Epitaxial Nanostructures across Length and Time Scales

Dimitri D. Vvedensky
The Blackett Laboratory, Imperial College, London

Contents

1 Introduction and Overview **1**

1.1 Epitaxial Systems across Length and Time Scales 1

1.2 Introduction to Epitaxial Phenomena 4

1.3 Molecular-Beam Epitaxy 7

1.4 In-Situ Observation of Growth 9

1.4.1 Reflection High-Energy Electron Diffraction . . . 9

1.4.2 Scanning Tunneling Microscopy 11

1.4.3 Atomic Force Microscopy 13

1.5 Epitaxial Growth Modes 14

1.6 Physics in Reduced Dimensions 16

1.6.1 The Coulomb Blockade 18

1.6.2 The Wigner Lattice 20

References 22

2 Quantum Theory of Electrons **25**

2.1 The Many-Body Equation 26

2.2 The Adiabatic Approximation 27

2.2.1 Separation of Variables 28

2.2.2 Adiabatic Wavefunctions 29

2.3 The Ion-Core Schrödinger Equation 31

2.3.1 The Structure of Solids 31

2.3.2 Lattice Vibrations 32

2.4 The Electron Schrödinger Equation 32

2.5 Density Functional Theory 33

2.5.1 The Hartree and Hartree–Fock Approximations . 33

2.5.2 Basic Density Functional Theory 35

2.5.3 The Kohn–Sham Equations 38

2.6 The Local Density Approximation 39

References 41

3 Molecular Dynamics and Kinetic Monte Carlo Simulations 43

3.1 Molecular Dynamics 43

3.2 Kinetic Monte Carlo Simulations 45

References 49

4 Analytic Theories of Morphological Evolution 51

4.1 Theory of Burton, Cabrera and Frank 51

4.2 Homogeneous Rate Equations 53

4.2.1 Irreversible Aggregation Kinetics 54

4.2.2 The Distribution of Island Sizes 57

4.3 Kinetic Roughening 58

4.4 Continuum Equations of Motion 59

4.4.1 Roughening by Random Deposition 60

4.4.2 The Villain Equation 62

References 64

Chapter 1

Introduction and Overview

1.1 Epitaxial Systems across Length and Time Scales

Most phenomena in materials science result from the interplay between processes that are operative over a wide range of length and time scales. For example, the formation of dislocations within a material (atomic scale) and their mobility across grain boundaries of the microstructure ("mesoscopic" scale) affect the deformation behavior of the material (macroscopic scale). A complete understanding of mechanical properties thus requires theoretical and computational tools that range from the atomic-scale detail of first principles density functional methods to the more coarse-grained picture provided by continuum elasticity theory.

At this level of discussion, epitaxial phenomena are no different from any other problem in materials science. Understanding the morphology and properties of epitaxial films requires accommodating the atomic-scale information about the movement of adatoms on surfaces and their various bonding configurations into the macroscopic evolution of the thin film. As Fig. 1.1 indicates this involves quite a large disparity of length and times scales, with quantum and classical molecular dynamics providing a resolution of the order of an atomic vibrational period (10^{-12}–10^{-15} s), while typical time scales for the formation of an atomic layer are of the order of 1 s–1 min. This precludes the direct simulation of epitaxy with these methods, so one of the central problems of describing epitaxial phenomena is finding a way of systematically incorporating the atomistic information provided by first

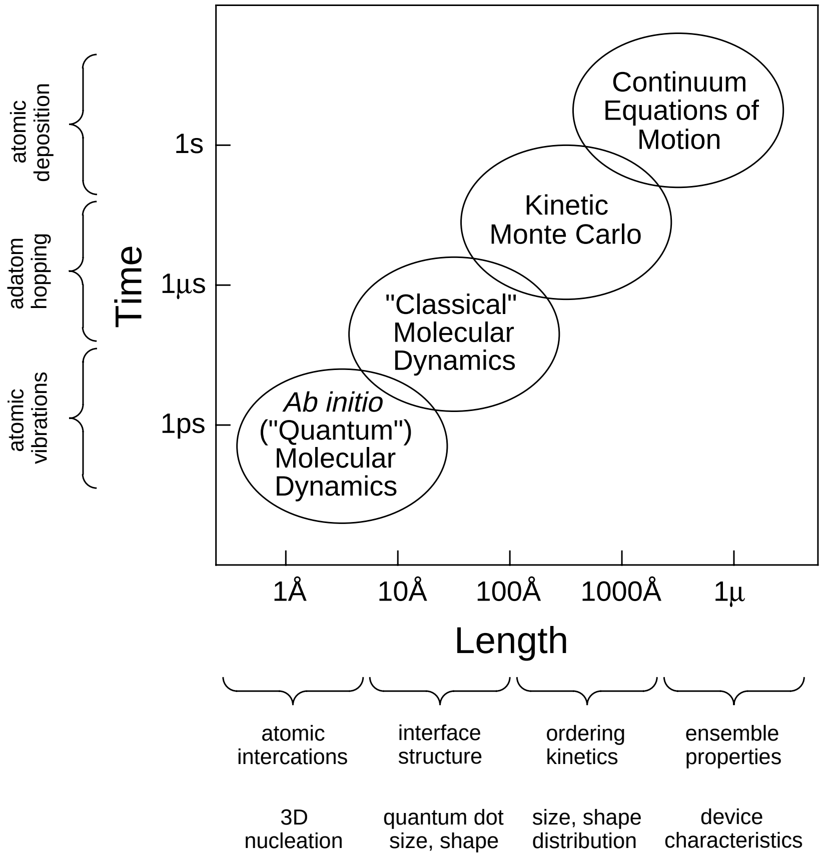

Figure 1.1: Schematic illustration of the types of theoretical methods available for kinetic problems in materials science along with the length and time scales over which these methods provide information. Specific phenomena at the various length and time scales that are required for describing the performance of devices based on quantum dots are shown for comparison.

principles methods into computational schemes that are appropriate for macroscopic, or at least mesoscopic, length and time scales.

In these tutorials we review the phenomenology of epitaxy across the length and time scales shown in Fig. 1.1 and the theoretical and modelling approaches that have been used to explain various experimental results, using quantum dots as a case study. After a brief discussion of the experimental apparatus used for the realization and analysis of epitaxial growth, we survey the theoretical methods that have been applied to studying epitaxial phenomena. Included will first-principles density functional methods, classical molecular dynamics, kinetic Monte Carlo simulations, and continuum equations of motion.

Ab initio techniques provide detailed information regarding specific atomic configurations, and thus are best suited to characterizing barriers and pathways to diffusion and other kinetic processes, and to determining the stability of relatively small collections of atoms. The molecular dynamics method shares with *ab initio* total energy calculations the common feature that the nuclear and electronic coordinates are separated to obtain an effective Hamiltonian for the electronic coordinates. This Hamiltonian can then be used to obtain a potential energy surface for the nuclei as a function of their positions. In total energy calculations, this is used to identify local minima in the total energy to obtain stable structures for a given configuration of atoms.

In the molecular dynamics method, the expression for the total energy of the system as a function of the positions of the atoms is written as an expansion in terms of potentials, and the subsequent motion of the atoms is determined by the forces acting on the atoms. In the molecular dynamics method information concerning energy barriers for particular kinetic processes and the relative likelihood of different events is a natural outcome of choosing a particular potential. Molecular dynamics in principle provides the most accurate way of modelling epitaxial growth and other dynamical processes, but suffers from the interaction potentials not being easily determined and from the fact that the basic time step does not permit especially lengthy simulations on large systems.

In the Monte Carlo method, the rate-determining events must be identified and rate constants must be estimated. The simplicity of the Monte Carlo method means that the details of local interatomic interactions are not explicitly incorporated into the model, but various processes are included on average through effective kinetic parameters. Thus, although this methods cannot be used to address effects that are too specific, comparisons with experiments are easier to make, because the simulations can be run under a greater variety of conditions. Furthermore, the Monte Carlo method provides a framework within which to identify the consequences of particular aspects of model potentials. Monte Carlo simulations alleviate many of the problems due to the discrepancy between simulated and real time scales in molecular dynamics by including explicitly only the rate-limiting steps. The main disadvantage of this method is that each process must be individually identified and included 'by hand.'

Continuum equations of motion have been used in theoretical descriptions since the pioneering work of work Burton, Cabrera and Frank. This approach does not contain the atomic-scale detail of molecular dynamics or Monte Carlo simulations, but does provide a way to making connections with thermodynamic quantities and is useful for examining the coarse-grained behavior of a kinetic system. Proposed continuum equations for various regimes of epitaxial growth will be discussed and, where possible, comparisons between continuum equations and lattice simulations will be made.

1.2 Introduction to Epitaxial Phenomena

Epitaxial growth is a process during which a crystal is formed on an underlying crystalline surface as the result of deposition of new material. The study of this process dates back over one hundred fifty years, but it was not until the work of Louis Royer in the 1920s that the systematics of epitaxial growth began to be revealed (Royer, 1928). Royer carried out an extensive study of the growth of ionic crystals on one another and on mica, mainly from aqueous solution and, using optical microscopy, summarized his observations with a set of rules based on crystal structure. These rules led Royer to coin the term 'epitaxy', which is a combination of the Greek words *epi*, meaning 'upon', and *taxis*, meaning 'arrangement', to convey the notion of growing a new crystal whose orientation is determined by a crystalline substrate and to distinguish epitaxial growth from polycrystalline and amorphous growth.

The development of vacuum technology in the 1960s—an off-shoot of the American space program—opened the way to the deposition of materials on well-characterized substrates in a controlled environment. Epitaxial growth techniques are now used fabricate thin films of essentially all materials types. The motivation for this is twofold. Epitaxial thin films can exhibit properties and structures that have no bulk counterparts. Examples include magnetic properties of metallic structures and electronic, transport, and optical properties of semiconductor structures. Thus, epitaxial films are a fertile arena for the study of fundamental properties in reduced dimensions. However, the overriding reason for the recent rapid expansion of the study of epitaxial phenomena is information technology. For semiconductors, this is a natural result of the drive toward increasing electronic miniaturization that was ushered in by the invention of the integrated circuit and planar fabrication technology. Although epitaxial techniques have not yet had an impact on Moore's law[1], there are several widespread commercial applications of semiconductor epitaxial structures, such as high-electron-mobility transistors, which find application in satellite television receivers and mobile telephones, and lasers, which are used in compact disk players. Magnetic thin film structures are viewed as being central to meeting the expanding needs of long-term data storage, particularly from graphics-intensive applications, and ferroelectric materials are being studied for possible use as non-volatile storage media. Since the production of semiconductor thin film structures has been the dominant application of epitaxial techniques, we begin by reviewing the recent history of their development.

The modern era of the epitaxial growth of semiconductors was founded on a suggestion in the late 1960s by Leo Esaki and Raphael Tsu (1970), then working at the IBM Research Laboratories in Yorktown Heights, New York. They proposed that structures composed of layered regions of semiconductors with different band gaps would have a spatially-varying potential energy surface that would confine carriers to the narrower band-gap material. If there were few enough adjacent layers of this material, then the carriers could be confined within regions comparable to their de Broglie wavelength—the natural length scale that governs their quantum mechanical behavior. For this reason, these narrow regions are now called '*quantum wells.*' Electrons (and holes) in quantum wells were predicted to exhibit remarkable optical and transport properties that could be controlled by varying the width of the wells and the materials forming the heterogeneous interfaces surrounding the well.

At the time that Esaki and Tsu made their proposal, the available technology could not produce materials of sufficient quality to verify the predicted effects. However, the first observation of confinement effects in a quantum well triggered a world-wide effort aimed at improving and extending the basic idea of Esaki and Tsu to other carrier-confining semiconductors, which are collectively referred to as '*quantum heterostructures.*' With many major subsequent developments, epitaxial growth techniques have matured to the point where atomic-scale control over interface quality has become a matter of routine. The control over interface definition and doping profiles has also made planar

[1] Moore's law (Moore, 1965) is the observation made by Gordon Moore, one of the founders of Intel, that the densities of semiconductor components on integrated circuits had and would continue to double on a regular basis (every 18 months).

heterostructures a popular testing ground for many fundamental ideas in condensed matter physics and has led to the discovery of new physical phenomena such as the quantum Hall and fractional quantum Hall effects.

During the late 1970s epitaxial growth techniques began to be applied to metal epitaxy, magnetic metal epitaxy, and eventually, in the mid 1980s, to the preparation of high-quality epitaxial magnetic rare-earth superlattices. The driving force for this was the expectation that, in analogy with the growth of low-dimensional semiconductor heterostructures, epitaxial technology could provide high-quality, epitaxial magnetic metallic structures which might exhibit new magnetic phenomena. This expectation was, in fact, realized by several discoveries in the late 1980s, one of the most surprising of which was giant magnetoresistance (GMR).

Subsequently, Parkin *et al.* (1991), using a system incorporating UHV design features, discovered that magnetron-sputtered polycrystalline multilayers (Fe/Cr, Co/Cr, Co/Ru) exhibited interlayer exchange coupling which oscillated from antiferromagnetic to ferromagnetic as a function of the nonmagnetic spacer thickness. Moreover, the magnetoresistance was oscillatory and its magnitude comparable with that in epitaxial structures (e.g. in Fe/Cr multilayers) prepared by MBE. This discovery had several major implications. It showed that *polycrystalline* magnetic multilayers, prepared by the widespread technique of sputtering, had properties similar to those of *single-crystal* multilayers prepared by epitaxial growth. This also had technological significance, since sputtering is a manufacturing technique used for producing magnetic storage devices. But it also raised questions about the effect of crystalline orientation, interface roughness, and structural quality of the multilayers on interlayer coupling and GMR. This stimulated widespread research in this area, including *in situ* studies of multilayer growth and interface formation.

In this chapter, we will introduce the basic epitaxial growth methods and describe the surface analytical techniques that are used to monitor and characterize the surface of the growing material. We then turn our attention to the types of growth morphologies that can occur when one material is deposited onto another material. Although the categorization that has been developed is based largely on observations using optical and electron microscopy, it remains a useful starting point from which to understand the morphology in light of the atomistic information provided by current methods of observing epitaxial systems, most notably, the scanning tunneling microscope.

Several textbooks (Stringfellow, 1989; Tsao, 1993; Yang *et al.* 1993; Barabási and Stanley, 1995; Markov, 1995; Villain and Pimpinelli, 1998) have appeared in recent years that cover various aspects of epitaxial growth. These should be consulted for more detailed discussions than those provided here.

1.3 Molecular-Beam Epitaxy

The simplest way conceptually of realizing epitaxial growth is with a process known as *molecular-beam epitaxy* (MBE) (Joyce, 1984). This technique has its origins in a series of experiments, based on silicon, carried out by Bruce Joyce and his colleagues in the mid-1960s. Major developments, particularly in the application to III–V compound semiconductors, took place at Bell Laboratories in Murray Hill, New Jersey some three to four years later, inspired by Al Cho and John Arthur. A historical review based on many of the seminal papers has been compiled by Cho (1994).

MBE is essentially a two-step process carried out in an ultra-high vacuum (UHV) environment (Fig. 1.2). In the first step, atoms or simple molecules that are the constituents of the growing material (e.g. atomic Ga and either As_2 or As_4 for GaAs, and atomic Si for Si) are evaporated from solid sources in heated cells, known as *Knudsen cells*, collimated into beams and directed toward a heated substrate which is typically a few centimeters in size. The particles within these beams neither collide with one another nor undergo chemical reactions, i.e. the deposition onto the substrate is *ballistic* and particles are said to undergo *molecular* flow—hence the name *molecular-beam* epitaxy. The substrate is often rotated for more uniform deposition rates across the substrate.

The second step of MBE is the migration of the deposited species on the surface prior to their incorporation into the growing material. The movement of these species across the surface and the resulting surface profile, or morphology, are among the central issues of epitaxial growth and depends on many factors, including deposition rates, the surface temperature, the surface material and its crystallographic orientation, just to name a few. The explicit dependence of the morphology on the deposition rate of new material means that MBE (and other epitaxial growth techniques) are inherently *nonequilibrium*, or *driven*, processes.

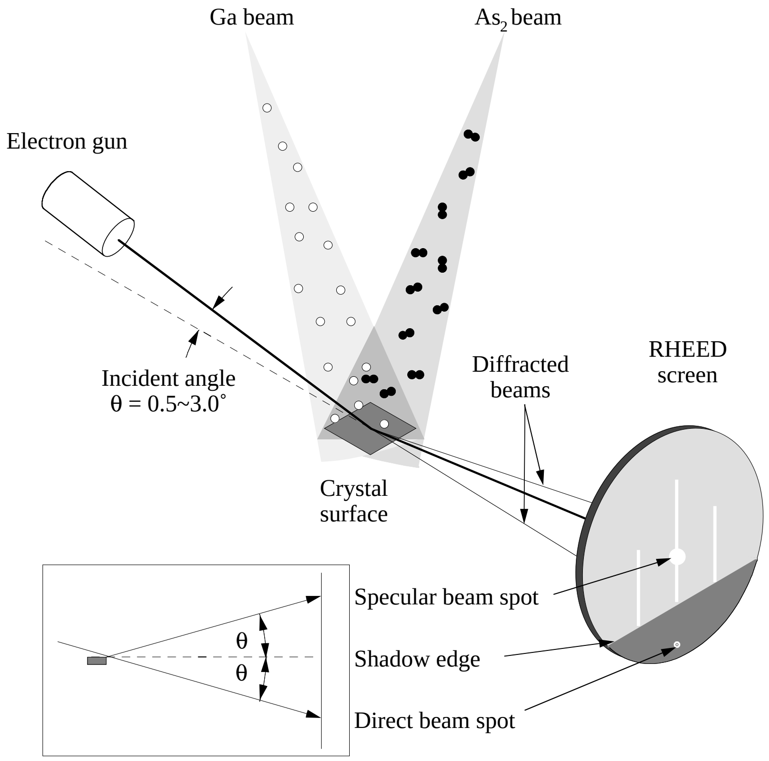

Figure 1.2: The arrangement of the substrate, the RHEED measurement apparatus, and the deposition of material within the UHV environment of an MBE growth chamber (Shitara, 1992).

This provides an important distinction from crystal growth from solution, where the supply of material to the growing crystal takes places by bulk diffusion through the surrounding solution, and is therefore a *near-equilibrium* process. Growth near equilibrium is governed almost exclusively by *thermodynamic* considerations. For epitaxial growth, thermodynamics still provides the overall driving force for the morphological evolution of the surface, but the extent to which equilibrium is attained even locally is the result of *kinetics*, i.e. the rates of processes that determine how a system evolves under a given set of external conditions (Madhukar, 1983).

A major strength of MBE is that the UHV environment enables the application of *in situ* analytical techniques to characterize the evolution of the growing material at various levels of resolution—from microns down to the arrangement of atoms. Particular techniques and the information they provide will be discussed in Section 1.4.

1.4 In-Situ Observation of Growth

An important advantage of performing growth experiments within a UHV environment is the wealth of surface analytic techniques available to examine the growing surface *in situ*. The most prevalent of these are based on diffraction and real-space imaging. Diffraction techniques include reflection high-energy electron diffraction (RHEED), low-energy electron diffraction, helium-atom scattering, and grazing-incidence x-ray diffraction. Real-space imaging techniques include the scanning tunneling microscope (STM), the atomic-force microscope (AFM), low-energy electron microscopy and reflection electron microscopy. Notable advances have also been made with optical techniques, with applications to both MBE and MOVPE, but these have not yet had the widespread impact of other methods. In this section, we will describe the most commonly-used techniques: RHEED, the STM and the AFM.

1.4.1 Reflection High-Energy Electron Diffraction

Surface electron diffraction is a standard method for examining the growth of thin films *in situ* (Larsen and Dobson, 1988) and dates back to the early days of electron diffraction. A RHEED measurement is carried out by directing a high energy (10–20 keV) beam of electrons at a glancing angle ($\simeq 0.5°$–$3°$) toward the surface (Fig. 1.2). The electrons penetrate a few layers into the surface and those that emerge are recorded on a phosphorescent screen. There are three principal reasons why RHEED is so suitable as a diagnostic tool for MBE: (i) it is a relatively simple measurement to set up, requiring only an electron gun and a collector screen, (ii) it is geometrically compatible with the molecular beams emanating from the Knudsen cells and so does not interfere with the growth process, and thus (iii) it can be carried out during growth. The primary disadvantage of RHEED is that the 'images' of the surface are diffraction patterns. These are difficult to interpret quantitatively in real-space terms because the strong interaction between the electrons and the atoms causes the incident electrons to be scattered several times before emerging from the crystal. This 'multiple scattering' means that RHEED diffraction patterns, unlike kinematic diffraction patterns, cannot be 'inverted' by performing a Fourier transform.

The RHEED diffraction pattern provides several types of information about a surface, including the crystallographic symmetry and

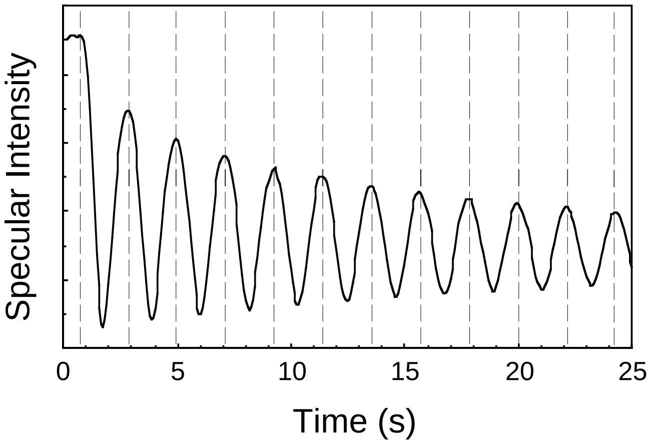

Figure 1.3: Specular RHEED oscillations on GaAs(001) in the temperature range 560–590°. The incident azimuth of the electron beam is [010], the incident polar angle is 1°, and the beam energy is 14 keV (Shitara, 1992).

the extent of long-range order. But the most common application of RHEED is based on measuring the intensity of the specular beam (equal incident and reflected angles). A typical example taken during growth on GaAs(001) is shown in Fig. 1.3. Most apparent in this trace are the oscillations. These oscillations, which are due to the repeated formation of bi-atomic Ga-As layers, provided the first direct evidence of layer-by-layer epitaxial growth in this system (Neave *et al.* 1983; Van Hove *et al.* 1983). The period of the oscillations indicates that the time required to form a complete bi-layer is of the order of seconds. Since the molecular beams can be turned on and off mechanically with a shutter, the amount of material deposited can be controlled to within a fraction of a layer. Thus, a prescribed amount of one material (e.g. GaAs) can then be deposited onto a flat surface, followed by a prescribed amount of a second material (e.g. AlAs). This process can be repeated to form a *superlattice*. The electronic properties of superlattices and other quantum heterostructures can be controlled by the amount and type of materials deposited. For semiconductor heterostructures, these characteristics determine the lateral size of the quantum well and the depth of the confining potential.

Another feature to notice about the oscillations in Fig. 1.4 is the decaying envelope. The reason for this decay will be discussed in detail later in this course, so for the moment we simply mention that this envelope is due to the layer-by-layer growth being imperfect, i.e. subsequent layers begin to form before the preceding layers are complete.

1.4.2 Scanning Tunneling Microscopy

The scanning tunneling microscope, invented in 1982 by Gerd Binnig and Heinrich Rohrer (Binnig and Rohrer, 1985) at the IBM Research Laboratories in Zürich, Switzerland, uses an atomically sharp tip placed sufficiently close (a few Ångstroms) to a surface to produce an electron tunneling current. By measuring this current as a function of position, images are obtained which reflect the electronic density near the surface. Under favorable circumstances, these images have a lateral resolution of ≈1 Å and a vertical resolution of ≈0.1 Å.

The basic principle of the STM can be understood with the model introduced by Tersoff and Hamann (1983) some years ago. The tip is represented by a spherical potential well within which the Schrödinger equation is solved. By retaining only the spherically-symmetric solutions, a simple expression is obtained for the tunneling current I at low bias voltage V: $I{\sim}eV\varrho(r_0, E_f)$, where $\varrho(r_0, E_f)$ is the local density of states at the Fermi energy, E_f, of the scanned surface at the position r_0 of the tip. Thus, scans taken at constant current measure contours of constant Fermi-level charge density of the sample. Although this expression ignores the properties of the tip, which modifies the tunneling current in several ways, it does show that the STM is sensitive to *charge densities*, rather than simply *atomic positions*.

The STM revolutionized the field of surface science and has seen applications that extend far beyond traditional boundaries of condensed matter physics. Its impact on fundamental studies of epitaxial growth has also been immediate and far-reaching, but the inherently kinetic nature of growth does introduce some technical complications that are absent in studies of static surface structure. If an STM is placed in a growth chamber, the tip shadows the incoming molecular beam. Thus, the application of the STM to image growing surfaces has had to rely on one of two indirect strategies. The most common is to image a surface that has been quenched after a prescribed period of growth, thereby providing a ‘snapshot’ of the surface. But recently it has become possible to arrange scan and growth rates to image the *same* region of a surface *during* growth (Voightländer and Zinner, 1993; Pearson *et al.* 1996). Though technically more demanding, this approach is the more desirable in principle because specific kinetic events can be tracked and no quenching is required, thus providing a more faithful record of

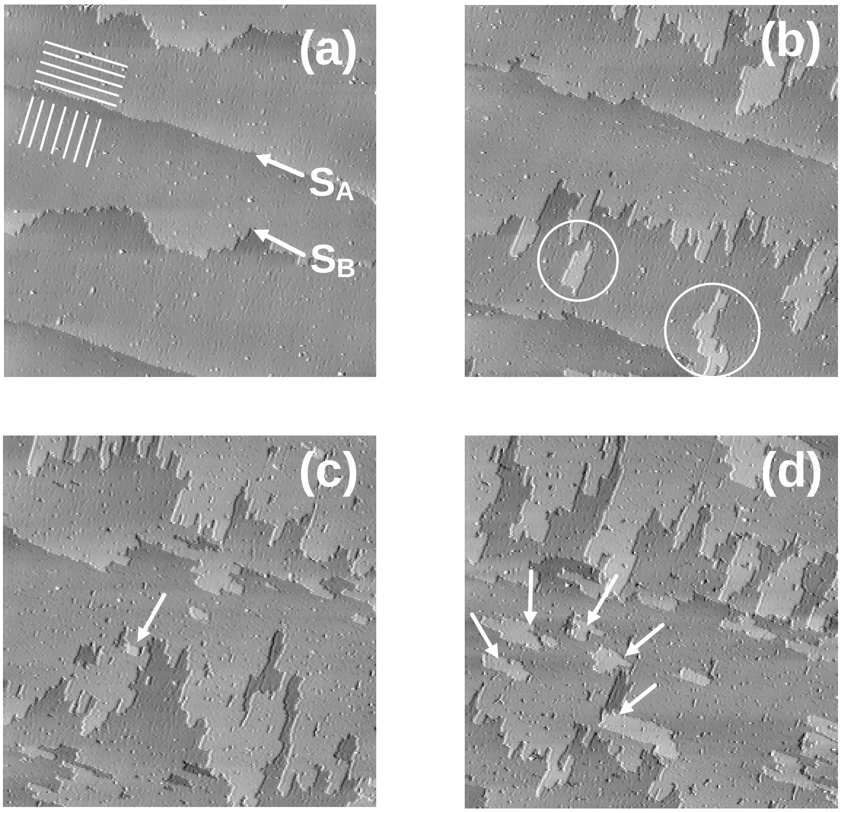

Figure 1.4: STM images (2900 Å × 2900 Å) during growth on vicinal Si(001) at 725 K at 8 ML/h (Voigtländer *et al.* 1997). The starting surface is shown in (a) and the same region of the surface after the deposition of 0.22, 0.53, and 0.94 ML is shown in images (b)–(d), respectively. The straight S_A step and rough S_B step are shown (a), with the white lines indicating the directions along which the dimer rows run along the two types of terrace. Islands formed during deposition are enclosed within the circles in (b) and islands of the next layer after the islands in (b) have coalesced with the advancing step are indicated by arrows in (c) and (d). (Courtesy B. Voigtländer)

surface evolution. However, because of the very slow growth rates required in current implementations of this '*in vivo*' method, the growing surface is exposed for relatively long times to the ambient impurities which are always present in the growth chamber. This can affect the growth in several ways, so care must be taken when interpreting these images to insure that they reflect the intrinsic growth characteristics of the material.

STM images of the (001) surface of Si are shown in Fig. 1.4. These images reveal an important feature that is typical of semiconductor surfaces (and surfaces of many other materials). Because the local coordination of surface atoms is lower than that in the bulk, there are broken, or dangling, bonds which leave the surface in a high-energy state. The formation of new bonds to lower the surface free energy results in a rearrangement of surface atoms. We will distinguish between two types of such rearrangement: *relaxation* and *reconstruction.* A relaxation preserves the symmetry and periodicity of the bulk unit cell. Expressed in units of the 2D primitive lattice vectors, such a structure is said to be 1×1. This is the typical case for non-polar semiconductor surfaces. A reconstruction involves more complex atomic distortions that modify the size and symmetry of the unit cell, leading generically to an $n \times m$ structure. In the images in Fig. 1.4, adjacent atoms on the surface of Si(001) form *dimers*, which produces a doubling of the unit cell along the axis of these dimers, i.e. a 2×1 reconstruction.

1.4.3 Atomic Force Microscopy

When an STM is brought close to a surface, the atoms near the apex of the tip exert a force on that surface which is of the same order of magnitude as the interatomic forces within the surfaces. This effect is the principle behind the atomic force microscope (Binnig *et al.* 1986). An STM tip, mounted on a flexible beam, is brought just above a surface. The force between the surface and the tip causes a small deflection of the beam. The surface is then scanned while maintaining a constant force between the tip and the surface with a feedback loop similar to that used in the operation of an STM.

The AFM complements the STM in several ways. Because the STM relies on a tunneling current for its operation, it is sensitive mainly to the density of electronic states near the Fermi level of the sample, as discussed in the preceding section. Thus, this density of states must be non-zero, i.e. the sample being scanned must be conducting. However, since the AFM tip responds to interatomic forces, which include contributions from *all* electrons, the sample need not be a conductor. Additionally, since the tunneling current decreases exponentially with the tip-sample distance, the STM tip must be placed a few Ångstroms from the surface to maximize the resolution of the image. The AFM most commonly operates in this mode (the ‘contact’ mode) as well, but it can also operate at much larger distances from the surface (50–150 Å) for samples susceptible to damage or alteration by being in close proximity to the tip (the ‘non-contact’ mode). Although achieving lat-

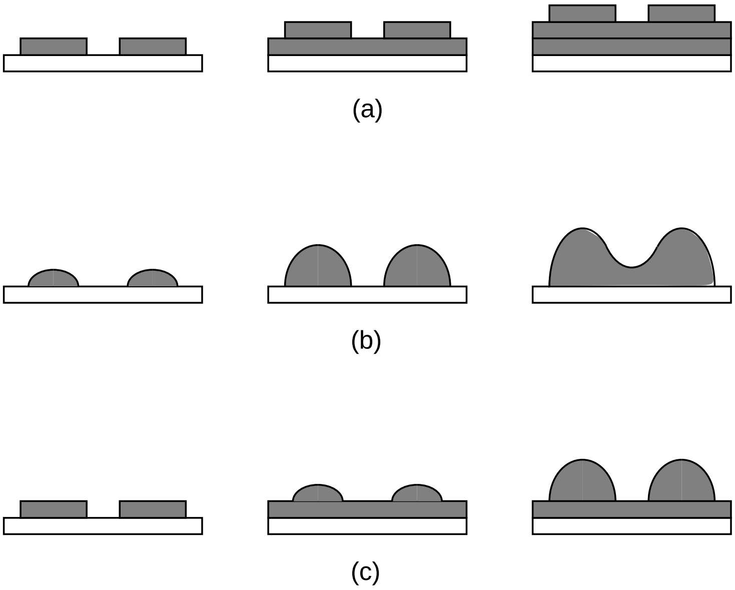

Figure 1.5: Schematic evolution of the (a) Frank–van der Merwe, (b) Volmer–Weber, and (c) Stranski–Krastanov heteroepitaxial growth modes.

eral atomic resolution with the STM is now commonplace, it is much more demanding technically with the AFM. Thus, many applications of the AFM involve scanning large areas to image the gross features of the morphology of the sample. This has the advantage of not requiring a UHV environment and AFMs often operate in ambient atmosphere or in a liquid (Quate, 1994).

1.5 Epitaxial Growth Modes

Numerous experiments (Kern *et al.* 1979, Venables *et al.* 1984) have revealed that, for small amounts of deposited material, the epitaxial growth morphology is one of three distinct types. By convention (Bauer 1958, Le Lay and Kern 1978), these are referred to as: *Frank–van der Merwe* morphology, with flat single crystal films consisting of successive complete layers, *Volmer–Weber* morphology, with three-dimensional (3D) islands that leave part of the substrate exposed, and *Stranski–Krastanov* morphology, with 3D islands atop a thin flat film that completely covers the substrate. These morphologies are illustrated schematically in Fig. 1.5.

For lattice-matched systems, the Frank–van der Merwe and Volmer–

Weber morphologies can be understood from thermodynamic wetting arguments based on the interfacial free energies (Bauer 1958). We denote the free energy of the epilayer/vacuum interface by γ_e, that of the epilayer/substrate interface by γ_i, and that of the substrate/vacuum interface by γ_s. Then Frank–van der Merwe growth is favored if the free energies of the epilayer and the interface between the epilayer and the substrate is less than that of the substrate:

$$\gamma_e + \gamma_i < \gamma_s \tag{1.1}$$

In this case, as the epilayers are formed, the free energy *decreases* initially before attaining a steady-state value for thicker films. Alternatively, if

$$\gamma_e + \gamma_i > \gamma_s \tag{1.2}$$

then Volmer–Weber growth is favored. In this case the free energy *increases* if epilayers are formed on the substrate, rendering a uniform layer thermodynamically unstable against a break-up into regions where the substrate is covered and those where it is uncovered.

Stranski–Krastanov growth may be viewed as the transition from the Frank–van der Merwe to the Volmer–Weber growth mode. This growth mode is not well understood but is thought to be closely related to the accommodation of elastic energy associated with lattice misfit between the epilayer and the substrate. Growth in the first monolayer or so initially proceeds in a layer-by-layer manner, but the epilayer is strained to match the lattice constant of the substrate. As the epilayer thickens the strain energy increases and reaches a point where it can be lowered through the formation of isolated 3D islands in which strain is relaxed by misfit dislocations. But there is another scenario within the Stranski–Krastanov morphology: the formation of islands *without* dislocations—called *coherent* islands—atop one or more wetting layers (Eaglesham and Cerullo, 1990; Madhukar and Rajkumar, 1990). Such islands have been observed for a number of systems (Petroff and DenBaars, 1994; Seifert *et al.* 1996); in Fig.1.6 we show an example of such an island of InP on GaInP(001).

Figure 1.7 shows a sequence of STM images taken during the formation and evolution of 3D InAs islands on GaAs(001) at 420°C. The growth of InAs on GaAs(001) proceeds first by the nucleation of 2D islands which coalesce into coherently strained layers. These are the 'wetting' layers in the conventional Stranski–Krastanov description. The 3D islands are first observed just after 1.7 monolayers. The transition

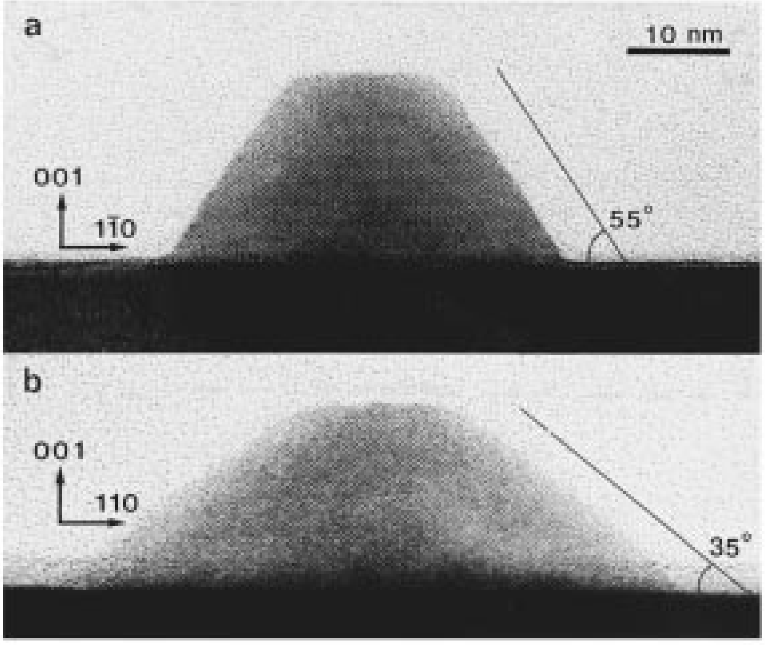

Figure 1.6: High-resolution cross-section micrograph of an uncapped InP island on GaInP grown by MOVPE at 580° C along the (a) [110] and (b) [$\bar{1}$10] directions (Georgsson *et al.* 1995). Note that the islands are elongated along [110] and that the planes of atoms are appreciably curved toward the center of the island near the substrate caused by the compressive strain, but there is no evidence of any dislocations.

to growth by 3D islands is quite abrupt, occurring over less than 0.1 monolayers. This transition can be followed by RHEED, which shows a change from a streaky pattern, characteristic of layer-by-layer growth, to a spotty pattern that corresponds to the transmission of the electrons through the 3D islands. As more material is deposited, the islands grow in number, but soon reach a saturation density.

1.6 Physics in Reduced Dimensions

The properties of artificially structured materials provide new opportunities for technological applications. Many of these applications are driven by the expanding, and seemingly insatiable, requirements of information processing, information transfer (communication), and in-

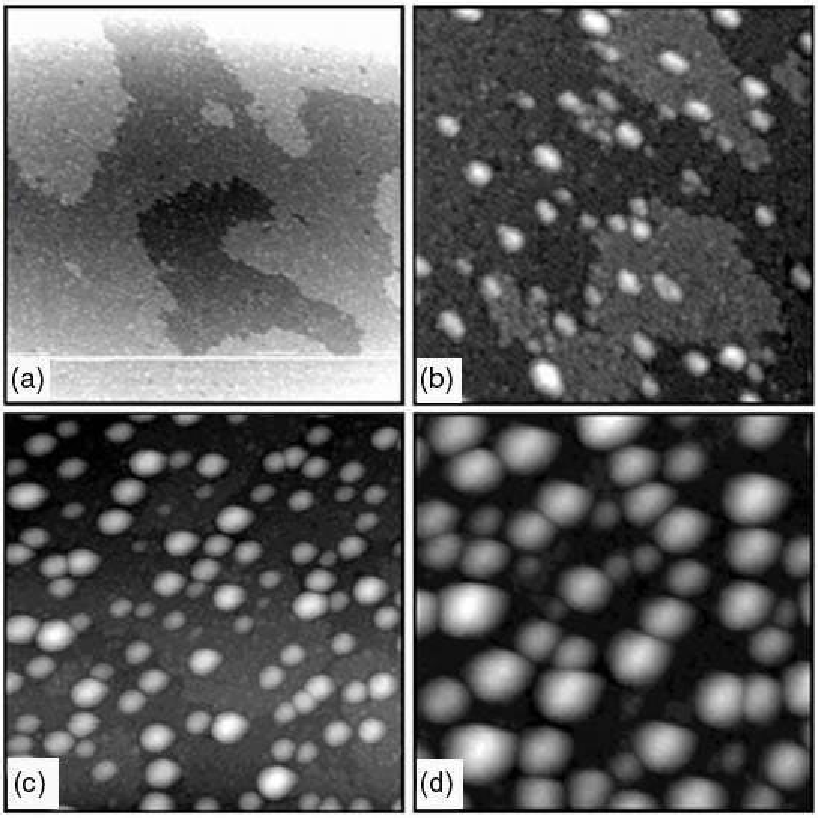

Figure 1.7: Filled states STM images (2000 Å × 2000 Å) of InAs deposited on GaAs(001)–c(4×4) at 420°C at coverages of (a) 1.7, (b) 2.0, (c) 2.5 and (d) 5.5 monolayers. (Courtesy G. R. Bell)

formation storage. For semiconductors, the biggest impact of quantum heterostructures has been in the area of *optoelectronics*, i.e. the generation, control, and detection of light. The advent of quantum wells, in particular, has revolutionized this field and has already seen several devices put into commercial production. There is also considerable ongoing research into extending the practical utility of low-dimensional structures to quantum wires and especially quantum dots, where the optical properties are predicted to be even further enhanced over those of quantum wells. However, these lower-dimensional structures must await further developments in processing before they can aspire to having the impact of quantum wells.

One of the most important factors for the physics of low-dimensional structures is that the small sizes of structures which can now be produced means that the motion of electrons (and holes) is severely restricted, or confined, in at least one direction. This confinement reduces the effective dimensionality of the carriers, which is manifested in remarkable optical, transport and magnetic properties of the materials forming the heterostructure. Some of these properties, which can be traced to changes in the density of single-particle states, are straightforward to understand, being simply the result of the geometrical effects of reduced dimensionality. Other phenomena, however, such as the Coulomb blockade and the Wigner crystal, require that interactions between electrons be taken into account to obtain a complete explanation of the observed behavior. Many issues related of these effects have not yet been fully resolved and remain under active investigation.

Quantum effects arise in systems which confine electrons to regions comparable to their de Broglie wavelength. When such confinement occurs in one dimension only (say, by a restriction on the motion of the electron in the z direction), with free motion in the x and y directions, a 'two-dimensional electron gas' (2DEG) is created. Confinement in two directions (y and z, say), with free motion in the x-direction, gives a 'one-dimensional electron gas' (1DEG) and confinement of its x, y, and z motions at once gives a 'zero-dimensional electron gas' (0DEG). The density of electronic states is a strong function of the spatial dimension. This has a strong influence on the transitions between different energy states, an effect which can be exploited in a number of ways, most notably in optical and transport properties in quantum heterostructures.

1.6.1 The Coulomb Blockade

Consider the classical description of what happens when one tries to charge an isolated conductor with a single electron. The increase in the energy by adding the electron is just the charging energy, $e^2/2C$, where C is the capacitance of the body being charged. The capacitance of a macroscopic conductor is large enough so that this energy penalty is negligible compared with the thermal energy at room temperature, $kT \approx 1/40$ eV, so for this situation there is no measurable barrier for this process. However, for very small conductors at low temperature it is possible for the charging energy to *exceed* the thermal energy. As a result it is energetically unfavorable for an electron to charge the conductor until the external driving force is sufficient to supply the extra energy. This is the regime of the *Coulomb blockade*, where no current can flow through the conductor.

Many systems are capable of displaying a Coulomb blockade at low enough temperatures. Some of the earliest observations of this phenomenon date back to the 1960s where zero bias anomalies in the current flowing through a large array of small tin particles were explained in terms of the charging energy of the particles (Zellar and Giaever, 1969). In these original experiments the current flowed through a large number of tin islands and only the average or dominant properties could be observed. Nanofabrication technology has now developed to the stage where the Coulomb blockade can be observed in a variety of settings. It is of special interest in low-dimensional semiconductor systems because of the fact that the discreteness energies within heterostructures makes the on/off nature of the conductance very precise. This forms the basis of the *single-electron transistor* (Kastner, 1992), which will be discussed in the next chapter.

There are three ingredients that conspire to form a Coulomb blockade: the quantization of the electronic charge, the small size of the structures (and, therefore, the low electron densities), and low temperatures. Consider what happens when one tries to send a current along a one-dimensional quantum wire containing a quantum dot at a very low temperature ($\simeq$100 mK). We can regard this situation as corresponding to the dots connected by leads (the quantum wire). The conductance is a measure of how easily current can flow through the dot. But adding one electron to the charges already in the dot takes energy; how much is determined from elementary considerations, since the quantum dot is essentially a capacitor. To add an amount of charge Q to a capacitor whose capacitance is C requires an energy $E = Q^2/2C$. Thus, to put one more electron into the dot costs an energy $e^2/2C$. Similarly for a hole to tunnel into the 1DEG (i.e. for the electron to leave the dot) takes energy $-e^2/2C$. This means that electrons at the Fermi energy of the wire can get into the dot only if this energy is $e^2/2C$ higher than the lowest available electron state in the 1DEG and, once it is there, can only get out again if it can lose at least $e^2/2C$ on the other side. This leads to a gap of e^2/C in the tunneling density of states.

If thermal fluctuations are not to mask the charging energy of the Coulomb blockade the temperature has to be low enough to ensure that the inequality $kT \ll e^2/2C$ is satisfied. This condition represents the greatest challenge to the manufacture of single-electron devices which would operate at room temperature. At $T = 300$K the thermal energy is 25.8meV, corresponding to a total capacitance $C \approx 3$ aF ($1a = 10^{-18}$). For robust operation the capacitance should be 10–

100 times smaller than this value, leading to total device capacitances of the order of 10^{-20} F. At present most single-electron structures have values of $C > 1$ aF and only operate at cryogenic temperatures. Room temperature operation will require devices in which the charging islands are less than 100Å in size. Although this represents a significant challenge to microfabrication techniques, recent progress has been quite encouraging.

In addition to a small total capacitance, the charging region of a single electron device must be connected to the outside world via leads whose resistance exceeds approximately 26 kΩ; otherwise, the Coulomb blockade will be masked by quantum fluctuations (Averin and Likharev, 1992). To see why, consider the Heisenberg uncertainty relation in the form $\Delta E \Delta t \simeq h$. Quantum fluctuations will destroy the Coulomb blockade if the uncertainty in the energy ΔE exceeds the charging energy. To ensure this is not the case an electron must stay on the charging region for a time $\Delta t > hC/e^2$. We can equate the charge/discharge time τ of the region to its RC time constant, i.e. $\Delta t \approx \tau \approx RC$, where R is the total resistance through which the island is charged. This leads to the result that the resistance of any junction in the single electron device must be greater than $R_{\text{min}} = h/e^2 = 25.8$ kΩ.

1.6.2 The Wigner Lattice

A gas of electrons behaves very differently from a gas composed of neutral weakly-interacting particles. One of the most striking differences is the behavior of these two types of gases as a function of the density. At large densities, interactions between the particles in atomic and polyatomic gases become increasingly important. But for an electron gas, the phenomenon of *screening* leads to behavior that for many purposes may be regarded as that of free electrons. Thus, a high-density electron gas behaves essentially like an ideal gas of Fermions.As the density of an atomic or polyatomic gas is lowered, the interactions diminish in importance and the gas approaches ideal behavior. For an electron gas, however, decreasing the density *increases* the effect of the Coulomb potential because the screening effect becomes much less effective. This is the physical basis of the Coulomb blockade discussed in the preceding section.

It was precisely such observations that led Eugene Wigner (Wigner, 1934) to propose the existence of a lattice of electrons as the ground state of an interacting gas—what is now called a *Wigner crystal.* Wigner argued that below a certain critical density the kinetic energy will be negligible in comparison to the potential energy. Thus, at low enough temperatures the energy of a system of electrons would be dominated by the pair-wise Coulomb potential between the particles and the behavior of the gas will be determined by the configuration that minimizes the potential energy. Since the potential of a random array is higher than that of an ordered array, then in this regime the electrons will form a crystal. In three dimensions, the case that Wigner considered, the lowest potential energy is obtained for a body-centered cubic crystal.

There are two regimes to consider: the *quantum* regime, where $k_BT \ll E_\mathrm{F}$, and the *classical* regime, where $k_BT \gg E_\mathrm{F}$. The classical regime of Wigner crystallization is relatively easy to achieve when the density n_s of electrons is small, since $E_\mathrm{F} \propto n_s$. The potential energy V per electron can then be estimated by

$$V \simeq \frac{e^2}{4\pi\varepsilon_0 r} \propto n_s^{1/2} \tag{1.3}$$

The average kinetic energy can be obtained from the equipartition theorem, so the crossover temperature where the kinetic and potential energies are of comparable magnitude is $T \propto n_s^{1/2}$. The first observation of a Wigner crystal was, in fact, in the classical regime for electrons on the surface of liquid helium (Grimes and Adams, 1979).

The higher densities n_s (and lower effective masses) of 2DEGs in semiconductors means that $E_\mathrm{F} \gg k_BT$. In this (quantum) regime, the kinetic energy of the electrons remains nonzero even at the lowest temperatures, being of order E_F. Thus, the kinetic and potential energies are comparable, so electrons in most semiconductors remain in a 'liquid' state even at the lowest temperatures. Achieving lower densities is not yet technically feasible, so an alternative approach is to apply a large ($\sim$ 10 T) magnetic field perpendicular to the 2DEG which has the effect of confining electrons to small ($\sim$ 5 nm) orbits. This makes the 2DEG easier to solidify and there have been a number of experiments carried out that support the notion 2DEGs in GaAs crystallize in very high magnetic fields and low temperatures (Goldman *et al.* 1990).

References

D.V. Averin and K.K. Likharev, in *Single Charge Tunneling*, H. Grabert and M.H. Devoret, eds., (Plenum, New York, 1992).

A.–L. Barabási and H. E. Stanley, *Fractal Concepts in Surface Growth* (Cambridge University Press, Cambridge, 1995).

E. Bauer, *Z. Krist.* **110**, 372 (1958).

G. Binnig and H. Rohrer, *Sci. Am.* **253**, 50 (1985).

G. Binnig, C.F. Quate and Ch. Gerber, *Phys. Rev. Lett.* **56**, 930 (1986).

D.J. Eaglesham and M. Cerullo, *Phys. Rev. Lett.* **64**, 1943 (1990).

A. Cho, ed., *Molecular Beam Epitaxy* (American Institute of Physics, New York, 1994).

L. Esaki and R. Tsu, *IBM J. Res. Develop.* **14**, 61 (1970).

K. Georgsson, N. Carlsson, L. Samuelson, W. Seifert, L.R. Wallenberg, *Appl. Phys. Lett.* **67**, 2981 (1995).

V.J. Goldmann, M. Santos, M. Shayegan, and J.E. Cunningham, *Phys. Rev. Lett.* **65**, 2189 (1990).

C.C. Grimes and G. Adams, *Phys. Rev. Lett.* **42**, 795 (1979).

B.A. Joyce, *Rep. Prog. Phys.* **37**, 363 (1984).

M.A. Kastner, *Rev. Mod. Phys.* **64**, 849 (1992).

R. Kern, G. Le Lay and J.J. Metois, in *Current Topics in Materials Science*, E. Kaldis, ed. (North–Holland, Amsterdam, 1979), Vol. 3.

P.K. Larsen and P.J. Dobson, eds., *Reflection High-Energy Electron Diffraction and Reflection Electron Imaging of Surfaces* (Plenum, New York, 1988).

G. Le Lay and R. Kern, *J. Cryst. Growth* **44**, 197 (1978).

A. Madhukar, *Surf. Sci.* **132**, 344 (1983).

A. Madhukar and K.C. Rajkumar, *Appl. Phys. Lett.* **57**, 2110 (1990).

I.V. Markov, *Crystal Growth for Beginners: Fundamentals of Nucleation, Crystal Growth and Epitaxy* (World Scientific, Singapore, 1995).

G.E. Moore, *Electronics* **38**(8), 114 (1965).

J.H. Neave, B.A. Joyce, P.J. Dobson, and N. Norton, *Appl. Phys. A* **31**, 1 (1983).

S.S.P. Parkin, *Phys. Rev. Lett.* **71**, 1641 (1993).

C. Pearson, M. Krueger, and E. Ganz, *Phys. Rev. Lett.* **76**, 2306 (1996).

P.M. Petroff and S.P. DenBaars, *Superlat. and Microst.* **15**, 15 (1994).

C.F. Quate, *Surf. Sci.* **299/300**, 980 (1994).

L. Royer, *Bull. Soc. Fr. Mineralog. Crystallogr.* **51**, 7 (1928).

W. Seifert, N. Carlsson, M. Miller, M.–E. Pistol, L. Samuelson and L. R. Wallenberg, *Prog. Crystal Growth and Charact.* **33**, 423 (1996).

T. Shitara, *Growth Mechanisms of GaAs(001) during Molecular Beam Epitaxy*, Ph.D. thesis (University of London, 1992).

G.B. Stringfellow, *Organometallic Vapor–Phase Epitaxy* (Academic, Boston, 1989).

J. Tersoff and D.R. Hamann, *Phys. Rev. Lett.* **50**, 1998 (1983).

J.Y. Tsao, *Materials Fundamentals of Molecular Beam Epitaxy* (Academic, Boston, 1993).

J.M. Van Hove, C.S. Lent, P.R. Pukite and P.I. Cohen, *J. Vac. Sci. Technol. B* **1**, 741 (1983).

J.A. Venables, G.D.T. Spiller and M. Hanbucken, *Rep. Prog. Phys.* **47**, 399 (1984).

J. Villain and A. Pimpinelli, *Physics of Crystal Growth* (Cambridge University Press, Cambridge, 1998).

B. Voightländer and A. Zinner, *Appl. Phys. Lett.* **63**, 3055 (1993).

Y. Wang and R.J. Hamers, *Chem. Rev.* **96**, 1261 (1996).

E. Wigner, Phys. Rev. **46**, 1004 (1934).

H.–N. Yang, G.–C. Wang and T.–M. Lu, *Diffraction from Rough Surfaces and Dynamic Growth Fronts* (World Scientific, Singapore, 1993).

G. Zellar and I. Giaever, *Phys. Rev.* **181**, 798 (1969).

Chapter 2

Quantum Theory of Electrons

The observable properties of all forms of matter, and solids in particular, are determined completely by quantum mechanics from solutions of a many-body Schrödinger equation for the motion of the electrons and the nuclei. This includes all equilibrium properties and nonequilibrium, or response, properties. Equilibrium properties encompass all thermodynamic behavior, including the equations of state and phase diagrams, and quantities such as the specific heat and compressibility. Nonequilibrium properties include responses to various perturbations, such as electromagnetic fields and mechanical impulses, which determine the optical, transport, and mechanical properties of materials. However, because of the inherent difficulty of obtaining even grossly approximate solutions of the full many-body Schrödinger equation, one typically focusses on (sometimes *ad hoc*) approximations to this equation which are believed to capture the essential energetics of the problem of interest. This has resulted in a number of parallel strands in Condensed Matter Theory and is, in part, responsible for the richness of the subject as a whole.

In this Chapter, we carry out a systematic reduction of the full quantum mechanical description of solids to obtain a more conceptually and computationally manageable set of equations which can be applied to specific materials. What emerges are separate equations for the degrees of freedom of the electrons and nuclei and the conditions under which this partitioning is appropriate. We then discuss briefly the basic information that is obtained by solving these equations, but our main focus, to be covered in later lectures, is the solution of the Schrödinger equation for electrons.

2.1 The Many-Body Equation

The logical formulation of the quantum theory of solids begins with the exact problem which is then made tractable by making several systematic approximations. We are seeking solutions of the Schrödinger equation:

$$\mathcal{H}\Psi(\{\boldsymbol{r}_i\}, \{\boldsymbol{R}_\alpha\}, t) = i\hbar\frac{\partial\Psi}{\partial t} \tag{2.1}$$

where $\mathcal{H}$ is the exact many-body Hamiltonian and the wavefunction Ψ is a function of the all of the electronic and nuclear coordinates, which we denote by $\boldsymbol{r}_i$ and $\boldsymbol{R}_\alpha$, respectively. A solid typically contains of the order of 10^{25} electrons which are mutually interacting and moving in the electromagnetic fields of $\sim 10^{24}$ positively-charged ion cores, which are also mutually interacting. The solid as a whole is, of course, electrically neutral. Under ordinary circumstances, neither the electrons nor ion cores move at velocities anywhere near the speed of light ($\boldsymbol{v}_i$, $\boldsymbol{V}_\alpha \ll c$) so, as a first approximation, we can take the Hamiltonian to be the sum of the nonrelativistic kinetic energies and Coulomb interactions of the electrons and ion cores:

$$\begin{aligned}\mathcal{H} = \sum_i \frac{\boldsymbol{p}_i^2}{2m} + \sum_\alpha \frac{\boldsymbol{P}_\alpha^2}{2M_\alpha} + \tfrac{1}{2}{\sum_{i,j}}' \frac{e^2}{|\boldsymbol{r}_i - \boldsymbol{r}_j|} \\ + \tfrac{1}{2}{\sum_{\alpha,\beta}}' \frac{Z_\alpha Z_\beta e^2}{|\boldsymbol{R}_\alpha - \boldsymbol{R}_\beta|} - \sum_{i,\alpha} \frac{Z_\alpha e^2}{|\boldsymbol{r}_i - \boldsymbol{R}_\alpha|}\end{aligned} \tag{2.2}$$

where $\boldsymbol{p}_i$ and $\boldsymbol{P}_\alpha$ are the momenta of the electrons and ion cores, respectively, m is the electron mass, M_α is the mass of the ion core at position $\boldsymbol{R}_\alpha$, and Z_α is the charge of that ion core. The sums over i and j run over all of the electrons and the sums over α and β run over all of the ion cores. The terms on the left-hand side of this equation represent, respectively, the kinetic energies of the electrons and ion cores, the (repulsive) Coulomb potential energy between the electrons and the corresponding term for the ion cores, and the (attractive) Coulomb potential energy between the electrons and the ion cores. This Hamiltonian should be looked upon as a zeroth-order expression; it omits spin-orbit coupling, magnetic effects, and mass-velocity effects, all resulting from relativistic corrections, and, if we do not consider *all* of the electrons explicitly, core polarization effects. All of these corrections can be handled within perturbation theory, provided that the zero-order wavefunction can be obtained.

The first problem we face, already referred to, is how to break up the electrons within each atom into *outer*, or *valence*, electrons (which are given the explicit coordinates $\boldsymbol{r}_i$ in the formulae above), and *core* electrons, which are part of the ion core and are assumed to move with the nucleus at all times. The fewer electrons that we consider explicitly, the more we have to correct the Hamiltonian for polarization effects by introducing many-ion interaction terms. The decision depends on two considerations: (i) the nature of the atoms in the solid, and (ii) the type of solid (ionic, covalent, metallic, molecular, etc.). In general, for covalent and metallic solids, we begin by considering the material as a collection of *atoms*. We can arbitrarily specify that all electrons on these atoms whose binding energy is greater than, say, 30 eV will be taken as part of the ion core and *not* treated explicitly. Operationally, this means that all electrons outside of filled electronic shells are considered explicitly. For molecular solids, we can apply a similar guide to the binding energies of electrons to the *molecule*; for ionic solids, we begin with the appropriate ions (e.g., Na^+ and Cl^-).

How do we know *a priori* whether a given solid is metallic, ionic, covalent or something else? In principle, we should be able to solve the problem and find out. But the magnitude of doing so is revealed when one considers that (2.2) is the Hamiltonian not only for all possible allotropic forms of the solid under consideration, but also applicable to all other phases, including the liquid, gas, and even plasma phases. All of these should emerge from a complete solution to the problem. At the moment, even solving this problem at zero temperature is a hopeless task, so we are forced to take a much more empirical approach.

2.2 The Adiabatic Approximation

We now proceed to attempt to solve the Schrödinger equation (2.1) with the Hamiltonian (2.2) in as systematic a manner as possible. The first approximation that we introduce can be analyzed in terms of a perturbation expansion. The small parameter is the ratio of the electron mass to the mass of the ion core, m/M_α. This ratio is always less than 5×10^{-4} and ordinarily is less than 10^{-5} (for atoms heavier than calcium).

2.2.1 Separation of Variables

We begin by following a procedure motivated by the separation of variables method used to solve linear partial differential equations. We first write the wavefunction $\Psi(\{\boldsymbol{r}_i\}, \{\boldsymbol{R}_\alpha\}, t)$ as

$$\Psi(\{\boldsymbol{r}_i\}, \{\boldsymbol{R}_\alpha\}, t) = \varphi(\{\boldsymbol{r}_i\}; \{\boldsymbol{R}_\alpha\})\psi(\{\boldsymbol{R}_\alpha\})\, \mathrm{e}^{-iEt/\hbar}$$

where the quantities $\boldsymbol{R}_\alpha$ in φ are to be regarded as a set of *parameters*, in a sense to be made clear below. Substituting this expression into (2.1) and (2.2) yields Schrödinger equations for φ and ψ. The equation for φ is

$$\left\{\sum_i \frac{\boldsymbol{p}_i^2}{2m} + \tfrac{1}{2}\sum_{i,j}{}' \frac{e^2}{|\boldsymbol{r}_i - \boldsymbol{r}_j|} - \sum_{i,\alpha} \frac{Z_\alpha e^2}{|\boldsymbol{r}_i - \boldsymbol{R}_\alpha|}\right\}\varphi_n = E_n\varphi_n \tag{2.3}$$

This is a Schrödinger equation in the *electron coordinates only*, which can, in principle, be solved for the eigenvalues $E_n(\{\boldsymbol{R}_\alpha\})$ and eigenfunctions $\varphi_n(\{\boldsymbol{r}_i\}; \{\boldsymbol{R}_\alpha\})$ for a *fixed* set of ion-core positions $\boldsymbol{R}_\alpha$. Thus, the potential energy in this equation is derived from the mutual Coulomb repulsion of the electrons and Coulomb attraction between the electrons and the ions at their fixed positions. For *each* set of $\boldsymbol{R}_\alpha$, the solution of (2.3) is a complete orthonormal set. The orthonormality is expressed as

$$\int \cdots \int \prod_j \mathrm{d}\boldsymbol{r}_j\, \varphi_{n'}^*(\{\boldsymbol{r}_i\}; \{\boldsymbol{R}_\alpha\})\varphi_n(\{\boldsymbol{r}_i\}; \{\boldsymbol{R}_\alpha\}) = \delta_{n,n'}$$

and the completion relation is

$$\sum_n \varphi_n^*(\{\boldsymbol{r}_i'\}; \{\boldsymbol{R}_\alpha\})\varphi_n(\{\boldsymbol{r}_i\}; \{\boldsymbol{R}_\alpha\}) = \delta(\boldsymbol{r}_i - \boldsymbol{r}_i')$$

The Schrödinger equation for ψ is

$$\left\{\sum_\alpha \frac{\boldsymbol{P}_\alpha^2}{2M_\alpha} + \tfrac{1}{2}\sum_{\alpha,\beta}{}' \frac{Z_\alpha Z_\beta e^2}{|\boldsymbol{R}_\alpha - \boldsymbol{R}_\beta|} + E_n(\{\boldsymbol{R}_\alpha\})\right\}\psi_{n,\lambda} = E_{n,\lambda}\psi_{n,\lambda} \tag{2.4}$$

This is a Schrödinger equation in the *nuclear coordinates only.* The electronic coordinates do not *explicitly* enter this equation at all. But they enter *implicitly* through the term $E_n(\{\boldsymbol{R}_\alpha\})$ in the potential energy, which is the energy eigenvalue obtained by solving (2.3) as a function of the $\boldsymbol{R}_\alpha$. Each value of n gives a *different* potential energy in (2.4) and, thus, different different eigenvalues $E_{n,\lambda}$ and eigenfunctions $\psi_{n,\lambda}(\{\boldsymbol{R}_\alpha\})$. For each value of n, the solutions to (2.4) are orthonormal,

$$\int \cdots \int \prod_\beta \mathrm{d}\boldsymbol{R}_\beta \, \psi^*_{n,\lambda'}(\{\boldsymbol{R}_\alpha\})\psi_{n,\lambda}(\{\boldsymbol{R}_\alpha\}) = \delta_{\lambda,\lambda'}$$

and complete,

$$\sum_\lambda \psi^*_{n,\lambda}(\{\boldsymbol{R}'_\alpha\})\psi_{n,\lambda}(\{\boldsymbol{R}_\alpha\}) = \delta(\boldsymbol{R}_\alpha - \boldsymbol{R}'_\alpha)$$

2.2.2 Adiabatic Wavefunctions

Neither (2.3) nor (2.4) alone represent solutions of the original problem posed in (2.1) and (2.2). However, the product functions

$$\Psi_{n,\lambda}(\{\boldsymbol{r}_i\},\{\boldsymbol{R}_\alpha\}) = \varphi_n(\{\boldsymbol{r}_i\};\{\boldsymbol{R}_\alpha\})\psi_{n,\lambda}(\{\boldsymbol{R}_\alpha\}) \tag{2.5}$$

do represent approximate solutions to this problem. The orthonormality and completeness of the φ_n and the $\psi_{n,\lambda}$ discussed above imply that the $\Psi_{n,\lambda}$ are themselves orthonormal,

$$\int \cdots \int \prod_j \mathrm{d}\boldsymbol{r}_j \int \cdots \int \prod_\beta \mathrm{d}\boldsymbol{R}_\beta \Psi^*_{n',\lambda'}(\{\boldsymbol{r}_i\},\{\boldsymbol{R}_\alpha\})\Psi_{n,\lambda}(\{\boldsymbol{r}_i\},\{\boldsymbol{R}_\alpha\})$$
$$= \delta_{n,n'}\delta_{\lambda,\lambda'}$$

and complete,

$$\sum_{n,\lambda} \Psi^*_{n,\lambda}(\{\boldsymbol{r}'_i\},\{\boldsymbol{R}'_\alpha\})\Psi_{n,\lambda}(\{\boldsymbol{r}_i\},\{\boldsymbol{R}_\alpha\}) = \delta(\boldsymbol{R}_\alpha - \boldsymbol{R}'_\alpha)\delta(\boldsymbol{r}_i - \boldsymbol{r}'_i)$$

Thus, the set of functions (2.5) form a valid basis—called the *adiabatic basis*—for *any* problem involving *both* electronic and ionic coordinates: any eigenfunction of $\mathcal{H}$ in (2.2) can be constructed from a linear combination of the $\Psi_{n,\lambda}$:

$$\Psi(\{\boldsymbol{r}_i\},\{\boldsymbol{R}_\alpha\}) = \sum_{n,\lambda} a_{n,\lambda}\Psi_{n,\lambda}(\{\boldsymbol{r}_i\},\{\boldsymbol{R}_\alpha\})$$

where the $a_{n,\lambda}$ are constants determined by the initial conditions.

The *adiabatic approximation*, also known as the *Born–Oppenheimer* approximation (Born and Oppenheimer, 1927), is the assertion that the $\Psi_{n,\lambda}$ are *themselves* approximate eigenfunctions of (2.2). Physically, the adiabatic approximation assumes that the moving ion cores *continuously* deform the electronic wavefunctions (rather than causing any sudden changes in the eigenstate), but that the electrons just provide a potential energy for the ion-core motion. The latter is the change in electronic energy brought about by the necessity of the electrons following the motion of the ion cores.

The Born–Oppenheimer approximation would be exact if the Hamiltonian were diagonal in both n and λ. However, the matrix elements between different ionic states λ and λ' are

$$(n\lambda'|\mathcal{H}|n\lambda) \sim E_{n,\lambda}\delta_{\lambda,\lambda'} + E_{n,\lambda}\frac{m}{M} \tag{2.6}$$

where M is the weighted average of the ion-core masses M_α (weighted by the fraction of electrons bound to the cores), and those between different electronic states are

$$(n'\lambda'|\mathcal{H}|n\lambda) \sim E_{n,\lambda}\delta_{\lambda,\lambda'}\delta_{n,n'} + 2E_{n,\lambda}\left(\frac{m}{M}\right)^{1/2} \tag{2.7}$$

Equation (2.6) indicates that the ratio of the off-diagonal to diagonal ionic matrix elements is of the order of $m/M \leq 10^{-4}$. This indicates that corrections to the Born–Oppenheimer approximation are not significant, even for the lightest atoms. However, Eq. (2.7) shows that the corresponding ratio for electronic matrix elements is only a factor of the order of $(m/M)^{1/2} \sim 10^{-2}$. If this term is appreciable, then different electronic states can be coupled, and we cannot assume, for example, that the electrons are always in their ground state when we solve the ion-core problem for the vibrational modes of the system. The matrix elements are lead to the *electron-phonon interaction*, whose most striking manifestation is the formation of Cooper pairs (Cooper, 1956) (a pair of electrons "bound" by the exchange of a virtual phonon) which leads one mechanism for superconductivity (Bardeen *et al.*, 1957), and corresponds to sudden (i.e., nonadiabatic) transitions between different electronic states without a large change in the ion-core positions. These off-diagonal matrix elements of the full Hamiltonian can be considered as inducing a *scattering* of the electrons by the motion of the ion cores. Only when this term is small do the electronic energies change smoothly as the ions move.

2.3 The Ion-Core Schrödinger Equation

Suppose, for the moment, that the electronic problem in Eq. (2.3) can be solved for every arrangement of ion cores. We must then solve the associated ion-core problem in Eq. (2.4), which can be written as

$$\left[\sum_{\alpha}\frac{\boldsymbol{P}_{\alpha}^{2}}{2M_{\alpha}}+V_{n}(\{\boldsymbol{R}_{\alpha}\})\right]\psi_{n,\lambda}=E_{n,\lambda}\psi_{n,\lambda} \tag{2.8}$$

where the total potential of the cores is given by the sum of their mutual Coulomb repulsion and the potential due to the electronic motion:

$$V_{n}(\{\boldsymbol{R}_{\alpha}\})=\tfrac{1}{2}\sum_{\alpha,\beta}{}'\frac{Z_{\alpha}Z_{\beta}e^{2}}{|\boldsymbol{R}_{\alpha}-\boldsymbol{R}_{\beta}|}+E_{n}(\{\boldsymbol{R}_{\alpha}\}) \tag{2.9}$$

2.3.1 The Structure of Solids

Consider first the case where the electrons are in their ground state ($n = 0$) for every set of ionic positions. Then the solution of (2.8) yields a series of eigenvalues $E_{0,\lambda}$. The ground-state energy is $E_{0,0}$, corresponding to the eigenfunction $\psi_{0,0}$.

If the adiabatic approximation is valid at all, the function V_0 must have absolute minima with respect to each coordinate $\boldsymbol{R}_\alpha$ (Born and Huang, 1954):

$$\left.\frac{\partial V_0}{\partial \boldsymbol{R}_\alpha}\right|_{\{\boldsymbol{R}_{\beta,0}\}}=0 \tag{2.10}$$

where $\boldsymbol{R}_{\beta,0}$ is the equilibrium value of the $\boldsymbol{R}_\beta$. This means that there is an equilibrium configuration for the ion cores, known as the *structure of the solid.* For most, if not all materials, many other *local* minima exist: these alternative structures give the allotropic forms of the solid, which can be stable provided the Gibbs free energy are absolute minima over a finite range of the temperature T and pressure P.

For most simple solids, the equilibrium positions as determined from (x-ray, neutron, or electron) diffraction measurements show almost perfect periodicity. Such solids are called *crystalline* and their structure is described by a *lattice* (a periodic array of positions) and a *basis* (an arrangement of atoms associated with each lattice point). It has not been possible to prove that such a periodic array *must* be the ground state of a large collection of atoms, although the term

$$\tfrac{1}{2}\sum_{\alpha,\beta}{}'\frac{Z_{\alpha}Z_{\beta}e^{2}}{|\boldsymbol{R}_{\alpha}-\boldsymbol{R}_{\beta}|}$$

is minimized *at fixed density* for a periodic structure (i.e., one that keeps all ion cores as far away from each other as possible.

2.3.2 Lattice Vibrations

The potential (2.9) is also used to describe the dynamics of *small* deviations of the ion-cores from their equilibrium positions. This is the problem of *lattice dynamics*. In the simplest theory of this type, called the *harmonic approximation*, the potential is expanded to second-order in the these deviations, $\boldsymbol{x}_\alpha = \boldsymbol{R}_\alpha - \boldsymbol{R}_{\alpha,0}$,

$$V_0(\{\boldsymbol{R}_\alpha\}) = V_0(\{\boldsymbol{R}_{\alpha,0}\}) + \tfrac{1}{2}\sum_{\alpha,\beta} \frac{\partial^2 V_0}{\partial \boldsymbol{R}_\alpha \partial \boldsymbol{R}_\beta}\bigg|_{\{\boldsymbol{R}_{\gamma,0}\}} \boldsymbol{x}_\alpha \boldsymbol{x}_\beta$$

where the first-order term vanishes on account of (2.10). The resulting problem is equivalent to a set of coupled harmonic oscillators. When quantized, these modes are referred to as *phonons*, which obey Bose–Einstein statistics. Cubic and higher-order terms, which can be considered within the adiabatic approximation lead to phonon scattering.

2.4 The Electron Schrödinger Equation

We now return to the electronic problem in Eq. (2.3). The solution of this equation yields a series of eigenvalues E_n and eigenfunctions φ_n which completely characterize the electronic behavior of the system for specified positions $\boldsymbol{R}_\alpha$ of the ion cores. Once such a solution has been obtained, the forces on the nuclei can be related to changes in the quantum mechanical total energy $E_n(\{\boldsymbol{R}_\alpha\})$ by combing the virial theorem (Slater, 1933) with the Hellmann–Feynman theorem (Hellmann, 1937; Feynman, 1939):

$$\langle T\rangle_\varphi = -\tfrac{1}{2}\langle V\rangle_\varphi - \tfrac{1}{2}\sum_\alpha \boldsymbol{R}_\alpha \cdot \boldsymbol{\nabla} E_n(\{\boldsymbol{R}_\alpha\})$$

where $\langle\cdot\rangle_\varphi$ denotes the quantum-mechanical average of the indicated quantity in the state φ, and T and V are the kinetic and potential energy terms, respectively, in (2.3).

A direct solution of Eq. (2.3) is impractical, since apart from there being typically 10^{25} electrons interacting through a strong r^{-1} Coulomb potential, the tabulation of the solution would not be a convenient way of understanding the behavior of the electrons. Accordingly, a variety of approximate methods have been developed for solving the Schrödinger equation for interacting electron systems. These will be taken up in the next section.

2.5 Density Functional Theory

Since the formulation of quantum mechanics in the 1920s, two major approaches have emerged for the computation of the properties of atoms, molecules and solids: Hartree–Fock theory and density functional theory. The Hartree–Fock and related methods have been most popular in the quantum chemistry community, while density functional theory has been the dominant method used for calculations of solids. In this Chapter we discuss the basic concepts of density functional theory and its implementation for the computation of the properties of solids. To set the stage for the discussion of the impact of quantum mechanics and dynamical correlations on the motion of electrons, we begin with a discussion of the Hartree and Hartree–Fock approximations, which date back to the early days of quantum mechanics.

2.5.1 The Hartree and Hartree–Fock Approximations

To appreciate the role that quantum mechanics plays in electronic properties, we adopt an approach to solving the Schrödinger equation (2.3) that is based on a variational principle. We suppose that the wavefunction φ of the system can be written as a *product* of wavefunctions ϕ, one for each of the n particles in the system:

$$\varphi(\{\boldsymbol{r}_i\}) = \phi_1(\boldsymbol{r}_1)\phi_2(\boldsymbol{r}_2)\cdots\phi_n(\boldsymbol{r}_n) \tag{2.11}$$

We then minimize the energy of the Hamiltonian in (2.3) with respect variations in the ϕ_i. This yields an effective Schrödinger equation for each of the ϕ_i:

$$\left[-\frac{\hbar^2}{2m}\nabla_i^2 + e^2\sum_{j\neq i}\int\frac{|\phi_j(\boldsymbol{r})|^2}{|\boldsymbol{r}_i-\boldsymbol{r}|}\,\mathrm{d}\boldsymbol{r} - e^2\sum_\alpha\frac{Z_\alpha}{|\boldsymbol{r}_i-\boldsymbol{R}_\alpha|}\right]\phi_i(\boldsymbol{r}_i) = \varepsilon\phi_i(\boldsymbol{r}_i) \tag{2.12}$$

where the integration in the Coulomb term is understood to include the spin inner product. The first term on the left-hand side of this equation is the kinetic energy, the second term represents the Coulomb potential generated by all the other electrons, and the third term is the attractive Coulomb potential generated by the ion cores. The use of the product wavefunction (2.11), which leads to the effective Schrödinger equation (2.12) is known as the *Hartree approximation.* The obvious drawback of this approximation is that the trial wavefunction (2.11) does not respect the antisymmetric statistics of electrons. Moreover, the Hartree approximation drastically underestimates the tendency for cohesion because there is too much overlap of electronic wavefunctions in regions of repulsive Coulomb potential. Indeed, for metals, the Hartree approximation predicts that there is no cohesion at all! We now consider the effect of incorporating statistics into the trial wavefunction.

Given a basis of wavefunctions ϕ_i for n particles, an n-electron wavefunction which is antisymmetric under interchange of particles, is obtained in the form of a *Slater determinant*:

$$\varphi(\{\boldsymbol{r}_i\}) = \frac{1}{\sqrt{n!}} \begin{vmatrix} \phi_1(\boldsymbol{r}_1) & \phi_1(\boldsymbol{r}_2) & \cdots & \phi_1(\boldsymbol{r}_n) \\ \phi_2(\boldsymbol{r}_1) & \phi_2(\boldsymbol{r}_2) & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \phi_n(\boldsymbol{r}_1) & \cdots & \cdots & \phi_n(\boldsymbol{r}_n) \end{vmatrix} \tag{2.13}$$

Minimizing the energy of the Hamiltonian (2.3) with respect to variations in the ϕ_i, yields a somewhat different effective Hamiltonian from that in (2.12):

$$\begin{aligned} &\left[-\frac{\hbar^2}{2m}\nabla_i^2 + e^2 \sum_j \int \frac{|\phi_j(\boldsymbol{r}_j)|^2}{|\boldsymbol{r}_i - \boldsymbol{r}_i|}\, \mathrm{d}\boldsymbol{r}_j - e^2 \sum_\alpha \frac{Z_\alpha}{|\boldsymbol{r}_i - \boldsymbol{R}_\alpha|} \right] \phi_i(\boldsymbol{r}_i) \\ &\qquad -e^2 \sum_j \int \frac{\phi_j^*(\boldsymbol{r})\phi_j(\boldsymbol{r}_i)\phi_i(\boldsymbol{r})}{|\boldsymbol{r}_i - \boldsymbol{r}_j|}\, \mathrm{d}\boldsymbol{r}_j = \varepsilon \phi_i(\boldsymbol{r}_i) \end{aligned} \tag{2.14}$$

where the integration in the Coulomb term is again understood to include the spin inner product. The first three terms on the left-hand side of this equation are the same as those of the Hartree approximation. The fourth term, which is a direct result of the antisymmetrized trial wavefunction, but still originates in the electronic Coulomb interaction of the Hamiltonian in (2.3), is called the *exchange term.* There are several interesting features about this term. Notice first that the summations over j in the Coulomb and exchange terms are unrestricted.

This is because the terms corresponding to $j = i$ cancel. Secondly, the negative sign indicates that this term lowers the energy of the system in comparison with the Hartree approximation. Third, taking the spin inner product has no effect on the Coulomb term: the sum is over all states and spin orientations, just as in the Hartree approximation. But, in the exchange term, the spin of state j must be *parallel* to the spin of state i. This, together with the negative sign are a direct result of the Pauli exclusion principle: the overlap of spatial wavefunctions with the same spin is minimized, resulting in a decrease of their repulsive Coulomb interaction energies. Finally, the Coulomb term is *local* in that its effect on the $\varphi_i(\boldsymbol{r})$ depends only on the behavior of φ near $\boldsymbol{r}$. In the exchange term, on the other hand, *all* values of φ_i enter the integral, so this is a *nonlocal* term.

The cohesive energy of solids is somewhat improved in the Hartree–Fock approximation by the tendency to keep electrons with parallel spins apart, thus reducing their Coulomb repulsion. But the absence of the same effect for electrons with *anti-parallel* spins means that the tendency toward cohesion is still underestimated. This is essentially a *dynamical* correlation effect because it cannot be accounted for by the antisymmetrization of the wavefunction. The *correlation energy* is therefore *defined* as the difference between the *exact* energy of a system and the energy calculated in the Hartree–Fock approximation.

Although the Hartree–Fock approach has not been widely used for calculating the electronic properties of periodic systems, it has been very successfully applied to calculating the electronic structures and total energies of organic molecules (Hehre *et al.*, 1986). Moreover, in the quantum chemistry community, a variant of the Hartree-Fock approach, based on an expansion in a complete basis of Slater determinants, is used to calculate the properties of molecules. This approach, called *configuration-interaction*, or simply *CI*, is formally exact, but computationally very intensive because the convergence is notoriously slow. In the next section, we describe an altogether different approach to solving the Schrödinger equation in (2.3) that is in principle applicable to all types of quantum mechanical systems (atoms, molecules, and solids).

2.5.2 Basic Density Functional Theory

Hohenberg and Kohn (1964) and Kohn and Sham (1965) formulated a theorem that enabled the solution of the Schrödinger equation in (2.3) to be placed on a sound mathematical basis. This theorem states that the total energy E of a quantum mechanical system depends only on the electron density ϱ of its ground state, i.e., E is a functional of ϱ:

$$E = E[\varrho(\boldsymbol{r})]$$

and that the ground state energy minimizes this functional:

$$\left.\frac{\partial E[\varrho]}{\partial \varrho}\right|_{\varrho_0} = 0 \tag{2.15}$$

where ϱ_0 is the exact electron density of the many-body ground state.

This provides an enormous conceptual simplification to the problem of solving (2.3) because its reduces the number of degrees of freedom from $3N$, where $N \sim 10^{24}$, to the degrees of freedom of a scalar function in three-dimensional space, i.e., 3. The idea of using the electron density as a fundamental quantity in the quantum theory of atoms, molecules, and solids originated in the early days of quantum mechanics with the work of Thomas (1926) and Fermi (1928). As a simple example, all of the thermodynamic properties of an ideal electron gas (e.g., energy, chemical potential, compressibility) are determined completely by its density. For the problem at hand, where the electron density is determined also by the positions of the ion cores, the Kohn–Sham–Hohenberg theorem allows us to write

$$E = E[\varrho(\boldsymbol{r}; \{\boldsymbol{R}_\alpha\})] \tag{2.16}$$

where the $\boldsymbol{R}_\alpha$ are still to be regarded as *parameters*, rather than variables (i.e., degrees of freedom), since the theorem applied to every set of fixed ion-core positions. Equation (2.16) is the basis of density function theory.

The basic idea of how the Kohn–Sham–Hohenberg theorem is applied to solve (2.3) is that each electron is viewed as moving in some average effective potential V_{eff} which is generated by the other electrons and ion cores. This potential must be found self-consistently, since the wavefunction for each electron is included in the effective potential of all other electrons as is seen, for example, in the Hartree and Hartree-Fock approximations. Notice that in this picture, the "real" electrons are replaced by "effective" electrons with the same total density which move as *independent* particles in the effective potential. The Schrödinger equation that determines the wavefunctions of the effective electrons is thus of the general form

$$\left[-\frac{\hbar^2}{2m}\nabla^2 + V_{\text{eff}}\right]\psi_i(\boldsymbol{r}) = \varepsilon_i\psi_i(\boldsymbol{r}) \tag{2.17}$$

where the ψ_i produce the exact charge density:

$$\varrho(\boldsymbol{r}) = \sum_i n_i|\psi_i(\boldsymbol{r})|^2 \tag{2.18}$$

with the occupation number ($n_i = 0$ or $n_i = 1$) of the ith state. The ψ_i do *not* constitute a single-particle approximation to the exact theory; they are simply a way of representing the total electronic charge density.

In density functional theory, the total energy is first decomposed as

$$E = T + e^2\int\frac{\varrho(\boldsymbol{r})\varrho(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|}\,\mathrm{d}\boldsymbol{r}\,\mathrm{d}\boldsymbol{r}' - e^2\sum_\alpha\int\frac{Z_\alpha\varrho(\boldsymbol{r})}{|\boldsymbol{R}_\alpha-\boldsymbol{r}|}\,\mathrm{d}\boldsymbol{r} + E_{\text{xc}}[\varrho] \tag{2.19}$$

The simplest terms to understand are the second and third terms on the right-hand side of this equation. The second term is the Coulomb repulsion between the electrons and the third term is the Coulomb attraction between the electrons and the ion cores. Both of these terms are essentially classical in origin, a characteristics that will be used explicitly below. The term T is the sum of the kinetic energies of all of the "effective" electrons moving as *independent* in an effective potential. With the wavefunctions of these particles given in (2.18), T is given by

$$T = -\frac{\hbar^2}{2m}\sum_i n_i\int\psi_i^*(\boldsymbol{r})\nabla^2\psi_i(\boldsymbol{r})\,\mathrm{d}\boldsymbol{r}$$

Since this term considers the electrons as moving independently, dynamical correlations are excluded by construction.

The last term on the right-hand side of (2.19) includes all of the exchange and correlation contributions to the total energy and is called the exchange-correlation energy. The exchange energy, as discussed for the Hartree–Fock approximation, acts to reduce the Coulomb repulsion for electrons with parallel spins. The correlation energy is due to the same effect for electrons with anti-parallel spins and is a result of dynamical correlations between the electrons.

2.5.3 The Kohn–Sham Equations

Given the expression (2.19) for the energy, a procedure is now required for implementing density functional theory in practical calculations. In addition to this energy, the key equations are the expression (2.18) for the many-body density in terms of single-particle wavefunctions, and the stationarity (2.15) of the energy with respect to first-order variations of ϱ about its ground-state value. Notice that any change in the single-particle wavefunctions induces a corresponding change in ϱ. Thus, the variational condition (2.17) can be used to derive the conditions that the ψ_i produce the ground-stare density. These are obtained by substituting (2.19) and (2.18) into (2.15) and interpreting the variation with with respect to ϱ as a variation of each of the ψ_i, and we obtain equations of the form (2.17)

$$\left[-\frac{\hbar^2}{2m}\nabla^2 + V_{\rm eff}\right]\psi_i(\boldsymbol{r}) = \varepsilon_i\psi_i(\boldsymbol{r}) \tag{2.20}$$

where the effective potential is given by

$$V_{\rm eff}(\boldsymbol{r}) = V_{\rm C}(\boldsymbol{r}) + V_{\rm xc}[\varrho(\boldsymbol{r})]$$

which is the sum of Coulomb ($V_{\rm C}$) and exchange-correlation ($V_{\rm xc}$) contributions. These are called the *Kohn–Sham* equations. The wavefunctions obtained by solving these equations yield the ground-state density which minimizes the total energy and form an orthonormal basis,

$$\int \psi_i^*(\boldsymbol{r})\psi_j(\boldsymbol{r})\,\mathrm{d}\boldsymbol{r} = \delta_{i,j}$$

This condition is insured by the Lagrange multipliers ε_i, which appear as eigenvalues of the ψ_i.

The Coulomb potential in (2.20) is obtained from the energy in (2.19) as

$$V_{\rm C}(\boldsymbol{r}) = -e^2\int\frac{\varrho(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|}\,\mathrm{d}\boldsymbol{r} + e^2\sum_\alpha\frac{Z_\alpha}{|\boldsymbol{R}_\alpha-\boldsymbol{r}|}$$

This has a purely classical origin, since $V_{\rm C}$ is a solution of Poisson's equation

$$\nabla^2 V_{\rm C}(\boldsymbol{r}) = -4\pi e^2 q(\boldsymbol{r})$$

where $q(\boldsymbol{r})$ is the sum of the electronic and ion-core charge densities:

$$q(\boldsymbol{r}) = \varrho(\boldsymbol{r}) + \sum_{\alpha} Z_{\alpha}\delta(\boldsymbol{r} - \boldsymbol{R}_{\alpha})$$

The exchange-correlation potential V_{xc} is related to the exchange-correlation energy E_{xc} by

$$V_{\mathrm{xc}} = \frac{\partial E_{\mathrm{xc}}}{\partial \varrho}$$

This equation is formally exact in the sense that no approximations have been made in its derivation. However, while expressions for the kinetic energy and Coulomb potential energies are known, there is no know way of obtaining the exchange-correlation energy and, thus, the exchange-correlation potential. Therefore, the utilization of the Kohn–Sham equations in any other but a purely formal manner means that a particular form must be assumed for the exchange-correlation term. One such approach is discussed in the next section.

2.6 The Local Density Approximation

The Kohn–Sham theorem requires the exchange-correlation energy and potential to be functionals of the total electron density ϱ. One approximate form of these functions that has been widely adopted is obtained by assuming that E_{xc} depends only on the *local* value of ϱ:

$$E_{\mathrm{xc}}^{\mathrm{LDA}} = \int \varrho(\boldsymbol{r})\varepsilon_{\mathrm{xc}}(\boldsymbol{r})\,\mathrm{d}\boldsymbol{r}$$

This is called the local density approximation (LDA) and its validity rests on two assumptions:(i) exchange and correlation are dominated by the density in the immediate vicinity of a point $\boldsymbol{r}$, and (ii) these effects do not vary strongly with position. The LDA has been found to work well for many metals, but fails in systems with strongly varying electron densities, such as those involving f-electrons. The LDA is thus exact for an interact system with a constant density (see below), but becomes less accurate as the variations of the density increase.

The implementation of the LDA requires still requires a functional form for $\varepsilon_{\mathrm{xc}}$, i.e., the exchange-correlation energy per electron as a function of the electron density. This quantity has been studied in a system of interacting electrons with a constant density because of a homogeneous background of positive charge to render the system electrically neutral, called the *homogeneous electron gas*, with a number approaches, including many-body perturbation theory (Hedin and Lundqvist, 1972) and quantum Monte Carlo methods (Ceperley and Alder, 1980), and is now well established (Perdew and Wang, 1992). Earlier studies (Pines, 1962) used many-body perturbation theory in the limit of high density to calculate the correlation energy of this system. As a result of this work, $\varepsilon_{\mathrm{xc}}$ is known accurately for a range of densities. For the homogeneous electron gas, we can write $V_{\mathrm{xc}} = V_{\mathrm{x}} + V_{\mathrm{c}}$, where the exchange potential, V_x, is given by (Gáspár, 1954; Kohn and Sham, 1965)

$$V_{\mathrm{x}} = -2\left(\frac{3}{\pi}\varrho\right)^{1/3}$$

and the correlation potential, V_{c}, is (Hedin and Lundqvist, 1972)

$$V_{\mathrm{c}} = -c\ln\left(1 + \frac{1}{x}\right)$$

where

$$c = 0.0225, \qquad x = \frac{r_s}{21}, \qquad r_s = \left(\frac{3}{4\pi\varrho}\right)^{1/3}$$

In these equations, the energies are in units of Hartrees (1 Hartree = 27.21165 eV) and the units for electron density are number of electrons per Bohr radius cubed.

A huge body of calculations over the past 40 years have revealed some systematic trends in LDA calculations in comparison with experiment, where available, and with other calculations, such as CI. For a large variety of systems, including solids, surfaces, and even molecules, calculations of total energies have produces interatomic bond lengths to within ±0.05 Å of measured values and, in the most favorable conditions, to within ±0.02 Å. But, such calculations have also found systematic errors with results produced by the LDA: (i) weak bonds tend to be too short, and (ii) binding energies are too large, sometimes with errors of 50%. In the next chapter, we will describe one way of correcting the LDA to alleviate these discrepancies.

References

J. Bardeen, L. Cooper, and J. Schrieffer, *Phys. Rev.* **108**, 1175 (1957).

M. Born and R. Oppenheimer, *Ann. Physik* **84**, 457 (1927).

D.M. Ceperley and B.J. Alder, *Phys. Rev. Lett.* **45**, 566 (1980).

L. Cooper, *Phys. Rev.* **104**, 1189 (1956).

M. Born and K. Huang, *Dynamical Theory of Crystal Lattices* (Clarendon Press, Oxford, 1954), pp. 166–169.

E. Fermi, *Z. Phys.* **48**, 73 (1928).

R.P. Feynman, *Phys. Rev.* **56**, 340 (1939).

R. Gáspár, *Acta Phys. Acad. Sci. Hung.* **3**, 263 (1954).

L. Hedin and S. Lundqvist, *J. Phys.* (France) **33**, C3 (1972).

W.J. Hehre, L. Radom, P. Schleyer, and J.A. Pople, *Ab Initio Molecular Orbital Theory* (Wiley, New York, 1986).

H. Hellmann, *Einführung in die Quantenchemie* (Deuticke, Leipzig, 1937).

P. Hohenberg and W. Kohn, *Phys. Rev.* **136**, B864 (1964).

W. Kohn and L.J. Sham, *Phys. Rev.* **140**, A1133 (1965).

J.P. Perdew and Y. Wang, *Phys. Rev. B* **45**, 13244 (1992).

D. Pines, *The Many Body Problem* (Benjamin, London, 1962).

J.C. Slater, *J. Chem. Phys.* **1**, 687 (1933).

L.H. Thomas, *Proc. Cam. Phil. Soc.* **23**, 542 (1926).

Chapter 3

Molecular Dynamics and Kinetic Monte Carlo Simulations

Computer simulation has played a central role in the development of our understanding of epitaxial systems. Twenty years ago, Monte Carlo simulations of the so-called solid-on-solid (SOS) model (Weeks and Gilmer, 1979) provided essential data for the test of then current analytic predictions for the *macroscopic* growth rate. This same methodology and model are being used today with considerable success to provide a *microscopic* interpretation for various diagnostic measurements of the growth process. This type of result, combined with other developments in the intervening years such as Monte Carlo simulations of more realistic models (Madhkar and Ghaisas, 1987) and molecular dynamics simulations (Dodson, 1990) helps explain the growing popularity of computer methods in the study of epitaxial growth.

3.1 Molecular Dynamics

In the molecular dynamics method (Schneider *et al.*, 1987; Dodson, 1990), real-space trajectories of the atoms are determined by numerical integration of Newton's equations of motion. All of the physics is contained in the forces acting upon each particle in the system, which are determined by the interatomic potentials for the atoms in the system. Other constraints may be placed on the system, such as a fixed temperature and the restriction of particle activity to a fixed volume of space.

The expression for the total energy E of the system as a function of the positions of the atoms, $\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots$ is written as an expansion in terms of n-body potentials:

$$E(\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots) = \frac{1}{2}\sum_{i\neq j} V_2(r_{ij}) + \frac{1}{3!}\sum_{i\neq j\neq k} V_3(\boldsymbol{r}_{ij}, \boldsymbol{r}_{jk}, \boldsymbol{r}_{ik}) + \cdots \qquad (3.1)$$

Here, $\boldsymbol{r}_{ij} = \boldsymbol{r}_i - \boldsymbol{r}_j$ are interatomic separations, and V_n represents an n-body interaction energy. We have assumed in writing (9) that we are dealing with a single-component system, so the interaction energies do not depend on the relative separations, not the atomic type. The two-body term represents the interaction of two atoms at positions $\boldsymbol{r}_i$ and $\boldsymbol{r}_j$ and depends only on the separation $\boldsymbol{r}_{ij}$. The three-body term V_3 depends upon the relative orientations of triplets of atoms, i.e., no simply upon interatomic distances but on bond angles as well. For the expansion of the potential in terms of n-body potentials to be meaningful, it must converge rapidly with increasing n; in fact, virtually all studies truncate this expansion at either the second or the third order. Given the potential, the Hamiltonian of the system can be constructed from which the time-development of the system can be calculated from Hamilton's equations

Despite the evident appeal of applying molecular dynamics, there are two impediments to the practical application this method. The first concerns the choice of potential. The potential is determined by a combination of semi-empirical fitting to observed behavior combined with physically-motivated characteristics. For example, a two-body potential is appropriate for systems where the formation of chemical bonds is not an important feature of the dynamics, such as the interaction of He with a metal. However, for most systems, a three-body component is also required to describe the interatomic interactions responsible for bond-formation. Once the potential has been chosen, and the interaction energy determined, the forces acting on each atom is calculated and, at each time step, the atoms are moved in the direction of the forces. This method can be used both to calculate the structure of a collection of atoms by determining the minimum energy as a function of the atomic positions and to determine the dynamics of a system, such as the melting or freezing of a surface.

The commitment to a fixed form of interatomic potential has both advantages and disadvantages. On the positive side, the description of dynamic phenomena is enormously simplified in comparison with *ab initio* methods, since the individual potentials need not be recalculated as the atomic configuration is varied. For systems where hybridization is not a strong function of the local environment, this is a good approximation. On the other hand, the calculated behavior of systems which do show strong effects of rehybridization, such as C and Si, different potentials will produce different relative energies among different structures, which then can lead to different descriptions of the dynamics.

The second important feature of the molecular dynamics method is the time step used between successive evaluations of the forces acting on the atoms and the corresponding adjustments of the atomic positions. In common with the quantum molecular dynamics method, classical molecular dynamics still suffers from the “time gap” of the number of incremental time steps needed to obtain macroscopic time scales. This is especially problematic for infrequent events, such as surface diffusion, though there has been progress recently in dealing with such situations (Voter, 1997). Consequently, most implementations of molecular dynamics simulations of epitaxial growth have used unrealistically high growth rates in order to deposit a significant amount of material during the course of the simulation. Nevertheless, when used judiciously, this method can also be used to identify and quantify important kinetic processes.

3.2 Kinetic Monte Carlo Simulations

The Monte Carlo method is an additional level of abstraction over the molecular dynamics method. The effect of fast dynamical events is taken in account phenomenologically through transition rates for slower events. For example, in describing the mobility of surface adatoms, the diffusion can often be approximated as a nearest-neighbor hopping process with a transition rate given by the product of an attempt rate, which is typically of the order of the atomic vibrational frequency, and the probability of success per attempt, which is represented as an exponential involving the energy barrier to the process. The term “Monte Carlo” refers to the random sampling of numbers, in analogy with a roulette wheel. A Monte Carlo simulation proceeds by calculating the probability distribution of a physical event or series of events. A random number is then generated from a uniform distribution in the interval [0,1] and compared with the probability of the event occurring. If the random number is greater than or equal to the probability, the event occurs, otherwise not. For example, the hopping of atoms is then described by comparing the probability of the hopping, p, with random number, n, chosen from the interval $[0, 1]$: the hopping occurs only if $n \leq p$. Other kinetic events are similarly treated.

While the details of the underlying mechanism for the hopping are lost, the effect of the fast processes is correct on average. Thus, if the large-scale features of the dynamics of a system are of interest, rather than the details of the structure, the Monte Carlo method can offer considerable advantages over the molecular dynamics method, both in terms of the "real time" over which the simulation evolves, as well as the number of atoms included in the simulation. It must be emphasized that the construction of a model for a Monte Carlo simulations can often be greatly simplified and justified by appealing to a related molecular dynamics simulation. This includes the identification of the important physical process, as well as the numerical values of the corresponding kinetic barriers.

Much of the simulation work carried out in connection with crystal growth has been based upon the solid-on-solid (SOS) model. This model was explored widely for its applicability to crystal growth near equilibrium by Weeks and Gilmer (1979) in the 1970s [2]. The distinguishing feature among the various implementations of the SOS model, some examples of which are described in the following subsections, is in the number and type of processes that are considered explicitly. An Arrhenius expression (1) is associated with each process, which requires assigning values to the attempt frequency and to the barrier. If it was possible to isolate the effects of an individual process, then these parameters could be determined directly from experiment and used in the simulation. However, since these models always include only the processes that are expected to be the most important, a direct measurement of K_0 and E is not possible, so values must be assigned either based upon physical arguments, or fitted by comparing some result from the simulation with experiment. It is important to emphasize, the that omission of fast processes that are not rate-determining means that the attempt frequency and the barrier are to be regarded as *effective* parameters, since the neglected processes could influence the values of these parameters without affecting the qualitative features of the model.

In the SOS model, growth is initiated by the random deposition of atoms onto the substrate. The subsequent migration of surface adatoms is taken as a nearest-neighbor hopping process whose rate is

$$k(T) = k_0 \exp(-E_D/kT) \tag{3.2}$$

Here, the k_0 corresponds to an adatom vibrational frequency and E_D is the hopping barrier. The prefactor is usually taken either as $k_0 = 2kT/h$ or sometimes simply assigned the constant value 10^{13}s^{-1}. The hopping barrier comprised of two terms, a term, E_S, due to the substrate, and a contribution, E_N, from each nearest neighbor along the substrate. Thus, the barrier to hopping of an n-fold coordinated atom ($n = 0, \ldots, 4$) is given by $E_D = E_S + nE_N$. The barrier is assumed to depend only on the *initial* environment of the migrating atom. The quantities E_S and E_N are the only free parameters of the model.

A test of this model was carried out by Shitara et al. (1992) and Šmilauer and Vvedensky (1993). Reflection high-energy electron-diffraction measurements were carried out on vicinal surfaces with misorientations of 2° and 3° for a range of temperatures near the temperature T_c where growth became dominated by step advancement. The As/Ga ratio was maintained at approximately 2.5 to avoid variations in the effect of the As on the growth kinetics as a function of each Ga flux. Comparisons between the measured and simulated values of T_c for the two misorientations and several Ga fluxes produced the following optimized energy barriers:

$$E_S = 1.58\text{eV} \pm 0.02\text{eV}, \qquad E_N/E_S \approx 0.15\text{eV} \tag{3.3}$$

It must again be stressed that these values are *effective* migration barriers and include all of the effects not included explicitly; in particular, there is an implicit As dependence, so these are the appropriate barriers only for the specified As/Ga flux ratio.

In fact the correspondence between the RHEED measurements and the simulations go much deeper than simply the determination of T_c. In Fig. 2 is shown a comparison between the measured RHEED specular intensity and simulations of the step density on a vicinal surface with a misorientation of 2° and for the indicated Ga flux. The incident azimuthal angle of the electron beam such that the beam direction is perpendicular to the staircase of terraces and steps, and the polar was chosen to insure that the maxima of the RHEED intensity corresponds to the deposition of increments of a monolayer of material.

The simulated step density is seen to reproduce several features of the measured RHEED specular intensities, including the decay of the

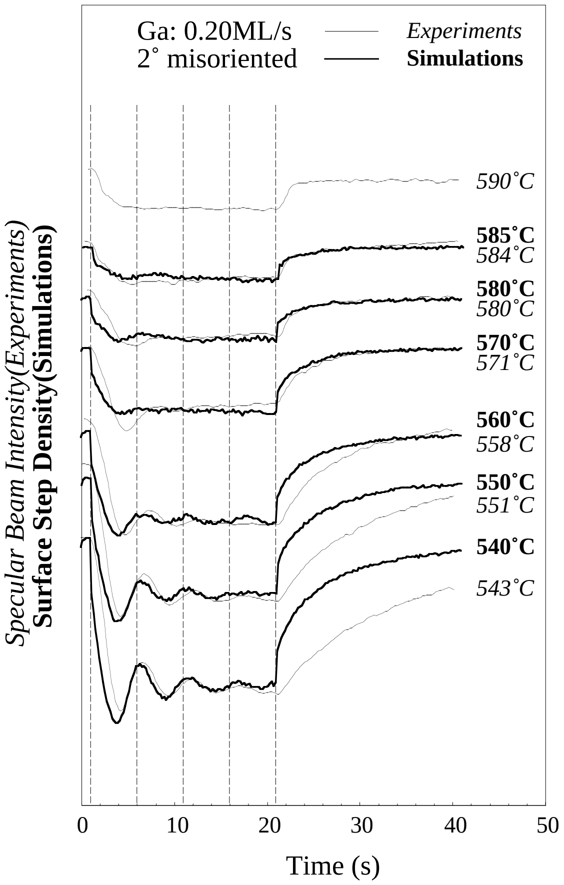

Figure 3.1: Comparison of measured RHEED specular intensities and step densities from simulated surfaces during growth on vicinal GaAs(001) with a misorientation of 2° for the indicated growth conditions (Shitara *et al.*, 1992). The scale of the step densities increases downward and the data for successively higher temperatures are shifted for ease of comparison.

oscillations at lower temperatures, the gradual decrease in the number of oscillations with increasing temperature and in the difference between the pre-growth and post-growth amplitudes, a slight shift of the first maximum with increasing temperature. In addition, there is even a degree of *quantitative* agreement between the two sets of data in that the relative changes of the amplitudes of the two quantities with temperature are the same. This provides strong support for the suggestion that the basic processes of diffraction are the same for a surface before growth and during, but that the disorder simply reduces the efficiency of these processes.

References

B.W. Dodson, *CRC Crit. Rev. Solid State Mater. Sci.* **16**, 115 (1990).

A. Madhukar and S.V. Ghaisas, *CRC Crit. Rev. Solid State Mater. Sci.* **13**, 1434 (1987).

M. Schneider, I.K. Schuller, and A. Rahman, *Phys. Rev. B* **36**, 1340 (1987).

T. Shitara, D.D. Vvedensky, M.R. Wilby, J. Zhang, J.H. Neave, and B.A. Joyce, *Appl. Phys. Lett.* **60**, 1504 (1992).

P. Šmilauer and D.D. Vvedensky, *Phys. Rev. B* **48**, 17603 (1993).

A.F. Voter, *Phys. Rev. Lett.* **78**, 3908 (1997).

J.D. Weeks and G.H. Gilmer, *Adv. Chem. Phys.* **40**, 157 (1979).

Chapter 4

Analytic Theories of Morphological Evolution

The use of analytic methods is complementary to computer simulations in that they do not incorporate the detail of simulations but instead attempt to include the essential features to describe a particular aspect of growth kinetics. The best known examples of this approach are the Burton, Cabrera, and Frank (BCF) theory, homogeneous rate equations, and stochastic equations of motion for the profile of the growing surface. Each of these addresses either a specific regime of growth or is concerned with a description on particular length and time scales. All of these approaches are largely phenomenological in that the precise connection with atomistic processes is not always apparent, so comparisons with atomistic simulations are seldom unequivocal. In this chapter, we describe the main types of analytic approaches and indicate the type of information that can be provided by these studies, and their regime of applicability.

4.1 Theory of Burton, Cabrera and Frank

A common starting point for modeling epitaxial growth is based on the work of Burton, Cabrera and Frank (BCF) (Burton *et al.*, 1951). The BCF theory describes growth on a monatomic vicinal surface (Fig. 4.1) by the deposition of single atoms. The central quantity in this theory is the adatom concentration $c(\boldsymbol{x}, t)$ at position $\boldsymbol{x}$ and time t. This quantity varies with time because of atomic surface diffusion (with diffusion

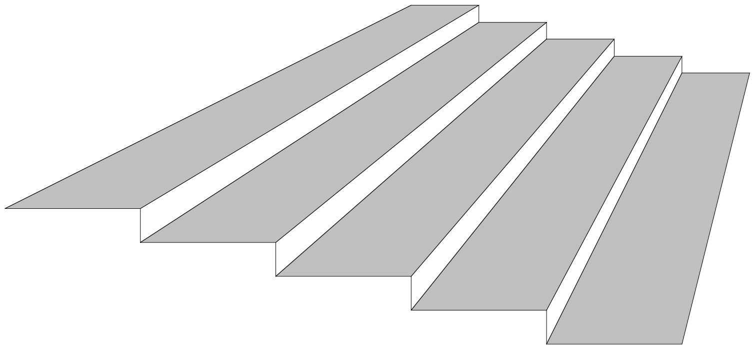

Figure 4.1: Schematic diagram of an ideal stepped surface showing a regular array of terraces (indicated by shading) and straight steps.

constant D) and the deposition of atoms by the molecular beam (with flux J). We will assume that the desorption of the atoms from the surface can be neglected, but this can be readily included in the theory if required. In the simplest form of the BCF theory, the equation determining $c(\boldsymbol{x}, t)$ is a one-dimensional diffusion equation with a source term:

$$\frac{\partial c}{\partial t} = D\frac{\partial^2 c}{\partial x^2} + J \tag{4.1}$$

This equation is supplemented by boundary conditions at the ends of a terrace, e.g.,

$$c(0,t) = 0, \qquad c(L,t) = 0 \tag{4.2}$$

where L is the terrace length and the range of x is $0 \leq x \leq L$. These boundary conditions, which are called ‘absorbing’ boundary conditions, stipulate that adatoms are absorbed at a step edge and immediately incorporated into the growing crystal with no possibility of subsequent detachment. Other boundary conditions can also be chosen, as discussed by Ghez and Iyer (1988).

We will focus here on the steady-state (time-independent) solution of equation (4.1). By setting the right-hand side of this equation equal to zero and invoking the boundary conditions in (4.2), we obtain

$$c(x) = \frac{J}{2D}x(L - x) \tag{4.3}$$

This expression is a parabola that attains its maximum at the center of the terrace and vanishes at the terrace edges, as required by the absorbing boundary conditions. This solution shows that as J/D

increases, which corresponds to decreasing temperature (through D) and/or increasing J, the concentration of atoms builds up on the terraces. Since the BCF theory neglects interactions between atoms, the growth conditions must be chosen to insure that the adatom concentration is maintained low enough to render their interactions unimportant. Thus, the BCF theory is appropriate only for small values of J/D, where growth is expected to occur by step flow.

Since the dimensions of D and J are

$$[D]=\frac{\text{length}^2}{\text{time}}, \qquad [J]=\frac{1}{\text{length}^2\times\text{time}} \tag{4.4}$$

we can use these quantities to form a characteristic length:

$$\ell=(D/J)^{1/4} \tag{4.5}$$

Suppose that on a vicinal surface the mean terrace width is ha, where a is the nearest-neighbor distance. If we now write the diffusion constant in the Arrhenius form,

$$D=a^2k_0\exp(-E_D/kT) \tag{4.6}$$

where k_0 is the attempt frequency for the diffusion process, E_D is the energy barrier to diffusion, and k is Boltzmann's constant, then setting $\ell=ha$ yields

$$kT_c=E_D\left[\ln\left(\frac{k_0}{a^2h^4J}\right)\right]^{-1} \tag{4.7}$$

This expression gives a surprisingly account of the temperature at which growth becomes dominated by step advancement as a function of misorientation h and flux J, as measured by the disappearance of RHEED oscillations.

4.2 Homogeneous Rate Equations

The BCF theory describes a surface growing by the advancement of steps. As the temperature is lowered or the deposition flux raised, growth by the formation, accretion and coalescence of clusters on the terraces becomes more likely and the BCF picture is no longer appropriate. One way of providing a theoretical description of this growth mode within an analytic framework is with rate equations (Venables *et al.*, 1984).

Rate equations have provided a conceptual and computational framework for examining many aspects of coagulation and aggregation phenomena since the early parts of this century (Smoluchowski, 1916). But it has been the advent of the scanning tunneling microscope that has led to the recent resurgence in the application and refinement of rate equations for describing island kinetics and island morphologies during epitaxial growth. By allowing as-grown island morphologies to be imaged directly in real space with atomic resolution, the *mechanisms* of particular atomistic processes can often be identified and their *rates* estimated from comparisons between experimentally measured quantities and those obtained from rate equations and simulations. This has spawned a huge experimental and theoretical effort aimed at characterizing islands in the submonolayer regime of epitaxial growth prior to significant coalescence, where the statistical properties of islands can be isolated, analyzed, and interpreted in terms of atomistic diffusion, nucleation, and growth kinetics.

In the rate equation approach to epitaxial growth, the dynamical variables are the densities of adatoms and islands on the surface. These densities are taken to be spatially homogeneous, so their governing equations are referred to as *homogeneous* rate equations. Such rate equations are constructed on the basis of a phenomenological identification of the processes that cause adatom and island densities to change. In this section, we will first consider the simplest rate equation description of growth, for which the formation of islands and their subsequent growth proceeds by the *irreversible* capture of atoms, i.e. atoms which attach to islands cannot subsequently detach. We then examine the information which can be obtained from island-size distributions and discuss the origin of the scaling for of these distribution functions.

4.2.1 Irreversible Aggregation Kinetics

We will signify the density of surface atoms by $n_1(t)$ and the density of s-atom islands by $n_s(t)$, where $s > 1$. The rate equation for n_1 is

$$\frac{\mathrm{d}n_1}{\mathrm{d}t} = J - 2D\sigma_1 n_1^2 - Dn_1 \sum_{s=2}^{\infty} \sigma_s n_s \tag{4.8}$$

The left-hand side of this equation is the total rate of change of the adatom density and on the right-hand side are the rates of individual processes which either increase or decrease this quantity. The first term on the right-hand side is the deposition of atoms onto the substrate, which increases the adatom density, and so has a positive sign. The next term is the formation of a two-atom island by the irreversible attachment of two migrating atoms. This term decreases the number of adatoms and thus has a negative sign. The rate for this process is proportional to the *square* of the adatom density because *two* adatoms are required to form a two-atom island and to D, the adatom diffusion constant. The third term is the rate of depletion of adatoms by their capture by islands. This term is proportional to the product of the adatom and total island densities and must also have a negative sign. The quantities σ_i in (4.8), called 'capture numbers,' account for the diffusional flow of atoms into the islands (Venables *et al.*, 1984; Bales and Chzan, 1994; Bartelt and Evans, 1996). We will discuss these quantities below.

The rate equations for the density of an s-atom island $n_s(t)$ is

$$\frac{\mathrm{d}n_s}{\mathrm{d}t} = Dn_1\sigma_{s-1}n_{s-1} - Dn_1\sigma_s n_s \tag{4.9}$$

The first term on the right-hand side of this equation is the rate of increase of n_s by the attachment of adatoms to $(s-1)$-atom islands. Similarly, the second term is the rate of decrease of n_s by the attachment of adatoms to s-atom islands to form a $(s+1)$-atom islands.

To illustrate the calculus of rate equations, we set all of the capture numbers equal to unity. The hierarchy of coupled equations in (4.9) can then be contracted into a single equation by introducing the total island density, $N = \sum_{s>1} n_s$. Then, by using this definition in (4.8) and summing the equations in (4.9) over s, we obtain a closed set of two equations for n_1 and N alone:

$$\frac{\mathrm{d}n_1}{\mathrm{d}\theta} = 1 - 2Rn_1^2 - Rn_1N \tag{4.10}$$

$$\frac{\mathrm{d}N}{\mathrm{d}\theta} = Rn_1^2 \tag{4.11}$$

where $R = D/J$ and we have used the relation between the coverage and the flux in the absence of desorption, $\theta = Jt$, to replace the time t by the coverage θ as the independent variable. This replacement is made because the coverage is the important quantity and it can be measured directly from an STM image.

These equations are straightforward to integrate numerically and the results are shown in Fig. 4.2. There are several important features

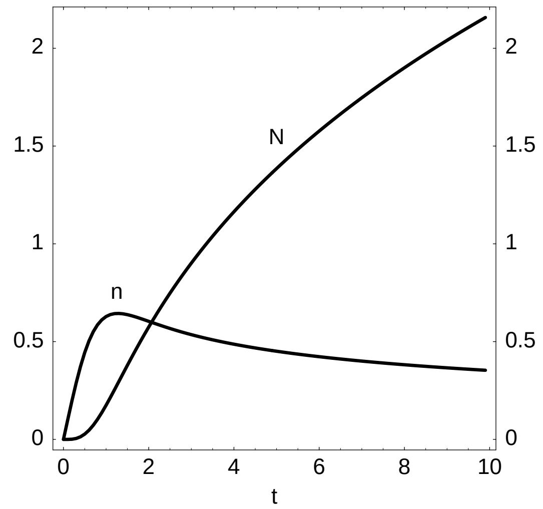

Figure 4.2: The solution of the rate equations (4.10) and (4.11) for the dimensionless quantities $\mathsf{n}_1 = (D/J)^{1/2} n_1$ and $\mathsf{N} = (D/J)^{1/2} N$ as a function of $\mathsf{t} = (DJ)^{1/2} t$.

to note. The concentration n_1 increases more steeply than N initially, by N continues to increase, while n_1 approaches an almost stationary value. We can obtain analytic solutions of (4.10) and (4.11) in these limiting regimes with relatively little effort. The solution at short times ($\theta \ll 1$) of these equations is easily determined:

$$n_1 \sim \theta, \qquad N \sim \theta^3 \tag{4.12}$$

The density of atoms initially shows a linear increase with coverage (or time), which is due entirely to the deposition flux. The islands are somewhat slower in their early development, showing a cubic dependence on the time because the low surface atom density is not sufficient for appreciable island formation. Equation (4.11) shows that N continues to increase for all later times, but equation (4.10) indicates that although n_1 increases initially, when the right-hand side becomes negative (as it must, since N always increases), it starts to decrease. This decrease continues as N increases until we reach a *steady-state* regime where $n_1 \ll N$ and $\mathrm{d}n_1/\mathrm{d}\theta \ll 1$. In this regime, we obtain the *scaling laws* for the adatom and island densities:

$$n_1 \sim \theta^{-1/3}(D/J)^{-2/3}, \qquad N \sim \theta^{1/3}(D/J)^{-1/3} \tag{4.13}$$

Notice that, just as in equation (4.3), the ratio D/J is a controlling parameter for a quantity that characterizes the growing surface. The equation for N indicates that increasing the temperature (i.e. increasing D) and/or decreasing the flux J causes the island density to decrease. This results in a surface with fewer but larger islands.

The results in equation (4.13), which were obtained by setting all of the captures equal to unity, display the correct scaling of N with D/J, but not with θ, nor does this approximation produce the correct distribution of island sizes. This can be traced to the assumption of constant capture numbers, which treats the islands as though they have no lateral extent, i.e. as 'point islands' (Bartelt and Evans, 1996). The next level of approximation is to include the spatial extent of the islands in an average way by assuming that the local environment of each island is independent of its size and shape (Bales and Chzan, 1994). This produces the correct scaling of N with both D/J and θ, but still not the correct island size distribution. To obtain a complete description of the island morphology, it is necessary to proceed one step further by including spatial information in the capture numbers which accounts for the correlations between neighboring nucleation centers and the differences in the local environment of individual islands (Bartelt and Evans, 1996).

4.2.2 The Distribution of Island Sizes

The morphology of a surface in the submonolayer regime, where islands have formed but have not yet begun to coalesce, is rich in information about the atomistics processes that are responsible for the formation and growth of these islands. The submonolayer island morphology also provides important signatures about processes that are operative in the *multi*layer regime. Apart from processes that are intrinsic to a clean homoepitaxial system, there is the effect of various surface impurities that are introduced either deliberately ("surfactants") or are unavoidably present because of the polyatomic molecules used to deliver the atoms of the growing material. Strain can also affect the morphologies of heteroepitaxial islands by causing island-size- and island-shape-dependent changes to both attachment and detachment barriers at island edges. An important practical application of these ideas is the growth of three-dimensional islands during Stranski-Krastanov growth, which have promising properties for applications as quantum dots.

The ability to image and acquire statistics about submonolayer islands with the STM has made this a very active area of research. One of the most far-reaching result of this work (Bartelt and Evans, 1992) is that the density $n(\theta, s)$ of s-atom islands at a coverage θ can be written as

$$n_s = \frac{\theta}{s_{\rm av}^2} f(s/s_{\rm av}) \tag{4.14}$$

where $s_{\rm av}$ is the average island size and f is a scaling function. This function is 'universal' in the sense that its dependence on the coverage, deposition rate, and substrate temperature is contained entirely in $s_{\rm av}$, which acts as the characteristic size for this problem. The scaling form (4.14) was suggested originally on the basis of a dynamical scaling ansatz (Viscek and Family, 1984) and was supported by extensive KMC simulations (Bartelt and Evans, 1992). Scanning tunneling microscopy measurements on metal (Stroscio and Pierce, 1994; Müller *et al.* 1996) and semiconductor (Bressler–Hill *et al.* 1995; Avery *et al.* 1997) surfaces, together with theoretical and simulational studies, are consistent with equation (4.14) and have shown how f is affected by various processes, such as adatom attachment and detachment (Ratsch *et al.*, 1994, 1995; Bartelt *et al.*, 1995), magic island sizes (Schroeder and Wolf, 1995), and adatom exchange (Chambliss and Johnson, 1994; Zangwill and Kaxiras, 1995).

4.3 Kinetic Roughening

In our discussion of the solution of the BCF equation in (4.3), we focussed on the roles of D and J in determining if the growth of a vicinal proceeds by step flow. However, there is another quantity that is equally important in determining the growth mode: the terrace length L. Suppose we fix the temperature (i.e. D) and J. Then if L is small enough, the adatom density will be corresponding small, and growth proceeds by step flow. But for surfaces with larger terraces, the adatom concentration on the terrace increases until at some terrace width L^*, adatom interactions are no longer negligible, and the growth of islands becomes appreciable. This simple observation is the basis of understanding multilayer growth on singular surfaces.

Consider a singular surface (or a surface with a very small misorientation angle). Then we are in a regime where $L^* \ll L$, so the probability of atoms encountering one another on a terrace is large and the presence of the steps does not significantly affect the growth of the crystal. The growth of the first surface layer is initiated by the formation of small 2D clusters, which grow laterally by capturing migrating atoms. Thus, to an electron beam the surface appears rough, which causes the specular intensity of the beam to decrease. This roughness continues to increase until the clusters begin to coalesce, at which point the surface appears to smoothen, causing the specular intensity to increase. Once the new layer is formed this process is repeated, resulting in intensity oscillations of the RHEED specular beam.

What is the origin of the decaying envelope seen in the RHEED oscillations shown in Fig. 1.3? The layer-by-layer process just described is not perfect. Once the lateral size of an island becomes large enough, atoms deposited on top of this island can collide and initiate the growth of the next layer. This is easy to understand given our earlier observations. If we regard the top of an island as a terrace of length $L(t)$, then at early times, when the island is small, we have $L^* \gg L(t)$. Thus, the growth of the island proceeds by 'step flow' in the sense that atoms which are deposited on top of the island migrate to the edge of the island, where they are incorporated into the lower layer. As the island grows laterally, however, the condition $L^* \ll L(t)$ is eventually reached. In this case, atoms deposited on top of the island are more likely to encounter one another to form a new island before migrating to the edge of the island. Thus, the next layer begins to form before the current layer is complete and the surface undergoes a gradual and progressive roughening—called *kinetic roughening*—whereby an increasing number of incomplete layers is exposed. The decay of the RHEED oscillations is indicative of this roughening. Kinetic roughening is an intrinsic aspect of the epitaxial growth process and is due to the randomness of the deposition process. Comprehensive discussions of the theory of kinetic roughening and its experimental characterization may be found in the books by Yang *et al.*, (1993) and Barabási and Stanley (1995). A brief discussion will be given in the next section.

4.4 Continuum Equations of Motion

The dynamics of surfaces during epitaxial growth in the multilayer growth regime can often be described by relatively simple evolution equations. These evolution equations typically take the form of partial differential equations for the height of the surface with stochastic noise that accounts for the randomness in the deposition and other processes. The deterministic terms in such equations represent the relaxation of the surface to thermal equilibrium, while the stochastic terms represent fluctuations, especially in the deposition flux, which drive the surface away from equilibrium and are responsible for kinetic roughening. Thus, kinetic roughening is an intrinsically nonequilibrium effect which is caused by the system being driven by the flux from the molecular beam.

4.4.1 Roughening by Random Deposition

Although the stochastic partial differential differential equations that are used to describe kinetic roughening are usually *nonlinear*, there are circumstances when the nonlinearities are so weak that a *linear* model may adequately describe the dynamics of the surface. We consider a simple example. Suppose the dynamics of a growing surface can be described only by deposition of atoms onto the surface, i.e. there *no relaxation processes at all.* The deposition is described by an *average* flux J and a stochastic component $\eta(\boldsymbol{x}, t)$, which accounts for the fact that the deposition is not completely uniform—there are fluctuations in both space and time. In other words, although the deposition is *macroscopically* uniform, there are fluctuations on short space and time scales. These fluctuations are specified in terms of the statistical properties of a fluctuating quantity η. If J is the total average deposition flux, then the fluctuations about this quantity must average to zero:

$$\langle \eta(\boldsymbol{x}, t) \rangle = 0 \tag{4.15}$$

We now suppose that the deposition is totally random, i.e. the deposition of atoms is an uncorrelated process. Such a process is characterized in terms of products if the η:

$$\langle \eta(\boldsymbol{x}, t)\eta(\boldsymbol{x}', t') \rangle = J\delta(\boldsymbol{x} - \boldsymbol{x}')\delta(t - t') \tag{4.16}$$

This expression says that the *average* of two deposition events vanishes unless these events are the same, i.e. there is a total absence of any correlations in the deposition. Averages of higher products of the η are taken to vanish. Fluctuations described only by (4.15) and (4.16) are called *Gaussian white-noise.* The continuum equation of motion for a surface which evolves only by such random deposition is

$$\frac{\partial h}{\partial t} = J + \eta(\boldsymbol{x}, t) \tag{4.17}$$

where $h(\boldsymbol{x}, t)$ is the height of the surface at the position $\boldsymbol{x}$ at time t.

The roughness of a surface is usually characterized by the surface width $W(t)$, which is defined by the standard deviation of the height across the surface:

$$W(t) = \Big[\langle h^2(\boldsymbol{x}, t)\rangle - \langle h(\boldsymbol{x}, t)\rangle^2\Big]^{1/2} \tag{4.18}$$

Thus, W can be calculated from the equation of motion for the dynamics of a surface. For the model in (4.17) this is an especially easy task, since the equation is linear, so we can integrate it directly:

$$h(\boldsymbol{x}, t) = h(\boldsymbol{x}, 0) + Jt + \int_0^t \eta(\boldsymbol{x}, s)\,\mathrm{d}s \tag{4.19}$$

where $h(\boldsymbol{x}, 0)$ is the initial condition. For example, for a flat initial surface all the heights are taken to vanish: $h(\boldsymbol{x}, 0) = 0$. Thus, by taking the average of this solution and using (4.15), we obtain

$$\langle h(\boldsymbol{x}, t)\rangle = h(\boldsymbol{x}, 0) + Jt \tag{4.20}$$

that is, the average height at the position $\boldsymbol{x}$ at time t is just the sum of the initial height at $\boldsymbol{x}$ and the average accumulated material, Jt. Thus, the mean evolution of the surface is just the initial profile translated by Jt. In particular, the only roughness of the surface is that which was present initially.

The calculation of $\langle h^2(\boldsymbol{x}, t)\rangle$ proceeds similarly. By using (4.3), (4.15) and (4.16), we obtain

$$\begin{aligned}\langle h^2(\boldsymbol{x}, t)\rangle &= [h(\boldsymbol{x}, 0) + Jt]^2 + \int_0^t \int_0^t \langle \eta(\boldsymbol{x}, s)\eta(\boldsymbol{x}, s')\rangle\,\mathrm{d}s\,\mathrm{d}s' \\ &= [h(\boldsymbol{x}, 0) + Jt]^2 + Jt \end{aligned}\tag{4.21}$$

Substituting (4.20) and (4.21) into (4.18), we find that

$$W(t) \propto t^{1/2} \tag{4.22}$$

i.e. the width of the surface is seen to increase as the square root of the deposition time. In other words, the surface *roughens*. We would expect that if we added any terms that describe relaxation mechanisms to the right-hand side of (4.17) that the surface would not roughen as rapidly. A more comprehensive of models for surface roughening may be found in the book by Barabási and Stanley (1995).

4.4.2 The Villain Equation

While this discussion provides a useful background to stochastic equations of motion, there is still the question of how to represent real epitaxial processes within this framework. A useful discussion of this point has been given by Villain (1991), who suggests that the long time and long wavelength morphological evolution of a growing epitaxial film is best described by the following nonlinear stochastic partial differential equation:

$$\frac{\partial h}{\partial t} = \nu \nabla^2 h + \lambda (\nabla h)^2 + K \nabla^2 (\nabla^2 h) + \sigma \nabla^2 (\nabla h)^2 + F + \eta. \qquad (4.23)$$

Here, F is net average deposition flux (average deposition flux minus average desorption flux) and $\eta(x,t)$ is a Gaussian random variable with zero mean and shot-noise-type covariance. This formula contains the Edwards-Wilkinson ($\lambda = K = \sigma$=0) and the Kardar-Parisi-Zhang equations ($K = \sigma = 0$) as special cases. One expects that the presence or absence of the various terms in Eq. (4.23) depends on the presence or absence of various physical processes. For example, there is broad agreement that, during growth, the terms proportional to ν and λ are present whenever thermal desorption is operative and that the terms proportional to K and σ can arise from surface diffusion.

On the other hand, in the absence of desorption, it is fair to say that the status of the coefficient ν remains an unsettled issue. Villain (1991) argues that $\nu \neq 0$ if asymmetric energy barriers are present in the vicinity of step edges. In that case, one generates the Laplacian term in (4.23) with a coefficient proportional to the flux F. On the basis of simulation studies that employ Metropolis-type kinetics, it has been claimed (Yan, 1992; Kessler *et al.*, 1992) that similar behavior is found for ν even for pure surface diffusion without special step edge barriers. Evans and Kang (1991, 1992) suggest that this term arises whenever there is 'lateral coupling due to realistic adsorption site geometries and deposition dynamics'. As an example of the latter, they cite so-called *knock-out* processes that involve the replacement of an existing step edge atom by a freshly deposited atom and thus have the effect of local downward relaxation. The resolution of this point is of some theoretical interest because the asymptotic scaling behavior of the surface roughness will be dominated by this term (and λ) if present. On the other hand, experiments likely will be dominated by crossover effects (Tang and Nattermann, 1991: Das Sarma *et al.* 1992) so that there is a need to at least estimate the sign and relative magnitude of the various coefficients that enter (4.23).

References

A.R. Avery, H.T. Dobbs, D.M. Holmes, B.A. Joyce, and D.D. Vvedensky, *Phys. Rev. Lett.* **79**, 3938 (1997).

G.S. Bales and D.C. Chzan, Phys. Rev. B **50**, 6057 (1994).

A.–L. Barabási and H. E. Stanley, *Fractal Concepts in Surface Growth* (Cambridge University Press, Cambridge, 1995).

M.C. Bartelt and J.W. Evans, *Phys. Rev. B* **46**, 12 675 (1992)

M.C. Bartelt and J.W. Evans, *Phys. Rev. B* **54**, R17 359 (1996).

M.C. Bartelt, L.S. Perkins, and J.W. Evans, *Surf. Sci.* **344**, L1193 (1995).

V. Bressler–Hill, S. Varma, A. Lorke, B.Z. Nosho, P.M. Petroff, and W.H. Weinberg, *Phys. Rev. Lett.* **74**, 3209 (1995).

W.K. Burton, N. Cabrera and F.C. Frank, *Philos. Trans. Roy. Soc. London*, Sect. A **243**, 299 (1951).

D.D. Chambliss and K.E. Johnson, *Phys. Rev. B* **50**, 5012 (1994).

S. Das Sarma, Z.–W. Lai and P.I. Tamborenea, *Surf. Sci.* **268**, L311 (1992).

H.C. Kang and J.W. Evans, *Phys. Rev. A* **44**, 2335 (1991).

H.C. Kang and J.W. Evans, *Surf. Sci.* **269–270**, 784 (1992).

R. Ghez and S.S. Iyer, *IBM J. Res. Develop.* **32**, 804 (1988).

B. Müller, L. Nedelmann, B. Fischer, H. Brune, and K. Kern, Phys. Rev. B **54**, 17 858 (1996).

C. Ratsch, A. Zangwill, P. Šmilauer, and D.D. Vvedensky, *Phys. Rev. Lett.* **72**, 3194 (1994).

C. Ratsch, P. Šmilauer, A. Zangwill, and D.D. Vvedensky, *Surf. Sci.* **329**, L599 (1995).

M. Schroeder and D.E. Wolf, *Phys. Rev. Lett.* **74**, 2062 (1995).

M. von Smoluchowski, Phys. Z. **17**, 557, 585 (1916).

J.A. Stroscio and D.T. Pierce, *Phys. Rev. B* **49**, 8522 (1994).

L.-H. Tang and T. Nattermann, *Phys. Rev. Lett.* **66**, 2899 (1991).

J.A. Venables, G.D.T. Spiller and M. Hanbucken, *Rep. Prog. Phys.* **47**, 399 (1984).

J. Villain, *J. Phys. I* **1**, 19 (1991).

T. Viscek and F. Family, *Phys. Rev. Lett.* **52**, 1669 (1984).

H. Yan, *Phys. Rev. Lett.* **68**, 3048 (1992).

H.–N. Yang, G.–C. Wang and T.–M. Lu, *Diffraction from Rough Surfaces and Dynamic Growth Fronts* (World Scientific, Singapore, 1993).

A. Zangwill and E. Kaxiras, *Surf. Sci.* **326**, L483 (1995).