Methods for Calculating Rates of Transitions with Application to Catalysis and Crystal Growth

Graeme Henkelman

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

University of Washington

2001

Program Authorized to Offer Degree: Chemistry

University of Washington

Graduate School

This is to certify that I have examined this copy of a doctoral dissertation by

Graeme Henkelman

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Chair of Supervisory Committee:

_____

Hannes Jónsson

Reading Committee:

_____

Hannes Jónsson

_____

William Reinhardt

_____

Charles Campbell

Date: _____

University of Washington

Abstract

Methods for Calculating Rates of Transitions with
Application to Catalysis and Crystal Growth

by Graeme Henkelman

Chair of Supervisory Committee

Professor Hannes Jónsson
Chemistry

An important problem in theoretical chemistry is the calculation of reaction rates. This is challenging because most interesting chemical reactions occur much slower than the vibrational period of atoms. Even though it is in principle possible to directly simulate the classical dynamics of atomic systems, reactions of interest can almost never be observed this way. Fortunately it is possible to focus on just those slow events in which the system passes from one basin on the potential energy surface to another. If the bottleneck, or transition state, through which the system must pass in order to undergo a transition can be found, transition state theory can be used to calculate the reaction rate. For tightly bound systems the problem of finding transition states reduces to the problem of finding saddle points on the potential surface.

Two methods are developed for finding saddle points efficiently – the dimer method which is used when only the initial state of the reaction is known and the climbing image nudged elastic band method which is used when the final state is also known. Both methods only require first derivatives of the potential energy, and can therefore be used with *ab-initio* calculations like density functional theory. The methods are applied to a variety of systems including the dissociation of methane on the Ir(111) surface and the diffusion of aluminum atoms on the Al(100) surface.

In many cases, finding the rate of one process is not enough, rather, many transitions are required in order to see interesting phenomena. For these types of problems, the dimer method is used to find many different possible processes available to the system. When many processes are found, they are weighted by their rates at the simulation temperature, and one is chosen through which the system is advanced. In this way, the dimer method is combined with kinetic Monte Carlo, so that the traditional limitation of having to know all possible processes *a priori* is relaxed. This method for determining the long time scale dynamics of a system was applied to island formation and growth of the Al(100) surface.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

AB-INITIO: From first principles. This expression can mean different things in different fields. In the field of quantum chemistry, it means that a calculation was done by solving Schrödinger's equation for the electrons in the system without using any empirical data. In solid state physics the term also includes methods such as density functional theory, which contains approximations that have been fit to experiment.

CINEB: Climbing image nudged elastic band.

CLIMBING IMAGE NUDGED ELASTIC BAND: A modification to the NEB method in which the highest energy image along the band is moved to the saddle point.

CLASSICAL DYNAMICS: The time evolution of Newton's equations for a classical system.

DENSITY FUNCTIONAL THEORY: An approximate method of solving Schrödinger's equation for a system of electrons. The energy and force of the system is calculated from the electron density. DFT scales better than many-body orbital based quantum chemistry methods and can be used with larger systems.

DIMER METHOD: A method for finding saddle points that requires energy and force but not second derivatives. The dimer method can be used to find many different saddle points on the potential energy rim surrounding a potential minimum.

DFT: Density functional theory.

EAM: Embedded atom method.

EMBEDDED ATOM METHOD: A form of empirical potential function which quite accurately describes some metals including aluminum, nickel, silver and copper.

HARMONIC TRANSITION STATE THEORY: A simplified form of transition state theory in which the potential is assumed to be of harmonic form both at the minimum and at the saddle point. This is a good approximation at low enough temperature, unless quantum effects become important. It typically works well for metals at room temperature.

HTST: Harmonic transition state theory.

HESSIAN MATRIX: The matrix of force constants (second derivatives of the potential energy). When this matrix is divided by the masses of the atoms, the eigenvectors are the normal modes, and the eigenvalues are the square of the normal mode frequencies.

HYPERDYNAMICS: A method developed by Art Voter for accelerating the classical dynamics of a system. A bias is added to the potential energy which fills in the wells but does not affect transition states.

MEP: Minimum energy path.

MINIMUM ENERGY PATH: A path between two points on a potential surface of the lowest possible energy. This path follows the direction of steepest descent.

NEB: Nudged elastic band.

NUDGED ELASTIC BAND: A method for finding the minimum energy paths between two points on a potential surface. A path of discrete images of a system are connected by springs (elastic band) and allowed to collectively relax. The 'nudging' refers to the fact that the spring forces act only along the band, and the potential forces act only perpendicular to the band. Typically this method is used to find the MEP and saddle point(s) between two potential minima.

PES: Potential energy surface.

POTENTIAL ENERGY SURFACE: Each point in configuration space represents one configuration or position of the atoms in the system. For this position, there is a potential energy. The potential energy surface is the surface defined by the value of the potential energy at each point in configuration space.

SADDLE POINT: A point on a potential surface at which the force is zero and at which there is one negative curvature or unstable mode in the Hessian matrix.

TRANSITION STATE: A bottle neck region which a system must cross in order to undergo a transition from a given initial state. The transition state has dimension one less than the full system.

TRANSITION STATE THEORY: A theory for calculating the rate at which a system leaves a given initial state (potential energy basin) though a bottle neck region.

TAD: Temperature assisted dynamics.

TEMPERATURE ASSISTED DYNAMICS: A method developed by Mads Sørensen and Art Voter for accelerating the classical dynamics of a system. The system is heated and classical dynamics are run during which the system is confined to a single potential basin. The time at which transitions are observed are extrapolated back to the low temperature so that the correct dynamics can be determined.

TST: Transition state theory.

ACKNOWLEDGMENTS

Most of all, I want to acknowledge Hannes Jónsson. Hannes has been an excellent advisor and good friend. This work has been a true collaboration with him. Beth Reyburn has also been very important to me during this time both for her emotional support and because she proof read and added her aesthetic touch to each paper. Blas Uberuaga has been an excellent person to work with, both during the many computing issues we dealt with and our boron clustering project. Gísli Jóhannesson was my congenial office mate and a coauthor on our saddle point finding review chapter. Finally, I want to thank Art Voter for having me down to Los Alamos on several occasions for many excellent discussions and some fantastic mountain biking.

Chapter 1

INTRODUCTION

This thesis is a compilation of the work I did with my advisor, Hannes Jónsson, during my PhD in physical chemistry at the University of Washington. Each of the following chapters contains a paper written with Hannes and in two cases, with an additional author, either Blas Uberuaga (Ch. 4) or Gísli Jóhannesson (Ch. 5). Each chapter is both self contained and is a piece of the larger story about calculating reaction rates with transition state theory. The introduction contains some general background and a brief description of each chapter.

1.1 Transition State Theory

In order to understand the important role of transition state theory in theoretical chemistry [1], it is first necessary to have a picture of what is going on in chemical or solid state systems at the atomic scale. At the scale of $1\text{Å} = 1^{-10}\text{m}$, solid systems are composed of atoms. The atoms have some thermal energy which causes them to vibrate and bump into each other. In a typical solid, atoms oscillate on a time scale of femtoseconds (10^{-15}s). Most of the oscillations do not lead to interesting chemical reactions. The atoms typically move in a somewhat chaotic way around their average positions. Only in very rare moments will circumstances conspire to make a reaction occur.

This can be illustrated with a practical example. Consider a computer chip, which is primarily made of up silicon. The atoms in the chip are held in a lattice by the strong silicon bonds. At room temperature, the silicon atoms move around their lattice sites with a small amount of thermal energy. But, nothing ensures that the atoms stay put indefinitely. Energy is free to move from one atom to another, and sometimes a particular atom or group of atoms will get a relatively large amount of energy. If this energy is directed in just the right degrees of freedom, it can cause bonds to break, and for example, an atom

can move to a new lattice site within the solid. Old glass windows show the results of this phenomena. Over time, as the atoms move around, they prefer positions in which they are lower to the ground so that the pane of glass becomes thicker at the bottom than the top. Another example is a block of ice in a freezer. Every once in a while an ice molecule gets enough energy to break free of the block so that after some time the ice in the block will be deposited on the walls of the freezer.

Processes of this type, and most other interesting reactions, take place on a timescale which is much longer than the vibrational timescale of femtoseconds. Even a reaction that a person would consider quick, say taking on average some milliseconds, is incredibly slow compared with vibrations. The atoms will vibrate on average 10^{12} times in between such reactions. Putting this in practical terms, if a person were to try to simulate the motion of the atoms on a computer, using the simplest description of the forces on the atoms, the calculation would take thousands of years before a single event could be seen.

Transition state theory (TST) is a concept that provides some relief from this problem. Within TST only those rare events that lead to reactions are described, and they are described in a statistical way. The cost is that the details of the atomic vibration are lost. TST can only be applied in certain circumstances. These are best described with reference to the potential energy surface (PES).

The PES is a surface defined by the potential energy at each point in configuration space. In general this is a high dimensional surface which is hard to visualize, so it can be useful to consider a two dimensional surface like that in Fig. 1.1. This figure can be thought of as a topographical map of the landscape. This analogy is very persistent, so that motion on the landscape is often described as climbing up or sliding down the potential. The system can be thought of as a point or a small ball rolling on this landscape.

If the system were isolated and held at a constant energy, the dynamics of the system would correspond to the ball rolling undisturbed indefinitely. If the system is started in basin A for example, and given some initial energy but not enough so that it can move from basin A to another basin, then the system will oscillate in basin A forever. TST does not describe constant energy systems like this. Instead, it can predict the rate at which the system moves from one basin to another when held at a constant temperature. Temperature in the ball analogy can be thought of as small random kicks, where the average strength
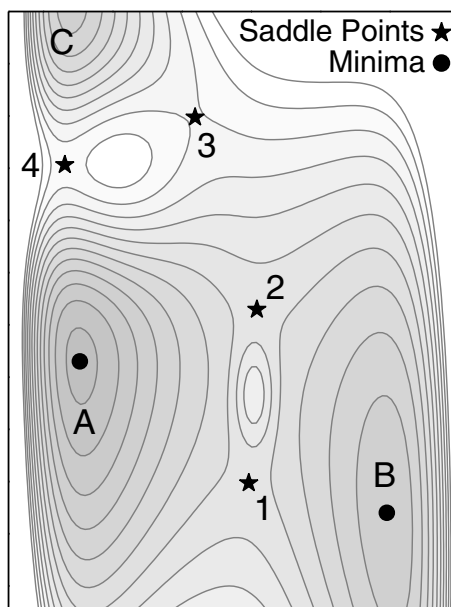
Figure 1.1: A two dimensional potential landscape. Basins A, B and C are shown in darker gray with their local minima (•) at the lowest point in the basin. Saddle points (⋆) 1 through 4 are the lowest points on the potential landscape which connect adjacent basins. In other words, descent paths on either side of a saddle point lead to different but adjacent basins.

of the kicks corresponds to the temperature. There is always some chance that the system will be kicked out of a basin.

An important aspect of TST is that it is used to calculate the rate of rare events. If the system in basin A were given a large thermal energy with respect to the height of the saddle points (1-4), it would move quickly between basins, and TST should not be used to calculated the rates of these transitions. If, on the other hand, the thermal energy is small compared to the height of the barriers, then such an event will happen only rarely and TST can be used to calculate its rate. Another way of saying this is that the system must spend a long time on average in each basin before moving to an adjacent basin.

The hardest part of calculating a rate with TST is identifying bottle neck regions or transition states. A transition state is a surface through which the system must pass to get to another basin. In tightly bound solid systems rates can be found using harmonic transition state theory (hTST) and the problem of finding transitions states reduces to the problem of finding saddle points on the potential surface. In the two dimensional potential shown in Fig. 1.1, finding saddles can be as simple as making the figure, and pointing to

the saddle points. But in general, systems of interest have many more dimensions (three for each atom in the system), so that it is impossible to visualize the entire PES. In order to apply hTST to these systems, it is important to have efficient methods for finding saddle points.

1.2   Saddle Point Finding Methods

Chapters 2-5 are primarily about methods to find saddle points. Chapter 2, which is taken from Ref. [2], describes a technique called the dimer method. Chapters 3 from Ref. [3], and Ch. 4 from Ref. [4], describe some extensions to a method called the nudged elastic band (NEB) method, which was developed in the Jónsson group several years ago. These two methods are compared with a collection of other methods from the literature in Ch. 5 which is taken from Ref. [5].

1.2.1   The Dimer Method

The dimer method was designed to find saddle points which lead from a known basin (minimum) to adjacent basins. A saddle point is the highest energy point on the minimum energy path connecting two basins on the potential energy surface. Or, put another way, the saddle point is the highest energy point that must be crossed to get from one basin to another while staying as low in energy as possible. With respect to landscapes, the saddle point is a mountain pass. Each saddle point corresponds to a process by which the system can undergo a transition. Once a saddle point is found, hTST (see Sec. 1.1) can be used to calculate the rate of the process.

There are several important features of the dimer method. The first is that the dimer method only needs to know from which minimum it is starting. This is in contrast to most methods which require both the initial and the final state of a process in order to find a connecting saddle point. The dimer method can address the question: given that a system is in a particular state, by what process can the system evolve and what are the rates of these possible processes? A second important feature of the dimer method is that it only requires first derivatives of the potential energy, ie. the force on the system. For many simple empirical potentials it is possible to directly calculate second derivatives of

the potential energy. But, for *ab-initio* calculations this is very expensive because second and higher derivatives of the potential require taking derivatives of the high dimensional electronic wavefunction. The fact that the dimer method only requires the energy and the force on the system means that it can be used in *ab-initio* calculations such as density functional theory.

Although the second derivatives (Hessian) matrix is not needed for the dimer method, the lowest curvature mode is. A saddle point is a point at which the potential has been maximized along the direction of lowest curvature, and minimized in all other directions. The core of the dimer method is a procedure, first proposed by Voter[6], in which the lowest curvature mode is found iteratively. The curvature along some direction is estimated with a finite difference scheme by evaluating the potential at two points which are close together. If this dimer is allowed to freely rotate according to the forces on it, it will align itself along the lowest curvature mode. When there is a good estimate of this direction, the dimer is moved up the potential along the lowest curvature mode, and down the potential in all other directions. When the dimer comes to rest, it is at a maximum along the lowest curvature mode (the reaction coordinate), and a minimum in all other directions, which is the definition of a first order saddle point.

1.2.2 The Nudged Elastic Band Method

The nudged elastic band (NEB) method is in many ways complimentary to the dimer method. Instead of just finding saddle points, the NEB method finds the entire minimum energy path (MEP) between two minima. Where as the dimer can find saddle points from an initial state, the NEB can find a reaction path between an initial state and a known final state. The NEB works by creating a string or band of images connecting the initial and final states. These images are connected by springs to keep them equally spaced. The path of springs acts like an elastic band. The images on the band feel the forces due to the springs, but they also feel the forces due to the potential so that each image tries to minimize its potential energy. An important part of the method is the nudging or force projection scheme. The spring forces are only allowed to act along the band so that constant image spacing is ensured, and the potential forces are only allowed to act in all directions perpendicular to the band, ensuring that the band comes to rest on the MEP between the

minima. The NEB method was developed in the Jónsson group several years ago.

Two improvements to the basic NEB method are described in Ch. 3 and Ch. 4. In some cases, it was noticed that the NEB method would develop kinks along the band. Instead of forming a smooth path between the initial and final states it would develop oscillations in which one image would be on one side of the MEP, and the next would be on the other. The problem was dealt with in the original method by relaxing some of the nudging constraints on the springs, so that when kinks were detected, the band would have a tendency toward the shortest path between initial and final states. This solution is not satisfactory because it adds a parameter to the method and it caused the NEB to find a path which deviates from the real MEP. Chapter 3 describes a better way of defining the direction along the band which removes the tendency for the NEB to develop kinks, and removes the extra parameter.

Another downside to the NEB is that while it finds a MEP, it does not find the exact saddle point along the path, which is used to calculate the activation energy of the process. It is possible to put down many images in the band so that the saddle point can be found by interpolation between the images, but having many images adds to the cost. Chapter 4 describes a way of moving the highest energy image on the band right to the saddle point. In the spirit of the dimer method, this image is moved up the potential along the band and down the potential in all other directions. This image is also freed of the spring forces so that it can move exactly to the saddle point. This modification, called the climbing image NEB, extends the NEB so that saddle points can be found with less images in the band.

These saddle point finding methods were developed with empirical potentials, for example an embedded atom method (EAM) potential to describe aluminum, because empirical potentials can be evaluated quickly. But, as mentioned, the methods were developed to be efficient and only rely on the energy and force of the system so they can be used with *ab-initio* calculations. Both the dimer method and the climbing image NEB were implemented in a density function theory code so that rates could be calculated in complicated systems more accurately than is possible with empirical potentials.

1.3   Methane Dissociation on Ir(111) using Density Functional Theory

Chapter 6, which is taken from Ref. [7], is a short description of a calculation of the dissociation of methane on the (111) surface of iridium. Iridium is an expensive transition metal which is an efficient catalyst for this reaction. The (111) indices describe which way the crystal is cut to form the surface upon which the methane reacts. Ir(111) is a close packed surface with the surface atoms arranged in a triangular lattice. The methane on iridium reaction was chosen for several reasons. First of all, it is an important reaction for the hydrocarbon industry. There are as many known methane reserves in the earth as all other hydrocarbons combined. The problem with methane is that it is a gas which is prohibitively expensive to liquefy, and therefore hard to transport. At most oil fields, methane (natural gas) is simply burned. Methane is also an important resource because oil reserves are being consumed at an incredible rate. Personally, I think conservation and energy efficiency should be the highest priority, but it would also be good to consider alternative hydrocarbon sources such as methane. In order to use methane as a hydrocarbon source, other than for burning, it has to be chemically converted from its gaseous form to larger hydrocarbon liquids or plastics. The most expensive part of this process is the initial activation of methane (CH_4), in which one hydrogen is removed as it dissociates onto a catalyst.

Another important reason for choosing this system is that the methane dissociation reaction is fairly well characterized experimentally, so the calculation can be used to test the theoretical methods. There is a particularly nice relationship between experiment and theory in this field. The experimentalist can more easily measure the right rate of a particular reaction, but it can be very hard to experimentally determine the mechanism of the reaction. Calculations on the other hand can determine mechanisms of a reaction, but it is important to compare with a known experimental rate to make sure the right mechanism is being calculated, and at an appropriate level of theory. With respect to the methane dissociation reaction, there are two experimental groups who have different interpretations of the molecular mechanism, and in particular why at low CH_4 incident energies the rate of reaction can increase with decreasing incident energy. This is a surprising result because in general if reactants have more energy, they can react faster. In the Mullins group [8],

the dissociation experiments are interpreted to suggest that a methane molecule incident on an Ir(111) surface first enters a trapping mediated state. That is, the molecule can be physically adsorbed to the surface while still intact as CH_4. They argue that if the incident molecule approaches the surface with less energy, it is more likely to get trapped at the surface. This gives the methane molecule far more time to react which makes the reaction more likely than if the molecule were to have a higher incident energy. A different interpretation is held by the King group [9] which argues that if a molecule approaches the surface with less energy, it has more time to get oriented in the right way so that it can pass over the saddle point region. If the molecule has too much energy, it has less time to get positioned over the right part of the surface or at the right orientation so the reaction can occur with the least activation energy. This, they argue, is why the reaction is more likely to occur as the incident energy is lowered.

The DFT calculations were shown to match experiment when interpreted with the trapping mediated model of the Mullins group. But, there was another interesting result from the calculation. DFT calculations are very expensive, and they get more expensive very quickly as the system gets larger. It is good practice to start a calculation with a small system, and then increase the size to see if the results are converged. For the methane dissociation reaction, an initial system was chosen with 4 layers in the iridium surface with 4 atoms per layer. This is a typical system size for results currently published in the literature. With this small system, the saddle point energy was an order of magnitude higher than experiment. As the system size was increased, to a maximum of 6 layers with 12 atoms per layer, the barrier height dropped to the experimental value. One reason for this drastic change with system size is that the metal surface was not just playing a spectator role in the reaction. A conventional way to think about a catalyst is that it provides a good electronic environment for a reaction to take place, but the atoms in the catalyst do not really participate in the reaction. For the methane dissociation reaction, this is not the case. As the methane approaches the iridium surface, one of the metal atoms lifts out of the surface by as much as 0.5Å to greet the incoming molecule. The dissociation reaction takes places on this raised surface atom, and then the products move back into the surface. At least one reason that such a large system was required to accurately model this reaction is that the surface has to have enough atoms to absorb the elastic strain energy caused by

this large displacement.

1.4   Long Time Scale Dynamics on the Al(100) Surface

In the final chapter the dimer method is combined with traditional kinetic Monte Carlo to make a hybrid method which is used to calculate the long time scale dynamics of a system. The emphasis in this chapter is somewhat orthogonal to that of the methane dissociation study in which a single rate is calculated as efficiently and accurately as possible. Chapter 7, which is taken from Ref. [10], focuses on problems in which the property of interest can not be determined from one process, but instead requires many sequential processes. An example of this is the growth of a surface by deposition of atoms. Depending upon the surface temperature and the rate at which atoms are deposited on the surface, the surface can grow smoothly as a single crystal, or it can form a rough crystal surface in an amorphous solid. If the surface is held at a very low temperature, atoms will stick to the surface where they land, and a rougher surface is expected. As the surface temperature is increased, the atoms on the surface have a chance to rearrange and get into low energy lattice sites and a smoother surface can form. Under typical experimental conditions, the collective property of rough or smooth, crystalline or amorphous, can only be determined from long time scale simulations.

In a traditional kinetic Monte Carlo (KMC) simulation a list of all possible rates in the system needs to be generated before the simulation starts. In a metal growth simulation for example, all possible surface processes and their rates have to be known in advance. This is a challenge for even simple processes such as an adatom hoping on the Al(100) surface, which is studied with the dimer method in Ch. 2. Furthermore, if the adatom is not on a perfectly flat surface, there will be a different rate depending upon the environment of the adatom. The number of processes grows exponentially as more neighbors are considered. To complicate the situation even more, in the simulation presented in Ch. 7, aluminum atoms on Al(100) can undergo multi-atom exchange processes in which several atoms can move at once. A further problem with KMC is that it is all but impossible to treat systems in which the atoms do not always minimize into lattice sites. This occurs occasionally in the Al(100) growth simulations, but it happens all the time in composite materials or when

one material is grown on another. The point is that it can be very hard to anticipate and make a complete table of all possible processes before the simulation is started.

Given that all processes and corresponding rates can be generated, the system can be advanced in time in a straight forward manner. When the system is in some state at a particular time, a table can be made from the complete list of events which contains all possible process that the system can undergo. These processes are weighted by their rate, or the probability of the process occurring with respect to all others, and then one is chosen with a random number generator. This is the Monte Carlo part of the algorithm. The system is moved via the chosen process, and the simulation time is advanced according to the sum of all possible rates. The table of rates is updated for the new configuration and the process is repeated.

The idea of the dimer kinetic Monte Carlo method is quite simple. Instead of trying to determine all possible processes that the system can undergo before the simulation starts, the dimer method can be used to find them as the simulation progresses. The hybrid algorithm works just like KMC, but the table of possible rates available to the system is generated on the fly. From each new basin the system visits, a swarm of dimer searches is sent out looking for saddle points. Each saddle point that connects to an adjacent basin corresponds to a possible process for the system. The energy of the saddle is used to determine the rate of that processes. Then, just as in KMC, one process is chosen, the system is moved according to that process, the clock is advanced, and the method is repeated with a new swarm of dimer searches.

The method is demonstrated with two types of simulations on the Al(100) surface. The first is ripening in which a random collection of atoms on the surface slowly form a large island. Islands of twenty atoms were shown to form at room temperature in a millisecond. As discussed in Sec. 1.1, this is a very long period of time for atomic processes. In the second type of simulation, atoms were deposited onto the surface to form new layers of material. Typically a layer was grown in a millisecond. It was found that aluminum will grow layer by layer on a length scale of tens of atoms even at surface temperatures as low as 30K.

Chapter 2

A DIMER METHOD FOR FINDING SADDLE POINTS ON HIGH DIMENSIONAL POTENTIAL SURFACES USING ONLY FIRST DERIVATIVES

2.1   Abstract

The problem of determining which activated (and slow) transitions can occur from a given initial state at a finite temperature is addressed. In the harmonic approximation to transition state theory this problem reduces to finding the set of low lying saddle points at the boundary of the potential energy basin associated with the initial state, as well as the relevant vibrational frequencies. Also, when full transition state theory calculations are carried out, it can be useful to know the location of the saddle points on the potential energy surface. A method for finding saddle points without knowledge of the final state of the transition is described. The method only makes use of first derivatives of the potential energy and is, therefore, applicable in situations where second derivatives are too costly or too tedious to evaluate, for example, in plane wave based density functional theory calculations. It is also designed to scale efficiently with the dimensionality of the system and can be applied to very large systems when empirical or semiempirical methods are used to obtain the atomic forces. The method can be started from the potential minimum representing the initial state, or from an initial guess closer to the saddle point. An application to Al adatom diffusion on an Al(100) surface described by an embedded atom method potential is presented. A large number of saddle points were found for adatom diffusion and dimer/vacancy formation. A surprisingly low energy four atom exchange process was found as well as processes indicative of local hex reconstruction of the surface layer.

2.2 Introduction

Most atomic scale transitions in the condensed phase, such as chemical reactions and diffusion, are activated processes, i.e. require surmounting a significant energy barrier. While under typical conditions the thermal energy is on the order of $k_B T = 0.025$ eV, the barriers for such transitions are typically on the order of 0.5 eV or higher. A transition that occurs thousands of times per second is so slow on the scale of atomic vibrations that it would typically take several thousands of years of computational time on the fastest modern computer to simulate a classical trajectory with reasonable chance of observing a single transition. As a result, classical dynamics simulations of activated transitions are faced with an impossible time scale problem. It is not possible to observe such transitions by simply simulating the classical dynamics of the system. Arbitrarily raising the temperature of the system can lead to crossover to a different transition mechanisms. The problem is to find a simulation algorithm that can be used to find which transition would occur and at what rate, if the classical dynamics could be simulated for a long enough time.

Within transition state theory (TST) the problem becomes that of finding the free energy barrier for the transition [1]. This is a very challenging problem, especially when the mechanism of the transition is unknown. Within the harmonic approximation to transition state theory (hTST) [11, 12] the problem becomes that of finding the saddle point on the potential energy surface corresponding to a maximum along a minimum energy path that takes the system from one potential energy minimum to another. This is still a difficult problem when dealing with condensed matter systems because of the high dimensionality of the potential energy surface. When both the initial and final states of the transition are known, minimum energy path(s) for the transition can be found quite readily [13] (the problem of making sure the path with lowest activation energy has been found is still a difficult problem). When only the initial state of the transition is known, the problem of finding the relevant saddle point(s) becomes that of navigating in high dimensional space – a very challenging task. If the set of all relevant, low lying saddle points for transitions from a given initial state could be found reliably and the prefactor in the hTST rate constant expression evaluated, then long time activated dynamics of the system could, in principle, be simulated. This may be impossible for all but the simplest systems.

Recently, significant progress has been made towards this goal. In the activation-relaxation technique developed by Barkema and Mousseau, the system is driven from one potential energy basin to another by inverting the component of the force acting on the system along a line drawn from the instantaneous configuration to the initial configuration [14, 15] (or to a trailing image of the system [16]). The new potential energy basin is then accepted or rejected based on Monte Carlo sampling. This has enabled equilibration of supercooled liquids down to much lower temperature than could be achieved with direct classical dynamics simulations. In principle the method could be used to estimate the *rate* of the transitions observed using harmonic transition state theory, but the algorithm does not always take the system through the close vicinity of the saddle point, making the estimate of the activation energy uncertain. Furthermore, there is no guarantee the activation-relaxation technique will give the same transition as a long classical trajectory, i.e. the transition with the lowest saddle point.

A very different approach to long time simulations has been developed by Voter, the socalled Hyperdynamics method [17, 6]. There, a classical trajectory is generated for a modified potential with a reduced well depth. The activated transitions are thus made more probable and may be observed during the short time interval accessible by classical dynamics simulations. The relative rate of different transition mechanisms is preserved in the modified system (within the transition state theory approximation) so the Hyperdynamics trajectory should reveal the most probable activated transitions. It can, futhermore, provide an estimate of the transition rate within the full, anharmonic transition state theory approximation. The Hyperdynamics trajectory simulation still requires, in general, a large number of force evaluations, more than can be handled at the present time with *ab initio* methods, and the acceleration of the transitions becomes smaller as the system gets larger.

Powerful methods have been developed for climbing up potential energy surfaces from minima to saddle points in *ab initio* calculations of molecules. These methods have become part of the standard tool kit for molecular calculations, but they require the evaluation and inversion of the Hessian matrix at each point along the search so as to find the local normal modes of the potential energy surface. The strategy of following local normal modes to find saddle points was apparently first described by Crippen and Scheraga [18], and later by Hilderbrandt [19]. In these early algorithms, a small step is taken up the potential

along a particular mode, followed by a step towards lower potential energy along all other modes. In the early 1980's, these methods were replaced by quasi-Newton methods, in which the eigenvalues of the Hessian matrix are shifted to ensure that the potential is maximized along one chosen mode and minimized along all others. The shift parameters, or Lagrange multipliers, were introduced by Cerjan and Miller [20] and later modified by Simons et al. [21, 22], and by Wales [23]. A summary of the early developments is given in Ref. [24]. These methods have been used extensively in *ab initio* calculations of molecules and empirical potential calculations of atomic and molecular clusters [25, 26]. We will refer to these methods collectively as mode following methods. They are derived by expanding the potential in a local quadratic form, and selecting one of the local harmonic modes as the direction for the climb. If the softest mode is chosen, this is analogous to walking up the slowest ascent of a valley. This does not necessarily lead to a saddle point (as will be illustrated in an example below), so an important property of these methods is the ability to search for a saddle point along different orthogonal modes leading away from a given initial configuration. But, since the mode following methods require the evaluation and inversion of the second derivative Hessian matrix, they scale poorly with the number of degrees of freedom in the system. Furthermore, second derivatives are only available at rather low levels of *ab initio* calculations, and are also not available in plane wave based density functional theory (DFT) calculations. The dimer method presented here captures the most important qualities of the mode following algorithms, while using only first derivatives of the potential energy.

An illustration of the importance of having a method that can be used to systematically identify saddle points leading from a given initial state, is the discovery by Feibelman in 1990 that an Al adatom does not diffuse on the Al(100) surface by repeated hops from one site to another, as had previously been assumed, but rather by a concerted exchange mechanism involving concerted displacement of two atoms [27]. This illustrates well how the preconceived notion of a transition mechanism can be incorrect even for a simple system. With the rapid increase in computational power and increased sophistication of simulation software, the complexity of simulated systems has increased greatly. For many systems that are actively being studied, even with *ab initio* methods, it is difficult and, in any case risky to simply guess what the transition mechanism is.
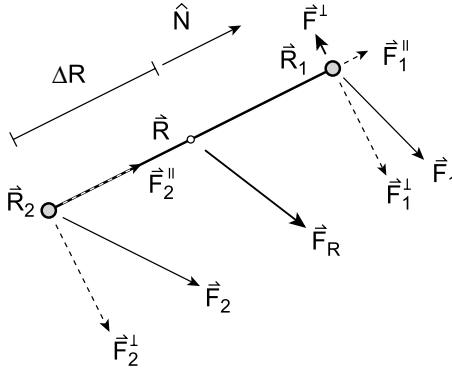
Figure 2.1: Definition of the various position and force vectors of the dimer. The rotational force on the dimer, $\mathbf{F}^{\perp}$, is the net force acting on image 1 perpendicular to the direction of the dimer.

2.3 The Dimer Method

The method presented here for finding saddle points involves working with two images (or two different replicas) of the system. We will refer to this pair of images as the 'dimer'. If the system has n atoms, each one of the images is specified by $3n$ coordinates. The two replicas have almost the same set of $3n$ coordinates, but are displaced slightly by a fixed distance. The saddle point search algorithm involves moving the dimer uphill on the potential energy surface, from the vicinity of the potential energy minimum of the initial state up towards a saddle point. Along the way, the dimer is rotated in order to find the lowest curvature mode of the potential energy at the point where the dimer is located. The strategy of estimating the lowest curvature mode at a point without having to evaluate the Hessian matrix was presented by Voter in his Hyperdynamics method. There, it is used to construct a repulsive bias potential so as to accelerate classical dynamics of activated processes [6]. In the method presented here, we use his dimer strategy to make the search for saddle points more efficient.

2.3.1 Forces and Energies

The dimer, depicted in Fig. 2.1, is a pair of images separated from their common midpoint $\mathbf{R}$ by a distance ΔR. The vector $\hat{\mathbf{N}}$ which defines the dimer orientation is a unit vector pointing from one image at $\mathbf{R}_2$ to the other image at $\mathbf{R}_1$. When a transition state search is launched from an initial configuration, with no prior knowledge of what $\hat{\mathbf{N}}$ might be, a

random unit vector is assigned to $\hat{\mathbf{N}}$ and the corresponding dimer images are formed

$$\mathbf{R}_1 = \mathbf{R} + \Delta R \hat{\mathbf{N}} \quad \text{and} \quad \mathbf{R}_2 = \mathbf{R} - \Delta R \hat{\mathbf{N}}. \tag{2.1}$$

Initially, and whenever the dimer is moved to a new location, the forces acting on the dimer and the energy of the dimer are evaluated. These quantities are calculated from the energy and the force $(E_1, \mathbf{F}_1, E_2, \text{ and } \mathbf{F}_2)$ acting on the two images. The energy of the dimer $E = E_1 + E_2$ is the sum of the energy of the images. The energy and the force acting on the midpoint of the dimer are labeled as E_0 and $\mathbf{F}_R$ and are calculated by interpolating between the images. The force $\mathbf{F}_R$ is simply the average force $(\mathbf{F}_1 + \mathbf{F}_2)/2$. The energy of the midpoint is estimated by using both the force and the energy of the two images. A relation for E_0 can be derived from the finite difference formula for the curvature of the potential C along the dimer.

$$C = \frac{(\mathbf{F}_2 - \mathbf{F}_1) \cdot \hat{\mathbf{N}}}{2 \, \Delta R} = \frac{E - 2E_0}{(\Delta R)^2}, \tag{2.2}$$

E_0 can be isolated from this expression in terms of the known forces on the images

$$E_0 = \frac{E}{2} + \frac{\Delta R}{4} \left(\mathbf{F}_1 - \mathbf{F}_2 \right) \cdot \hat{\mathbf{N}}. \tag{2.3}$$

It should be emphasized that all the properties of the dimer are derived from the forces and energy of the two images. There is no need to evaluate energy and force at the midpoint between the two images. This is important for minimizing the total number of force evaluations required to find saddle points. An additional benefit of this strategy is that the method can be efficiently parallelized over two processors, the energy and force on each image being calculated on a separate processor. For *ab initio* calculations, in which force evaluations typically take a very long time compared with communication time, the execution time for each transition state search is effectively halved if two processors are used.

2.3.2 Rotating the Dimer

Each time the dimer is displaced, it is also rotated with a single iteration towards the minimum energy configuration. The practicality of the dimer method relies heavily on
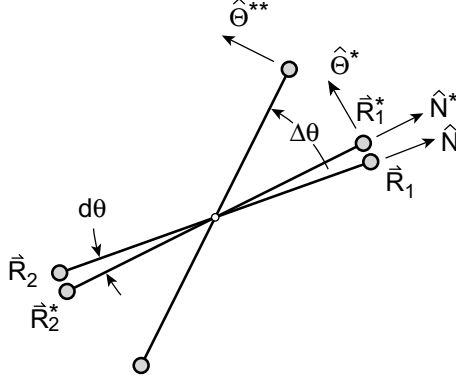
Figure 2.2: Definition of the various quantities involved in rotating the dimer. All vectors are in the plane of rotation. The dimer is first rotated about a small angle $d\Theta$ to give a finite difference estimate of F' [given by Eq. (2.5)]. The dimer is then rotated by a calculated angle $\Delta\Theta$ [given by Eq. (2.13)] to zero the force within the plane of rotation.

using an efficient algorithm for the rotation. Minimizing the dimer energy, E, is equivalent to finding the lowest curvature mode at $\mathbf{R}$. The energy, E_0, at the fixed midpoint of the dimer is constant during the rotation. Since ΔR is also constant, Eq. (2.2) shows that the dimer energy, E, is linearly related to the curvature, C, along the dimer. Therefore, the direction which minimizes E is along the minimum curvature mode at the midpoint $\mathbf{R}$.

Modified Newton Method for Rotation within a Plane

The dimer rotation will first be discussed in the context of a modified Newton method. In the next section, the method will be extended to incorporate also a conjugate gradient approach. The dimer is rotated along the rotational force, $\mathbf{F}^{\perp} = \mathbf{F}_1^{\perp} - \mathbf{F}_2^{\perp}$, where $\mathbf{F}_i^{\perp} \equiv \mathbf{F}_i - \left(\mathbf{F}_i \cdot \hat{\mathbf{N}} \right) \hat{\mathbf{N}}$ for $i = 1, 2$. The rotational force is taken to be the net force acting on image 1 (see Fig. 2.1). The rotation plane is spanned by $\mathbf{F}^{\perp}$ and the dimer orientation $\hat{\mathbf{N}}$. It is useful to define a unit vector, $\hat{\mathbf{\Theta}}$, within the plane of rotation, perpendicular to $\hat{\mathbf{N}}$. For the modified Newton method, $\hat{\mathbf{\Theta}}$ is just a unit vector parallel to $\mathbf{F}^{\perp}$. The vectors $\hat{\mathbf{\Theta}}$ and $\hat{\mathbf{N}}$ form an orthonormal basis which spans the rotation plane. Given an angle of rotation, $d\theta$, image 1 moves from $\mathbf{R}_1$ to $\mathbf{R}_1^{\star}$ (see Fig. 2.2)

$$\mathbf{R}_1^{\star} = \mathbf{R} + \left(\hat{\mathbf{N}} \cos d\theta + \hat{\mathbf{\Theta}} \sin d\theta \right) \Delta R. \tag{2.4}$$

After image 1 is moved to the new point $\mathbf{R}_1^\star$, the new dimer orientation $\hat{\mathbf{N}}^\star$ is calculated, and image 2 is positioned at $\mathbf{R}_2^\star$ according to Eq. (2.1). The forces $\mathbf{F}_1^\star$, $\mathbf{F}_2^\star$, and $\mathbf{F}^\star = \mathbf{F}_1^\star - \mathbf{F}_2^\star$ are then computed. A scalar rotational force $F = \mathbf{F}^\perp \cdot \hat{\boldsymbol{\Theta}}/\Delta R$ is used to describe the magnitude of the rotational force along the direction of rotation. Dividing by ΔR scales the magnitude of F so that it is independent of the dimer separation. A finite difference approximation to the change in the rotational force, F, as the dimer rotates through the angle $d\theta$ is given by

$$F' = \frac{dF}{d\theta} \approx \left. \frac{\mathbf{F}^\star \cdot \hat{\boldsymbol{\Theta}}^\star - \mathbf{F} \cdot \hat{\boldsymbol{\Theta}}}{d\theta} \right|_{\theta = \frac{d\theta}{2}}. \tag{2.5}$$

This approximation most accurately estimates the derivative for the midpoint of the finite rotation at $\theta = d\theta/2$.

A reasonable estimate of the rotation, $\Delta\boldsymbol{\Theta}$, required to bring F to zero can be obtained from Newton's method

$$\Delta\boldsymbol{\Theta} \approx \frac{\left(\mathbf{F} \cdot \hat{\boldsymbol{\Theta}} + \mathbf{F}^\star \cdot \hat{\boldsymbol{\Theta}}^\star \right)}{-2F'}. \tag{2.6}$$

The rotation angles are illustrated in Fig. 2.2. Unfortunately Newton's method systematically overestimates the angle $\Delta\boldsymbol{\Theta}$ required to rotate the dimer to the minimum energy. An improvement on Eq. (2.6) can be made if the form of $E(\theta)$, the dimer energy as a function of rotation angle, is known. This can be accomplished in the following way. The first term in a Taylor expansion of the potential U in the neighborhood of $\mathbf{R}$ is a hyperplane through the point $U(\mathbf{R}) = E_0$. This term alone produces no rotational force on the dimer because the dimer energy in this case is independent of orientation, $E(\theta) = 2E_0$. The quadratic term in the Taylor expansion introduces a rotational force. In order to write an analytic form of the quadratic approximation to the potential, $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ are defined to be the normal modes of the potential U within the plane of rotation. The plane is the two dimensional subspace spanned by $\hat{\boldsymbol{\Theta}}$ and $\hat{\mathbf{N}}$. Including terms up to second order in the Taylor expansion gives

$$U = E_0 - (F_x\,x + F_y\,y) + \frac{1}{2}\left(c_x\,x^2 + c_y\,y^2\right). \tag{2.7}$$

The forces F_x and F_y are $-\partial U/\partial x$ and $-\partial U/\partial y$ where x and y are distances along $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$. The curvature of the potential along $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ are c_x and c_y respectively. The dimer energy

E can be expressed within this quadratic approximation as a function of θ :

$$E = 2E_0 + (\Delta R)^2 \left(c_x \cos^2 (\theta - \theta_0) + c_y \sin^2 (\theta - \theta_0) \right), \tag{2.8}$$

where θ_0 is some reference angle. As expected, F_x and F_y which define the linear change in U do not contribute to the energy of the dimer. Equation (2.8) can be rearranged using a trigonometric identity.

$$E = 2E_0 + \frac{1}{2}(\Delta R)^2 \left((c_x - c_y) \cos (2 (\theta - \theta_0)) + (c_x + c_y) \right). \tag{2.9}$$

The derivative of this potential yields an analytical expression for the scalar rotational force on the dimer

$$F = A \sin (2 (\theta - \theta_0)). \tag{2.10}$$

The constant $A = (c_x - c_y)$ does not, in practice, have to be evaluated. The energy and the rotational force on the dimer are invariant to rotations of π. Equation 2.10 shows that θ_0 can be interpreted as the angle at which the force on the dimer is zero within the rotation plane. The difference $\theta - \theta_0$ is the necessary angle of rotation required to reach a zero force. It is now possible to obtain an analytic form of the derivative F' defined in equation 2.5 within this harmonic approximation to the energy

$$F' = \frac{d F}{d\theta} = 2A \cos (2 (\theta - \theta_0)). \tag{2.11}$$

In a simulation, F and F' are evaluated at some orientation of the dimer. If this point is labeled as $\theta = 0$, then the angle through which the dimer should be rotated to reach a zero in the force is θ_0. Let F_0 and F_0' be the values of F and F' evaluated at $\theta = 0$. Then the ratio of Eqs. (2.10) and (2.11),

$$\frac{F_0}{F_0'} = \frac{1}{2} \tan (-2\theta_0), \tag{2.12}$$

yields a simple expression in which the desired rotation angle can be isolated in terms of

known quantities

$$\Delta\Theta = \theta_0 = -\frac{1}{2}\arctan\left(\frac{2F_0}{F_0'}\right).\qquad(2.13)$$

Equation (2.13) has better behavior than Eq. (2.6) for rotation in the limits of $F \to 0$ and $F' \to 0$. This approach to the minimization is similar to a method used to minimize the electronic degrees of freedom in plane wave based density functional theory (DFT) calculations [28].

As an example, we will discuss an application of the modified Newton method to a system representing an aluminum adatom on an Al(100) surface. Here we focus on the properties of the rotation only. The results of the saddle point searches for this system are presented in Sec. 2.4.2. A dimer is placed on the potential surface and incrementally rotated through an angle of 2π. Figure 2.3 shows the force and energy of the dimer. The energy shows the expected sinusoidal behavior in the local quadratic approximation to the potential. The period of π is due to the symmetry of the dimer. The sinusoidal curve evaluated using Eq. (2.9) agrees well with the energy data. The force is also well represented by the sinusoidal Eq. (2.10). This is to be expected because F is simply proportional to the derivative of the dimer energy with respect to θ. A fit to these data (shown in Fig. 2.3) gives a value of $\theta_0 \approx 0.64937$. This value, indicated by the vertical dashed line, corresponds well with the minimum in the dimer energy within the rotation plane. In a simulation, $\Delta\Theta$ is determined from Eq. (2.13). For this example, F_0 was found to be 4.45312, and F_0' was -2.4847, from which $\Delta\Theta$ was calculated as 0.64936, in good agreement with the observed value.

In summary, the quadratic approximation to the potential provides a formula for rotating the dimer and zeroing the rotational force within the plane. This is done by evaluating the magnitude of the rotational force F, the curvature of the dimer energy F', and evaluating $\Delta\Theta$ by substituting into Eq. (2.13). Figure 2.3 shows the magnitude of the total rotational force. Once the dimer is rotated by $\Delta\Theta$, the rotational force has essentially no component within the plane of rotation. It is a bit disconcerting to see that the magnitude of the total force drops by only 35% in the first iteration. This ratio is typical of the modified Newton method. This can be improved upon by considering conjugate gradients.
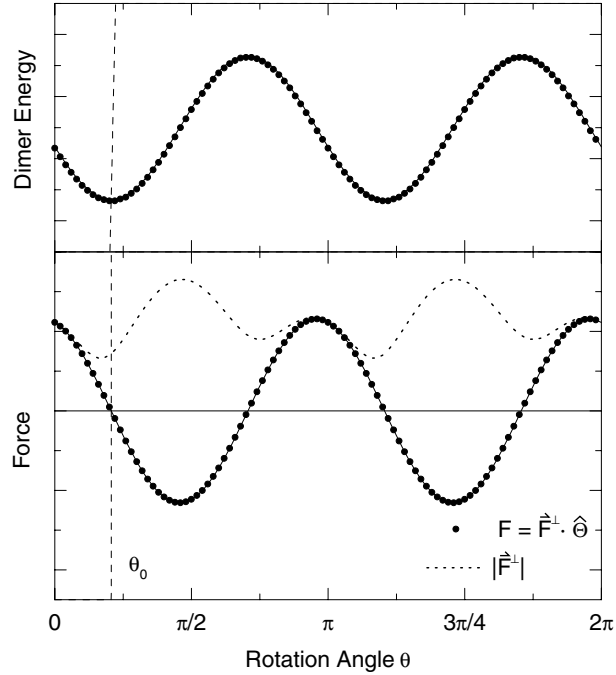
Figure 2.3: Illustration of the modified Newton's method for orienting the dimer. The force and the energy of the dimer for an Al adatom on Al(100) are shown for a full rotation. The success of a sinusoidal fit to the dimer energy indicates that a quadratic approximation [Eq. (2.9)] is a good approximation. A fit [Eq. (2.10)] to the force acting on the dimer yields a minimum dimer energy within the plane at $\theta_0 = 0.64937$, in good agreement with that obtained from Eq. (2.13) using only F and F' calculated at $\theta = 0$. The dashed line shows the magnitude of the total rotational force. At $\theta = \theta_0$ this force has no component in the plane of rotation.

Conjugate Gradient Choice of Rotation Plane

Conjugate gradient methods tend to be more efficient than steepest descent methods because the force at both the current iteration and the previous iteration are used to determine an optimal direction of minimization [29]. We use a conjugate gradient algorithm for choosing the plane of rotation, while the minimization of the force on the dimer within a plane is carried out using the modified Newton method described in the previous section. The conjugate gradient method as described in Ref. [29] cannot be applied directly to the dimer energy minimization. For rotation, the dimer's midpoint and separation must be held fixed, which adds an additional constraint on the system. In this section, the traditional conjugate gradient method is first reviewed, and then a modification for the constrained dimer rotation is described.

The first step in the conjugate gradient minimization is a steepest descent step in which

the direction of displacement is given by the gradient, (or force), $\mathbf{F}$. The energy along $\mathbf{F}$ is then minimized. For subsequent iterations, the direction of displacements, $\mathbf{G}$ is taken to be a linear combination of the current force, $\mathbf{F}_i$, and the force at the previous iteration $\mathbf{F}_{i-1}$. The vector $\mathbf{G}$ at iteration i is defined recursively:

$$\mathbf{G}_i = \mathbf{F}_i + \gamma_i\,\mathbf{G}_{i-1}, \tag{2.14}$$

where γ_i is the weighting factor

$$\gamma_i = \frac{(\mathbf{F}_i - \mathbf{F}_{i-1}) \cdot \mathbf{F}_i}{\mathbf{F}_i \cdot \mathbf{F}_i}. \tag{2.15}$$

In the dimer method, the traditional conjugate gradient method is modified in several ways to accommodate the constraints implicit in the dimer rotation. Each minimization direction, $\mathbf{G}$, becomes a plane of rotation spanned by the unit vector $\hat{\mathbf{\Theta}}$ which is parallel to $\mathbf{G}^{\perp}$, and the dimer orientation $\hat{\mathbf{N}}$. The line minimization step is implemented with the modified Newton's method of the previous section. The difference is that, for every step other than the first, $\hat{\mathbf{\Theta}}$ is not along the force $\mathbf{F}^{\perp}$ as it was, but rather along the conjugate vector $\mathbf{G}^{\perp}$. Equations analogous to Eqs. (2.14) and (2.15) are used in the conjugate gradient rotation scheme. For the first iteration $\mathbf{G}^{\perp} = \mathbf{F}^{\perp}$. For the ith iteration

$$\mathbf{G}_i^{\perp} = \mathbf{F}_i^{\perp} + \gamma_i\,|\mathbf{G}_{i-1}^{\perp}|\,\hat{\mathbf{\Theta}}_{i-1}^{\star\star} \tag{2.16}$$

where

$$\gamma_i = \frac{\left(\mathbf{F}_i^{\perp} - \mathbf{F}_{i-1}^{\perp}\right) \cdot \mathbf{F}_i^{\perp}}{\mathbf{F}_i^{\perp} \cdot \mathbf{F}_i^{\perp}}, \tag{2.17}$$

is the weighting factor between the rotational force $\mathbf{F}_i^{\perp}$ and the old modified force vector $\mathbf{G}_{i-1}^{\perp}$. The vector $\mathbf{G}_{i-1}$ in Eq. (2.14) has been replaced by the vector $|\mathbf{G}_{i-1}^{\perp}|\,\hat{\mathbf{\Theta}}_{i-1}^{\star\star}$ (Fig. 2.2 shows how $\hat{\mathbf{\Theta}}^{\star\star}$ is found). This difference is due to the fact that $\mathbf{G}_{i-1}^{\perp}$ was aligned along $\hat{\mathbf{\Theta}}_{i-1}$. As described in the previous paragraph, a vector which is perpendicular to $\hat{\mathbf{N}}_i$ within the old rotation plane is needed. This is simply a vector along $\hat{\mathbf{\Theta}}_{i-1}^{\star\star}$, with a magnitude equal to $\mathbf{G}_{i-1}^{\perp}$. The conjugate gradient approach (Eqs. (2.16) and (2.17)), including the modified Newton algorithm for line minimization [Eq. (2.13)], represents a significant improvement
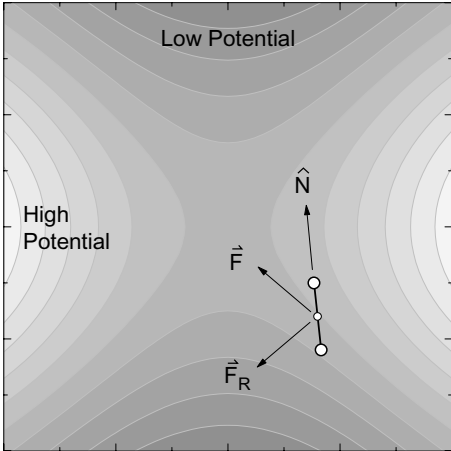
Figure 2.4: The effective force $\mathbf{F}^\dagger$ acting on the center of the dimer is the true force $\mathbf{F}_R$ with the component along the lowest curvature mode $\hat{\mathbf{N}}$ inverted. In the neighborhood of a saddle point, the effective force points towards the saddle point.

over a straightforward application of the standard conjugate gradient algorithm with a fixed bond length constraint for the dimer. The average number of force calls required to find a saddle point in the Al/Al(100) study was reduced by a factor of about 2.5.

2.3.3 Translating the Dimer

Compared to rotation, the translation of the dimer is relatively straightforward. The saddle point is a maximum along the lowest curvature mode, the reaction coordinate, and a minimum along all other modes. The dimer will orient itself along the lowest curvature mode when the energy of the dimer is minimized by rotation. The net translational force acting on the two images in the dimer, $\mathbf{F}_R$, tends to pull the dimer towards a minimum, however. Therefore, a modified force, $\mathbf{F}^\dagger$, is defined where the force component along the dimer is inverted:

$$\mathbf{F}^\dagger = \mathbf{F}_R - 2\mathbf{F}^\parallel. \tag{2.18}$$

Movement of the dimer along this modified force will bring it to a saddle point. This is illustrated in Fig. 2.4. In principle, any optimization algorithm depending only on first derivatives can be used to move the dimer along the effective force to the saddle point.

We have used two different algorithms to translate the dimer. The first is similar to an algorithm that has been used by others to find potential energy minima [30]. We will refer to

this algorithm as the 'quick-min' algorithm. A time step size, Δt, is selected. This should be as large as possible, while still allowing the system to reach the convergence criteria for the saddle point. The system is propagated from its initial position using a classical dynamics algorithm, with the modification that only the projection of the velocity at the previous step along the current force is kept. Additionally, if the dot product of the force and the velocity becomes negative, the cumulative velocity is set to zero

$$\Delta\mathbf{V}_i = \mathbf{F}_i^\dagger \Delta t/m \tag{2.19}$$

$$\mathbf{V}_i = \begin{cases} \Delta\mathbf{V}_i\left(1 + \Delta\mathbf{V}_i \cdot \mathbf{V}_{i-1}/\Delta\mathbf{V}_i^2\right) & \text{if } \mathbf{V}_i \cdot \mathbf{F}_i^\dagger > 0 \\ \Delta\mathbf{V}_i & \text{if } \mathbf{V}_i \cdot \mathbf{F}_i^\dagger < 0 \end{cases}. \tag{2.20}$$

There is a problem with the algorithm described thus far when the dimer is started from a shallow minimum. If the lowest curvature mode is along a contour of the potential energy basin, the dimer can take a very long time to leave the basin, or even possibly become trapped there forever. A solution to this problem is to treat regions where all modes have positive curvature, the *convex regions*, differently from the regions where at least one mode has a negative curvature, the *nonconvex regions*. The neighborhood of potential minima falls into the first category while the saddle point region falls into the second category. Equation 2.18 is modified in the following way to ensure that the dimer quickly leaves convex regions:

$$\mathbf{F}^\dagger = \begin{cases} -\mathbf{F}^{\parallel} & \text{if } C > 0 \\ \mathbf{F}_R - 2\mathbf{F}^{\parallel} & \text{if } C < 0 \end{cases}, \tag{2.21}$$

where C is the minimum curvature. In the convex regions $C > 0$, and the dimer follows this mode up the potential surface until the lowest curvature becomes negative. It is possible that C never becomes negative (an example of that in a two-dimensional case will be given below), in which case the dimer continues to climb up the potential forever. This problem is unlikely to occur in large atomic systems. We never encountered it in the Al/Al(100) calculations described below.

The second method we tried for translating the dimer was the conjugate gradient method. This was found to perform better than quick-min in the Al/Al(100) calculations. In the initial step, the system is minimized along a line defined by the initial force. Analo-

gous to the rotation algorithm, the system is moved a small distance along the line (keeping the dimer orientation fixed), and the derivative in the magnitude of the effective forces was calculated. Newton's method is used to estimate the zero in the effective force along the line and the dimer is moved to that point. If the effective force in the line increases in the small step, the dimer is still in the minimum region, and Newton's method calculates a step backwards against the effective force, pulling the dimer back into the minimum. In this case, the dimer is simply moved with the effective force a predefined step size. This algorithm tends to move the dimer out of the convex region quickly and in practice speeds up convergence to a saddle point. After each translation, the dimer is reoriented and then moved along a direction conjugate to the previous line minimization [29].

2.3.4   Selecting Initial Configurations

In most systems, there can be a large number of saddle points leading out of the potential energy basin of interest. A single saddle point search will generally not be enough to address the question of how the system tends to leave the basin. In general, it is necessary to know *all* low lying saddle points (to within a few $k_B T$ from the lowest energy saddle point) leading from a potential energy basin. While no existing method can guarantee that all relevant saddle points will be found, reasonable progress may be made if there is a way to search for new saddle points in a manner that minimizes the number of duplications.

One simple approach is to start with a collection of initial configurations, scattered about the potential energy minimum in the initial basin. In order to avoid high energy configurations, which might be spatially near the potential minimum, a system can be evolved by classical dynamics at some finite temperature and configurations of the atoms corresponding to maximal displacements from the potential energy minimum can be saved as initial configurations for saddle point searches. In other words, different saddle points can be found if the initial configurations are drawn from the high potential energy images within a thermal ensemble in the potential energy basin. This approach turned out to be quite successful. But, it is important to realize that some saddle points can be systematically excluded when only this method is used. The configurations generated tend to be along low energy modes around the minimum and the dimer searches from these configurations tend to converge to the same saddle points, the saddle point lying at the end of a low

curvature mode. These are, however, often the lowest energy saddle points. Starting with a random set of images displaced from the minimum, for example, a Gaussian distribution of displacements amounting to ≈ 0.1 Å in each coordinate, gave a greater variety of saddle points and therefore better sampling. The following section describes how different saddle points can be found when starting from the same initial configuration.

2.3.5   Orthogonalization

The various mode following algorithms can converge to a variety of saddle points starting from the same initial configuration by following different normal modes [20, 21, 22, 24]. There is, however, no inherent relationship between the number of normal modes and the number of saddle points. The important aspect of the mode following algorithms is the orthogonality of the modes, which tends to lead the system in different directions towards different saddle points. This kind of orthogonality constraint can be built into the dimer method quite easily and with little increase in computational cost.

From any initial configuration, the lowest curvature mode can be found with the dimer rotation method described in Sec. 2.3.2. When the dimer is rotated to its minimum energy configuration, the dimer orientation is along the lowest curvature mode $\hat{\mathbf{N}}_1$. Let the curvature along this mode be denoted by C_1. The first saddle point search will typically be launched with the dimer initially oriented along this mode. It is also possible to find the next lowest mode by again rotating a dimer to its minimum energy configuration but now maintaining orthogonality to the vector $\hat{\mathbf{N}}_1$. Orthogonalization is carried out by simply subtracting any component along $\hat{\mathbf{N}}_1$ from the vectors $\hat{\mathbf{N}}$, $\mathbf{F}_1$, and $\mathbf{F}_2$, while rotating the dimer to a minimum energy. A new saddle point search can then be launched from the same initial configuration with the dimer initially oriented along the second lowest mode $\hat{\mathbf{N}}_2$. In this second search, the orthogonality condition between the current orientation of the dimer $\hat{\mathbf{N}}$ and the initial orientation in the first search $\hat{\mathbf{N}}_1$ is maintained until the curvature C along the dimer becomes lower than the curvature measured along the direction $\hat{\mathbf{N}}_1$. This is not the same as requiring that the curvature along $\hat{\mathbf{N}}$ be lower than C_1, the initial curvature along $\hat{\mathbf{N}}_1$. The requirement is that $\hat{\mathbf{N}}$ gives the lowest curvature mode at the point on the potential energy surface when the orthogonality constraint is dropped. When $\hat{\mathbf{N}}$ gives the lowest curvature mode, there is no need to maintain the orthogonality

condition to the vector $\hat{\mathbf{N}}_1$ because the dimer has no tendency to rotate into this now higher curvature direction.

It is straightforward to continue this procedure to follow systematically different directions. As long as the curvature along the current lowest mode $\hat{\mathbf{N}}$ is greater than the curvature in one of the initial directions of earlier searches, then the orthogonality condition is maintained. These different saddle point searches can be carried out in parallel after the initial set of low lying modes ($\hat{\mathbf{N}}_1$, $\hat{\mathbf{N}}_2$, ...) have been found. The cost of these subsequent searches increases somewhat, because as long as the dimer does not lie along the lowest curvature mode, the force and dimer energy must be computed for every initially lower mode. Two things save this potentially poor scaling. First, there is no need to compare the curvatures very often, and second, it has turned out that the dimer tends to escape the region where previous modes have lower curvature quite quickly. The combination of a distribution of initial configurations and orthogonalization provides an efficient, highly parallelizable, method for searching for saddle points leading out of a given potential energy basin.

2.4   Results

The characteristics of the dimer method have been studied using two model potentials. The first is a two dimensional model potential. The second is a system representing an Al adatom on an Al(100) surface – a system containing 301 atoms.

2.4.1   Two Dimensional Test System

Simple potential surfaces which can be visualized easily are important test cases for any saddle point search method. We have chosen a LEPS potential coupled to a harmonic oscillator because it illuminates some of the problems that can be encountered in saddle point searches. The analytic form and justification for the potential are described elsewhere [13]. The slowest ascent direction along $-\hat{\mathbf{y}}$ from the initial basin centered at $(0.7655, 0.2490)$ does not lead to a saddle point if a steepest ascent search is carried out along this mode. Without modification, the potential has a single saddle point between the two large basins defined approximately by $x < 2$ and $x > 2$. We have added two Gaussian functions to this

Table 2.1: Parameters for the Gaussian functions added to the two-dimensional test potential.

i	1	2
A_i	1.5	6.0
x_{0_i}	2.02083	0.8
y_{0_i}	-0.172881	2.0
σ_{x_i}	0.1	5.0
σ_{y_i}	0.35	0.7

potential to increase the number of saddle points leading out of the initial basin from one to four

$$G_i(x,y) = A_i \; e^{\frac{-\left(x-x_{0_i}\right)^2}{2\,\sigma_{x_i}}} \; e^{\frac{-\left(y-y_{0_i}\right)^2}{2\,\sigma_{y_i}}}. \tag{2.22}$$

The parameters of the Gaussian functions are given in table 2.1. Figure 2.5 illustrates how the dimer method works on this potential surface.

A classical trajectory calculation was used to generate three starting configurations for the saddle point searches in Fig. 2.5a. Dimers are placed at these initial points (the dimer separation is much too small to be resolved), and the subsequent saddle point searches following the lowest mode and using the quick-min algorithm are shown. There is a fairly sharp change in the two paths initially following the nearly vertical directions. These are the points at which the curvature along the dimer has switched from positive to negative (the boundary of convex and nonconvex regions), as dictated by Eq. 2.21. To begin with, in the convex region, the dimer is only moved up the potential surface along the lowest curvature orientation. After entering the non-convex region, the dimer is also being moved down along the force perpendicular to the dimer orientation.

Figure 2.5b shows the results of saddle point searches using the orthogonalization algorithm. Two initial configurations were used and for each one the two lowest normal modes were used in saddle point searches. The two searches along $+\hat{\mathbf{x}}$ and $+\hat{\mathbf{y}}$ quickly converge to saddle points. The search following the softest mode along $-\hat{\mathbf{y}}$ goes down to about $y = -10$ before locking onto the negative curvature mode along $\hat{\mathbf{x}}$ and rotating by $90°$. The minimization perpendicular to the dimer brings it back towards the saddle point. The second search from this same initial point brings the dimer in the $-\hat{\mathbf{x}}$ direction, where no saddle point exists. It is convenient to specify both a maximum potential energy, and a maximum
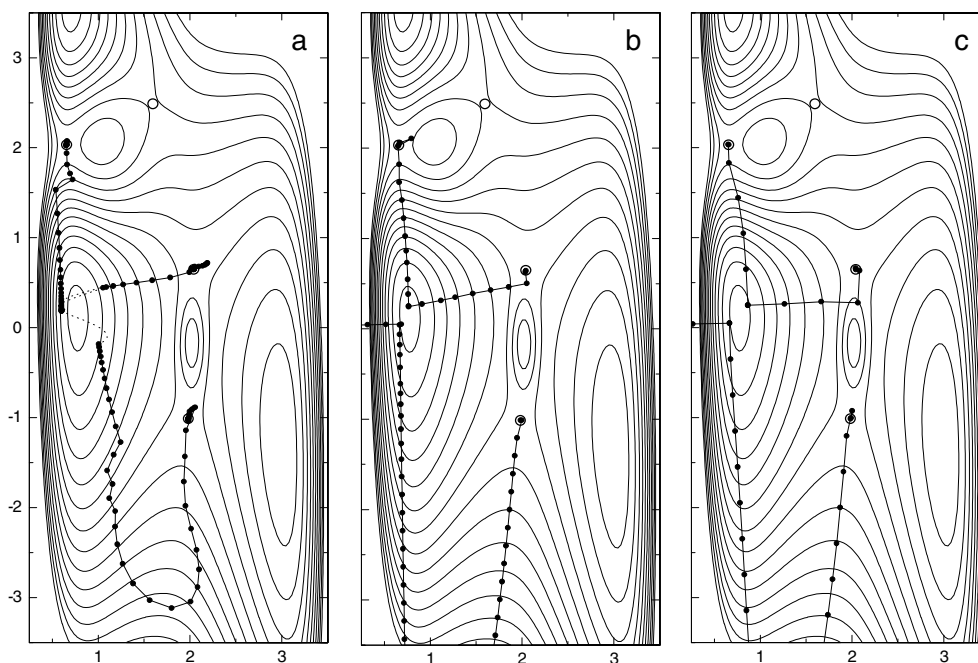
Figure 2.5: Two algorithms for moving the dimer are compared, along with a mode following algorithm (Ref. [24]). The left-hand side figure shows a short molecular dynamics trajectory (dashed line), and three initial configurations generated from the trajectory. The quick-min algorithm was used to move the dimer along the lowest curvature direction. In the middle figure, two initial configurations were chosen on opposite sides of the minimum. For each configuration, two dimer search calculations were carried out, in one case following the lowest curvature direction, in the other following the next lowest curvature. The dimer moving off the plot to the right-hand side did not converge to a saddle point. The dimer which moved off the bottom of the plot returned after moving to $y = -10$ and finally did converge to a saddle point. In this calculation, the conjugate gradient method was used to move the dimer. The right-hand side figure shows results of a mode following method requiring the diagonalization of the Hessian matrix. Qualitatively, the results of the dimer method with orthogonalization are very similar to the results of the mode following method. This comparison is only feasible in a few dimensions because the mode following algorithm requires inversion of the Hessian matrix. None of the methods located one of the saddle points (in the upper right-hand side corner of the basin).

number of allowed iterations so that a failed search, such as this one, can be aborted without wasting an unreasonable amount of effort.

Figure 2.5c shows the results of a mode following method [24] which relies on computing the Hessian matrix. It performs very efficiently on this test potential, finding three modes in an average of 16 moves each. In two dimensions, the cost of each move is small, but as the number of dimensions increase the cost of evaluating and inverting the Hessian matrix increases rapidly (as n^3). Even if the number of steps remains small, the total cost becomes prohibitive in large systems. It is reassuring to see that the orthogonalized dimer searches
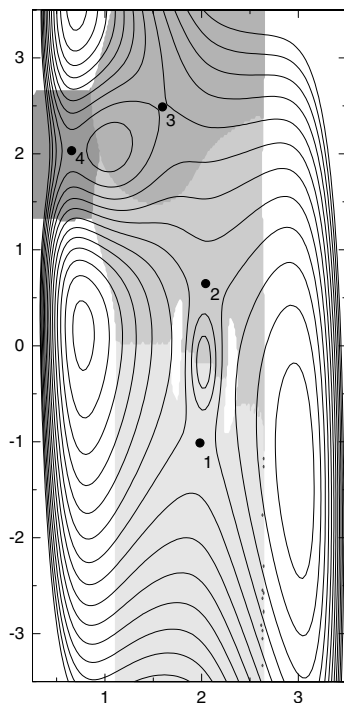
Figure 2.6: Regions of attraction around each saddle point. The shaded regions correspond to points with at least one negative curvature mode. The different shades of grey indicate which saddle point the dimer will converge to. A limitation of the method is apparent here. A dimer starting from the initial basin on the left-hand side will not be able to find saddle point 3, which represents one of the escape routes from the basin, without first visiting the basin around saddle point 2 or 4.

mimic the attractive features of the mode following algorithm without having to evaluate the Hessian matrix.

It is instructive to identify which initial points lead the dimer to a given saddle point. Figure 2.6 shows the basins of attraction for the various saddle points of the two-dimensional test problem when only the lowest mode is followed. A test dimer was placed at an array of initial points. The shaded areas are the regions of the potential in which the dimer rotated into a negative curvature mode, indicating that there is at least one negative curvature mode at that point. Each dimer was then moved in the direction of the effective force to a saddle point. The regions are shaded according to which saddle point the dimer moved to. In this way, the negative curvature region is further divided up into four basins of attraction, one for each saddle point. If the dimer is started outside of the shaded region, it can still converge to a saddle point as shown in Fig. 2.5, by moving up the potential along the lowest (positive) curvature mode until it reaches a basin of attraction. Once the dimer reaches

a basin of attraction of a saddle point, it is most likely to converge to that saddle point. Therefore, images which start around the left minimum will not reach saddle point 3.

It is fairly clear then why none of the images started around the left minimum reached saddle point 3. They would have to pass through basins 2 or 4 before getting to the basin around saddle point 3. This illustrates a limitation of the dimer method. If a saddle point has a basin which is surrounded or separated from the initial configuration by other saddle point basins, it is very unlikely for the dimer method to find it. Another limitation of the method can be seen by considering saddle point 3. This saddle point connects the two minima on the left by a rather curved minimum energy path. Dimers that are started from the minimum on the right following the $+\hat{\mathbf{y}}$ mode converge to saddle point 3. But this is not a saddle point leading out of that minimum. This is a problem when the goal is to find all saddle points leading out of a given minimum. It is easy enough to check if a given saddle point is relevant for a given initial state, but the cost of finding it can be a wasted effort. This effect is observed in the aluminum system as well as this two dimensional potential.

2.4.2   Al Adatom on an Al(100) Surface

We now turn to a study of transition mechanisms in a system where an Al adatom is initially sitting in the fourfold hollow site on Al(100) surface. The goal is to find all mechanisms by which the system can escape from this initial state with an activation energy less than about 1 eV. An embedded atom potential, similar to that of Voter and Chen [31], was used to model the atomic interactions. The energy of the two lowest saddle points predicted by this potential turn out to be close to the results of density functional theory calculations [27], indicating that the potential function is reasonably accurate. The substrate consists of 300 atoms, 50 per layer in 6 layers. The bottom two layers are held frozen and the top surface is left open to vacuum. A single Al adatom is placed on the surface bringing the total number of degrees of freedom to 603.

The dimer method was run starting from 1000 randomly chosen configurations around the minimum. A cluster including the adatom and 25 nearby substrate atoms was displaced according to a Gaussian probability distribution with a width of 0.1 Å along each coordinate. The conjugate gradient method was used both to rotate and translate the dimer. Values of the finite difference steps for rotation and translation were $d\theta = 10^{-4}$ rad and $dR = 10^{-3}$ Å
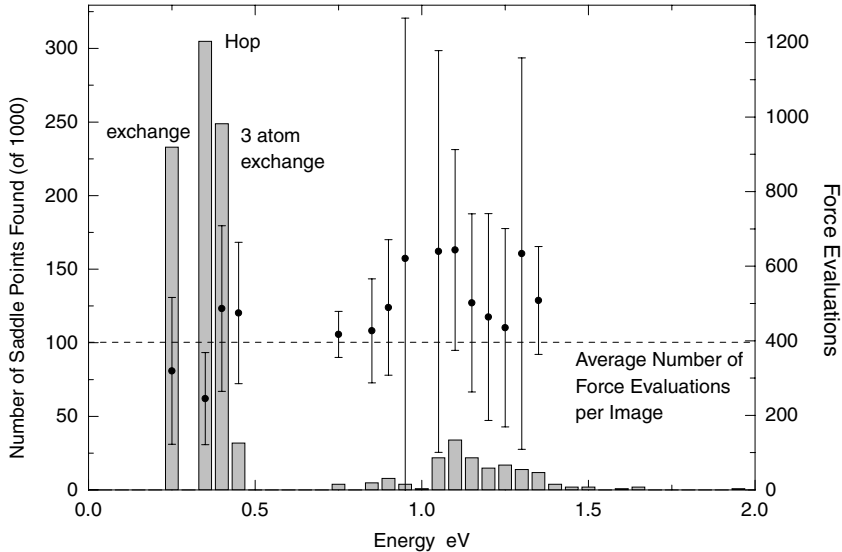
Figure 2.7: The result of 1000 saddle point searches using the dimer method for an Al adatom on an Al(100) surface. Each search starts from a point randomly displaced from the potential energy minimum, using a Gaussian distribution with a width of 0.1 Å in each coordinate of 25 atoms including the adatom and its neighbors. A large number of saddle points was found. The histogram shows how many of the dimers converged on saddle points in the given energy range. The processes corresponding to the ten lowest energy saddle points are shown in Fig. 2.8. A total of 60 processes below 2.0 eV were found, most of them in the range of 0.8 - 1.5 eV. The filled circle shows the average number of force evaluations required to converge to the saddle points within each bin. The bars show the range between the shortest and longest search. On average 400 force evaluations (200 per image in the dimer) were needed to converge to a saddle point (dashed line).

respectively. A maximum translation distance was set at 0.1 Å, and the dimer separation was set at $\Delta R = 10^{-2}$ Å. The tolerance for convergence to a saddle point was set at $|\mathbf{F}| < 10^{-4}$ eV/Å where $\mathbf{F}$ is the 3N dimensional force vector.

A summary of the results is shown in Fig. 2.7. Of the 1000 dimer searches, 990 converged to saddle points below 2 eV. Three searches failed to converge within the imposed limit of 2000 iterations. Seven of the searches converged to the long wavelength, high energy mode (with activation energy of ≈ 5 eV) corresponding to a concerted shift of the 51 surface atoms by one lattice constant. The energy of the saddle points were binned and the number of searches within each bin is shown in the histogram in Fig. 2.7. The average computational cost to find the saddle points within each bin is given in terms of the number of force evaluations. The average number of force evaluations needed to converge on a saddle point was 400, that is 200 per image in the dimer. The three lowest energy saddle points attracted 78% of the 1000 dimers (saddle points that are equivalent by symmetry are
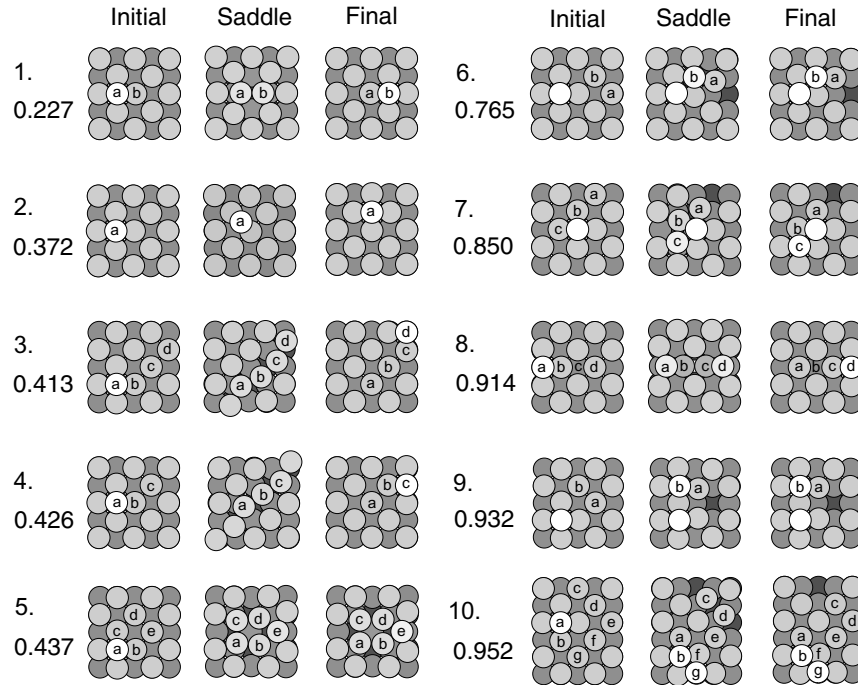
Figure 2.8: The ten lowest energy transition mechanisms found in 1000 searches using the dimer method. An on-top view of a small section of the surface is shown in the initial state (left), saddle point (middle) and final state (right). The atoms are shaded by depth, and the atoms that move the most are labeled. The energy of the saddle point configurations is given in eV with respect to an Al adatom in the fourfold hollow site on the flat Al(100) surface. In addition to the two atom exchange process (process 1, discovered by Feibelman) and the hop (process 2), a four atom exchange mechanism (process 3) and a three atom exchange mechanism (process 4) are found to be low energy diffusion mechanisms. Other processes correspond to an adatom getting buried in the surface (process 5), vacancy formation (processes 6,7,9 and 10) and a four atom exchange diffusion mechanism involving a second layer atom (process 8).

grouped together). These are the exchange process involving two atoms found by Feibelman [27], the hop, and a remarkably low *four* atom exchange process (see Fig. 2.8). A three atom exchange process has the fourth lowest saddle point. No transition mechanisms were found with saddle point energy in the range 0.44 – 0.76 eV, but above this energy gap there is a large number of different processes with saddle point energy up to about 1.5 eV. Figure 2.8 shows the ten lowest energy transitions. The initial state, saddle point and final state are shown. The energies listed below each transition number are the energy of the saddle point configuration with respect to a configuration of an adatom on a flat surface.

Once the dimer has converged to a saddle point, it is easy to trace out the minimum energy path. The dimer is first allowed to rotate into the lowest curvature mode, the unstable mode, so that it is aligned along the reaction coordinate. An image is placed on
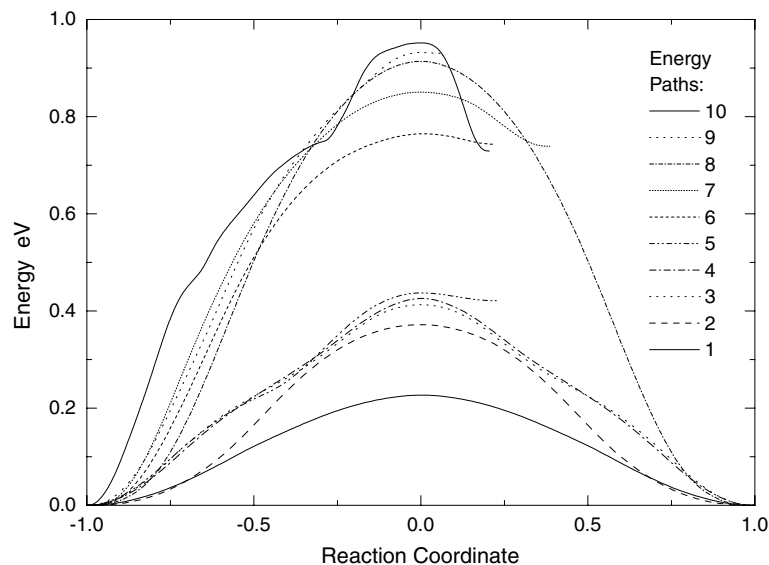
Figure 2.9: The minimum energy paths corresponding to the ten saddle points identified and shown in Fig. 2.8. The reaction coordinate has been scaled so that -1 represents the initial minimum and 0 the saddle point. Transitions 5, 6, 7, 9, and 10 lead to final state local minima which do not correspond to a single adatom on the Al(100) surface and are therefore asymmetric.

one side of the dimer along the direction $\hat{\mathbf{N}}_1$. The distance of the image from the midpoint of the dimer (the saddle point) is chosen according to the desired resolution of the path. In a manner very similar to the algorithm used to rotate the dimer, the energy of this image is minimized while keeping its distance from the previous image (in this case the saddle point) fixed. This procedure is repeated, each time placing a new image initially along the local path (the line between the two previous images) to minimize the number of function calls required to zero the tangential force on the new image. The process is stopped when the minimum energy of an image is greater than that of the previous image. After the path is traced out in one direction from the saddle point to a minimum, the opposite direction must be followed to complete the minimum energy path. This method was used for the ten saddle points of Fig. 2.8 and the energy paths are shown in Fig. 2.9. The reaction coordinate is scaled so that the distance between the minimum and the saddle point is 1 unit. The paths which do not terminate back at an energy of zero, the energy of a single Al adatom on the Al(100) surface, lead to other local minima on the potential energy surface. The final state of path 5 corresponds to a stable arrangement with the adatom buried in the surface. A group of four atoms in the surface layer has rotated by 45°. The final states of paths 6, 7,
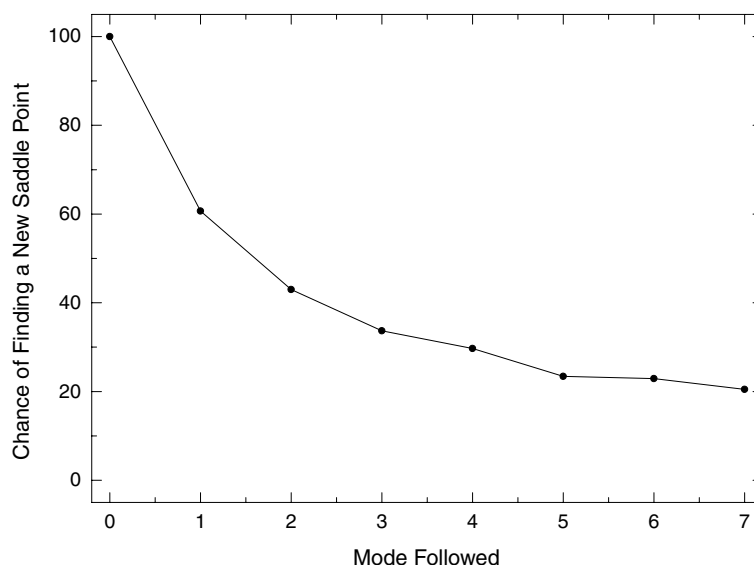
Figure 2.10: For each one of the 1000 random initial configurations, a total of eight dimer searches were carried out with each subsequent search orthogonalized to the initial orientation of the dimer in previous searches. The saddle point obtained in each run was compared to the saddle points found in previous searches started from the same initial configuration. The figure shows the average fraction of searches which lead to a new saddle point. For example, after a saddle point was found by following the lowest curvature direction, the chance of finding a new saddle point when a second lowest, orthogonal direction was followed is approximately 60%. The chance of finding a new saddle point when the third lowest mode is subsequently followed is approximately 40%.

and 10 at approximately 0.7 eV are arrangements in which an addimer/vacancy pair has been created. Path 8 corresponds to a four atom exchange diffusion mechanism involving a second layer atom.

Several orthogonal dimer searches were then carried out from the 1000 initial configurations described above. For each initial configuration, a total of eight orthogonal dimer searches were carried out. In each successive run, the dimer orientation is orthogonalized to the initial orientation of the dimer in the previous runs. After each search, the saddle point obtained was compared to those found in previous searches started from the same initial configuration. In this way, an average chance of finding a new saddle point in a subsequent, orthogonal search was estimated. The results are shown in Fig. 2.10. On average, the second search from a given initial configuration led to a new saddle point 60% of the time. A new saddle point was found in the third search about 40% of the time. During the course of these simulations, many new transition mechanisms for the Al/Al(100) system were found. Some of the more interesting low energy saddle points are shown in Fig. 2.11. A transition
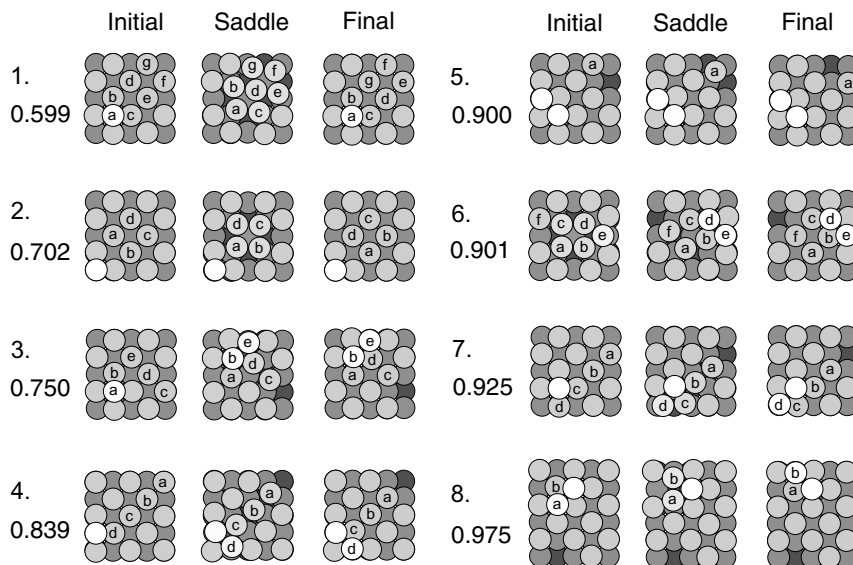
Figure 2.11: Searches using orthogonal directions led to many new transitions for the Al/Al(100) system. Some of the more interesting low energy transitions are shown in the figure. The first transition shows the formation of a local hex reconstruction in the surface plane at the saddle point. In the end, a rotation of a group of four surface atoms has occurred. A similar rotation also occurs in the second transition. An addimer/vacancy pair forms in transitions 3, 4, and 7. Neither the initial nor the final state in transitions 5, 6, and 8 corresponds to an adatom on a flat surface, illustrating that the dimer does not always converge on saddle points corresponding to escape routes for the potential energy basin where the initial point of the search is located.

involving the formation of a local hex reconstruction in the surface layer at the saddle point was found. In the final state of the transition, a group of four surface atoms has rotated so as to exchange places (process 1). A similar rotation of four surface atoms also occurs in the second transition shown in Fig. 2.11. An addimer/vacancy pair forms in several of the transitions (processes 3, 4, and 7). For some of the transitions, neither the initial nor the final state corresponds to an adatom on a flat surface (processes 5, 6, and 8), illustrating that the dimer does not always converge on saddle points corresponding to escape routes for the potential energy basin where the initial point of the search is located.

2.4.3 Scaling with System Size

The motivation behind the dimer method is to develop an algorithm which scales well with system size. Most of the savings over the mode following methods, such as the Cerjan-Miller method, is a result of not having to evaluate and invert the Hessian matrix. The dominant computational effort in the dimer method involves computation of the force on the atoms,

so the effort can effectively be measured by the number of force evaluations required to converge on a saddle point. The more degrees of freedom there are in the system, the more iterations are needed to orient and translate the dimer. An important question is how the computational effort scales with the number of degrees of freedom. The maximum number of degrees of freedom was obtained in the Al/Al(100) system by freezing only 55 atoms at the bottom of the substrate leaving the method to explore the remaining 738 degrees of freedom. In the other extreme, only the adatom was allowed to move, making 3 the minimum number of degrees of freedom. The average cost of finding an ensemble of saddle points was computed for eight configurations with more and more of the substrate atoms frozen. For the most restricted sytems with fewer than 70 degrees of freedom, the hop was the only process found. The dimer method converged with about 70 force evaluations in these simulations. For simulations with less frozen atoms the full range of saddle points was found. This distribution was fairly insensitive to the number of degrees of freedom beyond 70. The average number of force evaluations for these runs is approximately 400, the same as what was shown in Fig. 2.7. The data in Fig. 2.12 show that the method is relatively insensitive to increasing the number of dimensions in the problem, as long as all processes are available to the system. This is very encouraging since it indicates that the dimer method should be useful for finding saddle points in large and complex systems.

2.5   Discussion

Finding the mechanism and estimating the rate of activated transitions from a given initial state boils down to locating the low energy saddle points at the rim of the potential energy basin corresponding to the initial state (if the harmonic approximation to transition state theory is a good approximation). For problems involving diffusion of atoms in and on the surface of crystals this problem is tractable, but far from trivial. Preconceived notions of the transition mechanism can be incorrect, as exemplified by the diffusion of an adatom on an Al(100) surface, which was thought to occur by simple hops of the adatom from one surface site to another until Feibelman discovered that an exchange mechanism is significantly lower in energy [27]. A systematic procedure for finding saddle points is needed for such systems. *Ab initio* calculations of small molecules have over the last decade made use of
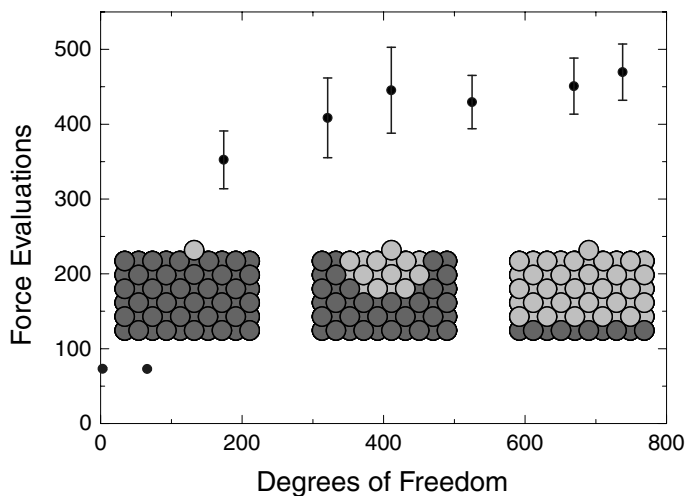
Figure 2.12: The scaling of the computational effort for the dimer method, measured by number of force evaluations, as a function of the system size. These data are taken from the Al/Al(100) system. The substrate consisted of 300 atoms, 50 atoms per layer. Starting with all movable atoms, the number of degrees of freedom was gradually reduced by freezing more and more of the substrate atoms so they became effectively removed from the calculation. If an atom is frozen the dimer cannot be oriented in these degrees of freedom, and the forces on frozen atoms do no affect the dimer. For the smallest number of degrees of freedom, the only mechanism found was the hop. When more than 20 atoms were unfrozen, the full range of saddle points was found. In this region, the plot shows a remarkably slow increase in cost with system size, especially if compared to the n^3 scaling of mode following algorithms which involve the inversion of the Hessian matrix.

an efficient mode-following algorithm [20, 21, 22, 24]. This method has also been applied to small clusters described by empirical potentials, but the mode following algorithm scales poorly with size, making it inefficient for solid state applications. It, furthermore, requires knowledge of second derivatives of the potential energy with respect to the coordinates of the atoms, preventing its use in the highly successful plane wave based DFT calculations of solids and surfaces of solids. (Some examples of DFT studies of surface diffusion are Refs. [27, 32, 33, 34, 35]). The dimer method presented here can be applied to large systems and since it only relies on first derivatives of the energy, it can be used in conjunction with plane wave based DFT calculations of the atomic forces. We are currently implementing the dimer method in a plane wave DFT code.

An essential aspect of the dimer method is a highly optimized algorithm for rotating the dimer into the lowest energy orientation. This makes it feasible to use the method in conjunction with *ab initio* atomic forces. The calculations for the Al/Al(100) system it took on average 400 force evaluations to converge on a saddle point. Since the force

calculations on the two images in the dimer are independent, the dimer method parallelizes almost perfectly over two processors if the force evaluation is computationally intensive as in *ab initio* calculations. This means the computational effort is 200 force evaluations per processor. In order to find the set of low lying saddle points, a few saddle point searches have to be carried out. In the Al/Al(100) system, approximately ten searches would suffice to find the three to four lowest saddle points. Using either a collection of randomly chosen initial points or orthogonal searches from a given initial point (or a combination of both), the different saddle point searches can be carried out in parallel. The algorithm will, therefore, be particularly useful when a cluster of processors is available for the computations.

Note added in proof. After submitting our manuscript we have learned of a new mthod by Munro and Wales [Phys. Rev. B **59**, 3969 (1999)] which also enables saddle point searches with only first derivatives.

2.6 Acknowledgments

Chapter 3

IMPROVED TANGENT ESTIMATE IN THE NUDGED ELASTIC BAND METHOD FOR FINDING MINIMUM ENERGY PATHS AND SADDLE POINTS

3.1 Abstract

An improved way of estimating the local tangent in the nudged elastic band method for finding minimum energy paths is presented. In systems where the force along the minimum energy path is large compared to the restoring force perpendicular to the path and when many images of the system are included in the elastic band, kinks can develop and prevent the band from converging to the minimum energy path. We show how the kinks arise and present an improved way of estimating the local tangent which solves the problem. The task of finding an accurate energy and configuration for the saddle point is also discussed and examples given where a complementary method, the dimer method, is used to efficiently converge to the saddle point. Both methods only require the first derivative of the energy and can, therefore, easily be applied in plane wave based density-functional theory calculations. Examples are given from studies of the exchange diffusion mechanism in a Si crystal, Al addimer formation on the Al(100) surface, and dissociative adsorption of CH_4 on an Ir(111) surface.

3.2 Introduction

The nudged elastic band (NEB) method is an efficient method for finding the minimum energy path (MEP) between a given initial and final state of a transition [36, 37, 13]. It has become widely used for estimating transition rates within the harmonic transition state theory (hTST) approximation. The method has been used both in conjunction with electronic structure calculations, in particular plane wave based density-functional theory (DFT) calculations (see, for example, Refs. [38, 39, 40, 41, 42]), and in combination with

empirical potentials [43, 44, 45, 46]. Studies of very large systems, including over a million atoms in the calculation, have been conducted [47]. The MEP is found by constructing a set of images (replicas) of the system, typically on the order of 4 to 20, between the initial and final state. A spring interaction between adjacent images is added to ensure continuity of the path, thus mimicking an elastic band. An optimization of the band, involving the minimization of the force acting on the images, brings the band to the MEP. An essential feature of the method, which distinguishes it from other elastic band methods [48, 49, 50, 51], is a force projection which ensures that the spring forces do not interfere with the convergence of the elastic band to the MEP, as well as ensuring that the true force does not affect the distribution of images along the MEP. It is necessary to estimate the tangent to the path at each image and every iteration during the minimization, in order to decompose the true force and the spring force into components parallel and perpendicular to the path. Only the perpendicular component of the true force is included, and only the parallel component of the spring force. This force projection is referred to as "nudging." The spring forces then only control the spacing of the images along the band. When this projection scheme is not used, the spring forces tend to prevent the band from following a curved MEP (because of "corner-cutting"), and the true force along the path causes the images to slide away from the high energy regions towards the minima, thereby reducing the density of images where they are most needed (the "sliding-down" problem). In the NEB method, there is no such competition between the true forces and the spring forces; the strength of the spring forces can be varied by several orders of magnitude without effecting the equilibrium position of the band.

The MEP can be used to estimate the activation energy barrier for transitions between the initial and final states within the hTST approximation. Any maximum along the MEP is a saddle point on the potential surface, and the energy of the highest saddle point gives the activation energy needed for the hTST rate estimate. It is important to ensure that the highest saddle point is found, and therefore some information about the shape of the MEP is needed. It is quite common to have MEPs with one or more intermediate minima and the saddle point closest to the initial state may not be the highest saddle point for the transition.

While the NEB method is robust and has been successful, there are situations where

the elastic band does not converge well to the MEP. When the force parallel to the MEP is large compared to the force perpendicular to the MEP, and when many images are used in the elastic band, kinks can form on the elastic band. As the minimization algorithm is applied, the kinks can continue to oscillate back and forth, preventing the band from converging to the MEP. Previously, this problem was reduced by including some fraction of the perpendicular spring force when the tangent was changing appreciably between adjacent images in the band [13]. A switching function was added which gradually introduced the parallel spring component as the change in the tangent increased. The problem with this approach is that the elastic band then tends to be pulled off the MEP when the path is curved. If the path is curved at the saddle point, this can lead to an overestimate of the saddle point energy. This problem was, in particular, observed in calculations of the exchange diffusion process in a Si crystal (discussed below).

In this paper we present an analysis of the origin of the kinks and give a simple solution to the problem. With a different way of estimating the local tangent to the elastic band, the tendency to form kinks disappears. We also address the issue of finding a good estimate of the saddle point given the converged elastic band. An estimate of the saddle point configuration is obtained from the pair of images near the saddle point by interpolation. The dimer method [2] is then used to converge on the saddle point. In cases where the energy barrier is narrow, and few images are located near the saddle point, the estimate obtained from the elastic band alone may not be reliable. When the saddle point is found rigorously with the dimer method, it is not necessary to converge the elastic band all the way to the MEP. The optimal combination of partial minimization of the elastic band and subsequent dimer method calculations is discussed. The methodology is presented with reference to simple two-dimensional model potentials. We also present results of calculations on three realistic systems: the exchange of Si atoms in a Si crystal, the formation of an Al addimer on the Al(100) surface, and the dissociation of a CH_4 molecule on the Ir(111) surface. We have used both empirical potentials and DFT calculations in these studies.

3.3   The Original Implementation of NEB

An elastic band with $N + 1$ images can be denoted by $[\mathbf{R}_0,\ \mathbf{R}_1,\ \mathbf{R}_2, \ldots, \mathbf{R}_N]$ where the endpoints, $\mathbf{R}_0$ and $\mathbf{R}_N$, are fixed and given by the energy minima corresponding to the initial and final states. The $N - 1$ intermediate images are adjusted by the optimization algorithm.

In the original formulation of the NEB method, the tangent at an image i was estimated from the two adjacent images along the path, $\mathbf{R}_{i+1}$ and $\mathbf{R}_{i-1}$. The simplest estimate is to use the normalized line segment between the two

$$\hat{\tau}_i = \frac{\mathbf{R}_{i+1} - \mathbf{R}_{i-1}}{|\mathbf{R}_{i+1} - \mathbf{R}_{i-1}|}, \tag{3.1}$$

but a slightly better way is to bisect the two unit vectors

$$\tau_i = \frac{\mathbf{R}_i - \mathbf{R}_{i-1}}{|\mathbf{R}_i - \mathbf{R}_{i-1}|} + \frac{\mathbf{R}_{i+1} - \mathbf{R}_i}{|\mathbf{R}_{i+1} - \mathbf{R}_i|}, \tag{3.2}$$

and then normalize $\hat{\tau} = \tau/|\tau|$. This latter way of defining the tangent ensures the images are equispaced (when the spring constant is the same) even in regions of large curvature. The total force acting on an image is the sum of the spring force along the tangent and the true force perpendicular to the tangent

$$\mathbf{F}_i = \mathbf{F}_i^s \mid_{\parallel} -\nabla V(\mathbf{R}_i) \mid_{\perp}, \tag{3.3}$$

where the true force is given by

$$\nabla V(\mathbf{R}_i) \mid_{\perp} = \nabla V(\mathbf{R}_i) - \nabla V(\mathbf{R}_i) \cdot \hat{\tau}_i, \tag{3.4}$$

and the spring force was, in the simplest version of the method, calculated as

$$\mathbf{F}_i^s \mid_{\parallel} = k \left[(\mathbf{R}_{i+1} - \mathbf{R}_i) - (\mathbf{R}_i - \mathbf{R}_{i-1}) \right] \cdot \hat{\tau}_i \hat{\tau}_i. \tag{3.5}$$

When the perpendicular component of the spring force is not included, as in Eq. (3.5), the path forms kinks in some cases. An illustration of this is given in Fig. 3.1. The
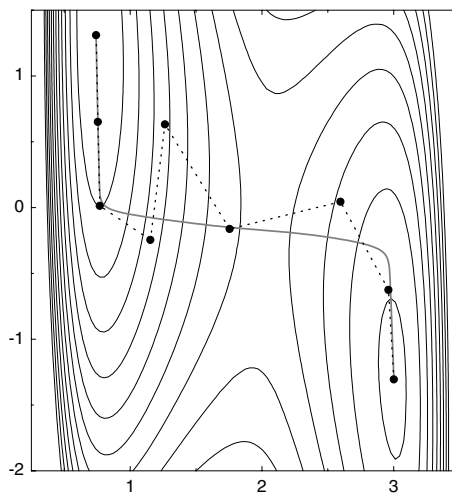
Figure 3.1: The original nudged elastic band method as described by Eqs. (3.2) and (3.3) can develop kinks along the path as illustrated here for a two-dimensional LEPS model potential. The kinks do not get smaller in this case, but fluctuate as the minimization is continued. The band does not converge to the minimum energy path (solid line).

potential surface is a two-dimensional combination of a LEPS potential [52] and a harmonic oscillator (the functional form and parameters are given in Ref. [13]). Initially, the seven movable images were placed equispaced along the straight line between the minima. A minimization algorithm was then applied, moving the images in the direction of the force given by Eq. (3.3). We use a projected velocity Verlet algorithm for the minimization [13]. After a few iterations the images are close to the MEP, but after a certain level of convergence is reached, the magnitude of the force on the images does not drop any further. Kinks form in regions where the parallel component of the force is large compared with the perpendicular component, often near the inflection points of the energy curve for the MEP. As the minimization algorithm is continued, the kinks simply fluctuate back and forth. The saddle point region, and thereby the estimate of the saddle point energy, were typically not affected significantly by the kinks. In some systems, however, this turned out to be a serious problem.

In order to reduce the kinks and enable the minimization to reduce the magnitude of the forces to a prescribed tolerance, some fraction of the perpendicular component of the spring force was included when the angle between the vectors $\mathbf{R}_i - \mathbf{R}_{i-1}$ and $\mathbf{R}_{i+1} - \mathbf{R}_i$ deviated appreciably from zero. A switching function of the angle was used to multiply the perpendicular spring force component [13]. At kinks the angle is large, and addition of some
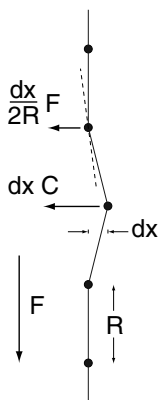
Figure 3.2: An illustration of the cause of the kinks that can develop on the elastic band in the original formulation when the force component parallel to the path is large while the perpendicular component is small. The sketch indicates the critical condition for stability as given by Eq. (3.6).

of the perpendicular spring force tends to straighten the elastic band out. The problem is that this leads to corner-cutting in regions where the MEP is curved. If the saddle point is in such a curved region, then the addition of a fraction of the perpendicular spring force will lead to an overestimate of the saddle point energy. In many cases this overestimate is not a serious problem. Nevertheless, it is still an inconvenience; the switching function means that there is an extra parameter in the implementation of the NEB and there is some uncontrollable error in the estimate of the activation energy. The exchange diffusion process in Si is one example where the MEP is so curved at the saddle point that the switching function leads to a significant overestimate of the activation energy, as discussed below. Before discussing a way to solve this problem, we analyze the origin of the kinks in the following section.

3.4   Analysis of Kinks on the Elastic Band

We now examine more closely the origin of the kinks on the elastic band, shown for example in Fig. 3.1. In many calculations the original NEB method converges very well, even without a switching function, but in some situations it is unstable. Figure 3.2 shows five images, separated by a distance R, in an idealized NEB calculation. It is assumed that additional images are located above and below the segment shown. The MEP is assumed to be straight in this region, and the top two and bottom two images shown are lying on the MEP. The potential energy surface is assumed to have a constant slope in the direction of the MEP and

the force in that direction is given by the constant F. The potential energy perpendicular to the MEP is assumed to be quadratic with a restoring force of $-x\,C$ where x is the perpendicular distance from the MEP, and C is the curvature. The middle image in Fig. 3.2 is displaced a small distance dx from the MEP. If the original way of estimating the tangent [Eq. (3.2)] is used, there are two competing effects acting on the middle image. The first is the tendency for the displaced image to move back under the restoring force $-dx\,C$. The second is for the higher energy neighboring image to move away from the MEP because the estimated tangent at the higher image is no longer along the MEP and the force F now has a small perpendicular component, $dx/2R$. The band becomes unstable when the restoring force $-dx\,C$ is less than the destabilizing force $dx/2R\,F$. This leads to the stability condition

$$F < 2\,C\,R. \tag{3.6}$$

Since the spacing between the images, R, depends on the number of images used in the band, this stability condition predicts that the band always becomes unstable if the number of images is made large enough (in the absence of the switching function).

The stability condition was tested on a simple two-dimensional system with a straight MEP. The potential energy function is

$$V(x, y) = A_x \cos(2\pi x) + A_y \cos(2\pi y). \tag{3.7}$$

An MEP lies along the straight line between (0,0) and (1,0). The maximum force along the path is $F = 2\pi A_x$ and the curvature perpendicular to the path is $C = 4\pi^2 A_y$. Choosing $A_x = A_y = 1$ yields the stability condition $R > \frac{1}{4}\pi \approx \frac{1}{13}$. This corresponds to a maximum of 12 images in the elastic band before it becomes unstable.

This prediction is born out by the calculations. A chain with 11 images with small initial perturbations from the MEP, proved to be stable and converged monotonically to the straight path as shown in Fig. 3.3(a). A chain with 13 images, on the other hand, took much longer to converge and developed large fluctuations, bigger than the initial displacement, before eventually converging to the MEP. Several snapshots of the calculation are shown in Fig. 3.3(b), and illustrate the typical magnitude of oscillations during the simulation.

When 25 images were used in the band, very large kinks developed so that images
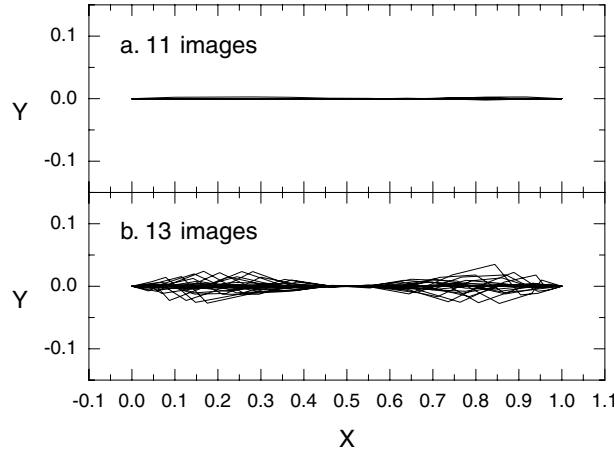
Figure 3.3: The effect of the number of images on the stability of the nudged elastic band in the original formulation. The calculations were performed on the cosine potential given in Eq. (3.7). (a) When 11 images are used, the stability condition given by Eq. (3.6) is satisfied and the elastic band converges to the minimum energy path. The initial configuration, which deviates slightly from the minimum energy path, is shown as well as a few configurations taken from the minimization. (b) With 13 images, the stability is borderline and the condition given by Eq. (3.7) violated. The small initial deviations of the band from the minimum energy path are magnified as large kinks form in the neighborhood of the inflection points. The kinks oscillate during the minimization until the band eventually settles to the minimum energy path.

were able to slide down into the minima at (0,0) and (1,0) during the minimization. This continued until the the number of images located between the minima had dropped and the spacing between them increased enough to satisfy the stability condition, Eq. (3.6). Then the images at the minima were slowly pulled into place by the spring force. In this way, chains with up to ∼80 images were able to (slowly) converge. For more images, the minimization would not converge and some images were always left in a jumble at the minima.

In order to further test the stability condition, the restoring force perpendicular to the path was doubled by setting $A_y = 2$. The stability condition predicts that a band with twice the number of images, 24, will remain stable. Calculations showed that the band becomes unstable at 21 images. The good agreement between the simple prediction of Eq. (3.6) and these simulations suggest that this is the correct origin of the kinks. The modified NEB method with the new tangent, presented in the next section, converges well for both small and large numbers of images.

3.5 The New Implementation of NEB

By using a different definition of the local tangent at image i, the kinks can be eliminated. Instead of using both the adjacent images, $i+1$ and $i-1$, only the image with higher energy and the image i are used in the estimate. The new tangent, which replaces Eq. (3.2), is

$$\tau_i = \begin{cases} \tau_i^+ & \text{if } V_{i+1} > V_i > V_{i-1} \\ \tau_i^- & \text{if } V_{i+1} < V_i < V_{i-1} \end{cases}, \tag{3.8}$$

where

$$\tau_i^+ = \mathbf{R}_{i+1} - \mathbf{R}_i, \quad \text{and} \quad \tau_i^- = \mathbf{R}_i - \mathbf{R}_{i-1}, \tag{3.9}$$

and V_i is the energy of image i, $V(\mathbf{R}_i)$. If both of the adjacent images are either lower in energy, or both are higher in energy than image i, the tangent is taken to be a weighted average of the vectors to the two neighboring images. The weight is determined from the energy. The weighted average only plays a role at extrema along the MEP and it serves to smoothly switch between the two possible tangents τ_i^+ and τ_i^-. Otherwise, there is an abrupt change in the tangent as one image becomes higher in energy than another and this can lead to convergence problems. If image i is at a minimum $V_{i+1} > V_i < V_{i-1}$ or at a maximum $V_{i+1} < V_i > V_{i-1}$, the tangent estimate becomes

$$\tau_i = \begin{cases} \tau_i^+ \Delta V_i^{\max} + \tau_i^- \Delta V_i^{\min} & \text{if } V_{i+1} > V_{i-1} \\ \tau_i^+ \Delta V_i^{\min} + \tau_i^- \Delta V_i^{\max} & \text{if } V_{i+1} < V_{i-1} \end{cases} \tag{3.10}$$

where

$$\begin{aligned} \Delta V_i^{\max} &= \max\left(|V_{i+1} - V_i|, |V_{i-1} - V_i|\right), \quad \text{and} \\ \Delta V_i^{\min} &= \min\left(|V_{i+1} - V_i|, |V_{i-1} - V_i|\right). \end{aligned} \tag{3.11}$$

Finally, the tangent vector needs to be normalized. With this modified tangent, the elastic band is well behaved and converges rigorously to the MEP if sufficient number of images are included in the band.

Another small change from the original implementation of the NEB is to evaluate the spring force as

$$\mathbf{F}_i^s \mid_{\parallel} = k\left(|\mathbf{R}_{i+1} - \mathbf{R}_i| - |\mathbf{R}_i - \mathbf{R}_{i-1}|\right) \hat{\tau}_i. \tag{3.12}$$
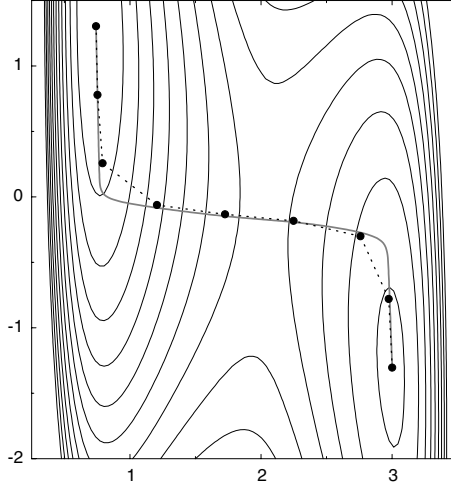
Figure 3.4: With the modified tangent, Eqs. (3.8)-(3.12), the nudged elastic band method does not develop kinks and converges smoothly to the minimum energy path (solid line).

instead of Eq. (3.5). This ensures equal spacing of the images (when the same spring constant, k, is used for the springs) even in regions of high curvature where the angle between $\mathbf{R}_i - \mathbf{R}_{i-1}$ and $\mathbf{R}_{i+1} - \mathbf{R}_i$ is large.

When this modified NEB method is applied to the system of Fig. 3.1 the band is well behaved as shown in Fig. 3.4. The energy and force of the NEB images is shown in Fig. 3.5 along with the exact MEP obtained by moving along the gradient from the saddle point. The thin line though the points is an interpolation which involves both the energy and the force along the elastic band. The details of the interpolation procedure is discussed in the Appendix.

The motivation for this modified tangent came from a stable method for finding the MEP from a given saddle point. A good way to do this is to displace the system by some increment from the saddle point and then minimize the energy while keeping the distance between the system and the saddle point configuration fixed. This gives one more point along the MEP, say M_1. Then, the system is displaced again by some increment and minimized keeping the distance to the point M_1 fixed, etc. The important fact is that the MEP can be found by following force lines down the potential from the saddle point, but never up from a minimum. If a force line is followed up from a minimum, it will most likely not go close to the saddle point. This motivated the choice for the local tangent to be determined by the higher energy neighboring image in the NEB method.
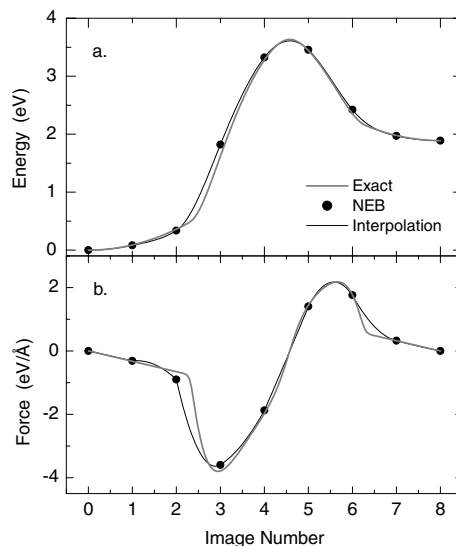
Figure 3.5: The energy and force of the converged elastic band for the two-dimensional LEPS potential shown in Fig. 3.4. An interpolation between the images (thin line), see Appendix, agrees well with the exact minimum energy path (thick, gray line).

3.5.1   Exchange Diffusion Process in Si Crystal

A particularly severe problem with kinks had been noticed in calculations of self-diffusion in a Si crystal. For example, one possible mechanism is the exchange of two atoms on adjacent lattice sites [53]. Both calculations using DFT and empirical potentials could not converge the force because of kinks. In calculations using 16 images and the Tersoff potential [54], the force fluctuations remained at a level of 0.5 eV/Å (the average total force on all the images). When a switching function was used to add a fraction of the perpendicular component of the spring force, the saddle point estimate was clearly strongly dependent on how much spring force was added, indicating a problem with corner-cutting. The reason why diffusion processes in Si are particularly difficult is that the energy changes rapidly as the covalent bonds get broken, but the lattice is quite open and there is a small restoring force for bending bonds. The problem becomes worse when the Tersoff potential function is used because the potential energy surface has several bumps and minima, presumably because of sharply varying cut-off functions. The MEP takes sharp detours into these local minima and this can lead to large angles between neighboring images.

Figure 3.6 shows a calculation carried out with the new formulation of the NEB. The new tangent was found to be resilient to the large angles on the MEP, some larger than
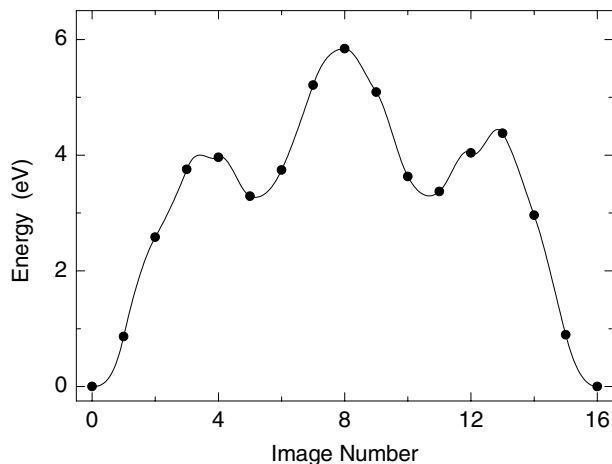
Figure 3.6: The minimum energy path for the Pandey exchange process in silicon crystal modeled with the Tersoff potential. While the original formulation of the nudged elastic band did not converge to the minimum energy path in this case unless a large fraction of the perpendicular spring force was included (leading to an overestimate of the activation energy), the modified formulation with the new tangent converges well. The potential energy surface has several small ripples and local minima. The local minima near images 4 and 12, suggested by the interpolated energy curve, are indeed present on the Tersoff potential energy surface.

90° between adjacent images. The calculation converged to <0.01 eV/Å in 500 iterations. The minima near images 5 and 11 in Fig. 3.6 are similar to results obtained by DFT, and probably correspond to true metastable configurations of the rotating pair of Si atoms. The small dips predicted by the interpolation near images 4 and 12 are indeed minima on the Tersoff potential surface, but do not show up in DFT calculations and are likely artifacts in the empirical potential. The presence of the intermediate minima illustrate the importance of getting information about the full MEP, rather than just converging on the saddle point nearest to the initial state of the transition.

3.6 Combining the Dimer and NEB Methods

When estimating rates in hTST, the activation energy is determined only by the highest energy along the MEP. The full MEP is only needed to locate the highest maximum, when two or more maxima are present. Finding the precise value of the energy at the maximum saddle point can be tedious with the NEB method. Enough images need to be included to get high enough resolution near the maximum for the interpolation to give an accurate value. Also, many force evaluations can be wasted converging images far from the saddle point, which in the end are not relevant. It can, therefore, be advantageous to first use only

a few iterations of the NEB method, enough to get a rough estimate of the shape of the MEP, and then turn to another method which can efficiently converge on the highest saddle point.

The dimer method [2] is a good candidate for this co-operative strategy. Similar to the NEB method, it requires only forces to find a saddle point from an initial configuration. We have used the interpolation scheme discussed in the Appendix to estimate the coordinates of the saddle point configuration from the two images closest to the maximum. This configuration was then used as a starting point for the dimer method. The tangent to the path at the estimated saddle point (obtained from the interpolation) was also used as the initial orientation of the dimer, i.e., as a guess for the direction of the lowest curvature normal mode. The dimer method was implemented as described in Ref. [2].

3.6.1   Formation of Al Addimer on Al(100) Surface

The strategy of combining the NEB and dimer methods was tested on a process where an addimer forms on the Al(100) surface. A surface atom gets pulled out of the surface layer and ends up forming an addimer with an adatom that is present in the initial state. The atomic interactions were described by an empirical embedded atom method (EAM)[31] potential. An extensive study of the various transitions involving an adatom on this surface have been calculated with the dimer method and are described elsewhere[2]. We chose this process because it is complex and the saddle point is quite far removed from a straight line interpolation between the initial and final states. The process involves an intermediate state that is barely stable (see Fig. 3.7). First, the adatom gets incorporated into the surface to form a shallow but stable structure where a group of four surface atoms forming a square has rotated by 45°. Then, an addimer is formed leaving a vacancy in the surface layer. Figure 3.7 shows a fully converged NEB with eight images, the interpolated energy and the energy obtained by descending from the saddle point along the MEP. The eight image NEB clearly gives a good representation of the MEP.

We carried out a systematic study of the efficiency of combining the NEB and dimer methods. Figure 3.8 shows the results. First, the NEB method was run to convergence, starting from a straight line interpolation between the initial and final states. This took 50 iterations. The computational effort is measured by the number of force evaluations, since
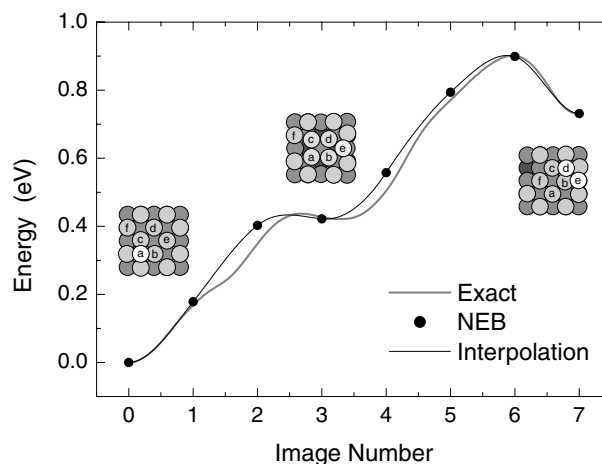
Figure 3.7: The minimum energy path for a process where an Al adatom on an Al(100) surface bonds with one of the surface atoms to form an addimer. An EAM potential is used to describe the interatomic interactions. The path is highly curved. The insets show the atomic configuration in initial, final and intermediate stable state. The best estimate of the minimum energy path, obtained by sliding down the steepest descent path from the saddle point is shown (thick, gray line) as well as the energy of an elastic band with six movable images (filled circles) and interpolation between the images (solid line).

over 90% of the computer time is used to evaluate the force on the various configurations. The straight line in Fig. 3.8 simply shows how the total number of force evaluations increases as more iterations are performed in the NEB calculation (since there are six movable images in the band, it takes six force evaluations to carry out one iteration of the NEB method). Configurations were saved every five iterations of the NEB calculation, and for each the interpolation was used to obtain an estimate of the saddle point configuration. From this configuration a dimer calculation was started and run until the saddle point was found to within the given tolerance of 1 meV/Å. The number of force evaluations needed in the dimer calculation is given in Fig. 3.8, each point located along the x axis at the number of NEB iterations that had been carried out before the dimer calculation was started. The first point, corresponding to zero NEB iterations, corresponds to a calculation where the energy and force were calculated for the eight images along the straight line interpolation from the initial to the final states (including end points). The saddle point configuration was estimated from this data and it took 420 force evaluations for the dimer method to converge to the saddle point. The total number of force evaluations for this calculation was, therefore, 428. This entire procedure was repeated starting from an elastic band configuration obtained after five NEB iterations, etc. The most efficient procedure for this particular case, corresponding to
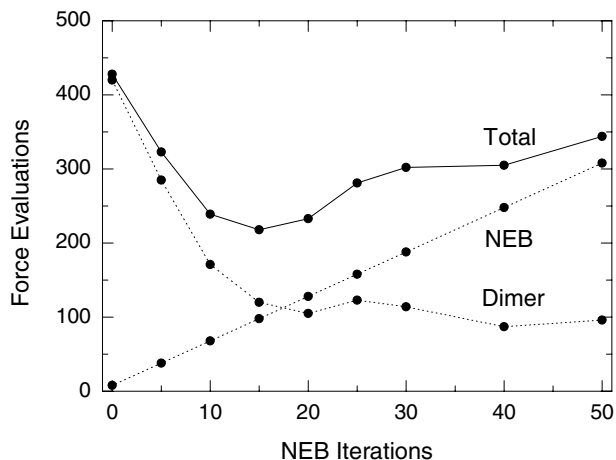
Figure 3.8: The computational effort (given as number of force evaluations) involved in finding the saddle point for the Al addimer formation process shown in Fig. 3.7. The x axis shows the number of iterations in a nudged elastic band calculation with six movable images, starting from a straight line interpolation between the initial and final states. The slope of the straight dashed line labeled "NEB" is the six force evaluations needed per iteration in the minimization of the elastic band. After every five iterations the interpolation procedure described in the Appendix was used to obtain an estimate of the saddle point energy and configuration, and this was used as a starting point for a search for the saddle point using the dimer method. The dashed curve labeled "dimer" gives the number of force evaluations needed to converge on the saddle point in each case. The total number of force evaluations, from the initial $5n$ steps of the elastic band minimization plus the dimer calculation, is given by the solid line labeled "Total". The results show that an optimal strategy for finding the saddle point in this case is to carry out 15 iterations of the elastic band minimization, and then a dimer search.

a little over 200 force evaluations, involved 15 NEB iterations and then convergence to the saddle point with the dimer method.

The two methods complement each other well. When the final state is known, this provides important information for the saddle point search. The NEB method incorporates this information. The dimer method can be used to search for a saddle point starting in the vicinity of the initial state minimum, but in a high-dimensional system with multiple intermediate minima, the calculation of a complete MEP with the dimer method is tedious. The dimer method would converge on the first saddle point near the initial state, then a search for the intermediate minimum has to be carried out, followed by another dimer search for the next saddle point, etc. When the final state is known, using the NEB method for at least a few iterations is better than solely using the dimer method from the initial or final state. The NEB gives a nonlocal picture of the potential energy landscape, and can reveal intermediate stable states which can be important for understanding the basic mechanism of the transition and possibly suggest different mechanisms which lead to alternate final

states. The interpolation using the force along the elastic band makes the NEB method able to reveal such intermediate minima even when very few images are included in the band. Ultimately, it is the energy of the highest saddle point that is needed, and unless one of the NEB images happens to land very close to the saddle point, the estimate of the saddle point energy from the NEB calculation relies on interpolation. Then, a more rigorous convergence to the saddle point can be obtained with the dimer method.

3.6.2 Dissociation of CH_4 on $Ir(111)$

We have used the combination of NEB and dimer methods in plane wave based DFT [55, 56, 57]calculations of the dissociative adsorption of a CH_4 molecule on an Ir(111) surface. The VASP code was used for the calculations [58, 59, 60, 61]. We used the PW91 functional [62] and ultrasoft pseudopotentials [63]. A parallel implementation of the original formulation of the NEB method is included in the current VASP code. We modified the code to incorporate the new formulation [Eqs. (3.8), (3.10) and Eq. (3.12)]. We have also implemented the dimer method in the VASP code, making use of the natural parallelization over two processors (one for each image in the dimer). An Ir slab with five close packed layers was simulated with the top three layers of the slab allowed to relax while the bottom two were kept fixed in bulk crystal configuration. Each layer contained 12 atoms.

The results of calculations using the original formulation of the NEB and the new formulation are shown in Fig. 3.9. The H^- and CH_3^- adsorbed at adjacent on-top sites correspond to reaction coordinate of 0.0 in the figure. The CH_4 molecule is 4.0 Å above the Ir surface when the reaction coordinate is 1.0. Moving the molecule further away from the surface did not change the energy. The DFT/PW91 calculations include correlation of the electrons only semilocally. The long range dispersion interaction at distances where the overlap of the electron density of the Ir surface and the CH_4 molecule is small and is not included in the DFT/PW91 calculation. There is, therefore, no physisorption well evident in the results shown in Fig. 3.9. The activation energy barrier, however, occurs when the electron clouds overlap strongly, and the semilocal description of correlation should be adequate there. The height of the activation energy barrier, 0.4 eV, is in reasonable agreement with the activation energy deduced from experimental measurements [8]. A more detailed analysis of the calculation and comparison with experiment will be given elsewhere [7].
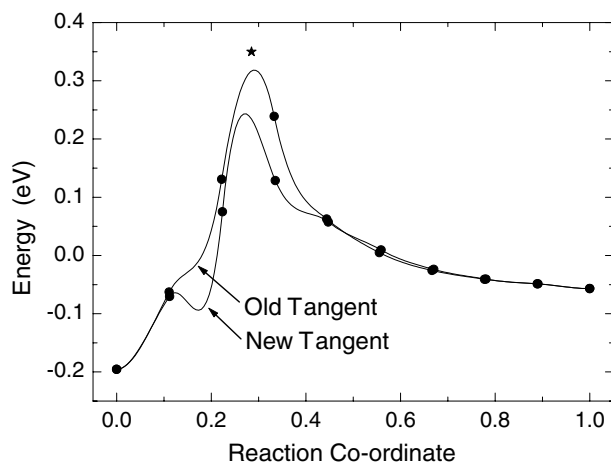
Figure 3.9: A density functional theory calculation of the dissociative adsorption of a CH_4 molecule on an Ir(111) surface. Converged elastic bands with 8 movable images are shown using both the original way of estimating the tangent, and using the new way of estimating the tangent. An interpolation using the method described in the Appendix is shown with a solid line. With the new tangent, the interpolation correctly predicts an intermediate minimum (near reaction coordinate of 0.2). The resolution of the barrier is, however, too poor and the estimate of the saddle point obtained from the interpolation is much too low. The saddle point energy ($\star$) was found with a dimer calculation started from the estimated saddle point configuration given by the interpolation.

A comparison of the results from the new and old formulation of the NEB is shown in Fig. 3.9. Since the barrier is quite narrow compared with the length of the path from the free molecule to the dissociated fragments, the number of images in the barrier region is low. The elastic band obtained from the new formulation predicts a stable local minimum in the region of reaction coordinate 0.2. There is a minimum in the interpolated MEP due to the value of the parallel force at the images. This minimum is indeed there, only deeper, and was verified by a separate calculation where the system coordinates where simply minimized starting from the prediction obtained from the interpolation between the elastic band images. We have carried out several calculations of the dissociation with varying number of layers in the slab and various numbers of atoms in each slab, both of which have significant effect on the barrier height, and in all those cases the minimum is predicted by the new formulation but not by the old one. This illustrates how much closer the elastic band with the new tangent follows the MEP.

The number of images in the barrier region is, however, too small to predict accurately the barrier height from the interpolation. This is a good example of where the combined NEB and dimer approach is advantageous. The dimer method gave a saddle point energy

that is higher than either interpolation predicts, as shown in Fig. 3.9.

3.7 Summary

We have presented a new formulation of the NEB method which is more stable than the previous formulation. With a different way of estimating the local tangent to the elastic band, the problems with kinks on the elastic band can be eliminated. While in many systems this is was not a serious problem, there were examples where the previous procedure of including part of the perpendicular spring force led to serious overestimate of the saddle point energy. The exchange of two Si atoms in the Si crystal is one example. With the new formulation, the calculation of the MEP for the exchange process is straight forward.

We have also described a strategy where the NEB is combined with the dimer method. A few iterations using the NEB, enough to get a rough estimate of the shape of the MEP, followed by interpolation between the images and then a search for the saddle point using the dimer method was found to cut the overall computational effort in half in a test problem involving dimer formation on an Al(100) surface. This strategy was also applied to a calculation of methane dissociation on an Ir(111) surface using plane wave based DFT calculations.

3.8 Appendix

The analysis of the results of an NEB calculation requires an interpolation between the images in order to get estimates of the coordinates of atoms and the energy at maxima and minima along the MEP. In addition to the energy at each image, the force parallel to the band is also known and provides useful information for the interpolation. From a good interpolation scheme, the presence of maxima and minima on the MEP can be deduced from remarkably few images. In some cases, the use of the force can lead to the identification of extrema in the MEP which could not be seen when only the energy data were used in the interpolation.

We have used a cubic polynomial to represent the MEP between each pair of adjacent images. There are four parameters in the polynomial which are determined by matching the energy and force at ends of the interval. Letting the length of the ith segment be R_i

and writing the polynomial as $a_i x^3 + b_i x^2 + c_i x + d_i$ the parameters are

$$
\begin{aligned}
a_i &= \frac{2V_{i+1} - V_i}{R^3} - \frac{F_i + F_{i+1}}{R^2} \\
b_i &= \frac{3V_{i+1} - V_i}{R^2} + \frac{2F_i + F_{i+1}}{R} \\
c_i &= -F_i \\
d_i &= V_i.
\end{aligned}
\tag{3.13}
$$

where the energy at the images is V_i and V_{i+1} and parallel force is F_i and F_{i+1}.

This interpolation is typically quite smooth. However, unlike traditional cubic spline interpolations [29], the second derivative is not continuous at the images. A close look at the interpolation in Fig. 3.5 shows the discontinuity in the slope of the force. A possible improvement in the scheme described above is to generate a quintic polynomial interpolation so that the second derivatives can also be made continuous (and set to zero at the end points for a natural spline). It was found that although this did generate a smoother interpolation, it sometimes added spurious oscillations in between the images. If a continuous second derivative is not essential, the simple cubic interpolation described above is probably preferred.

The interpolation scheme described above can also be used to estimate the configuration of the system at various points along the MEP. Each atomic coordinate, $\mathbf{R}_i^j$, is a function of the distance along the path, s. The slope $d\mathbf{R}_i/ds$ is taken to be along the force at the NEB images, $d\mathbf{R}_i^j/ds = \mathbf{F}_i^j/|\mathbf{F}_i|$, except for the endpoints. At the endpoints, the force is zero and the direction of the path is, therefore, not well defined. There, we assumed the path was pointing towards the neighboring images. The equations for the coordinate interpolation are analogous to those given above for the energy interpolation, with the modifications $V_i \to \mathbf{R}_i$, and $F_i \to -d\mathbf{R}_i/ds = -\mathbf{F}_i/|\mathbf{F}_i|$. Each line in Eq. (3.13) becomes a vector equation.

3.9   Acknowledgments

Chapter 4

A CLIMBING IMAGE NEB METHOD FOR FINDING SADDLE POINTS AND MINIMUM ENERGY PATHS

4.1 Abstract

A modification of the nudged elastic band method for finding minimum energy paths is presented. One of the images is made to climb up along the elastic band to converge rigorously on the highest saddle point. Also, variable spring constants are used to increase the density of images near the top of the energy barrier to get an improved estimate of the reaction coordinate near the saddle point. Applications to CH_4 dissociative adsorption on Ir(111) and H_2 on Si(100) using plane wave based density functional theory are given.

4.2 Introduction

An important problem in theoretical chemistry and condensed matter physics is the calculation of transition rates, for example rates of chemical reactions or diffusion events. Most often, it is sufficient to treat the motion of the atoms using classical mechanics, but the transitions of interest are typically many orders of magnitude slower than vibrations of the atoms, so a direct simulation of the classical dynamics is not feasible. For a process with a typical, low activation energy of 0.5 eV, the computer time required to simulate a classical trajectory long enough that a single transition event can be expected to occur is on the order of 10^4 years on present day computers. This "rare event" problem is devastating for direct dynamical simulations, but makes it possible to obtain accurate estimates of transition rates using a purely statistical approach, namely, transition state theory (TST) [64, 65, 66, 1]. Apart from the Born–Oppenheimer approximation, TST relies on two basic assumptions: (a) the rate is slow enough that a Boltzmann distribution is established and maintained in the reactant state, and (b) a dividing surface of dimensionality D-1, where D is the number degrees of freedom in the system, can be identified such that a reacting trajectory going

from the initial state to the final state only crosses the dividing surface once. The dividing surface must, therefore, represent a bottleneck for the transition.

Since atoms in crystals are usually tightly packed and the typical temperature of interest is low compared with the melting temperature, the harmonic approximation to TST (hTST) can typically be used in studies of diffusion and reactions in crystals or at crystal surfaces [67, 68]. This greatly simplifies the problem of estimating the rates. The search for the optimal transition state then becomes a search for the lowest few saddle points at the edge of the potential energy basin corresponding to the initial state. The rate constant for transition through the region around each one of the saddle points can be obtained from the energy and frequency of normal modes at the saddle point and the initial state [11, 12],

$$k^{\mathrm{hTST}} \;=\; \frac{\Pi_i^{3N} \nu_i^{\mathrm{init}}}{\Pi_i^{3N-1} \nu_i^{\ddagger}} \; e^{-(E^{\ddagger} - E^{\mathrm{init}})/k_B T}. \tag{4.1}$$

Here, $E^{\ddagger}$ is the energy of the saddle point, E^{init} is the local potential energy minimum corresponding to the initial state, and the ν_i are the corresponding normal mode frequencies. The symbol $\ddagger$ refers to the saddle point. All the quantities can be evaluated from the potential energy surface, at zero temperature, but entropic effects are included through the harmonic approximation. The most challenging part in this calculation is the search for the relevant saddle point.

A path connecting the initial and final states that typically has the greatest statistical weight is the minimum energy path (MEP). At any point along the path, the force acting on the atoms is only pointing along the path. The energy is stationary for any perpendicular degree of freedom. The maxima on the MEP are saddle points on the potential energy surface. The distance along the MEP is a natural choice for a reaction coordinate, and at the saddle point the direction of the reaction coordinate is given by the normal mode eigenvector corresponding to a negative curvature.

The MEP often has one or more minima in addition to the minima at the initial and final states. These correspond to stable intermediate configurations. The MEP will then have two or more maxima, each one corresponding to a saddle point. Assuming a Boltzmann population is reached for the intermediate (meta)stable configurations, the overall rate is determined by the *highest* saddle point. It is, therefore, not sufficient to find *a* saddle point.

One needs to have a good enough estimate of the shape of the MEP to be able to assign the highest saddle point as ‡ in Eq. (4.1) in order to get an accurate estimate of the rate.

Many different methods have been presented for finding MEPs and saddle points [69, 13, 5]. Since a first order saddle point is a maximum in one direction and a minimum in all other directions, methods for finding saddle points invariably involve some kind of maximization of one degree of freedom and minimization in other degrees of freedom. The critical issue is to find a good and inexpensive way to decide which degree of freedom should be maximized.

The nudged elastic band (NEB) method is an efficient method for finding the MEP between a given initial and final state of a transition [13, 36, 37]. It has become widely used for estimating transition rates within the hTST approximation. The method has been used both in conjunction with electronic structure calculations, in particular plane wave based DFT calculations (see, for example, Refs. [38, 39, 40, 41, 42]), and in combination with empirical potentials [43, 44, 45, 46]. Studies of very large systems, including over a million atoms in the calculation, have been conducted [47]. The MEP is found by constructing a set of images (replicas) of the system, typically on the order of 4–20, between the initial and final state. A spring interaction between adjacent images is added to ensure continuity of the path, thus mimicking an elastic band. An optimization of the band, involving the minimization of the force acting on the images, brings the band to the MEP.

An essential feature of the NEB method, which distinguishes it from other elastic band methods [48, 49, 50, 51], is a force projection which ensures that the spring forces do not interfere with the convergence of the elastic band to the MEP, as well as ensuring that the true force does not affect the distribution of images along the MEP. It is necessary to estimate the tangent to the path at each image and every iteration during the minimization, in order to decompose the true force and the spring force into components parallel and perpendicular to the path. Only the perpendicular component of the true force is included, and only the parallel component of the spring force. This force projection is referred to as "nudging". The spring forces then only control the spacing of the images along the band. When this projection scheme is not used, the spring forces tend to prevent the band from following a curved MEP (because of "corner-cutting"), and the true force along the path causes the images to slide away from the high energy regions towards the minima, thereby

reducing the density of images where they are most needed (the "sliding-down" problem). In the NEB method, there is no such competition between the true forces and the spring forces; the strength of the spring forces can be varied by several orders of magnitude without effecting the equilibrium position of the band.

Recently, an improved way of estimating the tangent to the elastic band at each image has been presented [3]. This eliminates a problem which occurred in systems where the force parallel to the MEP was very large compared with the restoring force perpendicular to the MEP [13]. In such situations kinks could form on the elastic band and prevent rigorous convergence to the MEP. We use this new way of estimating the tangent in the calculations presented here.

While the NEB method gives a discrete representation of the MEP, the energy of saddle points needs to be obtained by interpolation. When the energy barrier is narrow compared with the length of the MEP, few images land in the neighborhood of the saddle point and the interpolation can be inaccurate.

This communication describes a modification of the NEB method which gives a precise estimate of the saddle point at no extra cost as compared with the regular NEB.

4.3 DFT Calculations of Dissociative Adsorption

The method presented here has been applied to calculations of CH_4 dissociative adsorption on the Ir(111) surface and H_2 on the Si(100) surface using plane wave based density functional theory (DFT) [55, 56]. The PW91 functional [57, 62] was used in combination with ultrasoft pseudopotentials [63]. The energy cutoff was 350 eV in the CH_4/Ir(111) calculation and 200 eV in the H_2/Si(100) calculation. The calculations were carried out with the VASP code [58, 59, 60, 61] which we have extended to implement the new method presented here. The calculations were carried out in parallel on a cluster of workstations. The NEB method lends itself so well to parallel processing that it is sufficient to use a regular ethernet connection to transfer data between the nodes.

4.4 Regular NEB Method

An elastic band with $N + 1$ images can be denoted by $[\mathbf{R}_0, \ \mathbf{R}_1, \ \mathbf{R}_2, \ldots, \mathbf{R}_N]$, where the end points, $\mathbf{R}_0$ and $\mathbf{R}_N$, are fixed and given by the energy minima corresponding to the initial and final states. The $N - 1$ intermediate images are adjusted by the optimization algorithm.

In the NEB method [13, 3], the total force acting on an image is the sum of the spring force along the local tangent and the true force perpendicular to the local tangent

$$\mathbf{F}_i = \mathbf{F}_i^s \mid_\parallel - \nabla E(\mathbf{R}_i) \mid_\perp, \tag{4.2}$$

where the true force is given by

$$\nabla E(\mathbf{R}_i) \mid_\perp = \nabla E(\mathbf{R}_i) - \nabla E(\mathbf{R}_i) \cdot \hat{\tau}_i. \tag{4.3}$$

Here, E is the energy of the system, a function of all the atomic coordinates, and $\hat{\tau}_i$ is the normalized local tangent at image i. The spring force is

$$\mathbf{F}_i^s \mid_\parallel = k \left(|\mathbf{R}_{i+1} - \mathbf{R}_i| - |\mathbf{R}_i - \mathbf{R}_{i-1}| \right) \hat{\tau}_i, \tag{4.4}$$

where k is the spring constant. An optimization algorithm is then used to move the images according to the force in Eq. (4.2). We have used a projected velocity Verlet algorithm [13]. The images converge on the MEP with equal spacing if the spring constant is the same for all the springs. Typically none of the images lands at or even near the saddle point and the saddle point energy needs to be estimated by interpolation.

An example of a NEB calculation is shown in Fig. 4.1. The MEP for dissociative adsorption of CH_4 on an Ir(111) surface has a narrow barrier compared with the length of the MEP. The molecule is 4 Å above the surface when the reaction coordinate is 1.0. At the other end, at 0.0 the molecule has broken up into a H^- and a CH_3^- fragment sitting on adjacent on-top sites on the Ir(111) surface. The resolution of the MEP near the saddle point is poor and the estimate of the activation energy obtained from the interpolation is subject to large uncertainty. A force and energy based cubic polynomial interpolation
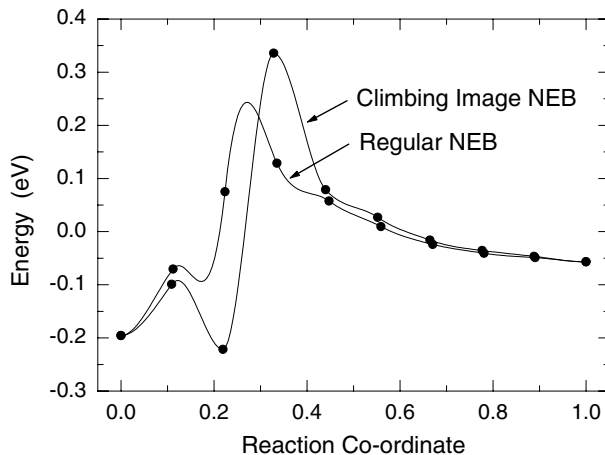
Figure 4.1: Density functional theory calculations of the minimum energy path for CH$_4$ dissociative adsorption on a Ir(111) surface. The dissociated H$^-$ and CH$_3^-$ fragments sitting on adjacent on-top sites correspond to reaction coordinate of 0.0. The CH$_4$ molecule 4 Å away from the surface corresponds to 1.0. A regular NEB calculation and a climbing image NEB calculation are compared, both involving 8 movable images. The regular NEB results in a low resolution of the barrier, and the interpolation gives an underestimate of the activation energy. The climbing image NEB brings one of the images right to the saddle point and gives the activation energy precisely with insignificant additional computational effort.

was used between each pair of adjacent images. This is an example of a system where an intermediate minimum is located along the MEP. In fact, it turns out that this minimum is deeper than the chemisorbed state at the 0.0 end point. The configuration corresponding to the intermediate minimum has the adsorbed H atom at a bridge site.

4.5 Climbing Image NEB Method

The climbing image NEB (CI-NEB) method constitutes a small modification to the NEB method. Information about the shape of the MEP is retained, but a rigorous convergence to a saddle point is also obtained. This additional feature does not add any significant computational effort. After a few iterations with the regular NEB, the image with the highest energy $i_{\max}$ is identified. The force on this one image is not given by Eq. (4.2) but rather by

$$\mathbf{F}_{i_{\max}} = -\nabla E(\mathbf{R}_{i_{\max}}) + 2\nabla E(\mathbf{R}_{i_{\max}}) \mid_{\parallel} = -\nabla E(\mathbf{R}_{i_{\max}}) + 2\nabla E(\mathbf{R}_{i_{\max}}) \cdot \hat{\tau}_{i_{\max}} \hat{\tau}_{i_{\max}}. \quad (4.5)$$

This is the full force due to the potential with the component along the elastic band inverted. The maximum energy image is not affected by the spring forces at all.

Qualitatively, the climbing image moves up the potential energy surface along the elastic band and down the potential surface perpendicular to the band. The other images in the band serve the purpose of defining the one degree of freedom for which a maximization of the energy is carried out. Since the images in the band eventually converge to the MEP, they give a good approximation to the reaction coordinate around the saddle point. As long as the CI-NEB method converges, the climbing image will converge to the saddle point. Since all the images are being relaxed simultaneously, there is no additional cost of turning one of the images into a climbing image.

The results of a CI-NEB calculation of the CH_4 dissociation on Ir(111) is shown in Fig. 4.1. A significantly higher estimate of the activation energy is obtained than with the regular NEB, using the same number of images. The computational effort is the same to within 10% (CI-NEB not necessarily being slower). Alternatively, one could have run a second elastic band between the two images adjacent to the barrier to get a better estimate of the saddle point energy from the regular NEB, but this would have required more force evaluations and, therefore, more computational effort. The activation energy predicted by the DFT/PW91 calculations is approximately 0.4 eV. This calculation sill needs to be corrected [7] for quantum zero point energy, dispersion, and system size effects before it can compared to the experimental value [8] of 0.28 eV. The MEP is nontrivial because it involves a large relaxation of the substrate. The Ir atom closest to the CH_4 molecule in the transition state is pulled out from the surface plane by 0.5 Å. This means that the saddle point does not lie close to the straight line interpolation between the two endpoints. A more detailed presentation of the DFT calculations and comparison with experimental results will be given elsewhere [7].

The climbing image is not affected by the spring forces. Therefore, the spacing of the images will be different on each side of the climbing image. As it moves up to the saddle point, images on one side will get compressed, and on the other side spread out. Two or more climbing images can be specified if the MEP appears to have two or more high maxima that are close in energy. The only issue is to have enough images close to the climbing image(s) to get a good estimate of the reaction coordinate, since this determines the climbing direction.

4.6   Variable Spring Constants

Since the saddle point is the most important point along the MEP, one would typically prefer to have more resolution in the MEP close to the saddle point than near the end points. The important issue is to get a good enough estimate of the tangent to the path near the saddle point, especially when a climbing image is included. As the images are brought closer to the saddle point, the approximation of the tangent will become more accurate. Dissociative adsorption of a molecule on a surface is an example of a process where the MEP is often highly asymmetric and the barrier region is only a small fraction of the MEP (see Figs. 4.1 and 4.2). In such cases, it is more efficient to distribute the images unevenly along the MEP.

This can be accomplished by using stronger springs near the saddle point. Because of the nudging, there is no interference between the spring forces that distribute the images along the MEP and the true force that brings the elastic band to the MEP. One is, therefore, free to choose different spring constants between different pairs of images without affecting the convergence of the band to the MEP, as long as the number of images is high enough. We have used a scheme where the spring constant depends linearly on the energy of the images, in such a way that images with low energy get connected by a weaker spring constant

$$
k_i' \;=\; \begin{cases} k_{\max} - \Delta k \ \left(\frac{E_{\max} - E_i}{E_{\max} - E_{\mathrm{ref}}} \right) & \text{if} \quad E_i > E_{\mathrm{ref}} \\ k_{\max} - \Delta k & \text{if} \quad E_i < E_{\mathrm{ref}} \end{cases} \tag{4.6}
$$

Here, $E_i = \max\{E_i, E_{i-1}\}$ is the higher energy of the two images connected by spring i, $E_{\max}$ is the maximum value of E_i for the whole elastic band, and E_{ref} is a reference value for the energy, defining a minimum value of the spring constant. We have chosen E_{ref} to be the energy of the higher energy end point of the MEP. This choice ensures that the density of images is roughly equal near the two endpoints, even for highly asymmetric MEPs. The spring constant is, therefore, linearly scaled from a maximum value $k_{\max}$ for highest energy images to a minimum value $k_{\max} - \Delta k$ for images with energy of E_{ref} or lower.

By choosing E_i to be the higher energy of the two images connected by the spring, the two images adjacent to the climbing image will tend to be symmetrically arranged around the saddle point. This is only approximately true because of the compression/stretching of the band on each side of the climbing image.
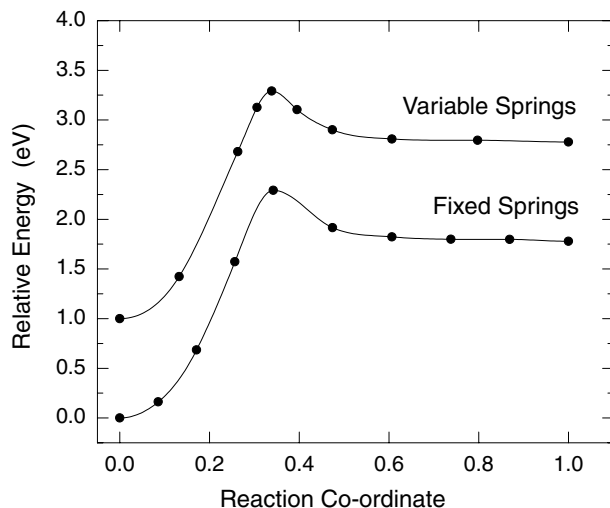
Figure 4.2: Density functional theory calculations of the minimum energy path for H_2 dissociative adsorption on a Si(100) surface. The H^- adatoms sitting on adjacent Si atoms in a surface dimer correspond to reaction coordinate of 0.0. The H_2 molecule 3.8 Å away from the surface corresponds to 1.0. A regular climbing image NEB calculation with equal spring constants (curve labeled "Fixed Springs") is compared with a calculation where the spring constants are scaled with the energy (curve labeled "Variable Springs", arbitrarily shifted by 1.0 eV). Both calculations involve 8 movable images. The variable spring calculation results in a higher resolution of the barrier with insignificant additional computational effort.

Figure 4.2 shows results of a calculation of the dissociation of a H_2 molecule on the Si(100) surface. This is an interesting system because of a long standing discrepancy [70] between experimental and theoretical measurements of adsorption and dissociation barriers. A CI-NEB calculation with equal spring constants is compared with a CI-NEB calculation with variable spring constants. The region where the H_2 molecule approaches the Si(100) surface is flat and rather uninteresting. The energy scaling of the spring constants results in images being pulled up towards the barrier region, thus increasing the resolution of the MEP near the saddle point at the expense of the less important regions.

The number of force evaluations required to reach convergence to within a tolerance of 0.03 eV/Å was 179 for the regular NEB method (first 13 iterations with a large time step until the magnitude of the force had dropped below 1 eV/Å and then 166 iterations with a larger time step in the projected velocity Verlet algorithm [13]). The number of force evaluations needed for the CI-NEB calculation with equal spring constants was 190, and the number of force evaluations needed for the CI-NEB calculation with variable spring constants was 178. The difference between these numbers is not significant but simply reflects slight variations in the way the system moves on the energy surface towards the

MEP.

4.7 Acknowledgments

Chapter 5

METHODS FOR FINDING SADDLE POINTS AND MINIMUM ENERGY PATHS

5.1 Abstract

The problem of finding minimum energy paths and, in particular, saddle points on high dimensional potential energy surfaces is discussed. Several different methods are reviewed and their efficiency compared on a test problem involving conformational transitions in an island of adatoms on a crystal surface. The focus is entirely on methods that only require the potential energy and its first derivative with respect to the atom coordinates. Such methods can be applied, for example, in plane wave based Density Functional Theory calculations, and the computational effort typically scales well with system size. When the final state of the transition is known, both the initial and final coordinates of the atoms can be used as boundary conditions in the search. Methods of this type include the Nudged Elastic Band, Ridge, Conjugate Peak Refinement, Drag method and the method of Dewar, Healy and Stewart. When only the initial state is known, the problem is more challenging and the search for the saddle point represents also a search for the optimal transition mechanism. We discuss a recently proposed method that can be used in such cases, the Dimer method.

5.2 Introduction

A common and important problem in theoretical chemistry and in condensed matter physics is the calculation of the rate of transitions, for example chemical reactions or diffusion events. In either case, the configuration of atoms is changed in some way during the transition. The interaction between the atoms can be obtained from an (approximate) solution of the Schrödinger equation describing the electrons, or from an otherwise determined potential energy function. Most often, it is sufficient to treat the motion of the atoms using classical mechanics, but the transitions of interest are typically many orders of magnitude slower

than vibrations of the atoms, so a direct simulation of the classical dynamics is not useful. This 'rare event' problem is best illustrated by an example. We will be describing below a study of configurational changes in a Pt island on a Pt(111) surface, relevant to the diffusion of the island over the surface. The approximate interaction potential predicts that the easiest configurational change has an activation energy barrier of 0.6 eV. This is a typical activation energy for diffusion on surfaces. Such an event occurs many times per second at room temperature and is, therefore, active on a typical laboratory time scale. But, there are on the order of 10^{10} vibrational periods in between such events. A direct classical dynamics simulation which necessarily has to faithfully track all this vibrational motion would take on the order of 10^5 years of computer calculations on the fastest present day computer before a single diffusion event can be expected to occur! It is clear that meaningful studies of these kinds of events cannot be carried out by simply simulating the classical dynamics of the atoms. It is essential to carry out the simulations on a much longer timescale. This time scale problem is one of the most important challenges in computational chemistry, materials science and condensed matter physics.

The time scale problem is devastating for direct dynamical simulations, but makes it possible to obtain accurate estimates of transition rates using purely statistical methods, namely Transition State Theory (TST) [64, 65, 66, 1, 71]. Apart from the Born-Oppenheimer approximation, TST relies on two basic assumptions: (a) the rate is slow enough that a Boltzmann distribution is established and maintained in the reactant state, and (b) a dividing surface of dimensionality D-1 where D is the number degrees of freedom in the system can be identified such that a reacting trajectory going from the initial state to the final state only crosses the dividing surface once. The dividing surface must, therefore, represent a bottleneck for the transition. The TST expression for the rate constant can be written as

$$k = \frac{\langle |v| \rangle}{2} \frac{Q^{\ddagger}}{Q_R} \tag{5.1}$$

where $\langle |v| \rangle$ is the average speed, $Q^{\ddagger}$ is the configurational integral for the transition state dividing surface, and Q_R is the configurational integral for the initial state. The bottleneck can be of purely entropic origin, but most often in crystal growth problems it is due to a potential energy barrier between the two local minima corresponding to the initial and

final states. It can be shown that TST always overestimates the rate of escape from a given initial state [65, 66] (a diffusion constant can be underestimated if multiple hops are not included in the analysis [67]). This leads to a variational principle which can be used to find the optimal dividing surface [66, 72]. The TST rate estimate gives an approximation for the rate of escape from the initial state, irrespective of the final state. The possible final states can be determined by short time simulations of the dynamics starting from the dividing surface. This can also give an estimate of the correction to transition state theory due to approximation (b), the so called dynamical corrections [73, 68].

Since atoms in crystals are usually tightly packed and the relevant temperatures are low compared with the melting temperature, the harmonic approximation to TST (hTST) can typically be used in studies of diffusion and reactions in crystals [68]. This greatly simplifies the problem of estimating the rates. The search for the optimal transition state then becomes a search for the lowest few saddle points at the edge of the potential energy basin corresponding to the initial state. The rate constant for transition through the region around each one of the saddle points can be obtained from the energy and frequency of normal modes at the saddle point and the initial state [11, 12],

$$k^{hTST} \; = \; \frac{\Pi_i^{3N} \nu_i^{init}}{\Pi_i^{3N-1} \nu_i^{\ddagger}} \; e^{-(E^{\ddagger} - E^{init})/k_B T}. \tag{5.2}$$

Here, $E^{\ddagger}$ is the energy of the saddle point, E^{init} is the local potential energy minimum corresponding to the initial state, and the ν_i are the corresponding normal mode frequencies. The symbol $\ddagger$ refers to the saddle point. The most challenging part in this calculation is the search for the relevant saddle points. Again, the mechanism of the transition is reflected in the saddle point. The reaction coordinate at the saddle point is the direction of the unstable mode (the normal mode with negative eigenvalue). After a saddle point has been found, one can follow the gradient of the energy downhill, both forward and backward, and map out the Minimum Energy Path (MEP), thereby establishing what initial and final state the saddle point corresponds to. The identification of saddle points ends up being one of the most challenging tasks in theoretical studies of transitions in condensed matter.

The MEP is frequently used to define a 'reaction coordinate' [74] for transitions. It can be an important concept for building in anharmonic effects, or even quantum corrections

[71]. The MEP may have one or more minima in between the endpoints corresponding to stable intermediate configurations. The MEP will then have two or more maxima, each one corresponding to a saddle point. Assuming a Boltzmann population is reached for the intermediate (meta)stable configurations, the overall rate is determined by the highest energy saddle point. It is, therefore, not sufficient to find *a* saddle point, but rather one needs to find the *highest* saddle point along the MEP, in order to get an accurate estimate of the rate from hTST.

For systems where one or more atoms need to be treated quantum mechanically, a quantum mechanical extension of TST, so called RAW-QTST, can be used [75, 76]. Zero point energy and tunneling are then taken into account by using Feynman Path Integrals [77]. Since RAW-QTST is a purely statistical theory analogous to classical TST, the path integrals are statistical (involve only imaginary time) and are easy to sample in computer simulations even for large systems. The definition of the transition state needs to be extended to higher dimensions, but otherwise the RAW-QTST calculation for quantum systems is quite similar to the TST calculations for classical systems. A central problem is finding a good reaction coordinate and a good transition state surface. In a harmonic approximation to RAW-QTST, the central problem becomes the identification of saddle points on an effective potential energy surface with higher dimensionality than the regular potential energy surface [75, 76]. The saddle points are often referred to as 'instantons' and the harmonic approximation to RAW-QTST is the so called Instanton Theory [78, 79, 80]. Any method that can be used to locate saddle points efficiently in high dimension, can, therefore, also be useful for calculating rates in quantum systems.

Many different methods have been presented for finding MEPs and saddle points [69, 13]. Since a first order saddle point is a maximum in one direction and a minimum in all other directions, methods for finding saddle points invariably involve some kind of maximization of one degree of freedom and minimization in other degrees of freedom. The critical issue is to find a good and inexpensive estimate of which degree of freedom should be maximized. Below, we give an overview of several commonly used methods in studies of transitions in condensed matter. We then compare their performance on the surface island test problem.

5.3 The Drag Method

The simplest and perhaps the most intuitive method of all is what we will refer to as the Drag method. It actually has many names because it keeps being reinvented. One degree of freedom, the drag coordinate, is chosen and is held fixed while all other D-1 degrees of freedom are relaxed, i.e. the energy of the system minimized in a D-1 dimensional hyperplane. In small, stepwise increments, the drag coordinate is increased and the system is dragged from reactants to products. The maximum energy obtained on the way is taken to be the saddle point energy. Sometimes, a guess for a good reaction coordinate is used as the choice for drag coordinate. This could be the distance between two atoms, for example, atoms that start out forming a bond which ends up being broken. In the absence of such an intuitive choice, the drag coordinate can be simply chosen to be the straight line interpolation between the initial and final state. This is a less biased way and all coordinates of the system then contribute in principle to the drag coordinate. We will follow this second approach, which is illustrated in figure 5.1. We have implemented the Drag method in such a way that the force acting on the system is inverted along the drag coordinate and the velocity Verlet algorithm [81] with a projected velocity is used to simulate the dynamics of the system. The velocity projection is carried out at each time step and ensures that only the component of the velocity parallel to the force is included in the dynamics. When the force and projected velocity point in the opposite direction (indicating that the system has gone over the energy ridge), the velocity is zeroed. This projected velocity Verlet algorithm has been found to be an efficient and simple minimization algorithm for many of the methods discussed here.

The problem with the Drag method is that both the intuitive, assumed reaction coordinate and the unbiased straight line interpolation can turn out to be bad reaction coordinates. They may be effective in distinguishing between reactants and products, but a reaction coordinate must do more than that. A good reaction coordinate should give the direction of the unstable normal mode at the saddle point. Only then does a minimization in all other degrees of freedom bring the system to the saddle point. Figure 5.1 shows a simple case where the drag method fails. As the drag coordinate is incremented, starting from the initial state, $\mathbf{R}$, the system climbs up close to the slowest ascent path. After climbing high
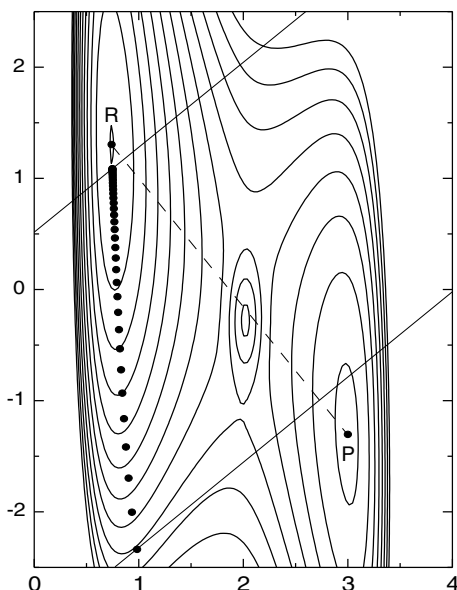
Figure 5.1: The 'drag' method. A drag coordinate is defined by interpolating from **R** to **P** with a straight line (dashed line). Starting from **R**, the drag coordinate is increased stepwise and held fixed while relaxing all other degrees of freedom in the system. In a two-dimensional system, the relaxation is along a line perpendicular to the **P** − **R** vector. The solid lines show the first and last relaxation line in the drag calculation. The final location of the system after relaxation is shown with filled circles. As the drag coordinate is increased, the system climbs up the potential surface close to the slowest ascent path, reaching a potential larger than the saddle point, and then, eventually, slipping over to the product well. In this simple test case, the drag method cannot locate the saddle point.

above the saddle point energy, the energy contours eventually stop confining the system in this energy valley and the system abruptly snaps into an adjacent valley (the product valley in the case of figure 5.1). The system is never confined to the vicinity of the saddle point because the direction of the drag coordinate is at a large angle to the direction of the unstable normal mode at the saddle point. While there certainly are cases where the drag method works, there are also many examples where it does not work [82, 83]. The method failed, for example, on half the saddle points in the surface island test problem described below. What seems to be a more intuitive reaction coordinate, such as the distance between two atoms, can also fail, for example if adjacent atoms also get displaced in going from the initial to final states. As the two atoms get dragged apart, the adjacent atoms can snap from one position to another, never visiting the saddle point configuration. As we will demonstrate below, much more reliable methods exist which are insignificantly more involved to implement or costly to use.

5.4 The NEB Method

In the Nudged Elastic Band (NEB) method [13, 36, 37] a string of replicas (or 'images') of the system are created and connected together with springs in such a way as to form a discrete representation of a path from the reactant configuration, $\mathbf{R}$, to the product configuration, $\mathbf{P}$. Initially, the images may be generated along the straight line interpolation between $\mathbf{R}$ and $\mathbf{P}$. An optimization algorithm is then applied to relax the images down towards the MEP. The NEB and the CPR method, are unique among the methods discussed here in that they do not just give an estimate of the saddle point, but also give a more global view of the energy landscape, for example, showing whether more than one saddle point is found along the MEP.

The string of images can be denoted by $[\mathbf{R}_0,\ \mathbf{R}_1,\ \mathbf{R}_2, \ldots, \mathbf{R}_N]$ where the endpoints are fixed and given by the initial and final states, $\mathbf{R}_0 = \mathbf{R}$ and $\mathbf{R}_N = \mathbf{P}$, but $N-1$ intermediate images are adjusted by the optimization algorithm. The most straightforward approach would be to construct an object function

$$S(\mathbf{R}_1, \ldots, \mathbf{R}_N) \ = \ \sum_{i=1}^{N-1} E(\mathbf{R}_i) \ + \ \sum_{i=1}^{N} \frac{k}{2}(\mathbf{R}_i - \mathbf{R}_{i-1})^2 \qquad (5.3)$$

and minimize with respect to the intermediate images, $\mathbf{R}_1, \ldots, \mathbf{R}_N$. This mimics an elastic band made up of $N-1$ beads and N springs with spring constant k. The band is strung between the two fixed endpoints. The problem with this formulation is that the elastic band tends to cut corners and gets pulled off the MEP by the spring forces in regions where the MEP is curved. Also, the images tend to slide down towards the endpoints, giving lowest resolution in the region of the saddle point, where it is most needed [13]. Both the corner-cutting and the sliding-down problems can be solved easily with a force projection. This is what is referred to as 'nudging'. The reason for corner-cutting is the component of the spring force perpendicular to the path, while the reason for the down-sliding is the parallel component of the true force coming from the interaction between atoms in the system. Given an estimate of the unit tangent to the path at each image (which will be discussed later), $\hat{\tau}_i$, the force on each image should only contain the parallel component of the spring

force, and perpendicular component of the true force

$$\mathbf{F}_i \;=\; -\nabla E(\mathbf{R}_i)|_\perp \;+\; \mathbf{F}_i^s \cdot \hat{\tau}_i \, \hat{\tau}_i \tag{5.4}$$

where $\nabla E(\mathbf{R}_i)$ is the gradient of the energy with respect to the atomic coordinates in the system at image i, and $\mathbf{F}_i^s$ is the spring force acting on image i. The perpendicular component of the gradient is obtained by subtracting out the parallel component

$$\nabla E(\mathbf{R}_i)|_\perp \;=\; \nabla E(\mathbf{R}_i) \;-\; \nabla E(\mathbf{R}_i) \cdot \hat{\tau}_\| \, \hat{\tau}_\| \tag{5.5}$$

In order to ensure equal spacing of the images (when the same spring constant, k, is used for all the springs), even in regions of high curvature where the angle between $\mathbf{R}_i - \mathbf{R}_{i-1}$ and $\mathbf{R}_{i+1} - \mathbf{R}_i$ deviates significantly from $180°$, the spring force should be evaluated as

$$\mathbf{F}_i^s \,|_\| \;=\; k\left(|\mathbf{R}_{i+1} - \mathbf{R}_i| - |\mathbf{R}_i - \mathbf{R}_{i-1}|\right)\hat{\tau}_i. \tag{5.6}$$

We now discuss the estimate of the tangent to the path. In the original formulation of the NEB method, the tangent at an image i was estimated from the two adjacent images along the path, $\mathbf{R}_{i+1}$ and $\mathbf{R}_{i-1}$. The simplest estimate is to use the normalized line segment between the two

$$\hat{\tau}_i = \frac{\mathbf{R}_{i+1} - \mathbf{R}_{i-1}}{|\mathbf{R}_{i+1} - \mathbf{R}_{i-1}|} \tag{5.7}$$

but a slightly better way is to bisect the two unit vectors

$$\tau_i \;=\; \frac{\mathbf{R}_i - \mathbf{R}_{i-1}}{|\mathbf{R}_i - \mathbf{R}_{i-1}|} + \frac{\mathbf{R}_{i+1} - \mathbf{R}_i}{|\mathbf{R}_{i+1} - \mathbf{R}_i|} \tag{5.8}$$

and then normalize $\hat{\tau} = \tau/|\tau|$. This latter way of defining the tangent ensures the images are equispaced even in regions of large curvature.

These estimates of the tangent have, however, turned out to be problematic in some cases [3]. When the energy of the system changes rapidly along the path, but the restoring force on the images perpendicular to the path is weak, as when covalent bonds are broken and formed, the paths can get 'kinky' and convergence to the MEP may never be reached. One way to aleviate the problem is to introduce a switching function that introduces a small

part of the perpendicular component of the spring force [13]. This, however, can introduce corner-cutting and lead to an overestimate of the saddle point energy. The kinkiness can be eliminated by using a better estimate of the tangent [3]. The tangent of the path at an image i is defined by the vector between the image and the neighboring image with higher energy. That is

$$\tau_i = \begin{cases} \tau_i^+ & \text{if } E_{i+1} > E_i > E_{i-1} \\ \tau_i^- & \text{if } E_{i+1} < E_i < E_{i-1} \end{cases} \tag{5.9}$$

where

$$\tau_i^+ = \mathbf{R}_{i+1} - \mathbf{R}_i, \quad \text{and} \quad \tau_i^- = \mathbf{R}_i - \mathbf{R}_{i-1}, \tag{5.10}$$

and $E_i = E(\mathbf{R}_i)$. If both of the adjacent images are either lower in energy, or both are higher in energy than image i, the tangent is taken to be a weighted average of the vectors to the two neighboring images. The weight is determined from the energy. The weighted average only plays a role at extrema along the MEP and it serves to smoothly switch between the two possible tangents τ_i^+ and τ_i^-. Otherwise, there is an abrupt change in the tangent as one image becomes higher in energy than another and this could lead to convergence problems. If image i is at a minimum $E_{i+1} > E_i < E_{i-1}$ or at a maximum $E_{i+1} < E_i > E_{i-1}$, the tangent estimate becomes

$$\tau_i = \begin{cases} \tau_i^+ \, \Delta E_i^{max} + \tau_i^- \, \Delta E_i^{min} & \text{if } E_{i+1} > E_{i-1} \\ \tau_i^+ \, \Delta E_i^{min} + \tau_i^- \, \Delta E_i^{max} & \text{if } E_{i+1} < E_{i-1} \end{cases} \tag{5.11}$$

where

$$\begin{aligned} \Delta E_i^{max} &= \max\left(|E_{i+1} - E_i|, \, |E_{i-1} - E_i|\right), \quad \text{and} \\ \Delta E_i^{min} &= \min\left(|E_{i+1} - E_i|, \, |E_{i-1} - E_i|\right). \end{aligned} \tag{5.12}$$

Finally, the tangent vector needs to be normalized. With this modified tangent, the elastic band is well behaved and converges rigorously to the MEP if sufficient number of images are included.

The implementation of the NEB method in a classical dynamics program is quite simple. First, the energy and gradient need to be evaluated for each image in the elastic band using some description of the energetics of the system (a first principles calculation or an empirical or semi-empirical force field). Then, for each image, the coordinates and energy of the two

adjacent images are required in order to estimate the local tangent to the path, project out the perpendicular component of the gradient and add the parallel component of the spring force. The computation of ∇V for the various images of the system can be done in parallel on a cluster of computers, for example with a separate node handling each one of the images. Each node then only needs to receive coordinates and energy of adjacent images to evaluate the spring force and to carry out the force projections. Various techniques can be used for the minimization. We have used projected velocity Verlet algorithm described above (see the section on Drag method).

To start the NEB calculation, an initial guess is required. We have found a simple linear interpolation between the initial and final point adequate in many cases. When multiple MEPs are present, the optimization leads to convergence to the MEP closest to the initial guess, as illustrated in figure 5.2. In order to find the optimal MEP in such a situation, some sampling of the various MEPs needs to be carried out, for example a simulated annealing procedure, or an algorithm which drives the system from one MEP to another, analogous to the search for a global minimum on a potential energy surface with many local minima [15].

It is important to eliminate overall translation and rotation of the system during the optimization of the path. A method for constraining the center of mass and the orientation of the system has been described, for example, by [49, 50]. Often, it is sufficient to fix six degrees of freedom in each image of the system, for example by fixing one of the atoms (zeroing all forces acting on one of the atoms in the system), constraining another atom to only move along a line (zeroing, for example, the x and y components of the force), and constraining a third atom to move only in a plane (zeroing, for example, the x component of the force).

In order to obtain an estimate of the saddle point and to sketch the MEP, it is important to interpolate between the images of the converged elastic band. In addition to the energy of the images, the force along the band provides important information and should be incorporated into the interpolation. By including the force, the presence of intermediate local minima can often be extracted from bands with as few as three images. The interpolation can be done with a cubic polynomial fit to each segment $[\mathbf{R}_i, \mathbf{R}_{i+1}]$ in which the four parameters of the cubic function can be chosen to enforce continuity in energy and force at

both ends. Writing the polynomial as $a_i x^3 + b_i x^2 + c_i x + d_i$, the parameters are [3]

$$
\begin{aligned}
a_x &= \frac{2E_{i+1} - E_i}{R^3} - \frac{F_i + F_{i+1}}{R^2} \\
b_x &= \frac{3E_{i+1} - E_i}{R^2} + \frac{2F_i + F_{i+1}}{R} \\
c_x &= -F_i \\
d_x &= E_i.
\end{aligned}
\tag{5.13}
$$

where E_i and E_{i+1} are the values of the energy at the endpoints, and F_i and F_{i+1} the value of the force along the path. This type of interpolation is usually quite smooth even though the second derivative is not forced to be continuous. A possible improvement is to generate a quintic polynomial interpolation so that the second derivatives can also be matched (and set to zero at the end points for a natural spline). This higher order polynomial can, however, add artificial wiggles in the path [3]

The NEB method has been applied successfully to a wide range of problems, for example studies of diffusion processes at metal surfaces[43], multiple atom exchange processes observed in sputter deposition simulations[44], dissociative adsorption of a molecule on a surface[37], diffusion of rigid water molecules on an ice Ih surface [45], contact formation between metal tip and a surface[46], cross-slip of screw dislocations in a metal (a simulation requiring over 100,000 atoms in the system, and a total of over 2,000,000 atoms in the MEP calculation)[47], and diffusion processes at and near semiconductor surfaces (using a plane wave based Density Functional Theory method to calculate the atomic forces) [38]. In the last two applications the calculation was carried out on a cluster of workstations with the force on each image calculated on a separate node.

The NEB method is an example of what has been called chain-of-states methods [84]. The common feature is that several images of the system are connected together to trace out a path of some sort. The simple object function for a chain (equation 5.3) is mathematically analogous to a Feynman path integral [77] for an off-diagonal element of a density matrix describing a quantum particle, which was used, for example, by Kuki and Wolynes to study electron tunneling in proteins [85]. Several chain-of-states methods have been formulated for finding transition paths that are optimal in one way or another [48, 49, 50, 86, 87, 88, 89, 51, 90]. The NEB method is the only one that converges to the MEP without having to

use second derivatives of the energy. Elber and Karplus [48] formulated an object function which is essentially similar to equation 5.3 although more complex. Czerminski and Elber presented an improved method with the Self-Penalty Walk algorithm (SPW) [49, 50] where a repulsion between images was added to the object function to prevent aggregation of images and crossings of the path with itself in regions near minima. Ulitsky and Elber [86], and Choi and Elber presented a quite different algorithm, the Locally Updated Planes (LUP) [87]. There, the optimization of the chain-of-states involves estimating a local tangent using equation 5.7 and then minimizing the energy of each image, i, within the hyperplane with normal q_i, i.e. relaxing the system according to

$$\frac{\partial \mathbf{R}_i}{\partial t} = -\nabla V(\mathbf{R}_i)[1 - \hat{\mathbf{q}}_i \hat{\mathbf{q}}_i]. \tag{5.14}$$

After every M steps (where M is on the order of 10) in the relaxation, the local tangents $\hat{\mathbf{q}}_i$ are updated. Since there is no interaction between the images (such as the spring force in the NEB), the LUP algorithm gives an uneven distribution of images along the path, and can even give a discontinuous path when two or more MEPs lie between the given initial and final states [87]. Also, the images do not converge rigorously to the MEP, but slide down slowly to the endpoint minima because of kinks that form spontaneously on the path and fluctuate as the minimization is carried out. Choi and Elber point out that it is important to start with a good initial guess to the MEP to minimize these problems. The NEB method is closely related to both the LUP method and the Elber-Karplus method. The NEB method incorporates the strong points of both of these approaches.

Smart [90] modified the Elber-Karplus-Czerminski formulation to get better convergence to the saddle point. The object function in his formulation involves a very high power (on the order of 100 to 1000) of the energy of the images to increase the weight of the highest energy image along the path.

Sevick, Bell and Theodorou [88] proposed a chain of states method for finding the MEP, but their optimization method, which includes explicit constraints for rigidly fixing the distance between images, requires evaluation of the matrix of second derivatives of the potential and is, therefore, not as applicable to large systems and complex interactions.

Chain-of-states methods have also been used for finding classical dynamical paths [89,

51]. Gillilan and Wilson [51] suggested using an object function similar to equation 5.3 for finding saddle points, but this suffers from the corner-cutting and down-sliding problems discussed above.

5.5   The CI-NEB Method

Recently, a modification of the NEB method has been developed, the Climbing Image - NEB [4]. There, one of the images, the one that turns out to have the highest energy after one, or possibly a few relaxation steps, is made to move uphill in energy along the elastic band. This is accomplished by zeroing the spring force on this one image completely and including only the inverted parallel component of the true force

$$\mathbf{F}_{imax}^{climb} \;=\; \nabla V(\mathbf{R}_{imax}) \cdot \hat{\tau}_{\parallel} \; \hat{\tau}_{\parallel} \tag{5.15}$$

The climbing image is dragged uphill, analogous to the drag method, but the essential difference is that the drag direction is determined by the location of the adjacent images in the band, not just $\mathbf{R}$ and $\mathbf{P}$ (unless the band only consists of one movable image). The tangent to the path is also weighted by the energy of the adjacent images as explained above. This turns out to be important in the surface island test problem.

Figure 5.2 shows the result of a CI-NEB calculations for the two dimensional test problem. Three movable images are included between the end points, and a straight line interpolation between $\mathbf{R}$ and $\mathbf{P}$ is used as a starting guess. The central image becomes the climbing image since it has the highest energy initially. Simultaneously, as the climbing image is pushed uphill, the other two images relax subject to the force projections of the nudging algorithm. After convergence is reached, a crude representation of the MEP has been obtained and one of the images is sitting at the saddle point to within the prescribed tolerance. An important aspect of the algorithm is that all movable images are adjusted simultaneously, and since only the position of adjacent images are needed for each step, the algorithm again parallelizes just as efficiently as the regular NEB.
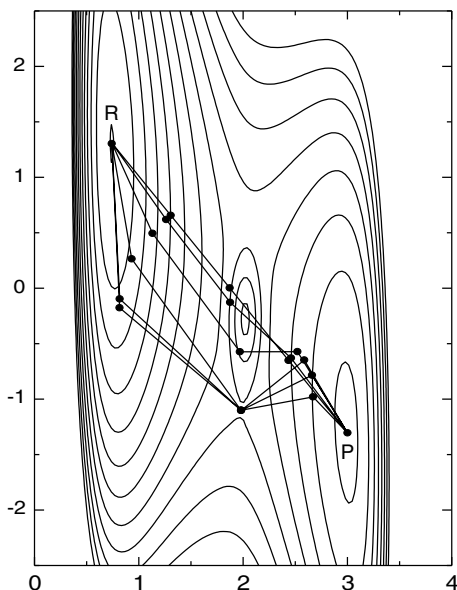
Figure 5.2: The Climbing Image Nudged Elastic Band method, CI-NEB. An elastic band is formed with three movable images of the system connected by springs and placed in between the fixed endpoints, **R** and **P**. The calculation is started by placing the three images along a straight line interpolation. The images are then relaxed keeping only the the component of the spring force parallel to the path and the component of the true force perpendicular to the path. The image with the highest energy is also forced to move uphill along the parallel component of the true force to the saddle point.

5.6 The CPR Method

For the conjugate peak refinement method [91], (CPR), a set of images is generated, one at a time, between the initial and final configurations, **R** and **P**. After the images are optimized, a line between the images constitutes a path that lies close to (but not at) the MEP. The maxima along the path will be at saddle points. Each point along the path is generated in a cycle of a line maximizations and a conjugate gradient minimizations. This is illustrated in figure 5.3. In the first cycle, the maximum along the vector $\mathbf{P} - \mathbf{R}$ is found, $\mathbf{y}_1$. Then, a minimization is carried out along the direction of each of the conjugate vectors (a total of D-1 dimensions) to give a new point $\mathbf{x}_1$.

In the second cycle the maximum along an estimated tangent to the $\mathbf{R} - \mathbf{x}_1 - \mathbf{P}$ path is found. The tangent is estimated using equation 5.8. This new maximum is denoted $\mathbf{y}_2$ in figure 5.3. The energy is then minimized along each of the conjugate vectors to give a new point that could potentially get incorporated into the path, etc. The rules for deciding whether a new point gets added to the path permanently are quite complicated and will not
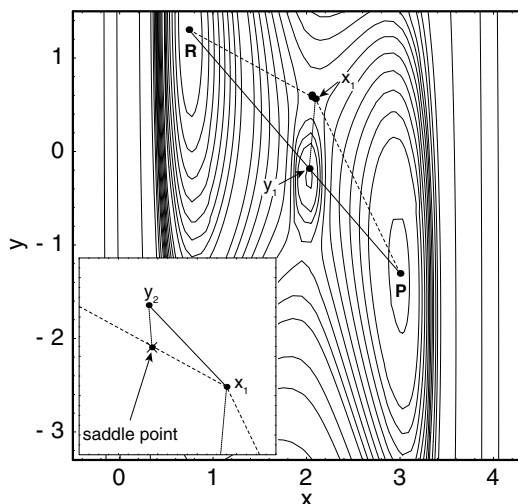
Figure 5.3: The conjugate peak refinement, CPR, method. Points along a path connecting **R** and **P** are generated, one point at a time through a cycle of maximization and then minimization. First, the maximum along the vector **P** − **R** is found, $\mathbf{y}_1$. Then, a minimization is carried out along a conjugate vector (small dashed line) to give location $\mathbf{x}_1$ on the path. In the second cycle (shown in inset) the maximum along an estimated tangent to the **R** − $\mathbf{x}_1$ − **P** path (solid line in inset) is found, $\mathbf{y}_2$, and then energy is minimized along a conjugate vector (small dashed line in inset) to give a fourth point along the path, etc.

be given here. The cycle of maximization along the tangent and then conjugate gradient line minimizations is repeated until a maximum along the path has a smaller gradient than the given tolerance for saddle points.

A detailed implementation of the CPR method, the TRAVEL algorithm, has been described by Fischer [92], providing values for all relevant parameters. We have used standard algorithms [29] for bracketing energy extrema and the line-optimizations.

We did not use the algorithm to generate a full path but stopped as soon as a point was found that satisfied our criterion for a saddle point (the magnitude of the gradient of the energy being less than a given tolerance).

5.7 The Ridge Method

The Ridge method of Ionova and Carter [93] involves advancing two images of the system, one on each side of the potential energy ridge, down towards the saddle point. The pair of images is moved in cycles of 'side steps' and 'downhill steps' in the following way. First, a straight line interpolation between products, **P**, and reactants, **R**, is formed and the maximum of energy along this line is found. The method is illustrated in figure 5.4, where
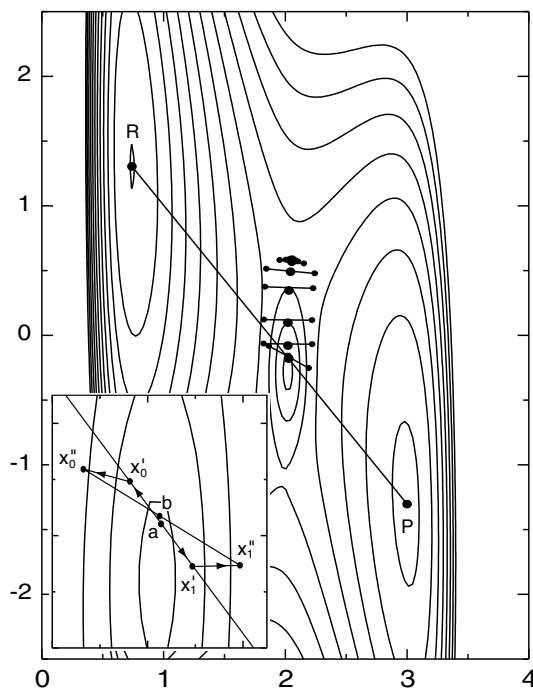
Figure 5.4: The Ridge method. A pair of images on each side of the potential energy ridge is moved towards the saddle point. First, the maximum along the vector $\mathbf{P} - \mathbf{R}$ is found, point $\mathbf{a}$ in the inset. Then the two images are formed on each side of the maximum, points $\mathbf{x}'_0$ and $\mathbf{x}'_1$, and they get displaced downhill along the gradient to points $\mathbf{x}''_0$ and $\mathbf{x}''_1$. This cycle of maximization between the two images, and then downhill move of the two images along the gradient is repeated, with smaller and smaller displacements until the saddle point is reached.

the maximum is found at point $\mathbf{a}$. We used the routine DBRENT from reference [29] to carry out the line maximizations, which makes use of the force, and typically takes a couple of force evaluations to converge to within 0.01 Å of the maximum).

Then, two replicas of the system are created on the line, one on each side of the maximum, $\mathbf{x}'_0$ and $\mathbf{x}'_1$ (see figure 5.4).

The magnitude of the displacement of the two images from the maximum needs to be chosen. This 'side-step' distance is typically chosen to be .1 Å in the first cycle. The force is now evaluated at the two images and they are moved in the direction of the force a certain distance, the 'downhill-step'. This generates points $\mathbf{x}''_0$ and $\mathbf{x}''_1$. The downhill distance is typically chosen to be 0.1 Å in the first cycle. This completes the first cycle. Then, a new cycle is started by maximizing along the line $[\mathbf{x}''_0, \mathbf{x}''_1]$ to obtain the point $\mathbf{b}$, etc.

The side-step and downhill-step of the images need to gradually decrease as the images get closer to the saddle point. It is possible that the energy of a point (in the sequence

$\mathbf{a}, \mathbf{b}, \mathbf{c}, \dots$) is higher than at the previous point. In such cases the downhill displacement is reduced by a half. Also, if the ratio of the side-step to downhill-step distance becomes larger than a certain, chosen ratio, the side step distance is also decreased by a half. This ratio is typically chosen to be some number in the range between 1 and 10. We found that the algorithm worked best for a ratio of 1.2 in the test cases we carried out. As the two images move and the size of the side-step to downhill-step is decreased, the sequence of points $\mathbf{a}, \mathbf{b}, \mathbf{c}, \dots$ should lead to a saddle point.

If the two images are almost equally displaced from the top of the energy ridge and the ridge is straight, it can be sufficient to evaluate the force only at the central point, rather than at the two images, thereby saving a factor of two in the number of force evaluations. This is implemented in such a way that if a new point in the sequence $\mathbf{a}, \mathbf{b}, \mathbf{c}, \dots$ is close to the center of the two images (not within 30% of either image), then the force in the next cycle is only evaluated at the central point and applied to both images in the downhill-step.

It turns out that most of the force evaluations are needed when the two images are rather close to the saddle point. Ionova and Carter [93] have discussed possible ways to improve the performance of the method in this final stage of the search.

5.8   The DHS Method

Dewar, Healy and Stewart [94] (DHS) have proposed a method which also involves two images of the system. First, the endpoints $\mathbf{R}$ and $\mathbf{P}$ are joined by a line segment. The two images are then systematically drawn toward each other until the distance between them is smaller than a given tolerance for finding the saddle point.

There are two steps in each cycle. First, the energy of both images is calculated. The one at lower energy is then pulled towards the one at higher energy along the line segment, typically about 5% of the way. Second, the energy of the lower energy image is minimized keeping the distance between the two fixed. An application of the method to the two-dimensional test problem is shown in figure 5.5. In the first cycle, the image at $\mathbf{P}$ is higher in energy than the one at $\mathbf{R}$, so the latter is brought in towards $\mathbf{P}$ by 5% and the allowed to relax with a fixed distance constraint. This repeats several times, causing the image that starts at $\mathbf{R}$ to climb up the potential energy valley leading up from $\mathbf{R}$. Eventually, the
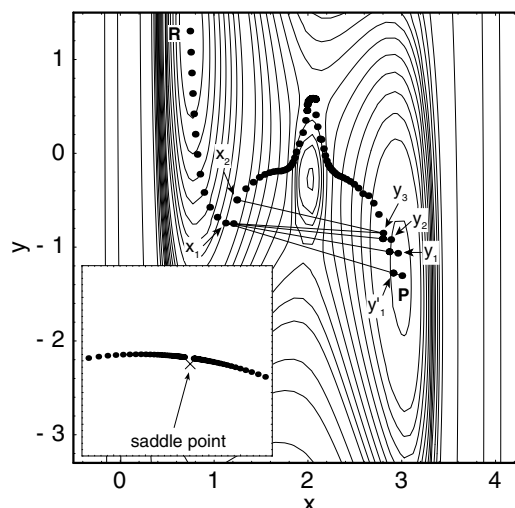
Figure 5.5: The method of Dewar, Healy and Stewart, DHS. Initially, a pair of images is created at **R** and **P**. In each cycle, the lower energy image is pulled towards the higher energy one and then let relax keeping the distance between the two fixed. Eventually, the two images straddle the energy ridge near the saddle point.

image at **P** becomes lower in energy. The five cycles following that are shown with solid lines in figure 5.5. Remarkable, the pair of images end up moving past the local maximum and converge on the saddle point on the other side.

The method can locate the neighboring region of the saddle point quite quickly, but does not converge close to the saddle point efficiently. If the images are pulled towards each other too quickly, the probability of both images ending on the same side of the ridge is increased. Eventually, as the pair of images gets close enough to the saddle point, such a slip over the ridge is bound to occur and both images will then settle into one of the minima **R** or **P**.

We chose to use a velocity Verlet type algorithm [81] for the minimization of the position of the lower energy image. At each step only the force perpendicular to the line segment connecting the two images was included. The velocity parallel to the force was included in the dynamics until the two pointed in the opposite direction, at which point the velocity was zeroed. This is the same kind of minimization algorithm we use with the Drag, NEB and CI-NEB methods.

5.9   The Dimer Method

When the final state of a transition is not known, the search for the saddle point is more challenging. A climb up from the initial state to the saddle point is more difficult than might at first appear. It is not sufficient to just follow the direction of slowest ascent. The two-dimensional test problem illustrated in figures 5.1 to 5.5 is an example of that. Several methods have been developed where information from second derivatives is built in to guide the climb [20, 95, 96, 97, 98, 25]. These methods have become widely used in studies of small molecules and clusters. Their disadvantage is that they require the second derivatives of the energy with respect to all the atomic coordinates, i.e. the full Hessian matrix, and then the matrix needs to be diagonalized to find the normal modes, an operation that scales as D^3. The evaluation of second derivatives is often very costly, for example in plane wave based Density Functional Theory calculations. Also, in large systems where empirical potentials are used, the D^3 scaling becomes a problem. For example, in a very interesting recent study of relaxation processes in Lennard-Jones glasses, a practical limit was reached at a couple of hundred atoms [99], while system size effects can be present in such systems even when up to 1000 atoms are included [100, 101].

A new method for finding saddle points was recently presented which has the essential qualities of the mode following methods, but only requires first derivatives of the energy and no diagonalization [2]. It can therefore be applied to plane wave DFT calculations and it can be applied to large systems with several hundred atoms, as illustrated below. The method involves two replicas of the system, a 'dimer', as illustrated in figure 5.6. The dimer is used to transform the force in such a way that optimization leads to convergence to a saddle point rather than a minimum. The force acting on the center of the dimer (obtained by interpolating the force on the two images) gets modified by inverting the component in the direction of the dimer. Before translating the dimer, the energy is minimized with respect to orientation. As pointed out by Voter [6], this gives the direction of the lowest frequency normal mode. This effective force will take the dimer to a saddle point when an optimization scheme is applied, for example conjugate gradients or the velocity Verlet algorithm with velocity damping. A detailed algorithm for finding the optimal orientation in an efficient way is described in ref. [2]. In a test problem involving Al adatom diffusion
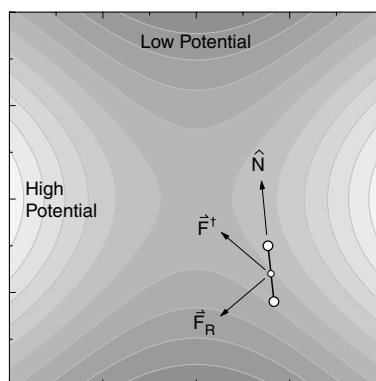
Figure 5.6: The calculation of the effective force in the Dimer method. A pair of images, spaced apart by a small distance, on the order of 0.1 Å, is rotated to minimize the energy. This gives the direction of the lowest frequency normal mode. The component of the force in the direction of the dimer is then inverted and the minimization of this effective force leads to convergence to a saddle point. No reference is made to the final state.

on the Al(100) surface, the Dimer method was found to converge preferably on the lowest saddle points (75% of the time the method converged on one of the lowest four saddle points) and the computational effort was found to increase only weakly as the number of degrees of freedom in the system was increased [2].

Figure 5.7 shows a Dimer calculation for the two-dimensional test problem. The initial configurations for the dimer searches were taken from the extrema of a short high temperature molecular dynamics trajectory (shown as a dashed line). The three initial points are different enough that the dimer searches converge to separate saddle points. In general the strategy for the Dimer method is to try many different initial configurations around a minimum, in order to find the saddle points that lead out of that minimum basin.

5.10 Configurational Change in an Island on FCC(111)

As a test problem for comparing the various methods described above, we have chosen a heptamer island on the (111) surface of an FCC crystal. Partly, this choice is made because it is relatively easy to visualize the saddle point configurations and partly because there is great interest in the atomic scale mechanism of island diffusion on surfaces (see for example [102]). The interaction potential is chosen to be a simple function to make it easy for others
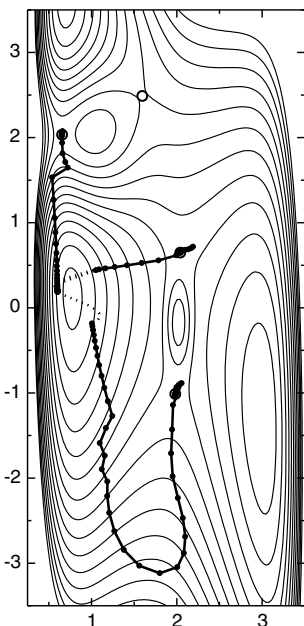
Figure 5.7: Application of the dimer method to a two-dimensional test problem. Three different starting points are generated in the reactant region by taking extrema along a high temperature dynamical trajectory. From each one of these, the dimer is first translated only in the direction of the lowest mode, but once the dimer is out of the convex region a full optimization of the effective force is carried out at each step (thus the kink in two of the paths). Each one of the three starting points leads to a different saddle point in this case.

to verify and extend our results. The atoms interact via a pairwise additive Morse potential

$$V(r) = A \left(e^{-2\alpha(r-r_0)} - 2e^{-\alpha(r-r_0)} \right) \tag{5.16}$$

with parameters chosen to reproduce diffusion barriers on Pt surfaces [103] (A=0.7102 eV, α=1.6047 Å^{-1}, r_0=2.8970 Å). The potential was cut and shifted at 9.5 Å. While exchange processes are not well reproduced with such a simple potential, the predicted activation energy for hop diffusion processes is quite similar to the predictions of more complex potential functions and in some cases in quite good agreement with experimental measurements [43, 103].

The surface is simulated with a 6 layer slab, each layer containing 56 atoms. The minimum energy lattice constant for the FCC solid is used, 2.74412 Å. The bottom three layers in the slab are held fixed. A total of $7 + 168 = 175$ atoms are allowed to move during the saddle point searches. This is 525 degrees of freedom. The displacements mainly involve
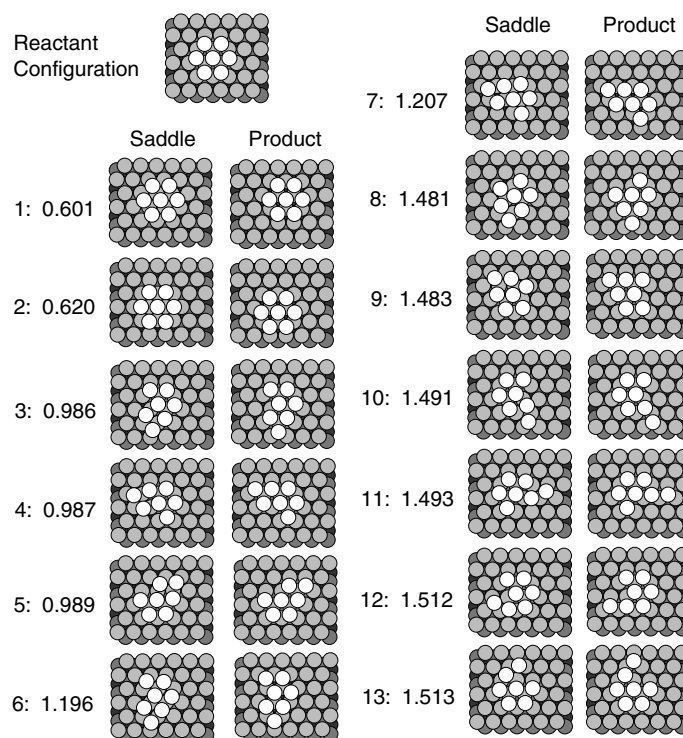
Figure 5.8: On-top view of the surface and the seven atom island used to test the various saddle point search methods. The shading indicates the height of the atoms. The initial state is shown on top. The saddle point configuration and the final state of the 13 transitions are also shown, with the energy of the saddle point (in eV) indicated to the left. The first two transitions correspond to a uniform translation of the intact island. Transitions 3-5 correspond to a pair of atoms sliding to adjacent FCC sites. In transitions 6 and 7 the pair of atoms slides to the adjacent HCP sites and the remaining 5 atoms slide in the opposite direction to HCP sites. In transitions 7 and 8, a row of three edge atoms slides into adjacent FCC sites. In transitions 10 and 11 a pair of edge atoms moves in such a way that one of the atoms is displaced away from the island while the other atom takes its place. In transitions 12 and 13 a single atom gets displaced.

some of the island atoms, but relaxation of the substrate atoms can also be important.

The initial configuration of the island is a compact heptamer as shown in figure 5.8. The question is how the island diffuses. We have focused on the initial stage of such a configurational transition, i.e. saddle points that are at the boundary of the potential basin corresponding to the compact heptamer state. A total of 13 processes were found with saddle point energy less than or equal to 1.513 eV. The lowest energy processes correspond to uniform translation of the island from FCC sites to HCP sites. There are two slightly different directions for the island to hop, and thus two slightly different saddle points, of energy 0.601 eV and 0.620 eV (see figure 5.8). The next three low energy saddle points, processes 3 to 5, correspond to a pair of edge atoms shifting to adjacent FCC sites. The

Table 5.1: Number of force elvaluations needed to reach saddle point to 0.01 eV/Å tolerance in the force.

saddle	Drag	CI-NEB(3)	CI-NEB(1)	Ridge	CPR	DHS	Dimer	ART
1	47	81	25	189	241	232	80	83
2	37	75	25	288	240	230	76	70
3	-	285	177	1369	1277	788	439	246
4	-	276	179	1129	1464	785	94	236
5	-	333	151	1165	1443	736	354	250
6	-	654	204	1369	2412	2434	449	-
7	-	735	206	1245	2426	2057	430	-
8	146	300	163	772	776	526	262	380
9	149	351	179	781	748	483	281	386
10	-	363	115	734	1551	736	510	-
11	-	282	126	869	2612	706	214	-
12	156	294	48	884	718	521	186	-
13	153	333	105	913	686	478	304	-
Average	115	336	131	901	1276	824	283	236
Std	56	184	64	368	810	662	149	125

three processes are quite similar, just three slightly inequivalent directions. Process 6 and 7 are quite interesting. Here, a pair of atoms is again shifted, but now only to the nearby HCP sites. The other 5 atoms in the cluster are also shifted to adjacent HCP sites but in the opposite direction. The final state has all island atoms sitting at HCP sites. Processes 8 and 9 involve a concerted move of three edge atoms. Process 10 and 11 involve an edge dimer where one of the atoms moves in a direction away from the island while the other one takes its place. This is a significantly higher energy final state, because of the low coordination of one of the displaced atoms. Finally, processes 12 and 13 involve the displacement of just one atom away from the island, again resulting in low coordination in the final state.

One common feature of processes 3 to 13 is that the final state is higher in energy than the initial state. The saddle point is typically late, i.e. close to the final state.

5.11   Results

The results of the calculations are given in tables 5.11 and 5.11. The number of force evaluations needed to reach a saddle point is given. We use this unit of computational effort because the evaluation of the force dominates the effort at each step, even with

Table 5.2: Number of force elvaluations needed to reach saddle point to 0.001 eV/Å tolerance in the force.

saddle	Drag	CI-NEB(3)	CI-NEB(1)	Ridge	CPR	DHS	Dimer	ART
1	324	372	122	3441	653	795	328	332
2	70	192	45	288	433	290	244	146
3	-	597	327	2382	1610	1295	746	336
4	-	585	246	2047	1729	1296	546	366
5	-	675	314	2112	1695	1258	570	377
6	-	999	274	2187	2821	4310	704	-
7	-	978	271	2144	2720	4076	588	-
8	323	573	309	4090	1197	1320	559	742
9	338	855	446	1995	1268	1342	553	754
10	-	648	174	1610	1739	1468	816	-
11	-	447	237	1859	2793	1474	308	-
12	299	687	150	1861	1038	1160	386	-
13	293	738	230	1901	969	1097	562	-
Average	275	642	242	2147	1590	1629	532	436
Std	102	228	103	890	788	1182	173	227

empirical potentials. We are particularly interested in plane wave based DFT calculations where the evaluation of just the energy and not the force presents insignificant savings. The computational effort is, therefore, simply characterized by the number of force evaluations. Table 1 gives the results obtained with convergence tolerance of 0.01 eV/Å in the magnitude of the force, i.e. the saddle point searches were stopped when the magnitude of the force on each degree of freedom had dropped below this value. This tolerance is small enough to get the saddle point energy to within 0.01 eV. To illustrate how fast the various methods home in on the saddle points, the number of force evaluations needed to satisfy a tighter tolerance, 0.001 eV/Å, is given in table 2 for comparison. In most cases, the saddle point energy obtained is different by less than 0.001 eV as the tolerance is reduced, but in some cases the difference is on the order of 0.01 eV.

The results show that the drag method fails for 7 out of the 13 processes. This is because the MEP has large curvature and the direction of the unstable normal mode at the saddle point is quite different from the direction of the vector **P-R**. The drag method should, therefore, not be used. When the drag method works, however, it is very efficient.

The CI-NEB method with three movable images, CI-NEB(3), is highly reliable, gets all the saddle points, and is less than three times more expensive than the drag method.

Since it is easy to paralellize the CI-NEB with one image per node, the number of force evaluations per node, and therefore the elapsed time until the calculation finishes on a three node cluster, would actually be just about the same or even less for CI-NEB(3) than for Drag.

It is interesting to push the elastic band method to the extreme and reduce the number of images to one. This is essentially the same as the Drag method except the direction of the drag is different. If the tangent in the CI-NEB were estimated using equation 5.7, then the two method would be identical. The fact that CI-NEB uses an estimate of the tangent, equations 5.9 and 5.11, where the weight of the adjacent points is a function of the energy, makes the CI-NEB(1) converge in these cases while the Drag method diverges. The saddle point is closer to the higher energy final state, and the tangent of the path is biased more towards the line segment to the final state than to the initial state. It is interesting that CI-NEB(1) is so successful in these test problems, but it cannot be expected to work in all cases.

The Ridge method is significantly more expensive than CI-NEB(3), a factor of 2.7 for the larger tolerance and a factor of 3.3 for the smaller tolerance. The method has relatively hard time converging rigorously on the saddle point, i.e. it uses a large number of force evaluations towards the end of the search. There are several parameters in the Ridge method that need to be chosen and the performance depends quite strongly on the choice of these parameters. We optimized for one of the saddle point searches and then used the same parameter set for all of them (the parameters are given in the discussion of the method above).

The CPR method is the most difficult method to implement, because of the complex rules for adding or rejecting points on the path. It is also the least efficient of the methods tested. It does, however, converge fast to the saddle point once it is close, as is evident from comparing table 1 and 2. This is probably because of the use of the conjugate gradient minimization which is quite efficient.

The DHS method of Dewar and coworkers is easy to implement and it does quite well, is the second best method, at the larger tolerance. But, as the Ridge method, it has hard time converging on the saddle point. A significant improvement in the timing might occur if a switch to a different method, for example the CI-NEB(1), is made once the two images
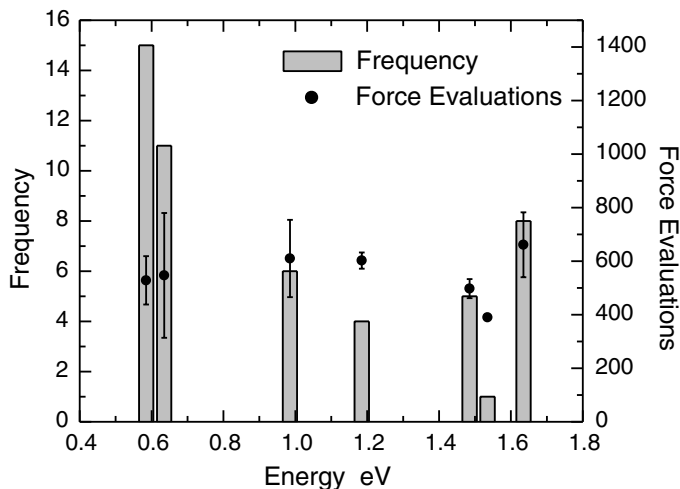
Figure 5.9: The frequency at which the various various saddle points for the surface island transitions (illustrated in figure 8) are found with the Dimer method. The lowest saddle points are found with the highest frequency. Also shown are the number of iterations required to go from the intial state up to the saddle point to within a force tolerance of 0.001 eV/Å. For the more practical 0.01 eV/Å tolerance, the average number of force evaluations was a little under 300. The error bars show the standard deviation.

are in the region of the saddle (for example, when the force has dropped to 0.1 eV/Å).

The Dimer method can be started from any point on the potential energy surface. While the method is designed to work without any knowledge of the final state, it is possible to make use of the final state in cases where it is known. Tables 1 and 2 are timings for the Dimer method where a line maximization along the $\mathbf{P} - \mathbf{R}$ line is first carried out, and then the Dimer search is started from the maximum. The dimer method is highly efficient, each saddle point search involves fewer force evaluations than CI-NEB(3). The advantage of CI-NEB(3) is that it gives some picture of the whole MEP in addition to the saddle point, as discussed below. The unique quality of the Dimer method is its ability to climb up the potential surface starting from the minimum. Results of 50 such runs are shown in figure 5.9. Here, the starting points were generated by random displacements of the atoms about the initial state minimum with maximum amplitude of 0.1 Å. The tighter tolerance, 0.001 eV/Å was used in these runs. It is surprising that the average number of force evaluations is not that much larger than when the search was started from the maximum along $\mathbf{P} - \mathbf{R}$ (590 force evaluations vs. 528). Of course, if one is only interested in a particular final state, the dimer method started from the minimum may converge on the 'wrong' saddle point and then needs to be repeated a few times.

For comparison, we have included in tables 1 and 2 the timings for a simpler algorithm, ART [15], a method which is mainly used to help equilibrate systems by finding final states rather than saddle points (and has proven to be highly successful in simulations of amorphous materials [14], for example). The method is analogous to the drag method except no reference is made to the final state, the drag coordinate is taken to be the direction from the initial state to the current location. The force was inverted along the drag coordinate and velocity Verlet algorithm with velocity projections used to home in on the saddle point. The method is very efficient and takes somewhat fewer iterations than the dimer method, but similar to the drag method, it does not find about half the saddle points.

5.12   Discussion

It is important to point out that all the timings given above are for a search of a single saddle point. In order to verify that the saddle point found is indeed the highest saddle point on the MEP for the process of interest, a calculation of the MEP needs to be carried out. Given the saddle point, it is rather straightforward to slide down along the MEP. One stable method is to displace the system downward and then minimize the energy with a fixed distance to the previous point higher up along the path. The CI-NEB(3) method provides three points along the MEP and with the interpolation where forces are included this is typically enough to see whether the path has more than one saddle point. The CI-NEB(3) timings in table 1 and 2 are, therefore, the total number of force evaluations needed to get both the saddle point energy and to get a reasonable idea of what the MEP looks like. If it is evident that additional saddle points are present, additional images can be introduced starting from the best estimate from the interpolation. The Ridge, CPR and DHS methods would all need to be followed by a calculation of the MEP starting from the saddle point. This would typically add a couple of hundred force evaluations to the numbers given for the Drag, Ridge, CPR and DHS methods in table 1 and 2.

5.13   Summary

An overview has been given of several methods used to find saddle points on energy surfaces when only the energy and first derivatives with respect to atomic positions are available.

Finding saddle points is the most challenging task when estimating rates of transitions within harmonic Transition State Theory. The high dimensionality of condensed matter systems makes this non-trivial. Several commonly used methods have been applied to a test problem involving configurational changes in an island on a crystal surface where the final state of the transition is known. The CI-NEB method turned out to be the most efficient method. In addition to the saddle point, it gives an idea of what the shape of the whole MEP is. This is necessary to determine whether more than one saddle points are present, and then which one is highest. When the final state is not known, the Dimer method can be used to climb up the potential energy surface starting from the initial state. The average number of force evaluations for a Dimer to converge on a saddle point is similar to a CI-NEB calculation with three movable images in the test problem studied here.

It is our hope that the test problem presented here will continue to be a useful standard for comparing methods for finding saddle points. Clearly, other test problems with different qualities should also be added. To make it easier for others to use this test problem, we have made configurations and other supplementary information available on the web at *http://www-theory.chem.washington.edu/~hannes/paperProgrInThChem*.

5.14 Appendix: The Two Dimensional Test Problem

This model includes a LEPS [52] potential contribution which mimics a reaction involving three atoms confined to motion along a line. Only one bond can be formed, either between atoms A and B or between atoms B and C. The potential function has the form

$$
\begin{aligned}
V^{LEPS}(r_{AB}, r_{BC}) = {} & \frac{Q_{AB}}{1+a} + \frac{Q_{BC}}{1+b} + \frac{Q_{AC}}{1+c} - \left[\frac{J_{AB}^2}{(1+a)^2} + \frac{J_{BC}^2}{(1+b)^2} + \frac{J_{AC}^2}{(1+c)^2} \right. \\
& \left. - \frac{J_{AB}J_{BC}}{(1+a)(1+b)} - \frac{J_{BC}J_{AC}}{(1+b)(1+c)} - \frac{J_{AB}J_{AC}}{(1+a)(1+c)} \right]^{\frac{1}{2}}
\end{aligned}
\tag{5.17}
$$

where the Q functions represent Coulomb interactions between the electron clouds and the nuclei and the J functions represent the quantum mechanical exchange interactions. The form of these functions is

$$
Q(r) = \frac{d}{2} \left(\frac{3}{2} e^{-2\alpha(r-r_0)} - e^{-\alpha(r-r_0)} \right)
\tag{5.18}
$$

98

and

$$J(r) = \frac{d}{4} \left(e^{-2\alpha(r-r_0)} - 6e^{-\alpha(r-r_0)} \right) . \tag{5.19}$$

The parameters were chosen to be $a = 0.05$, $b = 0.80$, $c = 0.05$, $d_{AB} = 4.746$, $d_{BC} = 4.746$, $d_{AC} = 3.445$, and for all three pairs we use $r_0 = 0.742$ and $\alpha = 1.942$.

In order to reduce the number of variables, the location of the end point atoms A and C is fixed and only atom B is allowed to move. A 'condensed phase environment' is represented by adding a harmonic oscillator degree of freedom coupled to atom B. This can be interpreted as a fourth atom which is coupled in a harmonic way to atom B

$$V(r_{AB}, x) = V^{LEPS}(r_{AB}, r_{AC} - r_{AB}) + 2k_c(r_{AB} - (r_{AC}/2 - x/c))^2 \tag{5.20}$$

where $r_{AC} = 3.742$, $k_c = 0.2025$, and $c = 1.154$. This type of model has frequently been used as a simple representation of an activated process coupled to a medium, such as a chemical reaction in a liquid or in a solid matrix.

In order to create two saddle points rather than just one, a Gaussian function is added to $V(r_{AB}, x)$ to give

$$V^{tot}(r_{AB}, x) = V(r_{AB}, x) + 1.5 \, G(r_{AB} - 2.02083, x + 0.272881) \tag{5.21}$$

where the Gaussian function is $G(a, b) = \exp(-0.5((a/0.1)^2 + (b/0.35)^2))$. A contour plot of this 2D potential surface is given in figures 5.1 to 5.5.

5.15 Acknowledgments

Chapter 6

THEORETICAL CALCULATIONS OF DISSOCIATIVE ADSORPTION OF CH$_4$ ON AN IR(111) SURFACE

6.1 *Abstract*

The activation energy for chemisorption of CH$_4$ at an Ir(111) surface is determined using density functional theory combined with an estimate of the long range dispersion interaction. The results are found to be in good agreement with published results of bulb and beam experiments analyzed with a precursor model. A surprisingly large surface relaxation is found where an Ir surface atom is displaced outwards by as much as 0.6 Å. A strongly bound *molecular state* at kinks and adatoms involving η^2-H,H bonding was also found.

6.2 *Introduction*

Alkanes are potentially an important source of raw material for the petrochemical industry. However, the severe reaction conditions currently required to activate the strong bonds in hydrocarbons make selective catalysis difficult. The dissociative adsorption of alkanes on transition metal surfaces is a key step in the catalytic process. Improved insight into the fundamental mechanism is important for development of selective conversion of small hydrocarbons to chemicals of greater value. Extensive experimental measurements have been carried out in the last decade on various transition metal surfaces [104, 105, 106, 107, 108, 8, 9]. Much of the work has focused on Ir(110) and Ir(111) surfaces [105, 107, 8]. Theoretical studies have also been done, both on the energetics [109, 110, 111] and on the dynamics [112, 113, 114], but it is still unclear what the basic mechanism of alkane dissociative adsorption is, especially under temperature and pressure conditions relevant to industrial catalysis. The interpretation of experimental measurements is in some cases also controversial.

Seets *et al.* have carried out extensive experimental studies of the initial, zero-coverage dissociative sticking of CH_4 on Ir(111), using both bulb experiments and molecular beam experiments [8]. The surface temperature was kept high, between 900 K and 1100 K. Since the dissociative adsorption probability is low, less than one in a thousand collisions of a CH_4 molecule with the surface leads to dissociative adsorption. The gas near the surface is, therefore, expected to heat up to the surface temperature before most of the dissociative sticking events. We will assume the methane gas is thermally equilibrated with the surface in our calculations. Under such conditions, transition state theory (TST) can be applied. For H_2 sticking on Cu(110) it has been shown that TST gives an excellent approximation (within a factor of two) of the exact, classical dynamical result [37].

In the high energy molecular beam experiments, Seets *et al.* observed the typical increase of the dissociative sticking probability with beam energy, indicating a direct mechanism where the initial kinetic energy of the molecule helps overcome the activation energy barrier to dissociation. The sticking probability increased gradually by four orders of magnitude as the energy was increased from 0.1 eV to 1.2 eV. It is unclear how to interpret such data in terms of the energy surface describing the interaction of the CH_4 with the surface. At high beam energy, the trajectories of the dissociating CH_4 molecules are probably not close to the optimal path characterizing dissociative adsorption in the low energy range relevant for thermal activation. A novel feature of the beam experiments of Seets *et al.* was their ability to lower the beam energy to very small values. At low energy, a drop in the sticking coefficient was observed with beam energy. Such a behavior can be ascribed to a precursor mechanism. Seets *et al.* analyzed their data in terms of a precursor model and extracted the ratio of the rate of desorption to the rate of chemisorption from the precursor state. More recently, an analogous drop in sticking probability with beam energy was observed in CH_4 dissociation on Pt(110), but there it was attributed to steering rather than precursor effects [9].

We calculated the minimum energy path (MEP) [37] for CH_4 dissociation on Ir(111) using density functional theory (DFT). The path connects the CH_4 molecule above the surface to the dissociated H and CH_3 fragments adsorbed on the surface. Within the harmonic approximation to TST, the activation energy for sticking is given by the highest maximum on the MEP, a saddle point on the energy surface. The calculated activation

energy can be compared directly with the activation energy deduced from the experiments. Our calculations do not include tunneling as a possible adsorption mechanism. For H_2 dissociation on Cu(110), it has been estimated that tunneling becomes important below 400 K [37, 76], a much lower temperature than used by Seets *et al.*

6.3 Calculations

The DFT calculations were carried out using the PW91 functional [62]. Ultrasoft pseudopotentials were used [63] with a 350 eV energy cutoff in the plane wave representation of the wave function. The calculations were done with the VASP code [58, 59, 60, 61]. The Ir metal was represented by a slab including up to 6 layers with 12 atoms per layer. It was found to be sufficient to use only the gamma point in the k-point sampling for this large simulation cell. The minimum energy path was calculated using the nudged elastic band (NEB) method [13, 3] with 8 images of the system forming a discretization of the path between the fixed end points. The exact location of the saddle point and, thereby, the value of the activation energy barrier was found by first interpolating the energy and force between the images, and then refining the estimate with the dimer method [3, 2].

Several calculations of dissociative adsorption of molecules on surfaces have been carried out in the last few years, using both wave function based methods and DFT. These calculations have provided a great deal of insight and helped answer long standing questions about the interpretation of experimental measurements. The surface has either been represented with a small metal cluster [109, 111] or by using slabs [110, 115, 116, 117]. A particularly great effort has been made to model both energetics and quantum dynamics of H_2 dissociative adsorption on Cu surfaces [118]. In these previous studies, only small surface relaxations have been observed. In several of the calculations the surface has, in fact, been kept rigid to reduce computational effort. Given that the dissociative sticking probability of CH_4 on Ir(111) has been observed to be strongly dependent on surface temperature, we have investigated the importance of substrate relaxations. This requires a large simulation cell and makes the calculations more tedious.

The optimal bonding of the dissociated fragments, H and CH_3 on the Ir(111) surface was found to be on top of Ir atoms, 2.88 eV and 2.27 eV, respectively. The H atoms can
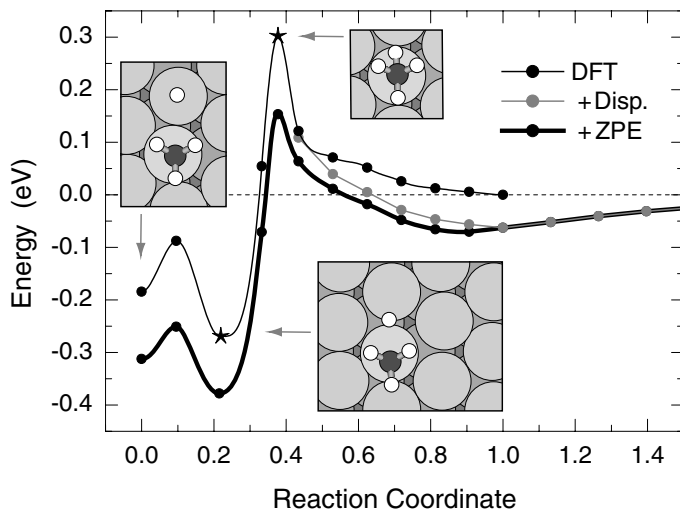
Figure 6.1: The minimum energy path for dissociation. Results of DFT/PW91 calculations (thin solid line), with an estimate of long range dispersion interaction added (gray line), and with zero point energy correction also added (thick solid line). The CH_4 molecule 4.3 Å above the surface plane corresponds to reaction coordinate of 1.0, while the adsorbed H and CH_3 fragments correspond to 0.0. The maximum ($\star$) was found by the dimer method.

also bind at bridge sites, but the binding there is 0.05 eV weaker.

We calculated the MEP from an initial state where the CH_4 molecule is 4.3 Å above the surface and a final state where the dissociated fragments are sitting on adjacent on-top sites. The results are shown in Fig. 6.1. An exponential rise in the DFT energy is observed as the molecule approaches the surface. The barrier for dissociation is found to be 0.3 eV with respect to the molecular state far from the surface. The saddle point configuration is shown in Fig. 6.2. The long range dispersion energy is not included in the DFT/PW91 calculations because of the semi-local description of correlation. A physisorption well is, therefore, not obtained from the DFT/PW91 results.

The height of the barrier turns out to be highly sensitive to the size of the simulation cell. The dependence on the number of layers in the slab is shown in Fig. 6.3. When only three layers are included in the slab, the calculated activation energy is three times higher. This is in part due to a large relaxation of the surface. The underlying Ir atom moves out normal to the surface and has been displaced by 0.41 Å at the saddle point. Further along the MEP, when the molecule is slightly closer to the surface, the Ir atom is displaced outwards by as much as 0.6 Å. Such a large surface relaxation has not been reported in previous calculations of dissociative adsorption. The energy required to distort just the
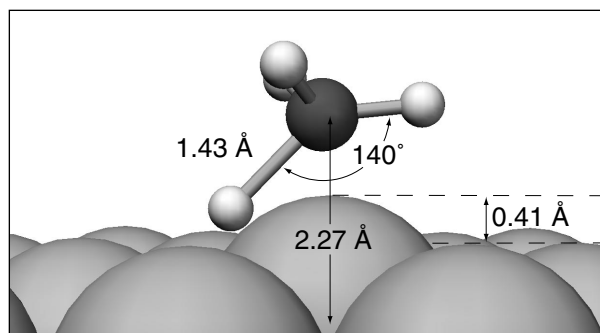
Figure 6.2: The configuration of atoms at the saddle point. Note the large displacement of the underlying Ir atom normal to the surface, 0.41 Å.

surface to the saddle point configuration is 0.30 eV. This is consistent with the large effect surface temperature has on the measured sticking probability, even in the higher energy regime where the dissociation is direct [8]. The reason for the large outward displacement is likely an upward shift of the d-electron band which strengthens the bonding with the dissociating molecule [116].

After the CH_4 molecule comes apart, a deep minimum is observed in the path. This was already evident from the NEB calculation, but further relaxation of the configuration obtained by interpolation between the images in the NEB gave a converged minimum at nearly -0.3 eV (see Fig. 6.1). This configuration is more stable than binding of the dissociated fragments on adjacent on-top sites.

Long range, zero overlap dispersion energy needs to be added to the DFT calculations. The leading term in the asymptotic expansion of the dispersion interaction is C_3/z^3 where z is the distance from the surface [119]. The C_3 coefficient can be estimated from the polarizability of the CH_4 molecule and the dielectric function of the metal using the Lifshitz formula [120]. In the simplest approximation C_3 is proportional to the static polarizability of the molecule. We have estimated the value of the C_3 coefficient for $CH_4/Ir(111)$ by scaling the value for $He/Au(111)$ [119] (taking Au to be an analog for Ir) with the ratio of the polarizability of CH_4 and He (a factor of 12.6). This gives $C_3 = 3.45$ eV/Å^3.

Close to the surface, where the electron wave function of the CH_4 and Ir start to overlap, the asymptotic expansion gives an overestimate of the dispersion energy. Also, due to the large displacement of one of the Ir surface atoms during the dissociation, the atomic structure of the surface needs to be taken into account when calculating the dispersion
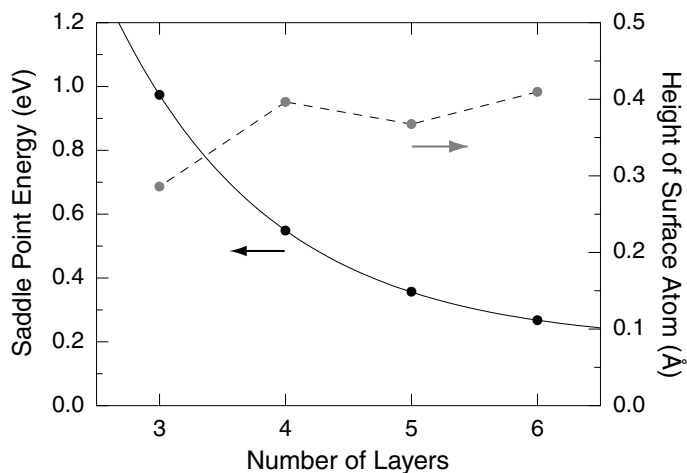
Figure 6.3: An exponential drop in the activation energy calculated with DFT occurs when the number of layers in the Ir slab is increased from 3 to 6 (solid line). The extent of the displacement of the underlying Ir atom normal to the surface also increases as more layers are added, although not monotonically (dashed line).

interaction. While the dielectric response of the Ir surface should contain contributions from both localized d-electrons and delocalized s-electrons [121], we have approximated the response entirely in terms of a localized oscillator at each Ir atom and cast the dispersion interaction in terms of a superposition of c_6/r^6 contributions (where r is C-Ir atom distance), as has been done previously for He/metal and semi-metal interactions [122]. By matching the long range behavior, the estimate of the C_3 coefficient gives a c_6 coefficient of 7.2 eV/Å^3 for the induced dipole-induced dipole interaction between the CH_4 and each one of the local oscillators. Higher multipole interactions, such as the induced quadrupole-induced dipole, etc., terms are effectively taken into account by simply increasing the c_6 coefficient by 20%. The dispersion interaction is turned off as the C-atom approaches the outermost Ir-atom by using a damping function which has proved to be successful in calculations of atom-atom and atom-molecule interactions[1,2] [123]. Figure 6.1 shows the estimated dispersion energy.

The vibrational zero point energy (ZPE) of CH_4 is large because of the small mass of the H-atoms and the high vibrational frequency. At the saddle point, one of the modes has

---

[1]Damped dispersion accounts for correlation and should be added to Hartree-Fock results. Here, the DFT calculations already include semi-local dispersion, so some double counting of correlation will occur in the region between the well minimum (where DFT/PW91 gives negligible correlation) and the saddle point (where the damped dispersion has become negligible.

[2]The parameter r_m in the damping function was taken to be 6 Å, analogous to Ref. [122]

negative curvature so the corresponding contribution to the ZPE is lost. This effectively lowers the activation energy for chemisorption. We have calculated the Hessian matrix and diagonalized to find the normal modes at each NEB image along the MEP. The change in the total ZPE as compared with the gas phase molecule is then added as a correction to the DFT calculation (see Fig. 6.1). The ZPE correction is largest at the saddle point (lowered by 1.4 eV) and beyond but is also significant when the molecule is quite far from the surface. The physisorption well has a depth of 0.07 eV with the minimum energy at $z = 4.1$ Å. The barrier to go from the physisorbed state to the chemisorbed state is estimated to be 0.23 eV. Typically DFT calculations of barriers are quoted to have an error bar of 0.1 eV. Our results are, therefore, in excellent agreement with the experimental estimate of 0.28 ± 0.04 eV. (A recent theoretical estimate based on cluster calculations gave a much higher barrier, 0.79 eV [111]). The good agreement with the results of Seets *et al.* lends support for their interpretation of the molecular beam data in terms of a precursor model.

6.4 Strongly Bound Molecular States

We have also studied the interaction of CH_4 with defects on the Ir(111) surface: a kink and an Ir adatom. The DFT results alone give an energy well corresponding to a strongly bound *molecular* state at the least coordinated Ir atom. The binding energy at the kink on the *B*-type step edge and at the adatom is 0.22 eV while the binding at the kink on the *A*-type edge is 0.18 eV. This binding is three times stronger than the physisorption well at the flat terrace. This suggests that chemical bonding is taking place between a CH_4 molecule and a highly under coordinated Ir atom. At the kink, the distance between the C-atom and the Ir-atom is only 2.5 Å, much closer to the chemisorbed methyl (2.1 Å) than the physisorbed methane (4.1 Å). The Ir-C distance is so short, in fact, that the long range dispersion interaction correction has become negligible. One of the H-C-H bond angles has opened up to 116°. Two of the four H-atoms are pointed towards the under coordinated Ir atom (at the *B*-kink the Ir-H distances are 2.0 Å and 2.3 Å).

Such chemical bonding of methane and other alkanes is well known from studies of transition metal complexes and is referred to as η^2-H,H bonding [124]. A similar strong molecular state has been seen experimentally and by DFT calculations of H_2 on a stepped

Cu surface[125, 126]. This strongly bound molecular state could play an important role in precursor mediated mechanisms at low temperature where the life time is long enough for the CH_4 molecules to sample the surface and find defects (the high surface temperature in the experiments of Seets *et al.* [8] would prevent this). We find, however, that the activation energy for dissociation at the adatom is not lower than at the terrace site.

The strong molecular binding at kinks is relevant to the possibility of using high index surfaces with periodic arrays of kinks for enantiomeric separation [127]. Our results indicate that larger, chiral alkanes and their derivatives could interact more strongly with kinks and with more specificity than if the attractive interaction were purely due to long range dispersion interaction.

6.5   Acknowledgments

6.6   Erratum

This chapter is the same as what appears in ref. [7] except for the sentences describing the experimental determination of the physisorption well depth by the Mullins group. In ref. [7] it is said that the well depth can be determined by comparing the bulb and beam experiments when in fact both techniques were used to determine the chemisorption barrier. The Mullins group inferred the presence of the physisorption well because the trapping mediated model could be used to accurately describe their data, but they did not measure it directly. A more complete discussion of these results is in progress [128].

Chapter 7

LONG TIME SCALE KINETIC MONTE CARLO SIMULATIONS WITHOUT LATTICE APPROXIMATION AND PREDEFINED EVENT TABLE

7.1 Abstract

We present a method for carrying out long time scale dynamics simulations within the harmonic transition state theory approximation. For each state of the system, characterized by a local minimum on the potential energy surface, multiple searches for saddle points are carried out using random initial directions. The dimer method is used for the saddle point searches and the rate for each transition mechanism is estimated using harmonic transition state theory. Transitions are selected and the clock advanced according to the kinetic Monte Carlo algorithm. Unlike traditional applications of kinetic Monte Carlo, the atoms are not assumed to sit on lattice sites and a list of all possible transitions need not be specified beforehand. Rather, the relevant transitions are found on the fly during the simulation. A multiple time scale simulation of Al(100) crystal growth is presented where the deposition event, occurring on the time scale of picoseconds, is simulated by ordinary classical dynamics, but the time interval in between deposition events, on the order of milliseconds, is simulated by the long time scale algorithm. The Al(100) surface is found to grow remarkably smooth, even at 30 K because of concerted displacements of multiple atoms with significantly lower activation energy than adatom diffusion on the flat terrace.

7.2 Introduction

A common problem in theoretical chemistry, condensed matter physics and materials science is the calculation of the time evolution of an atomic scale system where, for example, chemical reactions and/or diffusion occur. In either case, the configuration of atoms is changed in some way during as time evolves. The interaction between the atoms can be ob-

tained from an (approximate) solution of the Schrödinger equation describing the electrons, or from a potential energy function determined in some empirical way. Most often, it is sufficient to treat the motion of the atoms using classical mechanics. Quantum mechanical effects in the motion of atoms are important only in exceptional cases. Since the classical equations of motion can easily be solved numerically the simulation of dynamical evolution is, therefore, in principle quite simple. The problem is that the transitions of interest are typically many orders of magnitude slower than vibrations of the atoms, so a direct simulation of the classical dynamics ends up being of little use. This 'rare event' problem is best illustrated by an example. A typical, low activation energy for a chemical reaction or diffusion event is 0.5 eV. Such an event can occur thousands of times per second at room temperature and would typically be important in the time evolution of the system. But, the atoms vibrate on the order of 10^{10} times before a fluctuation of thermal energy occurs in the right degree of freedom for a transition to take place. A direct classical dynamics simulation which necessarily has to faithfully track all this vibrational motion would take thousands of years of computer calculations on the fastest present day computer before a single transition can be expected to occur! It is clear that meaningful studies of chemical reactions and/or diffusion cannot be carried out by simply simulating the classical dynamics of the atoms. It is essential to simulate the system on a much longer time scale. This time scale problem is one of the important challenges in computational research on atomic scale systems.

Fortunately, the separation of time scales between vibrations and transitions makes it possible to use a statistical approach to estimate transition rates. If a bottleneck region through which the system must pass in order to make the transition can be identified, the so-called transition state, then transition state theory (TST) [64, 65, 66, 1, 129, 67, 68] can be used to calculate the average amount of time the system will spend in a given state. Short time dynamical trajectories can subsequently be used both to correct for the approximations made in TST and to identify the possible final states of transitions. In general, the free energy of the transition state needs to be evaluated to estimate the rate. But, for solid state systems, where the atoms are relatively tightly held in place by their neighbors, it is often possible to assume that the transition state can be characterized by a few saddle points on the potential energy rim surrounding the initial state basin and that the partition function

of the system near each saddle point and near the energy minimum can be approximated by a product of harmonic partition functions. In this case, TST simplifies to harmonic transition state theory (hTST) and the rate of escape, k, through each of the saddle point regions can be related to properties of the initial state energy minimum and the saddle point [11, 12] as

$$k^{\text{hTST}} = \frac{\Pi_i^{3N} \nu_i^{\text{init}}}{\Pi_i^{3N-1} \nu_i^{\ddagger}} \, e^{-(E^{\ddagger} - E^{\text{init}})/k_B T}. \tag{7.1}$$

Here, $E^{\ddagger}$ is the energy of the saddle point, E^{init} is the energy of the local minimum corresponding to the initial state, and the ν_i are the corresponding normal mode frequencies. The symbol $\ddagger$ refers to the saddle point. All the quantities can be evaluated directly from the potential energy surface without dynamical calculations. Entropic and thermal effects are included through the harmonic partition functions.

With the use of TST or hTST, a long time scale simulation consists of identifying states of the system and finding the mechanism and rate of transitions from a current state to a new state. The key thing is to *find* the relevant mechanism and not just *assume* a mechanism. Often, preconceived notions of the transition mechanism have turned out to be incorrect. One example of that is the mechanism of adatom diffusion on Al(100) [27]. When hTST is used, the most challenging part of the calculation is the search for the low lying saddle points without knowledge of the possible final states. A typical simulation system includes a hundred atoms or more, which means that saddle points need to be found in a space with at least several hundred degrees of freedom. The large number of degrees of freedom makes this a challenging problem. We will now review briefly various approaches that have been taken to address this issue.

7.2.1   Bias Potentials

One approach is to increase the probability that the system escapes out of a given initial state by adding a repulsive potential energy, a bias potential, to the actual interaction potential in such a way as to reduce the activation energy of the transitions [130, 131, 17, 6, 132, 133, 134]. The most accurate and efficient formulation of such an approach is the hyperdynamics method developed by Voter [17, 6]. It is important to make sure the bias potential vanishes at the transition state. Within the hTST approximation, this means that the bias potential
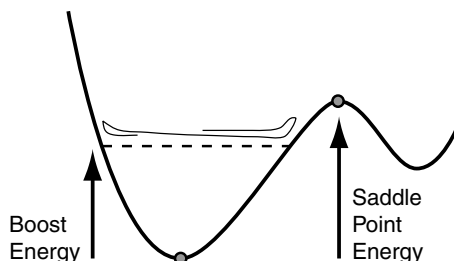
Figure 7.1: A simple bias potential can be constructed to accelerate the dynamics of a system. The true potential is replaced by a constant equal to the "boost energy" whenever the true potential drops below the boost energy. The boost energy has to be set less than the saddle point energy so as not to alter the transition state. A stylized trajectory is shown for illustration. With the bias potential, the potential energy of the system never drops below the boost energy.

needs to vanish at the potential energy rim near the relevant saddle points. Voter's bias potential is guaranteed to do this, even if it exceeds the saddle point energy within the initial state energy basin. It is, in fact, very important to allow the bias potential to greatly exceed the saddle point energy within the basin in order to get any appreciable acceleration of transitions in systems with many degrees of freedom. We will illustrate this with an example.

One might expect that just a slight filling of the potential well, reducing the energy difference between the bottom of the well and the saddle point, could lead to significant acceleration. This is true when the system has only very few degrees of freedom. But, each degree of freedom brings in $\frac{1}{2}kT$ of kinetic energy and a system with enough degrees of freedom almost always has more kinetic energy than the saddle point energy. It is still a rare event to find enough energy focused in the right degree of freedom to bring the system through the saddle point region. The effect of dimensionality is best illustrated by applying a bias potential that leads to a flat-bottom potential. This method has, in fact, been proposed as a viable way of accelerating dynamics simulations [132, 133]. When the true potential energy of the system is less than a certain value, the *boost energy*, the system evolves on a constant potential energy surface equal to the boost energy, but when the true potential energy of the system is higher than the boost energy, the system evolves on the unbiased potential surface. This type of bias potential is illustrated in Fig. 7.1. Results of calculations using this approach on a two-dimensional potential surface are shown in Fig. 7.2. The functional form and parameters of the potential function are given in the
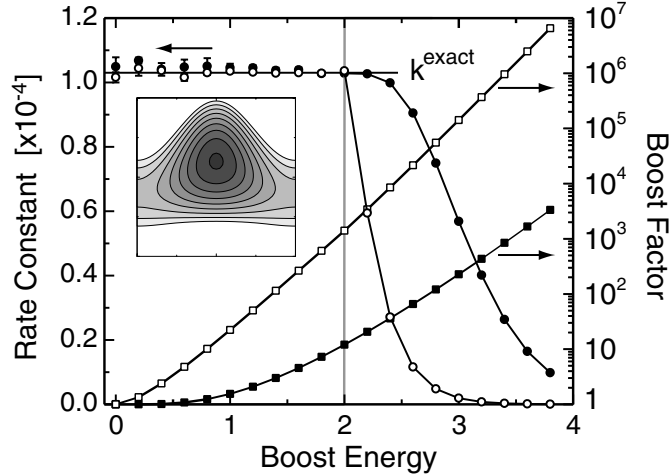
Figure 7.2: The escape rate from a two-dimensional potential well (shown in the inset) and a ten-dimensional well (see Appendix) calculated with accelerated dynamics using the bias potential illustrated in Fig. 7.1. The acceleration obtained, i.e. effective time in the evolution of the system divided by the time of the simulated, boosted trajectory is referred to as the boost factor. The escape rate for $k_B T = 0.2$ is shown with open circles for the two-dimensional system and filled circles for the ten-dimensional system. The calculated rate is the same as the exact value, k^{exact}, for boost energy up to the saddle point energy of 2.0 (gray vertical line). When the boost energy is increased above the saddle point energy, the calculated escape rate drops rapidly below the exact value. The boost factor is also shown with open squares for two-dimensional system and filled squares for ten-dimensional system. With the maximum allowable boost energy, equal to the saddle point energy, the bias potential is accelerating the dynamics by a factor of 1400 in the two-dimensional system, but only by a factor of 12 in the ten-dimensional system. This illustrates the poor scaling of the boost factor with the number of degrees of freedom when this kind of bias potential is used. In the higher-dimensional system, the total potential energy is larger than the saddle point energy most of the time so the system spends little time in the region of constant potential, resulting in small boost.

Appendix. In these simulations, the time step was adjusted when the system went from the biased to the unbiased regions, and vice versa, to ensure that energy was conserved, despite the discontinuity in the force. The escape rate for the system was calculated by counting the number of times the system left the basin shown in the inset in a fixed amount of simulation time. The figure shows how the escape rate varies with increasing boost energy. As the boost energy is raised up to the saddle point energy (gray line), the calculated escape rate remains constant and agrees well with the exact rate, k^{exact}, calculated by Voter [17]. The factor by which the dynamics are accelerated as compared with direct classical dynamics, the so-called boost factor [17], is 1400 when the boost energy equals the saddle point energy. Remarkably, the calculated escape rate is still valid for this aggressive bias potential. As soon as the boost energy is increased over the saddle point, however, incorrect results are obtained for the calculated escape rate.

While a boost factor of 1400 is quite respectable, the problem arises when the number of degrees of freedom increases. The probability of finding the potential energy of the system below a saddle point energy becomes vanishingly small. To demonstrate this, a potential energy surface for a system with ten degrees of freedom was constructed by adding up five of the two-dimensional potential surfaces described above (see the Appendix). The potential energy and the kinetic energy of this ten-dimensional system is on average five times higher than that of the two-dimensional system, but the activation energy, i.e. the saddle point energy minus the minimum energy, is the same. The calculated escape rate as a function of the boost energy for this larger system is also shown in Fig. 7.2 along with the boost factor. The higher-dimensional system spends less time in the region where the potential energy equals the bias and so the boost factor is smaller than for the two-dimensional case. The maximum boost factor is only twelve, and this is still a very small system for practical applications. For a twenty dimensional system, the maximum boost factor is less than two. This illustrates that a flat bias potential can only be effective in trivially small systems and that an effective bias potential must greatly exceed the saddle point energy in the interior of the potential energy basin but then drop to zero at the potential energy rim.

Voter's bias potentials are efficient in that the boost factor scales well with system size. However, the evaluation of his bias potentials is complicated and complex potential surfaces with small ripples reduce their effectiveness. Possibly, simpler bias potentials can be constructed that still scale well with dimensionality - that remains to be seen. Recently, a bias potential based on energy per atom has been proposed [134]. Here a strong bias is applied except in regions where the energy of any one of the atoms exceeds a predefined value. Although the energy per atom is not a well defined quantity for complicated potential functions (for example interaction between water molecules that include induction) and *ab initio* calculations, it can be defined for some simple potential functions. This method, however, seems to be applicable only to certain types of processes and simple systems. For example, in the Al/Al(100) system described in Sec. 7.4, the lowest energy process for adatom diffusion is via a concerted displacement process involving two atoms. In this process, the energy of the adatom does not increase from the minimum to the saddle point, but rather drops as the adatom moves into the surface. Surface and bulk atoms which are more highly coordinated than adatoms and have lower energy should not be subject to the
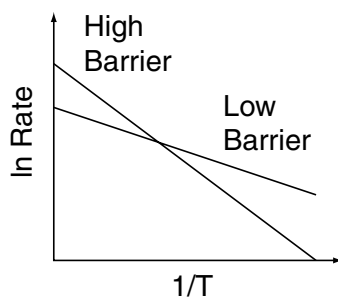
Figure 7.3: The temperature dependence of the rate of two different transitions from the same initial state is shown. The transition with the lower energy barrier corresponds to the line with smaller slope. A typical situation is shown in which the higher barrier process has a higher prefactor (intercept with vertical axis). At low temperature, on the right side of the graph, the rate of the low barrier process will be higher than that of the high barrier process. At high temperature, the rate of the higher energy, higher entropy process becomes larger. This illustrates why a system cannot simply be heated to use a short simulation to find out which events would occur at low temperature on a longer time scale.

same boost energy as the adatom. A different bias potential would, therefore, need to be applied to atoms in different environment and during a concerted displacement process the bias would need to be changed continuously as the environment of the atoms changes. This makes the simulation of all but the simplest systems impractical with this method. The point is that constructing bias potentials of general applicability is non-trivial.

7.2.2   Heating

Another, perhaps simpler approach for identifying transitions is to increase the rate of rare events by simply heating the system. If the atoms have more energy, they will more likely undergo transitions. One should, however, not expect the favored transition mechanism to be the same at the higher temperature. This situation is illustrated in Fig. 7.3. The temperature dependence of the rate of two possible processes the system can undergo from a given initial state is shown. One of the processes has a low activation energy and small prefactor while the other has a high activation energy and large prefactor. At low temperature, the low barrier process will have a higher rate and dominate the dynamics. At high temperature, entropy becomes more important and the process with higher prefactor dominates even though the energy barrier is higher. An even more serious problem arises when the thermodynamic state of the system changes as it is heated up high enough to make transitions observable in a short, classical dynamics simulation. An example of this

is diffusion of admolecules on an ice surface. If an ice slab is heated up high enough to make the diffusion events frequent enough, the surface melts and the diffusion mechanism becomes quite different from the low temperature, long time scale diffusion mechanism.

High temperature dynamics can, however, in favorable cases be used to search for the relevant mechanism if many searches are carried out. Repeated simulations from the same initial state can be used to identify several possible transition mechanisms. Within hTST the task is then to find the activation energy and prefactor for each mechanism and predict which one(s) would be relevant at the lower temperature of interest. Transitions can be detected from the high temperature dynamics by periodic minimization to the nearest local minimum on the energy surface and the nudged elastic band method [13, 3, 4] can then be used to find the minimum energy path connecting the given initial state with the final state found from the high temperature dynamics. The maximum energy on the minimum energy path gives the activation energy. This approach was used by Sørensen et al. to study the mechanism of contact formation as a metal tip is brought up to a metal surface on a laboratory time scale at room temperature [46]. While earlier classical dynamics simulations had seen a sudden jump-to-contact when the tip-surface distance became very small, the method described above could be used to show that a sequence of activated processes would occur at a larger tip-surface distance on the time scale of the experiments.

By assuming the system is harmonic also at the high temperature and given a minimum value for the prefactors, Sørensen and Voter formulated a concise prescription for the length of time for which the high temperature dynamics need to be run in order to safely advance the system to a new state [135]. This relieves the pressure of finding all possible transition mechanisms. In their method, the escape time calculated from hTST at the low temperature is recorded for each transition observed in the high temperature dynamics. After the simulation has been run long enough at the high temperature, the system is advanced by the first process that would have been observed at the low temperature. They refer to the method as temperature accelerated dynamics (TAD).

In this work we present a method that is quite different from the ones described above. Classical dynamics are not used in any form but instead the system is pushed up the potential energy surface using the so-called dimer method [2] to find saddle points. The rate of transitions through the vicinity of each saddle point is then estimated within hTST

and a kinetic Monte Carlo method used to simulate the evolution of the system over long time scales [136]. This method is easy to implement and, compared to existing methods, may require less computational time for small systems.

7.3   The Long Time Dynamics Method

Simulations of systems over long time scale are carried out by combining kinetic Monte Carlo (KMC) and saddle point searches. KMC relies on knowing the rate and mechanism of all the relevant transitions from a given initial state. Within hTST this corresponds to finding all the low energy saddle points on the rim of the potential energy basin corresponding to the initial state. The dimer method is used to search for these saddle points.

7.3.1   The Dimer Method

The dimer method has been described in detail in a previous publication [2]. Only a brief overview will be given here. The dimer is made up of two images (replicas) of the system. These images are separated in space by a finite distance displacement along a vector $\hat{\mathbf{N}}$. For an empirical potential this can be small. Here we have used 0.005 Å.

There are two parts to each dimer move. The first part is dimer rotation. The lowest energy orientation of the dimer is along the lowest curvature mode. If the dimer is free to rotate, the forces acting on the two images will pull the dimer to the lowest curvature mode. This is done by defining a rotational force which is the difference in the force at the two images. Minimizing the energy of the dimer with respect to this rotational force aligns the dimer with the lowest curvature mode (this feature was used by Voter in his construction of bias potentials in hyperdynamics [6]). A modified Newton's method can be used to make this rotation efficient [2]. An important aspect of the dimer method is that it only requires the first derivative of the energy, not the second derivatives.

The second part of the algorithm is translation of the dimer. A first order saddle point on a potential surface is at a maximum along the lowest curvature direction and a minimum in all other directions. In order to converge to a saddle point, the dimer is moved up the potential along the lowest curvature mode, and down the potential in all other directions. This is done by defining an effective force on the dimer, in which the true force due to the

potential acting at the center of the dimer has the component along the dimer inverted. Minimizing with respect to this effective force moves the dimer to a saddle point. With empirical potentials, minimization using the conjugate gradient method works well.

It is not necessary to fully converge the dimer orientation at each translational step. We have found it most efficient to pick a certain tolerance for the rotational force. This sets how precisely the dimer is oriented along the lowest curvature mode before it is moved. The dimer is rotated at least once and possibly a few times until the rotational force is less than the specified tolerance. In these calculations the maximum rotational force was set at 5 meV/Å.

For the Al adatom on an Al(100) surface, the dimer requires on average 400 force evaluations to converge on a saddle point [2]. A very important feature of the dimer method is the slow increase in the number of required force evaluations as more degrees of freedom are added to the system. This is to be contrasted with mode following approaches [18, 19, 20, 21, 24, 25, 22] where the matrix of second derivatives is constructed and diagonalized at each step in the search for a saddle point. with Other methods for finding saddle points using only first derivatives of the energy have been proposed [137, 138], but it is unclear how their efficiency compares with the dimer method.

In the long time scale calculations presented here, several dimer searches were launched for each state visited. Dimer searches were started from configurations close to the potential minimum. The easiest way to choose a starting position is to make a random displacement away from the minimum. If *all* the atoms in the system are included in the random displacement, the dimer search can be biased towards finding higher energy processes involving many atoms. We have found that it is better to displace only atoms within a local region. For each dimer search in the Al(100) ripening calculations (described in Sec. 7.5), an under-coordinated surface atom was chosen at random to be the center of the initial, local displacement. The displacements had a Gaussian distribution with a mean of 0.2 Å in each degree of freedom. The region consisted of the central atom and its first and second neighbors. We have found that a continuous distribution in the magnitude of the displacements is better than a fixed magnitude because it increases the variety of saddle points found. In principle, a scheme like this can eventually find all saddle points around a minimum simply because the starting points of the dimer search can be at any point in

space. In practice, there is no guarantee that a complete list of saddle points can be found in a reasonable amount of time. Our experience from studies of adatom diffusion processes [2] and island rearrangement processes [5] is that a dimer search started in a random direction most likely finds one of the low energy saddle points.

7.3.2   Kinetic Monte Carlo

Kinetic Monte Carlo [139, 140, 141, 142, 143] is a powerful method that can be used to extend the time scale of simulations far beyond the vibrational time scale. If a list of possible transitions for a given initial state is available, a random number can be used to choose one of the processes and evolve the system to a new state. The probability of choosing a certain transition is proportional to its rate, r_i. On average, the amount of time that would have elapsed in order for this process to occur is

$$\Delta t = \frac{1}{\sum r_i},$$ (7.2)

which is independent of the chosen transition. It may also be important to include the appropriate distribution of escape times. For random uncorrelated processes, this is a Poisson distribution. If μ is a random number from 0 to 1, the elapsed time for a particular transition is given by

$$\Delta t = \frac{-\ln \mu}{\sum r_i}.$$ (7.3)

The system is then advanced to the final state of the chosen transition and the process is repeated.

In a traditional kinetic Monte Carlo simulation, all transitions that can ever occur in the system, along with their rates, must be known before the simulation starts. Ideally, the rates are estimated from some description of the atomic interactions [144], such as an empirical interaction potential or ab initio calculations, but the problem is to know in advance the mechanism of the relevant transitions for each possible configuration of the atoms. The requirement of knowing and tabulating the relevant transitions ahead of time limits the method to simple systems. Systems which can undergo complicated transitions involving several atoms, such as the aluminum system described in Sec. 7.4, or where atoms do not sit at lattice sites are extremely difficult to model with traditional KMC.
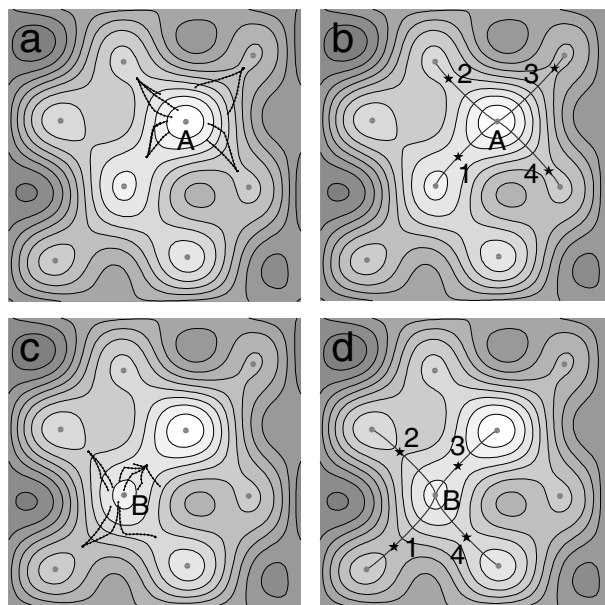
Figure 7.4: Application of the long time scale simulation method to a model two-dimensional potential surface. The system is initially in state A. (a) Ten dimer searches are started from random positions around the minimum. They converge on four distinct saddle points (two of the searches practically overlap). (b) The system is then made to slide down the minimum energy path (gray lines) on either side of the saddle points, indicated with ⋆. Here, all four saddle points have a minimum energy path starting at the initial state minimum A, but this does not have to be the case. The rate of each process is then calculated using harmonic transition state theory. A process is chosen at random using the kinetic Monte Carlo algorithm. In this case, process 1 gets chosen. The system is moved to the final state of this process, to minimum B. (c) Dimer searches are run from the new minimum, again four distinct saddle points are found. (d) Minimum energy paths are traced out, and the process repeated.

### 7.3.3 Combined Dimer and Kinetic Monte Carlo

The dimer method can be used to relax some of the limitations of the traditional implementation of the kinetic Monte Carlo scheme. If the dimer method is used to find possible transitions, there is little limitation on the complexity in terms of the number of atoms or the spatial extent of the transition. Also, the energy barriers do not need to be known before the calculation is started. Furthermore, the atoms do not need to be mapped onto a lattice and it is not necessary to anticipate all possible states of the system.

The method is illustrated for a two-dimensional model potential in Fig. 7.4. The system is started at a potential minimum, A. When a new state is visited, a swarm of dimer searches is sent out from the vicinity of the potential energy minimum. For the calculations described in Sec. 7.5, either 25 or 50 dimer searches were used. In this example, ten random

displacements from the position of the minimum were chosen as starting points of dimer searches. Figure 7.4a shows the path of the ten dimer searches. In this calculation, four distinct saddle points ($\star$) were found. The system is then quenched on either side of each saddle point in order to verify that it lies on a minimum energy path (shown in gray) from the given initial state minimum. All the saddle points found in this case did connect back to the initial minimum. If not, the saddle point is discarded from the list of possible transitions. In the same way as described in Sec. 7.3.2, a transition is chosen from the list, the system is advanced to the final state of that transition, and the time interval associated with the transition is added to the accumulated time. In this example, transition 1, which corresponds to the lowest barrier was chosen, and the system was advanced to state B. From the new minimum the process is repeated. New dimer searches are sent out (Fig. 7.4c), the saddle points verified (Fig. 7.4d) and then one is chosen for the next transition.
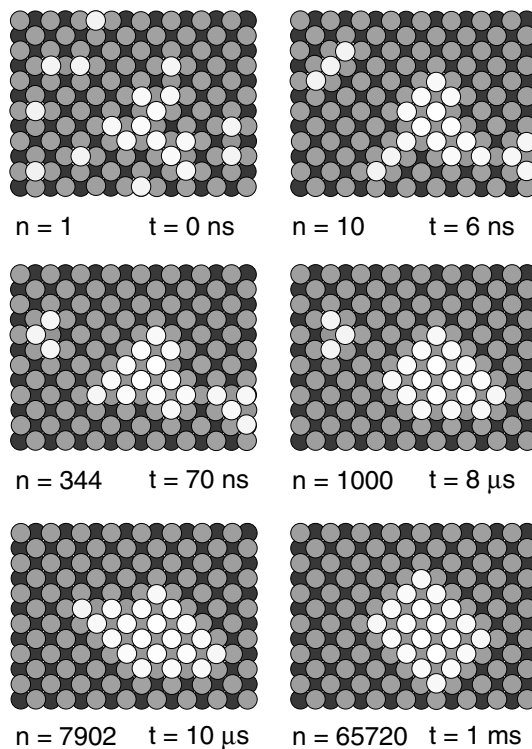
7.4   The Al(100) Surface

We have chosen dynamics of Al adatoms on an Al(100) surface as a test problem for the long time scale simulation method. There are two reasons for that. First, an accurate embedded atom potential of the Voter and Chen form exists for aluminum [31]. Secondly, Al(100) is a rich system because there are many different transitions with rather low energy barrier even for just a single adatom on the Al(100) surface. We have previously studied this system extensively with the dimer method [2]. The four lowest energy processes found are shown in Fig. 7.5. A particularly interesting aspect of the Al(100) system is that a concerted displacement process has a lower energy barrier than the direct hop process. This was shown by Feibelman with density functional theory calculations [27]. In these simulations, the system consists of a six layer slab with 50 atoms per layer. The bottom two layers are held frozen and the top surface is left open to vacuum. The dimer calculations revealed sixty different transitions for adatom diffusion in 1000 dimer searches [2]. On average the low energy transitions: the concerted displacement involving two, three and four atoms, and the hop, were found three quarters of the time. One quarter of the time a wide collection of higher energy processes were found. This is a very important result because it indicates that a dimer search started in a random direction has high probability of finding a saddle
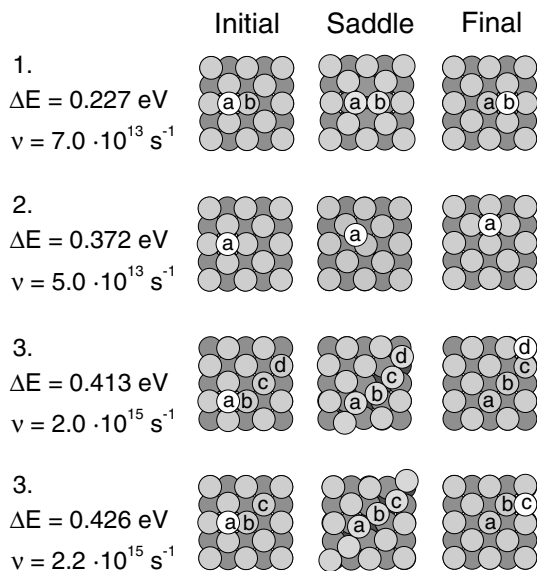
Figure 7.5: The four lowest energy transitions found with the dimer method for the diffusion of an Al adatom on Al(100). For each transition, the initial state, the saddle point configuration, and the final state are shown. Atoms are shaded by depth and the atoms that move the most in each transition are labeled. The energy of each transition is given in eV. The lowest energy transition is the two-atom concerted displacement (1). The hop (2) has similar activation energy as a thee-atom (3) and four-atom (4) concerted displacement processes. Because Al adatoms can so easily displace atoms in the surface, a growing Al(100) surface can undergo a great variety of transition which would be hard to find by guesswork.

point for one of the low barrier transitions.

In the KMC scheme, it is assumed that all relevant transitions have been found. Transitions that have such a high activation energy and/or low prefactor that the rate is insignificantly small are irrelevant. Given that the two-atom concerted displacement process is found a quarter of the time, there is an 80% certainty of finding all the four transitions that are equivalent by symmetry in 50 dimer searches. The fact that all of the equivalent processes may not be found is not so serious since only one gets chosen at random in the KMC simulation. The important thing is that a representative sample of the transitions is found over the relevant range of activation energy and that no category of transitions is excluded. The error mainly shows up in the time scale of the simulation which is dominated by the low barrier transitions. If only half the low barrier transitions are found, the simulation clock will be running two times too fast. Since the challenge of reaching long time scale is a matter of spanning several orders of magnitude, a factor of two is typically not a serious issue. With a modest number of dimer searches, the method can give a good

Figure 7.6: Snapshots from a simulation of ripening of Al adatoms on an Al(100) surface. Initially twenty atoms were deposited at random on the surface, a coverage of 0.4. After 6 ns (10 transitions) all the adatoms have merged to form clusters. After 70 ns (344 transitions) a large, compact island has formed, but there are still two outlying islands. The trimer in the upper left has four possible rearrangements with a low activation barrier. Many of the 344 transitions correspond to these rearrangements but old configurations are stored during the simulation so the repeat processes do not require new dimer searches. After 8 μs (1000 transitions) the large island has taken a more compact shape and merged with one of the smaller islands. Finally after 10 μs (7902 transitions) a single large island has formed, and at 1 ms (65720 transitions) the island arrives at its most compact shape.

qualitative idea of how the system will behave at low temperature over long time scale. For an accurate simulation, it is clear that many dimer searches need to be carried out. Fortunately, the different searches can easily be carried out in parallel on loosely connected cluster of computers.

7.5 Application to Island Ripening

The results of a kinetic Monte Carlo simulation coupled with dimer searches is shown in Fig 7.6. Initially 20 atoms, a coverage of 0.4, were randomly deposited on the Al(100) surface consisting of 50 atoms using classical dynamics. Then the system was quenched to

the nearest local energy minimum. This configuration is shown in the first panel (n=1). Then, the time evolution of the system at 300 K was simulated. On average, 17 distinct processes were found when 25 dimer searches were carried out. When 50 dimer searches were used, this number rose to 23, indicating that some, but not too many, transitions were missed with 25 searches. During the first ten transitions, the adatoms diffused via the concerted displacement or hop processes to form clusters. After 344 transitions, a large island has formed, but two smaller islands also exist. After 1000 transitions, one of the small clusters has merged with the large island which has taken a more compact shape. Finally after 7902 transitions, a single compact island is formed, and at 1 ms, (65720 transitions) the island reaches its lowest energy shape. The mechanism for atoms crossing a corner or filling in a kink site invariably involves a two-atom displacement process with a substrate atom.

Of the 100,000 transitions in this simulation, only 300 brought the system to a new state. Many processes involved rearrangements of the three atom cluster or an adatom moving along the edge of an island. After each transition, the new state is compared to a table of all the old configurations. If the new state has been seen before, no dimer searches are performed. Rather, a new transition is chosen from the old list of processes. In this way, low barriers which are seen frequently do not contribute significantly to computational time. The total computer time for the simulation was approximately one week on a PC.
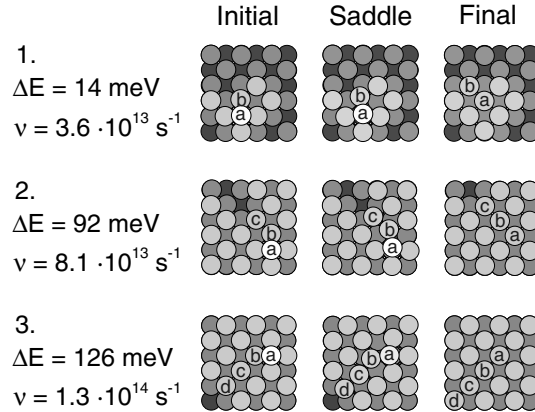
7.6   Deposition and Surface Growth

Multiple time scale simulations of Al(100) crystal growth were carried by combining classical dynamics of deposition events with the long time scale simulation of the time intervals between deposition events. Interesting processes can take place during the deposition event [145, 146], but the excess energy released as the incoming atom binds to the surface dissipates quickly, in one or two picoseconds [145, 146]. This period of time can easily be simulated by direct classical dynamics. After the 'hot-spot' has cooled down and the system has thermalized, transition state theory can be applied for the activated diffusion processes. The long time scale calculation was done in the same way as described in Sec. 7.5, but at each step a deposition process was added to the table of possible events. A smaller system was

used than in the ripening simulations, the cell consisted of 32 atoms per layer. Atoms were deposited at a rate of one monolayer per millisecond (32×10^3 s^{-1}). A deposition process was simulated by placing an aluminum atom at a random position 10Å above the surface and giving it an initial velocity towards the surface characteristic of the simulation temperature. Classical dynamics were run until the component of the velocity of the deposited atom perpendicular to the surface had changed sign twice. At this point the system was considered to be equilibrated and the KMC simulation took over. In this way, the short deposition time scale was simulated accurately as well as the longer time scale of thermally activated diffusion processes.

Growth at 100K was first simulated. A total of 9.5 layers were deposited in 9.8 ms. This took 500 events in the KMC simulation of which 302 were deposition events. An important question in crystal growth is how smoothly the surface grows, i.e. how well a layer gets completed before the next layers starts forming. A key issue is how an adatom that lands on top of an island manages to descend down to the more stable edge site. Typically, an adatom descents by a two-atom concerted displacement process, as is shown in Fig. 7.7, but the simulation also revealed interesting three- and four-atom descent processes where the adatom started out two or three sites away from the edge. The transitions involving three and four atoms have somewhat larger activation barrier, 92 meV and 126 meV, but were frequently observed in the simulation. The ease if these types of descent processes in the Al(100) system make it easy for aluminum to grow layer-by-layer, especially for this small 32 atom per layer cell.

The growth simulation was then repeated for a lower temperature, 30 K. A total of 7.3 layers were grown in 5.6 ms. The KMC simulation involved 264 events of which 232 were deposition events. Snapshots of the system during the growth simulation are shown in Fig. 7.8. The range of possible surface processes is then much more limited. The long range descent events observed at 100 K were not observed at 30 K. Since individual adatoms were less likely to descend, rougher surface morphology could form. But, after structures with several under-coordinated atoms had built up, some even not sitting at well defined lattice sites, concerted processes involving several atoms would typically occur and smoothen out the surface. Two of these transitions are shown in Fig. 7.9. In the initial state of the first process, an atom has been deposited on an island consisting of four atoms resulting in a

	Initial	Saddle	Final

1.
ΔE = 14 meV
$\nu = 3.6 \cdot 10^{13}$ s^{-1}

2.
ΔE = 92 meV
$\nu = 8.1 \cdot 10^{13}$ s^{-1}

3.
ΔE = 126 meV
$\nu = 1.3 \cdot 10^{14}$ s^{-1}

Figure 7.7: Three processes observed during a simulation of Al atom deposition on Al(100) at 100 K. The deposition rate was one monolayer per millisecond. Ten monolayers were grown, each consisting of 32 atoms. The processes involve descent of an adatom into a lower layer. Most often an adatom descends by a two-atom concerted displacement process as shown in (1). But, processes involving concerted displacement of three and four atoms where the adatom starts two or three sites away from the island edge were also often observed in the simulations. While the activation energy of these long range processes is higher, it is still small compared with the 0.22 eV activation energy for adatom diffusion on the flat (100) terrace. The remarkable ease by which an adatom can descend from atop an island on Al(100) leads to layer-by-layer growth even at temperature as low as 30 K.

structure where five atoms are not sitting on lattice sites. Then, a transition occurs which moves all five atoms simultaneously into lattice sites, two in the top layer and the other three in a lower layer. In the other process shown in Fig. 7.9, a deposited atom lands on top of an island, starting a third incomplete layer. In the subsequent transition, five atoms slide down in a concerted way. The third layer atom ends up in the second layer, the two second layer atoms end up in the first layer, and two first layer atoms get pushed out into different sites. The energy barrier of these two five-atom displacement processes is only 48 meV and 45 meV. Evidently, when the surface becomes rough, low barrier cascading events where many atoms move into more stable sites become possible. Even at this low temperature, the surface morphology remains smooth, although the basic adatom diffusion process on the flat Al(100) terrace, which has an activation energy of 0.2 eV, is not active.

A simulation was also carried out at an intermediate temperature, 70 K. As the temperature was lowered from 100 K to 30 K, the average barrier height of observed transitions also dropped, as expected. This is shown in Fig. 7.10. At 100 K the system was able to cross barriers of 0.18 eV, but at 30 K the highest barrier that was overcome was less than 50 meV.

n = 0   n = 88   n = 176   n = 232
t = 0 ms   t = 1.8 ms t = 3.8 ms t = 5.6 ms
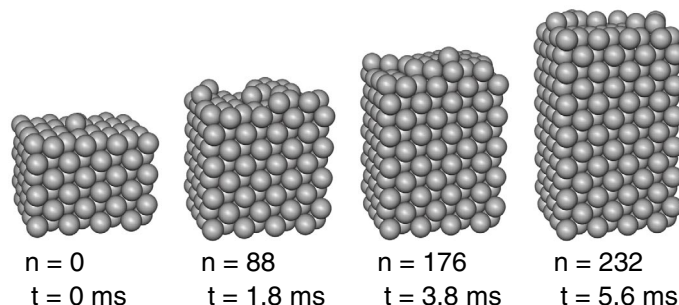
Figure 7.8: Seven monolayers of Al were grown at 30 K on the Al(100) substrate. The deposition rate was one monolayer per millisecond. Even at this low temperature, the layers grown are free of defects and the surface is smooth.

Animations of the growth and ripening simulations can be viewed on the web at http:// eon.chem.washington.edu.

7.7   Efficiency

It is important to know how a simulation method scales with system size. To determine this, we can consider how the number of force evaluations changes when the system is doubled in size. Force evaluations are considered rather than total computational time, because the evaluation of the force can scale differently depending upon the complexity of the interaction potential. In favorable situations, evaluation of the force scales linearly with system size.

Figure 7.11 shows an illustration of a system which has been doubled in size by joining two identical replicas of the system. In the larger system, there are twice as many possible transitions. In the crudest implementation of the simulation method, the number of dimer searches would then increase in proportion to the system size. If, however, the transitions are local, i.e. they only involve a small subset of the atoms, then large portions of the system will be unaffected by any one transition. Transitions that take place in these unaffected regions will not require repeated saddle point searches. Rather, only the transitions which are affected by the last transition need to be updated and new transitions need to be added only for the region where the last transition took place. With the dimer method, it is also relatively straight forward to reconverge a set of known saddle points which have changed only slightly because of a nearby transition. Therefore, if the transitions are local, the
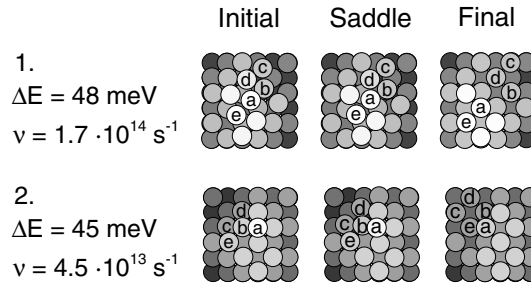
Figure 7.9: Two processes observed during the 30 K deposition simulation of Al atoms on Al(100) shown in Fig. 7.8. At this low temperature only very low energy processes can occur. The surface tends to grow rough temporarily and regions with islands stacked on top of islands form. The initial state of the first process resulted from classical dynamics simulations of the deposition of atom (a) on top of a four atom island consisting of atoms (b), (c), (d), and (e). The five atoms are not sitting at lattice sites. In a concerted, activated process all five atoms move to lattice sites, two in the top layer ((a) and (e)) and three in the lower layer ((b), (c), and (d)). In the second process, five atoms slide in a concerted way down the side of an island stack. The adatom (a) in the third layer moves down to the second layer, while two atoms in the second layer, (e) and (b), move to the first layer and two atoms in the first layer get pused into different sites in the first layer. The barrier for these two multi-atom, sliding processes is remarkably low, less than 0.05 eV, and is active at 30 K even though adatom diffusion on the flat (100) terrace, which has a barrier of 0.22 eV, is not active.

number of new dimer searches that need to be carried out after a transition has occurred will not change as a function of the system size.

Another aspect to the scaling is the time increment at each transition. When the system size is doubled, there are on average twice as many transitions possible. The time increment for each transition will then be half as long, as can be seen from Eq. 7.3. If the simulation is carried out in such a way as to cover a fixed time interval, then twice as many transitions need to be simulated in a system that is double in size. The number of force evaluations required to simulate a fixed length of time, therefore, scales linearly with system size.

This method for carrying out long time scale dynamics simulations is efficient enough for it to be implemented with first principles calculations of atomic interactions such as density functional theory. We have implemented the method in the VASP code [58, 59, 60, 61]. More information about the implementation can be found on the web site http://ikazki01. chem.washington.edu/vasp/. We are currently applying the technique to the formation and break-up of boron clusters in silicon [147] and the diffusion of self-trapped excitons in silica glass [148].
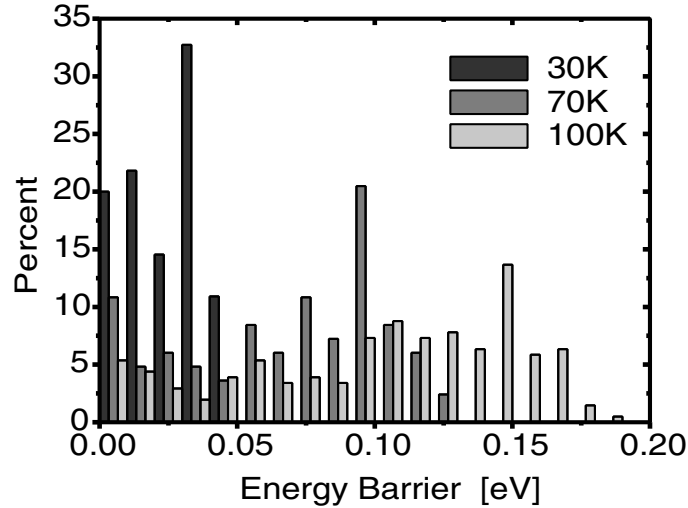
Figure 7.10: The average barrier height of events observed during the Al(100) growth simulations drops with temperature, as expected. In each simulation the deposition rate is one monolayer per millisecond. As the temperature is lowered, it gets harder and harder to surmount high barriers within the time interval between deposition events. At 100 K there is on average one surface diffusion process for every deposition event. At 30 K this ratio drops to one in eight. Even so, the Al(100) surface still grows smoothly at 30 K because as multi-atom cascade events such as those shown in Fig. 7.9.

7.8   Appendix: The Model Potentials

The bias potential algorithm described in the Introduction was first of all tested on the two dimensional potential described by Voter[17]

$$V(x, y) \;=\; \cos(2\pi x)(1 + 4y) + \frac{1}{2}(2\pi y)^2 + V_0 \tag{7.4}$$

This potential is periodic in x. A contour plot is shown in the inset in Fig. 7.2. We chose to set $V_0 = -1.203$, so that the minimum potential is zero and occurs at the point $(x, y) = (k + 0.5, 0.1013)$, where k is any integer. The saddle point has a potential $V_{SP} = 2$ and is located at $(k, -0.1013)$.

The ten-dimensional potential was constructed by adding up five of the two-dimensional potentials given by Eq. 7.4

$$\tilde{V}(x, y) \;=\; \sum_{i=1}^{5} V(x_{2i-1}, x_{2i}) \tag{7.5}$$

128
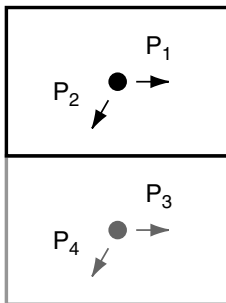
Figure 7.11: In a system of twice the size, there are on average twice as many possible transitions. If P_1 and P_2 are transitions in the original system, there are two more transitions, P_3 and P_4, possible in the larger system. In the simplest implementation of the simulation method, this means that the number of dimer searches increases linearly with system size. But, if the transitions are local, only a limited region of the system is affected by each transition and only a part of the system will require new dimer searches. Also, saddle points for transitions close to the transition region may only be slightly affected and previously found saddle point configurations can be reconverged with just a few dimer iterations. The number of dimer searches needed per transition does not, therefore, increase with system size. However, the time increment for each transition is only half as large in the larger system.

The results of the accelerated dynamics calculations where a repulsive bias potential is added to these potential functions so as to create a flat-bottom potential are shown in Fig. 7.2.

7.9 Acknowledgments

BIBLIOGRAPHY

[1] P. Pechukas, in *Dynamics of Molecular Collisions*, edited by W. Miller (Plenum Press, N.Y., 1976), part B.

[2] G. Henkelman and H. Jónsson, J. Chem. Phys. **111**, 7010 (1999).

[3] G. Henkelman and H. Jónsson, J. Chem. Phys. **113**, 9978 (2000).

[4] G. Henkelman and H. Jónsson, J. Chem. Phys. **113**, 9901 (2000).

[5] G. Henkelman, G. Jóhannesson, and H. Jónsson, in *Progress on Theoretical Chemistry and Physics*, edited by S. Schwartz (Kluwer Academic, New York, 2000).

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

[7] G. Henkelman and H. Jónsson, Phys. Rev. Lett. **86**, 664 (2001).

[8] D. C. Seets, C. T. Reeves, B. A. Ferguson, M. C. Wheeler, and C. B. Mullins, J. Chem. Phys. **107**, 10229 (1997).

[9] A. V. Walker and D. A. King, Phys. Rev. Lett. **82**, 5156 (1999).

[10] G. Henkelman and H. Jónsson, J. Chem. Phys. (submitted).

[11] C. Wert and C. Zener, Phys. Rev. **76**, 1169 (1949).

[12] G. H. Vineyard, J. Phys. Chem. Solids **3**, 145 (1957).

[13] H. Jónsson, G. Mills, and K. W. Jacobsen, in *Classical and Quantum Dynamics in Condensed Phase Simulations*, edited by B. J. Berne, G. Ciccotti, and D. F. Coker (World Scientific, Singapore, 1998).

[14] G. T. Barkema and N. Mousseau, Phys. Rev. Lett. **77**, 4358 (1996).

[15] N. Mousseau and G. T. Barkema, Phys. Rev. E **57**, 2419 (1998).

[16] G. T. Barkema and N. Mousseau, The trailing image, personal communication, 1999.

[17] A. F. Voter, J. Chem. Phys. **106**, 4665 (1997).

[18] G. M. Crippen and H. A. Scheraga, Archives of Biochemistry and Biophysics **144**, 462 (1971).

130

[19] R. L. Hilderbrandt, Computers & Chemistry **1**, 179 (1977).

[20] C. J. Cerjan and W. H. Miller, J. Chem. Phys **75**, 2800 (1981).

[21] J. Simons, P. Jørgensen, H. Taylor, and J. Ozment, J. Phys. Chem. **87**, 2745 (1983).

[22] J. Nichols, H. Taylor, P. Schmidt, and J. Simons, J. Chem. Phys. **92**, 940 (1990).

[23] D. J. Wales, J. Chem. Phys. **101**, 3750 (1994).

[24] J. Baker, J. Comp. Chem. **9**, 465 (1988).

[25] D. J. Wales, J. Chem. Phys. **91**, 7002 (1989).

[26] C. J. Tsai and K. D. Jordan, J. Phys. Chem. **97**, 11227 (1993).

[27] P. J. Feibelman, Phys. Rev. Lett. **65**, 729 (1990).

[28] M. C. Payne, M. P. Teter, D. C. Allen, T. A. Arias, and J. D. Joannopoulos, Rev. Mod. Phys. **64**, 1075 (1992).

[29] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C: The Art of Scientific Computation*, 2nd ed. (Cambridge University Press, Cambridge, 1992), p. 420.

[30] R. G. D. Valle and H. C. Andersen, J. Chem. Phys **97**, 2682 (1992).

[31] A. F. Voter and S. P. Chen, Mat. Res. Soc. Symp. Proc. **82**, 2384 (1987).

[32] G. Brocks, P. J. Kelly, and R. Car, Phys. Rev. Lett. **66**, 1729 (1991).

[33] A. P. Smith and H. Jónsson, Phys. Rev. Lett. **77**, 1326 (1996).

[34] R. Stumpf and M. Scheffler, Phys. Rev. B **53**, 4958 (1996).

[35] P. J. Feibelman, Phys. Rev. Lett. **81**, 168 (1998).

[36] G. Mills and H. Jónsson, Phys. Rev. Lett. **72**, 1124 (1994).

[37] G. Mills, H. Jónsson, and G. K. Schenter, Surface Science **324**, 305 (1995).

[38] B. Uberuaga, M. Levskovar, A. P. Smith, H. Jónsson, and M. Olmstead, Phys. Rev. Lett. **84**, 2441 (2000).

[39] J. Song, L. R. Corrales, G. Kresse, and H. Jónsson, Phys. Rev. B (in press).

[40] W. Windl, M. M. Bunea, R. Stumpf, S. T. Dunham, and M. P. Masquelier, Phys. Rev. Lett. **83**, 4345 (1999).

[41] C. T. R. Stumpf, C. L. Liu, Phys. Rev. B **59**, 16047 (1999).

[42] T. C. Shen, J. A. Steckel, and K. D. Jordan, Surf. Sci. **446**, 211 (2000).

[43] M. Villarba and H. Jónsson, Surf. Sci. **317**, 15 (1994).

[44] M. Villarba and H. Jónsson, Surf. Sci. **324**, 35 (1995).

[45] E. Batista and H. Jónsson, Comput. Mat. Sci. (in press).

[46] M. R. Sørensen, K. W. Jacobsen, and H. Jónsson, Phys. Rev. Lett. **77**, 5067 (1996).

[47] T. Rasmussen, K. W. Jacobsen, T. Leffers, O. B. Pedersen, S. G. Srinivasan, and H. Jónsson, Phys. Rev. Lett. **79**, 3676 (1997).

[48] R. Elber and M. Karplus, Chem. Phys. Lett. **139**, 375 (1987).

[49] R. Czerminski and R. Elber, Int. J. Quantum Chem. **24**, 167 (1990).

[50] R. Czerminski and R. Elber, J. Chem. Phys. **92**, 5580 (1990).

[51] R. E. Gillilan and K. R. Wilson, J. Chem. Phys. **97**, 1757 (1992).

[52] J. C. Polanyi and Wong, J. Chem. Phys **51**, 1439 (1969).

[53] K. C. Pandey, Phys. Rev. Lett. **57**, 2287 (1986).

[54] J. Tersoff, Phys. Rev. B **39**, 5566 (1989).

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

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

[57] W. Kohn, A. D. Becke, and R. G. Parr, J. Phys. Chem. **100**, 12974 (1996).

[58] G. Kresse and J. Hafner, Phys. Rev. B **47**, 558 (1993).

[59] G. Kresse and J. Hafner, Phys. Rev. B **49**, 14251 (1994).

[60] G. Kresse and J. Furthmüller, Comput. Mater. Sci. **6**, 16 (1996).

[61] G. Kresse and J. Furthmüller, Phys. Rev. B **54**, 11169 (1996).

[62] J. P. Perdew, in *Electronic Structure of Solids*, edited by P. Ziesche and H. Eschrig (Akademie Verlag, Berlin, 1991).

[63] D. Vanderbilt, Phys. Rev. B **41**, 7892 (1990).

[64] H. Eyring, J. Chem. Phys. **3**, 107 (1935).

[65] E. Wigner, Trans. Faraday Soc. **34**, 29 (1938).

[66] J. C. Keck, Adv. Chem. **13**, 85 (1967).

[67] A. F. Voter and D. Doll, J. Chem. Phys. **80**, 5832 (1984).

[68] A. F. Voter and D. Doll, J. Chem. Phys. **82**, 80 (1985).

[69] M. McKee and M. Page, in *Reviews in Computational Chemistry*, edited by K. B. Lipkowitz and D. B. Boyd (VCH, New York, 1993), Vol. IV.

[70] F. Zimmermann and X. Pan, Phys. Rev. Lett. **85**, 618 (2000).

[71] D. G. Truhlar, B. C. Garrett, and S. J. Klippenstein, J. Phys. Chem. B **100**, 12771 (1996).

[72] D. G. Truhlar and B. C. Garrett, Annu. Rev. Phys. Chem. **35**, 159 (1984).

[73] J. B. Anderson, J. Chem. Phys. **58**, 4684 (1973).

[74] R. Marcus, J. Chem. Phys. **45**, 4493 (1966).

[75] G. Mills, G. K. Schenter, D. Makarov, and H. Jónsson, Chem. Phys. Lett. **278**, 91 (1997).

[76] G. Mills, G. K. Schenter, D. E. Makarov, and H. Jónsson, in *Classical and Quantum Dynamics in Condensed Phase Simulations*, edited by B. J. Berne, G. Ciccotti, and D. F. Coker (World Scientific, Singapore, 1998).

[77] R. P. Feynman and A. R. Hibbs, *Quantum Mechanics and Path Integrals* (McGraw Hill, New York, 1965).

[78] W. H. Miller, J. Chem. Phys. **62**, 1899 (1975).

[79] S. Coleman, in *The Whys of Subnuclear Physics*, edited by A. Zichichi (Plenum, New York, 1979).

[80] V. A. Benderskii, D. E. Makarov, and C. A. Wight, *Chemical Dynamics at Low Temperature* (Whiley, New York, 1994).

[81] H. C. Andersen, J. Chem. Phys. **72**, 2384 (1980).

[82] T. A. Halgren and W. N. Lipscomb, Chem. Phys. Lett. **49**, 225 (1977).

[83] M. J. Rothman and L. L. Lohr, Chem. Phys. Lett. **70**, 405 (1980).

[84] L. R. Pratt, J. Chem. Phys. **85**, 5045 (1986).

[85] A. Kuki and P. G. Wolynes, Science **236**, 1647 (1986).

[86] A. Ulitsky and R. Elber, J. Chem. Phys. **92**, 1510 (1990).

[87] C. Choi and R. Elber, J. Chem. Phys. **94**, 751 (1991).

[88] E. M. Sevick, A. T. Bell, and D. N. Theodorou, J. Chem. Phys. **98**, 3196 (1993).

[89] T. L. Beck, J. D. Doll, and D. L. Freeman, J. Chem. Phys. **90**, 3183 (1989).

[90] O. S. Smart, Chem. Phys. Lett. **222**, 503 (1994).

[91] S. Fischer and M. Karplus, Chem. Phys. Lett. **194**, 252 (1992).

[92] S. Fischer, Ph.D. thesis, Harvard University, 1992.

[93] I. V. Ionova and E. A. Carter, J. Chem. Phys. **98**, 6377 (1993).

[94] M. J. S. Dewar, E. F. Healy, and J. J. P. Stewart, J. Chem. Soc., Faraday Trans. 2 **80**, 227 (1984).

[95] D. T. Nguyen and D. A. Case, J. Phys. Chem. **89**, 4020 (1985).

[96] W. Quapp, Chem. Phys. Lett. **253**, 286 (1996).

[97] H. Taylor and J. Simons, J. Phys. Chem. **89**, 684 (1985).

[98] J. Baker, J. Comput. Chem. **7**, 385 (1986).

[99] N. P. Kopsias and D. N. Theodorou, J. Chem. Phys. **109**, 8573 (1998).

[100] J. D. Honeycutt and H. C. Andersen, Chem. Phys. Lett. **108**, 535 (1984).

[101] J. D. Honeycutt and H. C. Andersen, J. Chem. Phys. **90**, 1585 (1986).

[102] G. Mills, T. R. Mattsson, I. Mollnitz, and H. Metiu, J. Chem. Phys. **111**, 8639 (1999).

[103] D. W. Bassett and P. R. Webber, Surf. Sci. **70**, 520 (1978).

134

[104] A. C. Luntz and D. S. Bethune, J. Chem. Phys. **90**, 1274 (1989).

[105] C. B. Mullins and W. H. Weinberg, J. Chem. Phys. **92**, 4508 (1990).

[106] M. C. McMaster and R. J. Madix, J. Chem. Phys. **98**, 9963 (1993).

[107] R. W. Verhoef, D. Kelly, C. B. Mullins, and W. H. Weinberg, Surf. Sci. **325**, 93 (1995).

[108] P. M. Holmblad, J. Wambach, and I. Chorkendorff, J. Chem. Phys. **102**, 8255 (1995).

[109] H. Burghgraef, A. P. J. Jansen, and R. A. van Santen, Surf. Sci. **324**, 345 (1995).

[110] P. Kratzer, B. Hammer, and J. K. Nørskov, J. Chem. Phys. **105**, 5595 (1996).

[111] C.-T. Au, C.-F. Ng, and M.-S. Liao, J. Catal. **185**, 12 (1999).

[112] J. Harris, J. Simon, A. C. Luntz, C. B. Mullins, and C. T. Rettner, Phys. Rev. Lett. **67**, 652 (1991).

[113] I. H. V. A. Ukraintsev, J. Chem. Phys. **101**, 1564 (1994).

[114] A. C. Luntz, J. Chem. Phys. **102**, 8264 (1995).

[115] S. Dahl, A. Logadottir, R. C. Egeberg, J. H. Larsen, I. Chorkendorff, E. Tornqvist, and J. Nørskov, Phys. Rev. Lett. **83**, 1814 (1999).

[116] M. Mavrikakis, B. Hammer, and J. Nørskov, Phys. Rev. Lett. **81**, 2819 (1998).

[117] I. M. Ciobîca, F. Frechard, R. A. van Santen, A. W. Kleyn, and J. Hafner, J. Phys. Chem. B **104**, 3364 (2000).

[118] D. A. McCormack, G. J. Kroes, R. A. Olsen, E. J. Berends, and R. C. Mowrey, Phys. Rev. Lett. **82**, 1410 (1999).

[119] E. Zaremba and W. Kohn, Phys. Rev. **B13**, 1769 (1977).

[120] G. Vidali, G. Ihm, H. Y. Kim, and M. W. Cole, Surf. Sci. Rep. **12**, 133 (1991).

[121] N. R. Hill, M. Haller, and V. Celli, Chem. Phys. **73**, 363 (1982).

[122] H. Jónsson and J. H. Weare, J. Chem. Phys. **86**, 3711 (1987).

[123] R. Ahlrichs, R. Penco, and G. Scoles, Chem. Phys. **19**, 119 (1977).

[124] C. Hall and R. N. Perutz, Chem. Rev. **96**, 3125 (1996).

[125] K. Svensson, L. Bengtsson, J. Bellman, M. Hassel, M. Persson, and S. Andersson, Phys. Rev. Lett. **83**, 124 (1999).

[126] L. Bengtsson, K. Svensson, M. Hassel, J. Bellman, M. Persson, and S. Andersson, Phys. Rev. B **61**, 16921 (2000).

[127] D. Sholl, Langmuir **14**, 862 (1998).

[128] G. Henkelman and H. Jónsson, J. Chem. Phys. (submitted).

[129] D. Chandler, J. Chem. Phys. **68**, 2959 (1978).

[130] J. C. T. E. K. Grimmelman and E. Helfand, J. Chem. Phys. **74**, 5300 (1981).

[131] H. Grubmuller, Phys. Rev. E **52**, 2893 (1995).

[132] M. M. Steiner, P. A. Genilloud, and J. W. Wilkins, Phys. Rev. B **57**, 10236 (1998).

[133] S. Pal and K. A. Fichthorn, Chem. Eng. J. **74**, 77 (1999).

[134] J. Wang, S. Pal, and K. A. Fichthorn, Phys. Rev. B **63**, 5403 (2001).

[135] M. R. Sørensen and A. F. Voter, J. Chem. Phys. **112**, 9599 (2000).

[136] G. Henkelman and H. Jónsson, Mat. Res. Soc. Symp. Proc. **677**, AA8.1 (2001).

[137] R. Malek and N. Mousseau, Phys. Rev. E **62**, 7723 (2000).

[138] L. J. Munro and D. J. Wales, Phys. Rev. B **59**, 3969 (1999).

[139] A. B. Bortz, M. H. Kalos, and J. L. Lebowitz, J. Comput. Phys. **17**, 10 (1975).

[140] D. T. Gillespie, J. Comp. Phys. **22**, 403 (1976).

[141] D. T. Gillespie, J. Phys. **81**, 2340 (1977).

[142] D. T. Gillespie, J. Comp. Phys. **28**, 395 (1978).

[143] G. H. Gilmer, Science **208**, 335 (1980).

[144] A. F. Voter, Phys. Rev. B **34**, 6819 (1986).

[145] M. Villarba and H. Jónsson, Phys. Rev. B **49**, 2208 (1994).

[146] H. Jónsson, Annual Review of Physical Chemistry **51**, 623 (2000).

136

[147] G. Henkelman, B. P. Uberuaga, and H. Jónsson (unpublished).

[148] K. Tsemekhman, G. Henkelman, and H. Jónsson (unpublished).

VITA

Graeme Henkelman was born in Vancouver BC, Canada in 1974. He received a BSc in Physics from Queen's University in 1996, and a PhD in Chemistry from the University of Washington in 2001.