M; Pollack, F H (1966). *Energy-Band Structure of Germanium and Silicon: The k·p Method*. Phys. Rev. 142, 530

Carter, D J et al. (2009). *Electronic structure models of phosphorous δ-doped silicon*. Phys. Rev. B 79, 033204

Ceperley, D M; Alder, B (1980). *Ground state of the electron gas by a stochastic method.* Phys. Rev. Lett. 45(7)

Chadi, D J (1987). Stabilities of Single layer and bilayer steps on Si(001) surface. Phys. Rev. Lett. 59, 1691

Colinge J.P. et al. (2010). *Nanowire transistors without junctions*. Nature Nanotech. 5, 225–229

Curson, N J, Schofield, S R et al. (2004). *STM characterization of the Si-P heterodimer*. Physical Review B 69, 195303

Dennard R H, et al. (1974). *Design of ion-implanted MOSFET's with very small physical dimensions.* IEEE J. Solid-State Circuits 5 256

Dürr, M; Höfer, U (2013). *Hydrogen diffusion on silicon surfaces*. Prog. Surf. Sci. 88 61

Evastorev, R A; Petrashen M I; Ledovskaya, E M. (1975). *The translational symmetry in the molecular model of solids*. Phys. Stat. Sol. B, 68:453-461

Eyring, H (1935). *The Activated Complex in Chemical Reactions.* J. Chem. Phys. 3, 107.

Feynman, R P (1939). *Forces in molecules.* Phys. Rev. 56:340

Feynman, R P (1960). *'There's plenty of room at the bottom'*. Talk given to the annual meeting of the American Physical Society and reprinted in *Miniaturization*, *Horace D Gilbert, Ed*., Van Nostrand Rheinhold

Franssila, S (2010). *Introduction to microfabrication* (2nd ed.). Chichester, West Sussex, UK: John Wiley & Sons

Fuechsle, M et al. (2012). *A single-atom transistor*. Nature Nanotechnology 7, 242-246

Fuster, F; Sevin, A; Silvi, B (2000). *Topological analysis of the Electron Localization Function (ELF) applied to the electrophilic aromatic substitution*. J. Phys. Chem. A, 104, 4, 852–858

Greenwood, N; Earnshaw, A (1997). *Chemistry of the Elements* (2nd ed.). Oxford, UK: Butterworth-Heinemann

Gillan, M J; Alf, D; Michaelides, A (2016). *Perspective: How good is DFT for water?* J. Chem. Phys. 144, 130901

Gladfelter, W L et al. (1989*). Trimethylamine complexes of alane as precursors for the low-pressure chemical vapor deposition of aluminum*. Chem. Mater. 1 339

Griffiths, D.j. (1999). *Introduction to Electrodynamics*. Page 183. Prentice Hall

Hamann, D R et al. (1979). *Norm-Conserving Pseudopotentials.* Phys. Rev. Lett. 43 1494

Hamers, R J; Wang, Y (1996). *Atomically resolved studies of the chemistry and bonding at silicon surfaces*. Chem. Rev. 96, 1261-1290

Hashizume, T et al. (1996). *Interaction of Ga adsorbates with dangling bonds on the hydrogen terminated Si(100) surface*. Jpn. J. Appl. Phys. Pt 2 35, L1085

He, Y et al. (2019). *A two-qubit gate between phosphorus donor electrons in silicon.* Nature 571**,** 371–375 https://doi.org/10.1038/s41586-019-1381-2

Hellman, H (1937). *Einfuhrung in die Quantumchemie.* Franz Duetsche, Leipzig.

Henkelman, G; Uberuaga, B P; Jonsson, H (2000). *A climbing image nudged elastic band method for finding saddle points and minimum energy paths.* J. Chem. Phys. 113 9901.

Hoddesdon, l et al. (1992). *Out of the Crystal Maze: Chapters from the History of Solid-State Physics.* Oxford, UK: Oxford University Press

Hofer W, (2003). *Challenges and errors: interpreting high resolution images in scanning tunneling microscopy*. Prog. Surf. Sci. 71 147

Hoyt J, et al. (2002). *Strained Silicon MOSFET Technology*. Microsystems Technology Laboratory, MIT, Cambridge, MA 02319, USA

Huang, Y et al. (2001). *Logic Gates and Computation from Assembled Nanowire Building Blocks*. Science Vol. 294, Issue 5545, pp. 1313-1317

Humphrey, W; Dalke A; Schulten K (1996). *VMD – Visual Molecular Dynamics*, J Molec. Graphics 14.1, 33-38

Itoh, K (2008). *The history of DRAM Circuit Designs.* IEEE SSCS News

Iwai H, et al. (2011). *Si nanowire FET and its modeling*. Sci China Inf Sci, 54: 1004–1011, doi: 10.1007/s11432-011-4220-0

Iwai, H (2012). *Si nanowire technology*. The Electrochemical Society. ECS Transactions 50 (4) 251-260 10.1149/05004.0251ecst

Jones A; Hitchman, M L (ed) (2009). *Chemical Vapour Deposition: Precursors, Processes and Applications.* (Cambridge: RSC Publications) (https://doi.org/10.1039/9781847558794)

Jonsson, H et al. (1998). *Nudged Elastic Band Method for finding minimum energy paths of transitions*. Classical and Quantum Dynamics in Condensed Phase Simulations edited by Berne B J, Cicotti G and Coker D F (Word Scientific, Singapore 1998) p. 385

Kane, B (1998). A silicon-based nuclear spin computer. Nature, Vol 393, 14 May 1998: Macmillan

Keeler, J; Wothers P (2003). *Why chemical reactions happen*. Oxford, UK: Oxford University Press

Kerker, G P (1981). *Efficient iteration scheme for self-consistent pseudopotential calculations*. Phys. Rev. B*,* 23(6):3082-3084

Kittel, C (2005). *Introduction to solid state physics* (8th ed.). New Delhi: John Wiley

Klimes, J et al. (2010*). A critical assessment of theoretical methods for finding reaction pathways and transition states of surface processes*. J. Phys.: Condens. Matter 22 074203

Kohn, W; Sham L J (1965). *Self-consistent equations including exchange and correlation effects.* Phys. Rev**.** 140, 1133

Kresse, G; Furthmüller, J (1996). *Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set.* Comput. Mat. Sci. 6, 15-50

Kresse, G; Joubert, D (1999). *From ultrasoft pseudopotentials to the projector augmented-wave method.* Phys. Rev. B, 59:1758-1775

Kresse, G et al. (2009). *Vienna Ab-initio Simulation Package (VASP the Guide)*. University of Vienna, Austria

Kohn, W (1999). *Nobel Lecture: Electronic structure of matter—wave functions and density functionals*. Rev. Mod. Phys. 71 1253

Kumarasinghe, C; Bowler, D R (2019). *DFT study of undoped and As-doped nanowires approaching the bulk limit.* J. Phys.: Condens. Matter 32 035304

Lent, C.S. et al. (1993). *Quantum cellular automata*. Nanotechnology 4, 49–57

Lieb, E H; Oxford, S (1981). *Improved lower bound on the indirect Coulomb energy*. Int. J. Quantum Chem., 19:427.439.

Lyding, J W et al. (1994). *Nanoscale patterning and selective chemistry of silicon surfaces by ultrahigh-vacuum scanning tunneling microscopy.* J. Vac. Sci. Technol. B 12 3735

Ma, D D et al. (2003). *Small-diameter silicon nanowire surfaces.* Science 299 1874

Marsh C M B et al. (1995). *Lack of Dative Bonds in bis(ammonia)alane.* J. Chem. Phys. 99 14309

Martin, C (2014). *Binary Challenge*. Nature Nanotech 9 89-90

Martin, R M (2004). *Electronic Structure. Basic Theory and Practical Methods*. Cambridge, UK: Cambridge University Press

Metcalfe, R (1995). *'I predict the internet … will soon go spectacularly supernova and in 1996 will catastrophically collapse'*. In InfoWorld magazine, December 1994. Metcalfe co-invented the Ethernet communications protocol, on which the internet is built.

Moll, N et al. (1995). *Application of generalized gradient approximations: The diamond -β-tin phase transition in Si and Ge*. Phys. Rev. B52, 2550

Monkhorst, H; Pack J (1976). *Special Points for Brillouin-zone integrations*. Phys. Rev. B 13, 5188

Ng, M-F et al. (2011). *First-Principles Study of Silicon Nanowire Approaching the Bulk Limit*. Nano Lett. 11, 4794–4799

Ng, M-F; Tong S W (2012). *Chemically Doped Radial Junction Characteristics in Silicon Nanowires.* Nano Lett. 12, 6133–6138

Owen, J H G et al. (2011). *Patterned Atomic Layer Epitaxy of Si/Si(001):H.* J. Vac. Sci. Technol. B29 06F201

Owen, J H G et al. (1996). *Hydrogen diffusion on Si(001)*. Phys. Rev. B 54 14153

Owen, J H G et al. (1997). *Gas-source growth of group IV semiconductors: I. Si(001) nucleation mechanisms.* Surf. Sci. 394 79, 91

Payne, M C et al. (1992). *Iterative minimization techniques for ab initio total-energy calculations: molecular dynamics and conjugate gradients*. Rev. Mod. Phys. 64:1045-1097

Peng, K; Lee, S (2011). *Silicon nanowires for photovoltaic energy conversion*. Adv. Mater.23 198

Perdew, J P (1991). *Unified theory of exchange and correlation beyond the local density approximation*. In Electronic Structure of Solids '91; Ziesche, P and Eschrig, H, editors. Berlin: Akademie Verlag GmbH.

Perdew, J P; Burke K; Ernzenhof M (1996). *Generalized Gradient Approximation made Simple*, Phys. Rev. Lett. 77, 3865

Perdew, J P; Zunger, A (1981). *Self-interaction correction to density-functional approximation for many-electron systems.* Phys. Rev. B, 23(10)

Press, W et al. (2007). *Numerical Recipes: The Art of Scientific Computing*, 3rd edition. Cambridge UK; Cambridge University Press

Randall, J N et al. (2008). *Patterned Atomic Layer Epitaxy*, US patent application US7326293B2

Reuβ, F J et al. (2006). *Narrow, highly P-doped, planar wires in silicon created by scanning probe microscopy*. Nanotechnology 18, 044023

Rinke, P et al. (2008). *Exciting prospects for solids: Exact-exchange functionals meet quasiparticle energy calculations*. Phys. Stat. Sol. (b) 245, No. 5 929-945

Roosgaard, C (2009). *The Projector Augmented-wave Method*. arXiv:0910.1921v2 [cond-mat.mtrl-sci]

Sagisaka, K et al. (2017). *Importance of bulk states for the electronic structure of semiconductor surfaces: implications for finite slabs*. J. Phys.: Condens. Matter 29 145502

Samsung (2019). *Exynos 9825c: first 7nm EUV*. Press release, Seoul, South Korea

Sarubbi, F et al. (2010). *Chemical Vapor Deposition of α-Boron Layers on Silicon for Controlled Nanometer-Deep pn Junction Formation*. J. Electron.Mater. 39 162

Savin A et al. (1997). *ELF: The Electron Localization Function.* Angew Chem. Int. Ed. Engl. 36 1808

Schofield, S R; Curson, N J et. al (2006). *Phosphine dissociation and diffusion on Si(001) observed at the atomic scale*. J. Phys. Chem B, 110, 3173-3179

Sholl, D; Steckel, J (2009): *Density Functional Theory - A Practical Introduction.* Wiley

Sheppard, D; Terrell, R; Henkelman, G (2008). *Optimization methods for finding minimum energy paths.* J. Chem. Phys. 1

Shen, T C et al. (1995). *Atomic-scale desorption through electronic and vibrational excitation mechanisms*. Science 268 1590

Shor, P W (1997). *Polynomial-Time Algorithms for Prime Factorization and Discrete Logarithms on a Quantum Computer.* SIAM J. Sci. Statist. Comput. 26 1484

Škereň, T et al. (2020). *Bipolar device fabrication using a scanning tunnelling microscope.* Nat Electon 3 524-530

Simmons, M Y; Fuechsle, M (2013). In *Single Atom Nanoelectronics Prati, E; Shinada T Eds*. Pan Stanford Publishing Pte. Ltd. pp 61-80

Simmons, M Y; Curson, N J (2003). *Towards the atomic-scale fabrication of a silicon-based solid state quantum computer.* Surf. Sci. 532-535

Smith, R; Bowler, D R (2017). *Alane adsorption and dissociation on the Si(001) surface*. J. Phys.: Condens. Matter 29 395001

Smith, R; Bowler D R (2017). *Figshare.* https://doi.org/10.6084/m9.figshare.c.3727687

Smith, R; Bowler, D R (2018). *Reaction paths of alane dissociation on the Si(001)* surface **J.** Phys.: Condens. Matter 30 105002

Smith, R; Brázdová, V; Bowler, D R (2014). *Hydrogen adsorption and diffusion around Si(001)/Si(110) corners in nanostructures*. J. Phys.: Condens. Matter 26 295301

Streetman, B; Banerjee, s (2015). *Solid State Electronic Devices*. Chapter 3, Pearson

Szabo, A; Ostlund, N (1996). *Modern Quantum Chemisty: Introduction to Advanced Electronic Structure Theory.* Chapter 2, Dover

Takahash, I P et al. (2002). *Silicon single-electron devices.* J. Phys.: Condens. Matter 14 R995

Tao, J et al. (2003*). Climbing the Density Functional Ladder: Nonempirical Meta–Generalized Gradient Approximation Designed for Molecules and Solids*. Phys. Rev. Lett. 91 146401

Tersoff, J; Hamann. D R (1985). *Theory of the scanning tunneling microscope.* Phys. Rev. B 031 805

Vanderbilt, D (1990). *Soft self-consistent pseudopotentials in a generalized eigenvalue formalism.* Phys. Rev. B, 41:7892-7895

Von Barth, U; Hedin L (1972). *A local exchange-correlation potential for the spin polarized case.* J. Phys. C: Solid State Phys. 5 1629

Wagner, R; Ellis, W (1964). *Vapor Liquid Solid Mechanism of single crystal growth*. Appl. Phys. Lett. 4, 89

Walsh, M; Hersam, M (2009). *Atomic-scale Templates Patterned by UHV STM on Silicon*, Ann. Rev. Phys. Chem. 60:193-216

Wang, Y; Hamers, R (1995*). Boron-induced reconstructions of Si(001) investigated by scanning tunneling microscopy* J. Vac. Sci. Technol. A 13 1431

Wang, Y et al. (1996). *Combined scanning tunneling microscopy and infrared spectroscopy study of the interaction of diborane with Si(001).* J. Vac. Sci. Technol. B14 1038

Wang, Y et al. (2018). *16-qubit IBM universal quantum computer can be fully entangled.* npj Quantum Inf **4,** 46. https://doi.org/10.1038/s41534-018-0095-x

Warschkow, O et al. (2005). *Phosphine adsorption and dissociation on the Si(001) surface: An ab initio survey of structures*. Phys. Rev. B 72 125328

Warschkow, O et al. (2016). *Reaction paths of phosphine dissociation on silicon (001).* The Journal of Chemical Physics 144 014705; doi: 10.1063/1/4939124

Wolkow, R A et al. (2014). In *Field-Coupled Nanocomputing: Paradigms, Progress, and Perspectives. Anderson, N G; Bhanja, S Eds.* Springer; pp 33–58

Woods, N D et al. (2019*). Computing the Self-Consistent Field in Kohn-Sham Density Functional Theory.* https://arxiv.org/abs/1905.02332

Wu Y et al. (2004). *Controlled growth and structures of molecular scale silicon nanowires.* Nano Lett. 2004 4 433