5-12-2016

Surface Chemistry with Machine Learning and Quantum Mechanics

Venkatesh Botu
University of Connecticut - Storrs, venkatesh.botu@uconn.edu

Surface Chemistry with Machine Learning and Quantum Mechanics

Venkatesh Botu, PhD

University of Connecticut, 2016

Surface chemistry is a phenomenon manifesting itself in several key areas; catalysis, materials fabrication, and emissions mitigation, to name a few. At the present time, atomistic computational driven efforts to study such processes are dominated by models based on quantum mechanics. Their flexibility in studying diverse chemistries, along with the ability to predict accurate thermodynamic and kinetic insights of surface processes, makes them increasingly popular. From ultra-low temperature and pressure to normal operating conditions these methods are now commonly utilized. Nevertheless, the computational burden inherent in the method renders it insufficient to keep up with the current need for quick discovery, i.e. predicting properties of millions of permutations of materials or the meticulous analysis of a chemical reaction on a material. Consequently, a push to go beyond traditional design and characterization practices to explain materials chemistry is becoming necessary.

In this thesis, a new framework that combines quantum mechanics with data-driven machine learning methods is put forth. The premise of such an approach is to mine and find patterns within data and in doing so come up with human fathomable relationships, to help accelerate discovery. Here, I focus on model development, which begins by generating data, identifying descriptors for a process, learning from the data and culminating with model validation. This then enables accelerated estimation of thermodynamic and kinetic properties of surface processes. Two detailed examples of this hybrid approach are discussed; (i) a guided and targeted catalyst design framework to identify optimal dopants to enhance thermochemical dissociation of H_2O, and (ii) a force predictive framework (commonly known as force field) to rapidly compute

forces on atoms, so as to extend dynamic simulations to length and time scales beyond current quantum mechanical methods.

Surface Chemistry with Machine Learning and Quantum Mechanics

By,

VENKATESH BOTU

B. Sc., Purdue University, 2009

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

at the

University of Connecticut

2016

APPROVAL PAGE

Doctor of Philosophy Dissertation

Surface Chemistry with Machine Learning and Quantum Mechanics

Presented by

Venkatesh Botu, B. Sc.

Major Advisor _____

Dr. Rampi Ramprasad

Associate Advisor _____

Dr. Ranjan Srivastava

Associate Advisor _____

Dr. Avinash M. Dongare

University of Connecticut

2016

To mum and dad .....

Acknowledgments

To my parents and Sravani for their continuous love and support.

To my graduate advisor, Dr. Rampi Ramprasad. For any graduate student, the advisor is instrumental in shaping their career. I am glad to have made the right choice. Dr. Ramprasad, epitomizes the meaning of ingenuity, ambitiousness, enthusiasm, dynamism and benevolence. In the past 5 years of working with him, I have had the privilege of being exposed to a glimpse of his methods. His out of the box thinking certainly paves way for a new realm of working in the scientific field, while never undermining the fundamental principles of science. One mantra that I learnt from him and never forget is "a good scientist always questions while others simply believe". During my time, he allowed me the freedom to work on projects of my choice while encouraging me to pursue internships, something that a lot faculty member do not. He is not only a great scientist, but an excellent manager. His inter-personal and planning skills are something I envy about and constantly strive to achieve. I am indebted for all you have done.

To my associate advisor Dr. Avinash Dongare. I thank you for the technical and career guidance we have had, and also the personal and cricket discussions.

To my previous advisor Dr. Ashish Mhadeshwar. I thank you for the career guidance and help in structuring the initial phase of my graduate career.

To my associate advisor, Dr. Ranjan Srivastava, and graduate panel Dr. Puxian Gao, Dr. Brian Willis, and Dr. Seok-Woo Lee.

To my previous group member Hong Zhu, for being my first mentor in graduate

school.

To the Office of Naval Research and National Science Foundation (NSF) for funding my research projects, and computational support from the NSF-Teragrid, Institute of Material Science, and School of Engineering allocation.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Surface chemistry overview

The interaction of gaseous (or liquid) matter with a solid surface has always been of great interest to society. A well known manifestation of this behavior being that of *rusting* of iron surfaces. Such heterogeneous phenomenon are not only ubiquitous in nature but are commonly found in the chemical and material sciences as well (c.f., Figure 1.1). For example, in the electro-chemical synthesis of H_2 fuel, or the fabrication of materials for micro-electronics, or the mitigation of harmful automobile exhaust emissions, etc., all of which have broad societal implications [1]. The sea of possible applications poses a significant challenge from a materials and chemistry perspective. In particular, the vast chemical expanse makes searching for and characterizing ideal material candidates or optimal chemical protocols extremely formidable. Consequently, a need for novel methods (both theoretical and experimental) that offer a more guided and rational explanation of materials behavior or chemical reactions, rather than serendipitously, are becoming increasingly necessary [2]. This is the focus of my thesis, mainly from a theoretical perspective.

The first step to building better surface chemistry models requires one to have a thorough understanding of the physics that governs such processes. To illustrate the key characteristics of a typical surface process I revisit the rusting of iron example.

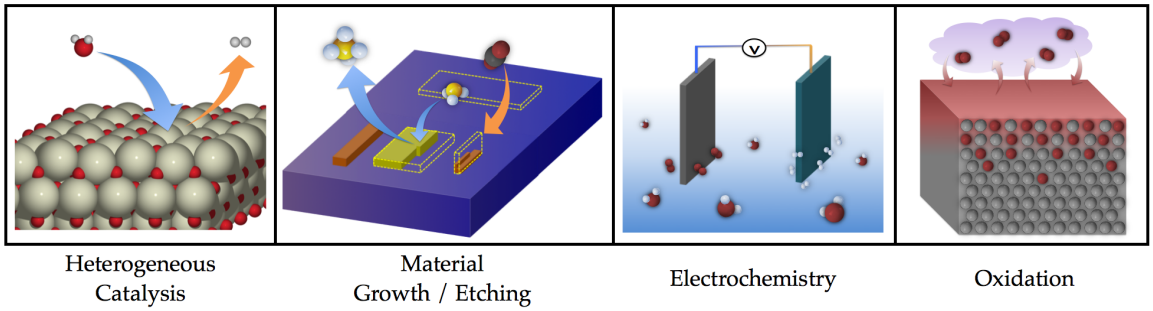

Figure 1.1: Examples of processes in the chemical and material sciences where surface chemistry plays a dominant role; catalysis, materials fabrication, electrochemistry, corrosion.

Here, the interaction of the metal surface with its surrounding gaseous environment results in a chemical transformation to a red-oxide material [3]. One can rewrite this phenomena (a simplified version of it) in terms of a chemical equation,

$$2Fe_{(s)} + \frac{3}{2}O_{2(g)} \longrightarrow Fe_2O_{3(s)} \quad \Delta E < 0, \tag{1.1}$$

where, Fe and Fe_2O_3 represent the metal and metal-oxide state at the surface, respectively. (i) *Why does the metal transform*, and (ii) *how long does it take*? In scientific parlance these questions are answered by understanding the *thermodynamics* and the *kinetics* governing the surface phenomenon [4]. In the case of rusting, at ambient conditions (room temperature and pressure) the oxide surface is lower in the energy scale ($\Delta E < 0$) than the clean metal surface, thereby, providing the necessary thermodynamic driving force for the transformation to occur. Nevertheless, the high kinetic barriers (E_a) underlying this mechanism results in a process that requires days or months to realize. In general, any surface chemistry process is comprised of several such transformations, similar to those given by Eq. 1.1, each with an associated ΔE and E_a. What is observed in reality is a complex interplay between the

thermodynamics and kinetics of the different processes. Therefore, a path to building better predictive models is by understanding the relevant thermodynamic and kinetic factors that govern such heterogeneous interfacial processes.

1.2 Role of computational methods

Towards this goal, *in silico* driven efforts, in particular those based on atomistic models, have been instrumental in helping unravel surface phenomena [5]. The atomic-level detail offered by such methods makes them a suitable choice to probe surface chemistry processes (amongst several others).

At the present time, a vast majority of the thermodynamic properties are modeled with the help of quantum mechanics (QM), either directly or indirectly [6, 7]. Such methods are commonly referred to as first principles or *ab initio* methods. At the core of this progress lies density functional theory (DFT) - a framework proposed by Kohn, Hohenberg, and Sham in the mid-1960s [8, 9]. In DFT, the well known multi-dimensional Schrödinger equation is reformulated into the Kohn-Sham equation, whereby one needs only solve a tractable 3-dimensional problem. A more extensive discussion on this topic is provided in Chapter 2. In short, DFT offered a practical prescription for computing the energy of a given configurations of atoms, which forms the basis for thermodynamic comparisons (i.e., ΔEs). Nevertheless, it was not until the early 90's when necessary approximations were introduced that allowed DFT to be more commonly used in the chemical, physical, and engineering communities [10]. Today, such *ab initio* methods have reached a level of sophistication that they offer a canvas for virtual surface science experiments, e.g. calculating binding energies of different species, identifying low energy surface terminations, finding the electronic properties, etc., - offering a capability to complement or sometimes

supplement traditional experiments.

Unfortunately, DFT is a 0 K based theory neglecting the contributions of finite temperatures and pressures, as is commonly encountered in nature. However, by merging *ab initio* methods, such as DFT, with statistical mechanics a new framework known as first principles thermodynamics (FPT) emerged. This offered a much needed "realistic" description (from a theoretical perspective) of equilibrium surface conditions at non-zero temperatures and pressures [6, 11, 12]. An example where such high-fidelity computational methods have been successful include identifying the role of different gaseous environments on observed surface phases (synonymous to the rusting problem) [13]. Similarly, in catalysis DFT methods have illuminated trends in reactivity, as well as selectivity, i.e. why some materials are more apt than others for a given chemical reaction [12]. This ability to meticulously explain and predict trends makes such methods immensely powerful, providing answers to the question "*why does it transform*?"

To answer the second question, "*how long does it take?*", one needs to delve deeper than just the thermodynamics. A surface is often not in a state of thermodynamic equilibrium, as portrayed by DFT or FPT, but is constantly evolving due to chemical reactions and/or migration of species on the surface, or between the surface and its environment. A capability to monitor this dynamic behavior is key to unraveling the rate at which different processes occur, thereby revealing the relevant kinetics. One way to handle this is by molecular dynamics (MD) [14]. In MD, the temporal-evolution of a configuration of atoms, in space, are integrated using classical Newton's equations of motion. To be able to do so a recipe to predict the force on an atom, as it interacts with its neighbors, is necessary and critical. Once again DFT comes to the rescue, providing an accurate prescription to compute atomic forces. This variant of MD is appropriately referred to as *ab initio* MD. Nevertheless, MD by itself is of

no purpose as it merely captures the distribution and motion of atoms. To correlate this microscopic information, such as atomic positions and velocities, to macroscopic observables, such as temperature, pressure, energy, heat capacities, etc., requires the mathematical expressions of statistical mechanics [15]. It is this combination that allowed scientists to explore a diverse class of problems, e.g. reconstruction of Si (111) surface, interaction of H_2O at silica interfaces, surface morphology under different growth conditions, etc [16, 17, 18].

Both the frameworks described thus far, be it to predict thermodynamic properties or explore the kinetic behavior, rely entirely on DFT, which has severe limitations at the present time. Firstly, owing to the enormous computational overhead involved with DFT predictions (such as energies), one is restricted to a handful of materials or chemical reactions to be analyzed meticulously. Developing alternative quicker predictive frameworks that build on DFT are necessary to enumerate through the vast permutations and combinations of materials. Secondly, once again owing to the computational overhead the length and time scales accessible by *ab initio* models to explore the relevant kinetic domains are severely restricted. For instance, the explicit dynamical evolution of surface processes, by *ab initio* MD, to timescales larger than a few picoseconds and length scales beyond a few angstrom are amongst the few challenges that cannot be handled. However, simulating surface chemical processes at these time- and length-scales is critical to understanding the non-equilibrium behavior of material surfaces during the course of a reaction.

In the past researchers have indeed come up with creative schemes to accelerate *ab initio* modeling of both the thermodynamic and kinetic aspect of materials. These broadly fall in two classes; (i) methods that speed up the evaluation of energy and forces in comparison to DFT, e.g. cluster expansion, tight-binding DFT or parameterized force-fields [19, 20, 21], and (ii) methods that coarsen simulations either

stochastically, by speeding up the clock or eliminating irrelevant degrees of freedom, e.g. Monte Carlo methods [22, 23], meta-dynamics [24, 25], temperature accelerated dynamics [26, 27, 28, 29] and hyperdynamics [27, 30, 28, 29]. These attempts though are not entirely satisfactory. Cluster expansion and tight-binding DFT remain computationally intensive for all practical purposes, while the transferability of force fields to situations not originally used in the parameterization is of constant inquiry. On the other hand coarsening a simulation requires some prior knowledge of the critical features encountered during the evolution process (and may involve artificial constraints and some loss of vital dynamical information), a task that is difficult to do a priori. These immense challenges prompted the development of an alternative data-driven approach, one that ties in more closely with the first class of methods, as shall be elaborated upon next.

1.3 An alternative: Machine learning

In todays world data is being generated and accumulated at an astronomical rate. This has led to a wave of sophisticated algorithms that makes use of the data to help build powerful predictive models to solve real-life problems, a field commonly known as machine learning (ML). Some inspiring examples include; autonomous cars, e-commerce shopping recommenders, predicting weather conditions or flu trends, personal voice assistants, amongst several others [31, 32, 33, 34, 35]. Recently, with Alphago [36]- an invention by Google's DeepMind team, we got a glimpse of ML's true potentiality when it competed directly with a human in the game of Go, and *won* convincingly. It comes as a surprise that with such sophisticated algorithms, the computers have not "taken over the world".

In the chemical and material sciences a similar accumulation of data has spurred

the use of ML methods in one of two ways; (i) to mine and discover hidden laws or rules buried in the data, and (ii) to develop more robust predictive models between some inputs and outputs. Such methods have indeed made significant inroads into various aspects of materials science [37, 38], e.g. an accelerated and accurate prediction of phase diagrams [39, 40], crystal structures [38, 41, 42, 43], mapping complex materials behavior to a set of process variables [44, 45, 46], analysis of high-throughput experiments [40, 47, 48], etc., have all been developed. Of particular relevance to the present contribution are recent successful efforts that exploit ML methods to develop models to predict material properties [49, 50, 51] and accurate force-fields (or inter-atomic potentials) [52, 53]. This paves room for optimism, whereby, one can harness the capability of ML methods to put in place models for surface chemistry applications.

In this thesis I demonstrate the role of ML capabilities to discover patterns in data and develop predictive models to help tackle the thermodynamic and kinetic modeling of surface chemical processes. My first contribution is a design framework for surface catalysis applications. The particular example targeted here explores the relationship between the surface composition of cerium oxides and its impact on surface reactivity, geared towards the thermochemical dissociation of H_2O. To do so, firstly one needs to understand the chemistry governing surface reactivity, which was studied using *ab initio* methods [54]. This allowed the identification of thermodynamic descriptors that correlated strongly with experimental observations of reactivity. Following which data-mining methods were then used to dig deeper into the attributes of the surface composition that led to better yields. By combining the two methods a rapid design framework to identify promising candidate materials based on purely a thermodynamic analysis, for the dissociation of H_2O is provided in a guided and rational manner, as we initially set out to do [55].

In the second contribution I provide a scheme that systematically *learns* in an interpolative manner to predict the forces on atoms quickly, by using a set of high-level calculations (based on DFT) as reference [56, 57]. With this high fidelity force field, several proof-of-concept atomistic simulations were explored, to validate that the proposed framework abides by the rules of statistical mechanics [58]. Once enough confidence was gained, simulations at large length and time scales, truly beyond existing methods, were used to study the surface ripening processes on an Al(111) surface [59]. Such simulations as shall be discussed are necessary to go beyond the current realm of *ab initio* methods to explore kinetically relevant domains of surface chemistry processes. Further, such a framework can be directly extended to other atomistic simulations, e.g. geometry optimization, identifying reaction barriers, predicting thermal properties, etc., amongst others.

The premise of both these frameworks remains the same; to harness data, particularly that computed by an accurate theory such as DFT, to discover rules or predict properties (e.g. energies or forces) much faster.

1.4 Thesis outline

Below is an outline describing the subsequent arrangement of chapters in this thesis.

In Chapter 2, I review the theoretical methods used in the present work, which include density functional theory, molecular dynamics, and machine learning. As the methods are well established and documented in literature, here I only provide sufficient background for reader to come to terms with the methods.

Chapter 3 contain a series of sections, put together from my publications. In this chapter, I demonstrate the development of a thermodynamic model for catalyst design. I start with the development of an *ab initio* generated surface phase

diagram for cerium oxides, and to identify suitable descriptors. Following which, a high-throughput screening framework exploring the role of dopants in cerium oxides is studied. This reference data along with machine learning methods are then used to identify key material properties of the dopant that governs reactivity. This offers a targeted solution to the discovery process, correlating surface composition to reactivity.

Chapter 4 also contain a series of sections, put together from my publications (past and some in progress). I start by discussing the construction, validation and estimation of uncertainty of machine learning force fields. This forms the basis for exploring kinetically relevant surface processes. Using one such developed model, I illustrate several simple examples of how such a framework can be used, e.g. geometry optimization, molecular dynamics, etc., followed by a long-time scale simulation that explores the ripening of adatoms on a surface. Lastly, a brief introduction into extending such a framework to multi-elemental systems is also discussed.

In Chapter 5, I provide a broad outlook on the promises and challenges facing materials design and discovery driven by machine learning.

Lastly, the Appendix contains additional model results, e.g. force fields developed for other elements.

Chapter 2

Methods

2.1 Density functional theory

The Schrödinger equation is known to be the master equation governing the interaction of matter at the quantum level, i.e. the behavior of nuclei and electrons that make up atoms [60]. Given that nuclei are much heavier than the electrons one can independently solve the equations describing the motion of electrons, whilst keeping the nuclei fixed, this is known as the Born-Oppenheimer approximation [61]. The equation (time-independent version) describing the behavior of electrons, with the nuclei fixed, is given by

$$\left[\frac{-\hbar^2}{2m}\sum_{i=1}^{N}\nabla_i^2+\sum_{i=1}^{N}V(\mathbf{r}_i)+\sum_{i=1}^{N}\sum_{j<1}U(\mathbf{r}_i,\mathbf{r}_j)\right]\psi(\mathbf{r}_1,.....,\mathbf{r}_N)=E\psi(\mathbf{r}_1,.....,\mathbf{r}_N). \quad (2.1)$$

The terms in the bracket denote the kinetic energy, interaction energy between the electrons and nuclei, and interaction energy between electrons. Here ψ describes the electronic wave function; related to the probability of finding an electron in space, and $\boldsymbol{r}_i$ is the electron coordinates in the Cartesian space. E is the ground state energy of the electrons. It is this quantity that we seek (more correctly, we seek $E_{tot} = E + E_{nuclei}$), and forms the basis for several thermodynamic comparisons in

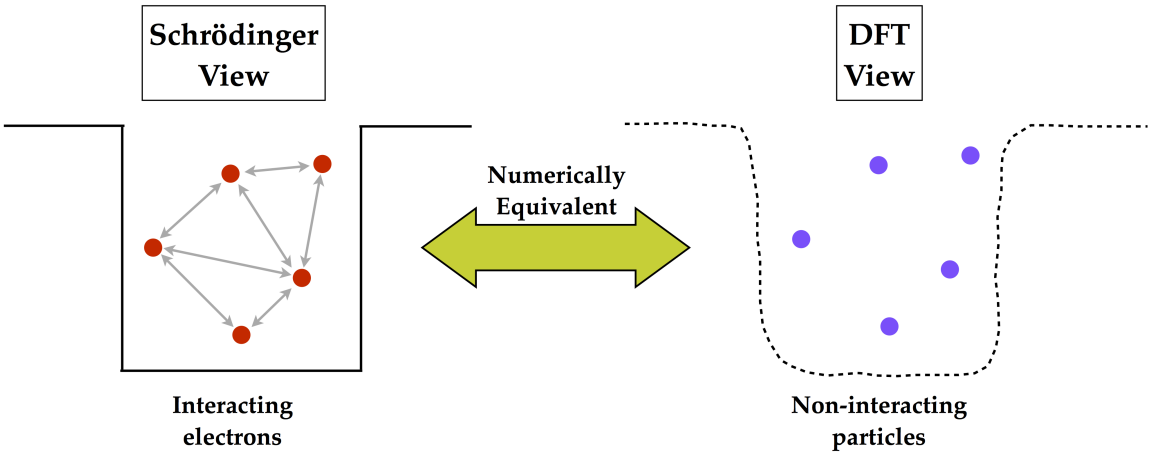

Figure 2.1: Comparison between Schrödinger's and the density functional theory (DFT) view point on the many-body electron problem. DFT transforms the many-body interacting problem to an equivalent non-interacting scenario, whereby the external potential is transformed to an effective potential, as indicated by the solid and dashed lines.

atomistic models. Unfortunately, Eq. 2.1 can only be solved exactly for the simplest of systems (at the most a 3-body problem). Even with approximate numerical methods, solving for ψ for several atoms is extremely challenging owing to its high dimensionality. For example in a H_2O molecule, the 10 e^- in 3-dimensional space results in a electronic wave function with 30 dimensions. Density functional theory offers a practical alternative by reformulating the Eq. 2.1 into one that is more tractable to solve [62, 63].

Density functional theory (DFT) is based upon two fundamental theorems, proposed by Hohenberg and Kohn [8]:

Theorem 1 *The ground state energy from Schrödingers equation is a unique functional of the e^- density.*

Theorem 2 *The e^- density that minimizes the energy of the overall functional is the true e^- density corresponding to the full solution of the Schrödingers equation*

However, the true leap came about when Kohn and Sham showed that the many-body electron problem, in the presence of the nuclei, can be solved self-consistently in terms of a set of non-interacting particles in an effective potential [9], as illustrated in Figure 2.1. This led to the Kohn-Sham one electron equation,

$$\left[\frac{-\hbar^2}{2m}\nabla^2 + V(\mathbf{r}) + V_H(\mathbf{r}) + V_{XC}(\mathbf{r})\right]\phi_i(\mathbf{r}) = \varepsilon_i\phi_i(\mathbf{r}). \tag{2.2}$$

Here, the terms in the bracket represent; kinetic energy of an electron, the interaction potential of an electron with surrounding nuclei, interaction fo the electron with surrounding electrons, and the exchange-correlation potential. The last term (V_{XC}) compiles the missing interactions upon transforming a many-body electron problem to a non-interacting single electron problem. Over the past decades, there have been several attempts to approximate the exchange-change correlation, and research on this subject still remains to be active. Upon comparing Eq. 2.1 with Eq. 2.2, the formidable all electron wave function, $\psi(\mathbf{r}_1, ..., \mathbf{r}_N)$, is replaced by one that is a function of only 3 variables, $\phi_i(\mathbf{r})$. This offered a means to compute the energy for a given configuration of atoms quickly, allowing for several thermodynamic properties of matter to be computed. For instance, stable or meta-stable phases of solids and alloys, equilibrium geometry, vibrational and phonon frequencies, elastic constants, bulk modulus, etc., are all properties that depend on this energy, which in turn is a function of the atomic positions.

Another quantity of interest is the force ($\boldsymbol{F}_i$) an atom experiences, as this can be directly used to study dynamic simulations as discussed in the next section, and is given by

$$-\boldsymbol{F}_i = \frac{\partial E}{\partial \boldsymbol{R}_i}. \tag{2.3}$$

Here, $\boldsymbol{R}_i$ represents the position of the nuclei. A non-sophisticated and inefficient way to access this derivative is by using a numerical finite difference approach. The *Hellmann-Feynman* theorem [64] on the other hand provides a quick, accurate and more direct means to access these forces - it is has now become the primary workhorse in modern atomistic methods. The beauty of this theorem is the ability to compute the force on nuclei using nothing but the nuclei Hamiltonian once an energy calculation is complete, as demonstrated below. Let us consider H and φ to be the nuclei Hamiltonian and wavefunction. Then it follows from Eq. 2.3 that,

$$-F_i = \frac{\partial \langle \varphi_i^* | H | \varphi_i \rangle}{\partial \boldsymbol{R}_i} = \langle \varphi_i^* | \frac{\partial H}{\partial \boldsymbol{R}_i} | \varphi_i \rangle + E \frac{\partial \langle \varphi_i^* | \varphi_i \rangle}{\partial \boldsymbol{R}_i} \tag{2.4}$$

Upon realizing that the last term in Eq. 2.4 is equal to 0, as the wavefunctions are normalized to 1, the force on a nucleus is then simply the sum of the classical electrostatic interaction between nuclei, and the interaction between nuclei and electron charge density. These are the only terms that depend on $\boldsymbol{R}_i$ in H.

2.2 Molecular dynamics

Molecular dynamics, as the name suggests pertains to exploring the dynamic evolution of molecular structures - composed of atoms [14]. A simple way to study this temporal landscape is by solving Newton's equations of motion for an atom, a subject taught in primary school. The two governing equations are

$$\ddot{\mathbf{r}}_i = \mathbf{a}_i, \quad \text{and} \tag{2.5}$$

$$\mathbf{a}_i = \frac{\mathbf{F}_i}{m_i}. \tag{2.6}$$

Where, $\ddot{\mathbf{r}}_i$, $\mathbf{a}_i$, $\mathbf{F}_i$ and m_i, are the second derivative of position with time, vectorial

acceleration and forces, and the mass of a given atom. To solve these coupled differential equations, an initial set of atomic positions and velocities need to be specified as input. This essentially governs how the configuration of atoms evolve, also known as an exploration of the phase space. A common approach for phase space exploration is to use numerical methods, wherein, the time dimension is discretized. This offers a practical alternative, whereby one can numerically integrate the equations of motion between timesteps (Figure 2.2), so long as the true trajectory is preserved, i.e., sympletic [65]. The choice of the timestep is purely governed by the underlying physics that is explored. A typical time-step is on the order of femtoseconds, as atomic vibrations occur at this timescale. A well known time-integration algorithm for this purpose is the *velocity-Verlet* method [15]. It allows us to solve the stiff coupled differential equations, is time reversible, and requires only one force evaluation for every timestep. This is the algorithm used here as well.

Though we have laid out the theory for MD simulations, we still lack the necessary tools to extract atomic level information and convert them to macroscopic observables. A direct MD simulation at the macroscopic scale is computationally implausible given the sheer number of atoms, e.g. in 1 cm^3 Cu there are $\approx 10^{22}$ atoms. For these reasons, we require mathematical laws that allow us to connect between the microscopic and macroscopic states. Statistical mechanics offers a solution, whereby, several microscopic simulations are averaged to recover the true macroscopic behavior. This is commonly referred to as *ensemble* modeling. The mathematical equation that describes and allows such an averaging is given by

$$\langle A \rangle_{ensemble} = \iint A(\boldsymbol{r}, \boldsymbol{p}).P(\boldsymbol{r}, \boldsymbol{p}).d\boldsymbol{r}d\boldsymbol{p}. \tag{2.7}$$

Here, $\boldsymbol{r}$ and $\boldsymbol{p}$ describe the phase space, i.e., any combination of position and mo-

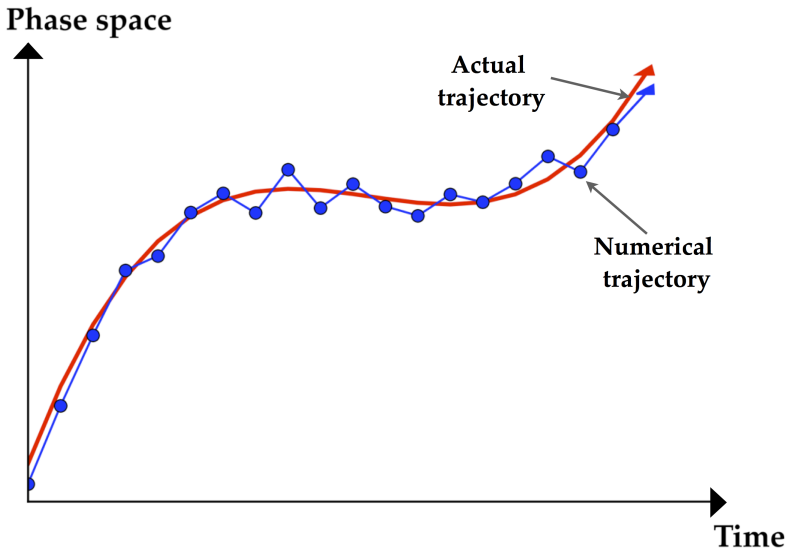

Figure 2.2: An actual phase space trajectory (red solid line) and an approximate numerical trajectory (blue solid line with markers) computed by a time integration algorithm. The blue circle markers indicate the discretization in the time domain, with the distance between two consecutive points the chosen timesetp.

mentum for a given set of atoms. $A(\boldsymbol{r}, \boldsymbol{p})$ and $P(\boldsymbol{r}, \boldsymbol{p})$ are the value of the interested property and the probability of existing at a given phase point, $(\boldsymbol{r}, \boldsymbol{p})$ [65]. Further, $P(\boldsymbol{r}, \boldsymbol{p})$, from statistical mechanics, can be deduced by knowing the energy associated with the given phase point,

$$P(\boldsymbol{r}, \boldsymbol{p}) = Q^{-1} e^{-E(\boldsymbol{r}, \boldsymbol{p})/k_B T}. \tag{2.8}$$

The challenge lies in determining Q the system partition function, as one needs to integrate across the entire phase space, which is extremely vast. MD simulations can be used to sequentially cover this phase space, however, this would require infinitely long simulations. A more practical alternative is by applying the *Ergodic* hypothesis [66], which states that an ensemble average is equivalent to a time average for a given property, as described below

$$\langle A \rangle_{ensemble} = \langle A \rangle_{time} = \lim_{\tau \to \infty} \frac{1}{\tau} \int_0^\tau A\left(\boldsymbol{r}(t), \boldsymbol{p}(t)\right) dt. \tag{2.9}$$

It is this relation that allows us to compute a spectrum of properties, e.g. temperature, pressure, potential energy, etc., for atomistic MD simulations.

The framework laid above is applicable only in the micro-canonical (NVE) ensemble, i.e. constant volume, energy, and number of atoms. Often one would like to study constant temperature or pressure simulations, as these mimic reality better. To introduce these ensembles into an MD simulation requires reformulation of the equation of motions. For example, to achieve a constant temperature simulation the microscopic system needs to exchange heat with its surrounding to achieve thermal equilibrium. A way to do this is by coupling it with a heat bath. One such thermostat regulation was introduced by Nosé-Hoover, on which several external discussions exist [67]. In a nutshell, it adds a frictional component that regulates the velocities of the atoms, based on a feedback-control between the system temperature versus the desired set point. Similar such feedback-control implementations exist for maintaining a constant pressure as well. It is this versatility of the MD framework that allows us to study a host of atomistic simulations, pushing atomistic models closer to reality.

2.3 Machine learning

Machine learning (ML) is the art of extracting insights or knowledge from raw data [68]. However, unlike traditional statistical methods, the models are often iteratively refined and adjusted with time as more data becomes available. Initially, a field brought about by marrying methods in computer science, engineering, and statistics, it is only recently that such capabilities are being exploited within the materials and

chemical sciences.

Generally, in the chemical and material sciences we tend to collect data on observations. Often this contains records of some inputs, with its attributes $\boldsymbol{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & .. & x_{1,p} \\ x_{2,1} & x_{2,2} & .. & x_{2,p} \\ x_{n,1} & x_{n,2} & .. & x_{n,p} \end{bmatrix}$ (also known as predictors or features), describing the process and their corresponding properties, $\boldsymbol{y} = [y_1, y_2, ..., y_n]$. Here, $\boldsymbol{y}$ can either be a continuous or discrete variable. n and p are the total number of observations and the dimensionality of the input attributes, respectively. From this data, there are often three kinds of problems that one looks to solve;

1. Establish a predictive model, by mapping $\boldsymbol{X}$ and $\boldsymbol{y}$ without the need to understand the underlying relation.

2. Infer the behavior or response in changing any of the p attributes ($\boldsymbol{X}_i$) on the corresponding property (y_i). This provides a more structured view of the data, rather than treating it as a *black box*.

3. Lastly, the data collected could be unstructured or unlabeled, i.e. for each observation, we only have access to $\boldsymbol{X}_i$ but lack the property value y_i. The goal is to extract meaningful insights or patterns by learning hidden relationships within the data.

Except for a few cases, these challenges often lie beyond human cognition, owing to the scale of the data as well as the complex interplay between the different attributes. It is under such situations that ML algorithms come to their element. The premise of such methods is to use data to drive discovery.

The general process of constructing a ML model follows the progression of; (i) acquiring data that is accurate and curated, (ii) representing the data in a machine fathomable manner, (iii) choosing a learning algorithm, and (iv) validation and verification of the developed model itself, as illustrated in Figure 2.3, irrespective of the

problem class. Often one needs to worry about the accuracy of the reference data but given that we use DFT as the source, this noise is minimal. For this reason, a more specific challenge is how to select amongst the data, and come up with relevant descriptors. In the chemical and material sciences, the particular choice of the de-

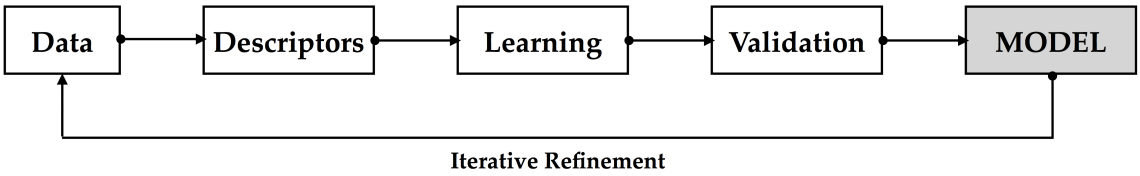

Figure 2.3: Key steps in constructing ML models; acquiring data, developing descriptors, choosing a learning algorithm, and model validation (and verification).

scriptors (or fingerprint) is generally driven by the underlying relation that is being explored. The set of descriptors used can vary from macroscopic/global properties all the way to the atomic-level, such as the coordination environment of an atom. For instance, in a model used to predict the macroscopic friction coefficient of ceramic materials, the representation comprised of high-level material properties such as the melting point, density, etc [44, 45]. Similarly, in another model developed to predict the electronic properties of polymers, a chemical motif-based representation that encodes information about groups of atoms, was used [50]. The learning algorithm is another important ingredient in developing accurate ML models. Depending upon the class of problem that one intend to solve the learning algorithms generally fall under two classes; (i) supervised learning [31, 69, 70], e.g., linear and non-linear regression, neural networks, random forests, classification, etc., and (ii) unsupervised learning [31], e.g. cluster analysis, principal component analysis, feature selection, etc. The former used when the data being studied consists of both the attributes and properties, while the later used when only data attributes are available.

Here, we make use of a combination of the methods discussed to help develop models that allow for quicker prediction of thermodynamic and kinetic properties in surface chemistry applications. The specific details entailing model construction are avoided here, and are discussed extensively in the respective chapters (Ch. 3 and 4) next.

Chapter 3

Modeling thermodynamic behavior

3.1 Introduction

Oxides of cerium (CeO_2 and Ce_2O_3) possess a peculiar characteristic of being able to regulate their oxygen content based on the operating environment [71]. This "buffering" capability makes them suitable for redox reactions, e.g., during the treatment of automotive emissions in fuel-rich or -lean environments. Another application of cerium oxides is in high temperature gas sensors, water-gas shift reaction, and thermochemical water splitters that rely on alternating oxidized and reduced states of cerium oxides [72, 73]. A governing factor in all these applications is the role of oxygen within the ceria lattice or at the surface. Therefore, developing a fundamental understanding of the oxygen interaction with ceria is of paramount importance in discerning its chemical traits. Knowing the surface/lattice chemistry allows one to rapidly tailor materials, say with dopants, based on the desired characteristics (e.g., reactivity, selectivity, and stability, to name a few). This has been exploited in many situations ranging from material strengthening to electronics to electrochemistry. However, the search and identification of suitable dopant candidates has been laborious though, and dominated either by lengthy trial-and-error strategies (guided by intuition) or plain serendipity.

We are entering an era where such Edisonian approaches are gradually being

augmented (and sometimes, replaced) by rational strategies based on advanced computational screening [74]. Often these strategies rely on first principles methods, that provide a reasonably accurate description of the underlying chemistry [75, 76, 77]. More recently, it has been shown that supplementing first principles investigations with data-driven approaches can help identify meaningful correlations within the data [37, 39, 38, 41, 42, 44, 45, 46, 78]. In the next few sections, the steps to building such data-driven frameworks is laid out. Firstly, a purely first-principles derived surface phase diagram is constructed, to better understand the thermodynamic driving forces that governs surface catalysis. This provides knowledge into the key descriptors that govern surface reactivity for undoped ceria. Following which, a screening framework, inspired by Sabatier's principle, is used to tailor the surface properties with a host dopants spanning the alkali and transition series metals. Data-driven approaches are then implemented to mine and correlate the between the descriptor and the surface chemistry, as it is modified by dopants. Lastly, a framework to quickly predict these thermodynamic descriptors, without resorting to expensive DFT calculations, is provided offering a framework that can be used to enhance the thermochemical splitting of water, or other reaction mechanisms.

3.2 Ceria in an oxygen environment: Surface phase equilibria and its descriptors

Regulating the O stoichiometry in ceria offers a promising prospect in catalysis given the existence of multivalent cationic states. In reactions involving a net transfer of O such as thermochemical dissocation of H_2O, ceria is believed to play an intermediary role of providing redox sites for the reactants [79, 80]. The current understanding of

the reaction mechanisms involving ceria is based upon experimental results and first principles calculations. Ample experimental data characterizing O non-stoichiometry of ceria exists, primarily from a few decades ago. In the 1980's, Bevan and Kordis equilibrated ceria in a mixture of either H_2O/H_2 or CO_2/CO and measured the partial pressures of the corresponding gases to determine the oxygen non-stoichiometry at 10^{-8} - 10^{-32} atm and 909–1443 K [81]. Panlener *et al.* performed similar non-stoichiometry studies using thermo-gravimetric measurements over 1023–1773 K and 10^{-2} - 10^{-26} atm oxygen partial pressure [82]. Additional experiments (mass spectrometry, effusion measurements, high temperature X-ray diffraction, thermal expansion measurements, and specific heat measurements) studying the non-stoichiometry of ceria, along with those discussed above, indicate that high temperatures and low oxygen partial pressures are required to observe any appreciable reduction of the undoped bulk CeO_2 or any appreciable changes in its equilibrium fluorite crystal structure (Figure 3.1a) [83, 84, 85, 86]. Within the context of catalysis, non-stoichiometry at the ceria surface introduced by point defects such as O vacancies is crucial for the creation of active reaction sites. These vacancies can be formed either in the surface or sub-surface O layer (Figure 3.1b), as elaborated in the work by Torbrugge and Reichling [87]. Using atomic and dynamic force microscopy, they revealed that the preferred position of the O vacancies is the sub-surface region for ceria annealed at $\sim$1200 K and 10^{-12} atm.

The availability of such experimental information on the stoichiometry of ceria and the importance of vacancies in processes involving ceria have provided the motivation and testing ground for parallel and complementary first principles computational studies. While extensive first principles density functional theory (DFT) studies exist for bulk ceria, there is limited understanding on the transitions between stoichiometric and non-stoichiometric surface phases. Past computational studies on ceria have led

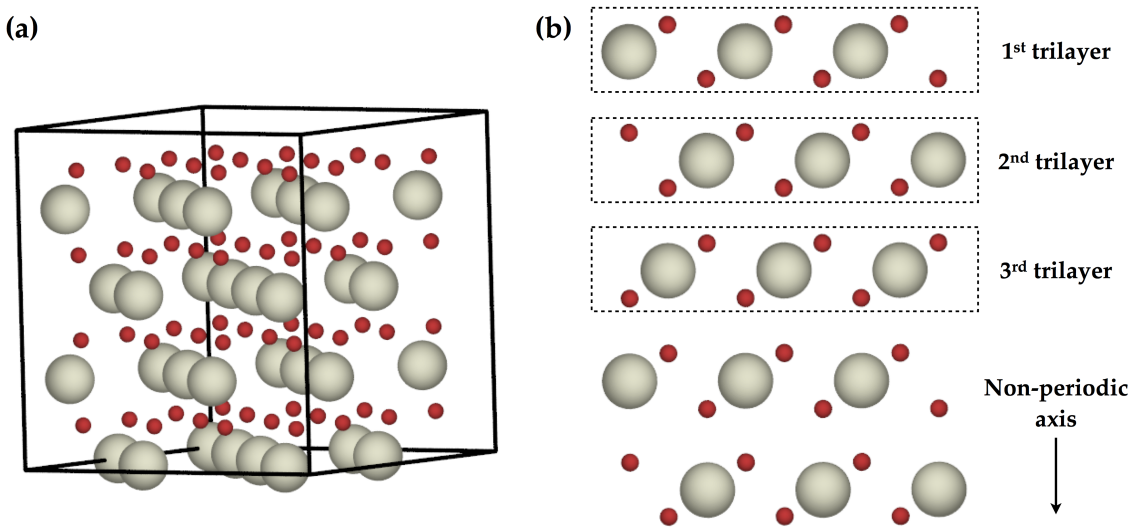

Figure 3.1: (a) Fluorite structure of ceria, and (b) a 5-trilayer slab model of the ceria surface. Grey atoms represent Ce and red atoms represent O. The dotted lined box represents one trilayer consisting of Ce and the top and bottom O atomic layers.

to a deeper understanding of its electronic and structural properties, stable surface orientation, and the role of oxygen diffusion and vacancies within the bulk lattice. These include identification of the O vacancy as a crucial factor in determining the activity of ceria in chemical reactions. It is suggested that O vacancies are preferably formed in the sub-surface layer rather than the surface layer (or within the bulk), which is consistent with the aforementioned experiments. Recent work by Janik and colleagues further indicates the significance of O vacancies in activating a ceria surface for hydrocarbon conversion [88]. Vacancies reduce the neighboring Ce atoms making them active for dissociation of gas phase molecules. Fronzi *et al.* studied the surface behavior of ceria in the presence of a water and oxygen environment [89]. They found that water remains adsorbed on the surface even under extremely low water partial pressures ($\sim 10^{-10}$ atm), at a temperature of 300 K and oxygen pressure of 1 atm. Under these conditions, reduction in the oxygen pressure induces O vacancies and

water dissociation into OH and H. Thus, O vacancies increase the surface reactivity of ceria for water dissociation.

Even though significant work has been conducted on ceria using first principles calculations, ceria still poses technical difficulties due to localization of electrons in the 4f orbitals leading to increased columbic repulsion as it undergoes reduction. These effects are not captured by conventional DFT that uses local or semi-local electronic exchange-correlation functionals [90, 91]. The general practice of modeling cerium oxides has evolved from the traditional functionals to functionals modified using the Hubbard parameter (U) [92, 93] and to hybrid functionals [94]. Furthermore, while most of the zero-temperature DFT studies have provided substantial understanding of the role of oxygen chemistry, a more quantitative connection with the available experimental data for ceria stoichiometry requires a treatment that includes non-zero temperatures and pressures. First principles thermodynamics (FPT) combines zero-temperature DFT results with statistical thermodynamics concepts, and offers a reliable, practical, and powerful prescription to address such factors [95, 96, 97].

3.2.1 First principles modeling

All first principles DFT calculations were performed using the plane-wave based Vienna Ab-Initio Simulation Package (VASP) [98]. Projector augmented wave (PAW) frozen core potentials with the O 2s, 2p, and the Ce 5s, 5p, 4f, 5d, 6s states treated as the valence states were employed [99, 100]. A 400 eV plane-wave cut-off energy was necessary to ensure converged results. The quantum mechanical part of the electron-electron interactions was represented using the Perdew-Burke-Ernzerhof (PBE) exchange correlation functional, and its Hubbard modified extension (PBE+U, $U = 5$ eV) along with the Heyd-Scuseria-Ernzerhof (HSE06) hybrid functional [10, 101, 102,

103]. While the PBE functional is a widely used semi-local functional, the HSE06 functional incorporates a certain amount of screened range-separated nonlocal exchange interaction, known to improve various properties including thermochemistry (i.e., energetics) and the electronic structure (e.g., the band gap of insulators) [94]. The Hubbard modified PBE functional, on the other hand, accounts for the increased Coulombic repulsion due to electron localization on a reduced surface. As the HSE06 functional requires significantly more computational time relative to the PBE functional, all geometry optimization calculations were performed using the PBE functional, followed by the evaluation of just the energies using the HSE06 functional at the PBE-optimized geometry. For completeness, geometry and electronic optimization was carried out with PBE+U functional; and it was included here primarily to compare and contrast the three levels of theory. A convergence criterion of 10^{-4} eV and 10^{-3} eV between consecutive electronic and ionic iterations was adopted. A 0.1 eV Gaussian smearing width was used to treat the band occupancies close to the Fermi level. Spin polarized calculations assured correct treatment of the magnetic components of the ceria system, particularly in treating the atomic O, molecular O_2, and reduced Ce atom states.

The bulk ceria fluorite structure in Figure 3.1a was optimized using a Γ-centered 6×6×6 k-point mesh and the PBE functional, yielding a lattice parameter of 5.47 Å, which is reasonably close to the experimental value of 5.41 Å [104]. Similarly optimization based on PBE+U functional yields a lattice parameter of 5.49 Å for the bulk structure. Using the PBE optimized bulk lattice parameters, a 2×2 ceria slab was created to model the surface with a total of 15 atomic layers (Figure 3.1b) and a vacuum of 15 Å between periodic images to minimize the finite size interactions. As shown in Figure 3.1b, stacking along the direction can be represented in terms of [O-Ce-O] trilayers, with equal number of atoms in each layer of a trilayer for

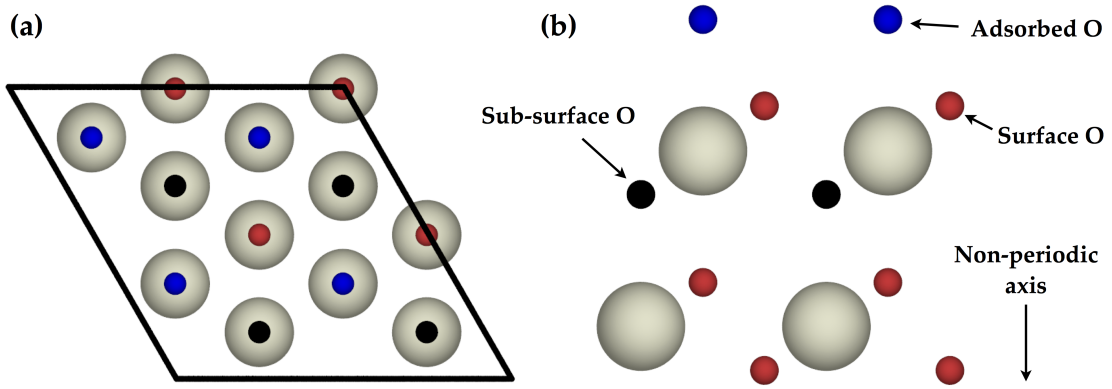

Figure 3.2: (a) Top view of the ceria 2×2 supercell, and (b) different O atoms in the slab model; red atoms - surface O, black atoms - sub-surface O, and blue atoms - adsorbed O. The grey atoms represent Ce in the 1^{st} and 2^{nd} trilayers, respectively.

a stoichiometric CeO_2 system. Therefore, the 15 atomic layers consist of 5 such trilayers. For the 1^{st} trilayer, we refer to the upper and lower O layers as the surface and sub-surface layers, respectively. The k-point grid was reduced to 3×3×1 for the slab supercell, and the middle 3 trilayers were fixed to create a bulk-like region, yielding a symmetric surface model. Figure 3.2 shows the top and the side views of the supercell used to model the crystal surface (only the top two trilayers of the surface are shown for clarity). The grey spheres represent Ce atoms in the first and second trilayers, respectively. The red, black, and blue atoms represent the surface, sub-surface, and adsorbed (adatom) O atoms, whose concentrations are denoted as α, β, and θ, respectively. O concentration in a given layer is defined as the ratio of the number of O atoms in that layer to the number of Ce atoms per trilayer. In the present study, α, β, and θ, were varied systematically and independently between 0 and 1 in increments of 0.25 (noting that an intact layer without vacancies contains 4 O atoms at the surface and sub-surface layers). This encompasses more than 20 configurations, ranging from an O depleted surface or sub-surface to a completely O adatom saturated surface.

We systematically start from a stoichiometric CeO_2 slab without any O adatoms ($\alpha = \beta = 1$, $\theta = 0$) and successively remove the surface or sub-surface O atoms from the 1^{st} trilayer until the surface or sub-surface is devoid of O. These conformations simulate the surface transition of ceria as it undergoes reduction in a highly oxygen-lean/reducing environment. In addition to removing O atoms from the surface *or* sub-surface layers, several additional cases were considered in which surface *and* sub-surface O atoms were removed simultaneously. These included situations leading to a reduced Ce_2O_3 stoichiometry in the surface trilayer ($\alpha = \beta = 0.75$, $\theta = 0$), or to the surface trilayer completely devoid of O ($\alpha = \beta = \theta = 0$) which may occur under extremely reducing conditions. On the other hand, to model O adsorption on ceria in oxygen-rich/oxidizing conditions, θ was varied from 0 to 1 monolayers (ML) while maintaining the ceria stoichiometry ($\alpha = \beta = 1$). For a given concentration of the O adatom(s), three different adsorption sites, viz., top, bridge, and hollow (3-fold), were considered. To prevent ambiguity, we use the general notation $CeO_x^{\alpha,\beta}$ + θ ML to represent the surface trilayer region including O adatoms. The trilayer stoichiometry is captured by x ($= \alpha + \beta$). For simplicity and to eliminate some redundancy, when $\theta = 0$, we represent the surface region by just $CeO_x^{\alpha,\beta}$; and when $\theta > 0$, we use the notation CeO_2+ θ ML (as $\alpha = \beta = 1$ when $\theta > 0$).

FPT has been extensively discussed in the past [95, 96, 97]; here we briefly remark on some key concepts only. The free energy for the formation of a certain concentration of O vacancies or adatoms ($G^f_{CeO_x^{\alpha,\beta}+\theta}$) is defined as

$$G^f_{CeO^{\alpha,\beta}{}_x+\theta} = \frac{E_{CeO^{\alpha,\beta}{}_x+\theta} - E_{CeO_2} - 4(\alpha+\beta+\theta-2)\mu_O}{|4(\alpha+\beta+\theta-2)|}. \tag{3.1}$$

Here, $E_{CeO_x^{\alpha,\beta}+\theta}$ and E_{CeO_2} represent the 0 K DFT energy of the non-stoichiometric slab and the stoichiometric ceria slab ($\alpha = \beta = 1$, $\theta = 0$), respectively; and $4(\alpha + \beta$

+ θ - 2) represents the net number of O atoms that have been added to or removed from the stoichiometric slab (while noting that there are 4 O atoms per layer in the 2×2 slab). μ_O is the oxygen chemical potential of the reservoir, which can be written in its temperature (T) and pressure (P) dependent form as

$$\mu_O(T, P_{O_2}) = \frac{1}{2}\left[E_{O_2} + \Delta\mu_{O_2}(T, P_{O_2})\right], \tag{3.2}$$

where,

$$\Delta\mu_{O_2}(T, P_{O_2}) = \Delta\mu'_{O_2}(T, P^o) + kTln\left(\frac{P_{O_2}}{P^o}\right). \tag{3.3}$$

E_{O_2} in Eq. 3.2 represents the 0 K energy of an isolated O_2 molecule including the zero-point harmonic vibrational energy. $\Delta\mu_{O_2}$contains the temperature and pressure dependent components, and can be separated into a purely temperature dependent part ($\Delta\mu'_{O_2}$) that can be determined using the JANAF thermochemical tables [105] or statistical thermodynamics [106]. $kTln\left(\frac{P_{O_2}}{P^o}\right)$ accounts for the pressure dependence with a chosen reference pressure of P^o (1 atm). In deriving Eq. 3.1, the entropic and pressure-volume contributions to the free energy of condensed phases were neglected given that these effects would likely cancel out between the non-stoichiometric and stoichiometric slabs, as has been verified before by Zhu *et al* [106].

Similarly, the surface energy corresponding to a slab with a certain concentration of O vacancies or adatoms, relative to the stoichiometric slab, is given by

$$\Delta\gamma_{CeO_x^{\alpha,\beta}+\theta} = \gamma_{CeO_x^{\alpha,\beta}+\theta} - \gamma_{CeO_2} = \frac{E_{CeO_x^{\alpha,\beta}+\theta} - E_{CeO_2} - 4(\alpha+\beta+\theta-2)\mu_O}{2\sigma}. \tag{3.4}$$

Here, $\Delta\gamma_{CeO_x^{\alpha,\beta}+\theta}$ is the relative surface energy; $\gamma_{CeO_x^{\alpha,\beta}+\theta}$ and $\gamma_{CeO_2}$are the surface energies of the non-stoichiometric and the stoichiometric slabs, respectively; and σ is the exposed surface area. The factor of 2 in the denominator accounts for the fact that our slab contains identical top and bottom surfaces. The relative surface energy is used to identify the stable phases observed for a range of oxygen potentials. The driving force for O exchange is governed by the difference in the chemical potential between the solid and its environment. Low μ_O indicates an oxygen-lean environment in which O from the solid would likely desorb and create vacancies. Similarly, high μ_O indicates an oxygen-rich environment in which O adatoms would adsorb on the stoichiometric surface. Thus, μ_O is bound by two limits: a critical maximum when molecular O_2 condensation occurs on the surface and a critical minimum when μ_O is low enough to promote decomposition of ceria into its constituent components (Ce metal and O_2 gas).

The aforementioned equations apply to a ceria surface in equilibrium with an explicit O_2 reservoir. However, O transfer to and from the surface can also indirectly be facilitated by the presence of other redox environments (e.g., NO/NO_2, H_2/H_2O, and CO/CO_2). In such cases, μ_O is governed by the ratio of the gas concentrations or pressures, i.e., $\frac{P_{NO_2}}{P_{NO}}$, $\frac{P_{H_2O}}{P_{H_2}}$, or $\frac{P_{CO_2}}{P_{CO}}$, based on the following relations:

$$\mu_O = \mu_{NO_2} - \mu_{NO} \text{ or } \mu_O = \mu_{H_2O} - \mu_{H_2} \text{or } \mu_O = \mu_{CO_2} - \mu_{CO}, \tag{3.5}$$

$$\mu_O(T, P_{NO_2}, P_{NO}) = \left[E_{NO_2} - E_{NO} + \Delta\mu'_{NO_2}(T, P^o) - \Delta\mu'_{NO}(T, P^o) + kTln\left(\frac{P_{NO_2}}{P_{NO}}\right)\right], \tag{3.6}$$

$$\mu_O(T, P_{H_2O}, P_{H_2})$$
$$= \left[E_{H_2O} - E_{H_2} + \Delta\mu'_{H_2O}(T, P^o) - \Delta\mu'_{H_2}(T, P^o) + kTln\left(\frac{P_{H_2O}}{P_{H_2}}\right)\right], \tag{3.7}$$

and

$$\mu_O(T, P_{CO_2}, P_{CO})$$
$$= \left[E_{CO_2} - E_{CO} + \Delta\mu'_{CO_2}(T, P^o) - \Delta\mu'_{CO}(T, P^o) + kTln\left(\frac{P_{CO_2}}{P_{CO}}\right)\right]. \tag{3.8}$$

Here, E_i represents 0 K energy, whereas the temperature dependent $\Delta\mu'_i$ terms for the different gases ($i = NO,\ NO_2,\ H_2,\ H_2O,\ CO,\ CO_2$) are obtained from the JANAF thermochemical tables. The other terms are similar to those defined in Eqs. 3.2 and 3.3.

All the above mentioned relations apply to a thermodynamically governed system. Thus the stability of a surface is based on minimizing its free energy. The surface energy relation along with the expression for the oxygen potential allows for a one-to-one mapping between the stable surface phases and the operating variables (temperature and pressure), thereby allowing for the creation of the ceria surface phase diagram.

3.2.2 0 K Energy of oxygen adatom(s) binding and vacancy formation

As alluded to earlier, the transition point when a surface readily switches between a stoichiometric and non-stoichiometric state is particularly important in understanding the potential catalytic activity of a particular material. In this section, in an attempt to understand such transitions, the dilute limit of O adatom binding or O vacancy formation in the stoichiometric ceria surface is explored. Table 3.1 shows the calculated 0 K formation energies for O adatoms on the surface, obtained by setting $\Delta\mu_{O_2}(T, P)$ to zero in Eq. 3.2. At $\theta = 0.25$ ML, which corresponds to a single O adatom on the 2×2 stoichiometric ceria slab, this simulates a dilute O adatom coverage on the ceria surface. For this dilute coverage scenario, the binding energies with respect to molecular O_2 at the 3 different adsorption sites, viz., top, hollow, and bridge, are 2.24, 1.52, and 0.34 eV, respectively, at the PBE level of theory (the corresponding values are -1.08, -1.80 and -2.99 eV if the atomic O energy is used as a reference). Given that HSE06 is computationally expensive, we determined the adatom binding energy only for the stable bridge site, and obtained a similar binding energy value of 0.34 eV with respect to molecular O_2. Similarly, for PBE+U, the adatom binding energy for the stable bridge site was 0.38 eV with respect to molecular O_2. At the bridge site, the O-Ce bond length is 1.95 Å, whereas the bond length increases to 2.40 Å and 2.35 Å for the hollow and top sites, respectively. The short O-Ce bond length at the bridge site increases its bond strength, making it the most favorable site. Similarly at the hollow site, which is the second most stable site, the increased bond length reduces the orbital overlap thereby weakening its bond. However, at the hollow site, the O atom concurrently bonds with three neighboring Ce atoms, thus inducing greater stability relative to the top site binding. Furthermore,

as the O adatom coverage increases from a dilute case ($\theta = 0.25$ ML) to one where the entire surface is covered with O adatoms ($\theta = 1$ ML), the bridge binding mode remains the most stable site followed by the hollow and top sites. Increasing the surface coverage of O adatoms results in larger adsorbate-adsorbate lateral repulsive interactions resulting in weaker O binding energies. These lateral interactions, however, are not significant even under high O coverage, given the large adatom separation in metal oxides compared to that in metals [107].

Table 3.1: Binding energy of O adatom(s) at different adsorption sites using molecular O_2 as reference. Values in parentheses are the binding energies with atomic O as the reference. $\bullet$ refers to vacancies or adatoms adjacent to each other along the same axis, and $\star$ indicates vacancies or adatoms not along the same axis in the 2×2 cell.

Coverage, θ ML	**Top (PBE)**	**Hollow (PBE)**	**Bridge (PBE)**	**Bridge (PBE+U)**	**Bridge (HSE06)**
0.25	2.24 (-1.08)	1.52 (-1.80)	0.34 (-2.99)	0.38 (-2.95)	0.34 (-2.25)
$0.5^{\bullet}$	2.28 (-1.05)	1.58 (-1.75)	0.38 (-2.95)	0.34(-2.99)	0.45 (-2.14)
$0.5^{\star}$	2.28 (-1.05)	1.58 (-1.74)	0.42 (-2.91)	0.45 (-2.87)	0.48 (-2.11)
0.75	2.30 (-1.02)	2.19 (-1.14)	0.51 (-2.81)	0.65 (-2.68)	0.58 (-2.01)
1	2.32 (-1.02)	2.43 (-0.89)	0.58 (-2.74)	0.68 (-2.65)	0.69 (-1.91)

Table 3.2: Vacancy formation energies for a surface and sub-surface O in the 1^{st} trilayer and at the upper and lower O layers of the 2^{nd} and 3^{rd} trilayers for a ceria surface. The O vacancy formation energy in bulk CeO_2 is also shown for comparison.

	Oxygen Vacancy Formation Energy, eV					
	PBE		**PBE+U**		**HSE06**	
Trilayer	**Upper**	**Lower**	**Upper**	**Lower**	**Upper**	**Lower**
1	3.27	2.98	3.25	2.90	3.60	3.04
2	3.97	3.91	-	-	-	-
3	3.85	3.85	-	-	-	-
Bulk	3.50		-		-	

Having discussed the adatom cases, we now move to mildly reducing conditions where O vacancies start forming in the surface or sub-surface layers. Table 3.2 shows the 0 K energy required to form such an O vacancy, at the surface or sub-surface layer. In the dilute limit represented by a single vacancy in the 1^{st} trilayer of the

2×2 slab, the energy required for an O vacancy formation at the surface or sub-surface layer is 3.27 and 2.98 eV, respectively. Therefore, a sub-surface O vacancy is more stable than a surface O vacancy by 0.29 eV. A similar DFT study comparing the relative stability of surface and sub-surface O vacancies in ceria indicates that a sub-surface vacancy is more stable than a surface vacancy by 0.29 eV due to lattice relaxations [108, 109, 110]. Nolan and colleagues also report a surface O vacancy formation energy of 3.30 eV [111, 112] The O vacancy formation calculations were repeated with the HSE06 level of theory, yielding a formation energy of 3.60 and 3.04 eV for a surface or sub-surface vacancy, respectively, in the 1^{st} trilayer. We observed a slightly higher O surface vacancy formation energy compared to Pirovanno *et al.*, mostly due to lack of geometry optimization [108]. The corresponding values with PBE+U were 3.25 and 2.90 eV, respectively. In all three levels of theory, the sub-surface site is the more stable location for the point defect. The phenomenon of a more stable sub-surface vacancy has also been observed experimentally by Torbrugge and Reichling [87], where the nucleation of sub-surface vacancies occurs prior to the formation of surface vacancies as resolved via atomic and dynamic force microscopy. Table 3.2 reports the O vacancy formation energy for the 2^{nd} and 3^{rd} trilayers as well, using PBE. For the 2^{nd} trilayer, creating an O vacancy requires 3.97 or 3.91 eV from an upper and lower O layer, respectively; whereas for the 3^{rd} trilayer, 3.85 eV is required to remove an O from either of the O layers. Finally, we also report that the O vacancy formation energy for bulk CeO_2 is 3.50 eV, consistent with the value of 3.62 eV reported by Jiang *et al* [113]. These results indicate that the bulk-like nature of the material is progressively recovered as one moves from the surface to the interior. It is also evident that removing O from the bulk-like internal trilayers requires significantly more energy and any underlying chemistry would likely be restricted to the surface (i.e., 1^{st} trilayer). Therefore, we have considered O non-stoichiometry only at the 1^{st}

trilayer in the subsequent discussion.

3.2.3 Relative surface energies of configurations

Before incorporating the finite temperature and pressure effects, we look at the 0 K relative surface energy ($\Delta\gamma$) values for all the configurations considered (Table 3.3). The 0 K $\Delta\gamma$ values were computed by setting $\Delta\mu_{O_2} = 0$ in Eq. 3.2. Comparing cases with similar configuration from Table 3.3, i.e., as β (for $\alpha = 1$, $\theta = 0$) and α (for $\beta = 1$, $\theta = 0$) vary from 0 to 1, we observe that a reduced sub-surface is again preferred over a reduced surface, regardless of the O vacancy concentration. Along with the 0 K $\Delta\gamma$ values, Table 3.3 also compares the three levels of theory.

Table 3.3: Relative surface energy, $\Delta\gamma$, for various O non-stoichiometry and adatom coverage configurations for ceria at 0 K, with reference to the stoichiometric slab ($\alpha = \beta = 1$, $\theta = 0$). For non-zero adatom coverages ($\alpha = \beta = 1$, $\theta > 0$), the adatom location is also mentioned as t, h, or b referring to top, hollow and bridge sites. $\bullet$ refers to vacancies or adatoms adjacent to each other along the same axis, $\star$ indicates vacancies or adatoms not along the same axis in the 2 $\times$ 2 cell, $\diamond$ indicate vacancy at the surface and sub-surface layer, with both the vacancies created by removing O atoms associated with a Ce atom or between a Ce atom and its nearest neighboring Ce atom, respectively. $^{\psi}$ $\Delta\gamma = 0$ eV/Åas CeO_2. Calculated γ with PBE functional = 0.56 J/m^2

$\boldsymbol{\theta}$	$\boldsymbol{\alpha}$	$\boldsymbol{\beta}$	$\boldsymbol{\Delta\gamma}$, eV/Å^2		
			PBE	**PBE+U**	**HSE06**
0	1^{ψ}	1^{ψ}	0	0	0
0	1	0.75	0.058	0.058	0.059
		0.5●	0.155	0.173	0.139
		0.5⋆	0.142	0.148	0.139
		0.25	0.218	0.219	0.198
		0	0.321	0.310	0.278
	0.75	1	0.063	0.063	0.069
	0.5●		0.188	0.208	0.173
	0.5⋆		0.149	0.148	0.146
	0.25		0.269	0.297	0.239
	0		0.342	0.356	0.307
	0	0.75	0.450	0.460	0.419
		0.5●	0.553	0.557	0.517
		0.5⋆	0.862	0.831	0.869
		0.25	0.935	0.980	0.974
		0	0.786	0.778	0.739
	0.75	0.75◇	0.195	-	-
	0.75◇	0.75	0.150	-	-
0.25-t	1	1	0.043	-	-
0.25-h	1	1	0.029	-	-
0.25-b	1	1	0.007	0.011	0.007
0.5●-t	1	1	0.088	-	-
0.5●-h	1	1	0.061	-	-
0.5●-b	1	1	0.015	0.019	0.017
0.5⋆-t	1	1	0.088	-	-
0.5⋆-h	1	1	0.061	-	-
0.5⋆-b	1	1	0.016	0.022	0.018
0.75-t	1	1	0.133	-	-
0.75-h	1	1	0.086	-	-
0.75-b	1	1	0.030	0.038	0.034
1-t	1	1	0.179	-	-
1-h	1	1	0.188	-	-
1-b	1	1	0.045	0.052	0.058

The agreement in $\Delta\gamma$ values among the three levels of theory is excellent, as indicated by the parity plot in Figure 3.3 or by comparing the tabular values in Table 3.3. Validation of our calculated values with literature is challenging given that almost no information exists on non-stoichiometric ceria surfaces to the best of our knowledge.

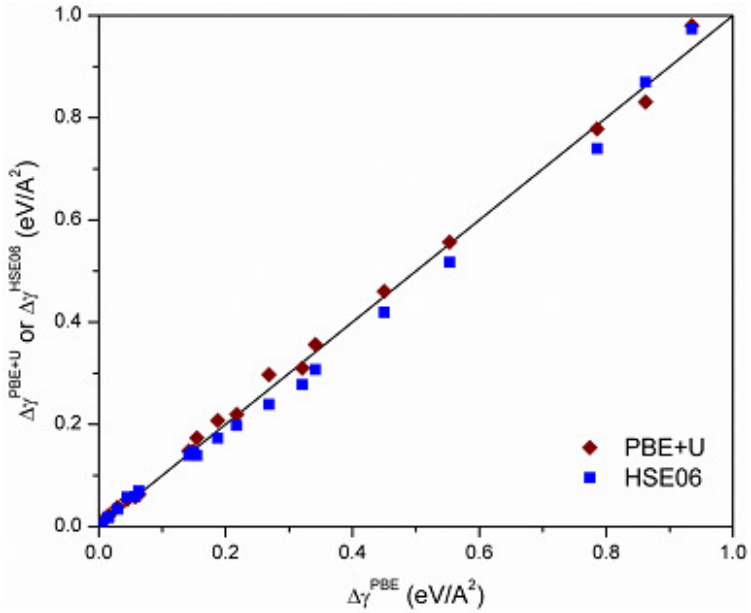

Figure 3.3: Parity plot of PBE+U vs. PBE and HSE06 vs. PBE relative surface energy ($\Delta\gamma$) values for the configurations in Table 3.3.

However, the ceria surface energy value is widely reported, and thus we have used it as a basis of comparison. For a relaxed stoichiometric surface, past data range from 0.45 to 0.77 J/m^2 [114, 115, 113], which compare well with our calculated value of 0.56 J/m^2. The 0 K data provide a generic understanding of the surface; but the finite temperature and pressure contributions must be considered in real operating environments.

Next, we incorporate non-zero temperature and pressure effects using Eq. 3.3 into the relative surface energy relations in Eq. 3.4. Incorporating these factors results in a plot of $\Delta\gamma$ as a function of $\Delta\mu_{O_2}$ as shown in Figure 3.4. With the thermodynamic relations defined, the surface configuration with the lowest $\Delta\gamma$ is the most stable one. In Figure 3.4, we observe that different surface configurations form the minimum trace line, as we progress gradually along the $\Delta\mu_{O_2}$ scale. The $\Delta\mu_{O_2}$ scale is indicative of the driving force for the exchange of O between the surface and its

environment. At highly negative $\Delta\mu_{O_2}$ values, the environment is extremely reducing due to the lack of oxygen, which creates a systematic imbalance that forces O out from the slab. Under such conditions, a highly reduced surface with no surface or sub-surface O atoms is observed, i.e., the surface is Ce terminated. As $\Delta\mu_{O_2}$ becomes more positive, the imbalance created between the oxygen reservoir and the surface diminishes, thereby reducing the extent of surface non-stoichiometry. Moving along the positive direction on the $\Delta\mu_{O_2}$axis in Figure 3.4, the heavily reduced Ce terminated surface configuration $CeO_0^{0,0}$ transitions to a $CeO_1^{1,0}$ configuration (sub-surface layer completely devoid of O), followed by a $CeO_{1.25}^{1,0.25}$configuration (sub-surface layer with high O vacancy concentration), followed by a marginally reduced $CeO_{1.75}^{1,0.75}$ configuration (sub-surface layer with low O vacancy concentration), and finally ending with a stoichiometric CeO_2 surface (no vacancies). With the current model, we do not observe the intermediate configuration of $CeO_{1.5}^{1,0.5}$. This particular stoichiometry is energetically unfavorable, and the stable phase switches between $CeO_{1.25}^{1,0.25}$ and $CeO_{1.75}^{1,0.75}$, as discussed in the sections to follow. At higher $\Delta\mu_{O_2}$ values (> 0.35 eV for PBE, > 0.4 eV for PBE+U, and > 0.3 eV for HSE06), the onset for O adatom adsorption is observed on the ceria surface. These potentials represent an oxygen-rich environment, where a role reversal between the environment and the surface allows for adatom covered surfaces to be thermodynamically favorable. Starting from a stoichiometric or clean surface ($\theta = 0$), the O adatom coverage (θ) increases from 0 ML to 0.25 ML to 0.5 ML to 0.75 ML and finally to 1 ML (complete saturation) as the O_2 content in the environment increases.

At this point, we also note the similarities and differences among the PBE, PBE+U, and HSE06 derived $\Delta\gamma$ plots (Figures 3.4a, 3.4b and 3.4c, respectively). In all the plots, the same stable phases are observed; the relative position of phase transformation however differs to some extent. Given that we solely optimized the

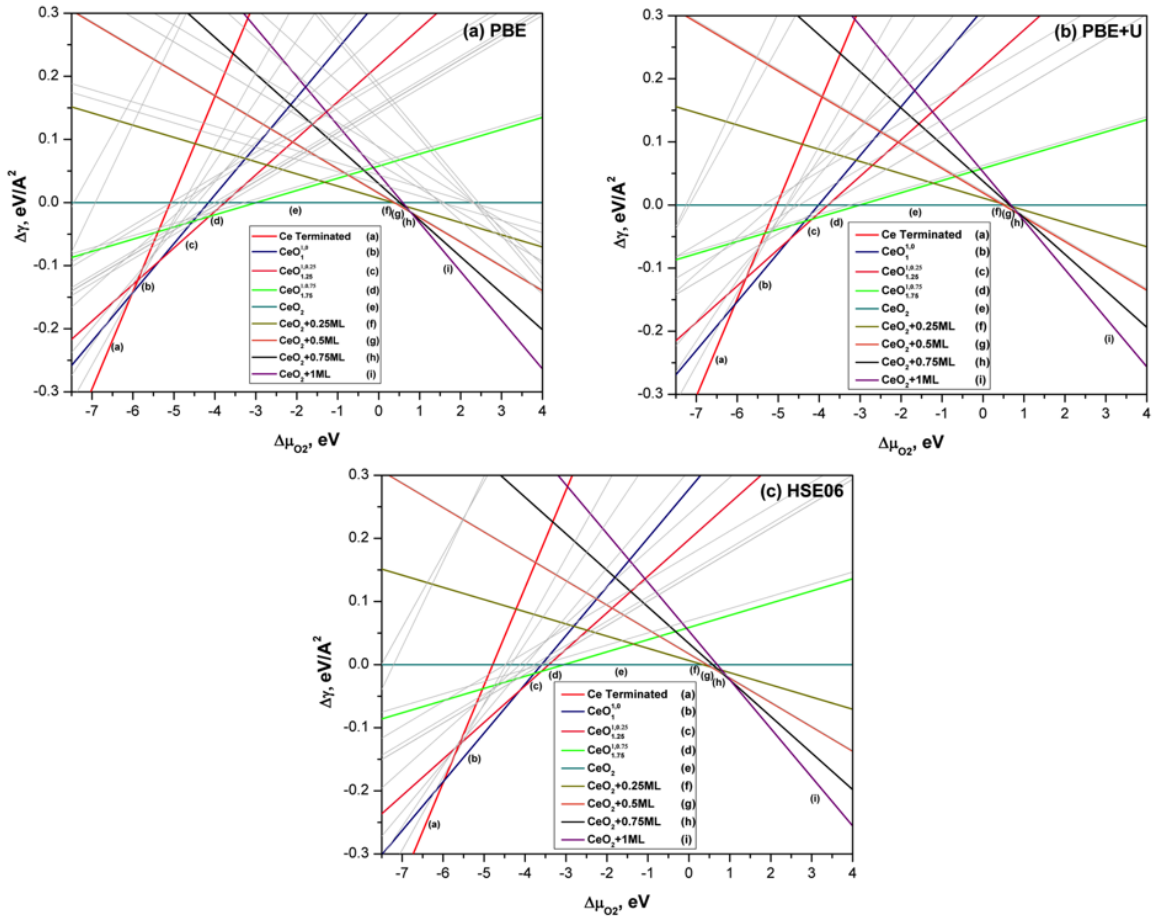

Figure 3.4: Relative surface energy ($\Delta\gamma$) as a function of oxygen potential ($\Delta\mu_{O_2}$) using the (a) PBE, (b) PBE+U, and (c) HSE06 functionals. Minimum energy line represents the most stable phases, whereas the intersection points depict phase transformation regions.

electronic structure in HSE06 calculations, the configurations with significant rearrangement (primarily the heavily reduced ones) show different oxygen potential locations for phase transition, and a geometric optimization could alter these boundaries. However, the initial reduction from CeO_2 to $CeO_{1.75}^{1,0.75}$ or the adatom adsorption from CeO_2 to $CeO_2 + 0.25\ ML$ occurs at similar $\Delta\mu_{O_2}$ conditions for all three levels of theory. Another key point, as described earlier, is that $\Delta\mu_{O_2}$ is bound by the two critical limits. In an oxygen-lean environment ($\Delta\mu_{O_2}$ < -5.23 eV), the onset for the decomposition of the bulk material into its constituent elements, i.e., bulk Ce metal and O_2 gas, occurs. Using the stable phases and the critical conditions observed under different $\Delta\mu_{O_2}$ values allows us to create the ceria surface phase diagram under various oxygen environments.

3.2.4 First principles derived phase diagram for ceria

A mapping of the stable phases observed in Figure 3.4 as a function of $\Delta\mu_{O_2}$ leads to the generation of the ceria surface phase diagram. Essentially, each $\Delta\mu_{O_2}$ value of Figure 3.4, corresponding to a transition from one configuration to another, manifests as a curve in the phase diagram (as prescribed by Eq. 3.3). Such a phase diagram is shown in Figure 3.5 (PBE, PBE+U, and HSE06 results in panels a, b, and c, respectively), reveals that under atmospheric pressure, a ceria surface remains in its stoichiometric state even at extremely high temperatures (up to 2000 K). Progressing from the top left (close to ambient conditions) downwards (lower oxygen pressure) and to the right (higher temperatures), an imbalance between the O_2 reservoir and the surface causes O to desorb from ceria, making a mildly non-stoichiometric surface in the sub-surface layer to be the most stable phase. A further decrease in pressure (and oxygen potential) leads to a high degree of O desorption creating a more reduced

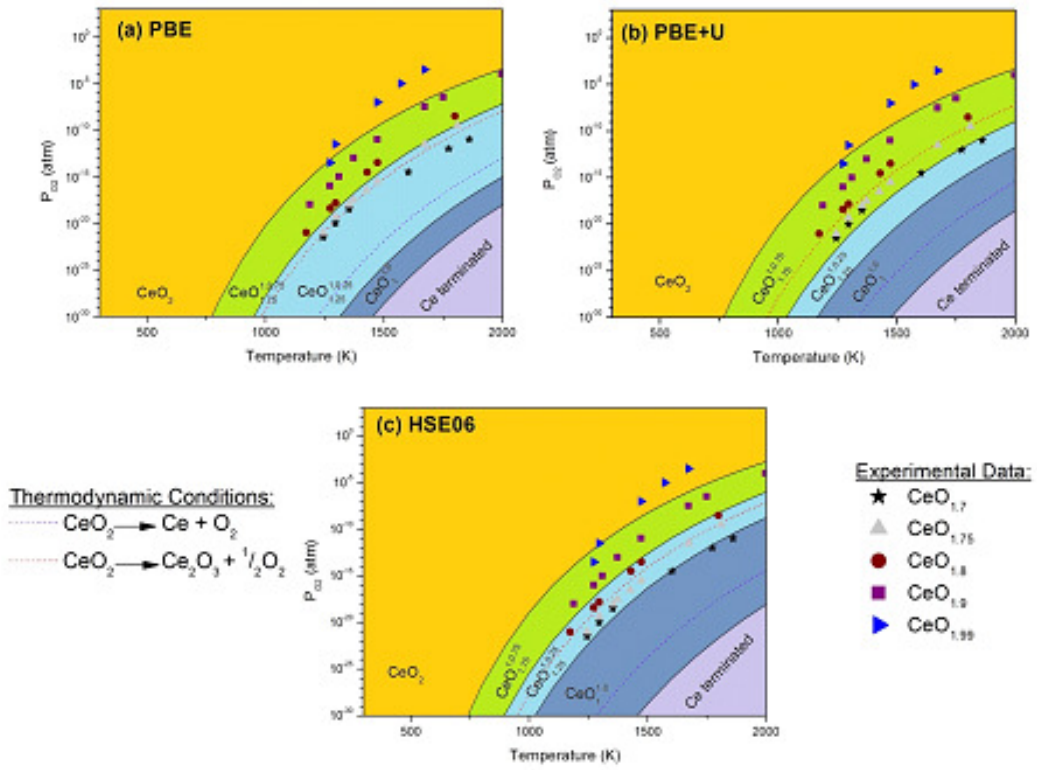

Figure 3.5: Predicted ceria surface phase diagram in an oxygen environment using the (a) PBE, (b) PBE+U, and (c) HSE06 levels of theory. Symbols represent experimental data, and the dashed lines indicate thermodynamically governed relations based on PBE.

sub-surface phase. Eventually under extremely high temperatures and low pressures (highly negative oxygen potentials), all the O atoms desorb from the sub-surface and surface layers, thereby creating a Ce terminated surface. All the manifestations observed in Figure 3.4 are directly translated onto the ceria surface phase diagram in Figure 3.5. The phases corresponding to adatom coverages do not appear under the conditions shown here, given that a practically unrealistic oxygen pressure is required ($P_{O_2} > 10^6$ atm) to observe any trace of adsorption.

After discussing the theoretical insights from the phase diagram, next we com-

pare the phase diagram features with experimental data in literature. Symbols in Figure 3.5 correspond to bulk ceria non-stoichiometry observed in an ensemble of experiments, such as thermo-gravimetric analysis, mass spectrometry, effusion measurements, high temperature X-ray diffraction, thermal expansion measurements, and specific heat measurements. The experimentally observed configurations include $CeO_{1.99}$, $CeO_{1.9}$, $CeO_{1.8}$, $CeO_{1.75}$, and $CeO_{1.7}$ [81, 82]. The presence of ample experimental data allows for a direct comparison with the predicted non-stoichiometry; and as seen in Figure 3.5, the predicted phases are in good coherence with the experimental data. In order to observe any appreciable reduction of bulk ceria, ultra-low oxygen pressures and high temperatures are required, which is consistent with our predicted trend. As mentioned earlier, creating surface non-stoichiometry is energetically less taxing compared to bulk non-stoichiometry; therefore, surface reduction occurs prior to bulk reduction. This phenomenon is also captured by our phase diagram, which indicates the presence of a $CeO_{1.75}^{1,0.75}$ surface phase prior to the experimentally observed bulk stoichiometry of $CeO_{1.75}$. The dashed red line represents the thermodynamic conditions derived from DFT energies for bulk CeO_2 transitioning to its reduced Ce_2O_3 state. Again, in order to observe this transformation, the overall bulk stoichiometry must correspond to that of Ce_2O_3. The dashed purple line represents the thermodynamically governed conditions for the decomposition of bulk ceria into its constituent elements, i.e., Ce metal and O_2 gas. The difference observed in the position of the dashed lines between panels (a) and (b) lies in the governing theory used to create the thermodynamic transformations. For the HSE06 derived phases to be consistent, it is necessary to allow for structural relaxation. This may be important for only those cases where a significant rearrangement of the surface occurs as described earlier. On the other hand, for minimal rearrangement such as the initial transition of stoichiometric ceria to $CeO_{1.75}^{1,0.75}$, using the PBE+U or HSE06 theory results in

only a marginal difference (the phase transformation occurs at a lower temperature by ∼5 K with PBE+U and ∼50 K with HSE06), indicating that the PBE level of theory is sufficient for practical purposes. Based on both theories, the region where ceria undergoes the initial phase transformation (transition of CeO_2 to $CeO_{1.75}^{1,0.75}$) is of utmost importance, as this governs the use of ceria as an O buffering material.

At this point, it is important to briefly mention the limitations under which the phase diagram was derived: (i) The observed stable phases are limited by the initial domain of the configurations considered in this study. Additional intermediate configurations are therefore buried within the phase transition boundary, but they could be discerned using larger unit cell calculations. Nonetheless, the overall predictions based on the considered configurations are in excellent agreement with the experimental and thermodynamic data. (ii) The vibrational entropy contribution of the condensed phases to the total free energy is neglected. While this may be a good assumption at low temperatures (due to favorable cancellation of this contribution in the system before and after O adsorption/desorption), this may have to be revisited at high temperatures close to the melting temperature [106]. Given that the upper temperature limit of the phase diagram is well below the melting point of the fluorite phase of ceria (∼2650 K), we do not expect these effects to drastically alter the predicted energetics. (iii) The abrupt transitions in the derived stoichiometries will not be observed experimentally, but they are rather a manifestation of ignoring the configurational entropy (and point (i) made above). (iv) The dynamics of vacancy filling and migration are not considered here. (v) Using a PBE based geometry as a baseline can induce the formation of Ce^{3+} ions more readily. (vi) Lastly, using a PBE functional results in a delocalization of electrons (charge density plots showing the extent of delocalization between the two functionals are provided in the supplementary information), but captures the initial phase transformation accurately given

that energetics is not significantly altered under a dilute vacancy limit. Despite these assumptions and limitations, we expect the predictions made using such strategies to be at least semi-quantitative in systems involving O chemistry, as has been pointed out earlier, and as is clearly borne out by the current work. The predicted phases not only validate experimentally observed transitions, but are also in excellent agreement with the experimental and thermodynamic data.

3.2.5 Phase diagram with indirect oxygen participation

As discussed earlier, in the presence of a reducing/oxidizing environment, O transfer can occur via an indirect redox reaction of ceria with various gas molecule pairs. FPT serves as a powerful tool to consider these situations of indirect O participation, as discussed by Eqs. 3.6 - 3.8. Given that the relative difference in surface stoichiometry changes marginally when using PBE+U or HSE06, we re-derived the ceria surface phase diagrams in three equilibrium redox environments - NO/NO_2, H_2/H_2O, and CO/CO_2 - (Figures 3.5a, 3.5b, and 3.5c, respectively) using PBE energetics. The critical assumptions made while deriving these phase diagrams are as follows: (i) Direct interactions of CO, CO_2, H, H_2O, NO, and NO_2 with the ceria surface are not included, i.e., the energetics of adsorption and surface reactions involving these species are not incorporated. (ii) The oxygen potential is defined based on a single redox reaction, whereas several side reactions could be possible in a real environment.

The phase diagrams in 3.5 were derived in terms of the ratio of the partial pressures, which governs the oxygen chemical potential as discussed earlier. A low ratio indicates an oxygen deficient (reducing) environment. Figures 3.5a, 3.5b, and 3.5c indicate that the ceria surface is readily reduced in the CO-rich and H_2-rich environments [116, 117], as compared to an NO-rich environment which requires higher

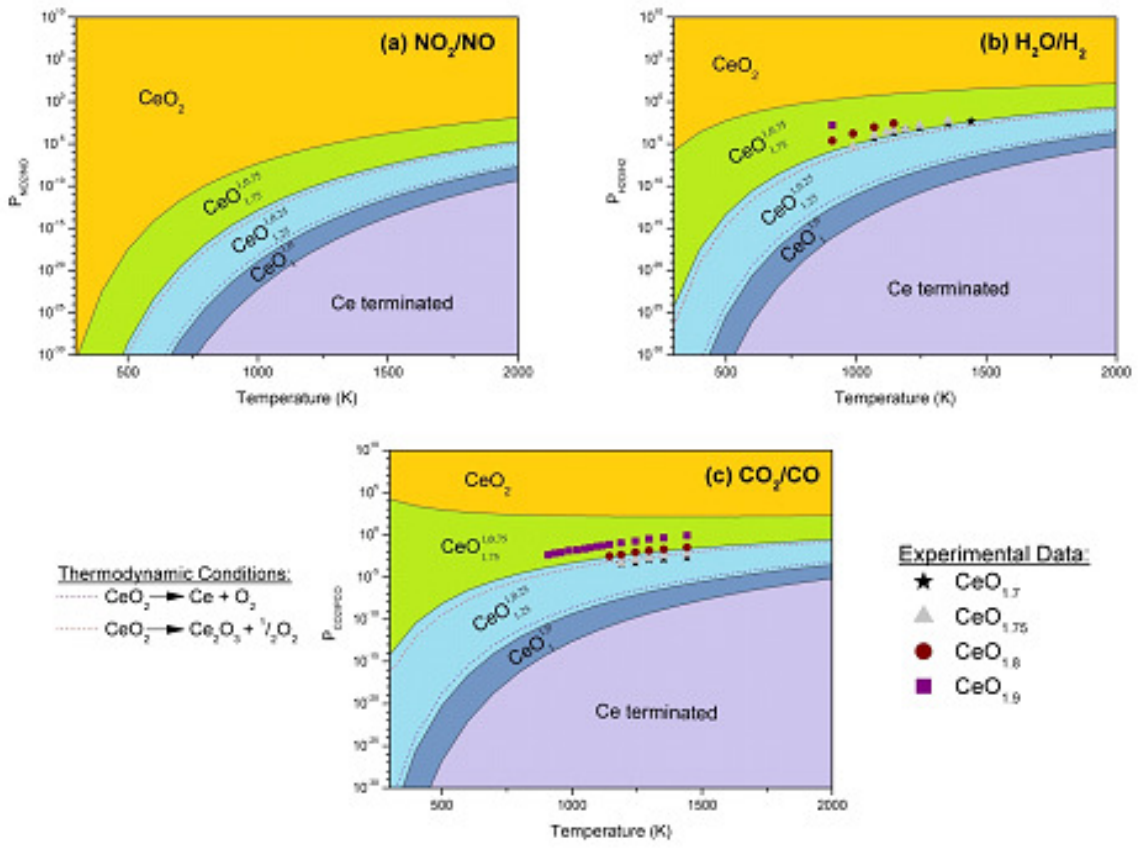

Figure 3.6: Ceria surface phase diagram derived in (a) NO/NO_2, (b) H_2/H_2O, and (c) CO/CO_2 redox environments using the PBE functional. Symbols represent experimental data.

temperatures for the same extent of non-stoichiometry. This is consistent with the notion that H_2 and CO are stronger reducing gases compared to NO [107]. Given assumptions (i) and (ii) above, the difference in the predicted phases in H_2/H_2O and CO/CO_2 environments (Figures 3.5b and 3.5c) is a direct manifestation of the thermodynamics governing their corresponding oxidation reactions. For any given temperature, a decrease in the ratio of partial pressures lowers the oxygen potential, therefore creating a reduced surface. The extent of non-stoichiometry increases as the temperature increases or the ratio of partial pressures decreases. Once again, the predicted phases are in good agreement with the experimentally observed phases of ceria under the considered redox environments [81].

3.2.6 Descriptors for catalyst design

The work presented thus far demonstrates our ability to make the high-fidelity predictions of phase transitions enabled by chemical processes on ceria surfaces using DFT-FPT. In order to use this scheme to tailor the activity of ceria for redox catalysis, e.g., through doping, it may be useful to identify the most important factor (or the *"descriptor"*) that controls the surface oxygen chemistry of ceria [12]. Based on our results, we believe that the O vacancy formation energy is such a descriptor as it largely governs the transition of stoichiometric ceria to a reduced phase, where active sites are created. In this work, even though the sub-surface O vacancy formation energy is identified as a descriptor, both the surface and sub-surface vacancies could play an active role after equilibration. Based on the experimental work by Torbugge *et al.* [87] and Esch *et al.* [90], both types of vacancies exist at $\sim$1300K, and the active rearrangement of the vacancies indicates the availability of defect sites for reactions (outside the scope of our work). O vacancies on a ceria surface also drive

water dissociation into OH and H – a pre-requisite for WGS reaction - as shown by Fronzi *et al* [89]. O vacancies on a ceria surface are known to be thermodynamically favorable during CO oxidation - another key component of the WGS reaction. The creation of such vacancies in a CO environment could provide nucleation sites for water dissociation, which could explain the improved performance of ceria in WGS [118, 119, 120, 79]. Recently, Janik and colleagues reported the role of O vacancies (ceria surface reduction) and metal doping in hydrocarbon conversion [88]. The O vacancies were shown to directly lower the dissociation energy of methane on a reduced surface. The collective evidence signifies the crucial role of O vacancies in chemical reactions involving ceria. As these vacancies play a key role in promoting surface reactions, measuring them and understanding their properties is critical to explaining the trends amongst new catalytic materials being developed or synthesized. Having identified the O vacancy formation energy as the *descriptor*, a framework for the predictive design of novel ceria-based materials for catalytic processes involving oxygen chemistry can now be established.

3.3 Optimal dopant selection for water splitting with cerium oxides

Complete gas phase thermolysis of water is highly endothermic (ΔH = +2.53 eV) requiring temperatures in excess of 4000 K to be thermodynamically favorable, making such reactions unviable for H_2 synthesis [105, 121]. On the other hand, partial thermolysis via a multistep process in the presence of MO catalysts provides an attractive practical alternative [121, 122]. The latter approach is performed at two distinct temperatures (both well below 4000 K): a high-temperature ($\approx$ 2200 K) re-

duction step that involves creation of O vacancies in the MO (and the consequent evolution of O_2 gas), and lower-temperature ($\approx$ 900 K) oxidation steps in the presence of steam, which lead to the filling up of O vacancy centers (resulting in the evolution of H_2 gas). Owing to this multistep procedure, an additional step to separate the H_2 and O_2 products is eliminated entirely. Equations (1)-(3) below represent a reordered version (for ease of subsequent discussion) of the multiple steps involved in this process.

$$\text{MO-V}_{\text{o (s)}} + \text{H}_2\text{O}_{\text{(g)}} \longrightarrow \text{MO-(H)(H)}_{\text{(s)}} \tag{3.9}$$

$$\text{MO-(H)(H)}_{\text{(s)}} \longrightarrow \text{MO}_{\text{(s)}} + \text{H}_{2\text{(g)}} \tag{3.10}$$

$$\text{MO}_{\text{(s)}} \longrightarrow \text{MO-V}_{\text{o (s)}} + (1/2)\text{O}_{2\text{(g)}} \tag{3.11}$$

The (s) and (g) subscripts represent solid and gas phases, respectively. Equations (1) and (2) are the low-temperature steps, with MO-V_o and MO-(H)(H) representing, respectively, the oxide containing an O vacancy and the oxide in which the O vacancy is filled up by a H_2O molecule (with '(H)(H)' indicating that the H atoms of H_2O are adsorbed on the oxide surface). Equation (3) is the high-temperature activation step that leads to the creation of MO-V_o.

Unfortunately, several MOs require temperatures in excess of 2700 K (leading to poor H_2 production efficiencies), leaving only a subset of oxides based on Zn, Fe and Ce to be the most promising [123, 124]. Oxides of Zn and Fe are prone to sintering, phase transformation or volatility due to the proximity of the high temperature step to their melting points [73]. CeO_2, on the other hand, displays high stability and high melting temperature ($\approx$ 2600 K), and is thus overwhelmingly favored [123].

Still, the efficiency of H_2 production with CeO_2 is quite low ($< 1\%$) [124]. This low

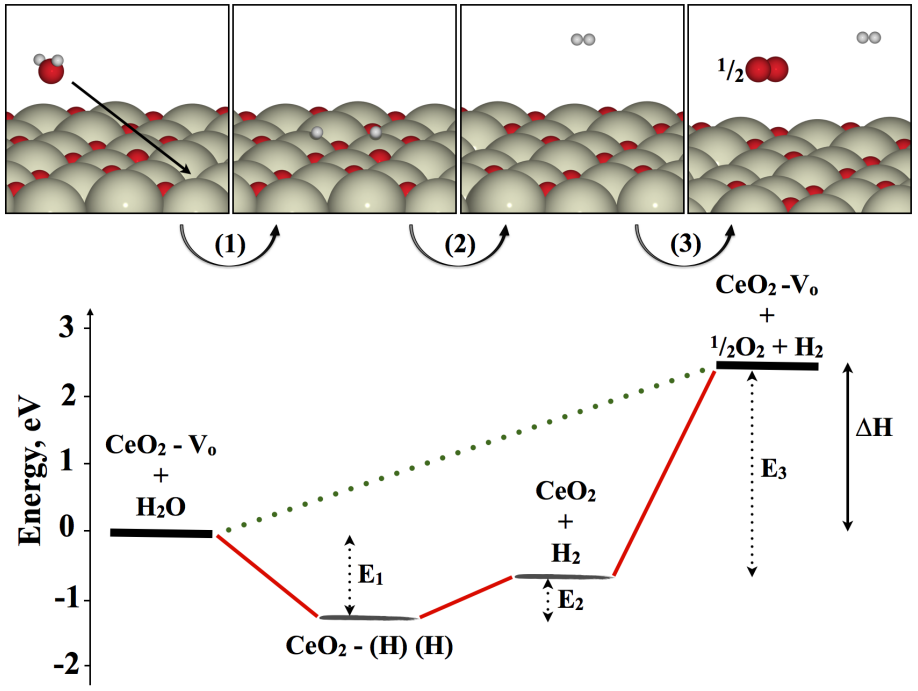

Figure 3.7: Reaction pathway and energetics (red solid line) for the dissociation of H_2O on an undoped ceria surface. CeO_2 - V_o is an oxide with a vacancy, CeO_2 - (H)(H) is an oxide with vacancy filled by a H_2O molecule and CeO_2 is a stoichiometric surface. The green dotted line shows the minimum energy pathway for dissociation. Ce, O and H are represented by beige, red and white colors respectively.

efficiency is rooted in the high temperatures (> 1900 K) required for the reduction step (Equation (3)), related directly to the large O vacancy formation energy of CeO_2, along with other operational difficulties [124, 125]. Figure 3.7 shows the energies E_1, E_2 and E_3 of Equations (1), (2) and (3), respectively, computed here using density functional theory (DFT) (details below), and helps identify the causes of the low efficiency. The dotted line indicates the uphill nature of the water splitting process. The ideal system should display E_1 and E_2 close to zero (for facile H_2 evolution at low temperatures), and small E_3 values (to alleviate the burden on the reduction step). In the case of CeO_2, E_1 is too negative and E_3 is too positive.

A pathway to circumvent these hurdles is to control the energetics of Equations (1)-(3) individually by the introduction of dopants (although, of course, the overall energetics of H_2O splitting cannot be altered). For instance, this strategy may be used to destabilize O in CeO_2 (and thus reduce the O vacancy formation energy) [123, 126, 127, 128, 129, 130, 89, 131]. Doping CeO_2 with a plethora of elements has been explored in the recent past [132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144], and many dopants (e.g., Zr, Cr, Sc) have been shown to help significantly increase the efficiency of H_2 production by reducing the temperatures required to accomplish Equation (3) [136, 138, 139]. *Nevertheless, a clear rationale for why a given dopant is desirable, and a framework for the systematic (non-Edisonian) selection of dopants is currently unavailable.* This work attempts to fill that gap.

3.3.1 Screening Framework

In the present first principles/data-driven based work, we consider a host of dopants in CeO_2, including 33 elements spanning the 4^{th}, 5^{th} and 6^{th} period of the Periodic Table (specifically the *alkali*, *alkaline earth* and *d* series elements). Assuming that

the energetics of Equations (1)-(3) determine whether a dopant is favorable or not, we define the following screening criteria to be used in a successive manner:

- *Criterion 1:* $0 \leq \mathrm{E}_3^{\mathrm{D}} \leq \mathrm{E}_3$
- *Criterion 2:* $0 \leq \mathrm{E}_1^{\mathrm{D}} \leq \delta$
- *Criterion 3:* $0 \leq \mathrm{E}_1^{\mathrm{D}} + \mathrm{E}_2^{\mathrm{D}} \leq \delta$

The superscripts D merely indicate that these are the energetics of doped ceria.

The rationale underlying this specific choice and sequence of screening criteria stems from insights derived from Sabatier's principle, and may be understood as follows (cf. Figure 3.7). *Criterion 1* merely states that the O vacancy formation energy (which is what $\mathrm{E}_3^{\mathrm{D}}$ represents) should not be too small to prevent further water dissociation nor too large (certainly not larger than that of undoped ceria (E_3)) to mandate higher activation temperatures. This criterion is listed first because $\mathrm{E}_3^{\mathrm{D}}$ appears to most strongly control the temperature requirement of the costly high-temperature step, and also because $\mathrm{E}_3^{\mathrm{D}}$ is the easiest quantity to compute (as it does not involve the $\mathrm{H_2O}$ species at all). *Criterion 2* states that $\mathrm{E}_1^{\mathrm{D}}$ should also be bracketed, but by a smaller range. Noting that overall dissociation of water for undoped ceria is too negative (see Figure 3.7), thus potentially adding an energy penalty to subsequent steps, we generously allow δ to be 1.5 eV, which is a reasonable choice considering energy uncertainties within DFT and the neglection of entropy. *Criterion 3* is specific to thermochemical water splitting and bounds the overall oxidation process within δ, ensuring that $\mathrm{E}_1^{\mathrm{D}}$ or $\mathrm{E}_2^{\mathrm{D}}$ occur at a lower temperature compared to $\mathrm{E}_3^{\mathrm{D}}$. In the case where this no longer holds, the process fails to fall within the realm of thermochemical water splitting.

3.3.2 First principles modeling

To measure the thermodynamic quantity, E_i^D, where i is Eq. 1, 2 or 3, DFT calculations were performed using the VASP code with the semi-local Perdew-Burke-Ernzerhof (PBE) exchange-correlation functional and a cutoff energy of 400 eV to accurately treat the valence O 2s, 2p and Ce 5s, 5p, 4f, 5d, 6s states [98, 100, 10]. The electron-core interactions were captured by projector-augmented (PAW) potentials, and all calculations were spin polarized to ensure the true electronic state of O and reduced Ce was captured [99]. The computed lattice parameter of bulk CeO_2 (5.47 Å) is in good agreement with the corresponding experimental value (5.41 Å) [142]. A 96-atom bulk 2×2×2 supercell model and a 60-atom (2×2) surface model (5 O-Ce-O trilayers) cleaved along the (111) plane were used in all calculations. The bottom 3 trilayers of the slab were fixed to recover the bulk nature of the material, and a vacuum of 15 Å along the c axis ensured minimal spurious interactions between periodic images. A Γ-centered k-point mesh of 3×3×3 and 3×3×1 were used for the bulk and surface calculations, respectively. The Hubbard (U) correction was not applied as no universal U value captures the true electronic state of all elements. Also, given that we consider a dilute vacancy limit, the effect of electron localization is insignificant as shown previously [54, 145].

Dopants were introduced by replacing a single Ce atom at the center of the bulk model and at the 1^{st} trilayer of the surface model. Our analysis indicated that the majority of the dopants favored the surface site to the bulk by $\approx$ 0.3 eV. Upon exploring the local coordination environment, a surface dopant was found to be 6-fold coordinated whereas a bulk dopant was 8-fold coordinated. Given the preference of a surface site, all dopants are assumed to occupy the surface unless specified otherwise.

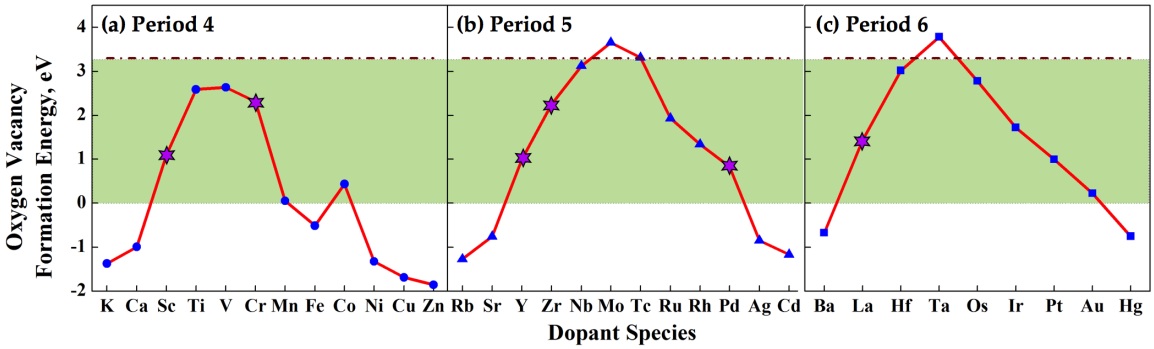

Figure 3.8: Oxygen vacancy formation energy (E_3^D) of doped ceria with elements from the (a) 4^{th}, (b) 5^{th} and (c) 6^{th} period of the Periodic Table. Dot-dashed maroon line indicates E_3^D for undoped ceria. Light green region indicates dopants that survived *Criterion 1*, while $\star$ identifies dopants that survived the 3 screening criteria.

3.3.3 Enforcing the screening criteria

The primary effect of introducing dopants is to induce a local perturbation to disrupt bonding between the metallic and O atoms, thereby altering its ability to form surface O vacancies, as measured by E_3^D (cf. Figure 3.7), computed here as

$$E_3^D = E^D_{CeO_2\text{-}V_o} - E^D_{CeO_2} + \frac{1}{2}\mu_{O_2} \tag{3.12}$$

where $E^D_{CeO_2\text{-}V_o}$ and $E^D_{CeO_2}$ are, respectively, the DFT energies of a doped surface with and without an O vacancy, and μ_{O_2} is the chemical potential of O, taken here to be the DFT energy of an isolated O_2 molecule. In all cases, the O vacancy is created adjacent to the dopant. Figure 3.8 shows E_3^D for various choices of the dopants, with the dot-dashed horizontal line indicating the corresponding value for the undoped case. Dopants adopting a low valence state compared to Ce (e.g., alkali, alkaline earth and late transition series metals) display low O vacancy formation energy, consistent with the observed high O_2 yield by ceria doped with Mn, Fe, Ni and Cu [146]. Conversely, dopants adopting a similar or higher valence state than Ce lead to high E_3^D values

(e.g., Mo, Tc, and Ta). These trends are not entirely surprising, and have been noted before in CeO_2 as well as $BaTiO_3$ [88, 147, 148].

E_1^D helps assess the impact of dopants on the dissociative adsorption of water on the doped surface, and is computed as

$$E_1^D = E^D_{CeO_2\text{-}(H)(H)} - E^D_{CeO_2\text{-}V_o} - \mu_{H_2O} \tag{3.13}$$

where $E^D_{CeO_2\text{-}(H)(H)}$ is the DFT energy of a doped surface upon the dissociative adsorption of water at the vacancy site. Upon dissociation, OH fills the vacancy site, while H has two possible adsorption sites; atop an adjacent O or a dopant atom. Interestingly, dopants exhibiting spontaneous vacancy formation ($E_3^D < 0$ eV) fail to accommodate a H atop a dopant, while those dopants that do facilitate H atop a dopant have an alternative lower energy pathway for dissociation. μ_{H_2O} is the chemical potential of water, taken here to be the DFT energy of an isolated H_2O molecule.

With E_1^D and E_3^D at hand (and E_2^D given by $\Delta H - E_1^D - E_3^D$), a plot that is equivalent to Figure 3.7 but for the case of doped ceria surfaces is shown in Figure 3.9. We now enforce *Criterion 1*, namely, $0 \leq E_3^D \leq E_3$, with $E_3 = 3.3$ eV (this value is consistent with past work [54]). Of the 33 dopants originally considered, 19 dopants (Sc, Ti, V, Cr, Mn, Co, Y, Zr, Nb, Ru, Rh, Pd, La, Hf, Re, Os, Ir, Pt and Au) satisfy this criterion (given by the dopants within the shaded region in Figure 3.8, which are also shown in the right part of Figure 3.9 by darkened horizontal lines). *Criterion 1* picks out those dopants that alter the surface reducibility in just the appropriate manner.

Next, we enforce *Criterion 2*, namely, $0 \leq E_1^D \leq \delta$, with $\delta = 1.5$ eV, on the 19 dopants that pass *Criterion 1*, resulting in the selection of Sc, V, Cr, Co, Y, Zr, Pd, La, Hf and Au. Lastly, enforcing *Criterion 3* on the 10 dopants results in the

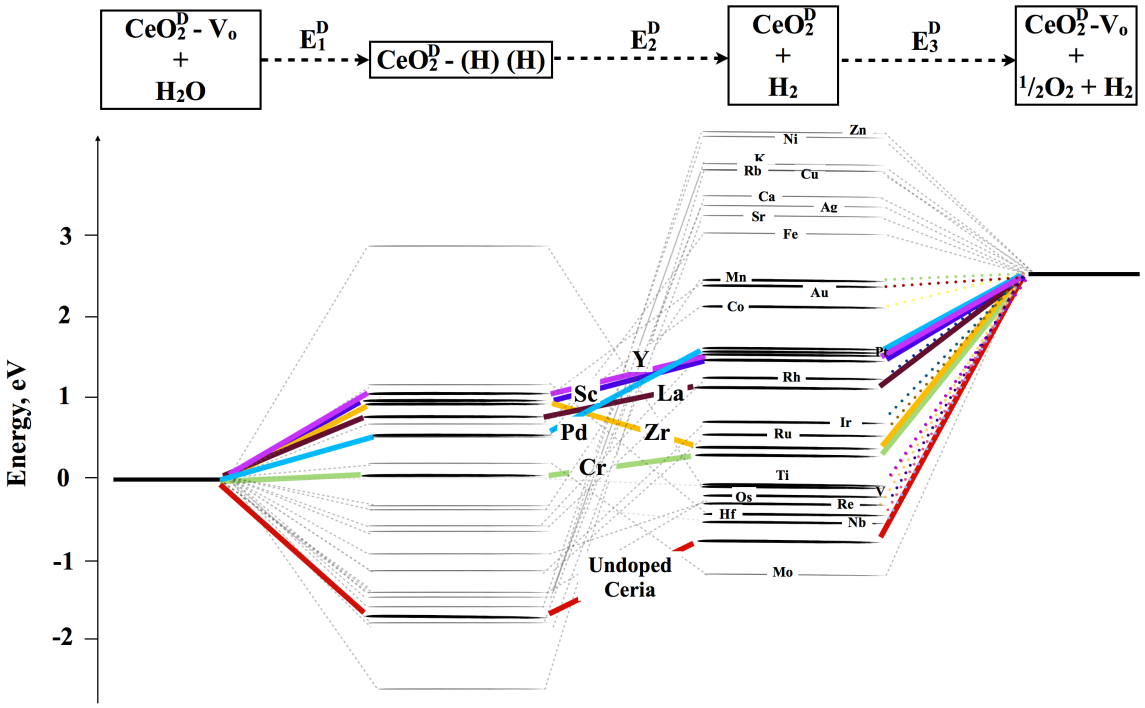

Figure 3.9: Reaction pathway and energetics for the multistep thermochemical splitting of H_2O on a doped ceria surface. CeO_2^D-V_o is a doped surface with vacancy, CeO_2^D-(H)(H) is a doped surface with vacancy filled by a H_2O molecule and CeO_2^D is a doped stoichiometric surface. Color solid lines identify the 4 promising dopants and undoped CeO_2. Grey dashed lines identifies the non feasible dopants, while partly colored and greyed dashed lines identifies dopants that pass Criterion 1.

down selection of 4 promising candidates (Sc, Cr, Zr and La). Inspection of Figure 3.9 shows that Pd and Y, although they do not pass *Criterion 3*, can be viewed as 'near misses'. These are hence included in our final list of favored candidates. Figure 3.10 summarizes the list of dopants that passed each stage of the screening process. The 6 dopants identified, namely, Sc, Cr, Zr, La, Pd and Y, lead to desired energetic profiles, with E_1^D and E_2^D low enough to allow for reasonable H_2O dissociation yields at moderate temperatures, and with E_3^D significantly smaller than undoped ceria allowing for low reduction temperatures (c.f., Figure 3.9). Dopants such as Mn, Fe, Ni, Cu, Sr, Ag, and Ca, which display small or negative E_3^D, do not pass our tests. Although low E_3^D values imply facile surface reduction (this is in fact what is observed experimentally for Mn and Fe) [146], such a tendency would not be appropriate for the multistep thermochemical water splitting process targeted here (lower yields were observed for Ni, Cu and Fe doped CeO_2 compared to undoped CeO_2) [132]. *Criterion 1*, as mentioned above, is imposed precisely to eliminate such candidates. However, dopants that lead to small or negative E_3^D may be appropriate for photocatalytic water splitting which require surface reduction to occur low temperatures ($\approx$ 300 K) [149].

Of the 6 promising dopants identified, experimental evidence exists for the enhanced performance of ceria when doped with Sc, Cr and Zr for the thermochemical water splitting process. Cr doped CeO_2 is known to lower the reduction and oxidation temperature to 750 K and 350 K, respectively [139]. Zr and Sc dopants increase the H_2 yield 4-fold and almost 2-fold, respectively, with respect to the undoped situation [142, 132, 133]. Lastly, although not conclusive, La doping appears to improve H_2 yield [143, 150]. The observed performances are strong functions of the synthesis, processing and measurement details. The present work ignores such complexities, and probes only the dominant and primary chemical factors that may control perfor-

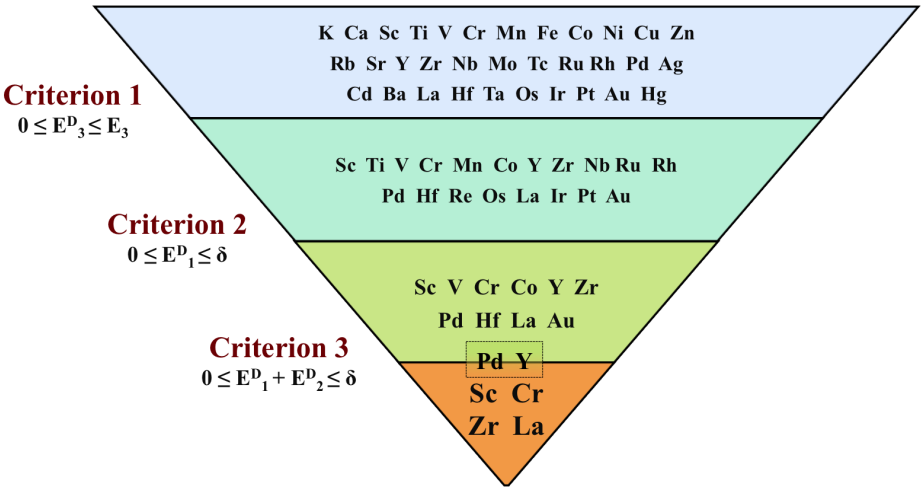

Figure 3.10: A hierarchical chart showing the list of dopants before and after each stage of the screening process. Sc, Cr, Zr and La were identified as the promising dopant elements, whilst Pd and Y can be viewed as the near miss cases.

mance.

Irrespective of these difficulties, such a guided screening strategy has led us to some promising candidates, shown as stars in Figure 3.8. Clearly, the best candidates display an O vacancy formation energy in the 1 - 2.5 eV range, i.e., neither too high nor too low, thereby respecting Sabatier′s principle. It thus appears that the O vacancy formation energy may be used as a 'descriptor' of the activity of doped ceria. This conclusion is consistent with an earlier similar proposal which was based on phase boundaries in surface phase diagrams of ceria exposed to an oxygen reservoir [54].

Thus far, by relying on first principles methods we are able to recognize whether a dopant increases or decreases the O vacancy formation energy, with respect to the undoped material, followed by its corresponding impact on the dissociation of water. However, an understanding of the complex dependence of the chemical attributes of a dopant and the O vacancy formation energy is absent. In the next chapter, with the help of machine learning methods, in particular data mining, we attempt

to understand the results of the first principles computations for the spectrum of dopants considered.

3.4 Mining *ab initio* data

The mining and extraction of information forms the core of the field of data analysis, which lies under a broader umbrella of methods known as machine learning (ML) [31]. Within data analysis a subset of methods, known as feature selection, allows us to unearth correlations between variables [151, 31, 152, 153, 44, 78]. In the context of this work, the variables are the chemical factors characterizing a dopant and the corresponding O vacancy formation energy of doped ceria. Given the strong correlation between the O vacancy formation energy and the activity, as discussed above, by identifying the key dopant factors that contribute to the O vacancy formation energy, a more educated guess on its impact on the corresponding thermodynamic activity can be made.

In order to discover such patterns, firstly, each dopant element needs to be represented numerically by a vector of numbers (also referred to as features or fingerprint in the ML community) that uniquely identifies the dopant element. Our choice of features stems from fundamental chemical factors, that are often used to describe elements in the periodic table. The 7 factors considered in this work are; atomic radius (AR), ionic radius (IR), covalent radius (CR), ionization energy (IE), electronegativity (EN), electron affinity (EA) and oxidation state (OS). To eliminate any bias induced by the spread of the feature values, the dataset was normalized to a mean of 0 and variance of 1. On these set of chemical factors we use two feature selection methods: (i) principal component analysis and (ii) random forests, to narrow down the dominant factors that govern the descriptor (O vacancy formation energy). In

the sections to follow we provide a brief overview of these methods and discuss the insights gained. We refer the readers to [154, 155, 70, 156, 157, 31] for a more exhaustive description. The data analysis routines used were implemented within the MATLAB statistical toolkit and Scikit-learn python module [158, 159].

3.4.1 Finding patterns: Principal component analysis

Principal component analysis (PCA) is a common dimensionality reduction technique, often used to identify the dominant subset of features from a larger pool. By transforming the original features into uncorrelated and orthogonal pseudo variables, that are a linear combination of the original features (as done in this work, although non-linear combinations have been recently developed), it allows us to pin point the dominant contributions [152, 153, 44, 154, 155]. The new transformed variables are referred to as *principal components* (PCs), which are solutions to the eigen-transformation of the covariance matrix. As with any eigen-transformation problem, the eigenvalues and eigenvectors play a critical role. The eigenvalue of a PC indicates the % of variance captured within the original dataset, whilst the eigenvector provides the coefficients that dictate the linear transformation. We shall make use of this information to down select the dominant chemical factors of a dopant.

First, we plot the transformation coefficient values of the 7 features for the first and second PCs in Figure 3.11a. Such a plot is referred to as the loadings plot, in which correlated features cluster together. Only the first and second PCs are used as it captures $\approx 80\%$ of the variance within the original dataset (c. f., inset of Figure 3.11a). Clearly, the dopant's OS is strongly correlated with the O vacancy formation energy. The CR, AR, IE and EN are close to orthogonal to the O vacancy formation energy, suggesting a negligible contribution to the descriptor. On the other

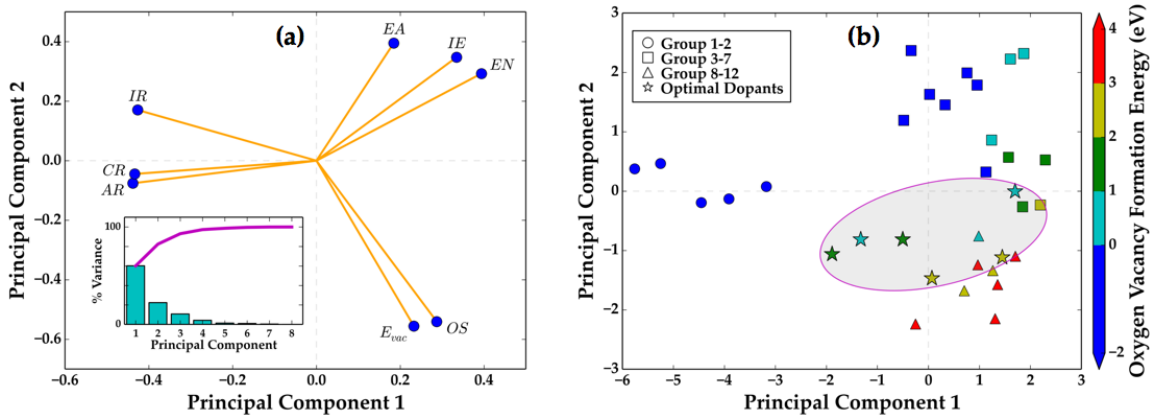

Figure 3.11: (a) PCA loadings plot showing the correlated dopant features. The features are; atomic radius (AR), ionic radius (IR), covalent radius (CR), ionization energy (IE), electronegativity (EN), electron affinity (EA) and oxidation state (OS). E_{vac} is the O vacancy formation energy. The inset shows the % contribution of each PC to the variance in the dataset. The oxidation state (OS) is the dominant feature governing the O vacancy formation energy. (b) PCA scores plot for the first and second principal components. The dopant elements group together based on their features and the O vacancy formation energy. $\star$ represents the final 6 dopants after the 3 step screening processes. The 6 dopants occupy a sub-space of the scores plot as highlighted by the grey region.

hand, the IR and EA are not truly orthogonal, thus their contribution towards the descriptor cannot be ignored. Another interesting phenomena is the congregation of subsets of the 7 features. This isn't entirely surprising, as one would recognize that the AR, CR are similar quantities, and their grouping in the loadings plot further validates this notion. Similarly, the IE and EN group together and appear negatively correlated to the AR and CR, given their $\approx 180^o$ separation. By looking at the relative position of all the features in Figure 3.11a, we can conclude that of the original 7 features considered only 3 are important; OS, IR and EA, in governing the O vacancy formation energy.

Next, we use the linear transformation coefficients of the PCs to transform the original dopant dataset (also referred to as the scores plot) and plot the first and second PCs in Figure 3.11b. Each dopant element in Figure 3.11b has further been

classified according to its relative location in the periodic table (as indicated by the different marker type) and the corresponding O vacancy formation energy (marker fill color). Firstly, dopants of similar type, groups 1-2, 3-7 and 8-12 can be seen aggregating together. In particular, dopants that adopt a low valence state lie predominantly in the top/left quadrants, whilst the high valence dopants lie in the bottom/right quadrants, giving rise to an increasing O vacancy formation energy in the direction of the bottom right quadrant. Not surprisingly, amongst the low valence dopants, the alkali and alkaline earth metals further segregate from the late transition series metals, based on their differences of atomic size, amongst others. Now, upon highlighting the location of the 6 promising candidates (Sc, Cr, Y, Zr, Pd and La), as indicated by the stars, they can be seen to occupy only a small subspace of the plot (highlighted by the grey region of Figure 3.11b). This suggests that in the high dimensional transformation these elements have similar traits, and equivalentaly a similar thermodynamic activity. Therefore, if one could identify other possible dopants that populate the grey region in Figure 3.11b, we can further extend the chemical space to achieve improved water dissociation.

3.4.2 Predictive model for the descriptor: Random forest

Another important class of regression and feature selection algorithms are random forests (RF). Unlike PCA, random forests work by constructing a regression (or classification) model first, in this case between the 7 features and the O vacancy formation energy, following which the important features are then extracted as a by-product. The framework is built upon an ensemble of individual regression models, also known as decision trees [31, 156, 157, 70]. The prediction of each individual tree is then averaged across the ensemble, resulting in the final or true predicted value. Given

our limited dataset size (based on 33 dopant elements), we selected a 75% split for training, with the remaining kept aside as validation/testing. Each decision tree in the model is then trained on a subset of the original training dataset, a procedure known as bootstrapping. The combination of bootstrapping and ensemble averaging makes RF models robust and devoid of overfitting, a common issue in ML. We generate a forest of 250 trees, based on the 7 dopant features described earlier and the O vacancy formation energy. The final regression model we obtained has an R^2 value of 0.95 (c. f., inset Figure 3.12), suggesting a good fit. Then by using mean decrease impurity metric, we estimate the relative importance of each feature in the regression model [157].

In Figure 3.12, we plot the relative importance of the 7 features in descending order. Clearly, the role of a dopant's OS supersedes all others. This observation is consistent with the PCA analysis above. Also, it can be seen that IR and EA rank 2nd and 3rd in feature importance in the regression model, once again suggesting a small contribution towards the descriptor.

Both the PCA and RF methods result in similar conclusions, leading us to believe that the dopant's OS primarily governs the role of the descriptor, i.e. O vacancy formation energy, followed by a much smaller contribution of the IR and the EA. Upon revisiting the OS of the 6 promising dopants, they adopt either a +3 or +4 state. Therefore as a first measure, by understanding the coordination environment of the dopant within the surface one can hazard a reasonable guess on its corresponding impact on the O vacancy formation energy. Even though many other elements such as Ti, V, Mn, Fe, Nb, Mo, Tc, Ru, Rh, Hf, Ta, Os, Ir adopt a similar OS state, the combination of the OS, IR and EA skews them out of the optimal regime.

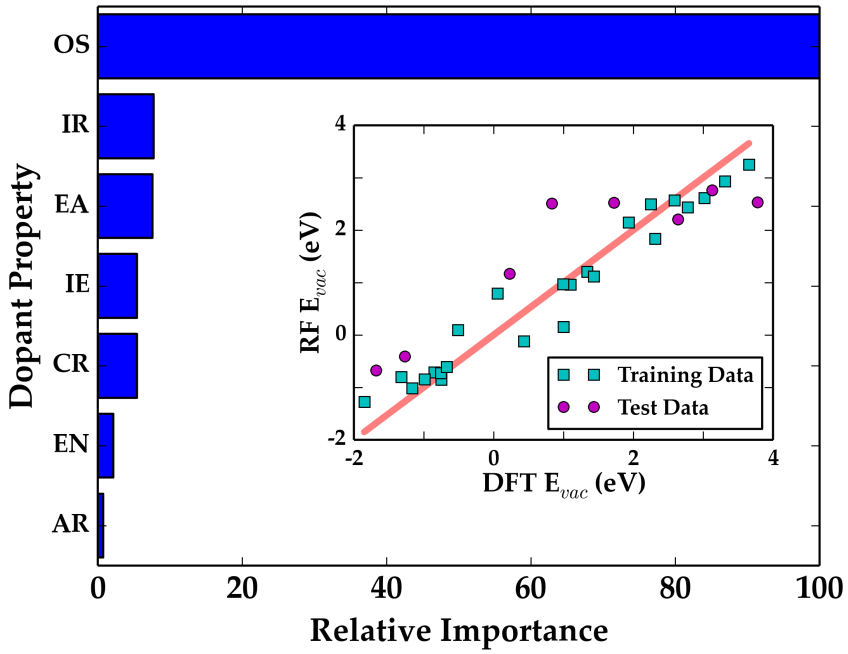

Figure 3.12: Relative feature importance arranged in descending order for the developed RF model. The features are; atomic radius (AR), ionic radius (IR), covalent radius (CR), ionization energy (IE), electronegativity (EN), electron affinity (EA) and oxidation state (OS). E_{vac} is the O vacancy formation energy. The inset shows a parity plot, comparing the density functional theory (DFT) and RF predicted O vacancy formation energy (E_{vac}). The regression model has an R^2 value of 0.94. The oxidation state (OS) is the dominant feature governing the O vacancy formation energy.

3.5 Summary

In this chapter, we started by deriving a surface phase diagram for ceria in four types of oxygen environments involving direct and indirect oxygen participation using first principles thermodynamics. Our results correlate well with literature experimental data in every case where such comparisons can be made. A stoichiometric ceria surface in a pure O_2 environment is highly stable; and any appreciable surface reduction occurs only under extreme temperatures and oxygen pressures. Surface reduction is however more facile in the presence of a redox environment (NO/NO_2, H_2/H_2O, or CO/CO_2). $H_2$and CO, being strong reducing agents, promote O capture from the sub-surface at any given temperature, as compared to an NO or a pure O_2environment. Transition from a stoichiometric to a reduced surface is a crucial factor in governing the surface reactivity of ceria in redox reactions. The O vacancy formation energy primarily drives this transition and can thus be viewed as a *descriptor* for the catalytic activity of ceria in redox reactions.

Following this discovery, we considered a host of dopants in cerium oxide, that span the 4^{th}, 5^{th} and 6^{th} period (specifically the *alkali*, *alkaline earth* and *d* series elements) of the Periodic Table, in order to understand the impact on the dissociation of water. Using a screening framework based on a first principles strategy augmented with data analysis methods, we successfully identified 6 promising dopants (Sc, Cr, Y, Zr, Pd and La), consistent with past experimental results, that are worthy of further inquiry. A dopant's oxidation state, ionic radius and electron affinity are found to be the dominant chemical factors that primarily govern the oxygen vacancy formation energy, which in turn governs the activity. The overall framework, we believe, can be easily extended for dopant selection in ceria and other oxides as well as for different chemical conversion processes (e.g. thermochemical CO_2 splitting, chemical looping,

etc.). This shows that such a descriptor-based learning framework can indeed be used to drive rational materials discovery.

Chapter 4

Modeling kinetic behavior

4.1 Introduction

The dynamic behavior of an atom in a molecule, liquid or solid is directly determined by the *local* force it experiences. Nevertheless, as already pointed out by Feynman [64], forces are generally viewed as secondary computed quantities and are obtained through the agency of the total potential energy - a *global* property of the entire system. In practice, forces on atoms are obtained either as by-products during a potential energy evaluation, or from the first derivative of the potential energy with respect to the atomic positions. Direct and rapid access to atomic forces, given just the atomic configuration of a system (molecule, liquid, or solid), immediately makes it possible to perform efficient geometry optimizations and molecular dynamics (MD) simulations, provided, of course, the predicted force is formally conservative, i.e., it is consistent with an underlying potential energy surface. If the capability to predict conservative forces preserves the fidelity of high-level quantum mechanics based methods, but comes at a minuscule fraction of the cost, and if this capability can be extended systematically and progressively to potentially all configurational and chemical environments that an atom may experience, we will have a powerful and adaptive materials simulation scheme.

In this chapter a recipe for the construction of a stand-alone data-driven force

model (devoid of any explicit functional form) is provided, that can also provide the underlying potential energy surface (through integration). Both the forces and the potential energy can be predicted with a high level of accuracy at speeds several orders of magnitude faster than the reference quantum mechanics based calculations. Moreover, this *force field* is adaptive (i.e., new configurational environments can be systematically incorporated as required), and generalizable (i.e., the scheme can be extended to any collection of elements for which reliable reference calculations can be performed). A practical scheme that exploits the rapid high-fidelity force prediction capability within a materials simulation framework is presented, and demonstrated for Al in several configurational environments and dynamical situations that go well beyond the reaches of conventional first principles simulations. Further, a preliminary analysis to extend this concept to handle multi-elemental systems is also proposed.

4.2 Machine learning force fields: Construction, validation and uncertainty quantification

Materials modeling approaches largely fall in two broad categories: one based on quantum mechanical methods (e.g., density functional theory, Hartree-Fock-based treatments), and the other based on semi-empirical analytical interatomic potentials or force fields (e.g., Stillinger-Weber potentials, embedded atom method) [160, 38, 77, 161, 162, 75, 163, 164, 165]. Choosing between the two approaches depends on which side of the cost-accuracy tradeoff ones wishes to be at (c.f., Figure 4.1). Quantum mechanics based methods (also referred to as *ab initio* or first principles methods) are versatile, and offer the capability to accurately model a range of chemistries and chemical environments. But these methods remain computationally very demand-

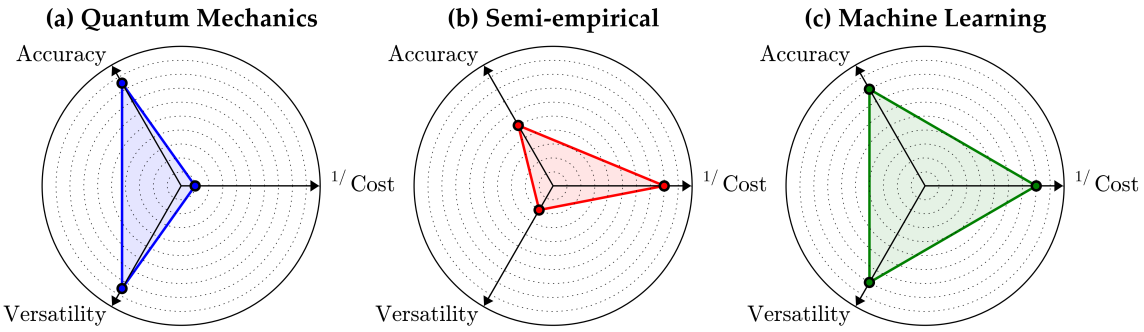

Figure 4.1: A qualitative estimate of the trade-off between the accuracy, cost, and versatility, in selecting (a) quantum mechanical and (b) semi-empirical methods. For comparison, we show the intended regime of the proposed machine learning method (c).

ing; limiting both the length and time scales of phenomena (to $\approx$ nanometers and $\approx$ picoseconds, respectively) that one may aim to treat in a practical and routine manner. Semi-empirical methods capture the essence of the interatomic interactions in a simple manner (via parameterized analytical functional forms), and are thus an inexpensive solution to the materials simulation problem. Nevertheless, their applicability is severely restricted to the specific chemistries and chemical environments intended, or considered during parameterization.

This section pertains to an alternative, data-driven, manner by which flexible and adaptive force fields may be developed [56, 58]. By using carefully created benchmark data (say, from quantum mechanics based materials simulations) as the starting point, non-linear associations between atomic configurations and potential energies (or forces, more pertinent to the present contribution) may be learned by induction [35, 31, 166]. This data-driven paradigm, popularly referred to as machine learning, has been shown by many groups to lead to viable pathways for the creation of interatomic potentials that; (1) surpass conventional interatomic potentials both in accuracy and versatility, (2) surpass quantum mechanical methods in cost (by orders of magnitude), and (3) rival quantum mechanics in accuracy [167, 53, 168], at least

within the configurational and chemical domains encompassed by the benchmark dataset used in the training of the potential.

A new recent development within the topic of machine learning based interatomic potentials is the realization that the vectorial force experienced by a particular atom may be learned and predicted directly given just a configuration of atoms [56, 58, 169]. This capability is particularly appealing as the atomic force is a local quantity purely determined by the local environment, in contrast to the total potential energy which is a global property of the system as a whole (note that partitioning the total potential energy into individual atomic contributions, conventionally adopted in semi-empirical interatomic potentials, is a matter of convenience of construction, rather than being a fundamental requirement). Moreover, a large body of materials simulations, such as geometry optimization and molecular dynamics simulations, require the atomic force as the sole necessary input ingredient [160]. This article deals specifically with using machine learning methods to create an atomic force prediction capability, i.e., a force field, which can also provide the underlying potential energy surface (through integration). As recently pointed out, this force field is *adaptive* (i.e., new configurational environments can be systematically added to improve the versatility of the force field, as required), *generalizable* (i.e., the scheme can be extended to any collection of elements for which reliable reference calculations can be performed), and is *neighborhood informed* (i.e., a numerical fingerprint that represents the atomic environment around the reference atom is mapped to the atomic force with chemical accuracy) [56, 58]. The force field is henceforth dubbed AGNI.

Here, we describe in detail the key steps involved in the construction of the AGNI force field. These include: (1) creation of a reference dataset derived from a plethora of diverse atomic environments of interest and the corresponding atomic forces computed using a chosen quantum mechanical method; (2) fingerprinting every atomic

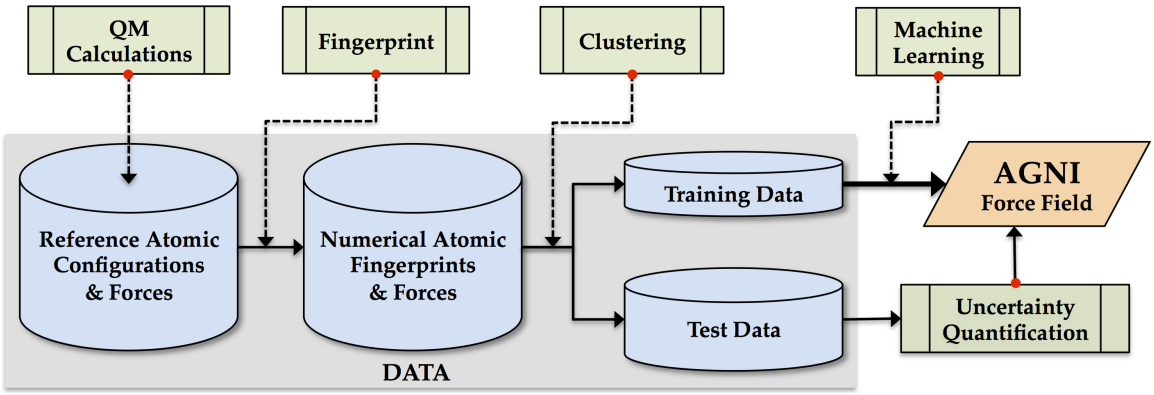

Figure 4.2: Flowchart illustrating key steps in constructing AGNI force fields; generating reference atomic configurations and forces, fingerprinting the atomic environments, selecting training and test datasets, learning the forces and quantifying uncertainty in predictions made.

environment in a manner that will allow the fingerprint to be mapped to atomic force components; (3) choosing a subset of the reference dataset (the "training" set) using clustering techniques to optimize the learning process while insuring that the training set represents the diversity encompassed by the original reference dataset; (4) learning from the training set, thus leading to a non-linear mapping between the training set fingerprints and the forces, followed by testing the learned model on the remainder of the dataset using best-statistical practices; and (5) estimation of the expected levels of uncertainty of each force prediction, so that one may determine when the force field is being used outside its domain of applicability. The entire workflow involved in the creation of the AGNI force field is portrayed schematically in Figure 4.2 , and is demonstrated step-by-step for the example of Al in the present article. The last point, namely, uncertainty quantification, is essential to systematically improve the force field, as atomic environments that lead to forces with unacceptable levels of uncertainty may be included in the reference dataset to create a refined force field, thus making the force field generation process adaptive - a component that should be at the heart of any machine learning paradigm, as originally proposed by Turing

[170].

4.2.1 Generating reference data

The first step in the construction of an accurate AGNI force field comprises the careful generation of reference atomic environments and the corresponding forces. Here, we begin with several periodical and non-periodical reference configurations (consisting of a few atoms, c.f., Figure 4.3), such as; (i) defect free bulk, (ii) surfaces, (iii) point defects - vacancies and adatoms, (iv) isolated clusters, (v) grain boundaries, (vi) lattice expansion and compression, and (vii) edge type dislocations, in order to compile a diverse set of atomic environments. For the configurations amassed, quantum mechanically accurate forces were then computed with DFT [9, 8], using the Vienna *ab initio* simulation package [98, 100]. Starting from the equilibrium configurations (where forces on the atoms are close to zero), constant temperature molecular dynamics (MD) simulations (at 800 K, with a timestep of 0.5 fs) were performed [171] - providing a spectrum of force values needed to learn/understand the underlying potential energy surface. This results in over a million reference atomic environments and forces(components). In all the calculations, the generalized gradient approximation functional parameterized by Perdew, Burke, and Ernzerhof to treat the electronic exchange-correlation interaction, the projector augmented wave potentials, and plane-wave basis functions up to a kinetic energy cutoff of 520 eV were used [10, 99]. A 14×14×14 Γ-centered k-point mesh was used for the primitive Al unit cell, and varied according to the unit cell.

From within the large pool of reference data, the choice of training environments plays a critical role in the generalizability of data-driven force fields. To better understand such limits imposed by data choices we construct four datasets, labeled as

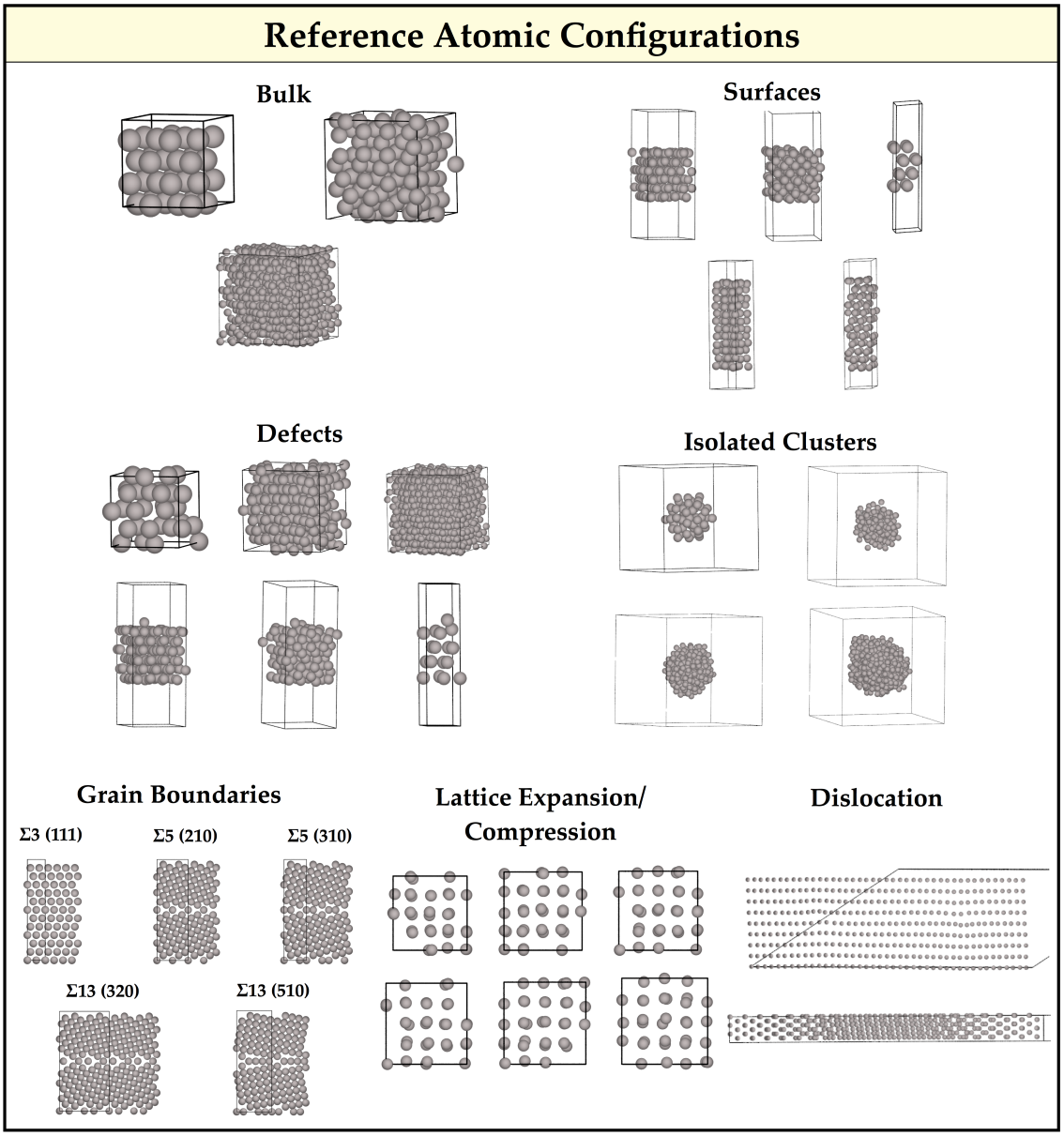

Figure 4.3: Reference atomic configurations used to construct and test AGNI force fields; (i) bulk, (ii) surfaces, (iii) defects (vacancies and adatoms), (iv) isolated clusters, (v) grain boundaries, (vi) lattice expansion and compression, and (vii) dislocation.

A, B, C, and D, with increasing complexity and diversity of atomic environments contained (c.f., Table 4.1). For each dataset a training and test set are compiled (the method used to gather the two sets is discussed in a later section) - the former used to construct the force field and the later used to validate its predictive prowess. Also a fifth dataset, E, consisting of configurations never used during AGNI construction are gathered, solely to further demonstrate the generalizability of AGNI force fields.

Table 4.1: Atomic environment makeup for the five datasets; A, B, C, D and E. For each dataset we generate a training and test set (except for dataset E, where only a test set is created) - the former used to construct the force field and the later to validate it. The number of new environments added is given in the last column.

Dataset	Atomic Envs. in Dataset		Number of Envs.
A	Defect free bulk fcc and bcc.		20385
B	Dataset A +	(100), (110), (111), (200), and (333) surfaces.	211255
C	Dataset B +	Defects in bulk fcc with 1, 2 and 6 randomly distributed vacancies and adatom on (100), (110) and (111) surfaces.	1502856
D	Dataset C +	Isolated clusters of 5Å, 8Å, 10Å, and 12Å.	586679
E	$\Sigma3$ (111), $\Sigma5$ (210), $\Sigma5$ (310), $\Sigma13$ (320), and $\Sigma13$ (510) grain boundaries, varying lattice vectors by ±7 % of equilibrium, edge dislocation along $(11\bar{2})$ direction.		394116

4.2.2 Fingerprinting atomic environments

Functional form

Another critical step in the proposed learning approach is to represent the chemistry and geometry of our system numerically (hopefully, uniquely), such that a mapping can be established between this numerical representation and the property of interest (namely, the forces). Such a representation is referred to here as a "fingerprint" (also

commonly referred to as the *feature vector* by the ML community). The atomic fingerprint chosen here describes the entire ensemble of atoms that are contained within a repeating unit cell (or a molecule), and is necessary to predict atomic properties (e.g., forces).

The atomic fingerprint is required to satisfy certain requirements [172, 173]. In order to adequately capture variations in energy and forces with geometry differences, the fingerprint has to be continuous with respect to slight changes in configuration. Moreover, transformations such as translations, rotations and permutations of atoms of the same type that lead to equivalent systems should not alter the fingerprint.

A natural first choice for the atomic fingerprint of an elemental system could be the radial distribution function (RDF) defined as follows for a particular atom i

$$R_i(r) = \sum_{j \neq i} \delta(r - r_{ij}) \tag{4.1}$$

where $\delta(r)$ is the Dirac delta function and $r_{ij} = |\boldsymbol{r}_i - \boldsymbol{r}_j|$, with $\boldsymbol{r}_i$ being the vectorial position of atom i. The sum runs over all the neighboring atoms within an arbitrarily large cutoff distance from atom i. Clearly, the RDF, $R_i(r)$, satisfies both the fingerprint requirements mentioned above, and has recently been used to establish structure-property mappings in materials [174]. The values of R_i in a radial grid can thus be viewed as a numerical fingerprint (or feature vector) describing the coordination environment. Moreover, $R_i(r)$ also captures the geometry in a visually appealing manner. This is demonstrated in Figure 4.4. Panel A contains three homonuclear diatomic molecules (labeled a, b and c) used here to illustrate our fingerprint choices, and Panel B shows the corresponding Gaussian smoothened RDFs. Clearly, the similarity between the bond distances of molecules a and b, and their dissimilarity with that of molecule c is reflected by the corresponding RDFs. Nevertheless, while these

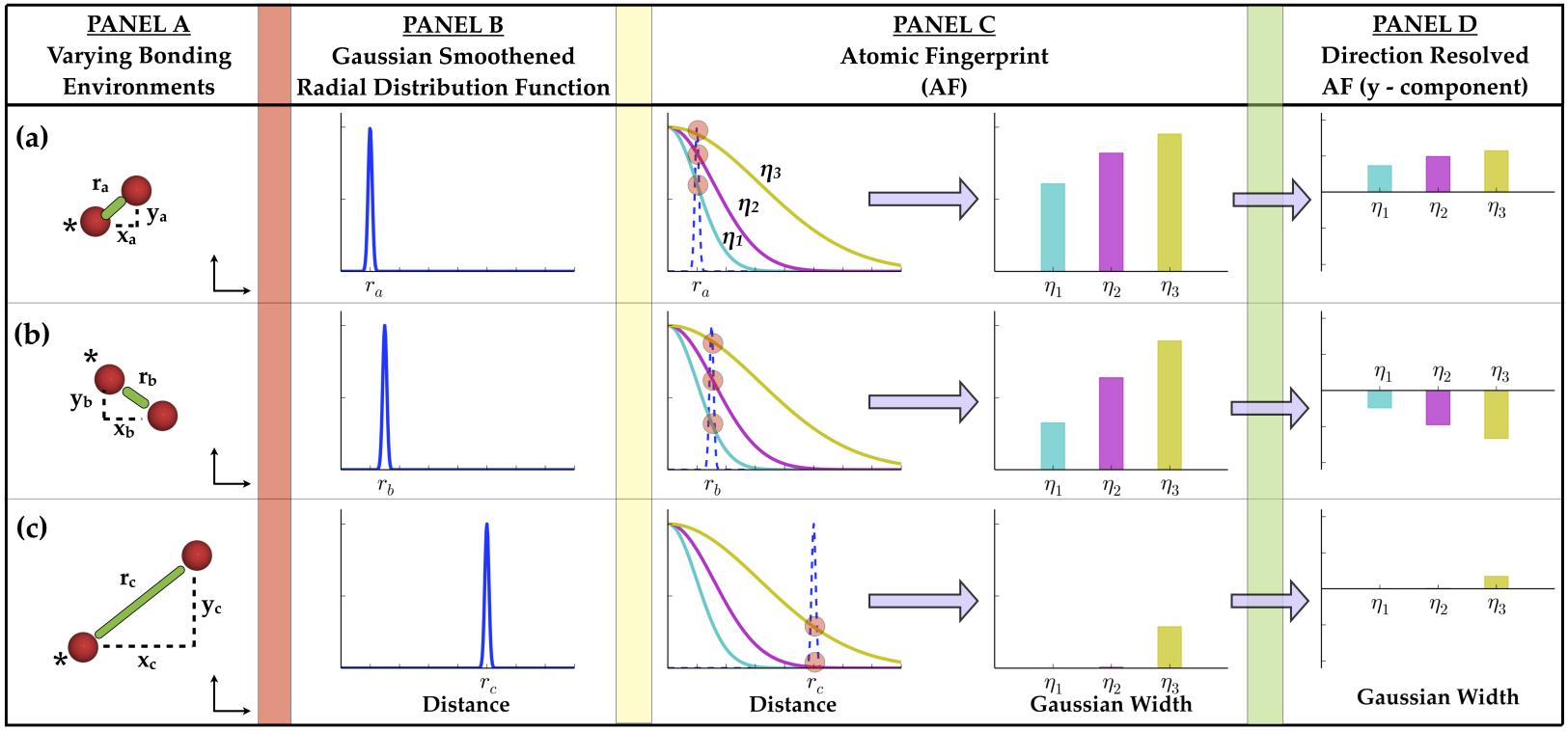

Figure 4.4: Panel A: A homonuclear diatomic molecule displaying three different bond lengths. Panel B: The corresponding Gaussian smoothened radial distribution function (RDF) for each of the bonding environments. Panel C: Transformation of the RDF using Gaussian functions on an eta-grid as indicated by the colored lines, into an atomic fingerprint. Panel D: The y-component of the direction resolved atomic fingerprint of an atom in the three bonding environments. The fingerprints generated are for the atom indicated by $\star$ in Panel A.

(dis)similarities are apparent to a human, it may not be so for a machine. Typical measures of (dis)similarity utilize the Euclidean norm of the difference between the fingerprint vectors or the dot product between the fingerprint vectors. Clearly, such measures will fail to capture the similarity between molecules a and b, and their dissimilarity with respect to molecule c (as the Euclidean norms of the difference between any pair of the three fingerprint vectors is the same constant value, and the dot products between any pair is zero).

Extending the RDF in a particular way can circumvent the above problem. Rather than using the RDF itself, a transformed quantity defined as the integral of the product of $R_i(r)$ and a Gaussian window function

$$G_i(\eta) = \int R_i(r)\ e^{-\left(\frac{r}{\eta}\right)^2}\, dr = \sum_{j \neq i} e^{-\left(\frac{r_{ij}}{\eta}\right)^2} \tag{4.2}$$

can be used, where η is a parameter that describes the extent of the window function. $G_i(\eta)$ is essentially a "cumulative" version of $R_i(r)$. This is visually demonstrated in Panel C of Figure 4.4, for three η values. While $R_i(r)$ is defined in a radial grid, $G_i(\eta)$ is defined in a η-grid. In order to account for the diminishing importance of atoms far away from the reference atom i, we multiply the summand of $G_i(\eta)$ by a cutoff function $f(r_{ij})$ that smoothly vanishes for large r_{ij} values, resulting in our choice of the atomic fingerprint (AF) function, $A_i(\eta)$, given by

$$A_i(\eta) = \sum_{j \neq i} e^{-\left(\frac{r_{ij}}{\eta}\right)^2} f(r_{ij}). \tag{4.3}$$

We note that $A_i(\eta)$ is essentially the radial symmetry function proposed earlier by Behler et al[167]. Following that previous work we define $f(r_{ij})$ as

$$f(r_{ij}) = \begin{cases} 0.5\left[\cos\left(\frac{\pi r_{ij}}{R_c}\right)+1\right] & \text{if } r_{ij} \leq R_c \\ 0 & \text{if } r_{ij} > R_c \end{cases} \tag{4.4}$$

where R_c is the cutoff radius, chosen here to be 8 Å. Interestingly, the η-grid does not have to be as fine as the radial grid. More importantly, $A_i(\eta)$ does not have the issues that $R_i(r)$ has, with respect to capturing the (dis)similarity between actual physical situations as defined by Euclidean norms. This can be ascertained by inspecting Panel C of Figure 4.4.

Finally, we consider the extension of the $A_i(\eta)$ definition so that it becomes applicable to represent vectorial atomic quantities such as forces. This can be simply done by *resolving* each term in the summation of $A_i(\eta)$ into directional components, leading to the direction-resolved atomic fingerprints, $\boldsymbol{V}_i^u(\eta)$ as follows

$$V_i^u(\eta) = \sum_{j \neq i} \frac{r_{ij}^u}{r_{ij}}\, e^{-\left(\frac{r_{ij}}{\eta}\right)^2} f(r_{ij}). \tag{4.5}$$

Here, r_{ij} is the distance between atoms i and j, while r_{ij}^u is a scalar projection of this distance along a direction $\hat{u}$ (c.f., Figure 4.5). Panel D of Figure 4.4 visually demonstrates the $V_i^y(\eta)$ function for the homonuclear diatomic molecular systems of Panel A. η is the Gaussian function width. $f(r_{ij}) = 0.5\left[\cos\left(\frac{\pi r_{ij}}{R_c}\right)+1\right]$, is a damping function for atoms within the cutoff distance (R_c), and is zero elsewhere. The summation in Eq. 4.5 runs over all neighboring atoms within an arbitrarily large R_c (8 Å, in the present work). To be able to construct the force on an atom, in 3-dimensional space, one requires any 3 non-parallel force components ($\hat{u}$). Depending on the direction chosen ($\hat{u}$) the force component along this direction (F^u) will vary, and the representation chosen should conform with directional changes in the individual force components as the local reference for the coordination changes.

The fingerprint described in Eq. 4.5 can be deconvoluted into three sub-components (as separated by the period symbol). The first term (Gaussian functions) describes coordination shells around an atom i, with η describing the extent of the shell. By using multiple such η values both nearby and distant coordination information is contained in the fingerprint. In this work, η's were sampled on a logarithmic grid between [0.8Å, 16Å], ensuring a sufficient description of the dominant nearest neighbor interactions. However, the optimal range remains to be system dependent. The middle term (a normalized scalar projection) introduces directionality to the fingerprint by selectively resolving the coordination information along the desired direction, $\hat{u}$. Lastly, the third term (a damping function) diminishes the influence of far away atoms smoothly. The combination of these three features makes this particular choice of representation suitable for mapping atomic force components. Similar such coordination based fingerprints were developed in the past, however, these were tailored for the purpose of mapping the total potential energy (a scalar quantity) for a given configuration of atoms, unlike the vectorial force components done here [167, 172].

Fingerprint properties: Invariance and Uniqueness

To demonstrate that the fingerprinting scheme proposed in Eq. 4.5 conforms to the basic invariance rules; translation, rotation and permutation of atoms, we refer the reader to Fig. 4.5. Each atom's position ($\mathbf{x}_i$) is defined in the $\mathbb{R}^3$-Euclidean space, with atom i as the reference. A translation operation ($\mathbf{t}$) on each atom shifts their positions as $\mathbf{x}_i' = \mathbf{x}_i + \mathbf{t}$. However, given that Eq. 4.5 only considers pair-wise distances, $r_{ij} = ||\mathbf{x}_i - \mathbf{x}_j|| = r_{ij}^{'}$, it does not alter the individual atomic fingerprint (V_i^u). Now consider rotating the sphere of atoms clock-wise, in Figure 4.5, about the z-axis by an angle θ (for the subsequent discussion, we drop the index i for the atomic force components and fingerprint). Upon rotation, both the force components and

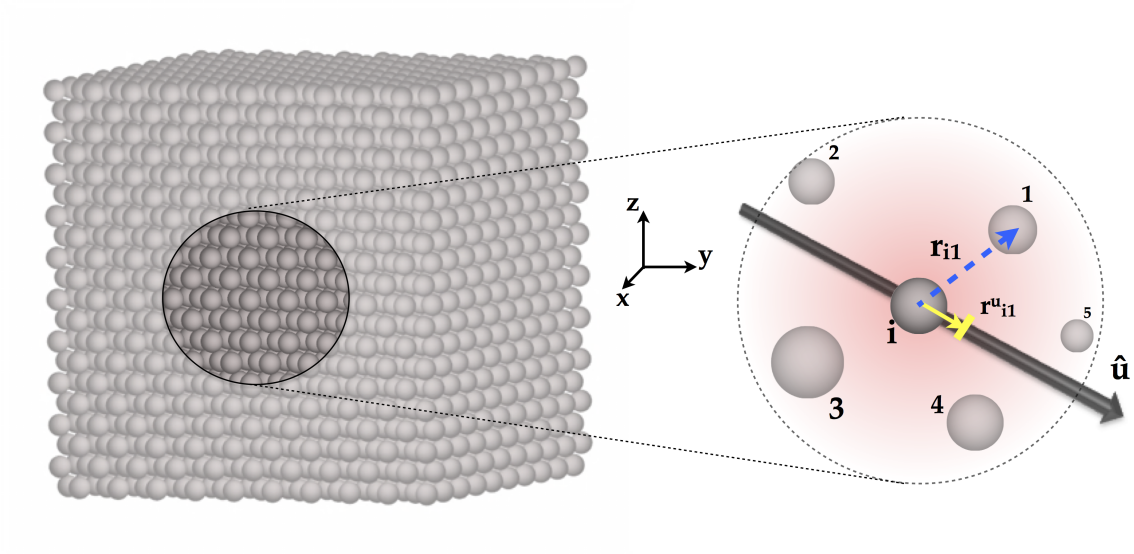

Figure 4.5: A schematic demonstrating the scalar projection for an atom (i is the reference atom) and one of its neighbor (atom 1) along a direction $\hat{u}$. To generate the final fingerprint for atom i, a summation over the atoms within the cutoff sphere, as indicated by the dashed line, are considered.

the corresponding fingerprints, along the Cartesian directions, change according to (shown here for forces but equally applicable for the fingerprint),

$$\begin{bmatrix} F^{x'} \\ F^{y'} \\ F^{z'} \end{bmatrix} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) & 0 \\ \sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} F^{x} \\ F^{y} \\ F^{z} \end{bmatrix}.$$

Nevertheless, the net force before (F) and after rotation ($F^{'}$) are identical, $F = F^{'} = \sqrt{(F^{x'})^2 + (F^{y'})^2 + (F^{z'})^2}$. As with the force components, the atomic fingerprints (V^u, $u \subset (x, y, z)$) change individually, but the net rotated quantity $V^{'} = \sqrt{(V^{x'})^2 + (V^{y'})^2 + (V^{z'})^2} = V$ is conserved, implying that the fingerprint transforms in a manner similar to the forces upon rotation. Lastly, permuting neighboring atoms only alters their indices, but given that the summand in Eq. 4.5 runs over all neighboring atoms within the cutoff sphere, the order of summation is unimportant

and doesn't alter V_i^u.

Another important aspect of the fingerprint is that it remains unique for the diverse set of environments considered. The premise of AGNI, as mentioned earlier, is that the force is governed primarily by an atom's local neighboring environment, as represented by its fingerprint. Numerically this implies that no two identical fingerprints should map to dissimilar forces. To identify the number of η values required to represent an atom, under which this hypothesis becomes true, we ask the question - *given an atom and its neighbors, can a different arrangement of the neighbors result in the same atomic fingerprint*? Consider two atomic arrangements, A and B with M and N atoms, respectively, with A being the reference. For a given η value, by freezing $N-1$ atom positions in B, the location of the free atom can varied such that fingerprint is non-unique. However, by using multiple η values, the probability of finding two high-dimensional identical fingerprints diminishes. In the limit that number of η values tends to ∞ we can ensure uniqueness. Nevertheless, for all practical purposes one can make do with a much smaller subset, as shall be seen shortly.

Using Eq. 4.5, we now compute a fingerprint along the Cartesian directions for each atomic environment within the database, as the DFT computed force components are along these directions. However, one is not restricted to these 3 directions only. By creating a spherical mesh around an atom, several arbitrary directions can be defined for which we reconstruct the atomic force and recompute the fingerprint. Adding these quantities to the pre-existing reference database expands upon the wealth of atomic environments, with no additional *ab initio* calculations. Though such an undertaking further ensures diversity in the reference database, it also builds in extensive redundancies. To train a force field on all the millions of atomic environments is impractical, computationally very demanding, and might lead to misbehaved models, therefore, further down-sampling from within this big pool of data

is an essential step in the construction process.

4.2.3 Clustering reference data

Visualizing the data

In order to select a few representative atomic environments for training, from within the millions, we first need to identify the redundancies that exist. Comparing amongst the fingerprints is an obvious place to start, however, given its high-dimensionality understanding or unraveling it directly is non-trivial. In order to handle the large quantities of data, better, we rely on dimensionality reduction techniques such as principal component analysis (PCA) to project V_i^u onto a lower dimension space [155]. In PCA, the original fingerprint components are linearly combined into uncorrelated and orthogonal pseudo variables, also known as principal components (PCs). Here, for all the reference atomic environments, we compute an 8-dimensional fingerprint (the rationale for which shall be discussed shortly) and transform them into PCs. Using the two relevant PCs (as majority of the variance, $> 99\%$ is captured by them) we then project the millions of transformed fingerprints onto this two-dimensional manifold known as a scores plot (c.f., Figure 4.6). Immediately, we observe clustering of different environment types; for clarity we labeled the atomic environments corresponding to a few cases, e.g. adatoms, surfaces, vacancies, etc., in Figure 4.6. Further, by color coding atoms according to the dataset they were sampled from, i.e. A, B, C, D or E, we observe the qualitative extent of their diversity. Dataset D (c.f., Figure 4.6) contains the most diverse set of atomic environments as it populates majority of the space, suggesting that isolated cluster configurations is a good starting point to sample reference data for AGNI construction. Interestingly, the atomic environments corresponding to dataset E also lie within the domain of dataset D.

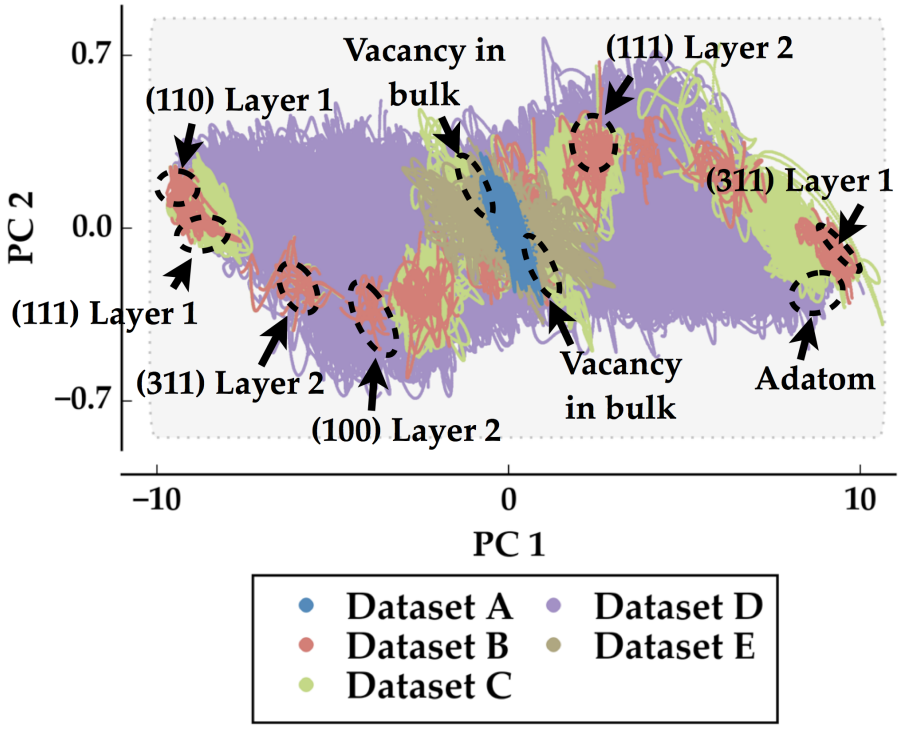

Figure 4.6: (a) Fraction of variance captured by the principal components (PCs) after a PCA projection of the atomic fingerprints. More than 99% of the variance is captured by the first two PCs. (b) A projection of the atomic fingerprints in dataset A, B, C and D, on the first two principal components. An 8-component fingerprint was used to represent each atom.

This suggests that a force field trained on dataset D should accurately predict the forces for environments in dataset E.

Selecting training and test data

The visualization tools described, thus far, provides a human-appealing method to identifying redundancies. To establish an expedited and efficient force field construction, we require automated sampling of the PCA transformed data to identify a smaller representative training dataset. An obvious first choice is to select data randomly. Unfortunately, this biases sampling according to the underlying probability distribution of the dataset and fails to sample sparsely populated regions. To avoid

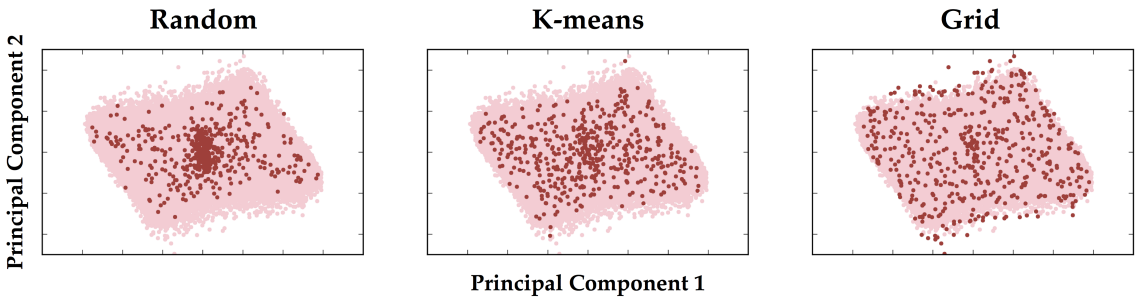

Figure 4.7: Sampling data on the PCA trasnformed data by three methods; randomly, k-means, and grid. The dark red points represent the chosen training data points, while the light red points indicate all the reference data.

such irregularities one could use k-means clustering methods to sparsify and identify a diverse set of atomic environments, as has been done in constructing machine learning force fields to predict total energies [175]. However, given the non-isotropic nature of our dataset k-means performs poorly. In this work, we adopt a simple grid-based sampling on the PCA space. Here, by splitting the clustered PCA data into uniform sub-grids, the bounds of which are determined by the minimum and maximum of the relevant PCs, data is then randomly sampled from within each sub-grid. By using a fine grid one can ensure uniform sampling from all regions of the PC space. In the limit the grid size becomes very large, this approach is equivalent to a random sampling approach. A pictorial comparison of the sampling methods is illustrated in Figure 4.7. Using such clustering methods to eliminate redundancies within the reference dataset, is necessary for tractable force field training as well as prediction, as the cost scales as $\mathcal{O}(n^3)$ and $\mathcal{O}(n)$, respectively (where n is the training dataset size). Lastly, the test sets, to validate the force field, are generated from the non-sampled training data.

4.2.4 Learning algorithm

The next vital ingredient required in putting together a predictive framework is the learning algorithm itself. Deep learning neural networks[52] and non-linear regression processes[53] have been the methods of choice for developing models to describe atomic interactions. Their capability to model highly non-linear relations, as is in the case of mapping an atom's environment to the force it experiences, makes them a suitable choice here. Here, we choose kernel ridge regression (KRR) as the machine learning workhorse [176, 166]. KRR works on the principle of (dis)similarity, wherein, by comparing an atom's fingerprint ($V_i^u(\eta)$) with a set of reference cases, an interpolative prediction of the u^{th} component of the force (F_i^u) can be made, and is given by

$$F_i^u = \sum_t^{N_t} \alpha_t \cdot \exp\left[-\frac{\left(d_{i,t}^u\right)^2}{2l^2}\right]. \tag{4.6}$$

Here, t labels each reference atomic environment, and $V_t^u(\eta)$ is its corresponding fingerprint. N_t is the total number of reference environments considered. $d_{i,t}^u = ||V_i^u(\eta) - V_t^u(\eta)||$, is the Euclidean distance between the two atomic fingerprints, though other distance metrics can be used. The weight coefficients $\boldsymbol{\alpha_t}$s and the length scale parameter l are determined during the training phase, whence the objective function $\sum_t (F_t^u - F_t^{u,\star})^2 \; + \; \lambda \left(\boldsymbol{\alpha_t}^T \boldsymbol{K} \boldsymbol{\alpha_t}\right)$ is minimized. $F_t^{u,\star}$ is the QM force value, $\boldsymbol{K}$ is the Gaussian kernel matrix of the cases within the training dataset, and λ is a regularization parameter that should be carefully chosen to avoid overfitting [49, 50, 31]. The parameters l and λ are determined by k-fold cross-validation (in this work k=5) on the training dataset. In this method, the training dataset is split into k bins. Each bin acts as a new test dataset, whilst the remaining k-1 bins are combined into a new training dataset. The process is repeated for every bin in the

k bins, and for every l and λ on a pre-selected logarithmically scaled fine grid. The optimal l and λ parameters (i.e., ones that lead to the lowest k-fold cross validation error) are then used in the final model development stage to determine the $\boldsymbol{\alpha_t}$ values for the entire training dataset, computed here as $\boldsymbol{\alpha_t} = (\boldsymbol{K} + \lambda \boldsymbol{I})^{-1} \boldsymbol{F_t^{u,\star}}$.

Finally, in order to evaluate the performance of a developed force field, three error metrics; mean absolute error (MAE), maximum absolute error (MAX), and the standard deviation (in particular 2σ), were used. Relying on multiple metrics reduces any bias, unknowingly, introduced during model selection as shall be discussed shortly.

4.2.5 Uncertainty quantification

The final component, to a successful predictive model, is to be able to quantify uncertainty in the predictions made. For instance, *given a force field can we estimate the prediction error on the atomic forces for a new observation*? Finding solutions to such questions, will help enable one to understand the domain of applicability of their respective models. As with any statistical model, the true value of the force ($F^\star$) for a given atom can be expressed as the predicted value (F) with some error (ε), given as

$$F^\star = F + \varepsilon. \tag{4.7}$$

If ε can be statistically estimated we then have a pathway to provide confidence bounds on the predicted atomic force.

In the KRR framework, for every new observation the distance, d_t (for brevity we only label the reference environment and dropped the observation and direction label, i and u, respectively), between its fingerprint and the reference training fingerprints is computed (resulting in a total of N_t distances). The final prediction is

then a weighted sum of the list of distances, making these distances an important metric on predictive accuracy. Amongst the list of distances, the minimum distance, $d_{min} = min\{d_1, d_2, ..., d_{N_t}\}$, in particular provides a measure of *closeness* of the new observation to the training cases, and can be thought of as a descriptor in estimating ε. To capture the correlation between d_{min} and ε, for every observation in the test dataset we compute d_{min} and ε $(= F^{\star} - F)$. By binning the range of d_{min}s observed into uniform and smaller sub-groups, a standard normal distribution function is fit to the observed ε. For each sub-group, collecting standard deviation (s) statistics of the normal distribution, ultimately, provides an estimate of ε as a function of d_{min} (with a confidence level of 68.2%, though higher confidence levels can be equally implemented).

4.2.6 Constructing the force field

At this stage we have laid out all the pieces required to construct AGNI force fields, as illustrated by the flowchart in Figure 4.2. Subsequently, we demonstrate the construction and validation of one such force field that is accurate and generalizable.

Convergence tests

The first step to attaining an optimal force field is to ensure convergence with respect two parameters: (i) the number of η values (or Gaussian window functions) used for the atomic fingerprint, and (ii) the training dataset size. As mentioned earlier the number of η values governs the resolution with which an atom's local environment is described, while, the size (and choice, as shall be elaborated in the next section) of training data governs AGNI's interpolative predictive capability. In order to identify this optimal parameter set, we systematically increase the fingerprint resolution from

2 to 16 η values and the training dataset size from 100 to 2000 atomic environments, while monitoring test set error (as measured by MAE). To remind the reader, η values were sampled on a logarithmic grid between [0.8Å, 16Å], while training data environments were sampled using the PCA projection followed by a grid-based sampling.

For the four training datasets (A, B, C and D), and for all combinations of the parameters, an AGNI force field was constructed and validated on the respective test datasets. The force fields are denoted as M_i^j, where i and j label the training and test environments used, respectively (the superscript is omitted when referring to the training environments only). In Figure 4.8 we illustrate heat maps of the test set error for all dataset and parameter combinations. Two key findings stand out: (i) by increasing the fingerprint resolution the error drops and quickly converges below ≈ 0.05 eV/Å (expected chemical accuracy), and (ii) increasing the training dataset size reduces error only beyond a reasonable fingerprint resolution. For example, in M_C^A increasing the training dataset size for a fingerprint with 2 or 4 η values has no effect on the predictive capability. Such a manifestation implies that 8 or more η values are required to "uniquely" discern amongst the atomic environments, in order for the learning algorithm to work. Nevertheless, this relation only holds for force fields used in an interpolative manner, as seen in the failure of M_A^B, M_A^C, M_A^D or M_A^E. Here, the diversity in the training data chosen plays a more prominent role in governing performance, as shall be elaborated in the next section. Overall, we find that a fingerprint of 8 η values and a training size of 1000 atomic environments is sufficient, beyond which the models exhibit diminishing returns, i.e. increased model training costs with no significant drop in model error, and are the parameters chosen for all subsequent discussions.

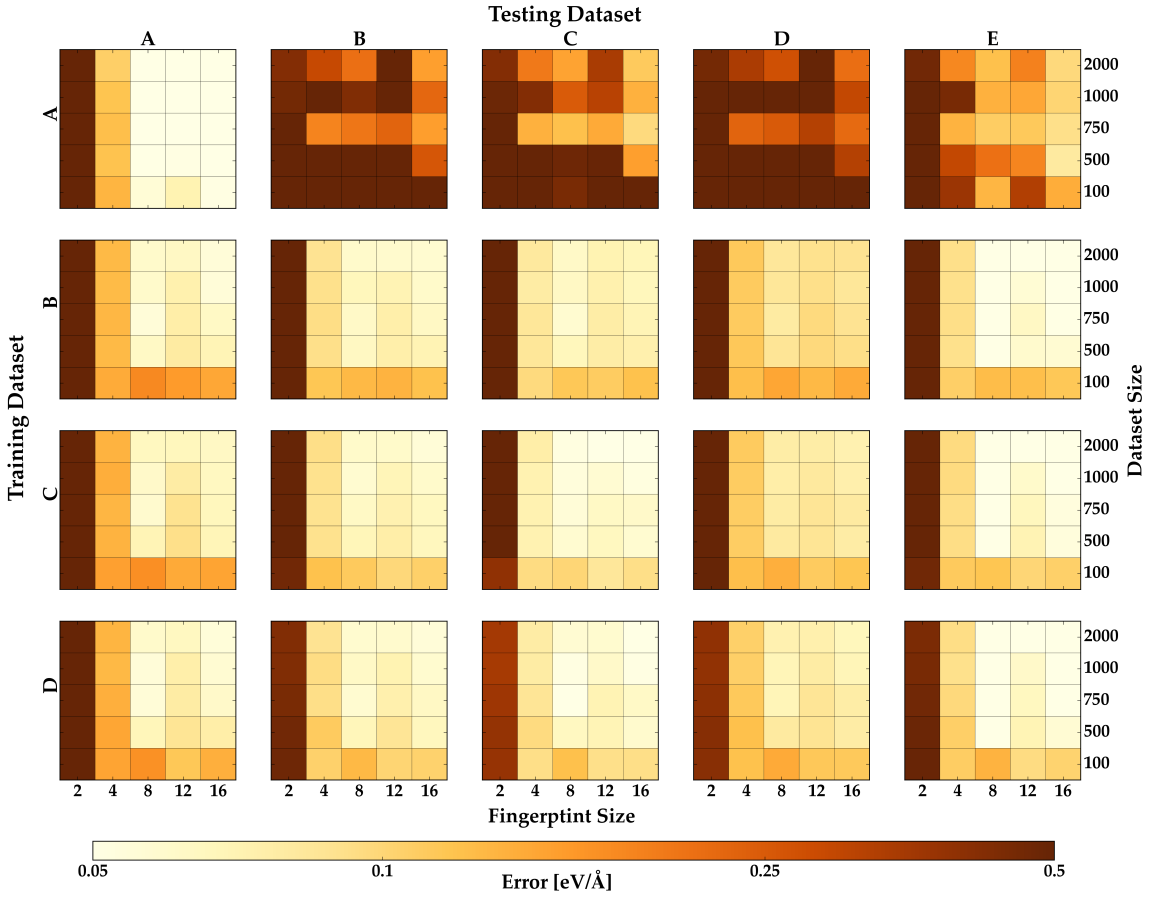

Figure 4.8: Heat maps illustrating model error (mean absolute error) as a function of fingerprint resolution and training dataset size. The fingerprint was varied from 2 to 16 η values, while the training dataset size was varied from 100 to 2000 environments. We report the error for models trained on each of the four datasets, and consecutively tested on all the test datasets. For example the top row corresponds to models trained on dataset A, while each column corresponds to a test datasets of the five cases. The errors quickly converge for a fingerprint with 8 η values and a training size of 1000 diverse environments.

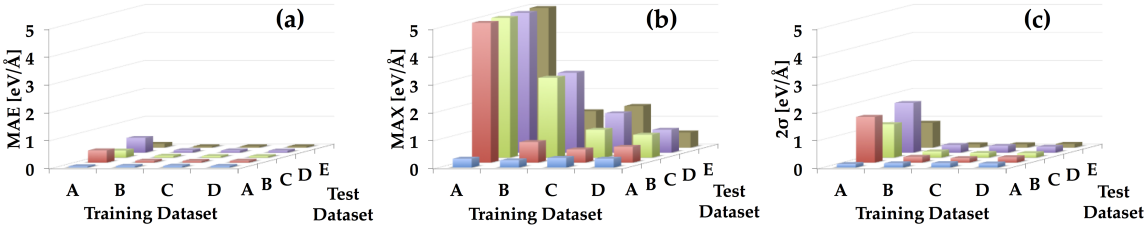

Figure 4.9: (a) Mean absolute error, (b) maximum absolute error, and (c) 2 * standard deviation error metric, for models trained on A, B, C and D and tested on dataset A, B, C, D and E. Here we use an 8-component fingerprint and a training set size of 1000 environments obtained with the PCA grid-based sampling.

Training data choice

The choice of atomic environments used for training is a crucial factor, as briefly alluded to in the previous section. Given that the learning algorithm is interpolative by nature, a force field trained say only on bulk type environments (M_A) cannot predict the forces corresponding to other environments types, e.g. datasets with surfaces and other features - M_A^B, M_A^C, M_A^D or M_A^E. By increasing the diversity in training environments, M_B, M_C and M_D, we make the force fields more generalizable once the optimal parameters are chosen, as given by their low test error in Figure 4.8. Surprisingly, it appears as though predictions made with M_B are equally as good as M_C or M_D. However, this is purely a manifestation of using the MAE as the error metric. Along with the MAE, we report test set errors computed with two other metrics - MAX and 2σ, as illustrated in Figure 4.9 (shown only for the optimal 8-component fingerprint and 1000 training atomic environments). For M_B, with MAX as the metric, the prediction error is high outside its domain of applicability (test set C, D or E), and a similar behavior is observed for M_C. It should be recognized that MAX reports the worst prediction made, while MAE reports a mean error skewed by test set size. By combining the two metrics with the actual variance in the errors,

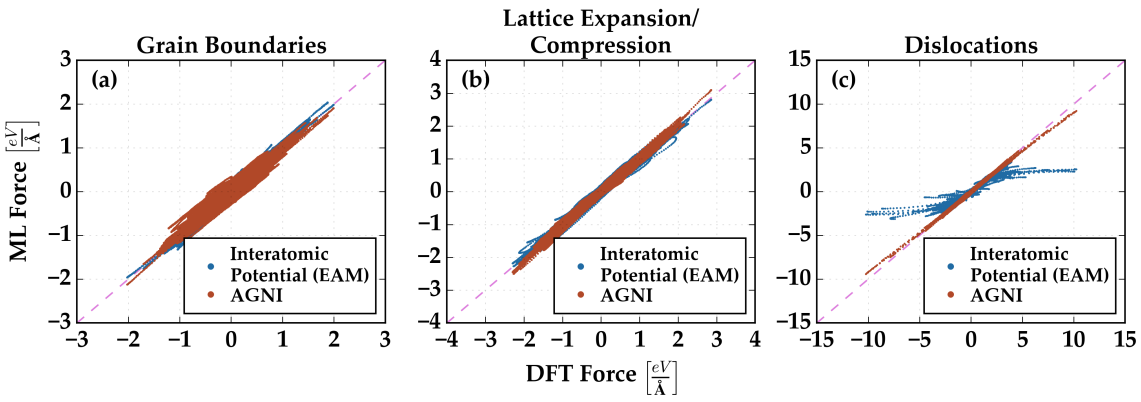

Figure 4.10: (a) A projection of the atomic fingerprints in validation configurations compared to the training data used in AGNI model, M_D. (b-d) Parity plots comparing the error in force prediction with the AGNI model, M_D, and the EAM interatomic potential with respect to DFT for the validation configurations, grain boundaries, lattice expansion/compression and dislocation, respectively.

as measured by the 2σ metric, we can ensure that the error is indeed under control. We observe that in M_D, by sampling atomic environments from a very diverse set of configurations all the error metrics are low, and the force field is highly generalizable, and is the force field used in subsequent discussions.

4.2.7 Validating the force field

To further validate that the developed force field, M_D, is indeed generalizable, we test its predictive limits for atomic environments in dataset E. Clearly, the configurations contained; grain boundaries, lattice expansion and compression, and dislocations, were never "observed" during the training phase. Being able to accurately predict the forces will further demonstrate the fidelity in using local-neighborhood based AGNI force fields.

The PCA scores plot, shown in Figure 4.6, provided a glimpse of what one could expect. Given that the transformed atomic fingerprints for dataset E lies within the domain of environments from dataset D, it qualitatively confirms that predic-

tions made by M_D are interpolative and thus should be accurate. However, a more stringent test is to predict forces on all the atoms in dataset E and compare them to those obtained by DFT methods. As is done and shown in Figure 4.10. Each AGNI prediction costs $\approx$ 0.1ms/atom/core, while DFT costs $\approx$ 1ks/atom/core. For all three cases, the AGNI predicted forces are in excellent agreement with DFT. This demonstrates, for the first time, the intended goal of AGNI force fields, i.e. to retain quantum mechanical accuracy, be computationally inexpensive, and remain generalizable. The last feature in particular, generalizability, is often lacking with traditional semi-empirical methods. For comparison, we recompute the forces for the atoms in dataset E using traditional semi-empirical potentials. Here, we particularly use an Al EAM potential[177], as it accurately captures interactions in close-packed metallic type systems. As with AGNI force field, EAM methods equally predict forces accurately for grain boundaries and lattice expansion/compression but fails for dislocation type of environments. This once again raises an important question in the realm of force field based simulations - *can one a priori judge the error in the forces predicted*? In the next section we provide one such attempt at estimating uncertainties in the force predictions made with AGNI.

4.2.8 Quantifying uncertainty with force field

Quantifying uncertainties is a challenging task that can at best be done probabilistically. Here, using such a framework laid out earlier, we discuss how one can generate uncertainty estimates for force predictions. Using M_D as the force field, for each test environment (in dataset A, B, C, D and E) we compute ε of the predicted force, as given in Eq. 4.7, and the corresponding d_{min}, i.e. minimum distance within the training dataset. The results are summarized in the scatter plot of Figure 4.11. Clearly,

as d_{min} increases, the variance within ε rises for a particular choice of d_{min}. Upon binning the data into smaller sub-groups as indicated by the dashed lines, a standard normal distribution function is fit. The histogram insets in Figure 4.11 demonstrate this for three such bins. $d_{min} < 10^{-3}$ were ignored as the data was too sparse for statistical interpretation. While, a cutoff $d_{min} > 10^{-1}$ was imposed, beyond which the variance in the force predictions is too high. These regions are indicated by the gray-filled rectangles in Figure 4.11. Access to the standard deviation (s) for each bin, allows us to provide a confidence estimate for the predictions made as a function of d_{min}. Upon plotting d_{min} and s for all the bins (red circle markers), as shown in Figure 4.12, we observe a trend, wherein, as d_{min} increases the uncertainty in the predicted force increases. At low d_{min}, a polynomial behavior with s is observed ($s = 49.1d_{min}^2 - 0.9d_{min} + 0.05$, as shown by the dashed blue line in Figure 4.12).

The proposed polynomial relation serves as a rudimentary estimate to the upper bound of prediction error. The relation in Figure 4.12 is symbolic of a typical interpolative model, whereby, if training data exists in the vicinity of a new observation confident predictions can be made. It is for these reasons; we employ diversification and filtering techniques to ensure that the model spans a diverse environment space uniformly in order to make reasonably accurate predictions. By quantifying uncertainty, it allows one to identify those atomic environments likely to result in high prediction errors. By flagging and accumulating such environments one can systematically retrain the force fields, resulting in an adaptive refinement of accuracy and generalizability over time.

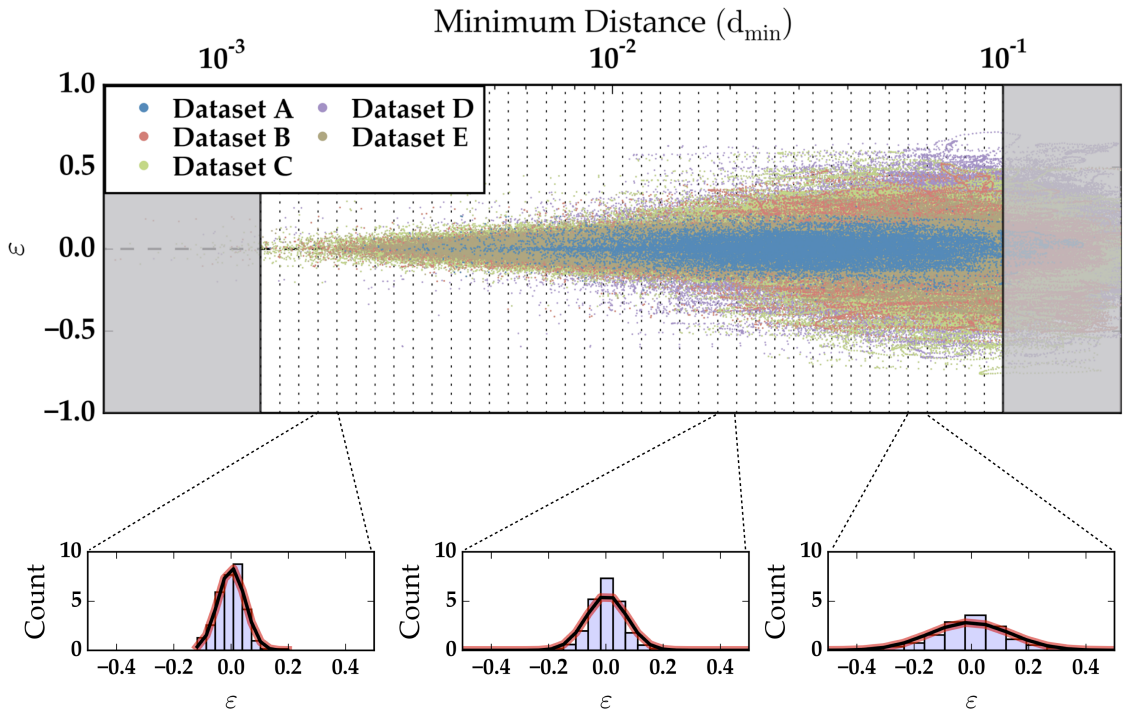

Figure 4.11: Top panel: a scatter plot of the minimum distance (d_{min}) vs. the predicted force error (ε). The range of d_{min} is further sub-divided into small groups for statistical analysis. The gray regions were not considered for any statistical purposes, due to the lack of sufficient data (left) and high errors (right). Bottom panel: a standard normal distribution fit for each sub-group (though only shown for three such bins), used to estimate the variance in model errors.

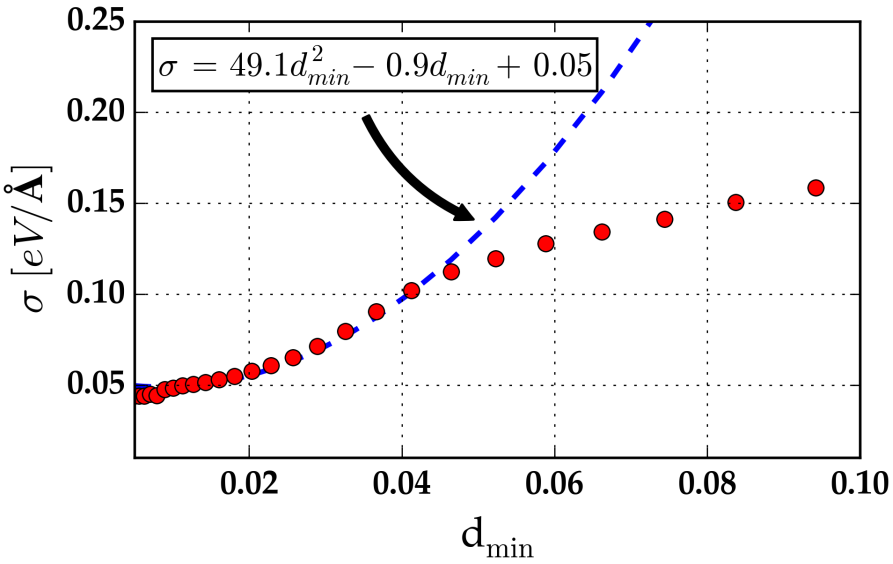

Figure 4.12: The uncertainty model, created for force field M_D, whereby d_{min} is used as a descriptor to measure the expected variance in the prediction made. The markers show the actual behavior, while the blue dashed line indicates a polynomial fit to the uncertainty.

4.3 Accelerating materials simulations

For many atomistic simulations, e.g. molecular dynamics, geometry optimization, identifying reaction barriers, materials properties, etc., determining the force on an atom is key. The AGNI framework prescribed in the previous section now provides us with a pathway to directly and quickly compute forces, at DFT accuracy. Here, using the above prescribed framework, a simple (not as elaborate in the previous section) AGNI force field for Al was developed using a smaller plethora of reference atomic environments accumulated from density functional theory (DFT) data. The subset of reference cases considered were sampled from defect-free bulk in the face centered cubic (fcc) phase, bulk fcc phase with vacancy, clean (111) surface, and the (111) surface with adatom, resulting in over 100,000 atomic environments [56]. Interestingly, a random sampling of of just 1000 atomic environments drawn from the accumulated

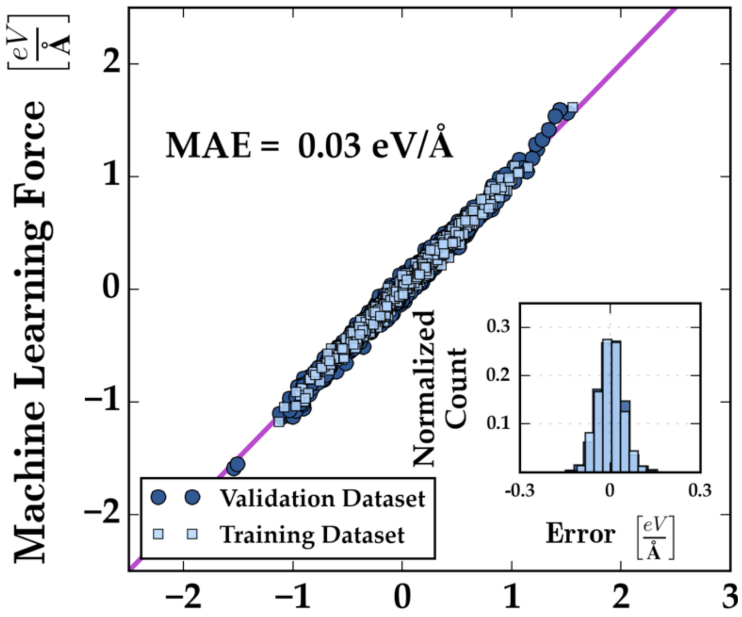

Figure 4.13: Comparison of the forces predicted using the ML force field with reference DFT results, for the training (light blue) and the validation dataset (dark blue) used.

environments proved sufficient to construct an accurate interpolative force prediction model. Figure 4.13 compares the predicted forces with the DFT forces (including the error distribution in the inset) for all accumulated configurations, i.e., those used in the training phase and the remaining configurations whose results were used for validation. The mean absolute error (MAE) of the force field was 0.03 eV/Å, of the order of the expected chemical and numerical accuracy of the reference DFT calculations. Furthermore, this procedure to predict atomic forces is also extremely expedient; it scales linearly with system size, and can be well over 8 orders of magnitude faster than a typical DFT calculation. Each force evaluation required less than 1 ms/atom in a single-core computer, comparable to typical molecular mechanics methods.

4.3.1 Molecular dynamics

As a first step towards validating this new approach, we consider non-zero temperature dynamical situations. For the force prescription to correctly capture dynamic processes with high-fidelity, ergodicity has to be preserved. In other words, the average behavior and time scales of elementary steps or processes should be correctly represented during a MD simulation using the force field. As a first example, we consider the self-diffusion of an Al adatom on an Al(111) surface, using a 6 Å × 6 Å surface unit cell containing a 4-layer thick Al slab. MD simulations were performed at 6 temperatures in the 50-300 K range for times up to 1 ns, with a timestep of 0.5 fs. By observing the dynamics of the vacancy, the average rate constant (k) for the migration process at each temperature was determined. k is given as $^{1}/t_{hop}$, where t_{hop} is the average time taken for an adatom to migrate to a neighboring site. To ensure that sufficient statistics are collected, k was averaged over 50 such hop events at each temperature. Figure 4.14(a) shows an Arrhenius plot of k versus the reciprocal temperature, whose slope yields the activation energy (E_a) for Al vacancy migration to be 0.03 eV. The corresponding DFT value for a similar, but static, migration process was determined to be 0.04 eV. Barrier "softening" is expected under dynamical conditions, relative to the results of static calculations in which entropic effects are neglected [178, 179].

Another elementary process we considered was the diffusion of a vacancy in bulk Al, using a unit cell containing 32 Al sites and an Al vacancy. Similar to the Al adatom example, by monitoring the dynamics of the adatom across a temperature range of 500-900 K, an E_a of 0.49 eV was predicted, as shown in Figure 4.14(b), whilst a static DFT calculation yielded an E_a of 0.59 eV. Both dynamical diffusion scenarios considered lead to the correct Arrhenius behavior indicating that the underlying

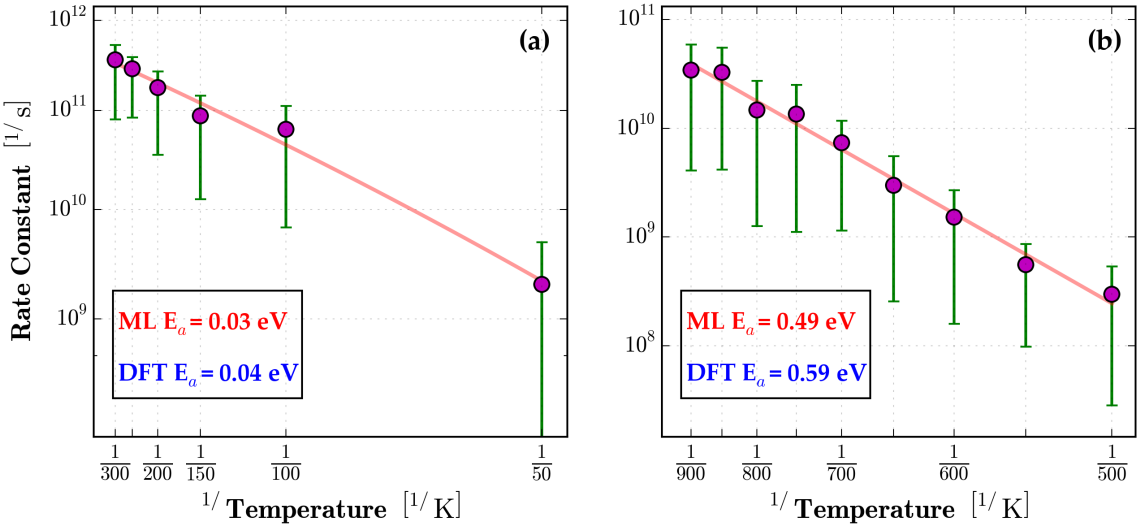

Figure 4.14: Arrhenius plots for (a) adatom diffusion on the Al (111) surface and (b) vacancy migration in bulk Al. For each temperature, the MD simulation time was extended so as to allow at least 50 hopping events (thus allowing estimation of an average hop rate, and the indicated error bar). A linear fit (solid red line) was used to determine the dynamic activation energy (ML E_a), and is compared with the static DFT activation energy (DFT E_a).

physics is properly captured in the ML force field based MD simulations. Also, the second example is a demonstration of the generalizability of such AGNI force fields to situations beyond surface chemistry.

4.3.2 Geometry optimization

Another immediate (and straight-forward) application of this fast high-fidelity capability to predict atomic forces is geometry optimization, including the prediction of potential energy minima and saddle points. Simulations involving hundreds of thousands of atoms (i.e., cases that are beyond the reaches of present day DFT computations) can be handled, provided the chemical environments encountered during the course of such optimizations are included in the force field. In order to understand the limits of the constructed ML force field for Al within the context of such

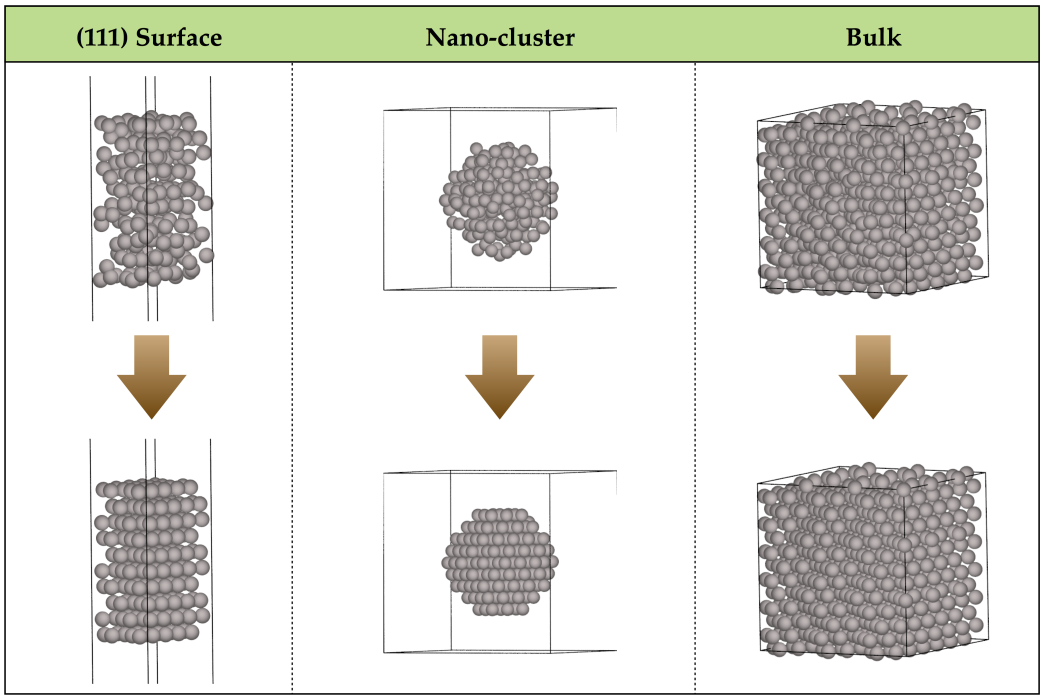

Figure 4.15: Geometry optimization for (a) a 111 surface, (b) an isolated nano-cluster, and (c) a bulk face-centered structures of Al. The top panel shows the intial perturbed state, while the later optimized structures are shown in the bottom. Atoms were perturbed randomly by a maximum of 0.3Å.

simulations, a few tests were performed. Starting from randomly perturbed atoms in several configurational arrangements; surface, clusters, bulk, the ML force field was used to optimize this perturbed structure. The correct equilibrium geometry was recovered, as ascertained by a separate DFT calculation starting with the same perturbed system (c.f., Figure 4.15). Although we restrict ourselves to modest sizes in this discussion (as we are constrained by the inability of DFT to provide validation for truly large unit cells), this example demonstrates that the force field is transferable to much larger systems, thus going significantly beyond previous efforts [56, 169].

4.3.3 Computing energy via force integration

The force field is aimed at predicting atomic forces. Nevertheless, in some situations access to the potential energy is a useful capability to possess. Firstly, it allows for a direct comparison with conventional semi-empirical approaches, all of which are energy based, but more importantly, it provides a pathway to validate individual elementary processes that often make up surface chemistry phenomena. The potential energy can be represented as a Taylor approximation,

$$E = E_o + \frac{dE}{dr}(r - r_o) + \frac{1}{2!}\frac{d^2E}{dr^2}(r - r_o)^2 + ... \tag{4.8}$$

Here, E is the potential energy, and r is the atomic coordinate in 1-dimension. The first derivative of the energy with respect to the atomic position is equal to negative of the force, as mentioned earlier. Thereby one can directly use the forces on atoms to recover the E. Nevertheless, to carry out this closed integration accurately, a pathway that connects the different phase space must exist (commonly known as the reaction coordinate). Therefore, by simply choosing any two arbitrary points in the phase space, one cannot predict the energy. Also, as with the time integrators in MD, the discretization of this reaction coordinate governs the accuracy by which we can predict the energy.

For instance, Figure 4.16 portrays pathway and the DFT energy profile for the rotation of an Al dimer on an Al(111) surface, as shown by the red atoms. Upon using Eq. 4.8 to integrate the forces predicted by the AGNI force field, we observe that the corresponding energy is in close agreement with the reference DFT method. This indicates that energies corresponding to critical parts of a trajectory may indeed be obtained from the forces through integration. More importantly, this demonstration

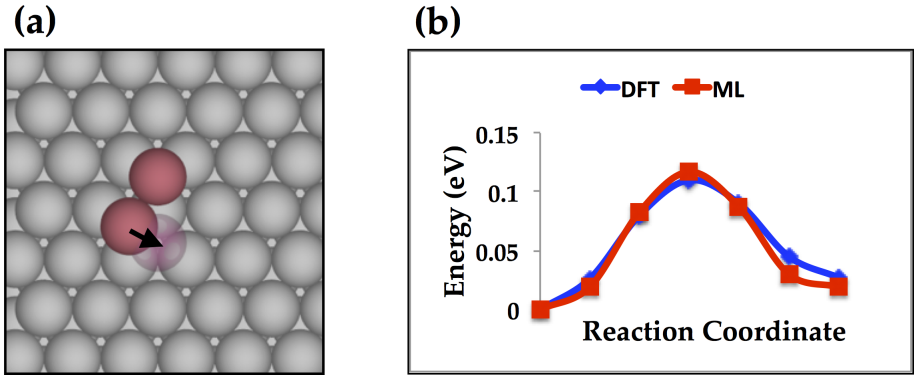

Figure 4.16: (a) A pictorial representation of dimer rotation on an Al(111) surface. The picture shows a top view of the (111) surface. The grey atoms correspond to surface Al, while the red atoms are adatoms. The shaded red atom indicates the final location of the dimer after rotation. (b) The potential energy computed via DFT (blue line) and the potential energy recomputed via integrating the forces predicted by an AGNI model, using Eq. 4.8

places the force prediction scheme in a formally solid framework as the predicted ML forces are shown to be consistent with the underlying potential energy.

4.3.4 Thermal properties

Lastly, we further expand upon the toolkit and evaluate the prospect of how well thermal behavior of materials can be simulated using the force-based framework. In particular, we focus on the vibrational (or phonon) density of states (DOS), which has to be properly captured to allow for accurate calculations of thermodynamic quantities, thermal expansion, thermal conductivity, etc. Figure 4.17(a) shows the phonon band structure as determined using the ML force field and using DFT, and in both cases, the finite displacement method was used [180]. Figure 4.17(b) shows the corresponding DOS, as well as the DOS computed using the Fourier transform of the velocity autocorrelation obtained from a MD simulation [181]. This latter approach implicitly includes anharmonicity to all orders (the first method, in contrast, includes

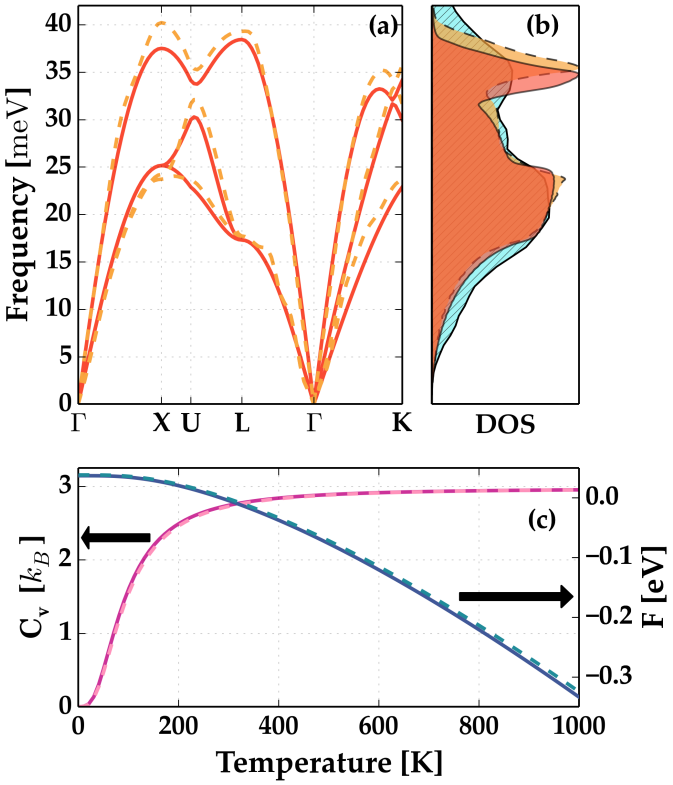

Figure 4.17: (a) Phonon band structure, (b) phonon density of states (DOS), and (c) Helmholtz free energy and constant volume heat capacity computed using the ML force field (solid lines) and DFT (dashed lines). The phonon band structure and DOS were computed using the finite atomic displacement method. Also included in (b) are the DOS results obtained from the Fourier transform of the velocity autocorrelation function (solid cyan hatched fill).

just the harmonic part). The MD simulation involved a 864 atom unit cell, and a simulation time of 5 ps at 700 K. Excellent agreement of the ML force field result with the reference DFT calculations can be seen. The deviations of the DOS computed using MD simulations relative to that obtained using the finite displacement scheme (especially at high frequencies) may be attributed to non-zero anharmonic effects. The DOS can be utilized to determine thermodynamic properties such as the Helmholtz free energy and the constant volume heat capacity. These properties, as a function of temperature, are compared with the corresponding DFT results in Figure 4.17(c). The ML force field and DFT results are nearly indistinguishable, indicating that even under the stringent test of small atomic perturbations encountered in these situations (as opposed to the larger length scale vacancy or adatom hops discussed earlier), the fidelity of the force prediction is preserved.

4.4 Island ripening on an Al(111) surface

As the fabrication of materials continually progresses towards the atomic-scale, an interest in layer by layer growth methods (such as molecular-beam epitaxy or atomic layer deposition) has risen. The high degree of control offered, allows for sub-nanometer scale precision in the morphological structure of the materials grown. Such precision has allowed tuning the chemical, electronic or mechanical properties of materials for use in micro-electronics, catalysis, and biomedical applications [1, 4, 165]. Consequently, the need to better characterize and understand such growth processes has risen. Towards this cause, *in silico* models have been instrumental in helping unravel the complex atomistic growth phenomena. Methods such as first-principles (*ab initio*) based density functional theory (DFT), with harmonic transition state theory, have commonly been used to map out all the constitutive elementary reaction

pathway energetics. By doing so, coarser stochastic approaches (e.g. kinetic Monte Carlo) could then be used to spatially and temporally evolve the state of a system well beyond the confines of a truly first-principles study. Nevertheless, building an a priori complete catalog of reaction pathways is often challenging and non-trivial for low symmetry systems. An alternative, and more natural, formalism is to use molecular dynamics (MD) simulations, whereby the temporal state of an atomistic system is evolved by solving Newton's equations of motion.

Here, by using AGNI force fields a demonstration of long time scale phenomenon is explored. In particular, we study the ripening of adatoms, a common feature that occurs during course of growth process, for the case of Al(111) surface. This goes well beyond exploring the dynamical behavior of a single adatom, as discussed in the previous section. The (111) surface was chosen in particular as the barriers for the elementary processes allow for the ripening phenomena to be explored by time scales achievable with conventional MD simulations. To do so, we start by briefly reviewing the construction of AGNI force fields. We then validate the force field by computing elementary reaction barriers often encountered on the surfaces. Using this force field, we then perform long time scale dynamic simulations exploring the ripening process, as it happens, as a function of time, temperature and adatom coverage.

The AGNI force field, was once again constructed using the procedure laid out earlier in the chapter. A plethora of reference atomic environments and their corresponding forces were generated from *ab initio* based MD runs, using the Vienna ab initio simulation package, at 700 K. To ensure a diverse set of reference cases, Al in different geometrical arrangements was considered(but each one with just a few tens of atoms per repeating unit cell), including environments from defect-free bulk in the face centered cubic (fcc) phase, clean surfaces, bulk and surfaces with defects such as vacancies and adatoms, and isolated clusters, resulting in over a million atomic

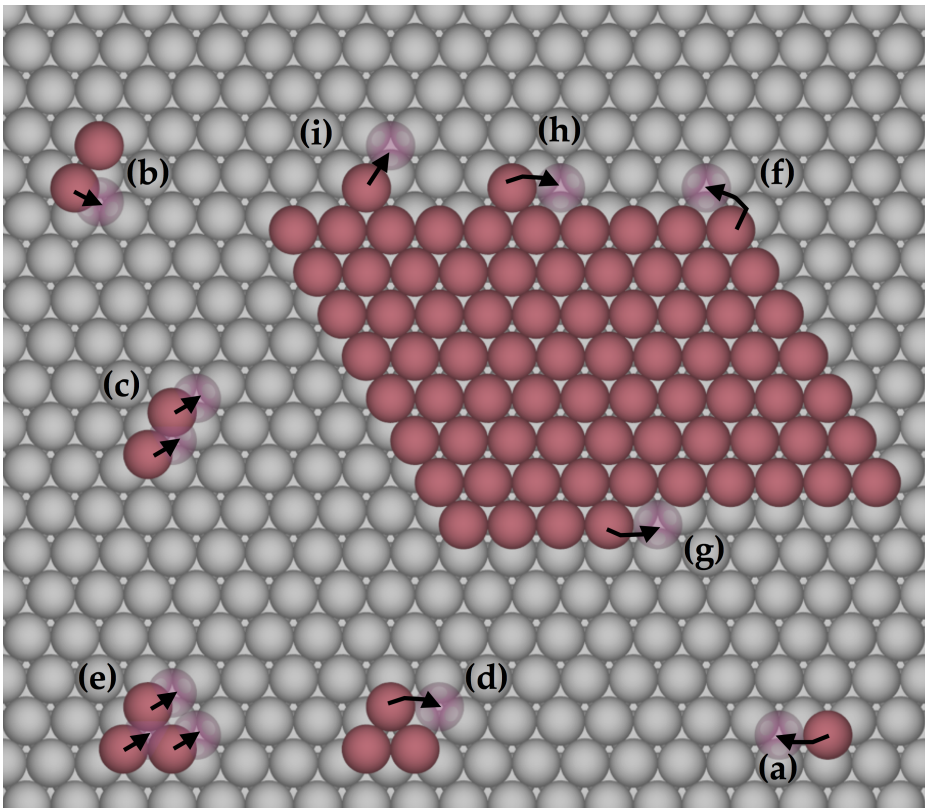

Figure 4.18: Elementary reaction pathways of monomers, dimers, trimers, and island features on the Al(111) surface that leads to ripening phenomena. Grey and red colored atoms correspond to the surface atoms and adatoms, respectively.

environments. Using a non-linear kernel ridge regression learning framework, to establish a mapping between the fingerprint and force, along with standard machine learning practices, we construct an AGNI force field with mean absolute prediction errors (MAE) $\approx$ 0.05 eV/Å. Of the order of the expected chemical and numerical accuracy of the reference DFT calculations.

4.4.1 Validating elementary processes

A first step to the realization of this ripening process is to ensure that the elementary reactions occurring on the surface, such as translation, rotation and diffusion of adspecies; monomer, dimer, trimer and beyond, as well as processes corresponding to re-arrangement of islands, such as corner breaking, kink breaking, terrace diffusion, edge evaporation etc., as illustrated in Figure 4.18, are correctly described. Here, we compute the reaction barriers for all pathways, (a)-(i), shown in Figure 4.18 using AGNI force fields (E_a^{AGNI}). The barriers are reported in Table 4.2. Given that AGNI force fields provide access to the forces only, energies are computed via thermodynamic integration of the forces. For comparison we report the corresponding DFT computed reaction barriers (E_a^{DFT}), using the climbing-image nudge elastic band method, and those reported in literature [182, 183]. On average the errors are within 5% of the DFT computed values. The correct energetics ensures true ergodic sampling of the different states with time.

Having demonstrated an accurate description of the elementary processes with the AGNI force field, we can be confident on exploring the long time ripening behavior of adatoms on the Al(111) surface. To do so, an asymmetric 35 $\times$ 30 Å^2 surface with 6 layers in the z-direction is constructed. Adatoms are randomly distributed on the surface, and their concentration is defined by a coverage (θ), given as the ratio of

adatoms on the surface to the maximum acceptable number (in a single layer). The dynamic simulations were performed in the canonical ensemble, with a timestep of 0.5 fs, using the popular LAMMPS MD code.

Table 4.2: Activation barrier for reaction pathways plotted in Figure 4.18. The activation barriers computed by AGNI (E_a^{AGNI}) were done so by integrating the forces, while DFT barriers (E_a^{DFT}) were computed using the climbing-image nudge elastic band method. Values indicated with $\star$ are literature reported DFT values.

Pathway	E_a^{AGNI}	E_a^{DFT}
(a) Monomer hopping	0.05	0.04
(b) Dimer rotation	0.12	0.11
(c) Dimer translation	0.13	0.07
(d) Trimer translation	0.19	0.21
(e) Trimer rotation	0.19	0.24
(f) Corner evaporation	0.71	0.60*
(g) Kink evaporation	0.67	0.65*
(h) Edge diffusion	0.48	0.45*
(i) Edge evaporation	0.91	0.80*

4.4.2 Role of time

To start with, we explore the temporal evolution of a system with $\theta = 0.14$ at 300 K. Snapshots during the course of this dynamic simulation are illustrated in Figure 4.19, up to a few nanoseconds. The randomly distributed adatoms quickly (in a few picoseconds) nucleate into small islands with a critical size of $\approx$ 4-5 atoms, and once formed remain intact consistent with past theoretical studies [184, 183, 185, 186]. Also, as the island size increases its mobility decreases as the diffusion process is now primarily governed by concerted displacements, which requires multiple bonds to be broken simultaneously. This increases the time required to observe any relevant diffusion. Nevertheless, at 300 K the thermal energy is sufficient to overcome these barriers and the individual clusters ripen to form an island after 2 ns.

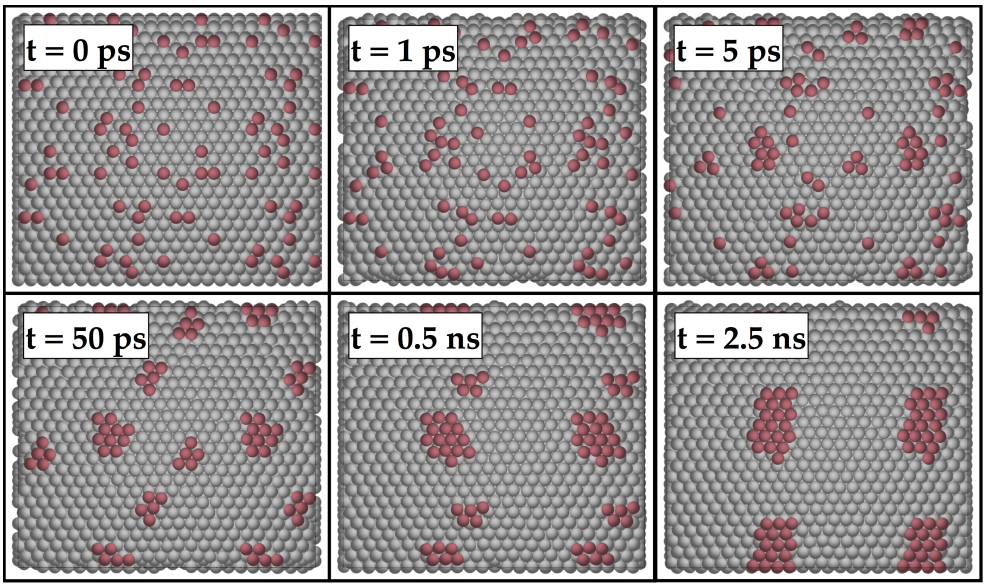

Figure 4.19: Snapshots of the time-evolution of adatoms on Al(111) surface using constant temperature (300 K) molecular dynamics simulation. Adatoms were randomly distributed on the surface as shown at t = 0 ps, $\theta = 0.14$. The images shown are a 2 $\times$ 2 repeat of the unit cell. Grey and red colored atoms correspond to the surface atoms and adatoms, respectively.

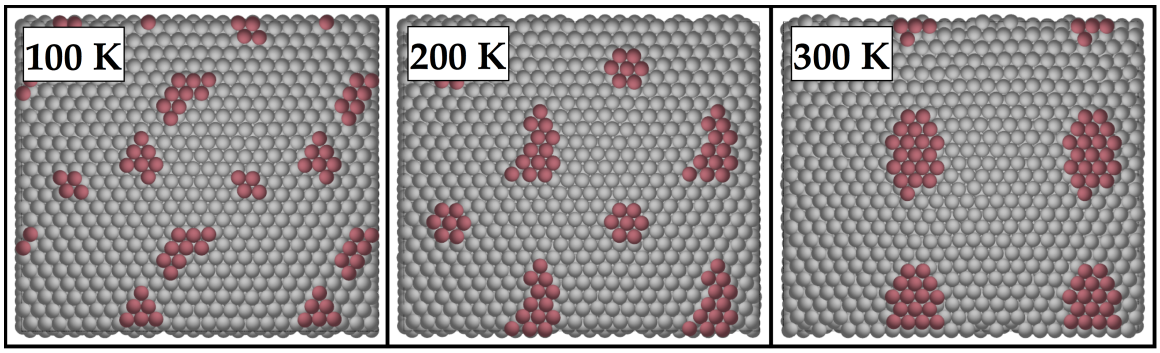

Figure 4.20: Island ripening as a function of temperature. Shown here is simulation at the end of 2.5 ns for 100 K, 200 K, 300 K. Clearly, as the temperature more compact islands start to form. $\theta = 0.14$. The images shown are a 2×2 repeat of the unit cell. Grey and red colored atoms correspond to the surface atoms and adatoms, respectively.

4.4.3 Role of temperature

Temperature plays a very critical role in governing the morphological shape and density of the islands (number of islands per unit area) formed. Here, to explore this morphological evolution we simulate the ripening behavior between 100 K - 300 K, as shown in Figure 4.20. At low temperatures ($\approx$ 100 K) several small islands nucleate, as a result of adatom or dimer diffusion, but remain immobile once a critical size is reached. At slightly elevated temperatures ($\approx$ 200 K) the onset of ripening, whereby multiple small islands coalesce (increasing the mean island size) begins to occur, but with no significant rearrangement in the shape of the larger islands formed. This is because the elementary processes that lead to island compactness; kink breaking, corner diffusion, etc., are only activated at higher temperatures. For the Al(111) surface this happens at temperatures $>$ 250 K, as seen by the simulations conducted at 300 K. A visual comparison between experimental STM topographic images or past kinetic Monte Carlo simulations and snapshots in Figure 4.20 reveal similar morphological shapes as one progresses along the temperature scale [187, 188, 189,

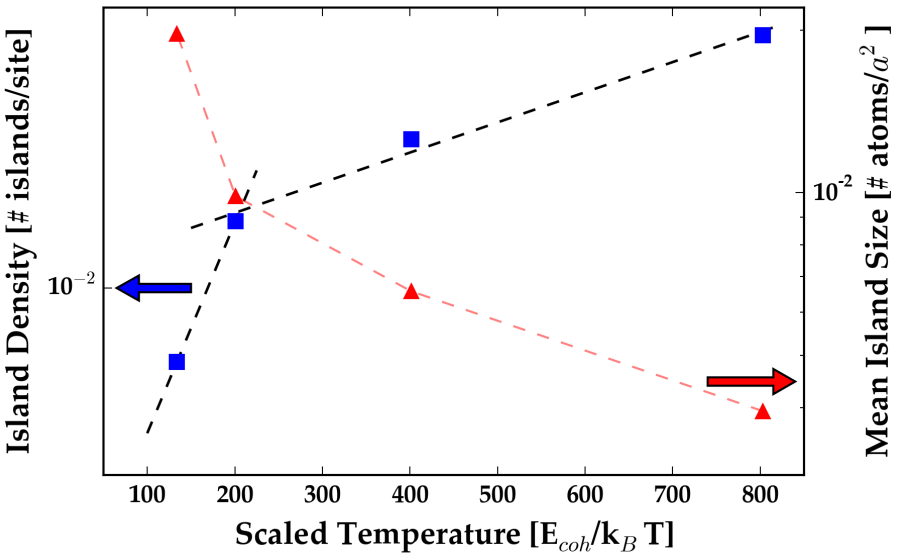

Figure 4.21: Mean island density and size as a function of temperature. Two growth regimes are observed, one at the low temperature and at the high temperature, with a transition point of $\approx$ 200 K.

182].

Given that the mobility small island nuclei is strictly governed by the underlying temperature, it is well known that two growth regimes dominate, with a transition temperature (T_t) at 200 K. For temperatures below T_t, islands grow by diffusion of monomers and dimers, while at temperatures above T_t clusters of larger sizes begin to be mobile. Therefore, by plotting the mean island density and size as a function of inverse temperature (c.f., in Figure 4.21), a clear distinction between the two growth regimes is observed. The slope was compute an estimated activation barrier of 2e-3 and 1e-2, while ultra-high vacuum sputtering studies, by Busse *et al.*, reveals values of 6e-3 and 4e-2, i.e., of the same order of magnitude [187].

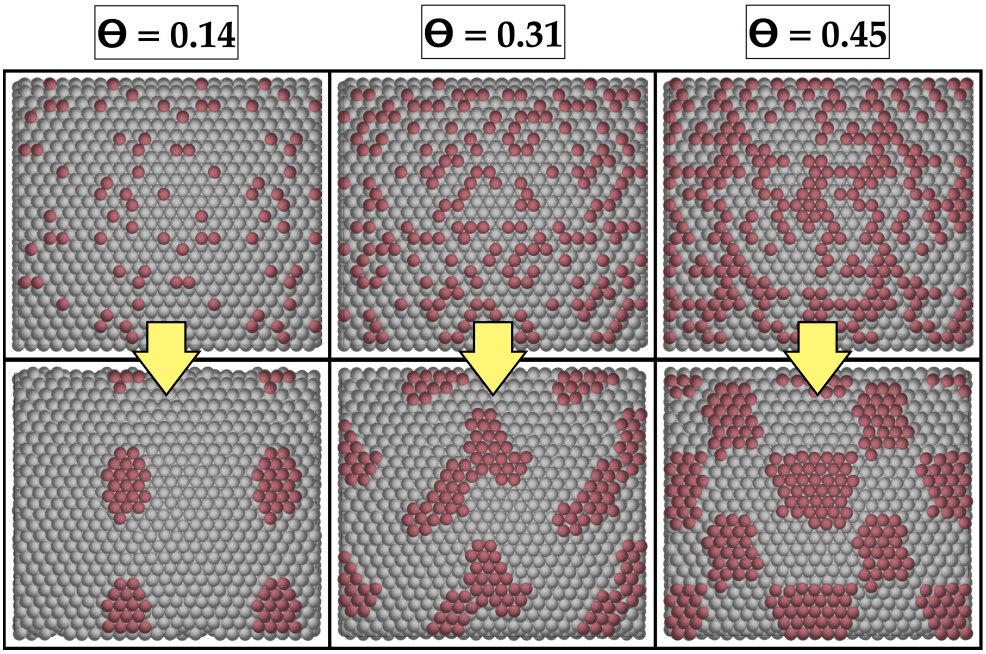

Figure 4.22: Island ripening for 3 different coverages; 0.14, 0.31, and 0.45. Top panel shows the starting configuration, where adatoms are randomly distributed.Bottom panel shows the island formation at the end of 7ns at a temperature of 300K. The images shown are a 2 $\times$ 2 repeat of the unit cell. Grey and red colored atoms correspond to the surface atoms and adatoms, respectively.

4.4.4 Role of surface coverage

Lastly, we study the role of surface coverage on the impact of island formation. At low coverages clusters of islands form, but as the coverage increases, the compact growth transitions more into a fractal type. Irrespective of these features, for a range of coverages $\theta = [0.14, 0.31, 0.45]$ at 300 K, islands still form. It should be noted that given the limited time scales achievable by conventional MD simulations, the islands do not necessarily form the minimum energy configurations. This would require using accelerated MD methods or even the stochastic kinetic Monte Carlo approach [165, 28, 22, 188, 189]. Here the simulations explore the beginnings of the ripening processes, where surface reaction process due to diffusion are prevalent. However, the goal is to demonstrate that AGNI force fields can indeed be used to explore phenomena beyond the *ab initio* scales.

4.5 O on an Al(111) surface

A natural question that arises at this point is how this force field paradigm may be extended to include multiple elements. The atomic fingerprint ($\boldsymbol{V}_i(\eta)$), described by Eq. 4.5, can be directly applied to non-elemental systems as well. Here, one would follow a similar approach whereby descriptors for each interaction pair between the different elements, one for each atom type, are considered. For example, given a binary system with elements m and n, the possible atomic neighbor pair interaction types are: mm, mn, nm and nn. The final multi-elemental atomic fingerprint is then generated by concatenating the independent atomic pair fingerprints, i.e., $V_i(\eta) = [V_i^{mm}, V_i^{mn}]$ for atom m, and $[V_i^{nm}, V_i^{nn}]$ for atom n. Then, by once again using the kernel ridge regression learning framework, along with the rigors of cross-validation, a mapping of

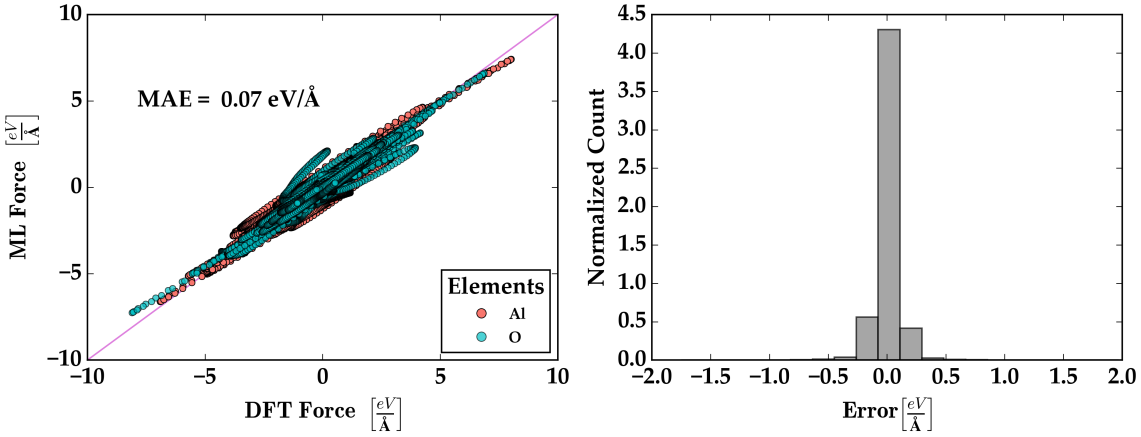

Figure 4.23: (a) Force prediction for O and Al force field in eV/Å. (b) A distribution of the errors. The mean absolute error (MAE) in prediction is ≈ 0.07 eV/Å.

the multi-elemental fingerprint to their corresponding atomic force is established.

Though this scheme requires further optimization, preliminary work shows significant promise. Here, as an extension to the elemental Al system, considered thus far, the interaction of a single O adatom on the Al(111) surface is explored. The first step in this process is to once again develop an accurate AGNI force field that can describe this class of system. To do so, we follow the procedure laid out in the beginning of the chapter; reference DFT data is sampled from diverse configurations in ,this case with O atoms on the surface, the atomic environments are fingerprinted, and the kernel ridge regression algorithm is once again used to map the fingerprint to the reference forces. A parity plot showing the force predictive prowess for this multi-elemental scenario is shown in Figure 4.23. Though, the errors are ≈ 0.07 eV/Å there is room for improvement. Nevertheless, it suggests that such a multi-elemental model can indeed be developed and used in a manner identical to the previous atomistic simulations. As a demonstration of the fidelity of the multi-elemental force field, reaction barriers were once again computed via integration of the forces, using Eq. 4.8. In this example the migration of an O adatom across adjacent surface sites (from

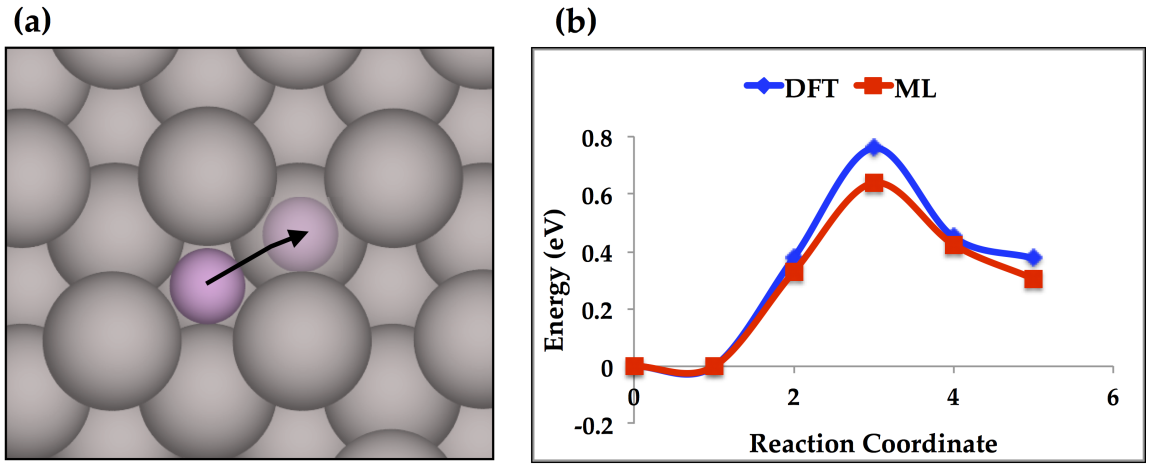

Figure 4.24: (a) A pictorial representation of O adatom migration from an fcc to hcp site on an Al(111) surface. The picture shows a top view of the (111) surface. The grey atoms correspond to surface Al, while the magenta atom is O. The shaded magenta atom indicates the final location of the O adatom. (b) The potential energy computed via DFT (blue line) and the potential energy recomputed via integrating the forces predicted by an AGNI model, using Eq. 4.8.

fcc to hcp) is considered, as illustrated pictorially in Figure 4.24(a). Upon computing and comparing the barrier with the DFT reference, we observe that the true value is underestimated, owing to the marginally higher errors in force predictions.

Interestingly, we observe that developing a force field to describe bulk α-Al_2O_3 is relatively simpler, as demonstrated in Appendix A. The periodic nature of such systems results in a smaller phase space, which then allows a direct extension of current methods to develop accurate force fields. For bulk α-Al_2O_3 we were able to explore geometry optimizations, compute phonon density of states, and do molecular dynamic simulations as well. Nevertheless, extending such simulations to surfaces still remains to be further tested. Owing to the vast increase in the configurational space, that arises due to the non-periodic nature of the surface, a larger phase space needs to be sampled and used while training the models. This poses a challenge for

the proposed kernel ridge regression learning framework. It is necessary to go beyond the learning methods described in this work, and is one of the current shortcomings of this approach, as shall be discussed next in Ch. 5.

4.6 Summary

A new machine learning framework to circumvent the accuracy, cost, and generalizability issues facing current atomistic models has been proposed. By directly mapping quantum mechanical derived force components to the numerical representation of an atom's local environment, accurate and computationally inexpensive force fields were developed. Here, a framework for their systematic construction and validation, with the example of Al, is discussed. Lastly, and more importantly, methods to quantify uncertainties in the force predictions made are proposed. This in turn allows one to identify the domain of applicability of such force fields, paving the way for their adaptive refinement. Such protocols are critical in keeping up with the current demands for atomistic based methods, given the ever increasing desire to study more complex materials and chemistries at the atomic scale.

Using the framework laid out, a force field for Al was developed and an exposé of materials simulation examples; dynamical evolution of defects over long time scales (vacancies and adatoms) to determine kinetic diffusion barriers, geometry optimization of configurations with several 100s of atoms, computing vibrational properties of materials to determine the phonon band spectrum and density of states, and estimating reaction barriers for dimer migration by force integration methods, were explored. To push the capability of modeling the dynamic behavior of surface processes with ML, a study of the ripening phenomena on an Al(111) surface was explored. Firstly, the relevant elementary processes that characterize such phenomena were validated,

followed by dynamical simulations as a function of time, temperature and surface coverage of adatoms. We see that at low temperatures the islands are more fractal like, and transitions into compact islands as temperature increases. Also, irrespective of the coverage, islands form at temperatures around 300K, agreeing well with past experimental and theoretical studies. Clearly, such methods do indeed provide room for optimism whereby one can now tackle atomistic simulations at much longer time and length scales in comparison to traditional *ab initio* methods. Lastly, an introduction to extending such simulations to multi-elemental systems was put forth, whereby, the behavior of an O-adatom on an Al(111) surface was explored. However, owing to the configurational diversity challenges pertaining to using larger amounts of training data arose.

Chapter 5

Summary and Future Outlook

Surface chemistry is an important topic in the chemical and material sciences. It is a phenomenon observed across diverse scenarios, e.g. catalysis, growth of materials, corrosion behavior, to name a few. For these reasons there is vested interest to better understand and characterize such processes, so as to design materials or chemical protocols in a more rational and guided fashion. A way to do this is by better understanding the thermodynamics and kinetics that govern such processes. Thus far, towards this cause *ab initio* models based on quantum mechanics have been instrumental to help probe surface chemistry processes at the the atomic level. Their capability to predict properties such as energy and forces, for diverse chemistries, makes them immensely powerful. These two quantities allow for thermodynamic comparisons, as well as to dynamically study the behavior of a system so as to reveal its kinetics. Nevertheless, the computational burden of *ab initio* methods severely restricts the extent of the generated models, even with current advances in computer hardware. For instance, enumeration of different material choices and/or exploring long time- and length-scale dynamical simulations are all beyond reach.

To overcome these challenges, here, methods based upon emerging machine learning techniques were used. By applying the sophisticated mathematical algorithms on data, generated with *ab initio* methods, predictive models as well as those to discover and mine patterns can be developed much quicker. This minimizes the need to

conduct future expensive *ab initio* calculations.

In this thesis, ML methods were used to both mine for patterns and develop predictive models targeted at surface chemistry applications. For the first case, a framework to systematically identify descriptors that describe surface reactivity of cerium oxides for the dissociation of H_2O were identified. This information, along with some high-throughput screening by *ab initio* methods for dopants in surfaces, allowed us to explore techniques such as principal component analysis and random forests to come up with a catalyst design framework to quickly identify promising dopants. Such guidelines for experimental design are critical to speed up the discovery process. In the second example, a predictive model to compute the forces on atoms was explored. Here by once again using *ab initio* data as reference, numerical descriptors that capture an atoms local atomic environment were proposed, both for elemental systems and beyond. These along with the kernel ridge regression framework were used to establish a mapping between the descriptors and the force on an atom. The framework was shown to be able to reproduce forces close to *ab initio* accuracy. Such a method was then used to study a host of simulations, e.g. molecular dynamics, geometry optimization, amongst several others.

Thus far, we witnessed a brief expose of the promises of machine learning in chemical and material sciences. Given the fidelity of such methods as demonstrated in this work and elsewhere, one can live with optimism that more complex phenomena such as transport (thermal and mass) behavior, phase transformations, chemical reactions, mechanical behavior, materials degradation and failure, etc., all lie within the framework of reality-mimicking non-zero temperature simulations. Nevertheless, to make such methods a mainstream tool for atomistic simulations a few challenges yet remain that need to be addressed, a few of which are discussed below.

1. *Big data approach.* As materials science or chemical systems become ever increasingly complex, the configuration space to be explored will increase exponentially. This poses a challenge in training the models. The continued realization of machine learning force fields in describing such problems will require adopting *big data* methodologies, wherein large quantities of data can be handled.

2. *Curating data.* The interpolative nature of ML methods, makes it necessary to a priori understand and sample the reference data chosen, smartly. This entails how the reference data is being sampled. Given that DFT data is expensive to generate, one could use sampling methods such as metadynamics or nested sampling, to explore relevant space.

3. *Adaptive models.* The first step to a true machine learning model is to identify regions where a model would possible fail. The uncertainty estimation methodology proposed here, is one such solution. The goal is that at some point in time, models self-learn areas of poor performance in a systematic manner and refine predictions automatically. Thus, hopefully eliminating the need to use expensive quantum mechanical methods entirely.

4. *More complex descriptors.* Given that machine learning methods lack a precise functional form, it is necessary that the physics be contained in the descriptors used to represent a phenomena or process. For example, in the catalyst design framework, the descriptors were merely atomic properties, but one could go beyond them and come up with varying combinations of the descriptors so as to capture more physics. Having said that, it is better to keep a model as simple as possible.

Though these challenges exist, a logical pathway to answering these questions is to learn from the progressed communities, such as computer science or chem-informatics. It is a matter of adapting such practices for the needs of surface chemistry applications, to hopefully pave the way for models that provide quick feedback on material and chemical behavior.

Appendices

Chapter A

Additional models details

A.1 Determining oxidation states of surface phases

To validate the stable phases predicted by FPT calculations, we performed a Bader analysis to determine the oxidation states of the Ce atoms under different O environments. Ce atoms in CeO_2 and Ce_2O_3 have a formal oxidation state of +4 and +3, respectively. However, the nominal oxidation states as recovered by the Bader analysis are expected to be different from these formal values due to incomplete charge transfer and the inability to unambiguously partition space in order to determine the formal atomic charge. In our calculations, bulk ceria in a nominal +4 and +3 state exhibits a Bader oxidation state of +2.23 and +1.98, respectively. These oxidation states are similar to the bulk values (+2.37 and +1.98) reported by Loschen *et al.*. Using these values as a benchmark, we compared the oxidation states of Ce atoms in each trilayer for all the configurations to distinguish whether the Ce atom is in an oxidized (+4) or reduced (+3) state. Figure A.1 shows the Bader charge of the Ce atoms as we progress from the 1^{st} trilayer down to the bulk-like internal trilayers. The two dotted lines in Figure A.1 represent the Bader oxidation states of the Ce atoms in bulk CeO_2 and Ce_2O_3. The $CeO_{1.75}^{1,0.75}$ configuration has an oxidation state (+2.37) similar to bulk CeO_2, whereas the $CeO_{1.25}^{1,0.25}$ configuration has an oxidation state (+1.98) similar to bulk Ce_2O_3. This indicates why the intermediate $CeO_{1.5}^{1,0.5}$ config-

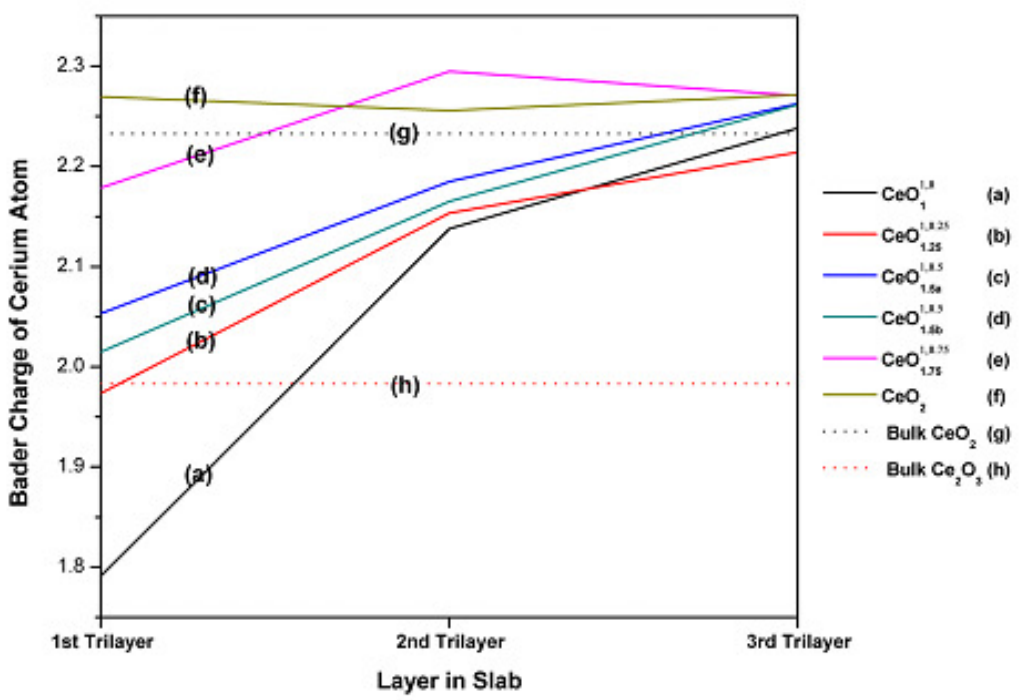

Figure A.1: Bader charge distribution of Ce atoms progressing from the 1^{st} trilayer at the surface to the internal bulk-like 3^{rd} trilayer.

uration does not show up as a stable phase in the relative surface energy plot. For any intermediate oxidation states, this configuration is not energetically favored. A similar reduction in the Bader charge ($\sim$0.1 eV) on progressing to non-stoichiometric cases has been reported with ceria nanoparticles whereby the reduced co-ordination affects bonding between the Ce and O ions, and also affecting the surface energy as a result[53]. Finally, as we move into the material (i.e., the inner trilayers), the Bader oxidation state of bulk CeO_2 is recovered; and thus the energy required for forming an O vacancy from the 2^{nd} and 3^{rd} trilayers is equivalent to that of an O vacancy from bulk CeO_2. This is also consistent with the conclusion drawn from the O vacancy formation energy.

A.2 Preliminary AGNI force fields for other elements

Preliminary machine learning force fields were developed for elements such as Si, W, Al2O3, and HfO2.

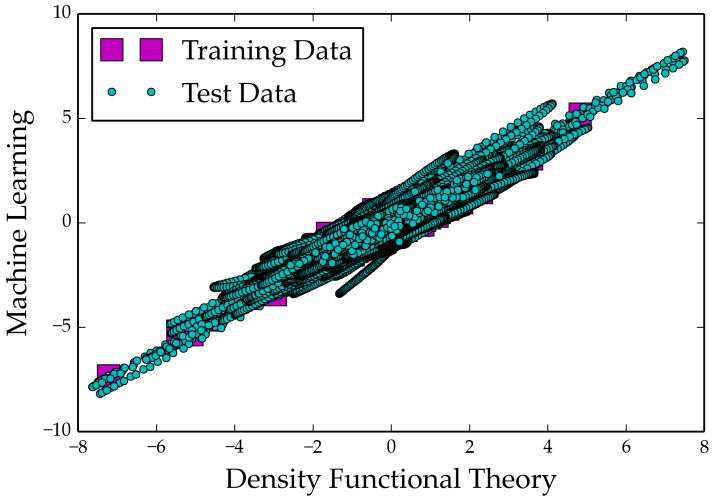

Figure A.2: Force prediction in eV/Åfor the element Si. The reference environments comprised of bulk, surfaces, and defects. The mean absolute error (MAE) in prediction is $\approx$ 0.05 eV/Å.

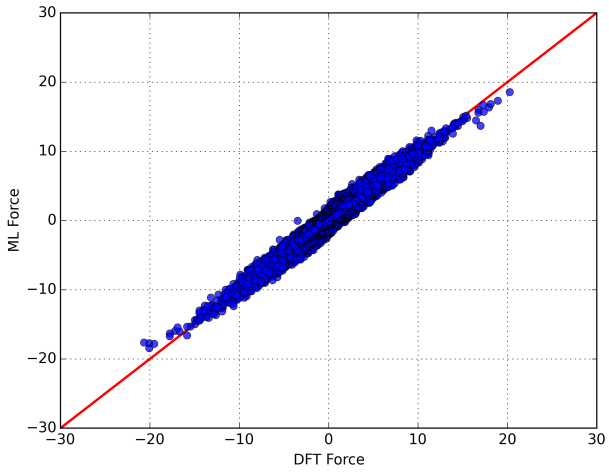

Figure A.3: Force prediction in eV/Åfor the element W. The reference environments comprised of bulk, surfaces, defects, and dislocations. The mean absolute error (MAE) in prediction is ≈ 0.05 eV/Å.

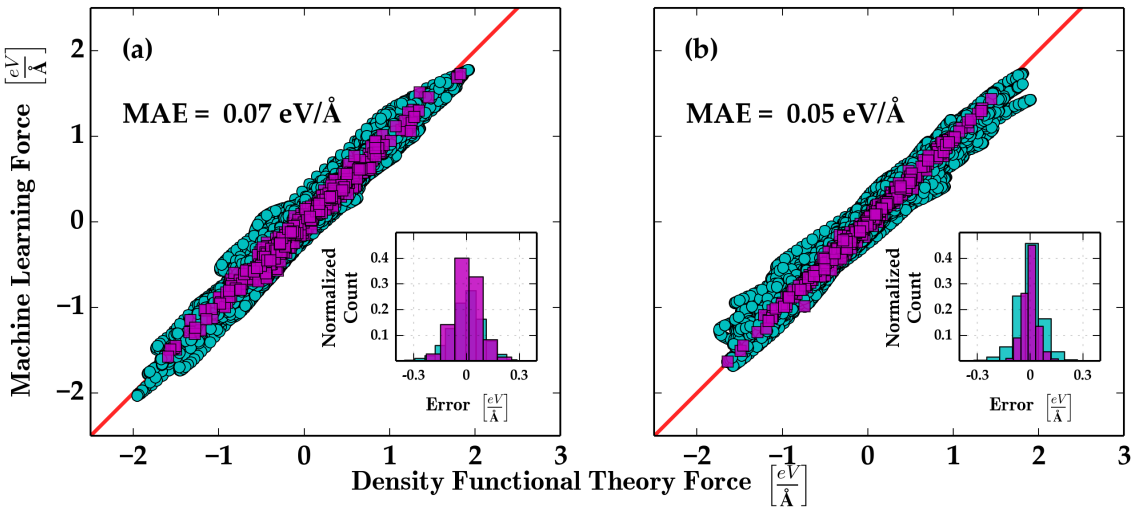

Figure A.4: Force prediction for (a) Al and (b) O type atoms in α-Al_2O_3 in eV/Å. The inset shows a distribution of the errors. The mean absolute error (MAE) in prediction is ≈ 0.05 eV/Å.

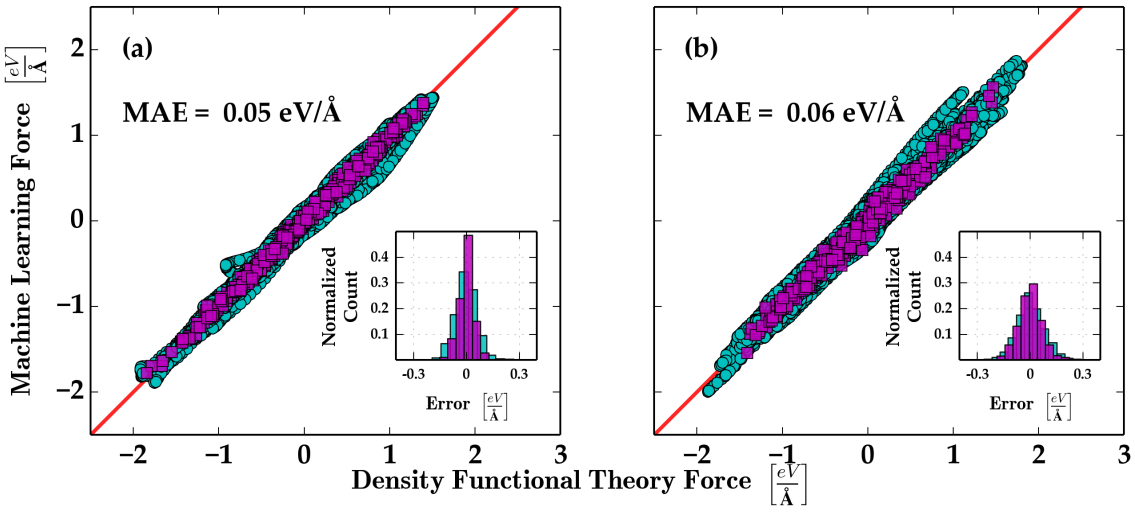

Figure A.5: Force prediction for (a) Hf and (b) O type atoms in monoclinic - HfO_2 in eV/Å. The inset shows a distribution of the errors. The mean absolute error (MAE) in prediction is ≈ 0.05 eV/Å.

Bibliography

[1] Gabor A. Somorjai and Yimin Li. Impact of surface chemistry. *Proceedings of the National Academy of Sciences*, 108(3):917–924, 2011.

[2] Catherine Stampfl, M. Veronica Ganduglia-Pirovano, Karsten Reuter, and Matthias Scheffler. Catalysis and corrosion: the theoretical surface-science context. *Surface Science*, 500(1–3):368 – 394, 2002.

[3] Wyndham Rowland Dunstan and John Richard Hill. Ccviii.-the aerial oxidation (rusting) of metals. *J. Chem. Soc. Trans.*, 99:1835–1853, 1911.

[4] Gabor A Somorjai and Yimin Li. *Introduction to surface chemistry and catalysis*. John Wiley & Sons, 2010.

[5] B. Hammer and J.K. Nørskov. Theoretical surface science and catalysis—calculations and concepts. In *Impact of Surface Science on Catalysis*, volume 45 of *Advances in Catalysis*, pages 71 – 129. Academic Press, 2000.

[6] J. Norskov, F. A. Pedersen, F. Studt, and T. Bligaard. Density functional theory in surface chemistry and catalysis. *P. Natl. Acad. Sci. USA*, 108(3):937, 2011.

[7] Jürgen Hafner, Christopher Wolverton, and Gerbrand Ceder. Toward computational materials design: The impact of density functional theory on materials research. *MRS Bulletin*, 31:659–668, 9 2006.

[8] P. Hohenberg and W. Kohn. Inhomogeneous electron gas. *Phys. Rev.*, 136:B864–B871, Nov 1964.

[9] W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Phys. Rev.*, 140:A1133–A1138, Nov 1965.

[10] J. P. Perdew, K. Burke, and Y. Wang. Generalized gradient approximation for the exchange-correlation hole of a many−electron system. *Phys. Rev. B*, 54:16533, 1996.

[11] K. Reuter, C. Stampfl, and M. Scheffler. *Handbook of Materials Modeling*, volume 1. Springer, Berlin, 2005.

[12] J. Norskov, J. Rossmeisl T. Bligaard, and C. H. Christensen. Towards the computational design of solid catalysts. *Nat. Chem.*, 1:37, 2009.

[13] J. Norskov, M. Scheffler, and H. Toulhoat. Density functional theory in surface science and heterogeneous catalysis. *Mat. Res. Soc. Bulletin*, 31:669, 2006.

[14] R. Petrenko and J. Meller. *Encyclopedia of Life Sciences*. John Wiley and Sons Ltd, 2010.

[15] Mike P Allen and Dominic J Tildesley. *Computer simulation of liquids*. Oxford university press, 1989.

[16] Karl D. Brommer, M. Needels, B. Larson, and J. D. Joannopoulos. *Ab initio* theory of the si(111)-(7×7) surface reconstruction: A challenge for massively parallel computation. *Phys. Rev. Lett.*, 68:1355–1358, Mar 1992.

[17] Joseph C. Fogarty, Hasan Metin Aktulga, Ananth Y. Grama, Adri C. T. van Duin, and Sagar A. Pandit. A reactive molecular dynamics simulation of the silica-water interface. *The Journal of Chemical Physics*, 132(17), 2010.

[18] Zhen Jiao, Changbao Song, Tiesong Lin, and Peng He. Molecular dynamics simulation of the effect of surface roughness and pore on linear friction welding between ni and al. *Computational Materials Science*, 50(12):3385 – 3389, 2011.

[19] Volker Blum, Gus LW Hart, Michael J Walorski, and Alex Zunger. Using genetic algorithms to map first-principles results to model hamiltonians: Application to the generalized ising model for alloys. *Physical Review B*, 72(16):165113, 2005.

[20] B. W. H. van Beest, G. J. Kramer, and R. A. van Santen. Force fields for silicas and aluminophosphates based on *ab initio* calculations. *Phys. Rev. Lett.*, 64:1955–1958, Apr 1990.

[21] G Seifert. Tight-binding density functional theory: an approximate kohn-sham dft scheme. *The Journal of Physical Chemistry A*, 111(26):5609–5613, 2007.

[22] G. Henkelman and H. Jonsson. Long time scale kinetic monte carlo simulations without lattice approximation and predefined event table. *J. Chem. Phys.*, 115:9657, 2001.

[23] A. Chatterjee and D. G. Vlachos. An overview of spatial microscopic and accelerated kinetic monte carlo methods. *J. Comput-Aided Mater.*, 14:253, 2007.

[24] A. Laio and M. Parrinello. Escaping free-energy minima. *Proc. Natl Acad. Sci.*, 99:12562, 2002.

[25] A. Laio and F. L. Gervasio. Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. *Rep. Progr. Phys.*, 71:126601, 2008.

[26] M. R. Sorensen and A. F. Voter. Temperature-accelerated dynamics for simulation of infrequent events. *J. Chem. Phys.*, 112:9599, 2000.

[27] F. Voter. A method for accelerating the molecular dynamics simulation of infrequent events. *J. Chem. Phys.*, 106:4665, 1997.

[28] F. Voter, F. Montalenti, and T. C. Germann. Extending the time scale in atomistic simulation of materials. *Annu. Rev. Mater. Res.*, 32:321, 2002.

[29] F. Voter and M. R. Sorensen. Accelerating atomistic simulations of defect dynamics: Hyperdynamics, parallel replica dynamics, and temperature-accelerated dynamics. *Mater. Res. Soc. Symp. Proc.*, 538:427, 1999.

[30] F. Voter. Hyperdynamics: accelerated molecular dynamics of infrequent events. *Phys. Rev. Lett.*, 78:3908, 1997.

[31] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, 2nd edition, 2009.

[32] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1684–1692, 2015.

[33] Herman Anthony Carneiro and Eleftherios Mylonakis. Google trends: a web-based tool for real-time surveillance of disease outbreaks. *Clinical infectious diseases*, 49(10):1557–1564, 2009.

[34] J Ben Schafer, Joseph A Konstan, and John Riedl. E-commerce recommendation applications. In *Applications of Data Mining to Electronic Commerce*, pages 115–153. Springer, 2001.

[35] I. H. Witten, E. Frank, and M. A. Hall. *Data mining: Practical machine learning tools and techniques*. Elsevier, 2011.

[36] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[37] T. Mueller, A. G. Kusne, and R. Ramprasad. Machine learning in materials science: Recent progress and emerging applications. In A. L. Parrill and K. B. Lipkowitz, editors, *Reviews in Computational Chemistry*. Wiley, 2016.

[38] Geoffroy Hautier, Anubhav Jain, and Shyue Ping Ong. From the computer to the laboratory: materials discovery and design using first-principles calculations. *J. Mater. Sci.*, 47(21):7317–7340, 2012.

[39] S. Srinivas and K. Rajan. "property phase diagrams" for compound semiconductors through data mining. *Materials*, 6:279, 2013.

[40] C. J. Long, J. Hattrick-Simpers, M. Murakami, R. C. Srivastava, I. Takeuchi, V. L. Karen, and X. Li. Rapid structural mapping of ternary metallic alloy systems using the combinatorial approach and cluster analysis. *Rev. Sci. Instrum.*, 78:072217, 2007.

[41] C. C. Fischer, K. J. Tibbetts, D. Morgan, and G. Ceder. Predicting crystal structure by merging data mining with quantum mechanics. *Nature Mat.*, 5:641, 2006.

[42] X. Zhang, L. Yu, A. Zakutayev, and A. Zunger. Sorting stable versus unstable

hypothetical compounds: The case of multi-functional abx half-heusler filled tetrahedral structures. *Adv. Funct. Mater.*, 22:1425, 2012.

[43] A. R. Oganov, Y. Ma, A. O. Lyakhov, M. Valle, and C. Gatti. Evolutionary crystal structure prediction as a method for the discovery of minerals and materials. *Rev. Mineral. Geochem.*, 71:271, 2010.

[44] P. V. Balachandran, S. R. Broderick, and K. Rajan. Identifying the 'inorganic gene' for high-temperature piezoelectric perovskites through statistical learning. *Proc. R. Soc. A*, 467:2271, 2011.

[45] E. W. Bucholtz, C. S. Kong, K. R. Marchman, W. G. Sawyer, S. R. Phillpot, S. B. Sinnot, and K. Rajan. Data-driven model for estimation of friction coefficient via informatics methods. *Tribol. Lett.*, 47:211, 2012.

[46] I. E. Castelli and K. W. Jacobsen. Designing rules and probabilistic weighting for fast materials discovery in the perovskite structure. *Modelling Simul. Mater. Sci. Eng.*, 22:055007, 2014.

[47] D. Morgan, G. Ceder, and S. Curtarolo. High-throughput and data mining with ab initio methods. *Meas. Sci. Technol.*, 16:296, 2005.

[48] A. G. Kusne, T. Gao, A. Mehta, L. Ke, M. C. Nguyen, K. M. Ho, V. Antropov, C. Z. Wang, M. J. Kramer, C. Long, and I. Takeuchi. On-the-fly machine-learning for high-throughput experiments: search for rare-earth-free permanent magnets. *Sci. Rep.*, 4:6367, 2014.

[49] Matthias Rupp. Machine learning for quantum mechanics in a nutshell. *Int. J. Quant. Chem.*, 115(16):1058–1073, 2015.

[50] G. Pilania, C. Wang, X. Jiang, S. Rajasekaran, and R. Ramprasad. Accelerating materials property predictions using machine learning. *Scientifc Reports*, 3:2810, 2013.

[51] G. Montavon, M. Rupp, V. Gobre, A. V. Mayagoitia an K. Hansen, A. Tkatchenko, K. R. Muller, and O. A. von Lilienfeld. Machine learning of molecular electronic properties in chemical compound space. *New J. Phys.*, 15:095003, 2013.

[52] J. Behler. Neural network potential-energy surfaces in chemistry: a tool for large-scale simulations. *Phys. Chem. Chem. Phys.*, 13:17930, 2011.

[53] Albert P. Bartók, Mike C. Payne, Risi Kondor, and Gábor Csányi. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.*, 104:136403, Apr 2010.

[54] V. Botu, R. Ramprasad, and A. B. Mhadeshwar. Ceria in an oxygen environment: Surface phase equilibria and its descriptors. *Surf. Sci.*, 619:49, 2014.

[55] V Botu, AB Mhadeshwar, SL Suib, and R Ramprasad. Optimal dopant selection for water splitting with cerium oxides: Mining and screening first principles data. In *Information Science for Materials Discovery and Design*, pages 157–171. Springer, 2016.

[56] Venkatesh Botu and Rampi Ramprasad. Adaptive machine learning framework to accelerate ab initio molecular dynamics. *Int. J. Quant. Chem.*, 115(16):1074–1083, 2015.

[57] Venkatesh Botu and Rampi Ramprasad. Machine learning force fields: Construction, validations and uncertainty estimation. *in progress*, 2016.

[58] V. Botu and R. Ramprasad. Learning scheme to predict atomic forces and accelerate materials simulations. *Phys. Rev. B*, 92:094306, Sep 2015.

[59] Venkatesh Botu and Rampi Ramprasad. Island ripening on an al(111) surface. *in progress*, 2016.

[60] Erwin Schrödinger. Quantisierung als eigenwertproblem. *Annalen der physik*, 385(13):437–490, 1926.

[61] Max Born. Born-oppenheimer approximation. *Ann. Physik*, 84:457, 1927.

[62] David Sholl and Janice A Steckel. *Density functional theory: a practical introduction*. John Wiley & Sons, 2011.

[63] Robert G. Parr and Weitao Yang. Density functional approach to the frontier-electron theory of chemical reactivity. *Journal of the American Chemical Society*, 106(14):4049–4050, 1984.

[64] R. P. Feynman. Forces in molecules. *Phys. Rev.*, 56:340–343, Aug 1939.

[65] Christopher J Cramer and FM Bickelhaupt. Essentials of computational chemistry. *ANGEWANDTE CHEMIE-INTERNATIONAL EDITION IN ENGLISH*, 42(4):381–381, 2003.

[66] Daan Frenkel and Berend Smit. *Understanding molecular simulation: from algorithms to applications*, volume 1. Academic press, 2001.

[67] Shūichi Nosé. A molecular dynamics method for simulations in the canonical ensemble. *Molecular physics*, 52(2):255–268, 1984.

[68] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[69] Sotiris B Kotsiantis, I Zaharakis, and P Pintelas. Supervised machine learning: A review of classification techniques, 2007.

[70] Antonio Criminisi, Jamie Shotton, and Ender Konukoglu. Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and Trends® in Computer Graphics and Vision*, 7(2–3):81–227, 2012.

[71] A. Trovarelli. *Catalysis by ceria and related materials*, volume 2. Imperial college press, London, 2002.

[72] Yixin Liu, Yu Ding, Lichun Zhang, Pu-Xian Gao, and Yu Lei. Ceo2 nanofibers for in situ o2 and co sensing in harsh environments. *RSC Adv.*, 2:5193–5198, 2012.

[73] W.C. Chueh and S. M. Haile. A thermochemical study of ceria: exploiting and old material for new modes of energy conversion and CO_2 mitigation. *Phil. Trans. R. Soc. A*, 368:3269, 2010.

[74] G. Ceder and K. Persson. The stuff of dreams. *Sci. Amer.*, 309:36, 2013.

[75] Jörg Neugebauer and Tilmann Hickel. Density functional theory in materials science. *Comp. Mol. Sci.*, 3(5):438–448, 2013.

[76] G. Hautier, A. Jain, and S. P. Ong. From the computer to the laboratory: materials discovery and design using first-principles calculations. *J. Mater. Sci.*, 47:7317, 2012.

[77] Axel D. Becke. Perspective: Fifty years of density-functional theory in chemical physics. *J. Chem. Phys.*, 140(18), 2014.

[78] J. Carrete, W. Li, N. Mingo, S. Wang, and S. Curtarolo. Finding unprecedentedly low-thermal-conductivity half-heusler semiconductors via high-throughput materials modeling. *Phys. Rev. X*, 4:011019, 2014.

[79] R. J. Gorte. Ceria in catalysis: From automotive applications to the water-gas shift reaction. *AIChE Journal*, 56(5):1126–1135, 2010.

[80] R. J. Gorte and S. Zhao. Studies of the water-gas-shift reaction with ceria-supported precious metals. *Catalysis Today*, 104:18–24, 2005.

[81] D.J.M. Bevan and J. Kordis. Mixed oxides of the type mo2 (fluorite)—m2o3—i oxygen dissociation pressures and phase relationships in the system ceo2-ce2o3 at high temperatures. *Journal of Inorganic and Nuclear Chemistry*, 26(9):1509 – 1523, 1964.

[82] R.J. Panlener, R.N. Blumenthal, and J.E. Garnier. A thermodynamic study of nonstoichiometric cerium dioxide. *Journal of Physics and Chemistry of Solids*, 36(11):1213 – 1222, 1975.

[83] R. J. Ackermann and E. G. Rauh. A high-temperature study of the stoichiometry, phase behavior, vaporization characteristics, and thermodynamic properties of the cerium + oxygen system. *The Journal of Chemical Thermodynamics*, 3:609–624, 1971.

[84] R. G. Schwab, R. A. Steiner, G. Mages, and H. J. Beie. Properties of ceo2 and ceo2-x films. part ii. high temperature properties. *Thin Film Solids*, 207:288–293, 1992.

[85] R. Korner and M. Nolting Ricken J. Phase tansformations in reduced ceria:

Determination by thermal expansion measurements. *Journal of Solid State Chemistry*, 78:136–147, 1989.

[86] M. Ricken and J. Nolting. Specific heat and phase diagram of nonstoichiometric ceria (ceo2-x). *Journal of Solid State Chemistry*, 54:89–99, 1984.

[87] S. Torbrugge and M. Reichling. Evidence of subsurface oxygen vacancy ordering on reduced ceo2 (111). *Physical Review Letters*, 99(056101):1–4, 2007.

[88] M. Krcha, A. D. Mayernick, and M. J. Janik. Periodic trends of oxygen vacancy formation and c−h bond activation over transition metal−doped CeO_2 (111) surfaces. *J. Catal.*, 293:103, 2012.

[89] M. Fronzi, S. Piccinin, B. Delley, E. Traversa, and C. Stampfl. Water adsorption on the stoichiometric and reduced CeO_2 (111) surface: a first-principles investigation. *Phys. Chem. Chem. Phys.*, 11:9188, 2009.

[90] F. Esch, S. Fabris, L. Zhou, T. Montini, C. Africh, P. Fornasiero, G. Comelli, and R. Rosei. Electron localization determines defect formation on ceria substrates. *Science*, 309:752–755, 2005.

[91] C. Zhang, A. Michaelides, D. A. King, and S. J. Jenkins. Oxygen vacancy clusters on ceria: Decisive role of cerium f electrons. *Physical Review B*, 79(075433):1–11, 2009.

[92] C. Loschen, J. Carrasco, K. M. Neyman, and F. Illas. First-principles lda+u and gga+u study of cerium oxides: Dependence on the effective u parameter. *Physical Review B*, 75(035115):1–8, 2007.

[93] S. Fabris, S. Gironcoli, S. Baroni, G. Vicario, and G. Balducci. Taming multiple

valency with density functionals: A case study of defective ceria. *Physical Review B*, 71(041102):1–4, 2005.

[94] J. L. F. Da Silva, M. V. G. Pirovano, J. Sauer, V. Bayer, and G. Kresse. Hybrid functionals applied to rare-earth oxides: The example of ceria. *Physical Review B*, 75(045121):1–10, 2007.

[95] K. Reuter, D. Frenkel, and M. Scheffler. The steady state of heterogeneous catalysis, studied by first-principles statistical mechanics. *Physical Review Letters*, 93(116105):1–4, 2004.

[96] K. Reuter and M. Scheffler. First-principles atomistic thermodynamics for oxidation catalysis: Surface phase diagrams and catalytically interesting regions. *Physical Review Letters*, 90(046103):1–4, 2003.

[97] J. Roga, K. Reuter, and M. Scheffler. First-principles statistical mechanics study of the stability of a subnanometer thin surface oxide in reactive environments: Co oxidation at pd(100). *Physical Review Letters*, 98(046101):1–4, 2007.

[98] G. Kresse and J. Furthmuller. Efficient iterative schemes for *ab initio* total−energy calculations using a plane−wave basis set. *Phys. Rev. B*, 54:11169, 1996.

[99] P. E. Blöchl. Projector augmented−wave method. *Phys. Rev. B*, 50:17953, 1994.

[100] G. Kresse and D. Joubert. From ultrasoft pseudopotentials to the projector augmented−wave method. *Phys. Rev. B*, 59:1758, 1999.

[101] J. Heyd, G. E. Scuseria, and M. Ernzerhof. Hybrid functionals based on a

screened coulumb potential. *Journal of Chemical Physics*, 118(18):8207–8215, 2003.

[102] J. F. Herbst, R. E. Watson, and J. W. Wilkins. Relativistic calculations of 4f excitation energies in the rare-earth metals: Further results. *Physical Review B*, 17(8):3089–3098, 1978. Cited By (since 1996): 35.

[103] V. I. Anisimov and O. Gunnarsson. Density-functional calculation of effective coulomb interactions in metals. *Physical Review B*, 43(10):7570–7574, 1991. Cited By (since 1996): 268.

[104] L. Gerward, J. S. Olsen, L. Petit, G. Vaitheeswaran, V. Kanchana, and A. Svane. Bulk modulus of ceo2 and pro2 - an experimental and theoretical study. *Journal of Alloys and Compounds*, 400:56–61, 2005.

[105] D. R. Hull and H. Prophet. Janaf thermochemical tables. Available at http://kinetics.nist.gov/janaf (01/15/2014).

[106] H. Zhu, C. Tang, and R. Ramprasad. Phase equilibria at si-hfo2 and pt-hfo2 interfaces from first principles thermodynamics. *Physical Review B*, 82(235413):1–10, 2010.

[107] R. Getman, Y. Xu, and W. Schneider. Thermodynamics of environment-dependent oxygen chemisorption on pt (111). *Journal of Physical Chemistry C*, 112:9559–9572, 2008.

[108] M. V. G. Pirovano, J. L. F. Da Silva, and J. Sauer. Density-functional calculations of the structure of near-surface oxygen vacancies and electron localization on ceo2 (111). *Physical Review Letters*, 102(026101):1–4, 2009.

[109] M. V. G. Pirovano, Alexander Hofmann, and Joachim Sauer. Oxygen vacancies in transition metal and rare earth oxides: Current state of understanding and remaining challenges. *Surface Science Reports*, 62(6):219–270, 6/30 2007.

[110] C. Loschen, A. Migani, S. T. Bromley, F. Illas, and K. M. Neyman. Density functional studies of model cerium oxide nanoparticles. *Physical Chemistry Chemical Physics*, 10:5730–5738, 2008.

[111] M. Nolan, J. E. Fearon, and G. W. Watson. Oxygen vacancy formation and migration in ceria. *Solid State Ionics*, 177:3069–3074, 2006.

[112] M. Nolan, S. Parker, and G. W. Watson. The electronic structure of oxygen vacancy defects at the low index surfaces of ceria. *Surface Science*, 595:223–232, 2005.

[113] Y. Jiang, J. B. Adams, and M. V. Schilfgaarde. Density-functional calculataion of ceo2 surfaces and prediction of effects of oxygen partial pressure and temperature on stabilities. *Journal of Chemical Physics*, 123(064701):1–10, 2005.

[114] M. Fronzi, A. Soon, B. Delley, E. Traversa, and C. Stampfl. Stability and morphology of cerium oxide surfaces in an oxidizing environment: A first-principles investigation. *Journal of Chemical Physics*, 131(104701):1–16, 2009.

[115] T. Desaunay, A. Ringuede, M. Cassir, F. Labat, and C. Adamo. Modeling basic components of solid oxide fuel cells using density functional theory: Bulk and surface properties of ceo2. *Surface Science*, 606:305–311, 2012.

[116] Bueno-Lopez A., Krishna K., Makkee M., and Moulijn J.A. Active oxygen from ceo2 and its role in catalysed soot oxidation. *Catalysis Letters*, 99(- 3-4):203, 2005.

[117] M. A. Henderson, C. L. Perkins, M. H. Engelhard, S. Thevuthasan, and C. H. F. Peden. Redox properties of water on the oxidized and reduced surfaces of ceo2(1 1 1). *Surface Science*, 526(1–2):1–18, 2/20 2003.

[118] S Hilaire, X Wang, T Luo, R.J Gorte, and J Wagner. A comparative study of water-gas-shift reaction over ceria supported metallic catalysts. *Applied Catalysis A: General*, 215(1–2):271 – 278, 2001.

[119] G. Jacobs, E. Chenu, P. M. Patterson, L. Williams, D. Sparks, G. Thomas, and B. H. Davis. Water-gas shift: comparative screening of metal promoters for metal/ceria systems and role of the metal. *Applied Catalysis A: General*, 258:203–214, 2004.

[120] C. Wheeler, A. Jhalani, E. J. Klein, S. Tummala, and L. D. Schmidt. The water-gas-shift reaction at short contact times. *Journal of Catalysis*, 223:191–199, 2004.

[121] S. Abanades, P. Charvin, G. Flamant, and P. Neveu. Screening of water-splitting thermochemical cycles potentially attractive for hydrogen production by concentrated solar energy. *Energy*, 31:2805, 2006.

[122] T. Nakamura. Hydrogen production from water utilizing solar heat at high temperatures. *Sol. Energy*, 19:467, 1977.

[123] S. Abanades and G. Flamant. Solar hydrogen production from the thermal splitting of methane in a high temperature solar chemical reactor. *Sol. Energy*, 80:1611, 2006.

[124] L. D'Souza. Thermochemical hydrogen production from water using reducible oxide materials: a critical review. *Mater. Renew. Sust. Energy*, 2:1, 2013.

[125] W.C. Chueh and S. M. Haile. Ceria as a thermochemical reaction medium for selectively generating syngas or methane from H_2O and CO_2. *Chem. Sus. Chem.*, 2:735, 2009.

[126] W. C. Chueh, C. Falter, M. Abbott, D. Scipio, P. Furler, S. M. Haile, and A. Steinfeld. High-flux solar-driven thermochemical dissociation of CO_2 and H_2O using nonstoichiometric ceria. *Science*, 330:1797, 2010.

[127] A. Trovarelli. *Catalysis by Ceria and Related Materials*. World Scientific, 2002.

[128] S. Kumar and P. K. Schelling. Density functional theory study of water adsorption at reduced and stoichiometric ceria (111) surfaces. *J. Chem. Phys.*, 125:204704, 2006.

[129] H. T. Chen, Y. M. Choi, M. Liu, and M. C. Lin. A theoretical study of surface reduction mechanisms of CeO_2(111) and (110) by H_2. *Chem. Phys. Chem.*, 8:849, 2007.

[130] Z. Yang, Q. Wang, S. Wei, D. Ma, and Q. Sun. The effect of environment on the reaction of water on the ceria(111) surface: A dft+u study. *J. Phys. Chem. C*, 114:14891, 2010.

[131] M. Molinari, S. C. Parker, D. C. Sayle, and M. S. Islam. Water adsorption and its effect on the stability of low index stoichiometric and reduced surfaces of ceria. *J. Phys. Chem. C*, 116:7073, 2012.

[132] Q. L. Meng, C. Lee, T. Ishihara, H. Kaneko, and Y. Tamaura. Reactivity of CeO_2-based ceramics for solar hydrogen production via a two-step water-splitting cycle with concentrated solar energy. *Int. J. Hyd. Energy*, 36:13435, 2011.

[133] C. Lee, Q. Meng, H. Kaneko, and Y. Tamaura. Solar hydrogen productivity of ceria–scandia solid solution using two-step water-splitting cycle. *J. Sol. Energy Eng.*, 1135:011062, 2013.

[134] C. Lee, Q. Meng, H. Kaneko, and Y. Tamaura. Dopant effect on hydrogen generation in two-step water splitting with CeO_2-ZrO_2MO_x reactive ceramics. *Int. J. Hydrogen Energy*, 38:15934, 2013.

[135] R. Bader, L. J. Venstrom, J. H. Davidson, and W. Lipinski. Thermodynamic analysis of isothermal redox cycling of ceria for solar fuel production. *Energy & Fuels*, 27:5533, 2013.

[136] L. J. Venstrom, N. Petkovich, S. Rudisill, A. Stein, and J. H. Davidson. The effects of morphology on the oxidation of ceria by water and carbon dioxide. *J. Sol. Energy Eng.*, 134:011005, 2012.

[137] G. Hua, L. Zhang, G. Fei, and M. Fang. Enhanced catalytic activity induced by defects in mesoporous ceria nanotubes. *J. Mater. Chem.*, 22:6851, 2012.

[138] J.Rossmeisl and W.G.Bessler. Trends in catalytic activity for sofc anode materials. *Solid State Ionics*, 178:1694, 2008.

[139] P. Singh and M. S. Hegde. $Ce_{0.67}Cr_{0.33}O_2$: A new low−temperature O_2 evolution material and H_2 generation catalyst by thermochemical splitting of water. *Chem. Mater.*, 22:762, 2010.

[140] Y. An, M. Shen, and J. Wang. Comparison of the microstructure and oxygen storage capacity modification of $Ce_{0.67}Zr_{0.33}O_2$. *J. Alloy Compd.*, 441:305, 2007.

[141] M. Zhao, M. Shen, X. Wen, and J. Wang. Ce−Zr−Sr ternary mixed ox-

ides structural characteristics and oxygen storage capacity. *J. Alloy Compd.*, 457:578, 2008.

[142] A. L. Gal, S. Abanades, N. Bion, T. L. Mercier, and V. Harle. Reactivity of doped ceria-based mixed oxides for solar thermochemical hydrogen generation via two-step water-splitting cycles. *Energy & Fuels*, 27:6068, 2013.

[143] A. L. Gal and S. Abanades. Dopant incorporation in ceria for enhanced water-splitting activity during solar thermochemical hydrogen generation. *J. Phys. Chem. C*, 116:13516, 2012.

[144] S. Abanades and A. L. Gal. CO_2 splitting by thermo-chemical looping based on $Zr_xCe_{1-x}O_2$ oxygen carriers for synthetic fuel generation. *Fuel*, 102:180, 2012.

[145] M. B. Watkins, A. S. Foster, and A. L. Shluger. Hydrogen cycle on CeO_2 (111) surfaces: Density functional theory calculations. *J. Phys. Chem. C*, 111:15337, 2007.

[146] H. Kaneko, T. Miura, H. Ishihara, S. Taku, T. Yokoyama, H. Nakajima, and Y. Tamaura. Reactive ceramics of $CeO_2{-}MO_x$ (M = Mn, Fe, Ni, Cu) for H_2 generation by two−step water splitting using concentrated solar thermal energy. *Energy*, 32:656, 2007.

[147] Z. Hu and H. Metiu. Effects of dopants on the energy of oxygen−vacancy formation at the surface of ceria: Local or global. *J. Phys. Chem. C*, 115:17898, 2011.

[148] V. Sharma, G. Pilania, G. A. Rossetti, K. Slenes, and R. Ramprasad. Comprehensive examination of dopants and defects in $BaTiO_3$. *Phys. Rev. B*, 87:134109, 2013.

[149] D. Channei, B. Inceesungvorn, N. Wetchakun, S. Phanichphant, A. Nakaruk, P. Koshy, and C. C. Sorrell. Photocatalytic activity under visible light of Fe−doped CeO_2 nanoparticles synthesized by flame spray pyrolysis. *Ceram. Int.*, 39:3129, 2013.

[150] T. Miki, T. Ogawa, M. Haneda, N. Kakuta, A. Ueno, S. Tateishi, S. Matsuura, and M. Sato. Enhanced oxygen storage capacity of cerium oxides in $CeO_2/La_2O_3/Al_2O_3$ containing precious metals. *J. Phys. Chem.*, 94:6464, 1990.

[151] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182, 2003.

[152] Eric W. Bucholz, Chang Sun Kong, Kellon R. Marchman, W. Gregory Sawyer, Simon R. Phillpot, Susan B. Sinnott, and Krishna Rajan. Data-driven model for estimation of friction coefficient via informatics methods. *Tribology Letters*, 47(2):211–221, 2012.

[153] Simone C Sieg, Changwon Suh, Timm Schmidt, Michael Stukowski, Krishna Rajan, and Wilhelm F Maier. Principal component analysis of catalytic functions in the composition space of heterogeneous catalysts. *QSAR & Combinatorial Science*, 26(4):528–535, 2007.

[154] J. E. Jackson. *A user's guide to principal components.* John Wiley and Sons, 1991.

[155] Ian Jolliffe. *Principal Component Analysis.* John Wiley and Sons, Inc., 2014.

[156] L. Breiman. Random forests. *Mach. Learn.*, 45:5–32, 2001.

[157] L. Breiman, J. Friedman, C.J. Stone, and R.A. Olshen. *Classification and*

Regression Trees. The Wadsworth and Brooks-Cole statistics-probability series. Taylor & Francis, 1984.

[158] MATLAB. *version 8.0.0.783 (R2012b)*. The MathWorks Inc., Natick, Massachusetts, 2012.

[159] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, R. Prettenhofer, P.and Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, 12:2825–2830, 2011.

[160] Ellad B. Tadmor and Ronald E. Miller. *Quantum Mechanics of Materials*, pages 153–300. Cambridge University Press, 2012.

[161] Kieron Burke. Perspective on density functional theory. *J. Chem. Phys.*, 136(15):150901, 2012.

[162] F. Matthias Bickelhaupt and Evert Jan Baerends. *Kohn-Sham Density Functional Theory: Predicting and Understanding Chemistry*, pages 1–86. John Wiley and Sons, Inc., 2007.

[163] Jörg-Rüdiger Hill, Clive M. Freeman, and Lalitha Subramanian. *Use of Force Fields in Materials Modeling*, pages 141–216. John Wiley and Sons, Inc., 2007.

[164] I. M. Torrens. *Interatomic potentials*. Academic Press Inc., 1972.

[165] J. A. Elliott. Novel approaches to multiscale modeling in materials science. *Int. Mat. Rev.*, 56:207, 2011.

[166] T. Hofmann, B. Scholkopf, and Alexander J. Smola. Kernel methods in machine learning. *Ann. Statist.*, 36(3):1171, 2008.

[167] J. Behler. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.*, 134:074106, 2011.

[168] Sönke Lorenz, Axel Groß, and Matthias Scheffler. Representing high-dimensional potential-energy surfaces for reactions at surfaces by neural networks. *Chem. Phys. Lett.*, 395(4–6):210 – 215, 2004.

[169] Zhenwei Li, James R. Kermode, and Alessandro De Vita. Molecular dynamics with on-the-fly machine learning of quantum-mechanical forces. *Phys. Rev. Lett.*, 114:096405, Mar 2015.

[170] A. M. Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950.

[171] R. Car and M. Parrinello. Unified approach for molecular dynamics and density-functional theory. *Phys. Rev. Lett.*, 55:2471–2474, Nov 1985.

[172] Albert P. Bartók, Risi Kondor, and Gábor Csányi. On representing chemical environments. *Phys. Rev. B*, 87:184115, May 2013.

[173] L. Yang, S. Dacek, and G. Ceder. Proposed definition of crystal substructure and substructural similarity. *Phys. Rev. B*, 90:054102, 2014.

[174] K. T. Schutt, H. Glawe, F. Brockherde, A. Sanna, K. R. Muller, and E. K. U. Gross. How to represent crystal structures for machine learning: Towards fast prediction of electronic properties. *Phys. Rev. B.*, 89:205118, 2014.

[175] Wojciech J. Szlachta, Albert P. Bartók, and Gábor Csányi. Accuracy and transferability of gaussian approximation potential models for tungsten. *Phys. Rev. B*, 90:104108, Sep 2014.

[176] K. R. Muller, S. Mika, G. Ratsch, K. Tsuda, and B. Scholkopf. An introduction to kernel-based learning algorithms. *IEEE Trans. Neural Networks*, 12:181, 2001.

[177] K. W. Jacobsen, J. K. Norskov, and M. J. Puska. Interatomic interactions in the effective-medium theory. *Phys. Rev. B*, 35:7423–7442, May 1987.

[178] Victor K. La Mer. Chemical kinetics. the temperature dependence of the energy of activation. the entropy and free energy of activation. *J. Chem. Phys.*, 1(5):289–296, 1933.

[179] S. W. Benson. *Thermochemical kinetics - methods for the estimation of thermochemical data and rate parameters*. Wiley, 2nd edition, 1976.

[180] Dario Alfe. Phon: A program to calculate phonons using the small displacement method. *Comp. Phys. Comm.*, 180(12):2622 – 2633, 2009.

[181] J. M. Dickey and Arthur Paskin. Computer simulation of the lattice dynamics of solids. *Phys. Rev.*, 188:1407–1418, Dec 1969.

[182] Alexander Bogicevic, Johan Strömquist, and Bengt I. Lundqvist. Low-symmetry diffusion barriers in homoepitaxial growth of al(111). *Phys. Rev. Lett.*, 81:637–640, Jul 1998.

[183] Alexander Bogicevic, Per Hyldgaard, Göran Wahnström, and Bengt I. Lundqvist. Al dimer dynamics on al(111). *Phys. Rev. Lett.*, 81:172–175, Jul 1998.

[184] Roland Stumpf and Matthias Scheffler. *Ab initio* calculations of energies and self-diffusion on flat and stepped surfaces of al and their implications on crystal growth. *Phys. Rev. B*, 53:4958–4973, Feb 1996.

[185] C.M Chang, C.M Wei, and S.P Chen. Structural and dynamical behavior of al trimer on al(111) surface. *Surface Science*, 465(1–2):65 – 75, 2000.

[186] C. M. Chang, C. M. Wei, and S. P. Chen. Self-diffusion of small clusters on fcc metal (111) surfaces. *Phys. Rev. Lett.*, 85:1044–1047, Jul 2000.

[187] Carsten Busse, Winfried Langenkamp, Celia Polop, Ansgar Petersen, Henri Hansen, Udo Linke, Peter J. Feibelman, and Thomas Michely. Dimer binding energies on fcc(1 1 1) metal surfaces. *Surface Science*, 539(1–3):L560 – L566, 2003.

[188] C. Ratsch, P. Ruggerone, and M. Scheffler. Density-functional theory of surface diffusion and epitaxial growth of metals. In M.C. Tringides, editor, *Surface Diffusion*, volume 360 of *NATO ASI Series*, pages 83–101. Springer US, 1997.

[189] Staffan Ovesson, Alexander Bogicevic, and Bengt I. Lundqvist. Origin of compact triangular islands in metal-on-metal growth. *Phys. Rev. Lett.*, 83:2608–2611, Sep 1999.