Physics-Informed Machine Learning for Modelling Defect-Driven Catalytic Phenomena

Amit Chaudhari

A thesis submitted for the degree of
Doctor of Philosophy

Cardiff University

September 2025

In loving memory of Professor Emeritus Vasant Agashe

Abstract

The growing global demand for clean and sustainable energy requires rapid advances in the atomic-level understanding of critical materials, especially transition metals, rare-earth metals and their oxides. These materials form the foundation of industrial catalysts; however, it is very challenging to construct predictive models to optimise their performance. Density functional theory (DFT) is a core method in modern computational chemistry but is limited in terms of accuracy and scalability, which prohibits the simulation of complex electronic effects, length scales and disorder of real materials.

This thesis aims to address these challenges through the application of computational methods and workflows that complement DFT and allow better understanding of experimentally observed metal oxide effects in catalysis.

In Chapter 3, I investigate the simulation of electron polarons in Nb- and W-doped TiO_2, which are important materials for photocatalysis and solar cells but poorly understood at the atomic level. I apply Hubbard corrected DFT+U to accurately simulate these materials and show that careful determination of both the Hubbard U value and projector is required to match experimental observations. Refinement of the Hubbard projector mitigates numerical instabilities and enables robust simulations across a wide range of materials, which is achieved using supervised machine learning in Chapter 4. In Chapter 5, I examine the role of metal oxide supports in enhancing the sulfur tolerance of Ni-based methane steam reforming catalysts. Combining DFT+U, grand canonical Monte Carlo sampling and machine learned interatomic potentials, I probe oxygen buffering and bulk phase transformations that govern sulfur oxidation and catalyst regeneration. The simulations rationalise experimental observations across different supports, linking atomic-scale defect chemistry with macroscopic catalytic performance.

Overall, this work demonstrates how refined DFT+U methods, machine learning and multiscale modelling can provide a pathway to more comprehensive simulations of complex catalytic materials that are far beyond current capabilities.

Acknowledgements

Completing this PhD has been one of the most challenging and rewarding endeavours of my life and it would not have been possible without the support of many incredible people. I am extremely fortunate to have such a supportive family. To my mum, Anjali; my dad, Milind; and my sister, Asmita- I am eternally grateful for your unwavering love, encouragement and belief in me, which has allowed me to navigate numerous challenges and shaped the person I have become.

I am extremely grateful to my supervisor, Dr Andrew Logsdail, from whom I have gained an immense amount of technical knowledge, together with steady mentorship and constant encouragement. I have thoroughly enjoyed the PhD and feel fortunate to have been part of such an excellent working environment, which is testament to Andrew's outstanding supervision- a sentiment I know is shared by all members of the group.

Over the past four years, I have been lucky to be surrounded by very talented researchers in the Theory and Molecular Modelling Group, and the wider community in the Cardiff Catalysis Institute. I am grateful for the friendship and collaboration of all the researchers, with too many to name here, but I would especially like to thank Dr Kushagra Agrawal, Dr Pavel Stishenko, Akash Hiregange, Dr Gabriel Bramley, Matthew Robinson, Oscar van Vuren and Zhongwei Lu. I am also very grateful to Dr Andrea Folli, who provided the experimental characterisation data (EPR spectroscopy) presented in Chapter 3.

I have had the pleasure of meeting and working with a lot of researchers across academic and industrial institutes as part of the Prosperity Partnership *Sustainable Catalysis for Clean Growth*, which has given me invaluable opportunities and inspiration as an early-career researcher hoping to make a societal-level impact through the development of science. I am thankful to all members of the Prosperity Partnership across the partner institutes of Cardiff University, the University of Manchester, Johnson Matthey and bp, particularly the fellow researchers on the *Advanced Methods* project bundle. I am thankful for mentorship from Dr James Paterson and Dr Corneliu Buda from bp, and Dr Rui Zhang from Johnson Matthey. I am grateful to Dr Misbah Sarwar, Dr Christopher Hawkins and Dr Stephen Poulston from Johnson Matthey, who supervised my secondment at Johnson Matthey Technology Centre (JMTC), during which we investigated strategies for mitigating catalyst poisoning, as outlined in Chapter 5. I would like to thank all the staff at JMTC for being so accommodating during my secondment, particularly Dr Andrew Steele for assisting with the experimental setup and testing, as well as Dr Gregory Goodlet, Dr Riho Green and Jason Raymond from the Advanced Characterisation Department for assisting with experimental characterisation (SEM, XPS and ICP, respectively), as outlined in Chapter 5.

I am thankful to the community of users and developers of the Fritz-Haber Institute *Ab Initio* Materials Simulation software (FHI-aims), who were always generous in addressing queries and hosted fascinating workshops that showcased the inspiring work of contributing researchers worldwide. I

am especially thankful to Professor Harald Oberhofer, Dr Matthias Kick and Maximilian Brand for valuable scientific discussions regarding the implementation of DFT+*U* in FHI-aims.

I am grateful for funding by the Prosperity Partnership project *Sustainable Catalysis for Clean Growth*, funded by the UK Engineering and Physical Sciences Research Council (EPSRC), bp through the bp International Centre for Advanced Materials (bp-ICAM) and Johnson Matthey plc in collaboration with Cardiff University and The University of Manchester (EPSRC grant number EP/V056565/1). I am grateful for funding by the Collaborative Computational Project Number 5 (CCP5) as part of a Postgraduate Industrial Secondment, which facilitated computational and experimental collaboration with partners at Johnson Matthey (EPSRC grant number EP/V028537/1). I am grateful for computational resources and support from the Supercomputing Wales project, which is part-funded by the European Regional Development Fund (ERDF) *via* the Welsh Government; and the UK National Supercomputing Services ARCHER and ARCHER2, accessed *via* membership of the Materials Chemistry Consortium, which is funded by Engineering and Physical Sciences Research Council (EP/L000202/1, EP/R029431/1, and EP/T022213/1).

"*The first gulp from the glass of natural sciences will turn you into an atheist, but at the bottom of the glass God is waiting for you.*"

Werner Heisenberg

Contents

List of Figures

List of Tables

List of Abbreviations

ASE	Atomic Simulation Environment
BO	Bayesian optimisation
DFT	Density functional theory
DFT+*U*	Hubbard corrected density functional theory
EPR	Electron paramagnetic resonance spectroscopy
FHI-aims	Fritz-Haber Institute *Ab Initio* Materials Simulation software
FLL	Fully localised limit
GCMC	Grand canonical Monte Carlo sampling
HI-SISSO	Hierarchical Sure Independence Screening and Sparsifying Operator
ICP	Inductively coupled plasma
MBTR	Many-body tensor representation
MLIP	Machine learned interatomic potential
MSR	Methane steam reforming
NAO	Numerical atom-centred orbital
NTO	Nb-doped TiO_2
OMC	Occupation matrix control
OMR	Occupation matrix release
PCA	Principal component analysis
REO	Rare-earth metal oxide
SCF	Self-consistent field
SEM	Scanning electron microscopy
SIE	Coulomb self-interaction error
SISSO	Sure Independence Screening and Sparsifying Operator
SR	Symbolic regression
SuSMoST	Surface Science Modelling and Simulation Toolkit
SVM	Support vector machine
TCO	Transparent conducting oxide
TMO	Transition metal oxide
TPD-MS	Temperature-programmed desorption- mass spectrometry
WTO	W-doped TiO_2
XPS	X-ray photoelectron spectroscopy
XRD	X-ray diffraction
ZORA	Zeroth order regular approximation

Chapter 1

Introduction

The following introduction aims to provide essential context for this thesis, outlining the motivations for constructing predictive models of catalytic materials and highlighting the key challenges that motivate the methods considered in the subsequent chapters.

Sustainable Catalysis and the Role of Defects

Human society continues to consume finite resources at an unsustainable rate, leading to rising energy costs and insecurity, as well as widespread destruction of the natural environment. To halt and reverse this trajectory, there is an urgent need for new technologies that enable the sustainable production of energy vectors (*i.e.*, fuels) from renewable feedstocks, alongside advanced technologies for energy conversion and storage. At the foundation of these technologies lie critical materials based on transition metals, rare-earth metals and their oxides, which demand an accurate atomic-level understanding to deliver the performance and economic viability required for a sustainable energy future.

Heterogeneous catalysis plays a central role in modern chemical processes, enabling the conversion of stable and otherwise hard-to-activate feedstocks into valuable products. Catalysts provide an alternative reaction pathway with a reduced activation energy barrier, thereby dramatically enhancing the efficiency and selectivity of chemical transformations. Importantly, catalysts are not consumed in the reaction and can therefore be used repeatedly, making them indispensable for both large-scale industrial applications and the development of sustainable energy technologies. Transition metal oxides (TMOs) and rare-earth metal oxides (REOs) are widely used as support materials in heterogeneous catalysis to enhance the reactivity of metal nanoparticles *via* complex metal-support interactions. Such interactions include the formation of point defects, *e.g.*, oxygen vacancies and substitutional dopants, which can result in charge transfer between the catalyst and support due to the formation of trapped electrons, *i.e.*, electron polarons, where excess electrons localise on metal cations causing a reduction in oxidation states and distortion of the surrounding lattice.

By controlling the formation of electron polarons, catalytic activity and selectivity can be tuned to achieve controllable CO_2 conversion using CeO_2-based catalysts, [1] H_2 production using MoO_3-based catalysts [2] and the production of sustainable aviation fuel *via* the Fischer Tropsch process using TiO_2-based catalysts. [3] The energetic favourability of defect formation in metal oxide supports can also aid the regeneration of poisoned catalysts, *e.g.*, *via* oxygen buffering from reducible support materials like CeO_2 and Y_2O_3, where lattice oxygen can migrate to catalyst active sites that are poisoned with S and C before removing them *via* oxidation to SO_2 and CO_2, respectively. [4–12]

Beyond catalysis, the role of defects in TMOs and REOs extends to a much wider range of energy technologies, where the tuneable conductivities of electrons and ions is crucial. For example, tuneable charge carrier mobility can be exploited for the development of photovoltaics and lithium ion battery cathodes using dopants such as Nb and Mg to optimise the electrical performance of TiO_2 and $LiCoO_2$, respectively. [13, 14]

Grand Challenges and Opportunities in Predictive Modelling

Constructing predictive models of defect-driven phenomena in TMOs and REOs that faithfully captures the behaviour of real materials remains extremely challenging. Density functional theory (DFT) is a core method in modern computational chemistry and materials modelling; however, its application for simulating defects in TMOs and REOs is hindered by well-known limitations. [15, 16] In particular, traditional exchange-correlation density functionals suffer from the Coulomb self-interaction error (SIE) when modelling electron correlation between localised d or f orbitals. The SIE manifests as systematic errors in the DFT-predicted properties of materials with partially filled d or f orbitals, *e.g.*, underestimated insulator band gaps, inaccurate lattice parameters [17] and inaccurate formation energies of point defects and electron polarons. [18, 19] These challenges motivate the need to understand how to efficiently correct DFT to achieve the accuracy of higher levels of theory, such as hybrid-DFT, which is generally more accurate but scales poorly at $O(N^4)$, where N is the number of electrons in the system (Figure 1.1). Even with these improvements in accuracy, the computational cost of DFT scales on the order of $O(N^3)$, which restricts its applicability to modelling large supercells (which are necessary to avoid artificial interactions between periodic images of defects) or sampling high-dimensional chemical space. The necessity for large simulations demands advances in the computational efficiency of DFT, *e.g.*, by parameterising highly efficient surrogate models such as classical interatomic potentials, which achieve linear scaling at $O(N)$ (Figure 1.1).

The limitations in accuracy and computational efficiency of DFT together expose a grand challenge in computational materials science that motivates this thesis: defect-driven phenomena in TMOs and REOs are governed by highly localised electronic structure effects that demand an accurate treatment of electron correlation, whilst their technological relevance requires simulation methods that are efficient and widely transferable across chemical space. The work presented in this thesis therefore develops and applies a sequence of complementary methods that progress from understanding and correcting electronic structure errors in model systems, to generalising these corrections across chemical space, and finally to exploiting them in simulation workflows to support experimental observations, uncover future optimisation strategies and demonstrate capabilities that extend far beyond the model systems selected.

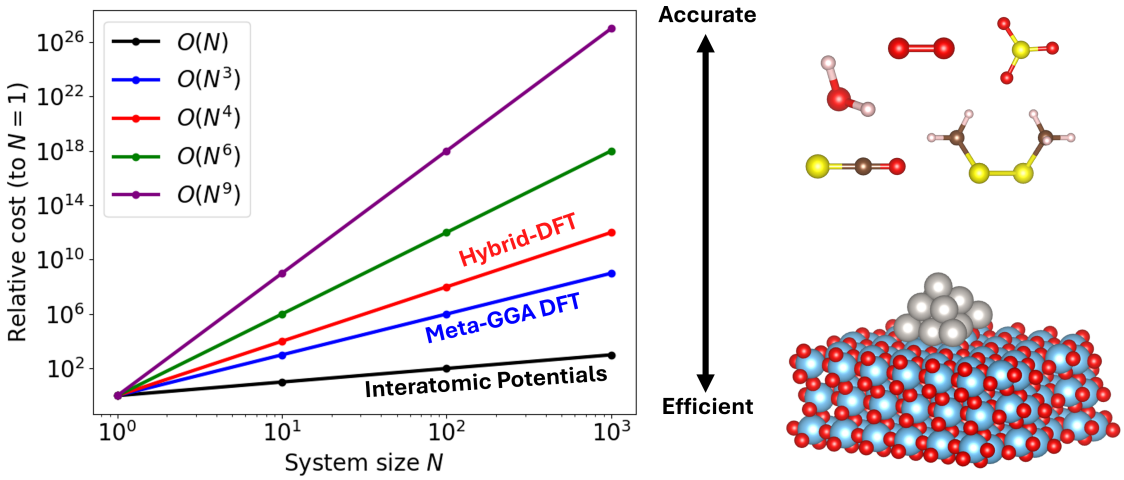


Figure 1.1: The relative computational cost as a function of system size N for methods with different formal scaling, including linear $O(N)$ with interatomic potentials, cubic $O(N^3)$ with meta-GGA density functional theory, quartic $O(N^4)$ with hybrid-DFT, sextic $O(N^6)$ and nonic $O(N^9)$ with wavefunction methods. The theoretical background for the different computational methods is outlined in Chapter 2. The steep growth in computational cost highlights the challenge of simulating defect-rich TMOs and REOs at realistic length scales. The high scaling methods are typically used for accurate simulations for small systems like molecules, whereas the lower scaling methods as typically used for large-scale simulations of heterogeneous catalysts.

I begin in Chapter 3 by investigating the challenges in simulating defects and polarons in Nb- and W-doped TiO_2, which is important knowledge for the design of transparent conducting oxides (TCOs) for high efficiency photovoltaics and photocatalysts with tuneable reaction selectivities. There is currently no clear explanation in the theoretical or experimental literature as to why these materials exhibit electronic properties; particularly dopant atom oxidation states and electronic conductivities, that vary with the TiO_2 polymorph, *i.e.*, anatase or rutile. Therefore, I apply Hubbard corrected density functional theory (DFT+U) to mitigate the SIE in these systems *via* a tuneable *ad-hoc* energy correction applied selectively to localised orbitals in the system. [20] DFT+U is popular because it adds minimal computational cost compared to standalone DFT, [21] whilst achieving the accuracy of higher levels of theory such as hybrid-DFT; however, this requires very careful determination of multiple simulation parameters including the Hubbard U value and projector. [22]

In Chapter 4, I expand on the challenge of determining accurate DFT+U simulation parameters to enable self-consistent simulations of defects in TMOs and REOs. Chapter 4 begins with an in-depth analysis of the deficiencies of default Hubbard projectors, before outlining the development of a novel first-principles method for DFT+U parameterisation that uses supervised machine learning to generalise predictive accuracy across a broad range of TMOs and REOs. Chapter 5 extends the concepts introduced in Chapters 3 and 4 by investigating the role of metal oxide supports in controlling the sulfur tolerance of industrially relevant Ni catalysts for H_2 production. The work highlights how accurate DFT+U predictions of the energetics of defect formation in catalyst support materials and scalable simulations of adsorption on catalyst surfaces using statistical sampling and machine learned interatomic potentials, can inform catalyst design strategies.

Together, the work demonstrates how advanced methods and workflows can complement DFT to expand the predictive power of first-principles simulations and provide essential insight into complex defect-driven phenomena that underpin sustainable energy technologies.

References

(1) K. Chang, H. Zhang, M.-j. Cheng and Q. Lu, Application of Ceria in CO_2 Conversion Catalysis, *ACS Catal.* 2020, **10** 613–631.

(2) A. Avani and E. Anila, Recent advances of MoO_3 based materials in energy catalysis: Applications in hydrogen evolution and oxygen evolution reactions, *Int. J. Hydrog. Energy* 2022, **47** 20475–20493.

(3) M. Lindley, P. Stishenko, J. W. M. Crawley, F. Tinkamanyire, M. Smith, J. Paterson, M. Peacock, Z. Xu, C. Hardacre, A. S. Walton, A. J. Logsdail and S. J. Haigh, Tuning the Size of TiO_2-Supported Co Nanoparticle Fischer–Tropsch Catalysts Using Mn Additions, *ACS Catal.* 2024, **14** 10648–10657.

(4) U. Oemar, K. Hidajat and S. Kawi, Pd–Ni catalyst over spherical nanostructured Y_2O_3 support for oxy-CO_2 reforming of methane: Role of surface oxygen mobility, *Int. J. Hydrogen Energy.* 2015, **40** 12227–12238.

(5) Z. Li and K. Sibudjing, Facile Synthesis of Multi-Ni-Core@Ni Phyllosilicate@CeO_2 Shell Hollow Spheres with High Oxygen Vacancy Concentration for Dry Reforming of CH_4, *ChemCatChem* 2018, **10** 2994–3001.

(6) D. Guo, Y. Lu, Y. Ruan, Y. Zhao, Y. Zhao, S. Wang and X. Ma, Effects of extrinsic defects originating from the interfacial reaction of CeO_{2-x}-nickel silicate on catalytic performance in methane dry reforming, *Appl. Catal. B: Environ.* 2020, **277** 119278.

(7) L. Pino, C. Italiano, A. Vita, M. Laganà and V. Recupero, $Ce_{0.70}La_{0.20}Ni_{0.10}O_{2-}$ catalyst for methane dry reforming: Influence of reduction temperature on the catalytic activity and stability, *Appl. Catal. B: Environ.* 2017, **218** 779–792.

(8) H. Wang, X. Dong, T. Zhao, H. Yu and M. Li, Dry reforming of methane over bimetallic Ni-Co catalyst prepared from $La(Co_xNi_{1-x})_{0.5}Fe_{0.5}O_3$ perovskite precursor: Catalytic activity and coking resistance, *Appl. Catal. B: Environ.* 2019, **245** 302–313.

(9) G.-R. Hong, K.-J. Kim, S.-Y. Ahn, B.-J. Kim, H.-R. Park, Y.-L. Lee, S. S. Lee, Y. Jeon and H.-S. Roh, Sulfur-Resistant CeO_2-Supported Pt Catalyst for Waste-to-Hydrogen: Effect of Catalyst Synthesis Method, *Catalysts* 2022, **12** 1670.

(10) Y.-L. Lee, K.-J. Kim, G.-R. Hong, S.-Y. Ahn, B.-J. Kim, H.-R. Park, S.-J. Yun, J. W. Bae, B.-H. Jeon and H.-S. Roh, Sulfur-Tolerant Pt/CeO_2 Catalyst with Enhanced Oxygen Storage Capacity by Controlling the Pt Content for the Waste-to-Hydrogen Processes, *ACS Sustain. Chem. Eng.* 2021, **9** 15287–15293.

(11) S. d. S. Eduardo, J. P. Mendonça, P. N. Romano, J. M. A. R. de Almeida, G. Machado and M. A. S. Garcia, Tailoring Ceria-Based Nanocatalysts for Enhanced Performance in Steam Reforming Processes: Exploring Fundamentals and Morphological Modulations, *Hydrogen* 2023, **4** 493–522.

(12) M. A. Ocsachoque, J. I. Eugenio Russman, B. Irigoyen, D. Gazzoli and M. G. González, Experimental and theoretical study about sulfur deactivation of Ni/CeO_2 and Rh/CeO_2 catalysts, *Mater. Chem. Phys.* 2016, **172** 69–76.

(13) A. Folli, J. Z. Bloh, A. Lecaplain, R. Walker and D. E. Macphee, Properties and photochemistry of valence-induced-Ti^{3+} enriched (Nb,N)-codoped anatase TiO_2 semiconductors, *Phys. Chem. Chem. Phys.* 2015, **17** 4849–4853.

(14) J. A. Santana, J. Kim, P. R. C. Kent and F. A. Reboredo, Successes and failures of Hubbard-corrected density functional theory: The case of Mg doped $LiCoO_2$, *J. Chem. Phys.* 2014, **141** 164706.

(15) W. Kohn and L. J. Sham, Self-Consistent Equations Including Exchange and Correlation Effects, *Phys. Rev.* 1965, **140** A1133–A1138.

(16) W. Kohn, A. D. Becke and R. G. Parr, Density Functional Theory of Electronic Structure, *J. Phys. Chem.* 1996, **100** 12974–12980.

(17) N. L. Nguyen, N. Colonna, A. Ferretti and N. Marzari, Koopmans-Compliant Spectral Functionals for Extended Systems, *Phys. Rev. X* 2018, **8** 021051.

(18) J. P. Perdew and A. Zunger, Self-interaction correction to density-functional approximations for many-electron systems, *Phys. Rev. B* 1981, **23** 5048–5079.

(19) M. Reticcioli, U. Diebold and C. Franchini, Modeling polarons in density functional theory: lessons learned from TiO_2, *J. Condens. Matter Phys.* 2022, **34** 204006.

(20) H. J. Kulik, Perspective: Treating electron over-delocalization with the DFT+U method, *J. Chem. Phys.* 2015, **142** 240901.

(21) M. Capdevila-Cortada, Z. Łodziana and N. López, Performance of DFT+U Approaches in the Study of Catalytic Materials, *ACS Catal.* 2016, **6** 8370–8379.

(22) D. S. Lambert and D. D. O'Regan, Use of DFT + U + J with linear response parameters to predict non-magnetic oxide band gaps with hybrid-functional accuracy, *Phys. Rev. Res.* 2023, **5** 013160.

Chapter 2

Theoretical Background

This chapter introduces the theoretical framework of this thesis, beginning with interatomic potentials and density functional theory (DFT), including the Hubbard corrected DFT+U method. The role of supervised machine learning is then outlined for electronic structure method development and a concluding overview is provided for multiscale modelling approaches that bridge atomistic simulations with experimental length scales.

2.1 Modelling Molecules and Materials from First-Principles

Accurately predicting the properties of molecules and solids is a central goal of theoretical and computational chemistry. Historically, atoms were represented as point masses that interact through empirically constructed interatomic potentials. Well-known examples include the Lennard-Jones potential, [1] which describes the potential energy between a pair of atoms, $V^{\mathrm{LJ}}(r)$, including attractive long-range van der Waals forces and repulsive short-range interactions. The Lennard-Jones potential is described by the functional form:

$$V^{\mathrm{LJ}}(r) = 4\varepsilon \left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} \right] \tag{2.1}$$

where r is the interatomic distance, ε is the depth of the potential well (corresponding to the strength of the attractive interaction) and σ is the distance at which the interatomic potential crosses zero. The r^{-12} term in Equation 2.1 represents repulsive short-range interactions and the r^{-6} term represents attractive long-range interactions (Figure 2.1). The total potential energy (E^{LJ}) is obtained by summing the pair potential $V^{\mathrm{LJ}}(r_{ij})$ over all distinct pairs of atoms (i, j), while the forces are evaluated as the analytical derivative of the energy with respect to the atomic positions, *i.e.*, the negative gradient of E^{LJ}.

Classical interatomic potentials remain widely used for large-scale molecular mechanics simulations due to the low cost of evaluating energies and forces; however, their transferability is inherently limited. The constants in these potentials are typically parameterised empirically for specific systems and cannot describe phenomena such as bond breaking and charge transfer that is foundational for the understanding of catalytic materials. These shortcomings necessitate approaches that explicitly incorporate the quantum mechanical behaviour of electrons, *i.e.*, *ab initio* approaches, where material properties are computed from first principles with no empirical fitting beyond fundamental physical constants. Nevertheless, the concept of interatomic potentials remains invaluable, as demonstrated

by recent advances in machine learned interatomic potentials (MLIPs), which leverage highly parameterised neural networks to move beyond the rigid analytical forms in Equation 2.1 and describe chemically reactive systems (discussed in Section 2.5).

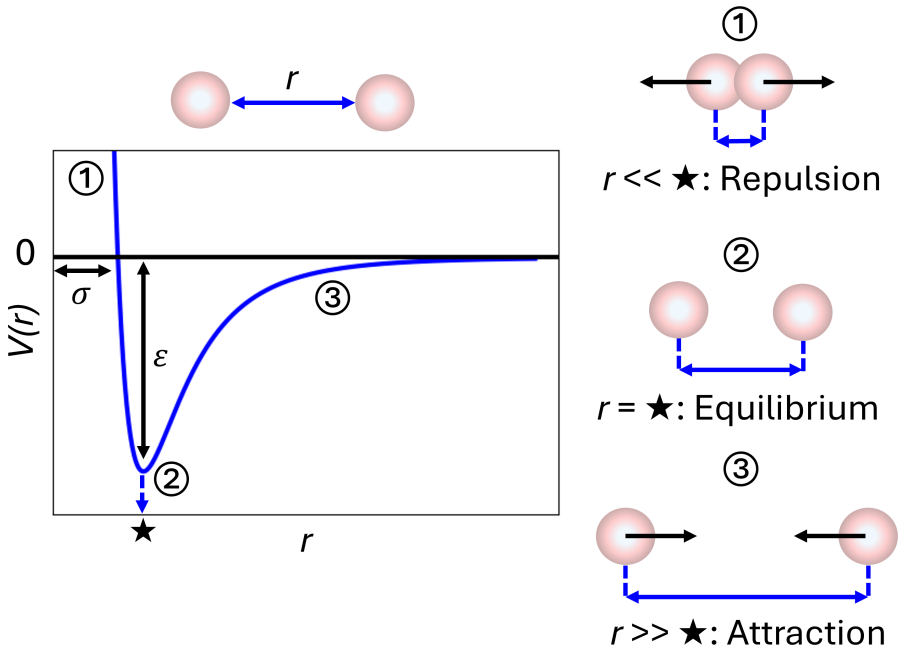


Figure 2.1: Schematic of a generic pair potential $V(r)$ as a function of interatomic distance r. At very short separations ($r \ll \star$), strong Pauli repulsion dominates (1). At the equilibrium separation ($r = \star$), the potential reaches its minimum, corresponding to a stable bond length (2). At large separations ($r \gg \star$), the interaction becomes attractive but weak, approaching zero as the atoms move apart (3).

At the foundation of all quantum mechanical descriptions lies the time-independent Schrödinger equation, [2] which is defined as:

$$\hat{H}\Psi(\mathbf{r}, \mathbf{R}) = E\Psi(\mathbf{r}, \mathbf{R}) \tag{2.2}$$

where $\hat{H}$ is the Hamiltonian operator, $\Psi(\mathbf{r}, \mathbf{R})$ is the many-body wavefunction that depends on the co-ordinates of all electrons $\mathbf{r}$ and nuclei $\mathbf{R}$, and E is the total energy of the system. The Hamiltonian can be expanded as a sum of several contributions in the form:

$$\hat{H} = \hat{T}_e + \hat{T}_n + \hat{V}_{en} + \hat{V}_{ee} + \hat{V}_{nn} \tag{2.3}$$

where $\hat{T}_e$ and $\hat{T}_n$ denote the electronic and nuclear kinetic energy operators, $\hat{V}_{en}$ describes electron-nuclear attraction, $\hat{V}_{ee}$ describes electron-electron repulsion and $\hat{V}_{nn}$ describes nuclear-nuclear repulsion. In principle, solving Equation 2.2 yields the exact ground state properties of any quantum system; but in practice, tremendous difficulty arises from the exponential complexity of the many-body wavefunction. For a system with N electrons, the wavefunction Ψ is defined over a $3N$-dimensional co-ordinate space. The Born-Oppenheimer approximation simplifies the formulation of the many-body wavefunction by treating the nuclei as stationary on the timescales of electronic motion, [3] therefore the Hamiltonian in Equation 2.3 can be reduced to the electronic Hamiltonian ($\hat{H}_e$), which

is defined as:

$$\hat{H}_e = \hat{T}_e + \hat{V}_{en} + \hat{V}_{ee} + V_{nn} \tag{2.4}$$

where V_{nn} is the classical nuclear-nuclear repulsion (constant), that depends only on the fixed nuclear co-ordinates. This formulation allows the electronic Schrödinger equation to be solved independently for a given set of nuclear positions, yielding the potential energy surface on which the nuclei move. However, even under this approximation, the electronic Schrödinger equation remains intractable for systems with more than a few electrons. This so-called *curse of dimensionality* underpins the challenge of electronic structure theory: the exact solution scales exponentially with N, while realistic catalytic materials contain an astronomical number of electrons, *e.g.*, $\sim 10^{26}$ in 1 mole of TiO_2.

Early strategies to approximate solutions of the electronic Schrödinger equation attempted to reduce the complexity of the many-body problem by introducing simplified models of electron-electron interactions. The Thomas-Fermi (TF) theory was one of the first density-based models, which expressed the total energy directly in terms of the electron density, with the kinetic energy approximated from the homogeneous electron gas. [4] This made the approach extremely simple with essentially linear scaling in system size. However, TF theory neglected electron exchange and correlation effects entirely and could not describe chemical bonding, making it qualitatively unreliable for molecules and solids. The Hartree method improved upon TF theory by introducing a mean-field description in which the many-electron wavefunction is represented as a product of single-electron orbitals. [5] This allowed a self-consistent treatment of electron-electron repulsion at relatively modest computational cost, with scaling of approximately $O(N^2)$. Despite this improvement, the Hartree method neglected the antisymmetry of the wavefunction required by the Pauli principle, leading to inaccurate electronic structures. The Hartree-Fock (HF) method resolved this limitation by enforcing antisymmetry, thus providing an exact description of electron exchange interactions and representing a major advancement over TF and Hartree approaches. [6] However, HF neglects electron correlation effects, leading to systematic errors in geometric, electronic and energetic properties. Moreover, the computational scaling of HF is significantly higher at $O(N^4)$, which restricts its applicability to relatively small systems.

The strengths and shortcomings of the TF, Hartree and HF methods motivated the development of new frameworks capable of incorporating electron exchange and correlation in a computationally efficient manner.

2.2 Density Functional Theory

Density functional theory (DFT) provides a computationally tractable framework for modelling the ground state properties of many-electron systems by recasting the problem of solving the electronic Schrödinger equation in terms of the electron density rather than the many-body wavefunction. The foundation of DFT was established by the Hohenberg-Kohn theorems (HKTs), which prove that the ground state electron density $\rho(\mathbf{r})$ uniquely determines all properties of a system and that there exists a variational principle to obtain the ground state energy as a functional of $\rho(\mathbf{r})$. [7] However, the HKTs do not specify the explicit form of the universal energy functional.

2.2.1 Kohn-Sham Formalism

The Kohn-Sham (KS) formalism introduces a fictitious system of non-interacting electrons that shares the same ground-state density as the interacting many-electron system. [8] Within this framework, the total energy of the real system is expressed as a functional of the electron density:

$$E[\rho] = T_s[\rho] + E_{\mathrm{H}}[\rho] + E_{\mathrm{ext}}[\rho] + E_{\mathrm{xc}}[\rho] \tag{2.5}$$

where $T_s[\rho]$ is the kinetic energy of the non-interacting reference system, $E_{\mathrm{H}}[\rho]$ is the classical Hartree electron-electron repulsion, $E_{\mathrm{ext}}[\rho]$ is the external potential energy due to the nuclei and $E_{\mathrm{xc}}[\rho]$ is the exchange-correlation energy. From the total energy functional, the KS effective potential is defined as:

$$v_{\mathrm{eff}}(\mathbf{r}) = v_{\mathrm{ext}}(\mathbf{r}) + v_{\mathrm{H}}(\mathbf{r}) + v_{\mathrm{xc}}(\mathbf{r}) \tag{2.6}$$

where $v_{\mathrm{ext}}(\mathbf{r})$ is the external potential due to the nuclei, $v_{\mathrm{H}}(\mathbf{r})$ is the Hartree potential and $v_{\mathrm{xc}}(\mathbf{r})$ is the exchange-correlation potential (discussed in Section 2.2.2). The total energy functional is minimised numerically with respect to the Kohn-Sham orbitals, which are used to construct the electron density using:

$$\rho(\mathbf{r}) = \sum_{i}^{\mathrm{occ}} |\phi_i(\mathbf{r})|^2 \tag{2.7}$$

where $\phi_i(\mathbf{r})$ denotes the i^{th} Kohn-Sham orbital, with their squared modulus describing the probability density of finding an electron at position $\mathbf{r}$. Minimising the total energy functional with respect to the Kohn-Sham orbitals leads to an effective single-particle Schrödinger equation:

$$\hat{H}_{\mathrm{KS}}\phi_i(\mathbf{r}) = \varepsilon_i\phi_i(\mathbf{r}), \qquad \hat{H}_{\mathrm{KS}} = -\frac{\hbar^2}{2m_e}\nabla^2 + v_{\mathrm{eff}}(\mathbf{r}) \tag{2.8}$$

where ε_i denotes the KS orbital eigenvalues and $-\frac{\hbar^2}{2m_e}\nabla^2$ denotes the kinetic energy operator, where $\hbar$ is the reduced Planck constant and m_e the electron mass. Because both v_{H} and v_{xc} depend on the electron density, Equation (2.8) is a non-linear partial differential equation requiring the determination of a self-consistent solution to compute the ground state density and total energy *via* iteratively optimising the electron density until input and output electron densities agree within a chosen threshold. In addition to the ground state energy, the same framework also allows the calculation of atomic forces *via* derivatives of the total energy with respect to nuclear positions.

2.2.2 Exchange-Correlation Density Functionals

The accuracy and applicability of a DFT calculation critically depends on the approximate exchange-correlation functional. Meta-GGA exchange-correlation functionals represent a critical advancement beyond local density approximation (LDA) and generalised gradient approximation (GGA) functionals, as they incorporate not only the electron density (ρ) and its gradient ($\nabla\rho$), but also the kinetic energy density (τ) in the evaluation of the exchange-correlation energy (E_{xc}) in Equation 2.5:

$$E_{\mathrm{xc}}[\rho,\, \nabla\rho,\, \tau] = \int \epsilon_{\mathrm{xc}}^{\mathrm{LDA}}[\rho] \cdot F_{\mathrm{xc}}[\rho, \nabla\rho,\, \tau]\, d\mathbf{r} \tag{2.9}$$

where $\epsilon_{\mathrm{xc}}^{\mathrm{LDA}}$ is the LDA exchange-correlation energy density per volume, F_{xc} is an exchange-correlation enhancement factor that defines the mapping from LDA to meta-GGA approximations and τ is a non-local property obtained *via* evaluating the gradient of Kohn-Sham orbitals (ϕ):

$$\tau = \frac{1}{2} \sum_i f_i \, |\nabla \phi_i|^2 \tag{2.10}$$

where f_i is the occupation number of the Kohn-Sham orbital ϕ_i. Once E_{xc} is evaluated for a given electron density, self-consistent optimisation towards the ground state is achieved using the exchange-correlation potential (v_{xc}), defined as the functional derivative of E_{xc} with respect to the electron density: [9]

$$v_{\mathrm{xc}} = \frac{\delta E_{\mathrm{xc}}[\rho, \, \nabla\rho, \, \tau]}{\delta \rho} = \frac{\partial \epsilon_{\mathrm{xc}}}{\partial \rho} - 2\nabla \cdot \left(\frac{\partial \epsilon_{\mathrm{xc}}}{\partial \sigma} \nabla \rho \right) + \frac{\partial \epsilon_{\mathrm{xc}}}{\partial \tau} \cdot \frac{\delta \tau}{\delta \rho} \tag{2.11}$$

where σ denotes the contracted gradient ($|\nabla\rho|^2$), ∂ denotes partial derivatives with respect to explicit variables of ϵ_{xc} and δ denotes functional derivatives. In theory, v_{xc} is a local multiplicative potential, *i.e.*, acting pointwise on the electron density; however, in practice the explicit dependence of ϵ_{xc} on the orbital-dependent τ^{DFT} makes the term $\frac{\partial \epsilon_{\mathrm{xc}}}{\partial \tau} \cdot \frac{\delta \tau}{\delta \rho}$ difficult to evaluate, resulting in a non-multiplicative, orbital-dependent potential which is evaluated using the generalised Kohn-Sham (gKS) scheme. [10] The additional τ-dependence of E_{xc} and v_{xc} allows meta-GGA exchange-correlation functionals to accurately capture key features such as non-local exchange effects and intermediate-range correlation, resulting in improved accuracy in the predicted geometric, electronic and energetic properties of molecules and solids. [11]

2.2.3 Numerical Implementation

In practice, solving the Kohn-Sham equations requires several numerical approximations, particularly in the representation of the Kohn-Sham orbitals that construct the electron density and the handling of periodic boundary conditions.

Basis Sets

The Kohn-Sham orbitals are expressed as linear combinations of basis functions:

$$\phi_i(\mathbf{r}) = \sum_j c_{ji} \, \chi_j(\mathbf{r}) \tag{2.12}$$

where $\chi_j(\mathbf{r})$ are the basis functions, c_{ji} are the linear expansion coefficients, i indexes the Kohn-Sham orbitals and j indexes the basis functions. The choice of basis set strongly influences both the accuracy and computational efficiency of a calculation. All work in this thesis uses a numerical atom-centred orbital (NAO) basis set, as implemented in the Fritz-Haber Institute *ab initio* materials simulation (FHI-aims) software [9], where basis functions are constructed from products of a radial function and a spherical harmonic. Each basis function takes the form:

$$\chi_{jlm}(\mathbf{r}) = \frac{u_{jl}(r)}{r} \, Y_{lm}(\theta, \phi) \tag{2.13}$$

where $\frac{u_{jl}(r)}{r}$ is the radial component of the basis function, with $u_{jl}(r)$ denoting the j-th radial function for angular momentum channel l. This radial part is numerically tabulated as an all-electron function and encodes the oscillatory behaviour characteristic of atomic orbitals (Figure 2.2).

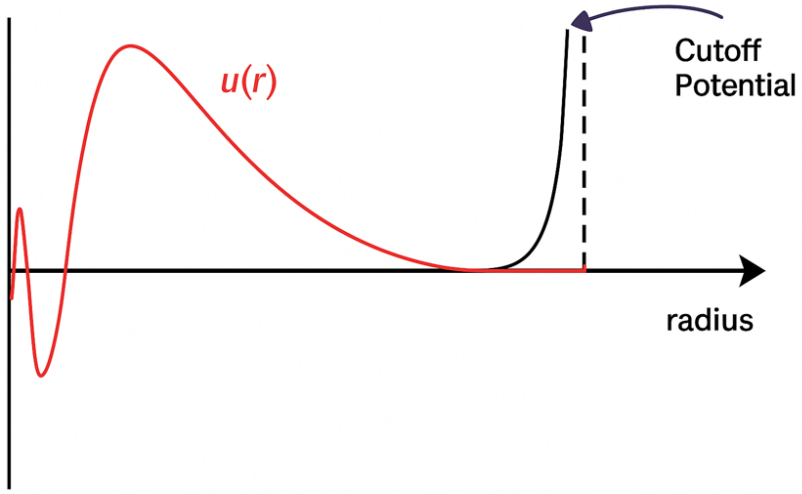


Figure 2.2: Schematic illustration of a numerical atom-centred orbital (NAO) basis function which is numerically tabulated as an all-electron function on a dense logarithmic grid, rather than assuming an approximate analytical form. [9] $u(r)$ denotes the radial component of the basis function. A cutoff potential is used to localise the basis function within a finite radius from the atomic centre, preventing long radial function tails and ensuring computational efficiency for simulating large systems. [9] *This figure is adapted from* [9].

The angular dependence of the basis function is described by the spherical harmonic $Y_{lm}(\theta, \phi)$, with degree l and order m. The degree l determines the angular character of the orbital: $l = 0$ corresponds to s-orbital-like functions, $l = 1$ to p-orbital-like functions, $l = 2$ to d-orbital-like functions and $l = 3$ to f-orbital-like functions, etc. For each l, the order m defines the magnetic quantum number, yielding $(2l + 1)$ functions for each l. The basis sets in FHI-aims are hierarchically constructed, providing a systematic path to improved accuracy. They are generated from non–spin-polarised DFT calculations of symmetric dimers *via* an iterative selection process in which additional basis functions are introduced and tested for convergence of the DFT total energy. [9] The outcome is the tiered basis sets and corresponding integration grids (light, intermediate and tight), which provide the user with easy-to-use default settings whilst enabling flexibility to systematically adjust accuracy and computational cost according to the requirements of a given study. [9]

Periodic Boundary Conditions

Periodic boundary conditions (PBCs) are used to model an infinite crystal using a finite unit cell, where the wavefunction satisfies Bloch's theorem, [12] which states that the Kohn-Sham orbitals can be written as:

$$\phi_{i\mathbf{k}}(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}} u_{i\mathbf{k}}(\mathbf{r}) \tag{2.14}$$

where $\mathbf{k}$ is a wavevector in the reciprocal lattice and $u_{i\mathbf{k}}(\mathbf{r})$ is a function with the periodicity of the crystal unit cell. The index i labels the band, while $\mathbf{k}$ labels the crystal momentum. The set of all

possible wavevectors **k** forms the reciprocal space of the crystal, with the primitive cell in this reciprocal lattice defined as the first Brillouin zone. Physical quantities such as the total energy and electron density require integration over the Brillouin zone; however, these integrals are approximated to save computational resources using a discrete sum over a finite grid of **k**-points. As a result, the accuracy of the calculation depends on how finely the Brillouin zone is sampled. A denser **k**-point grid yields more accurate results but at increased computational cost; therefore, it is necessary to systematically converge the predicted values of observable properties with respect to the **k**-point spacing, defined as the distance in reciprocal space between adjacent sampling points along a reciprocal lattice vector.

2.3 Beyond-DFT Methods

2.3.1 Coulomb Self-Interaction and Hybrid-DFT

A key limitation of local and semi-local DFT is the Coulomb self-interaction error (SIE), which originates from the fact that the Hartree energy counts the interaction of each electron with the total electron density, including its own charge distribution. [13] In exact theory, this unphysical self-repulsion is exactly cancelled by the exchange-correlation functional; however in reality, approximate exchange-correlation functionals do not fully cancel the electron self-repulsion. The SIE creates challenges for accurately simulating the properties of materials with partially filled, localised d or f orbitals such as transition metal oxides (TMOs) and rare-earth metal oxides (REOs). For these systems, the SIE causes systematic errors in predicted material properties, *e.g.*, underestimated insulator band gaps, inaccurate lattice parameters and inaccurate formation energies of point defects and electron polarons. [14–16] To overcome the limitations of DFT for strongly correlated materials, "*beyond-DFT*" methods can be applied that add corrective schemes to combat the SIE. Hybrid-DFT is one of the most popular schemes to mitigate the SIE *via* mixing a portion of the non-local exact exchange from Hartree-Fock theory with the local or semi-local DFT exchange-correlation energy:

$$E_{\mathrm{xc}}^{\mathrm{hybrid}} = \alpha E_{\mathrm{x}}^{\mathrm{HF}} + (1-\alpha) E_{\mathrm{x}}^{\mathrm{DFT}} + E_{\mathrm{c}}^{\mathrm{DFT}} \tag{2.15}$$

where E_x^{HF} is the exact exchange from Hartree-Fock, $E_{\mathrm{x}}^{\mathrm{DFT}}$ ($E_{\mathrm{c}}^{\mathrm{DFT}}$) are the exchange (correlation) contributions from a chosen semi-local functional and α is a mixing parameter (typically 20-25 %). Hybrid functionals generally predict more accurate band gaps, magnetic properties and defect energetics compared to LDA, GGA or meta-GGA DFT; however, the computational cost of evaluating exact exchange scales on the order of $O(N^4)$, which makes hybrid-DFT prohibitively expensive for large supercells or extended configurational sampling.

2.3.2 Hubbard Corrected DFT+U

Background and Formalism

Hubbard corrected density functional theory (DFT+U) is a popular alternative to hybrid-DFT for modelling strongly correlated metal oxides, involving a tuneable on-site Coulomb repulsion, *i.e.*, an energy penalty against electron delocalisation, that is applied selectively to localised orbitals in the system. [17] The physical intuition behind the DFT+U method is derived from the Hubbard model,

[18] which provides a simplified description of interacting electrons occupying lattice sites. In the case of a one-dimensional chain of atoms with each containing one electron that half fills one orbital, the absence of electron-electron interactions ($U = 0$) results in a metallic system with a single half-filled band (Figure 2.3). Upon applying an on-site Coulomb repulsion, the occupation of two electrons at the same site becomes increasingly unfavourable, causing the single band to split into two: a fully occupied lower Hubbard band (LHB) and an empty upper Hubbard band (UHB). The DFT+U method aims to treat the SIE in semi-local DFT in a similar way, by penalising the fractional occupancy of localised orbitals and thus driving the system towards integer occupations, whilst opening a gap between occupied and unoccupied localised states. Importantly, the DFT+U method incurs minimal added computational cost *vs.* semi-local DFT, and can be tuned to achieve the accuracy of higher levels of theory if parameterised correctly. [19]

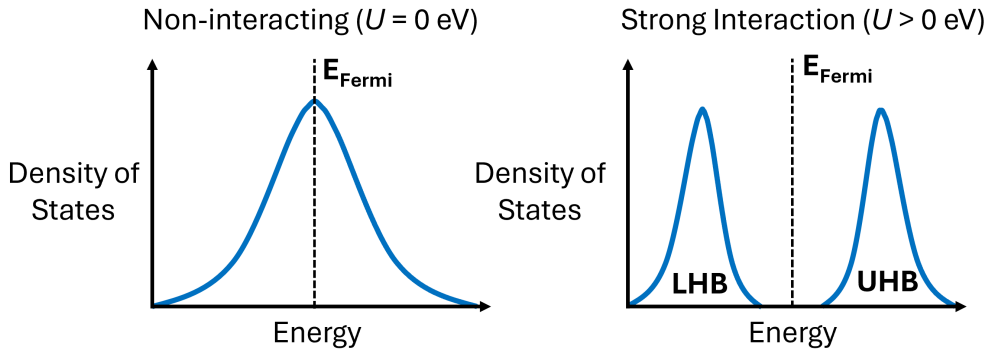


Figure 2.3: Schematic illustration of an on-site Coulomb repulsion in a half-filled 1-dimensional lattice splitting a single band into a lower Hubbard band (LHB) and upper Hubbard band (UHB), opening a gap at the Fermi level (E_{Fermi}).

To accurately parameterise DFT+U, one must account for both the magnitude and basis of the Hubbard correction, which are defined using the Hubbard U value and the Hubbard projector function (or Hubbard projector), respectively. These parameters are used to correct the DFT-predicted total energy (E_{DFT}) with a corrective Hubbard term that treats localised states only (E_U^0) and a double counting correction (E_U^{dc}) that prevents the double counting of localised states in both E_{DFT} and E_U^0:

$$E_{\mathrm{DFT}+U}[\rho(\mathbf{r}), \mathbf{n}_{I,m}^{\sigma}] = E_{\mathrm{DFT}}[\rho(\mathbf{r})] + E_U^0[\mathbf{n}_{I,m}^{\sigma}] - E_U^{\mathrm{dc}}[\mathbf{n}_{I,m}^{\sigma}] \tag{2.16}$$

where $\rho(\mathbf{r})$ is the electron density and $\mathbf{n}_{I,m}^{\sigma}$ is the occupation matrix, whose diagonal elements correspond to orbital occupation numbers for all atoms (I), orbital magnetic quantum numbers (m) and spin channels (σ). According to the rotationally invariant, spherically averaged implementation proposed by Dudarev *et al.*, [20] the corrective Hubbard term is calculated using the trace (Tr) of the occupation matrix and its square:

$$E_U^0[\mathbf{n}_{I,m}^{\sigma}] = \sum_{(\sigma,I)} U^I \left[\mathrm{Tr}(\mathbf{n}_{I,m}^{\sigma}) - \mathrm{Tr}(\mathbf{n}_{I,m}^{\sigma}\mathbf{n}_{I,m}^{\sigma})\right] \tag{2.17}$$

The occupation matrix is calculated by projecting all DFT-predicted Kohn-Sham states onto reference orbitals defined by the Hubbard projector, *i.e.*, calculating the overlap between Kohn-Sham

states and spatially localised orbitals, before an overlap-dependent assignment of the occupancy of each Kohn-Sham state to the localised orbitals. [21] After evaluating E_U^0, the corresponding Hubbard potential, *i.e.*, the added correction to the standard Kohn-Sham effective potential, is then obtained by taking the functional derivative of the corrective Hubbard energy with respect to the occupation matrix, yielding an orbital-dependent potential that acts only on the subspace defined by the chosen projector. [20] The Hubbard correction therefore acts as an occupancy-based bias potential that corrects the total energy using the Hubbard U value and the occupation matrix, which necessitates careful choice of both the Hubbard U value *and* the projector. Choosing an appropriate Hubbard projector is particularly important for accurate simulations of materials with strong covalent character, [21–24] and their specific representation prevents the transferability of Hubbard parameters across electronic structure codes that employ different types of Hubbard projector, *e.g.*, atomic orbitals, [21] Wannier functions, [25–27] projector augmented wave (PAW) projectors [28] and muffin-tin orbitals (MTOs). [22]

Modifying the Hubbard Projectors

In FHI-aims, the default Hubbard projector is defined as the atomic NAO basis function in the minimal basis set, which corresponds to the solution of the non-spin-polarised single atom Schrödinger equation. The use of localised atomic basis functions provides a reasonable initial guess for constructing the correlated subspace for the DFT+U correction, but is known to overestimate orbital occupation numbers leading to inaccurate predictions of oxidation states and energetic properties of complex oxides. [29–31] Alternative definitions of the Hubbard projector can provide more accurate predictions of orbital occupancies, such as maximally-localised Wannier functions; however, evaluating these projectors can introduce significant computational overhead. [32] To maximise computational efficiency, FHI-aims allows the Hubbard projector (Φ_{Im}) to be defined as a linear combination of NAO basis functions (χ^i_{Im}), specifically the atomic basis function in the mimimal basis set and auxiliary hydrogenic basis functions for the same atomic site I and orbital magnetic quantum number m:

$$\Phi_{Im} = \sum_i c_i \, \chi^i_{Im} \tag{2.18}$$

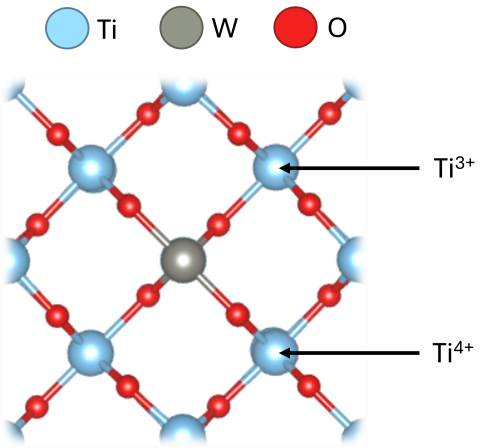
where c_i denotes the linear expansion coefficients for a maximum of four basis functions per projector. The use of basis sets with multiple basis functions for the same localised orbital can reportedly lead to erroneous ground state predictions due to the occupation of electronic states outside the correlated subspace; [21, 33] therefore, the DFT+U calculations in this thesis were restricted to the unmodified light basis set, which contains at most one auxiliary basis function per localised orbital listed in Table 2.1. With the available basis functions in Table 2.1, modified atomic-like Hubbard projectors were defined using the linear expansion coefficients c_1 and c_2, which correspond to the atomic and auxiliary basis functions, respectively. Before constructing modified projectors, the auxiliary basis functions are subject to a Gram-Schmidt orthogonalisation with respect to the corresponding atomic functions, which avoids double counting of the Hubbard correction from the overlap of basis functions with long tails of radial decay. [21, 34] Positive values of c_2 were avoided to prevent the mixing of the metal d or f auxiliary basis functions with O $2p$ states in an unphysical manner, as determined by Kick *et al.* from comparisons with hybrid-DFT. [21]

Table 2.1: Materials, crystal structures and localised orbitals to which a Hubbard correction is applied throughout Chapters 3-5. Materials used for training the first-principles machine learning approach in Chapter 4 are separated from the materials that are unseen from model training (denoted using *). The auxiliary basis functions in the light basis set, which are used for constructing modified atomic-like Hubbard projectors, are noted alongside their confinement parameters, which correspond to an effective core charge (Z_{val}, e) for the hydrogenic basis functions and the onset radius (Bohr) of the cutoff potential for the ionic basis functions. The light basis sets for Mo and W do not contain any auxiliary basis functions for 4d and 5d, respectively. For these systems, as well as Y_2O_3 and ZrO_2 with ionic auxiliary basis functions for 4d, only the first linear expansion coefficient c_1 in the linear combination affects the outcomes of the DFT+U calculations, as FHI-aims allows only hydrogenic basis functions in the linear combination. [33]

Material	Crystal Structure	Corrected Orbitals	Auxiliary Basis Function Type	Confinement Parameter
TiO_2	Tetragonal	Ti 3d	Hydrogenic	2.7
Cu_2O	Cubic	Cu 3d	Hydrogenic	5.0
Y_2O_3	Cubic	Y 4d	Ionic	4.0
ZrO_2	Orthorhombic	Zr 4d	Ionic	3.5
MoO_3	Orthorhombic	Mo 4d	N/A	N/A
CeO_2	Cubic	Ce 4f	Hydrogenic	7.6
WO_3	Monoclinic	W 5d	N/A	N/A
* $LiFePO_4$	Orthorhombic	Fe 3d	Hydrogenic	3.1
* $LiCoO_2$	Trigonal	Co 3d	Hydrogenic	5.4

Occupation Matrix Control

DFT+U is known to introduce challenges in the self-consistent determination of the ground state, due to the existence of metastable states in the potential energy surface with respect to orbital occupancies. The "*occupation matrix control*" (OMC) method is a popular approach to aid the identification of the ground state, by carefully controlling the orbital occupancies *via* constrained and/or self-consistent approaches. [21, 35, 36] In this work, constrained DFT+U calculations are performed using the OMC method to *fix* polarons at specific atoms by modifying the corresponding occupation matrices (which remain fixed for the entire calculation), *e.g.*, fixing the d_{z^2} orbital occupation number as 1 for a nearest neighbour Ti atom relative to a substitutional W dopant to simulate the formation of Ti^{3+} in W-doped TiO_2 (Figure 2.4). [21] Self-consistent defect calculations are also performed using the "*occupation matrix release*" (OMR) method to *initialise* polaron(s), then the DFT+U-predicted total energy (E) is pre-converged using OMC until $\Delta E \leq 0.001$ eV, the OMC constraint is then relaxed and the orbital occupancies are calculated self-consistently. [21]



Ti 3*d* Occupation Matrix

0.16098	0.00000	0.05529	0.00000	0.00000
0.00000	0.15932	0.00000	-0.06435	0.00000
0.05529	0.00000	**1.00000**	0.00000	0.00000
-0.00000	-0.06435	0.00000	0.15932	0.00000
0.00000	0.00000	0.00000	0.00000	0.09900

0.16098	0.00000	0.05529	0.00000	0.00000
0.00000	0.15932	0.00000	-0.06435	0.00000
0.05529	0.00000	**0.10456**	0.00000	0.00000
-0.00000	-0.06435	0.00000	0.15932	0.00000
0.00000	0.00000	0.00000	0.00000	0.09900

Figure 2.4: Simulating an electron polaron in W-doped TiO_2 at a nearest neighbour Ti atom, denoted Ti^{3+} by setting the $3d_{z^2}$ orbital occupation number to 1. The electron polaron can be *fixed* using the occupation matrix control (OMC) method or *initialised* using the occupation matrix release (OMR) method. The diagonal elements of the occupation matrix correspond to orbital occupancies for a given magnetic quantum number ($3d_m$) in the order (from top left to bottom right) $3d_{-2}$, $3d_{-1}$, $3d_0$, $3d_1$ and $3d_2$ corresponding to the $3d_{xy}$, $3d_{yz}$, $3d_{z^2}$, $3d_{xz}$ and $3d_{x^2-y^2}$ orbitals, respectively. [36] Off-diagonals elements in the occupation matrix reflect orbital hybridisation. In this work, the diagonal elements of the occupation matrix are used as a quantitative measure of local chemical bonding environments and to construct workflows for Hubbard parameter optimisation by assessing how the occupation matrix varies with the chosen simulation method (*i.e.*, DFT, DFT+U or hybrid-DFT, detailed in Chapter 4), Hubbard parameters and atomic properties of different materials.

2.4 Supervised Machine Learning

Supervised machine learning methods are increasingly applied in computational materials modelling to construct surrogate models of complex physical relationships and bypass the requirement for computationally expensive evaluations. Such methods are based on learning the mapping between input features $\mathbf{x}$ and target outputs $\mathbf{y}$ from labelled training data, *i.e.*, approximating the conditional probability distribution $P(\mathbf{y} \mid \mathbf{x})$, which describes the probability of observing the output $\mathbf{y}$ given the input $\mathbf{x}$. In regression tasks, $\mathbf{y}$ is continuous (*e.g.*, predicting the band gap of a material), whilst in classification tasks, $\mathbf{y}$ is discrete (*e.g.*, a calculation is stable or unstable). In both cases, the aim is to learn a mapping $f : \mathbf{x} \mapsto \mathbf{y}$ that generalises to unseen data, which can be combined with optimisation algorithms to efficiently search multi-dimensional feature spaces for potential solutions. In Chapter 4, regression, classification and optimisation algorithms are applied to parameterise Hubbard U values and projectors in a manner that generalises across different materials. The workflows constructed involve three complementary methods: symbolic regression, support vector machines and Bayesian optimisation.

Symbolic Regression

Symbolic regression (SR) is a powerful non-linear regression algorithm, which searches over a space of mathematical expressions rather than reparameterising constants of a predefined model form. In Chapter 4, SR is applied using the Sure Independence Screening and Sparsifying Operator (SISSO)

algorithm, [37, 38] which recursively combines primary features using a defined set of mathematical operators (*e.g.*, $+, -, \times, \div, \sin, \exp, \log$) to form secondary features. From a large pool of generated secondary features, sparse regression is then used to select a minimal number of secondary features and linear regression is used to optimise the cofficients of the final expression, which is a linear combination of secondary features. SR is particularly advantageous for our use case in DFT+U parameterisation as it is effective with small datasets and capturing non-linear mappings from $\mathbf{x} \mapsto \mathbf{y}$.

Support Vector Machines

Support vector machines (SVMs) are a popular supervised classification algorithm that find the optimal hyperplane separating data points of different classes. Given labelled training data $(\mathbf{x}_i, \mathbf{y}_i)$ with $\mathbf{y}_i \in \{1, 0\}$, an SVM maximises the margin between the hyperplane and the nearest training points (support vectors). In the context of DFT+U parameterisation in Chapter 4, SVMs are applied using the *Scikit-learn* Python library [39] to classify regions of the Hubbard parameter space that lead to numerically stable *vs.* unstable defect calculations, or physical *vs.* unphysical defect energies. By fitting SVMs to DFT+U data, the resulting decision boundaries provide explicit constraints for parameter optimisation, ensuring the subsequent DFT+U calculations are robust.

Bayesian Optimisation

Bayesian optimisation (BO) is a probabilistic global optimisation algorithm designed for expensive, black-box cost functions. [40] It operates by constructing a surrogate model (commonly a Gaussian Process) of an unknown function, alongside an acquisition function that balances exploration (sampling uncertain regions of the parameter space) and exploitation (refining regions close to the predicted optimum). At each iteration, the next sampling point is selected by maximising the acquisition function, after which the surrogate model is updated with the new evaluation. In the context of DFT+U parameterisation in Chapter 4, BO is useful in two complementary ways. Firstly, for active learning *via* iterative DFT+U calculations, it can minimise the total number of expensive electronic structure calculations required to optimise the Hubbard parameters to achieve the accuracy of higher levels of theory. Secondly, when screening cost functions derived from symbolic regression, BO does not rely on gradient information and is therefore less susceptible to becoming trapped in local minima in the complex cost function landscape.

2.5 Multiscale Modelling Beyond Atomistic Regimes

Even with advances in DFT and beyond-DFT methods, the direct simulation of realistic catalytic materials at experimentally relevant scales remains computationally intractable. For example, the formation of oxygen vacancies in catalyst support materials can result in oxygen spillover towards supported catalysts; however, simulating oxygen adsorption on catalysts at experimentally relevant surface coverages is itself computationally demanding. For these challenges, multiscale modelling approaches extend the reach of atomistic simulation methods by integrating statistical sampling and machine learning for the simulation of catalytic phenomena on length scales beyond atomistic regimes.

Grand Canonical Monte Carlo Sampling

In Chapter 5, the application of multiscale modelling is discussed for simulating the adsorption of catalyst poisons (S) and regenerators (O). Constructing more experimentally relevant predictive models for S and O adsorption on Ni(111) requires extensive sampling of the large configurational space of adsorption complexes, which is computationally infeasible with DFT alone. Statistical sampling algorithms, such as grand canonical Monte Carlo (GCMC), must therefore be considered as they are well suited for exploring the configurational space of adsorption complexes on a lattice model of the surface, where adsorbates occupy predefined adsorption sites. [41, 42] In GCMC, the ground state of the system is estimated by stochastically sampling a DFT-parameterised Hamiltonian through adsorbate perturbations such as adsorption, desorption or diffusion. [43] The GCMC approach allows the system to explore a wide range of chemically relevant surface configurations, producing extended models that are beyond the atomistic length scales afforded by DFT, whilst ensuring all accessible states contribute to the statistical ensemble when determining surface properties at thermodynamic equilibrium.

Machine Learned Interatomic Potentials

Lattice models simplify the sampling of the configurational space of adsorption complexes but neglect off-lattice effects, such as many-body lateral interactions and surface reconstruction, which can be non-negligible under experimental reaction conditions. To account for off-lattice effects, extended GCMC-predicted adlayers can be refined using classical interatomic potentials (IPs) to perform geometry optimisation and/or molecular dynamics simulations. [44–46] Classical simulations are a computationally efficient approach for modelling materials at the length scales unaffordable using DFT, but the accuracy of these simulations is dependent on that of the underlying IP. Modern machine learned interatomic potentials (MLIPs) offer a promising approach for balancing accuracy and computational efficiency by avoiding the predefined functional forms used in traditional IPs, enabling MLIPs to capture complex potential energy surfaces with greater flexibility. Recent advancements in neural network (*e.g.*, SchNet, [47] PaiNN, [48] M3GNet, [49] CHGNet [50] and MACE [51]) and Gaussian process-based (*e.g.*, GAP [52]) MLIPs have enabled more accurate modelling of chemical reactivity on transition metal surfaces. [53–55]

In Chapter 5, I apply the MACE [51] architecture to perform large-scale geometry relaxations of atomic poisons and regenerators on a Ni(111) catalyst surface, allowing the investigation of the validity of GCMC predictions and the viability of different catalyst regeneration mechanisms *via* entropic disorder. To ensure the validity of the MACE simulations, a pre-trained foundation model is fine-tuned on a large collection of newly generated meta-GGA DFT data, achieving root-mean-square errors of 14.4 meV per atom (14.2 meV per atom on the validation set) in total energies and 16.3 meV $Å^{-1}$ (17.2 meV $Å^{-1}$ on the validation set) in atomic forces. The close agreement between training and validation errors indicates good generalisation, allowing us to leverage the data efficiency of the MACE architecture compared to other MLIPs, and to effectively simulate off-lattice effects in extended catalyst surfaces with near *ab initio* accuracy. [56] MACE replaces the rigid analytical forms of classical IPs in Equation 2.1 with a highly-parameterised message-passing graph neural network, where atoms are encoded as nodes and local chemical environments are encoded through inter-node

edges. During model training, atomic features are updated using information from neighbouring atoms within a finite cutoff radius, enabling the model to learn short- and medium-range interactions that are essential to model DFT-predicted energies and forces across a broad range of materials. MACE learns the DFT-predicted total energy as an expansion of equivariant Atomic Cluster Expansion (ACE) basis functions, [57] which encode local many-body atomic correlations whilst preserving the required rotational, translational and permutational symmetries. These architectural features enable a robust and transferable representation of chemical environments, whilst often outperforming the accuracy of classical IPs by an order of magnitude across a broad range of tasks. [55] The application of the MACE MLIP is discussed further in Chapter 5 for the validation of GCMC-predicted adlayers, with comparisons between the accuracy of pre-trained *vs*. finetuned MACE foundation models.

References

(1) J. E. Jones, On the determination of molecular fields.—II. From the equation of state of a gas, *Proc. R. soc. Lond. Ser. A-Contain.* 1924, **106** 463–477.

(2) E. Schrödinger, Quantisierung als eigenwertproblem, *Ann. Phys.* 1926, **385** 437–490.

(3) M. Oppenheimer, Zur Quantentheorie der Molekeln [On the quantum theory of molecules], *Ann. Phys.* 1927, **389** 457–484.

(4) E. H. Lieb, Thomas-fermi and related theories of atoms and molecules, *Rev. Mod. Phys.* 1981, **53** 603–641.

(5) D. R. Hartree, The wave mechanics of an atom with a non-Coulomb central field. Part I. Theory and methods, *Math. Proc. Camb. Philos. Soc.* 1928, **24** 89–110.

(6) J. C. Slater, A simplification of the Hartree-Fock method, *Phys. Rev.* 1951, **81** 385.

(7) P. Hohenberg and W. Kohn, Inhomogeneous electron gas, *Phys. Rev.* 1964, **136** B864.

(8) W. Kohn and L. J. Sham, Self-consistent equations including exchange and correlation effects, *Phys. Rev.* 1965, **140** A1133.

(9) V. Blum, R. Gehrke, F. Hanke, P. Havu, V. Havu, X. Ren, K. Reuter and M. Scheffler, *Ab initio* molecular simulations with numeric atom-centered orbitals, *Comput. Phys. Commun.* 2009, **180** 2175–2196.

(10) A. Seidl, A. Görling, P. Vogl, J. A. Majewski and M. Levy, Generalized Kohn-Sham schemes and the band-gap problem, *Phys. Rev. B* 1996, **53** 3764–3774.

(11) R. R. Brew, I. A. Nelson, M. Binayeva, A. S. Nayak, W. J. Simmons, J. J. Gair and C. C. Wagen, Wiggle150: Benchmarking Density Functionals and Neural Network Potentials on Highly Strained Conformers, *J. Chem. Theory Comput.* 2025, **21** 3922–3929.

(12) W. Zawadzki, in *EME Optics*, ed. R. D. Guenther, Elsevier, Oxford, 2005, pp. 432–438.

(13) D. R. Lonsdale and L. Goerigk, One-electron self-interaction error and its relationship to geometry and higher orbital occupation, *J. Chem. Phys.* 2023, **158**.

(14) N. L. Nguyen, N. Colonna, A. Ferretti and N. Marzari, Koopmans-Compliant Spectral Functionals for Extended Systems, *Phys. Rev. X* 2018, **8** 021051.

(15) J. P. Perdew and A. Zunger, Self-interaction correction to density-functional approximations for many-electron systems, *Phys. Rev. B* 1981, **23** 5048–5079.

(16) M. Reticcioli, U. Diebold and C. Franchini, Modeling polarons in density functional theory: lessons learned from TiO_2, *J. Condens. Matter Phys.* 2022, **34** 204006.

(17) H. J. Kulik, Perspective: Treating electron over-delocalization with the DFT+U method, *J. Chem. Phys.* 2015, **142** 240901.

(18) J. Hubbard, Electron Correlations in Narrow Energy Bands, *Proc. R. Soc. Lond. A.* 1963, **276** 238–257.

(19) D. S. Lambert and D. D. O'Regan, Use of DFT + U + J with linear response parameters to predict non-magnetic oxide band gaps with hybrid-functional accuracy, *Phys. Rev. Res.* 2023, **5** 013160.

(20) S. L. Dudarev, G. A. Botton, S. Y. Savrasov, C. J. Humphreys and A. P. Sutton, Electron-energy-loss spectra and the structural stability of nickel oxide: An LSDA+U study, *Phys. Rev. B* 1998, **57** 1505–1509.

(21) M. Kick, K. Reuter and H. Oberhofer, Intricacies of DFT+U, Not Only in a Numeric Atom Centered Orbital Framework, *J. Chem. Theory Comput.* 2019, **15** 1705–1718.

(22) K. Park, M. Raman, A.-J. Olatunbosun and J. Pohlmann, Revisiting DFT+U calculations of TiO_2 and the effect of the local-projection size, *AIP Adv.* 2024, **14** 065114.

(23) Z. Wang, C. Brock, A. Matt and K. H. Bevan, Implications of the DFT + U method on polaron properties in energy materials, *Phys. Rev. B* 2017, **96** 125150.

(24) I. Timrov, N. Marzari and M. Cococcioni, Self-consistent Hubbard parameters from density-functional perturbation theory in the ultrasoft and projector-augmented wave formulations, *Phys. Rev. B* 2021, **103** 045141.

(25) D. D. O'Regan, N. D. M. Hine, M. C. Payne and A. A. Mostofi, Projector self-consistent DFT + U using nonorthogonal generalized Wannier functions, *Phys. Rev. B* 2010, **82** 081102.

(26) Y.-Y. Ting and P. M. Kowalski, Refined DFT+U method for computation of layered oxide cathode materials, *Electrochim. Acta* 2023, **443** 141912.

(27) G. L. Murphy, Z. Zhang, R. Tesch, P. M. Kowalski, M. Avdeev, E. Y. Kuo, D. J. Gregg, P. Kegler, E. V. Alekseev and B. J. Kennedy, Tilting and Distortion in Rutile-Related Mixed Metal Ternary Uranium Oxides: A Structural, Spectroscopic, and Theoretical Investigation, *Inorg. Chem.* 2021, **60** 2246–2260.

(28) I. Timrov, F. Aquilante, L. Binci, M. Cococcioni and N. Marzari, Pulay forces in density-functional theory with extended Hubbard functionals: From nonorthogonalized to orthogonalized manifolds, *Phys. Rev. B* 2020, **102** 235159.

(29) G. L. Murphy, Z. Zhang, R. Tesch, P. M. Kowalski, M. Avdeev, E. Y. Kuo, D. J. Gregg, P. Kegler, E. V. Alekseev and B. J. Kennedy, Tilting and distortion in rutile-related mixed metal ternary uranium oxides: a structural, spectroscopic, and theoretical investigation, *Inorg. Chem.* 2021, **60** 2246–2260.

(30) K. O. Kvashnina, P. M. Kowalski, S. M. Butorin, G. Leinders, J. Pakarinen, R. Bès, H. Li and M. Verwerft, Trends in the valence band electronic structures of mixed uranium oxides, *Chem. Commun.* 2018, **54** 9757–9760.

(31) P. M. Kowalski, Z. He and O. Cheong, Electrode and electrolyte materials from atomistic simulations: properties of Li_xFePO_4 electrode and zircon-based ionic conductors, *Front. Energy Res.* 2021, **9** 653542.

(32) Y.-Y. Ting and P. M. Kowalski, Refined DFT+U method for computation of layered oxide cathode materials, *Electrochim. Acta.* 2023, **443** 141912.

(33) K. Jakob and H. Oberhofer, "Self-Consistency in the Hubbard-Corrected DFT+U Method", Master's thesis, Faculty of Chemistry, Technical University of Munich, 2021.

(34) W. B. Begna, G. S. Gurmesa and C. A. Geffe, Ortho-atomic projector assisted DFT+U study of room temperature Ferro- and antiferromagnetic Mn-doped TiO_2 diluted magnetic semiconductor, *Mater. Res. Express* 2022, **9** 076102.

(35) B. Dorado, B. Amadon, M. Freyss and M. Bertolus, DFT + U calculations of the ground state and metastable states of uranium dioxide, *Phys. Rev. B* 2009, **79** 235125.

(36) J. P. Allen and G. W. Watson, Occupation matrix control of d- and f-electron localisations using DFT+U, *Phys. Chem. Chem. Phys.* 2014, **16** 21016–21031.

(37) R. Ouyang, S. Curtarolo, E. Ahmetcik, M. Scheffler and L. M. Ghiringhelli, SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates, *Phys. Rev. Mater.* 2018, **2** 083802.

(38) T. A. R. Purcell, M. Scheffler, C. Carbogno and L. M. Ghiringhelli, SISSO++: A C++ Implementation of the Sure-Independence Screening and Sparsifying Operator Approach, *JOSS* 2022, **7** 3960.

(39) F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, Scikit-learn: Machine Learning in Python, *JMLR* 2011, **12** 2825–2830.

(40) J. B. Mockus and L. J. Mockus, Bayesian approach to global optimization and application to multiobjective and constrained problems, *J. Optim. Theory Appl.* 1991, **70** 157–172.

(41) C. Schwennicke and H. Pfnür, O/Ni(111): Lateral interactions and binding-energy difference between fcc and hcp sites, *Phys. Rev. B* 1997, **56** 10558–10566.

(42) C. Lazo and F. J. Keil, Phase diagram of oxygen adsorbed on Ni(111) and thermodynamic properties from first-principles, *Phys. Rev. B* 2009, **79** 245418.

(43) S. S. Akimenko, G. D. Anisimova, A. I. Fadeeva, V. F. Fefelov, V. A. Gorbunov, T. R. Kayumova, A. V. Myshlyavtsev, M. D. Myshlyavtseva and P. V. Stishenko, SuSMoST: Surf. Sci. Modeling and Simulation Toolkit, *J. Comput. Chem.* 2020, **41** 2084–2097.

(44) L. Gai, Y. K. Shin, M. Raju, A. C. T. van Duin and S. Raman, Atomistic Adsorption of Oxygen and Hydrogen on Platinum Catalysts by Hybrid Grand Canonical Monte Carlo/Reactive Molecular Dynamics, *J. Phys. Chem. C.* 2016, **120** 9780–9793.

(45) T. Demeyere, T. Ellaby, M. Sarwar, D. Thompsett and C.-K. Skylaris, Bridging Oxide Thermodynamics and Site-Blocking: A Computational Study of ORR Activity on Platinum Nanoparticles, *ACS Catal.* 2025, **15** 5674–5682.

(46) T. Demeyere, H. U. Islam, T. Ellaby, M. Sarwar, D. Thompsett and C.-K. Skylaris, Multi-scale modeling and experimental investigation of oxidation behavior in platinum nanoparticles, *Phys. Chem. Chem. Phys.* 2025.

(47) K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller and A. Tkatchenko, Quantum-chemical insights from deep tensor neural networks, *Nat. Commun.* 2017, **8** 13890.

(48) K. Schütt, O. Unke and M. Gastegger, Equivariant message passing for the prediction of tensorial properties and molecular spectra, *Int. Conf. Mach. Learn.* 2021, 9377–9388.

(49) C. Chen and S. P. Ong, A universal graph deep learning interatomic potential for the periodic table, *Nat. Comput. Sci.* 2022, **2** 718–728.

(50) B. Deng, P. Zhong, K. Jun, J. Riebesell, K. Han, C. J. Bartel and G. Ceder, CHGNet: Pretrained universal neural network potential for charge-informed atomistic modeling, *Nat. Comput. Sci.* 2023, **3** 192–202.

(51) I. Batatia, D. P. Kovacs, G. Simm, C. Ortner and G. Csányi, MACE: Higher order equivariant message passing neural networks for fast and accurate force fields, *Adv. Neural Inf. Process.* 2022, **35** 11423–11436.

(52) A. P. Bartók, M. C. Payne, R. Kondor and G. Csányi, Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons, *Phys. Rev. Lett.* 2010, **104** 136403.

(53) W. G. Stark, J. Westermayr, O. A. Douglas-Gallardo, J. Gardner, S. Habershon and R. J. Maurer, Machine Learning Interatomic Potentials for Reactive Hydrogen Dynamics at Metal Surfaces Based on Iterative Refinement of Reaction Probabilities, *J. Phys. Chem. C.* 2023, **127** 24168–24182.

(54) H. Jung, L. Sauerland, S. Stocker et al., Machine-learning driven global optimization of surface adsorbate geometries, *npj Comput. Mater.* 2023, **9** 114.

(55) I. Batatia, P. Benner, Y. Chiang, A. M. Elena, D. P. Kovács, J. Riebesell, X. R. Advincula, M. Asta, M. Avaylon, W. J. Baldwin et al., A foundation model for atomistic materials chemistry, *arXiv preprint: 2401.00096* 2023.

(56) D. P. Kovács, I. Batatia, E. S. Arany and G. Csányi, Evaluation of the MACE force field architecture: From medicinal chemistry to materials science, *J. Chem. Phys.* 2023, **159** 044118.

(57) R. Drautz, Atomic cluster expansion for accurate and transferable interatomic potentials, *Phys. Rev. B* 2019, **99** 014104.

Chapter 3

Polymorph-Induced Reducibility and Electron Trapping Energetics of Nb and W Dopants in TiO_2

This chapter is based on the published work *Polymorph-Induced Reducibility and Electron Trapping Energetics of Nb and W Dopants in TiO_2* in *The Journal of Physical Chemistry C*, which is co-authored by Dr Andrew Logsdail (Cardiff University, CU) and Dr Andrea Folli (CU). [1]

The work introduces the challenge of accurately simulating defects and polarons in TiO_2 with sufficient accuracy to rationalise advanced material characterisation using electron paramagnetic resonance (EPR) spectroscopy. In this work, I performed the electronic structure calculations and Dr Andrea Folli performed all the experimental work, including materials synthesis and EPR spectroscopy. The input and output files of all electronic structure calculations have been uploaded as a dataset to the NOMAD repository at the DOI: 10.17172/NOMAD/2024.09.04-1.

3.1 Introduction

Transparent conducting oxides (TCOs) underpin modern consumer electronics, photovoltaics and light emitting devices (LED and OLED). For example, anatase Nb-doped TiO_2 (NTO) has emerged [2–11] as a more sustainable alternative to the widely used indium tin oxide (ITO), [12–15] capable of a resistivity of $2 \times 10^{-4}\,\Omega\,\text{cm}$ to $3 \times 10^{-4}\,\Omega\,\text{cm}$ and 97 % internal transmittance under visible light at room temperature for a 40 nm-thick film of anatase NTO with 3 $\%_{at}$ of Nb. [2] In contrast, rutile NTO is more resistive, [3] which allows for tailored applications based on the choice of TiO_2 polymorph. Similarly, anatase W-doped TiO_2 (WTO) exhibits n-type metallic behaviour [16, 17] with a resistivity of $1.5 \times 10^{-2}\,\Omega\,\text{cm}$ at room temperature for films with a doping concentration of 6.3 $\%_{at}$.[16] WTO has potential for TCO applications in electron transport layers (ETLs) in perovskite solar cells, with a demonstrated 28× greater efficiency compared to undoped TiO_2. [18] NTO and WTO have also attracted attention as promising photocatalysts with a 65 % increase in photocurrent demonstrated for NTO nanorod photoelectrodes doped with 0.25 $\%_{at}$ of Nb when compared to undoped TiO_2 [19]. NTO nanostructures, ranging from rutile nanorods to anatase nanosheets, also show improved adsorption of molecular O_2 and the formation of superoxide radicals $O_2^{-\cdot}$ under irradiation, when compared to pristine TiO_2 with the same morphology [20]. Furthermore, both NTO and WTO nanostructures

show enhanced dye photodegradation, [21, 22] photooxidation of nitrogen oxides (NO_x) to nitrates, [23] and ozone gas sensing, [21] when compared to undoped TiO_2.

To understand and further optimise the performance of NTO and WTO, it is necessary to understand the behaviour of electrons and holes that contribute to material conductivity and chemistry. This includes the mobility of charge carriers within the bands (conduction and valence, respectively), their recombination, their trapping (forming electron polarons) and the charge carrier transfer mechanisms that drive redox processes when these materials are used as semiconductor photocatalysts. Given the electronic spin associated with electrons and holes, electron paramagnetic resonance (EPR) spectroscopy is a powerful tool for the precise identification and characterisation of the dynamics, lifetimes and spatial distribution of excitons and their trapping states within NTO and WTO, including the paramagnetic species following insertion of Nb and W dopants in the TiO_2 lattice. For example, electron trapping states are mostly associated with Nb^{4+} and W^{5+} species, which are both paramagnetic d^1 species that can be detected and interrogated by EPR spectroscopy. The formation of Nb^{4+} and W^{5+} *vs.* Nb^{5+} and W^{6+} is particularly important, as it directly influences conductivity and photocatalytic efficiency.

The literature is conflicted with respect to the lack of Nb^{4+} and W^{5+} EPR signals when substitutional Nb and W are introduced in anatase TiO_2, in contrast to the presence of Nb^{4+} and W^{5+} EPR signals in doped rutile TiO_2. There is currently no clear explanation of this observation and so far first-principles atomistic modelling methods like density functional theory (DFT) have not completely clarified these experimental observations. For example, geometry optimisation with semi-local DFT followed by a single point calculation using a screened exchange hybrid functional (sX) predicts shallow donor states that are largely delocalised over Ti sites in anatase NTO, [6] supporting resonant photoemission experiments which confirm the absence of mid-gap states for anatase NTO. [3, 24] However, the same DFT calculations also predict a deep localised state in rutile NTO that is 0.9 eV below the conduction band edge, involving Ti $3d_{xy}$ orbitals with a small contribution from Nb 4*d* orbitals, thus contradicting the EPR observations. [6] Hubbard corrected density functional theory (DFT+*U*) calculations predict Ti 3*d* mid-gap states in rutile NTO, [25] but reports vary with some predicting shallow donor states in rutile NTO [26] and deep Ti 3*d* states in both anatase NTO [26, 27] and anatase WTO [27] *i.e.*, contradicting results. The application of these computational results based on DFT+*U* using planewave basis sets, fails to provide an unambiguous interpretation/prediction of the experimental observations. However, hybrid-DFT in a linear augmented plane wave basis has been demonstrated to successfully predict W 5*d* mid-gap states in rutile WTO, which gives promise for the application of alternative basis representations, particularly those based on all electron atom-centred basis functions. [28]

In this chapter, the challenges associated with accurately modelling polarons in anatase and rutile NTO and WTO are investigated using DFT+*U* in an all electron numerical atom-centred orbital (NAO) framework, [29] including the effects of the Ti 3*d* Hubbard projector and the use of the "*occupation matrix control*" method [30] for the identification of polaronic ground states. The DFT+*U* simulations are compared with experimental EPR spectra for powder anatase and rutile NTO and WTO, including the magnetic tensors characterising Nb^{4+} and W^{5+} polarons in doped rutile, providing accurate data to benchmark the validity of the DFT+*U* approach. The combination of theory and experiment improves our fundamental understanding of the nature and formation of reduced species in semiconductor metal

oxides, whilst providing a computational platform for simulating related systems in Chapters 4 and 5.

3.2 Methodology

3.2.1 Electronic Structure Calculations

DFT

All electronic structure calculations were performed using the Fritz-Haber Institute *ab initio* materials simulation (FHI-aims) software, [31] which uses an all electron numerical atom-centred orbital (NAO) basis set, interfaced with the Python based Atomic Simulation Environment (ASE). [32] The standard light basis set (2020) was used, with equivalent accuracy to the TZVP Gaussian-type orbital basis set, [33] as decided after benchmarking the TiO_2 formation energy, which was calculated using the energies of bulk Ti (in the hexagonal close packed, HCP, crystal structure) and an isolated O_2 molecule:

$$\Delta E_{\text{Form}} = E_{\text{TiO}_2} - E_{\text{Ti}} - E_{\text{O}_2} \tag{3.1}$$

The light basis set was chosen based on the negligible difference in the DFT-predicted ΔE_{Form}, as shown in Figure A.1, whilst dramatically reducing the computational cost. Relativistic effects were accounted for using the zeroth order regular approximation (ZORA) [31] as a scalar correction, whilst the system charge and spin was set to zero. Periodic boundary conditions were applied using converged **k**-point spacing for the optimised anatase and rutile unit cell respectively. The mBEEF meta-GGA exchange-correlation density functional was used, [34, 35] as defined in Libxc, [36] providing the best balance of accuracy and cost compared to other local, semi-local and hybrid functionals, which was defined using the DFT-predicted ΔE_{Form}, band gap (E_{bg}), unit cell equilibrium volume (V_0) and CPU time per SCF cycle for bulk geometry optimisation (Table A.1 and Figure A.2). Self-consistent field (SCF) optimisation of the electronic structure was achieved using a convergence criteria of 1×10^{-6} eV for the change in total energy, 1×10^{-4} eV for the change in the sum of eigenvalues and 1×10^{-6} e a_0^{-3} for the change in charge density. Geometry optimisation used the quasi-Newton BFGS algorithm [37–40] with a force convergence criteria of 0.01 eV/Å. Point defect calculations were performed using a $3 \times 3 \times 3$ TiO_2 supercell containing 324 and 162 atoms for anatase and rutile, respectively. The supercell size avoids spurious long-range defect-defect interactions between periodic images whilst simulating at low defect concentrations of 0.308% (anatase) and 0.617% (rutile) following substitution of a Ti atom with a Nb or W atom. Defect energies were calculated as:

$$\Delta E_{\text{Defect}} = E_{\text{Defective Bulk TiO}_2} + \mu_{\text{Ti}} - E_{\text{Stoichiometric Bulk TiO}_2} - \mu_{\text{Dopant}} \tag{3.2}$$

where the chemical potentials μ were calculated using the energy of bulk Ti (hexagonal close packed structure), Nb (body-centred cubic structure) and W (body-centred cubic structure).

DFT+*U*

All DFT+U calculations were performed using the on-site definition of the occupation matrix and the Fully Localised Limit (FLL) double counting correction. [29] A Hubbard correction was applied to treat the Coulomb self-interaction of Ti 3d orbital electrons only. Using the default atomic Ti

3*d* Hubbard projector presented challenges in identifying the ground state electronic structures of anatase and rutile NTO and WTO due to numerical instability; therefore, constrained and self-consistent DFT+*U* calculations were performed with an atomic and modified Ti 3*d* Hubbard projector, respectively. Constrained DFT+*U* calculations were performed using the default atomic Ti 3*d* Hubbard projector and a Ti 3*d* Hubbard *U* value of 3 eV in anatase TiO_2 and 4 eV in rutile TiO_2. These Hubbard *U* values were chosen to minimise the average error in the DFT+*U*-predicted band gap (E_{bg}) and unit cell equilibrium volume (V_0), defined as:

$$\text{Average Error} = \left\| \frac{100 \times (E_{bg}^{\text{DFT+}U} - E_{bg}^{\text{Exp}})}{E_{bg}^{\text{Exp}}}, \frac{100 \times (V_0^{\text{DFT+}U} - V_0^{\text{Exp}})}{V_0^{\text{Exp}}} \right\| \tag{3.3}$$

where V_0 is calculated by fitting to the Birch-Murnaghan equation of state using ASE, [41] and the experimental references are taken from the literature versus experimental references (Figure 3.1). [42, 43]

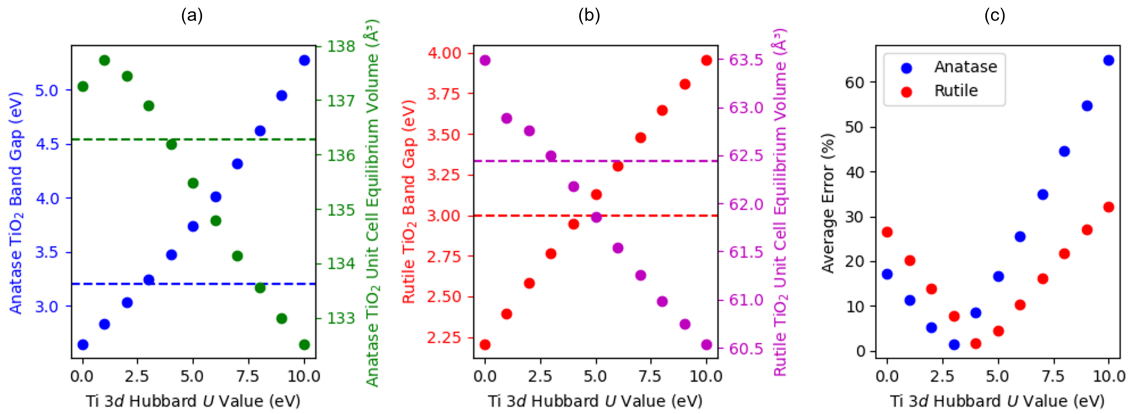


Figure 3.1: Benchmarking the Ti 3*d* Hubbard *U* value (using the default atomic Ti 3*d* Hubbard projector function) by comparing the DFT+*U*-predicted band gap and unit cell equilibrium volume of bulk (a) anatase and (b) rutile TiO_2 with experimental references [42, 43] (dashed lines). (c) shows the average error in (a) and (b) at each *U* value.

DFT+*U* calculations were performed using the "*occupation matrix control*" (OMC) method [30] to *fix* polaron(s) at specific atom(s) by modifying the corresponding atomic orbital occupation matrix (as outlined in Section 2.3.2). In these calculations, the inclusion of a Hubbard correction for Nb 4*d* or W 5*d* orbital electrons was found to result in geometric instability due to forced overlocalisation of polarons in the system; therefore a Hubbard correction was not applied to these orbitals. Self-consistent DFT+*U* calculations were performed with a refined atomic-like Ti 3*d* Hubbard projector. Here, the "*occupation matrix release*" (OMR) method was used to *initialise* the polaron(s) at specific atom(s) before self-consistent determination of the system occupation matrices (as outlined in in Section 2.3.2). [29] A refined atomic-like Ti 3*d* Hubbard projector was defined as a linear combination of the atomic Ti 3*d* and hydrogenic auxiliary basis function in the light basis set, where the auxiliary function is subject to a Gram–Schmidt orthogonalisation with respect to the atomic function, with the corresponding linear combination expansion coefficients c_1= 0.828 and c_2= -0.561. The values of c_1 and c_2 were chosen based on the work of Jakob and Oberhofer who computed a Ti 3*d* Hubbard

projector for bulk rutile TiO_2 in FHI-aims from first-principles. [44] These coefficients enabled successful convergence to the ground state using a Ti 3*d* Hubbard *U* value of 3 eV for both anatase and rutile NTO and WTO.

3.2.2 Materials Synthesis and Characterisation

The following experimental work was carried out by Dr Andrea Folli.

Materials Synthesis

The NTO and WTO materials studied in this Chapter were synthesised *via* a sol-gel route. 10 mL of titanium isopropoxide (≥97 %, Sigma-Aldrich) was dissolved in 10 mL of anhydrous ethanol. After thorough mixing, 5 mL of deionized water (18 MΩ cm) was slowly added to the solution. The resulting white precipitate redissolved upon further stirring. In the next step, 20 mL of a pH 10 ammonia/ammonium chloride buffer (5% ammonia, Sigma-Aldrich) was added to the solution. Finally, the desired amount of ammonium tungstate (BDH Chemicals) or ammonium niobate (V) oxalate hydrate (BDH Chemicals) to achieve a nominal 0.1 or 1.0 atomic % was dissolved in 10 mL of warm deionized water and subsequently added to the solution. After thorough stirring for at least 4 h, the solution was filtered, washed several times with deionized water, and then dried at 60 °C for 4 h. The dry powders were ground in an agate mortar and transferred into a crucible for calcination. The samples were calcined at 600 °C for 4 h, and then ground again afterwards.

Powder X-ray Diffraction

To confirm mineralogy and crystallinity, X-ray diffraction (XRD) patterns were obtained using a Bruker D8 Advance diffractometer equipped to deliver CuKα_1 X-ray radiation (1.54 Å) at room temperature. Refinement of the powder XRD patterns was carried out using the Profex suite for X-ray diffraction. [45]

EPR Spectroscopy

X-band continuous wave (CW) EPR spectra were recorded on a Bruker Elexsys E500 spectrometer equipped with an Oxford Instruments liquid-helium cryostat and a Bruker ER4122 SHQE-W1 super-high Q resonator operating at 50 K. Before each measurement, the samples were evacuated for at least 12 h at 393 K and under a dynamic vacuum at *ca.* 1×10^{-4} bar. Experimental spectra were simulated using the EasySpin toolbox [46] for Mathworks Matlab.

3.3 Results and Discussion

3.3.1 Experimentally Detected Polarons in NTO and WTO

The refined X-ray powder diffraction patterns (Appendix A) show that 0.1 $\%_{at.}$ of Nb or W in TiO_2, synthesised *via* a sol-gel route and calcined at 600 K, allows for the formation of mixed anatase and rutile polymorphs. This is in line with what happens for undoped TiO_2 exhibiting a rutilisation

temperature higher than 500 K (calcination of sol-gel TiO_2 precursors with 0.1 $\%_{at.}$ of Nb or W at $T < 500$ K generates anatase only polymorphs [10, 47, 48]). Refinement of the XRD patterns (Appendix A) revealed 92 % anatase and 8 % rutile for NTO whilst 72 % anatase and 28 % rutile for WTO. These samples are henceforth referred to as NTO-AR and WTO-AR. On the contrary, 1.0 $\%_{at.}$ of the same dopants in conjunction with the samples calcined at 600 K allows for the formation of anatase only NTO and WTO (Appendix A). These samples are henceforth referred to as NTO-A and WTO-A.

Electron and hole trapping are normally single electron transfer (SET) events occurring within the material and so they can be followed experimentally by detecting the formation, or disappearance, of paramagnetic species using EPR spectroscopy. Nb^{5+} ($[Kr]4d^0$) and W^{6+} ($[Xe]5d^0$) replace Ti^{4+} ($[Ar]3d^0$) in the TiO_2 lattice aliovalently and isomorphically, as Nb^{5+}, W^{6+} and Ti^{4+} have almost identical octahedral co-ordination environments with ionic radii of 78 pm, 74 pm, 74.5 pm, respectively. The dopant incorporation reactions for extrinsic defects using the Kröger-Vink notation can be written as follow:

$$Nb_2O_5 \xrightarrow{2TiO_2} 2\,Nb_{Ti}^{\bullet} + 4\,O_O^x + 2\,e' + \frac{1}{2}O_2 \tag{3.4}$$

$$2\,Nb_2O_5 \xrightarrow{5TiO_2} 4\,Nb_{Ti}^{\bullet} + V_{Ti}^{''''} + 10\,O_O^x \tag{3.5}$$

$$WO_3 \xrightarrow{TiO_2} W_{Ti}^{\bullet\bullet} + 2\,O_O^x + 2\,e' + \frac{1}{2}O_2 \tag{3.6}$$

$$2\,WO_3 \xrightarrow{3TiO_2} 2\,W_{Ti}^{\bullet\bullet} + V_{Ti}^{''''} + 6\,O_O^x \tag{3.7}$$

Valence-induced electron formation, as highlighted by Equations 3.4 and 3.6, increases the n-type character of NTO and WTO compared to undoped TiO_2. All EPR attempts at identifying *substitutional* and *isolated* Nb^{4+} and W^{5+} in NTO-A and WTO-A failed, which corroborates with previous experiments by De Trizio *et al.*, which could not detect Nb^{4+} in Nb-doped colloidal anatase nanocrystals even at at liquid helium temperature. [7] Giamello *et al.* [9] and Folli *et al.* [10] independently showed that in the case of Nb doping in anatase only TiO_2, Ti^{3+} is mostly formed as a result of valence induction when Nb^{5+} aliovalently replaces Ti^{4+} in the anatase lattice:

$$Ti_{Ti}^x + e' \longrightarrow Ti_{Ti}' \tag{3.8}$$

The resulting Ti^{3+} exhibits an anisotropic EPR spectrum characterised by a $\boldsymbol{g}$ tensor with axial symmetry and principal values equal to $g_\perp = 1.988$ and $g_{//} = 1.957$, [9, 10] which is consistent with a highly delocalised bulk species that is responsible for causing an increased conductivity of the doped anatase TiO_2. [9] This Ti^{3+} is structurally and magnetically different from surface-localised Ti^{3+} that forms following chemical or chemo/thermal reduction of undoped TiO_2. [9] The amount of delocalised bulk Ti^{3+} can also be augmented by photo-injection of extra conduction electrons [9, 10] followed by trapping:

$$\gamma \xrightarrow{TiO_2} h_{vb}^+ + e_{cb}^- \tag{3.9}$$

The situation appears completely different in the case of NTO-AR and WTO-AR, as demonstrated

by the respective X band continuous wave EPR spectra reported in Figure 3.2(a) (NTO-AR) and Figure 3.2(b) (WTO-AR). The clear axial signal in the spectrum of NTO-AR in Figure 3.2(a) can be attributed to bulk Ti^{3+} as previously described for the case of solely anatase polymorph. A very broad and low intensity signal is also visible at 50 K as shown by the arrows in Figure 3.2(a). We propose that this broad signal is associated with Nb^{4+} in rutile, as discussed further in Section A.3. Moving from single crystal to powder samples, Kiwi *et al.* [49] showed that in a mixed anantase and rutile powder sample a broad signal could be found at 4.2 K matching the $\boldsymbol{g}$ tensor reported by Zimmermann. [50] The broad signal in Figure 3.2(a) matches the signal reported by Kiwi *et al.*, [49] although it is much broader due to the much higher temperature of our measurement, *i.e.*, 50 K (according to Zimmermann [50] the signal completely vanishes above 77 K).

$$Nb_{Ti}^{\cdot} + e^{'} \longrightarrow Nb_{Ti}^{x} \tag{3.10}$$

In the case of WTO-AR, the situation is very similar to that described above for NTO-AR. Figure 3.2(b) shows a clear anisotropic EPR spectrum characterised by a rhombic $\boldsymbol{g}$ tensor with principal values reported in Table A.2. Small intensity doublets ($m_I = \pm 1/2$ lines) are visible at each side of the three principal resonances due to hyperfine interaction of the $5d^1$ unpaired electron to the ^{183}W nucleus ($I(^{183}W) = 1/2$, 14.3 % natural abundance). The other naturally occurring isotopes of W are ^{180}W, ^{182}W, ^{184}W and ^{186}W, all with nuclear spin quantum number $I = 0$, and these account for the three principal $m_I = 0$ resonance lines. W^{5+} in TiO_2 is not affected by the same fast relaxation issues as Nb^{4+} and therefore well defined EPR spectra can be easily obtained at 50 K as visible in Figure 3.2(b). The values of the magnetic tensors are in good agreement with Chang [51] for W^{5+} centres in rutile single crystals. The angular dependency has been here simulated and reported in Figures A.7(d) and A.7(e) at X and Q band, respectively.

This combined evidence on NTO-AR and WTO-AR indicates that the reduction of Nb^{5+} and W^{6+} dopants in the TiO_2 host lattice occurs only in the rutile polymorph.

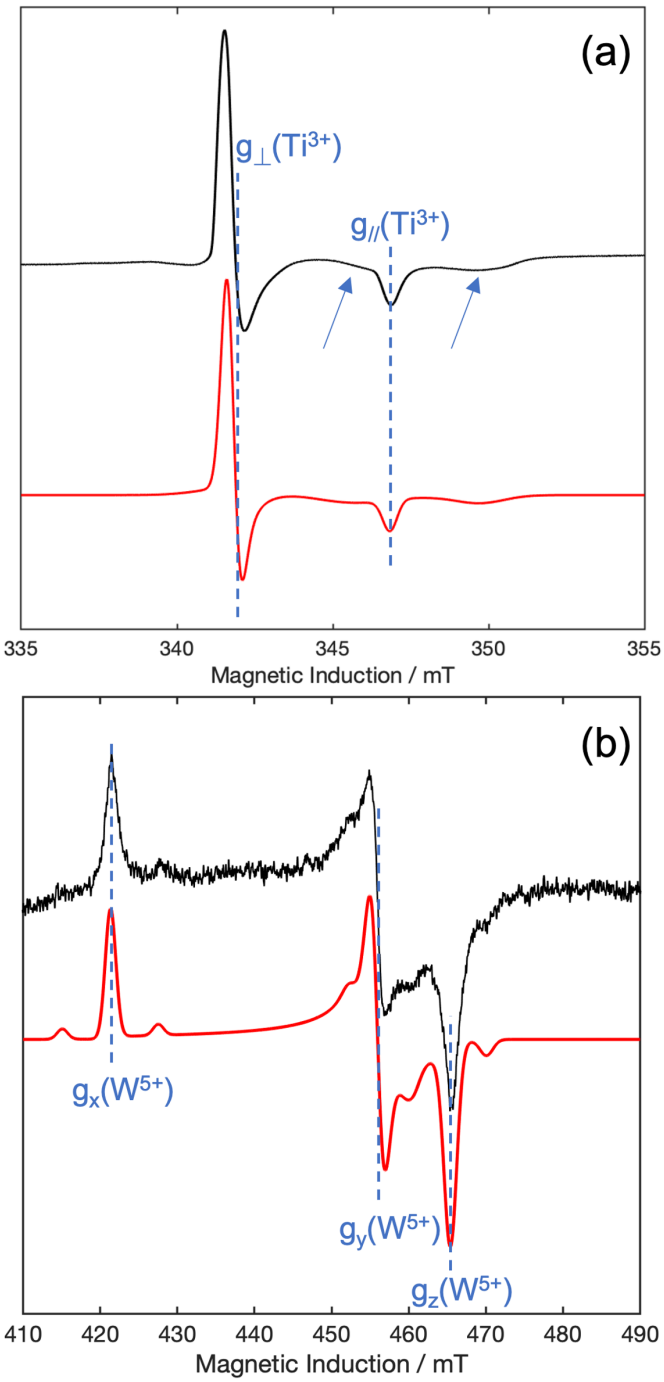


Figure 3.2: X band continuous wave EPR spectra at $50\,\mathrm{K}$ for (a) Nb-doped and (b) W-doped mixed polymorph TiO_2 nanoparticles with a doping concentration of $0.1\,\%_{\mathrm{at.}}$. *This EPR data was collected by Dr Andrea Folli.*

3.3.2 DFT+*U* Simulated Polarons in NTO and WTO

Self-Consistent DFT+*U* with a Refined Hubbard Projector

DFT+*U* calculations in a NAO framework were performed to rationalise the magnetic resonance observations in Section 3.3.1; however, self-consistent calculations resulted in significant numerical

instability with the default atomic Ti 3*d* Hubbard projector. The reasons why this occurred are discussed in detail in Chapter 4. Constrained DFT+*U* calculations with the atomic Ti 3*d* Hubbard projector could restore numerical stability to the simulations, but could not rationalise the experimental observations in Section 3.3.1. The use of a refined Ti 3*d* Hubbard projector, defined by a linear combination of NAO basis functions as outlined in Section 3.2.1, was then tested for anatase and rutile NTO and WTO. The total density of states reported in Figure 3.3(a) and Figure 3.3(c) shows defect states pinned to the bottom of the TiO_2 conduction band for both anatase NTO and WTO, respectively, in perfect agreement with the EPR observations.

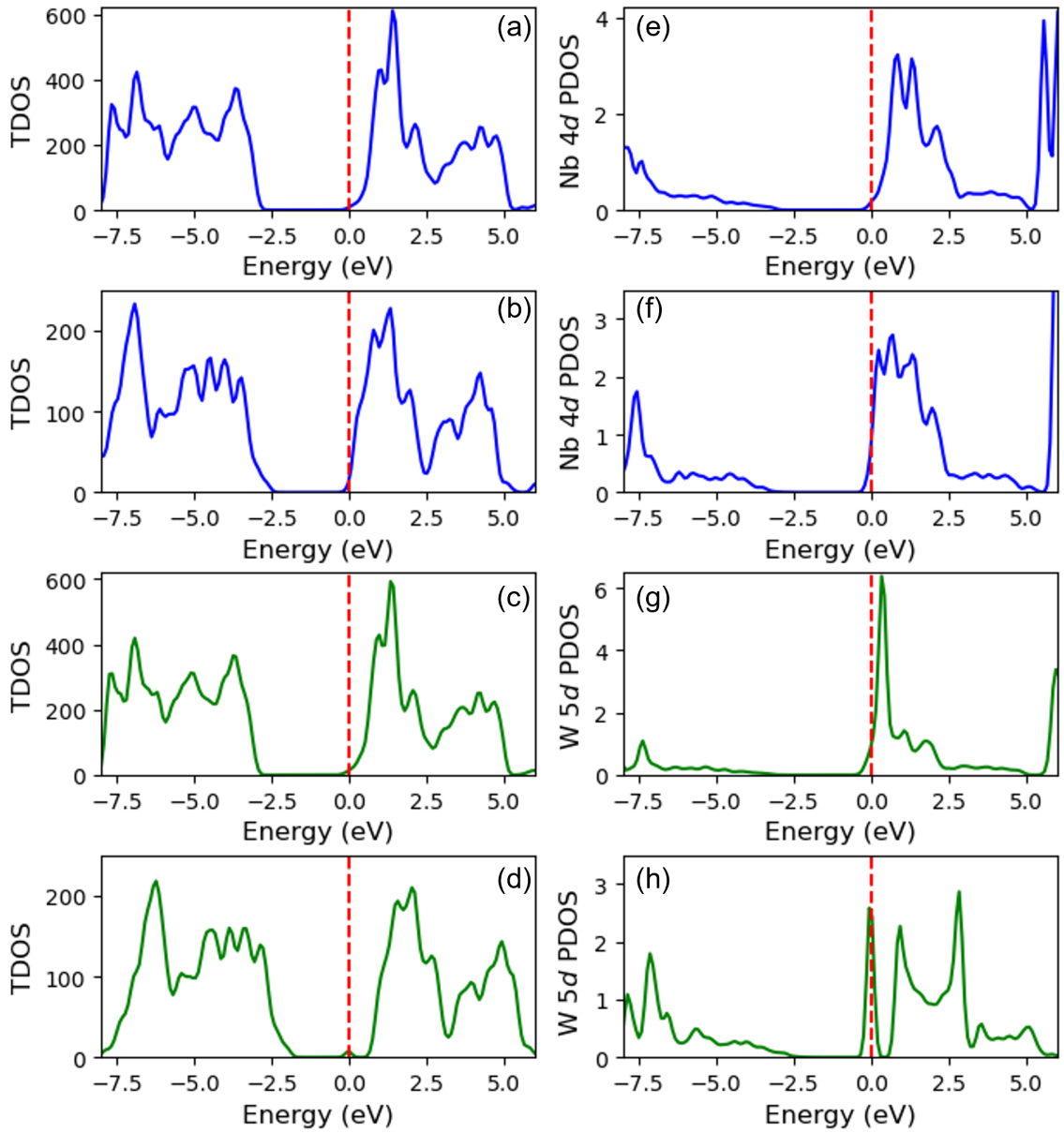


Figure 3.3: Self-consistent DFT+*U*-predicted total density of states (TDOS) and projected density of states (PDOS) for anatase NTO ((a) and (e) respectively), anatase WTO ((b) and (f) respectively), rutile NTO ((c) and (g) respectively) and rutile WTO ((d) and (h) respectively). All TDOS and PDOS are plotted relative to the Fermi level indicated by the red dashed line.

From the corresponding projected density of states in Figure 3.3(e) and Figure 3.3(g), there are

small Nb 4*d* and W 5*d* signatures at the Fermi level indicating these states are partially delocalised over Ti sites, contributing to metallic type behaviour. On the other hand, self-consistent DFT+*U* calculations show that the Nb 4*d* signature in rutile NTO is an order of magnitude greater at the Fermi level than in anatase NTO, when normalising with respect to the different defect concentrations in the simulation supercells, as shown in the projected density of states in Figure 3.3(e) and Figure 3.3(f). These results can be attributed to differences in the filling of the five Nb 4*d* orbitals, notably the greater occupancy of the three t_{2g} orbitals in rutile NTO, which correspond to orbital magnetic quantum numbers m = -2, -1 and 1 [30] in Figure 3.4. There is a negligible difference in the trace of the Nb 4*d* occupation matrix (*i.e.*, the total Nb 4*d* subshell occupancy) in anatase NTO (1.48) compared with rutile NTO (1.49). Figures 3.3(d) and 3.3(h) show the total density of states and W 5*d* projected density of states for rutile WTO, respectively. Here, a localised W 5*d* mid-gap state *ca.* 0.7 eV below the TiO_2 conduction band is predicted. The character of the mid-gap state is W $5d_{z^2}$, corresponding to a large occupation number of 0.93 for the m = 0 orbital in the W 5*d* occupation matrix (Figure 3.4). The other diagonal terms of the W 5*d* occupation matrix are of similar magnitude in anatase WTO and rutile WTO, which suggests the formation of W^{5+} in rutile but not anatase.

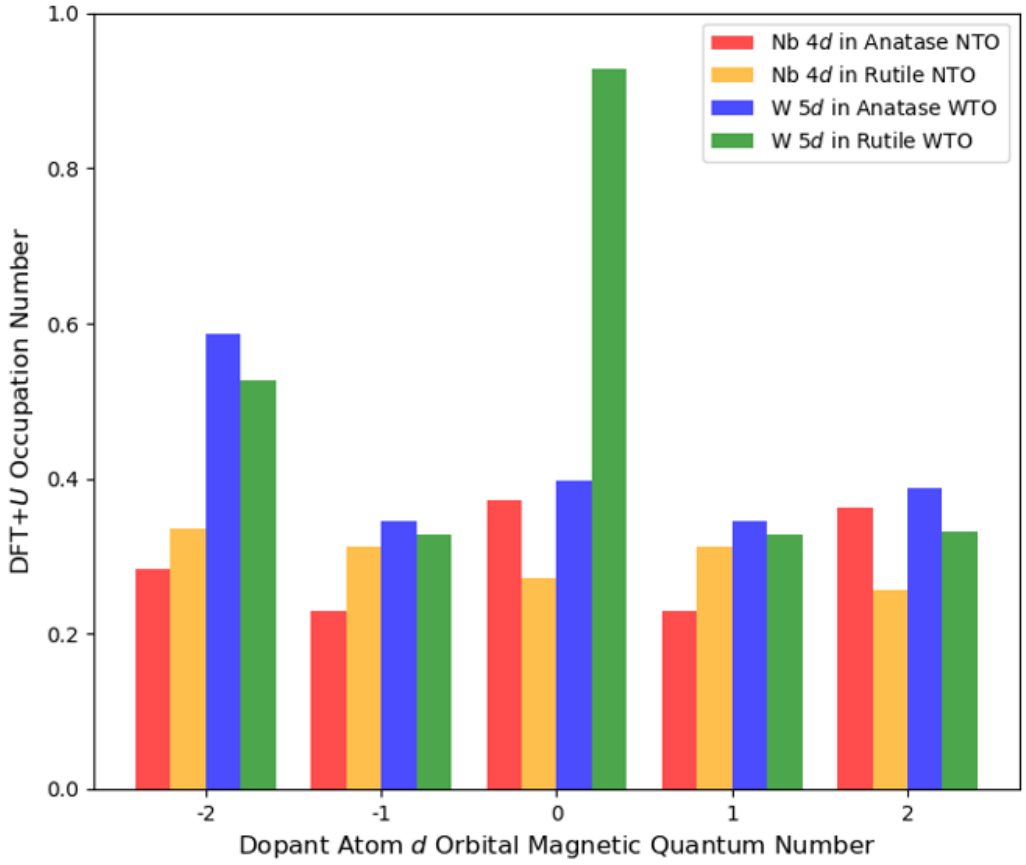


Figure 3.4: Ground state orbital occupation numbers (for orbital magnetic quantum number m) for Nb 4*d* and W 5*d* in doped anatase and rutile TiO_2 calculated using self-consistent DFT+*U*.

The formation of W^{5+} in rutile WTO is also suggested based on the local lattice distortion surrounding the W dopant, which is associated with localised polaronic states in defective TiO_2. [52, 53] This is shown in Figure 3.5, which plots the change in the bond distance between the dopant atom

and six neighbouring Ti atoms relative to the average Ti-Ti bond distance in bulk anatase and rutile TiO_2. Figure 3.5(c) shows *symmetric* geometric relaxation around the substitutional defect in anatase NTO and WTO, where the change in bond length between the dopant atom and Ti atoms A-D is almost constant for both materials, as is the change in bond length with Ti atoms E and F. In Figure 3.5(d), there is a stronger *asymmetric* local lattice distortion around the dopant atom in rutile WTO compared to rutile NTO, as shown by the differences in the change in bond lengths between the dopant atom and Ti atoms G-L.

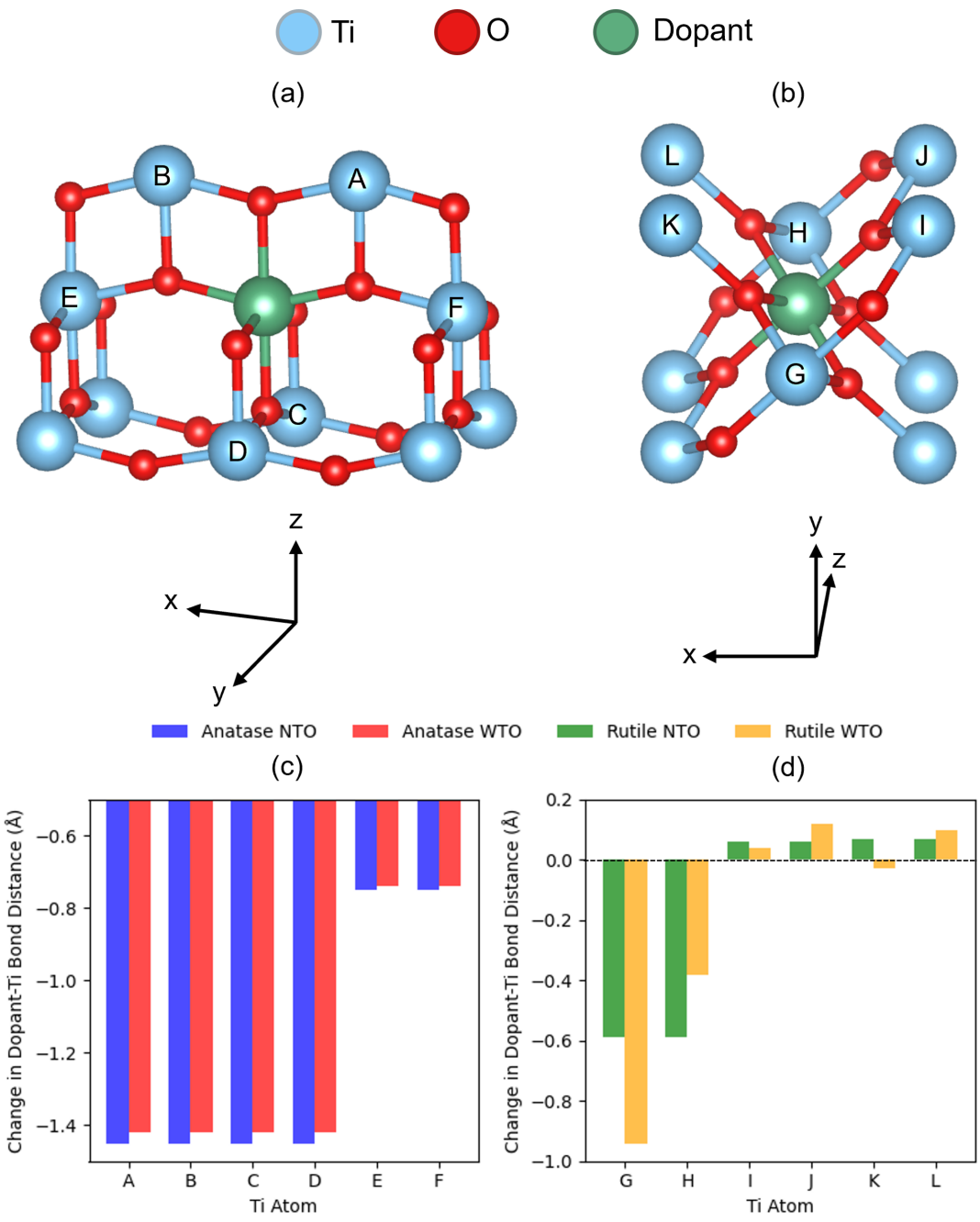


Figure 3.5: Change in the self-consistent DFT+*U* calculated bond distances between the dopant atom and surrounding Ti atoms in doped anatase (atoms A-F in (a)) and doped rutile (atoms G-L in (b)) calculated relative to the average Ti-Ti bond distance in bulk anatase (c) and rutile (d) TiO_2.

Constrained DFT+U with an Atomic Hubbard Projector

Constrained DFT+U simulations with the default atomic atomic Ti 3d Hubbard projector were not able to rationalise the EPR observations. Figures 3.6(a) and (b) show the total and projected density of states for rutile NTO, respectively, with occupation matrices initialised to reflect Nb^{4+}(Nb $4d^1$). Here, there is a localised defect state of Ti 3d character, as shown by Figures 3.6(b) which plots a reduced energy window around the Fermi level. Figures 3.6(c) and (d) show the total and projected density of states for rutile WTO, respectively, with occupation matrices initialised to reflect W^{5+}(W $5d^1$). Here, the Fermi level is pinned to the bottom of the conduction band, indicating the delocalisation of both Ti 3d and W 5d states.

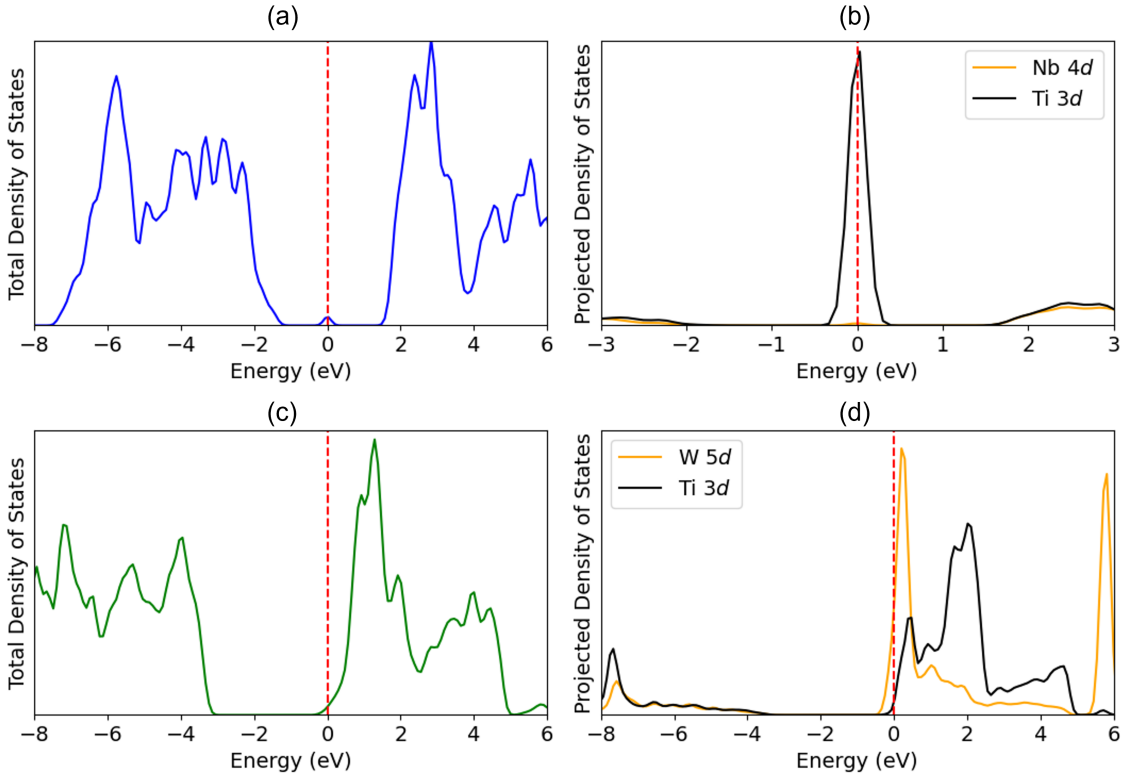


Figure 3.6: Total and projected density of states for rutile NTO ((a) and (b) respectively) and rutile WTO ((c) and (d) respectively) calculated using constrained DFT+U with the default atomic Ti 3d Hubbard projector (U = 3 eV for anatase and 4 eV for rutile, c_1 = 1 and c_2 = 0).

Constrained DFT+U simulations with the atomic Ti 3d Hubbard projector further resulted in unphysically low defect energies, as illustrated in Figure 3.7 for rutile NTO, where a defect energy of -0.19 eV indicates spontaneous dissolution of Nb into the TiO_2 lattice, in contrast to the experimental requirement for thermal activation (calcination) during catalyst preparation. Self-consistent DFT+U simulations with the refined Ti 3d Hubbard projector restore the defect energies for both NTO and WTO to positive values, indicating the thermally activated dissolution, in agreement with the well-established experimental synthesis protocol.

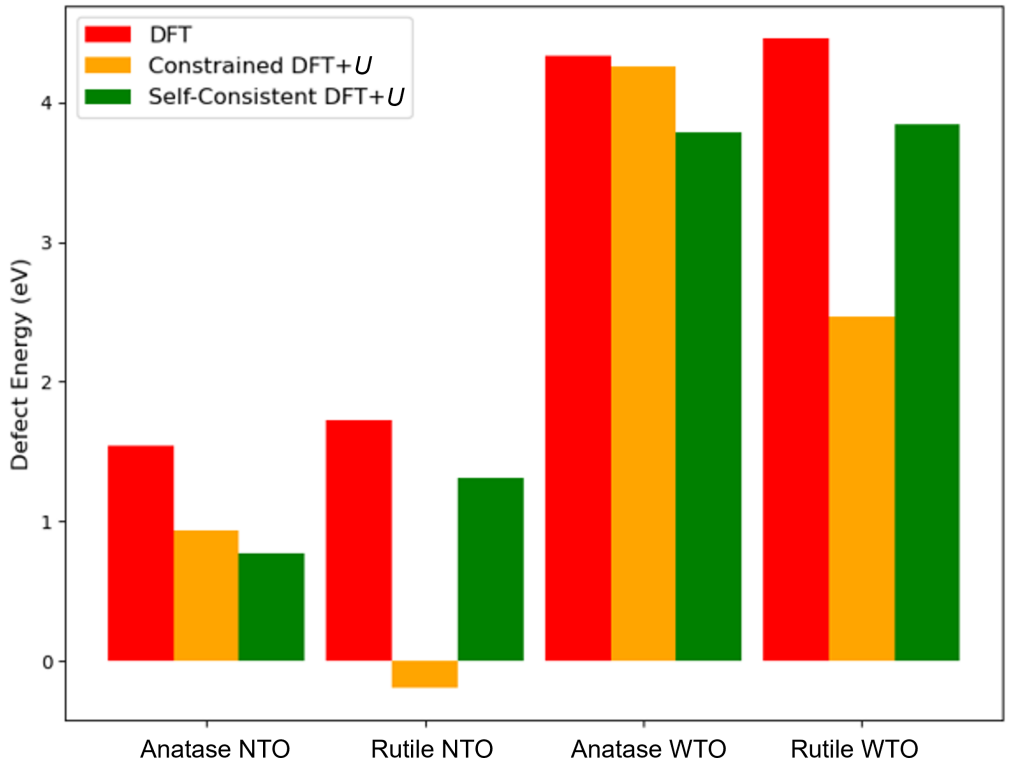


Figure 3.7: Defect energies for anatase and rutile NTO and WTO predicted using DFT, constrained DFT+U (U = 3 eV for anatase and 4 eV for rutile, c_1 = 1 and c_2 = 0) and self-consistent DFT+U (U = 3 eV for both anatase and rutile, c_1 = 0.828 and c_2 = -0.561).

The observed sensitivities of numerical stability, polaron localisation and defect energies with the definition of the Ti 3d Hubbard projector motivates a detailed investigation into the root causes of projector sensitivities of DFT+U simulations, as well as the development of advanced DFT+U parameterisation schemes for the accurate determination of the Hubbard projector. These challenges are discussed in Chapter 4.

3.4 Conclusions

NTO and WTO are promising TCOs with applications in heterogeneous photocatalysis and high efficiency photovoltaics and electronics. However, there remains uncertainty of the atomistic mechanisms that govern charge compensation in these materials, which prevents the development of accurate structure-property models for material optimisation. Using EPR spectroscopy, charge compensation is shown as highly sensitive to the TiO_2 polymorph, with Nb^{4+} and W^{5+} signals present in substitutionally doped rutile but not in doped anatase. These observations are challenging to rationalise theoretically due to the Coulomb self-interaction error in DFT, necessitating DFT+U which applies an *ad-hoc* energy correction to localised orbitals. Both anatase and rutile NTO and WTO are simulated using DFT+U calculations in an all electron numerical atom-centred orbital framework, where self-consistent resolution of the Ti 3d, Nb 4d and W 5d orbital occupancies is crucial to rationalise the EPR observations. Self-consistent DFT+U predicts favourability of Nb^{4+} in rutile NTO through

greater filling of the Nb 4*d* t_{2g} orbitals and reduced filling of the e_g orbitals compared to anatase NTO. Self-consistent DFT+*U* also predicts W^{5+} in rutile WTO through the formation of a localised mid-gap state of $5d_{z^2}$ character that is not formed in anatase WTO.

The combination of theory and experiments provides a coherent view on the reducibility of metal centres in semiconducting TiO_2 without apparent disagreement; whilst also providing a clear understanding on how the reducibility of metal centres and electron trapping energetics in TiO_2 are polymorph-dependent. The results also show the critical influence of the definition of the Ti 3*d* Hubbard projector on the validity of DFT+*U*-predicted geometric, electronic and energetic properties, which motivates further study into advanced DFT+*U* parameterisation strategies for simultaneously optimising *U* values and projectors.

References

(1) A. Chaudhari, A. J. Logsdail and A. Folli, Polymorph-Induced Reducibility and Electron Trapping Energetics of Nb and W Dopants in TiO_2, *J. Phys. Chem. C* 2025, **129** 15453–15461.

(2) Y. Furubayashi, T. Hitosugi, Y. Yamamoto, K. Inaba, G. Kinoda, Y. Hirose, T. Shimada and T. Hasegawa, A transparent metal: Nb-doped anatase TiO_2, *App. Phys. Lett.* 2005, **86** 252101.

(3) S. X. Zhang, D. C. Kundaliya, W. Yu, S. Dhar, S. Y. Young, L. G. Salamanca-Riba, S. B. Ogale, R. D. Vispute and T. Venkatesan, Niobium doped TiO_2: Intrinsic transparent metallic anatase versus highly resistive rutile phase, *J. App. Phys.* 2007, **102** 013701.

(4) S. Lee, J. H. Noh, H. S. Han, D. K. Yim, D. H. Kim, J. kun Lee, J. Y. Kim, H. S. Jung and K. S. Hong, Nb-Doped TiO_2: A New Compact Layer Material for TiO_2 Dye-Sensitized Solar Cells, *J. Phys. Chem. C* 2009, **113** 6878–6882.

(5) C. Tasaki, N. Oka, T. Yagi, N. Taketoshi, T. Baba, T. Kamiyama, S. ichi Nakamura and Y. Shigesato, Thermophysical Properties of Transparent Conductive Nb-Doped TiO_2 Films, *Jpn. J. App. Phys.* 2012, **51** 035802.

(6) H. Y. Lee and J. Robertson, Doping and compensation in Nb-doped anatase and rutile TiO_2, *J. App. Phys.* 2013, **113** 213706.

(7) L. D. Trizio, R. Buonsanti, A. M. Schimpf, A. Llordes, D. R. Gamelin, R. Simonutti and D. J. Milliron, Nb-Doped Colloidal TiO_2 Nanocrystals with Tunable Infrared Absorption, *Chem. Mater.* 2013, **25** 3383–3390.

(8) J. T. Park, W. S. Chi, H. Jeon and J. H. Kim, Improved electron transfer and plasmonic effect in dye-sensitized solar cells with bi-functional Nb-doped TiO_2/Ag ternary nanostructures. *Nanoscale* 2014, **6** 2718–29.

(9) J. Biedrzycki, S. Livraghi, E. Giamello, S. Agnoli and G. Granozzi, Fluorine- and Niobium-Doped TiO_2: Chemical and Spectroscopic Properties of Polycrystalline n-Type-Doped Anatase, *J. Phys. Chem. C* 2014, **118** 8462–8473.

(10) A. Folli, J. Z. Bloh, R. Walker, A. Lecaplain and D. E. Macphee, Properties and Photochemistry of Valence-Induced-Ti^{3+} Enriched (Nb,N)-Codoped Anatase TiO_2 Semiconductors, *Phys. Chem. Chem. Phys.* 2015, **17** 4849–4853.

(11) J. Yue, C. Suchomski, P. Voepel, R. Ellinghaus, M. Rohnke, T. Leichtweiss, M. T. Elm and B. M. Smarsly, Mesoporous niobium-doped titanium dioxide films from the assembly of crystalline nanoparticles: study on the relationship between the band structure, conductivity and charge storage mechanism, *J. Mater. Chem. A* 2017, **5** 1978–1988.

(12) G. Phipps, C. Mikolajczak and T. Guckes, Indium and Gallium: long-term supply, *Renew. Energy Focus* 2008, **9** 56–59.

(13) M. A. Green, The Path to 25% Silicon Solar Cell Efficiency: History of Silicon Cell Evolution, *Prog. Photovolt: Res. Appl.* 2009, **17** 183–189.

(14) P. C. K. Vesborg and T. F. Jaramillo, Addressing the terawatt challenge: scalability in the supply of chemical elements for renewable energy, *RSC Adv.* 2012, **2** 7933.

(15) M. Lokanc, R. Eggert and M. Redlinger, *The availability of indium: the present, medium term, and long term*, tech. rep., National Renewable Energy Lab (NREL), Golden, CO (United States), 2015.

(16) D.-m. Chen, G. Xu, L. Miao, L.-h. Chen, S. Nakao and P. Jin, W-doped anatase TiO_2 transparent conductive oxide films: Theory and experiment, *J. Appl. Phys.* 2010, **107** 063707.

(17) Q. Hou, C. Zhao, S. Guo, F. Mao and Y. Zhang, Effect on electron structure and magneto-optic property of heavy W-doped anatase TiO_2, *PLoS ONE* 2015, **10** 1–14.

(18) H. Wang, C. Zhao, L. Yin, X. Li, X. Tu, G. Lim, Y. Liu and Z. Zhao, W-doped TiO_2 as electron transport layer for high performance solution-processed perovskite solar cells, *Appl. Surf. Sci.* 2021, **563** 150298.

(19) H. Y. Wang, H. Yang, L. Zhang, J. Chen and B. Liu, Niobium Doping Enhances Charge Transport in TiO_2 Nanorods, *ChemNanoMat* 2016, **2** 660–664.

(20) H. Y. Wang, J. Chen, F. X. Xiao, J. Zheng and B. Liu, Doping-induced structural evolution from rutile to anatase: Formation of Nb-doped anatase TiO_2 nanosheets with high photocatalytic activity, *J. Mater. Chem. A* 2016, **4** 6926–6932.

(21) E. Santos, A. C. Catto, A. F. Peterline and W. Avansi, Transition metal (Nb and W) doped TiO_2 nanostructures: The role of metal doping in their photocatalytic activity and ozone gas-sensing performance, *Appl. Surf. Sci.* 2022, **579** 152146.

(22) S. Sathasivam, D. S. Bhachu, Y. Lu, N. Chadwick, S. A. Althabaiti, A. O. Alyoubi, S. N. Basahel, C. J. Carmalt and I. P. Parkin, Tungsten Doped TiO_2 with Enhanced Photocatalytic and Optoelectrical Properties via Aerosol Assisted Chemical Vapor Deposition, *Sci. Rep.* 2015, **5** 10952.

(23) A. Folli, J. Z. Bloh, K. Armstrong, E. Richards, D. M. Murphy, L. Lu, C. J. Kiely, D. J. Morgan, R. I. Smith, A. C. Mclaughlin and D. E. Macphee, Improving the Selectivity of Photocatalytic NO_x Abatement through Improved O_2 Reduction Pathways Using $Ti_{0.909}W_{0.091}O_2Nb_x$ Semiconductor Nanoparticles: From Characterization to Photocatalytic Performanc, *ACS Catal.* 2018, **8** 6927–6938.

(24) T. Hitosugi, H. Kamisaka, K. Yamashita, H. Nogawa, Y. Furubayashi, S. Nakao, N. Yamada, A. Chikamatsu, H. Kumigashira, M. Oshima, Y. Hirose, T. Shimada and T. Hasegawa, Electronic band structure of transparent conductor: Nb-doped anatase TiO_2, *Appl. Phys. Express* 2008, **1** 111203.

(25) K. K. Ghuman and C. V. Singh, A DFT+U study of (Rh, Nb)-codoped rutile TiO_2, *J. Phys.: Condens. Matter.* 2013, **25** 085501.

(26) B. J. Morgan, D. O. Scanlon and G. W. Watson, Small polarons in Nb- and Ta-doped rutile and anatase TiO_2, *J. Mater. Chem.* 2009, **19** 5175–5178.

(27) A. Raghav, K. Hongo, R. Maezono and E. Panda, Electronic structure and effective mass analysis of doped TiO_2 (anatase) systems using DFT+U, *Comput. Mater. Sci.* 2022, **214** 111714.

(28) J. Belošević-Čavor, V. Koteski, A. Umićević and V. Ivanovski, Effect of 5d transition metals doping on the photocatalytic properties of rutile TiO_2, *Comput. Mater. Sci.* 2018, **151** 328–337.

(29) M. Kick, K. Reuter and H. Oberhofer, Intricacies of DFT+U, Not Only in a Numeric Atom Centered Orbital Framework, *J. Chem. Theory Comput.* 2019, **15** 1705–1718.

(30) J. P. Allen and G. W. Watson, Occupation matrix control of d- and f-electron localisations using DFT+U, *Phys. Chem. Chem. Phys.* 2014, **16** 21016–21031.

(31) V. Blum, R. Gehrke, F. Hanke, P. Havu, V. Havu, X. Ren, K. Reuter and M. Scheffler, *Ab initio* molecular simulations with numeric atom-centered orbitals, *Comput. Phys. Commun.* 2009, **180** 2175–2196.

(32) A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, E. D. Hermes, P. C. Jennings, P. B. Jensen, J. Kermode, J. R. Kitchin, E. L. Kolsbjerg, J. Kubal, K. Kaasbjerg, S. Lysgaard, J. B. Maronsson, T. Maxson, T. Olsen, L. Pastewka, A. Peterson, C. Rostgaard, J. Schiøtz, O. Schütt, M. Strange, K. S. Thygesen, T. Vegge, L. Vilhelmsen, M. Walter, Z. Zeng and K. W. Jacobsen, The atomic simulation environment—a Python library for working with atoms, *J. Phys. Condens. Matter.* 2017, **29** 273002.

(33) O. Lamiel-Garcia, K. C. Ko, J. Y. Lee, S. T. Bromley and F. Illas, When Anatase Nanoparticles Become Bulklike: Properties of Realistic TiO_2 Nanoparticles in the 1–6 nm Size Range from All Electron Relativistic Density Functional Theory Based Calculations, *J. Chem. Theory Comput.* 2017, **13** 1785–1793.

(34) J. Wellendorff, K. T. Lundgaard, K. W. Jacobsen and T. Bligaard, mBEEF: An accurate semi-local Bayesian error estimation density functional, *J. Chem. Phys.* 2014, **140** 144107.

(35) J. P. Perdew, A. Ruzsinszky, G. I. Csonka, O. A. Vydrov, G. E. Scuseria, L. A. Constantin, X. Zhou and K. Burke, Restoring the Density-Gradient Expansion for Exchange in Solids and Surfaces, *Phys. Rev. Lett.* 2008, **100** 136406.

(36) S. Lehtola, C. Steigemann, M. J. Oliveira and M. A. Marques, Recent developments in libxc — A comprehensive library of functionals for density functional theory, *SoftwareX* 2018, **7** 1–5.

(37) C. G. Broyden, The Convergence of a Class of Double-rank Minimization Algorithms 1. General Considerations, *IMA J. Appl.* 1970, **6** 76–90.

(38) R. Fletcher, A new approach to variable metric algorithms, *Comput. J.* 1970, **13** 317–322.

(39) D. F. Shanno, Conditioning of Quasi-Newton Methods for Function Minimization, *Math. Comput.* 1970, **24** 647–656.

(40) D. Goldfarb, A Family of Variable-Metric Methods Derived by Variational Means, *Math. Comput.* 1970, **24** 23–26.

(41) F. Birch, Finite Elastic Strain of Cubic Crystals, *Phys. Rev.* 1947, **71** 809–824.

(42) L. Kavan, M. Grätzel, S. E. Gilbert, C. Klemenz and H. J. Scheel, Electrochemical and Photoelectrochemical Investigation of Single-Crystal Anatase, *JACS* 1996, **118** 6716–6723.

(43) T. Arlt, M. Bermejo, M. A. Blanco, L. Gerward, J. Z. Jiang, J. Staun Olsen and J. M. Recio, High-pressure polymorphs of anatase TiO_2, *Phys. Rev. B* 2000, **61** 14414–14419.

(44) K. Jakob and H. Oberhofer, "Self-Consistency in the Hubbard-Corrected DFT+*U* Method", Master's thesis, Faculty of Chemistry, Technical University of Munich, 2021.

(45) N. Doebelin and R. Kleeberg, Profex: A graphical user interface for the Rietveld refinement program BGMN, *J. Appl. Cryst.* 2015, **48** 1573–1580.

(46) S. Stoll and A. Schweiger, EasySpin, a Comprehensive Software Package for Spectral Simulation and Analysis in EPR, *J. Magn. Reson.* 2006, **178** 42–55.

(47) A. Folli, J. Bloh, E.-P. Beukes, R. Howe and D. MacPhee, Photogenerated charge carriers and paramagnetic species in (W,N)-codoped TiO_2 photocatalysts under visible-light irradiation: An EPR study, *J. Phys. Chem. C.* 2013, **117**.

(48) J. Z. Bloh, A. Folli and D. E. Macphee, Adjusting Nitrogen Doping Level in Titanium Dioxide by Codoping with Tungsten: Properties and Band Structure of the Resulting Materials, *J. Phys. Chem. C* 2014, **118** 21281–21292.

(49) J Kiwi, J. T. Suss and S Szapiro, EPR spectra of niobium-doped TiO_2 and implications for water photocleavage processes, *Chem. Phys. Lett.* 1984, **106** 3–6.

(50) P. H. Zimmermann, Temperature Dependence of the EPR Spectra of Niobium-Doped TiO_2, *Phys. Rev. B* 1973, **8** 3917–3927.

(51) T. T. Chang, Paramagnetic-Resonance Spectrum of W^{5+} in Rutile (TiO_2), *Phys. Rev.* 1966, **147** 264–267.

(52) C. Lin, D. Shin and A. A. Demkov, Localized states induced by an oxygen vacancy in rutile TiO_2, *J. Appl. Phys.* 2015, **117** 225703.

(53) C. M. Yim, M. B. Watkins, M. J. Wolf, C. L. Pang, K. Hermansson and G. Thornton, Engineering Polarons at a Metal Oxide Surface, *Phys. Rev. Lett.* 2016, **117** 116402.

Chapter 4

Machine Learning Generalised DFT+U Projectors in a Numerical Atom-Centred Orbital Framework

This chapter is based on the published work *Machine learning generalised DFT+U projectors in a numerical atom-centred orbital framework* in *Digital Discovery*, which is co-authored by Dr Kushagra Agrawal (Cardiff University, CU) and Dr Andrew Logsdail (CU). [1]

The work builds upon Chapter 3, which highlights the necessity of self-consistent DFT+U simulations to achieve experimentally accurate predictions of defects and polarons in TiO_2. However, the difficulties in performing such simulations with atomic Hubbard projectors necessitates the development of more advanced strategies for DFT+U parameterisation. Dr Kushagra Agrawal and I formulated the method and I performed the electronic structure and supervised machine learning calculations. All Python scripts for global optimisation, datasets for regression/classification and input/output files for electronic structure calculations available open-source in the GitHub repository *https://github.com/amitmc1/Hubbardprojectors*.

4.1 Introduction

As discussed in Section 2.3.2 and Chapter 3, DFT+U can help mitigate the Coulomb self-interaction error in simulations of defects in TMOs and REOs, whilst maintaining the computational efficiency of standalone semi-local DFT. However, Chapter 3 highlights the difficulties in simulating polarons in TiO_2 with an accuracy that matches experimental observations, using the default atomic Ti $3d$ Hubbard projector (in either constrained or self-consistent DFT+U calculations). [2] Typically, the Hubbard U value is parameterised with no consideration of the Hubbard projector, using semi-empirical benchmarking of DFT+U-predicted electronic, geometric and energetic properties against reference data from experiments or higher levels of theory. [3, 4] These semi-empirical approaches are reliant on accurate and available reference data, which prevents the high-throughput optimisation of Hubbard U values for vast numbers of materials. Several first-principles approaches for computing Hubbard U values have been developed, including the linear response approach based on constrained DFT (LR-cDFT), [5] density functional perturbation theory (DFPT), [6] constrained random phase approximation (cRPA) [7] and Hartree Fock based approaches (*e.g.*, UHF and ACBN0). [8, 9] Of

these methods, LR-cDFT is most popular and has been applied in a planewave basis for the high-throughput optimisation of Hubbard U values for over 1000 transition metal oxides. [10] However, LR-cDFT can yield unphysical U values, suffers from numerical instability for closed-shell systems [11, 12] and is computationally expensive due to the requirement of calculations using large supercells. [13] In addition, Hubbard U values computed using LR-cDFT are site-specific and therefore not transferable from stoichiometric systems with symmetry-equivalent sites to defective systems with broken symmetry. [14] The lack of transferability can prevent the simulation of experimentally observed electrical conductivities of defective TMOs, including the Li-ion battery cathode material $LiCo_{1-x}Mg_xO_2$, where deep defect states are predicted using Hubbard U values from LR-cDFT in a planewave basis, thus incorrectly suggesting material resistivity. [15]

To minimise the cost and instability of first-principles methods for computing Hubbard U values, active learning methods have been combined with global optimisation algorithms, such as Bayesian optimisation (BO) and Monte Carlo sampling, to minimise a cost function based on reproducing geometric and electronic properties predicted using higher levels of theory. [16–18] In active learning, the cost function is refined by comparison with the output of successive DFT+U calculations, which means there is no *a priori* knowledge of the DFT+U potential energy surface and the entire approach must be repeated for different systems. Supervised learning approaches have been used to improve the transferability of active learning methods for computing Hubbard U values, by attempting to learn the DFT+U potential energy surface for different materials. For example, BO and Random Forest Regression have been used to determine the structure-dependence of the Mn 3d Hubbard U value required to reproduce the electronic band structures of various MnO_x polymorphs computed using hybrid-DFT. [19] Similarly, equivariant neural networks have been used to estimate DFPT-predicted Hubbard U values using ground state atomic occupation matrices and interatomic distances for a range of materials, including Li_xFePO_4 and MnO_2. [20] However, none of these methods address the parameterisation of Hubbard projectors, as well as differences in the numerical stability of DFT+U calculations with different Hubbard parameters.

The aforementioned challenges in DFT+U parameterisation are considered in this chapter in the context of simulations of stoichiometric and defective TMOs and REOs in a NAO framework, in which there is a strong dependence of the accuracy and numerical stability of DFT+U calculations on both the Hubbard U value *and* the projector. Relying on semi-empirically derived Hubbard U values and the default atomic Hubbard projector can result in inaccurate, unphysical (calculations terminate due to excessive polaron localisation) and unstable (SCF cycle does not converge) simulations of common TMOs, *e.g.*, TiO_2, and REOs, *e.g.*, CeO_2. Thus, simultaneous optimisation of the Hubbard U value *and* the projector is demonstrated for Ti 3d orbitals in anatase TiO_2 using BO with a cost function defined using symbolic regression (SR), to minimise the errors of target properties relative to experimental references. BO is also subject to constraints on the DFT+U-predicted covalency, to ensure the numerical stability of point defect calculations, as determined using support vector machines (SVMs). Combining SR, SVMs and BO in this manner avoids the need for multiple successive DFT+U calculations, which significantly reduces the overall computational cost for Hubbard parameter optimisation compared to the existing first-principles and active learning approaches.

The method is then extended across materials (beyond TiO_2) by expanding the primary feature space for SR to include DFT-predicted orbital occupancies, basis set parameters and atomic material

descriptors for a diverse training set of TMOs and REOs; enabling the use of hierarchical SR [21] to optimise Hubbard U values and projectors from first-principles by targeting orbital occupancies calculated using hybrid-DFT. The outcome is a transferable approach for the one-shot computation of Hubbard U values and projectors, with good accuracy that is also achieved for unseen materials. Overall, the work demonstrates the development of cost-effective and transferable machine learning-based workflows for more complete DFT+U parameterisation, enabling more accurate and efficient simulations of complex energy materials.

4.2 Methodology

4.2.1 Electronic Structure Calculations

DFT

All electronic structure calculations were performed using the Fritz-Haber Institute *ab initio* materials simulation (FHI-aims) software, [22] with the same DFT parameters as discussed in Section 3.2.1. Unit cell equilibrium volumes (V_0) were calculated by fitting to the Birch-Murnaghan equation of state using ASE. [23] Where presented, formation energies (ΔE_{Form}) for TiO_2, CeO_2 and $LiCoO_2$ were calculated using the energies of bulk Ti (in the hexagonal close packed, HCP, crystal structure), bulk Ce, Li and Co (all in the cubic crystal structure) and an isolated O_2 molecule using:

$$\Delta E_{\text{Form}} = E_{\text{Compound}} - \sum_i n_i E_i - \frac{n_{\text{O}}}{2} E_{\text{O}_2} \tag{4.1}$$

where i denotes the metal species index in each compound and n_i (n_O) is the number of metal (oxygen) atoms in the formula unit.

DFT+U

All DFT+U calculations were performed using the on-site definition of the occupation matrix and the Fully Localised Limit (FLL) double counting correction.[24] In Section 4.3.1, a Hubbard correction is applied to treat the Coulomb self-interaction of Ti 3d and Ce 4f orbital electrons in tetragonal TiO_2 and cubic CeO_2, respectively, using the corresponding default atomic Hubbard projector. Constrained DFT+U calculations were performed using the "*occupation matrix control*" (OMC) method [25] to *fix* polaron(s) at specific atom(s) by modifying the corresponding atomic orbital occupation matrix (as outlined in Section 2.3.2). [24–26] In Section 4.3.2, self-consistent DFT+U calculations are performed for TiO_2 using a modified atomic-like Ti 3d Hubbard projector, determined using the semi-empirical machine learning approach detailed in Section 4.2.2. This is expanded in Section 4.3.3, where self-consistent DFT+U calculations are performed for all materials in Table 2.1 in Section 2.3.2, using modified atomic-like Hubbard projectors that were determined using the first-principles machine learning approach detailed in Section 4.2.3. Modified atomic-like Hubbard projectors were defined as a linear combination of the atomic and hydrogenic auxiliary NAO basis functions, if present in the light basis set, as outlined in Section 2.3.2. The linear expansion coefficients c_1 and c_2 correspond to the atomic and auxiliary functions, respectively.

Self-consistent DFT+U calculations were performed with a refined atomic-like Ti 3d Hubbard projector using the "*occupation matrix release*" (OMR) method to *initialise* polaron(s) at specific atom(s) before self-consistent determination of the system occupation matrices (as outlined in in Section 2.3.2). [24] All point defect calculations in this work were performed in a 3×3×3 supercell, which avoids spurious long-range defect-defect interactions between periodic images. The oxygen vacancy formation energy (ΔE_{OV}) and the defect energies ($\Delta E_{\mathrm{Defect}}$) following substitution of a host metal atom (Ti in TiO_2 and Co in $LiCoO_2$) with a Nb, W, Co, Mn, Pt, Au, Pd or Mg atom, are calculated as:

$$\Delta E_{\mathrm{OV}} = E_{\text{Oxygen Deficient Bulk}} + \mu_{\mathrm{O}} - E_{\text{Stoichiometric Bulk}} \tag{4.2}$$

$$\Delta E_{\mathrm{Defect}} = E_{\text{Doped Bulk}} + \mu_{\mathrm{Host}} - E_{\text{Stoichiometric Bulk}} - \mu_{\mathrm{Dopant}} \tag{4.3}$$

where the chemical potentials were calculated using the energy of half an isolated O_2 molecule (μ_{O}), bulk Ti in the HCP crystal structure (μ_{Ti}) and the bulk dopant species (μ_{Dopant}) in the cubic crystal structure except Mn (tetragonal) and Mg (HCP). No Hubbard correction was applied to the dopant atoms.

Both the semi-empirical and first-principles machine learning approaches in Sections 4.3.2 and 4.3.3, respectively, require the evaluation of metal d or f and O 2p orbital occupancies for DFT (mBEEF), DFT+U (mBEEF+U) and hybrid-DFT (PBE0). For DFT and hybrid-DFT, orbital occupancies were evaluated using DFT+U, with a Hubbard U value of 0 eV for metal d or f and O 2p states, as well as the default atomic projector ($c_1 = 1$ and $c_2 = 0$). For DFT+U, the parameters U, c_1 and c_2 were varied for metal d or f states only, whilst O 2p states were again treated with a Hubbard U value of 0 eV and the atomic O 2p Hubbard projector. Consequently, O 2p orbital occupancies can be directly compared across all methods, as they are consistently derived from the atomic O 2p Hubbard projector. In contrast, discrepancies between the orbital occupancies of metal d or f states between methods arise from the different definitions of the Hubbard projectors. For this reason, the cost function for the first-principles machine learning approach in Section 4.2.3 is based on O 2p orbital occupancies calculated using DFT, DFT+U and hybrid-DFT, which can be compared in a like-for-like manner. All DFT+U predicted properties and orbital occupancies correspond to the outputs of geometry optimisation calculations, with the DFT+U contributions to atomic forces provided in the code without any modifications required. [24]

4.2.2 Semi-Empirical Machine Learning Approach

A semi-empirical approach was adopted to optimise the Ti 3d Hubbard U value and projector to enable accurate and numerically stable simulations of anatase TiO_2, using DFT+U-calculated properties of the TiO_2 unit cell (targets for regression) and the classified outcome of point defect calculations in a TiO_2 supercell using the OMR method (targets for classification), as illustrated in Figure 4.1.

U	c_1	c_2	Optimised Unit Cell					Optimised Supercell
			E_{bg} (eV)	V_0 (Å^3)	Tr[**n**(Ti 3*d*)]	Tr[**n**(O 2*p*)]	E (eV)	Classification
1	1	0	2.72	137.76	1.06	5.06	…	Converges
3	1	0	3.13	137.07	0.79	5.39	…	Terminates
3	0.9	-0.5	3.06	137.79	0.99	4.68	…	Doesn't Converge

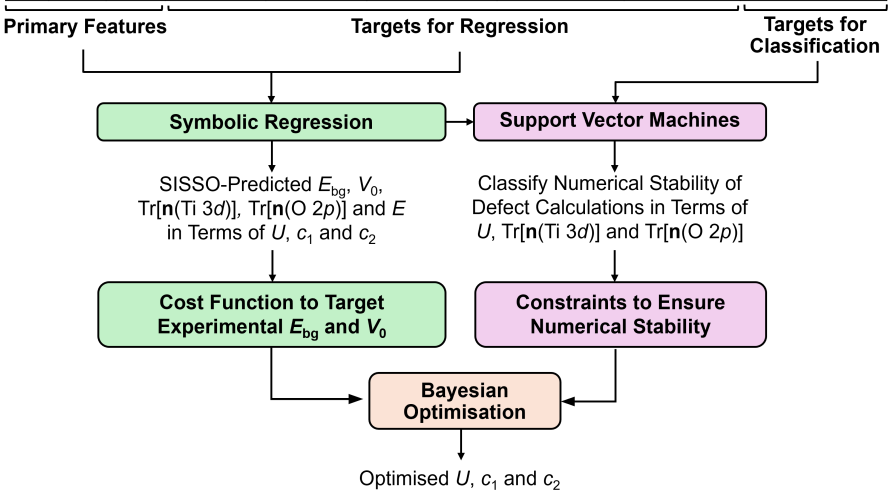


Figure 4.1: Semi-empirical approach for simultaneously optimising the Ti 3*d* Hubbard *U* value and projector for anatase TiO_2, using the DFT+*U*-predicted band gap (E_{bg}), unit cell equilibrium volume (V_0), occupation matrix trace for Ti 3*d* (Tr[**n**(Ti 3*d*)]) and O 2*p* (Tr[**n**(O 2*p*)]) orbitals, total energy (E) and the classified results of bulk oxygen vacancy calculations using OMR.

Symbolic Regression

SR was performed using the Sure Independence Screening and Sparsifying Operator (SISSO) algorithm, [27] as implemented in the SISSO++ package, [28, 29] to fit empirical correlations for target properties in terms of the primary features U, c_1 and c_2. Empirical correlations were constructed by searching a non-linear secondary feature space by recursively combining the primary features using the algebraic operators $+$, $-$, $\times$, $\div$, x^2, x^3, $\sqrt{x}$, $\sqrt[3]{x}$, $\exp(x)$, $\log(x)$, $\sin(x)$, and $\frac{1}{x}$, before using sparse regression techniques to select a minimal set of secondary features and linear regression to optimise the coefficients of the final expression. Empirical correlations were fitted with up to three terms, using a recursive depth of three, yielding a linear combination of non-linear terms, *e.g.*, F_1, F_2 and F_3 using the constants a_0, a_1, a_2 and a_3 for a three term correlation:

$$\text{Target} = a_0 + a_1 \times F_1 + a_1 \times F_2 + a_3 \times F_3 \tag{4.4}$$

The DFT+*U*-predicted band gap (E_{bg}), V_0 and traces of the Ti 3*d* (Tr[**n**(Ti 3*d*)]) and O 2*p* (Tr[**n**(O 2*p*)]) occupation matrices were used as target properties. The accuracy of each SISSO correlation was evaluated using the Pearson's coefficient of determination (R^2) and root mean squared error (RMSE). The SISSO-computed empirical correlations for the DFT+*U*-predicted E_{bg} (eV), $\bar{V}_0$ (normalised, hence unitless), Tr[**n**(Ti 3*d*)] (unitless), Tr[**n**(O 2*p*)] (unitless) and E (eV) are listed below, in terms of the primary features U (eV), c_1 (unitless) and c_2 (unitless), where the chosen number of terms for each empirical correlation gave the best model accuracy and a constant $\alpha = 1$

eV^{-1} is introduced to ensure dimensional consistency:

$$E_{\mathrm{bg}}^{\mathrm{SISSO}} = \mathrm{a}_0 + \mathrm{a}_1 \left(\left(c_1^6 \times \ln(c_1) \right) \times \left(e^{\alpha U} \times \sin(c_2) \right) \right) + \mathrm{a}_2 \left(\frac{c_2^6 \times \alpha U}{c_1^6 - \sin(c_1)} \right) + \mathrm{a}_3 \left(e^{c_1^3} \times \left(\frac{\alpha U}{c_1} + c_2^3 \right) \right) \tag{4.5}$$

$$\bar{V}_0^{\mathrm{SISSO}} = \mathrm{a}_0 + \mathrm{a}_1 (\bar{c}_1^6 \times \bar{c}_2 \times \bar{U}) + \mathrm{a}_2 \left(|\bar{c}_1 + \bar{U} - \bar{c}_2^6| \right) - \mathrm{a}_3 (\bar{c}_2^3 \times \bar{U}_2^2) \tag{4.6}$$

$$\mathrm{Tr}[\mathbf{n}(\mathrm{Ti}\ 3d)]^{\mathrm{SISSO}} = \mathrm{a}_0 + \mathrm{a}_1 \left(\frac{1.0}{\alpha U} - c_2 \times \alpha U \right) \times (\cos(c_1)^6) + \mathrm{a}_2 \left(e^{c_2 \times \alpha U} - (c_1^3 - \sqrt[3]{\alpha U}) \right) \tag{4.7}$$

$$\mathrm{Tr}[\mathbf{n}(\mathrm{O}\ 2p)]^{\mathrm{SISSO}} = \mathrm{a}_0 + \mathrm{a}_1 \left(\cos(\sqrt{\alpha U}) - \frac{\alpha U}{c_1} \sin(c_2) \right) \tag{4.8}$$

$$E^{\mathrm{SISSO}} = \mathrm{a}_0 + \mathrm{a}_1 \times \frac{\sqrt{e^{\alpha U}}}{(c_2 + c_1) + \sin(c_2)} + \mathrm{a}_2 \times \frac{(e^{c_1} - c_2)}{\left(\frac{1}{\alpha U} + (c_2^6) \right)} \tag{4.9}$$

where the corresponding SISSO-computed constants and accuracy metrics are listed in Table 4.1.

Table 4.1: Constants for all SISSO correlations used in the semi-empirical approach for Hubbard projector optimisation, with corresponding model accuracy metrics including the Pearson's coefficient of determination (R^2) and the root mean squared error (RMSE). All constants are unitless except those with associated units noted in brackets, thus ensuring dimensional consistency with the target properties, which are unitless except $E_{\mathrm{bg}}^{\mathrm{SISSO}}$ and E^{SISSO} (both eV)

Target Property	a_0	a_1	a_2	a_3	R^2	RMSE
$E_{\mathrm{bg}}^{\mathrm{SISSO}}$	2.53583 (eV)	-0.04786 (eV)	-0.47434 (eV)	0.06854 (eV)	0.99684	0.00987
$\bar{V}_0^{\mathrm{SISSO}}$	0.75501	0.78908	0.09107	-0.72346	0.95916	0.03155
$\mathrm{Tr}[n(\mathrm{Ti}\ (3d))]^{\mathrm{SISSO}}$	1.67456	-1.25868	-0.59587	N/A	0.98895	0.01911
$\mathrm{Tr}[n(\mathrm{O}\ (2p))]^{\mathrm{SISSO}}$	5.29651	-0.42443	N/A	N/A	0.99003	0.01625
E^{SISSO}	-109824 (eV)	-0.00396 (eV)	0.58472 (eV)	N/A	0.99817	0.06243

Min-max feature scaling was performed for all primary features and target properties *i.e.*, $\bar{x}$, for the SISSO correlation for V_0, using Equation 4.10 and the maximum and minimum values in Table 4.2:

$$\bar{x} = \frac{x - x_{\mathrm{min}}}{x_{\mathrm{max}} - x_{\mathrm{min}}} \tag{4.10}$$

The SISSO correlations were then used to evaluate a unitless, regularised cost function (J^{SE}), which is defined as the Euclidean norm of two terms J_1^{SE} and J_2^{SE}. J_1^{SE} is itself the Euclidean norm of the percentage errors of the DFT+U-predicted E_{bg} and V_0 of bulk anatase TiO_2 versus experimental references from the literature (E_{bg}^{Exp} and V_0^{Exp}, respectively). [30, 31] J_2^{SE} is an additional regularisation term to bias J^{SE} towards larger values of U and c_1, which favours stronger polaron localisation at point defects that is consistent with the experimentally observed formation of mid-gap states within

Table 4.2: Constants for normalising the primary features and target properties for the $\bar{V_0}$ SISSO correlation, where all properties are unitless except U (eV) and V_0 (Å^3).

Property	Maximum	Minimum
U	3.0	0.5
c_1	1.1	0.9
c_2	0.0	-0.5
V_0	137.96	137.07

the TiO_2 band gap. [2, 24, 32] J^{SE} therefore takes the form:

$$J^{\mathrm{SE}} = \left\| J_1^{SE},\ J_2^{SE} \right\| = \left\| \left\| \frac{100 \times (E_{bg}^{\mathrm{DFT}+U} - E_{bg}^{\mathrm{Exp}})}{E_{bg}^{\mathrm{Exp}}},\ \frac{100 \times (V_0^{\mathrm{DFT}+U} - V_0^{\mathrm{Exp}})}{V_0^{\mathrm{Exp}}} \right\|,\ \left(\frac{1000}{\alpha U + c_1} \right) \right\| \tag{4.11}$$

where J_2^{SE} involves a constant $\alpha = 1\ \mathrm{eV}^{-1}$ to ensure dimensional consistency, whilst being scaled by a factor of 1000 to ensure normalisation with respect to J_1^{SE}.

Support Vector Machines

To investigate numerical instability in self-consistent point defect calculations in TiO_2, classification was performed with linear support vector machines (SVMs) using the *Scikit-learn* Python library. [33] The SVMs were used to determine the equations of the boundaries S_1 and S_2 separating regions in the feature space U, $\mathrm{Tr}[\mathbf{n}(\mathrm{Ti}\ 3d)]$ and $\mathrm{Tr}[\mathbf{n}(\mathrm{O}\ 2p)]$, where calculations rapidly terminated due to unphysical predictions or did not converge due to *"charge sloshing"* when simulating a bulk oxygen vacancy using the OMR method. SVM classification was also performed to investigate the relationship between the Hubbard parameters and erroneous oxygen vacancy formation energies, which ranged from -7.11 eV to 14.35 eV depending on the choice of U, c_1 and c_2. Here, classification was performed to determine the equation of the linear boundary S_3 separating regions of "*physical*" (4 eV $\leq \Delta E_{\mathrm{OV}} \leq$ 6 eV) and "*unphysical*" ($\Delta E_{\mathrm{OV}} <$ 4 eV or $\Delta E_{\mathrm{OV}} >$ 6 eV) oxygen vacancy formation energies, in terms of the partial derivatives of the SISSO-predicted total energy (E^{SISSO}) with respect to U, c_1 and c_2. Numerical partial derivatives of E^{SISSO} with respect to each Hubbard parameter U ($\frac{\partial E^{\mathrm{SISSO}}}{\partial U}$), c_1 ($\frac{\partial E^{\mathrm{SISSO}}}{\partial c_1}$) and c_2 ($\frac{\partial E^{\mathrm{SISSO}}}{\partial c_2}$) were calculated using the forward finite difference method with a step size of 0.01.

Both SR and SVM classification were performed using a training set of ≤ 60 geometry optimised unit cells and bulk oxygen vacancy calculations, where all computational settings are kept constant except the Ti 3d Hubbard parameters. Feature scaling was performed for all primary features and target properties (denoted using $\bar{x}$) for the SISSO correlation for V_0 and SVM boundaries S_1 and S_2. The normalisation constants are listed with the non-linear terms, constants and accuracy metrics for all SISSO correlations and SVM boundaries in Table 4.3.

Table 4.3: Constants for normalising the primary features and target properties for the linear SVM boundaries S_1 and S_2 using Equation 4.10, where all properties are unitless except U (eV) and V_0 (Å^3).

Property	Maximum	Minimum
U	3.0	0.5
c_1	1.1	0.9
c_2	0.0	-0.5
V_0	137.96	137.07
Tr[**n**(Ti 3d)]	1.51584	0.77094
Tr[**n**(O 2p)]	5.42824	4.67555

The SVM boundaries S_1 and S_2, which classify the validity of bulk oxygen vacancy calculations in anatase TiO_2 using the OMR method, were fitted using the equation below, where A, B, C and D are constants in Table 4.4:

$$S_i = \mathrm{A}_i \times \overline{U}^{\mathrm{SISSO}} + 1.13 \times \mathrm{B}_i \times \overline{\mathrm{Tr}}[\mathbf{n}(\mathrm{Ti}\ 3d)]^{\mathrm{SISSO}} + \mathrm{C}_i \times \overline{\mathrm{Tr}}[\mathbf{n}(\mathrm{O}\ 2p)]^{\mathrm{SISSO}} + \mathrm{D}_i \tag{4.12}$$

where an *ad-hoc* prefactor of 1.13 is introduced to improve the prediction accuracy of Hubbard parameters that lead to numerically stable point defect calculations, when validated on out-of-training data (representing the uncertainty in the linear boundaries S_1 and S_2). The SVM boundary S_3, which classifies "*physical*" and "*unphysical*" oxygen vacancy formation energies in anatase TiO_2 using the OMR method, was fitted using the equation below, where A, B, C and D are constants in Table 4.4:

$$S_3 = \mathrm{A}_3 \cdot \frac{\partial E^{\mathrm{SISSO}}}{\partial U} + \mathrm{B}_3 \cdot \frac{\partial E^{\mathrm{SISSO}}}{\partial c_1} + \mathrm{C}_3 \cdot \frac{\partial E^{\mathrm{SISSO}}}{\partial c_2} + \mathrm{D}_3 \tag{4.13}$$

Table 4.4: Constants for all linear SVM boundaries S_1, S_2 and S_3 used in the semi-empirical approach for Hubbard projector optimisation, as well as the corresponding proportion of misclassified data points (sum of false positive and negative classification predictions) when evaluated on the respective training sets. All constants are unitless except those with associated units noted in brackets, thus ensuring dimensional consistency with the unitless target properties.

S_i	A_i	B_i	C_i	D_i	Misclassified (%)
1	-4.01	-1.14	-6.29	-5.16	2.27
2	-5.65	-4.07	1.36	-4.65	0.00
3	0.034	-0.15 (eV^{-1})	-1.03 (eV^{-1})	1.01	0.00

Bayesian Optimisation

BO was performed using the *GPyOpt* Python library [34] to probabilistically search the Hubbard parameter space by minimising J^{SE} whilst satisfying three constraints, including those derived from S_1 and S_2. The Hubbard parameter space was searched using the standard Expected Improvement acquisition function, [35] with a sampling weight of 0.01 to encourage parameter exploration within bounds for U between 0.5 eV and 5 eV, c_1 between 0 and 1.3 and c_2 between -0.6 and 0; as beyond these values, numerical instability including calculation termination and non-convergence was

observed when simulating the TiO_2 unit cell. An initial population of 1000 combinations of Hubbard parameters were defined using Latin Hypercube Sampling, [36] using the *pyDOE* Python library, [37] to build a surrogate model of J^{SE} that was minimised further using BO for 350 iterations. Sampled Hubbard parameters that violated constraints were penalised by assigning them a high value of J^{SE} = 1000, therefore discouraging the exploration of the Hubbard parameter space that leads to numerically unstable calculations. The effectiveness of BO in identifying the global minimum J^{SE} was assessed by comparing the BO-sampled landscape of J^{SE} with that evaluated using random sampling for 50,000 iterations.

4.2.3 First-Principles Machine Learning Approach

Hierarchical Symbolic Regression

To extend the semi-empirical approach across different materials, a first-principles approach was adopted to parameterise Hubbard U values and projectors by targeting the O 2p orbital occupancies calculated using hybrid-DFT, which was identified as crucial to enable self-consistency of DFT+U simulations of intrinsic and extrinsic defects in TiO_2 using the OMR method. The first-principles approach was similar to the semi-empirical approach outlined in Section 4.2.2, by aiming to ensure the accuracy and numerical stability of DFT+U simulations of stoichiometric and defective TMOs and REOs using generalised symbolic regression and classification, as illustrated in Figure 4.2.

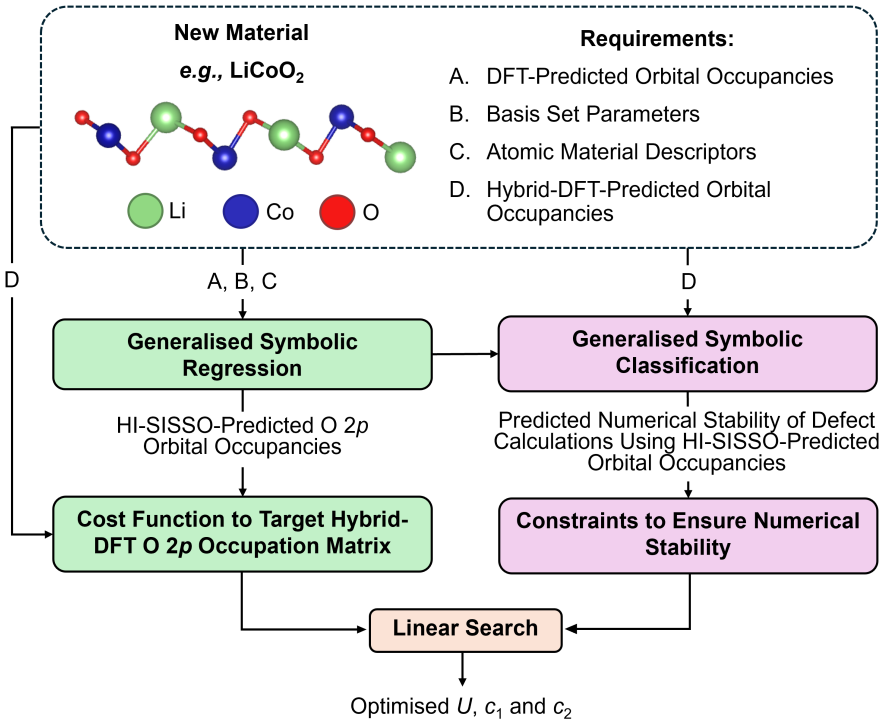


Figure 4.2: First-principles approach for optimising the Hubbard U value and the projector. Generalised symbolic regression is used to target the hybrid-DFT-predicted O 2p occupation matrix. Generalised symbolic classification is used to determine constraints on the Hubbard parameter space to ensure numerically stable point defect calculations.

To generalise SR across different materials, the primary feature space was expanded beyond the Hubbard parameters U, c_1 and c_2, to include data that can either be determined from a single reference DFT calculation or is widely available in the literature. The expanded primary feature space included (1) basis set parameters for the correlated subshell subject to the Hubbard correction, such as the type (Z_{Type}, *i.e.*, hydrogenic or ionic) and effective core charge (Z_{val}) of the auxiliary basis function, (2) DFT-predicted metal d or f and O $2p$ orbital occupancies averaged over all atoms in the unit cell and (3) atomic material descriptors, including the metal atom electronegativity (χ) and atomic radius (r), as well as the outer subshell type (S, encoding d or f subshells), principal quantum number (Q) and number of electrons in the ion (e^{Ion}). Both Z_{Type} and S were label-encoded as integers to enable their use in SR. The expanded set of 20 primary features makes the SISSO method computationally intractable due to the attempted exhaustive search of the secondary feature space, which scales exponentially with the number of primary features, therefore, a two-step hierarchical SISSO (HI-SISSO) [21] approach was adopted where the output from a first step using SISSO, with 13/20 of all primary features (Hubbard parameters and DFT-predicted orbital occupancies), was used as an input for a second step using HI-SISSO where the remaining primary features are included as inputs.

HI-SISSO was used to fit ten empirical correlations to estimate DFT+U-predicted orbital occupancies across the full range of magnetic quantum numbers (m), with seven correlations for m= -3 to +3 (grouping together d and f orbitals) and three correlations for m= -1 to 1 for O $2p$ orbitals. The training set used, contained 197 sets of DFT+U-calculated orbital occupancies from optimised unit cells of anatase and rutile TiO_2 (46%), CeO_2 (13%), Cu_2O (11%), Y_2O_3 (11%), ZrO_2 (11%), WO_3 (7%) and MoO_3 (1%), where the percentages correspond to the proportion of each material in the dataset. $LiCoO_2$ and $LiFePO_4$ were used as blind test cases with no training data. As illustrated by the parity plots in Figure 4.3 for O $2p$ orbitals, the HI-SISSO approach (including basis set parameters and atomic descriptors) improved the predictive accuracy of all three correlations compared to the single step SISSO approach, which is equivalent to Δ-machine learning with respect to DFT-predicted orbital occupancies. [38]

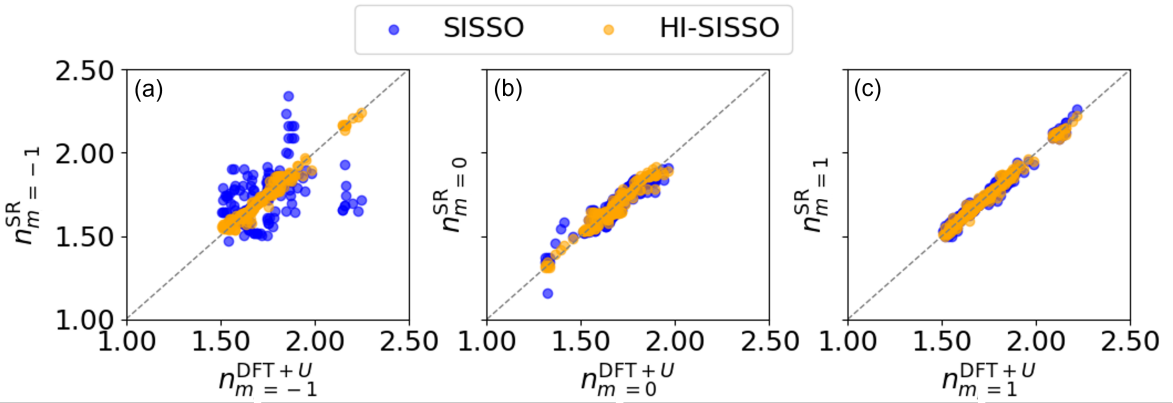


Figure 4.3: Parity plots for the DFT+U- and SR-predicted O $2p$ orbital occupancies for (a) n_m = -1, (b) n_m = 0 and (c) n_m = 1. Blue markers show the predictions using a single step SISSO fitting using the primary features U, c_1, c_2 and all DFT predicted metal d or f and O $2p$ orbital occupancies. Orange markers show the predictions after a second HI-SISSO fitting, where the outputs from the first step are included within a new set of primary features including S, Z_{Type}, Z_{val}, Q, χ, e^{Ion} and r.

The HI-SISSO-predicted O 2*p* occupancies were then used to construct a cost function for optimising Hubbard *U* values and projectors from first-principles, J^{FP}, by targeting O 2*p* orbital occupancies calculated using hybrid-DFT, as outlined in Section 4.3.3. All SISSO and HI-SISSO correlations were fitted using empirical correlations of two or three terms, using the constants in Table 4.5, which are used to predict the occupancies of the x_i orbitals, corresponding to to m = (i-3) for d orbitals and m = (i-4) for f orbitals, and the p_i orbitals, corresponding to m = (i-2) for O 2*p* orbitals.

Table 4.5: Constants for all SISSO and HI-SISSO correlations used in the first-principles approach for Hubbard projector optimisation, with corresponding model accuracy metrics including the Pearson's coefficient of determination (R^2) and the root mean squared error (RMSE). All constants are unitless to ensure dimensional consistency with the unitless target properties.

Subshell	x_i	Approach	a_0	a_1	a_2	a_3	R^2	RMSE
d or f	1	SISSO	-0.0195124	0.0512845	1.7850342	N/A	0.99586	0.03775
d or f	1	HI-SISSO	0.0081740	0.0078837	-0.0000002	-161.1072667	0.99884	0.01998
d or f	2	SISSO	-0.0744573	0.0221677	-0.0000306	1.0795195	0.99804	0.02572
d or f	2	HI-SISSO	0.0034645	-0.0363467	-0.0163977	2.0072180	0.99874	0.02057
d or f	3	SISSO	0.3208064	0.0000259	0.0088471	-0.3785612	0.99734	0.03114
d or f	3	HI-SISSO	0.0049139	0.0274422	-0.0012958	12.4632588	0.99851	0.02329
d or f	4	SISSO	-0.0390992	0.3477937	0.0007987	1.7937530	0.99749	0.02931
d or f	4	HI-SISSO	0.0188731	-0.0115786	0.0625755	151.4973711	0.99856	0.02226
d or f	5	SISSO	0.3430127	0.0077543	0.0062046	-0.3795066	0.99803	0.02682
d or f	5	HI-SISSO	0.0274316	-0.0001409	-0.0315249	197.0158617	0.99915	0.01767
d or f	6	SISSO	-0.0000009	0.0193878	-0.0237643	0.1396925	0.99951	0.00027
d or f	6	HI-SISSO	-0.0000402	0.0000143	-0.0001629	0.0000241	0.99972	0.00021
d or f	7	SISSO	-0.0000013	0.1736143	0.7761145	1.7892233	0.99986	0.00026
d or f	7	HI-SISSO	-0.0003830	0.0004469	-0.0006041	-0.0005738	0.99993	0.00019
Subshell	p_i	Approach	a_0	a_1	a_2	a_3	R^2	RMSE
O 2*p*	1	SISSO	0.9079618	-0.0749980	-0.0121296	0.1284832	0.90194	0.04482
O 2*p*	1	HI-SISSO	1.7699613	0.0001488	-0.0133692	-0.1984506	0.96764	0.02575
O 2*p*	2	SISSO	1.8894035	0.1009336	0.0193631	-1.3224441	0.88321	0.04323
O 2*p*	2	HI-SISSO	1.7700081	0.0252095	-0.0065591	-1.2267782	0.94510	0.02964
O 2*p*	3	SISSO	1.2184210	0.0936608	0.0002472	0.1183200	0.97732	0.02464
O 2*p*	3	HI-SISSO	1.7716178	-0.0139624	-0.0094436	-5.8280428	0.98742	0.01835

The general form of the empirical correlations is a linear combination of non-linear terms, *e.g.*, F_1, F_2 and F_3 using the constants a_0, a_1, a_2 and a_3 for a three term correlation:

$$\text{Target} = a_0 + a_1 \times F_1 + a_2 \times F_2 + a_3 \times F_3 \tag{4.14}$$

where the terms F_1, F_2 and F_3, used in the SISSO and HI-SISSO correlations are listed in Table 4.6. The primary features are the Hubbard parameters (`U` in eV, `c1` and `c2`), DFT- and SISSO-predicted orbital occupancies (`DFT_xi` and `DFT_pi`, and `SISSO_xi` and `SISSO_pi`, for metal *d* or *f* and O 2*p* orbitals, respectively), basis set parameters (`Auxiliary_type` *i.e.*, hydrogenic or ionic and `Auxiliary_zval` in *e*) and atomic descriptors of the metal species in the oxide, including the electronegativity (χ), atomic radius (`r` in pm), outer subshell type (i.e., *d* or *f*, `S`), principal quantum number (`Q`) and number

of outer subshell electrons in the ion (`Ion_OE`). Both `Auxiliary_type` and `S` were label-encoded as integers to ensure compatibility with SISSO. The constants $\alpha = 1$ eV^{-1}, $\beta = 1$ pm^{-1} and $\gamma = 1$ e^{-1} ensure dimensional consistency between all unitless constants in Table 4.5, terms and target properties.

Table 4.6: Summary of the non-linear terms F_1, F_2, and F_3 used in the SISSO and HI-SISSO correlations for each orbital magnetic quantum number m in the metal d or f subshells. All terms are made unitless using the constants $\alpha = 1$ eV^{-1}, $\beta = 1$ pm^{-1} and $\gamma = 1$ e^{-1}, ensuring dimensional consistency with the unitless target properties. The SISSO-predicted orbital occupancies (`SISSO_xi` for metal d or f and O 2*p* orbitals) are used as inputs in the HI-SISSO correlations.

Subshell	x_i	Approach	F_i	Expression
d or f	1	SISSO	F_1	(αU / ln(DFT_x1)) * (cbrt(c2) + ln(DFT_p1))
d or f	1	SISSO	F_2	((DFT_x1 / DFT_x3) * DFT_x5) * (sin(c1)^3)
d or f	1	HI-SISSO	F_1	(ln(χ) * (γAuxiliary_zval - Ion_OE)) - sin(γAuxiliary_zval^3)
d or f	1	HI-SISSO	F_2	((SISSO_x1 * γAuxiliary_zval)^3) / (cbrt(βr) - (SISSO_x1^2))
d or f	1	HI-SISSO	F_3	sin(exp(Auxiliary_type)) / (sin(βr) - (βr / SISSO_x1))
d or f	2	SISSO	F_1	((DFT_x3 + αU) - (DFT_p2^3)) / ln(DFT_x4)
d or f	2	SISSO	F_2	(((αU)^2)^2) / (sin(c2) + (DFT_p1 - c1))
d or f	2	SISSO	F_3	((DFT_x7 + DFT_x1) + DFT_x7) * (c1^2)
d or f	2	HI-SISSO	F_1	sin((S - Q) * (SISSO_x2^3))
d or f	2	HI-SISSO	F_2	(cbrt(γAuxiliary_zval) - Auxiliary_type) / (sin(SISSO_x2) - ln(SISSO_x2))
d or f	2	HI-SISSO	F_3	sqrt(exp(Q)) * ((SISSO_x2 / Q) / Q)
d or f	3	SISSO	F_1	(((αU)^2)^2) / ((DFT_p2 + c2) - c1)
d or f	3	SISSO	F_2	((ln(DFT_x4) + αU)) / (ln(DFT_x3) + sin(c1)))
d or f	3	SISSO	F_3	sqrt(exp(c2)) - ((DFT_x5 * c1) * exp(c1))
d or f	3	HI-SISSO	F_1	sin((SISSO_x3^3)) + ((Auxiliary_type^2) / (SISSO_x3 * βr))
d or f	3	HI-SISSO	F_2	sin(Q) / (ln(Q) + (SISSO_x3 - χ))
d or f	3	HI-SISSO	F_3	sin(exp(Auxiliary_type)) * ((SISSO_x3 / S) / sqrt(S))
d or f	4	SISSO	F_1	((c2^2)^2) / ((DFT_x5^2) - (c1 + (αU)))
d or f	4	SISSO	F_2	a1 * ((αU)^3) / ((DFT_x5 * c2) + ln(DFT_x4))
d or f	4	SISSO	F_3	((DFT_x6 + DFT_x1) * sin(DFT_p1)) * (sin(c1)^3)
d or f	4	HI-SISSO	F_1	sin(χ / SISSO_x4) / ((χ^3) - (1.0 / SISSO_x4))
d or f	4	HI-SISSO	F_2	sin(γAuxiliary_zval + S) * sin(SISSO_x4 * Ion_OE)
d or f	4	HI-SISSO	F_3	sin(exp(Auxiliary_type)) / ((βr / SISSO_x4) - (Auxiliary_type^2))
d or f	5	SISSO	F_1	((c2^3) * (DFT_x2^2)) / (exp(c1) - (αU))
d or f	5	SISSO	F_2	(exp(c2) * (αU)) / (ln(DFT_x3) + sin(c1))
d or f	5	SISSO	F_3	cbrt(exp(c2)) - ((DFT_x5 * c1) * exp(c1))
d or f	5	HI-SISSO	F_1	(exp(Auxiliary_type) - γAuxiliary_zval) / (ln(χ) - sin(SISSO_x5))
d or f	5	HI-SISSO	F_2	sin(sqrt(Q) * (SISSO_x5 * S))
d or f	5	HI-SISSO	F_3	sin(exp(Auxiliary_type)) / ((βr / SISSO_x5) * cbrt(χ))
d or f	6	SISSO	F_1	(((c2 / DFT_x2) + sin(αU))) * DFT_x7
d or f	6	SISSO	F_2	((c1^3)^3) * (DFT_x6 / (DFT_x2 - αU))
d or f	6	SISSO	F_3	(((c1 / DFT_x2) - cbrt(αU))) * DFT_x7
d or f	6	HI-SISSO	F_1	1.0 / ((SISSO_x6 * βr) + (SISSO_x6 - S))
d or f	6	HI-SISSO	F_2	1.0 / ((SISSO_x6 * βr) - (Auxiliary_type + Q))
d or f	6	HI-SISSO	F_3	((βr^2) * sin(SISSO_x6)) + sin(βr / SISSO_x6)
d or f	7	SISSO	F_1	sin(DFT_p3 * αU) * (ln(c1) * DFT_x7)
d or f	7	SISSO	F_2	(DFT_x6 / exp(c2)) * ((c2^3) + ln(c1))
d or f	7	SISSO	F_3	(DFT_x6 * c1) / exp(DFT_x1 * αU)
d or f	7	HI-SISSO	F_1	sin(exp(Q) * (SISSO_x7 * βr))
d or f	7	HI-SISSO	F_2	sin((S * βr) * (SISSO_x7 * Q))
d or f	7	HI-SISSO	F_3	sin(1.0 / SISSO_x7) - ((SISSO_x7 * Q) * (γAuxiliary_zval^3))

Table 4.7: Summary of the non-linear terms F_1, F_2, and F_3 used in the SISSO and HI-SISSO correlations for each orbital magnetic quantum number m of the O $2p$ subshell. All terms are made unitless using the constants α = 1 eV^{-1}, β = 1 pm^{-1} and γ = 1 e^{-1}, ensuring dimensional consistency with the unitless target properties. The SISSO-predicted orbital occupancies (`SISSO_pi`) are used as inputs in the HI-SISSO correlations.

S	p_i	Approach	F_i	Expression
O $2p$	1	SISSO	F_1	`((DFT_x2 + c2) / sqrt(DFT_x5)) * sin(1.0 / DFT_x5)`
O $2p$	1	SISSO	F_2	`sin(αU / DFT_x1) / ((DFT_p2^3) - DFT_p3)`
O $2p$	1	SISSO	F_3	`(sqrt(αU) + exp(DFT_p3)) + (c2 * αU)`
O $2p$	1	HI-SISSO	F_1	`exp(S / SISSO_p1) / (ln(SISSO_p1) - sin(Auxiliary_type))`
O $2p$	1	HI-SISSO	F_2	`sin(SISSO_p1^3) / (sin(SISSO_p1) - (SISSO_p1 - Auxiliary_type))`
O $2p$	1	HI-SISSO	F_3	`sin(SISSO_p1^2) * (ln(βr) + (SISSO_p1 - S))`
O $2p$	2	SISSO	F_1	`sin(DFT_p1 / DFT_x5) * sin(DFT_x4 * αU)`
O $2p$	2	SISSO	F_2	`sin(DFT_p1^2) / (cbrt(c2) + (DFT_x2 - c2))`
O $2p$	2	SISSO	F_3	`sin(cbrt(DFT_x5)) / ((DFT_p2^3) - (DFT_p3 - αU))`
O $2p$	2	HI-SISSO	F_1	`sin((Q - βr) / exp(SISSO_p2))`
O $2p$	2	HI-SISSO	F_2	`sin(exp(Q)) / ((SISSO_p2^2) - (SISSO_p2 + Auxiliary_type))`
O $2p$	2	HI-SISSO	F_3	`sin(SISSO_p2^2) / ((S - χ) + sin(βr))`
O $2p$	3	SISSO	F_1	`((c1 * αU) / cbrt(DFT_p1)) - exp(cbrt(αU))`
O $2p$	3	SISSO	F_2	`((c2^2) - (DFT_x7 * αU)) / (DFT_x1^3)`
O $2p$	3	SISSO	F_3	`(sqrt(αU) + (DFT_p3^3)) + (c2 * αU)`
O $2p$	3	HI-SISSO	F_1	`sin((Q^2) * (SISSO_p3 / χ))`
O $2p$	3	HI-SISSO	F_2	`sin(γAuxiliary_zval * Q) * ((χ^3) - (SISSO_p3^2))`
O $2p$	3	HI-SISSO	F_3	`(ln(SISSO_p3) - (1.0 / SISSO_p3)) / (sin(Ion_OE) - S)`

The HI-SISSO-predicted O $2p$ occupancies were then used to construct a cost function for optimising Hubbard U values and projectors from first-principles, J^{FP}, by targeting O $2p$ orbital occupancies calculated using hybrid-DFT, as outlined in Section 4.3.3.

$$J^{\mathrm{FP}} = \left\| \frac{100 \times \left(\boldsymbol{n}_m^{\text{O } 2p \text{ DFT+}U} - \boldsymbol{n}_m^{\text{O } 2p \text{ PBE0}} \right)}{\boldsymbol{n}_m^{\text{O } 2p \text{ PBE0}}} \right\|, \quad \boldsymbol{n}_m = (n_{-1}, n_0, n_1) \tag{4.15}$$

Symbolic Classification

The observed numerical instability of self-consistent DFT+U simulations of point defects in TiO_2, including excessive polaron localisation that causes calculations to terminate and charge sloshing preventing SCF convergence, was observed when modelling point defects across the broader range of TMOs and REOs in Table 2.1. Therefore, the use of SVMs for classifying the stability of point defect calculations in Section 4.2.2 was generalised across materials. Generalised classification was achieved by defining new primary features to account for the different degrees of covalent character across the materials in Table 2.1, which is key for the generalisation of predictive models across complex oxides.[39] The primary features for generalised classification were U, J^{FP}, the average error in the DFT+U-predicted metal d or f orbital occupancies relative to hybrid-DFT (E^{Metal}) and the ratio of the traces of the metal d or f and O $2p$ occupation matrices predicted using hybrid-DFT (R^{Hybrid}), as outlined in Section 4.3.3.

$$E^{\text{Metal}} = \left\| \frac{100 \times \left(\boldsymbol{n}_m^{\text{Metal } d \text{ or } f \text{ DFT+}U} - \boldsymbol{n}_m^{\text{Metal } d \text{ or } f \text{ PBE0}} \right)}{\boldsymbol{n}_m^{\text{Metal } d \text{ or } f \text{ PBE0}}} \right\|, \quad \boldsymbol{n}_m = (n_{-3}, n_{-2}, n_{-1}, n_0, n_1, n_2, n_3) \tag{4.16}$$

$$R^{\text{Hybrid}} = \frac{\text{Tr}[\mathbf{n}(\text{Metal } d \text{ or } f)]^{\text{PBE0}}}{\text{Tr}[\mathbf{n}(\text{O } 2p)]^{\text{PBE0}}} \tag{4.17}$$

Generalised constraints on the Hubbard parameter space were determined using two linear SVMs to determine the equations of the boundaries S_4 and S_5 that separated regions in the primary feature space leading to the termination and non-convergence of bulk oxygen vacancy calculations using the OMR method. The training set used, consisted of 86 DFT+U simulations across anatase and rutile TiO_2 (76%), CeO_2 (6%), ZrO_2 (6%), MoO_3 (6%), WO_3 (5%) and Cu_2O (2%), where the percentages correspond to the proportion of each material in the dataset. To reduce the number of misclassified data points associated with S_4 and S_5, classification was performed symbolically using the SISSO algorithm to recursively combine the primary features using the same algebraic operators as in SR, but with the objective of minimising the number of data points in the overlap region of a two-dimensional convex hull. [29] The secondary features generated from SISSO were then used as inputs for two linear SVMs, which identified simple boundaries to perform binary classifications in a highly non-linear feature space. The SVM boundaries S_4 and S_5, are defined as:

$$S_4 = \text{A}_4 \times \left(\frac{R^{\text{Hybrid}}}{\alpha U} + \sin\left(E^{\text{Metal}}\right) \right) + \text{B}_4 \times \left(\frac{R^{\text{Hybrid}} + J_{\text{FP}}}{\ln\left(R^{\text{Hybrid}}\right)} \right) + \text{C}_4 \tag{4.18}$$

$$S_5 = \text{A}_5 \times \sin\left(\exp(J_{\text{FP}})\right) + \text{B}_5 \times \left(\sin(J_{\text{FP}}) + \frac{\alpha U}{R_{\text{Hybrid}}} \right) + \text{C}_5 \tag{4.19}$$

where the constant $\alpha = 1$ eV^{-1} is introduced to ensure dimensional consistency and the associated constants and accuracy metrics are listed Table 4.8. The condition for numerically stable bulk oxygen vacancy calculations is therefore:

$$S_4 \geq 0 \quad \wedge \quad S_5 \geq 0 \tag{4.20}$$

Table 4.8: Constants for all linear SVM boundaries S_4 and S_5 used in the first-principles approach for Hubbard projector optimisation, as well as the corresponding proportion of misclassified data points (sum of false positive and negative classification predictions) when evaluated on the respective training sets. All constants are unitless

S_i	A_i	B_i	C_i	Misclassified (%)
4	0.86	0.97	9.76	8.06
5	-1.04	-0.75	5.18	9.26

One-Shot Optimisation of Hubbard Parameters

Optimisation of U, c_1 and c_2 from first-principles was performed by a linear search of the landscape of the HI-SISSO-predicted cost function ($J^{\mathrm{FP}}_{\mathrm{Predicted}}$) for each material, over the range of U between 0 eV and 5 eV, c_1 between 0.5 and 1 and c_2 between -0.6 and 0, with each feature split into 50 intervals. The output of each linear search was a family of candidate solutions for a material, with several combinations of U, c_1 and c_2 optimised to minimise $J^{\mathrm{FP}}_{\mathrm{Predicted}}$. The accuracy of the HI-SISSO correlations to predict O $2p$ orbital occupancies was evaluated by validating $J^{\mathrm{FP}}_{\mathrm{Predicted}}$ against the corresponding DFT+U-predicted cost function ($J^{\mathrm{FP}}_{\mathrm{Validated}}$) for up to ten different combinations of Hubbard U values and projectors per material (94 in total), before calculating the mean absolute error (MAE), which is averaged across all tested Hubbard parameters (N) for each material:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|J^{\mathrm{FP}}_{\mathrm{Predicted},i} - J^{\mathrm{FP}}_{\mathrm{Validated},i}\right| \tag{4.21}$$

The relationship between the accuracy of the one-shot approach for minimising $J^{\mathrm{FP}}_{\mathrm{Predicted}}$ and the training set size for each material was investigated using the Mahalanobis distance (D_{M}) [40] to quantify the distance of the primary feature vector for each material ($\boldsymbol{x}$), in a reduced two dimensional feature space determined using principal component analysis (PCA) with the *Scikit-learn* Python library, [33] from the mean vector of the training data ($\boldsymbol{\kappa}$), using the inverse covariance matrix of the training data ($\boldsymbol{C}^{-1}$):

$$D_{\mathrm{M}} = \sqrt{(\boldsymbol{x}-\boldsymbol{\kappa})^T\boldsymbol{C}^{-1}(\boldsymbol{x}-\boldsymbol{\kappa})} \tag{4.22}$$

An integrated one-shot approach for optimising Hubbard U values and projectors from first-principles was tested for computing the Co $3d$ Hubbard U value and projector for the simulation of stoichiometric, Mg-doped and oxygen defective $LiCoO_2$ (*i.e.*, $LiCo_{1-x}Mg_xO_{2-x}$) which is unseen by any of the trained regression or classification models. Here, the three HI-SISSO correlations to estimate the DFT+U-predicted O $2p$ orbital occupancies (for m= -1, 0 and 1) were used to screen the landscape of $J^{\mathrm{FP}}_{\mathrm{Predicted}}$ for stoichiometric $LiCoO_2$, before all Hubbard parameters that violate the generalised constraints S_4 and S_5, which are evaluated using all ten HI-SISSO correlations, are removed from the landscape. The remaining Hubbard parameters were reduced to a smaller set of candidates with K-means clustering, using the *Scikit-learn* Python library, [33] which uses unsupervised learning to partition the landscape of $J^{\mathrm{FP}}_{\mathrm{Predicted}}$ into smaller clusters by minimising intra-cluster variance. The centroids of these clusters were then used as screened Hubbard parameters for the simulation of stoichiometric and defective $LiCoO_2$, using the OMR method.

4.3 Results and Discussion

4.3.1 Projector Sensitivities in DFT+U Simulations

Stoichiometric Oxides

DFT+U is known to be non-trivial when simulating the ground state character (*i.e.,* metallic, semiconducting or insulating) of TMOs and REOs. For example, in a planewave basis, material-dependent

transitions in the DFT+U-predicted ground state from metallic to insulating can occur upon increasing the Hubbard U value, which can restore the experimentally observed electronic structures of Mott insulators such as NiO and Ce_2O_3 [41–44], and point defects in CeO_2 surfaces. [43] Furthermore, erroneous changes in the DFT+U-predicted hybridisation between metal d and O $2p$ orbitals can drive the predictions of ground state crystal structures and magnetic properties away from experimental observations, as is reported for $BaTiO_3$ and layered $AMoO_2$ (A = Li, Na, K). [45, 46] The existence of metastable states in the DFT+U potential energy surface, near integer orbital occupations, can also result in erroneous trapping in local energy minima using local optimisation algorithms, resulting in incorrect ground state predictions for point defects in actinide oxides such as UO_2. [26]

In a NAO framework, similar observations are made when modelling stoichiometric REOs using the default atomic Hubbard projector. For example, applying a Hubbard correction to Ce 4f electrons in stoichiometric CeO_2, using the atomic Ce 4f Hubbard projector, results in an insulator-metal transition (IMT) in the predicted ground state. Upon increasing the Ce 4f Hubbard U value, there is a monotonic change in the DFT+U-predicted E_{bg} in Figure 4.4(a) and ΔE_{Form} in Figure 4.4(b) between U values of 0 eV to 9 eV. Beyond U= 9 eV, which would be required if adopting the standard approach of benchmarking against the experimental E_{bg} of 3.2 eV, [47] there is a sudden deviation in these trends and DFT+U predicts strong electron localisation in the Ce 4f m = -2 orbital, corresponding to metallic behaviour with no band gap.

To trace the root cause of the observed IMT at U= 9.5 eV, constrained DFT+U calculations were performed using the ground state Ce 4f occupation matrix calculated using U= 9.5 eV, but with the m = -2 and m = -3 orbital occupancies systematically varied and all other occupancies fixed. As illustrated in Figure 4.4(c), there is a metallic global energy minimum and an insulating low-lying metastable state in the potential energy surface, therefore the observed IMT is caused by the small energy differences between these energy minima, which decreases as the Hubbard U value increases. For U = 9.5 eV, self-consistent determination of the Ce 4f occupation matrix leads to a metallic solution, irrespective of the initial occupation matrix, therefore, we conclude that there are no other insulating solutions with lower energy than the metallic solution, and that this represents an IMT in the potential energy surface. Beyond U= 11 eV, the insulating ground state character is seemingly restored in Figures 4.4(a) and (b), however these were found to be metastable states, which was confirmed using OMR with a modified initial Ce 4f occupation matrix that enabled convergence to the true metallic ground state that exists at all U values beyond the IMT at U= 9.5 eV. The IMT was also found to vary with the definition of the Hubbard projector, as shown in Figure 4.4(d), where the more localised projector in green brings the IMT forward to U > 5.5 eV and the more diffuse projector defers the IMT beyond U= 12 eV. The results exemplify the importance of developing new workflows for parameterising the Hubbard U value and projector to enable the accurate simulation of insulating ground states of REOs like CeO_2, whilst avoiding erroneous IMTs and metastable states; this goal cannot be achieved using the standard approach of semi-empirical benchmarking of the Hubbard U value to reproduce the experimental E_{bg}.

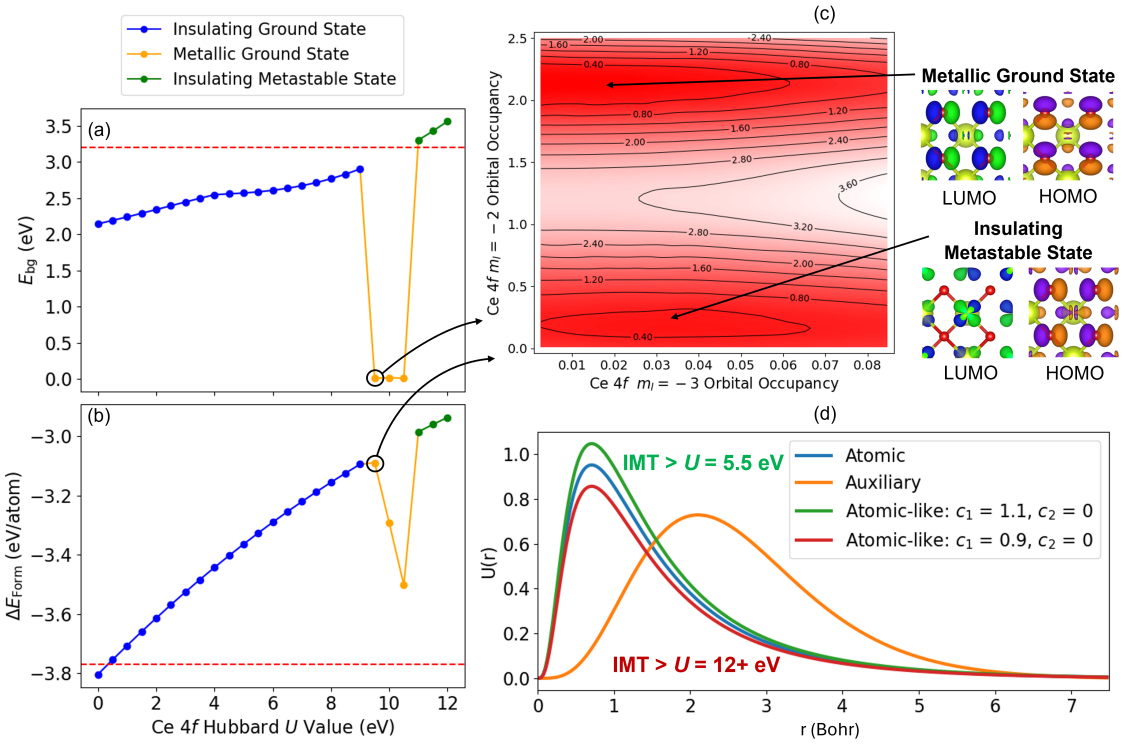


Figure 4.4: Overview of errors when modelling stoichiometric CeO_2 using the default Ce $4f$ atomic Hubbard projector, including the variation in the DFT+U-predicted (a) band gap and (b) formation energy with respect to the Ce $4f$ Hubbard U value, relative to experimental references denoted by the red dashed lines. [47, 48] Blue markers correspond to insulating ground states, yellow markers correspond to metallic ground states and green markers correspond to insulating metastable states. (c) Contour plot of the constrained DFT+U-predicted total energy relative to the ground state energy at U= 9.5 eV, after constraining the m = -2 and m = -3 orbital occupancies. The two regions in dark red correspond to global and low-lying local minima in the potential energy surface with respect to Ce $4f$ orbital occupancies. The metallic global minimum is 0.273 eV more stable than the insulating local minimum. (d) The radial functions corresponding to the atomic Ce $4f$ (blue) and hydrogenic auxiliary (orange) basis functions available for constructing a modified atomic-like Hubbard projector. The green and red radial functions correspond to modified projectors that do not include any contribution from the hydrogenic auxiliary function (*i.e.*, c_2 = 0) and are noted with the corresponding shift of the observed IMT.

Defective Oxides

Whilst DFT+U simulations using semi-empirically-derived Hubbard U values and the default atomic Hubbard projector can accurately model point defects in TMOs such as $Li_4Ti_5O_{12}$, [49, 50] the numerical stability of point defect calculations is generally highly sensitive to the choice of the Hubbard U value, as illustrated in Figure 4.5 for anatase TiO_2. Upon increasing the Ti $3d$ Hubbard U value, there is a monotonic change in the DFT+U-predicted E_{bg} in Figure 4.5(a) and ΔE_{Form} in Figure 4.5(b) of anatase TiO_2. An appropriate Hubbard U value could be chosen by considering the compromise in accuracy of these properties versus the experimental references denoted by the red dashed lines in each plot; however, the numerical stability of DFT+U simulations of a bulk oxygen vacancy in anatase TiO_2 were found to vary strongly with both the Hubbard U value and projector.

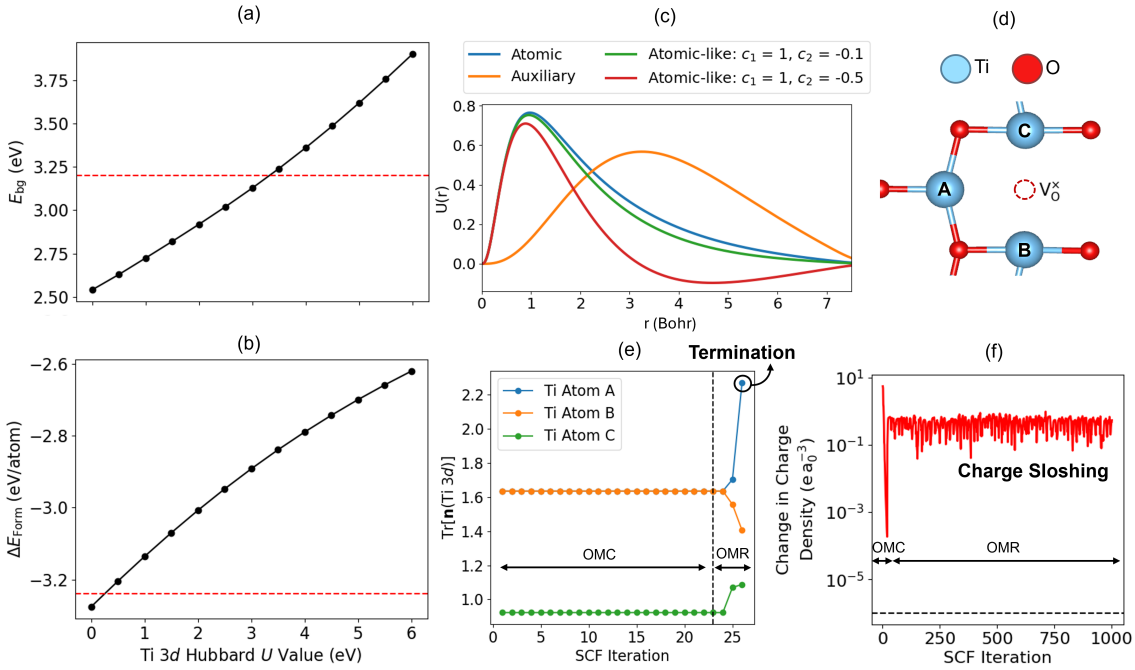


Figure 4.5: Overview of errors when modelling stoichiometric and defective TiO_2, including the variation of the DFT+U-predicted (a) band gap and (b) formation energy with respect to the Ti 3d Hubbard U value, using the default atomic Ti 3d Hubbard projector, relative to experimental references denoted by the red dashed lines. [30, 51] (c) The radial functions corresponding to the atomic Ti 3d (blue) and hydrogenic auxiliary (orange) basis functions available for constructing a modified atomic-like Hubbard projector. The green and red radial functions correspond to modified projectors that incorporate a contribution from hydrogenic auxiliary function given by the linear expansion coefficient c_2. (d) The nearest neighbour Ti atoms surrounding a bulk oxygen vacancy in anatase TiO_2. (e) The evolution of $\mathrm{Tr}[\mathbf{n}(\mathrm{Ti}\ 3d)]$ for Ti atoms A, B and C in (d) during an oxygen vacancy calculation using U= 3 eV, c_1= 1, c_2= -0.1, which leads to calculation termination due to excessive polaron localisation at atom A, after 3 SCF iterations of OMR (which begins after 23 SCF iterations of OMC). (f) The evolution of the change in charge density during SCF optimisation, during the 1st geometry optimisation step for an oxygen vacancy calculation using U= 3 eV, c_1= 1, c_2= -0.5, which does not converge to the convergence criteria of 1×10^{-6} e a_0^{-3}, denoted by the black dashed line, due to charge sloshing.

For example, Figure 4.5(c) shows the radial functions of two modified atomic-like Ti 3d Hubbard projectors (green and red functions), that are examples of a series of systematically tested Hubbard projectors for the simulation of a bulk oxygen vacancy using the OMR method. In each simulation, two Ti^{3+} polarons were initialised at Ti atoms A and B in Figure 4.5(d), which are nearest neighbours relative to the oxygen vacancy. With some combinations of Hubbard U values and projectors, full geometry optimisation successfully completed, whilst in other cases, the simulations were terminated within 2 to 5 iterations of self-consistent optimisation of the system occupation matrices, due to excessive polaron localistion at Ti atom A resulting in the predicted $3d_{z^2}$ orbital occupancy increasing beyond 4 (which is the condition for termination) in Figure 4.5(e). With other combinations of Hubbard U values and projectors, the OMR calculations did not converge due to strong oscillations in the charge density (Figure 4.5(f)), which is known as charge sloshing between partially filled, degenerate orbitals and is often associated with metallic systems. [52]

Tracing Numerical Instability to the Hubbard Projector

The termination of defect calculations in Figure 4.5(e) was observed for Hubbard U values > 1 eV using the default atomic Ti 3*d* Hubbard projector and occurred irrespective of initialising Ti^{3+} polarons further away from the oxygen vacancy as well as tuning available parameters such as SCF mixing parameters and initial occupation matrices. Initialising Ti^{3+} polarons in different Ti 3*d* orbitals (other than the $3d_{z^2}$ orbital) and performing local symmetry breaking *via* targeted bond distortions were also tested; however, these did not alleviate the termination of defect calculations in a NAO framework, despite their reported success for aiding SCF covergence in planewave codes. [25, 53, 54] The SCF non-convergence due to charge sloshing in Figure 4.5(f) was also very difficult to alleviate with common strategies. After extensive testing, no robust strategy to mitigate this sloshing was found, despite tuning available parameters including the basis set size, initial geometries and occupation numbers, SCF mixing parameters, Gaussian broadening parameters, Kerker preconditioning, [55] the OMR pre-convergence criteria, **k**-point spacing and using wavefunction restarts. Whilst the charge sloshing in Figure 4.5(f) appears insensitive to the aforementioned strategies to improve SCF convergence, other cases of SCF non-convergence such as plateauing of the electron density are more straightforward to address, as observed for Mn-doped rutile TiO_2 in Section 4.3.2, which requires an increased Pulay mixing history beyond the default value. The issue of charge sloshing therefore appears to be specific to DFT+U in the NAO framework, although planewave implementations are not without their own projector-related issues, as highlighted by Warda *et al.*, who discuss the challenge of spurious Hubbard forces and the resulting inaccuracies in the predicted phase stabilities of AUO_4 (A = Ni, Mg, Co, Mn) compounds when using atomic Hubbard projectors. [56]

We note three approaches that are observed to overcome the numerical instabilities when simulating point defects using DFT+U in a NAO framework. Kick *et al.* overcame SCF non-convergence when simulating Ti^{3+} polarons at oxygen vacancies in a rutile (110) TiO_2 surface by increasing the effective core charge (Z_{val}) of the *Tier 1* hydrogenic auxiliary basis function from 2.7 e to 4.4 e, before using an atomic Ti 3*d* Hubbard projector to define the basis for the Hubbard correction. [24] Using very small Hubbard U values, or employing constrained DFT+U using the OMC method, can also overcome the observed numerical instability; however, Chapter 3 shows that this comes at the expense of accuracy when modelling polarons in Nb- and W-doped bulk anatase and rutile TiO_2. [2] For these systems, self-consistent resolution of the system occupation matrices was achieved using a modified Ti 3*d* Hubbard projector, defined as a linear combination of the atomic and hydrogenic basis functions in Table 2.1 with the expansion coefficients c_1 = 0.828 and c_2 = -0.561 determined by Jakob and Oberhofer using a first-principles generalised LR-cDFT method in FHI-aims. [57] With a modified Ti 3*d* Hubbard projector, DFT+U successfully simulated the polymorph-dependent formation of Nb^{4+} and W^{5+} polarons as observed in electron paramagnetic resonance (EPR) spectroscopy, which cannot be rationalised by the current planewave DFT+U studies in the literature. [2] It is therefore essential to understand how to define an appropriate Hubbard projector to enable DFT+U simulations with appreciable Hubbard U values without sacrificing self-consistency and remaining robust with respect to the default basis sets. This challenge in investigated herein for TiO_2 before establishing a generalised understanding for a broader range of TMOs and REOs.

4.3.2 Bayesian Optimisation of the Ti 3*d* Hubbard Projector

Method Configuration

The semi-empirical cost function (J^{SE}) defined in Section 4.2.2 favours regions in the Hubbard parameter space that achieve a compromise between modelling localised polarons at point defects and the accurate geometric and electronic structure of stoichiometric anatase TiO_2. To rapidly sample the Hubbard parameter space without requiring multiple DFT+U calculations, the terms in J^{SE} were evaluted using SISSO-computed empirical correlations for the DFT+U-predicted E_{bg}, $\bar{V}_0$, $\mathrm{Tr}[\mathbf{n}(\mathrm{Ti}\ 3d)]$ and $\mathrm{Tr}[\mathbf{n}(\mathrm{O}\ 2p)]$ in terms of the primary features U, c_1 and c_2, as listed in Section 4.2.2. Three constraints for sampling the Hubbard parameter space were also defined to ensure physicality of the model and avoid numerical instability during point defect calculations using OMR, which result in either termination or non-convergence in Figure 4.5. Hubbard parameters that are predicted by the SISSO models to give a negative occupation matrix trace for Ti 3d or O 2p orbitals, or occupation matrix traces that deviate from the respective hybrid-DFT-predicted occupation matrix trace by over 50%, were excluded. The third constraint was applied using two linear SVMs that classify the validity of a bulk oxygen vacancy calculation in terms of the normalised $\overline{U}$, $\overline{\mathrm{Tr}}[\mathbf{n}(\mathrm{Ti}\ 3d)]$ and $\overline{\mathrm{Tr}}[\mathbf{n}(\mathrm{O}\ 2p)]$, as illustrated in Figure 4.6, which shows the boundaries S_1 and S_2 that separate regions in the Hubbard parameter space that lead to unphysical (termination), charge sloshing (preventing SCF convergence) and ground state (stable convergence) calculation outcomes.

Figure 4.6 illustrates how the numerical stability of defect calculations in TiO_2 is sensitive to the DFT+U-predicted covalency, with calculation termination occuring at high covalency, *i.e.*, $\overline{\mathrm{Tr}}[\mathbf{n}(\mathrm{O}\ 2p)] >> \overline{\mathrm{Tr}}[\mathbf{n}(\mathrm{Ti}\ 3d)]$ and charge sloshing occuring at high metallicity, *i.e.*, $\overline{\mathrm{Tr}}[\mathbf{n}(\mathrm{Ti}\ 3d)] >> \overline{\mathrm{Tr}}[\mathbf{n}(\mathrm{O}\ 2p)]$. To ensure the successful convergence of point defect calculations whilst mitigating the effects of the changing DFT+U-predicted covalency, the equations of the decision boundaries S_1 and S_2 (trained on DFT+U data, with all constants listed in Section 4.2.2) were used as constraints on the sampled Hubbard parameter space, based on the SISSO-computed cost function J^{SE}, with the critical values for S_1 and S_2 chosen based on the convex hull plots in Figure 4.6:

$$S_1 > -9.84 \quad \wedge \quad S_2 > -9.16 \tag{4.23}$$

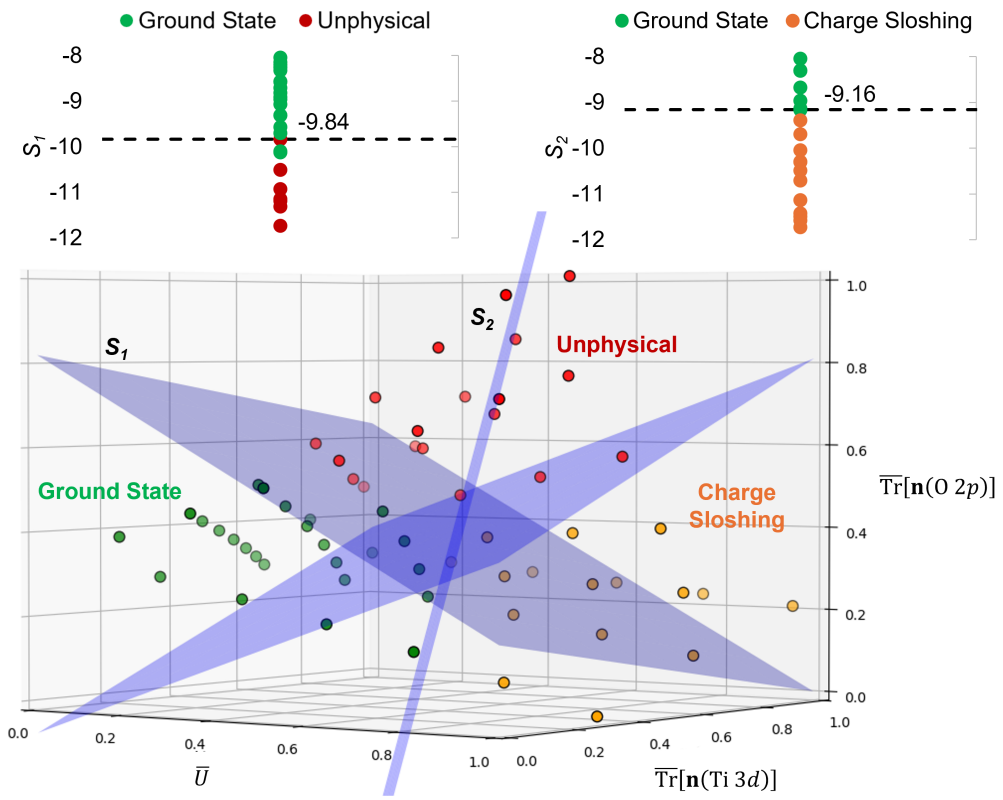


Figure 4.6: Illustration of the linear boundaries used to classify simulations of a bulk oxygen vacancy in anatase TiO_2. The boundaries separate successful convergence (green markers), termination due to an unphysical ground state (red markers) and charge sloshing preventing SCF convergence (orange markers). The convex hull associated with each binary classification S_1 and S_2 is shown to illustrate the basis for constructing the constraint in Equation (4.23).

Bayesian Optimisation and Defect Energy Corrections

BO was used to search the Hubbard parameter space by minimising J^{SE} whilst satisfying the constraints on the SISSO-predicted traces of the Ti $3d$ and O $2p$ occupation matrices, $\mathrm{Tr}[\mathbf{n}(\mathrm{Ti}\ 3d)]^{\mathrm{SISSO}}$ and $\mathrm{Tr}[\mathbf{n}(\mathrm{O}\ 2p)]^{\mathrm{SISSO}}$, respectively, and the SVM-derived constraints S_1 and S_2, as illustrated in Figure 4.7. Figures 4.7(a) and (b) show the results of BO and random sampling, respectively, where BO is able to efficiently optimise U, c_1 and c_2, yielding an almost equivalent set of optimised Hubbard parameters, to those obtained using random sampling, with minimal values of J^{SE}. This is further supported by Figure 4.7(c), which shows the values of J^{SE} corresponding to the sampled Hubbard parameters. The markers in pink show the first 1350 randomly sampled Hubbard parameters, where constraint violations do not inform subsequent sampling, making the approach highly inefficient. The markers in purple correspond to BO, where the 1000 randomly sampled points selected using Latin Hypercube Sampling are used to condition the BO prior distribution and enable efficient optimisation of U, c_1 and c_2. The Hubbard parameters U = 2.749 eV, c_1 = 0.758 and c_2 = -0.354, from the region of lowest J^{SE} in Figures 4.7(a) and (b), were tested for the simulation of both anatase and rutile TiO_2 polymorphs. With the refined atomic-like Ti $3d$ Hubbard projector, oxygen vacancies and the substitutional dopants Nb, W, Co, Mn, Pt, Au and Pd all successfully converged to the ground state, whilst 11 of these 16

calculations failed when using the same Hubbard U value with the default atomic Ti $3d$ Hubbard projector (listed in Section 4.3.2, Table 4.10).

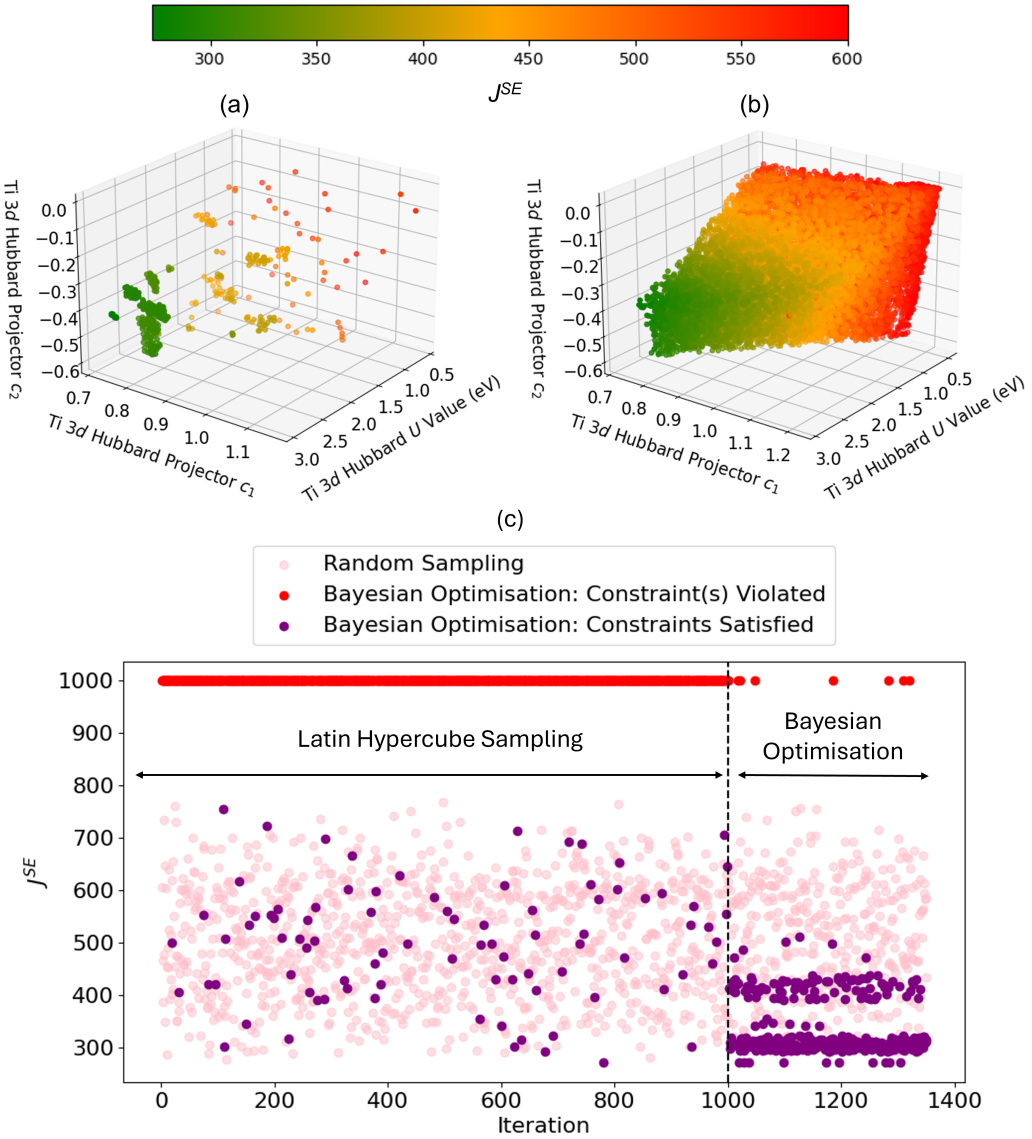


Figure 4.7: The sampled Hubbard parameter space for anatase TiO_2 using (a) BO and (b) random sampling, with markers coloured according to their value of the cost function J^{SE}. Hubbard parameters that violate the constraints on $\mathrm{Tr}[\mathbf{n}(\mathrm{Ti}\ 3d)]^{\mathrm{SISSO}}$, $\mathrm{Tr}[\mathbf{n}(\mathrm{O}\ 2p)]^{\mathrm{SISSO}}$, S_1 and S_2 are excluded. (c) The distribution of values of J^{SE} corresponding to the 1350 sampled Hubbard parameters using BO (red and purple markers) and the results of the first 1350 iterations using random sampling (pink markers). In BO, the prior distribution is conditioned using evaluations of 1000 randomly sampled Hubbard parameters selected using Latin Hypercube Sampling. During BO, any sampled Hubbard parameters that result in constraint violation are assigned a value of J^{SE}=1000. After 1000 iterations, BO is performed for 350 iterations to efficiently optimise U, c_1 and c_2.

Refining the Ti $3d$ Hubbard projector therefore enables numerically stable self-consistent defect calculations; however, the predictions for anatase TiO_2 have unphysical defect energies, ranging from -11.59 eV and -1.28 eV; the same is not observed for defective rutile TiO_2, with defect energies ranging

from -0.97 eV to 8.22 eV. These unphysical defect energies for anatase TiO_2, necessitate further study into the effect of tuning the Ti 3*d* Hubbard projector on the DFT+*U*-predicted total energy (E), as illustrated in Figure 4.8(a), which shows an interpolated surface plot of E, c_1 and c_2 for bulk stoichiometric anatase TiO_2, calculated using U = 0.5 eV and coloured according to the gradient norm of the partial derivatives of E with respect to c_1 and c_2. As evident by the blue to yellow transition in Figure 4.8(a), corresponding to increasing partial derivatives of E with respect to c_1 and c_2, there are large derivative discontinuities in E upon localising the Ti 3*d* Hubbard projector.

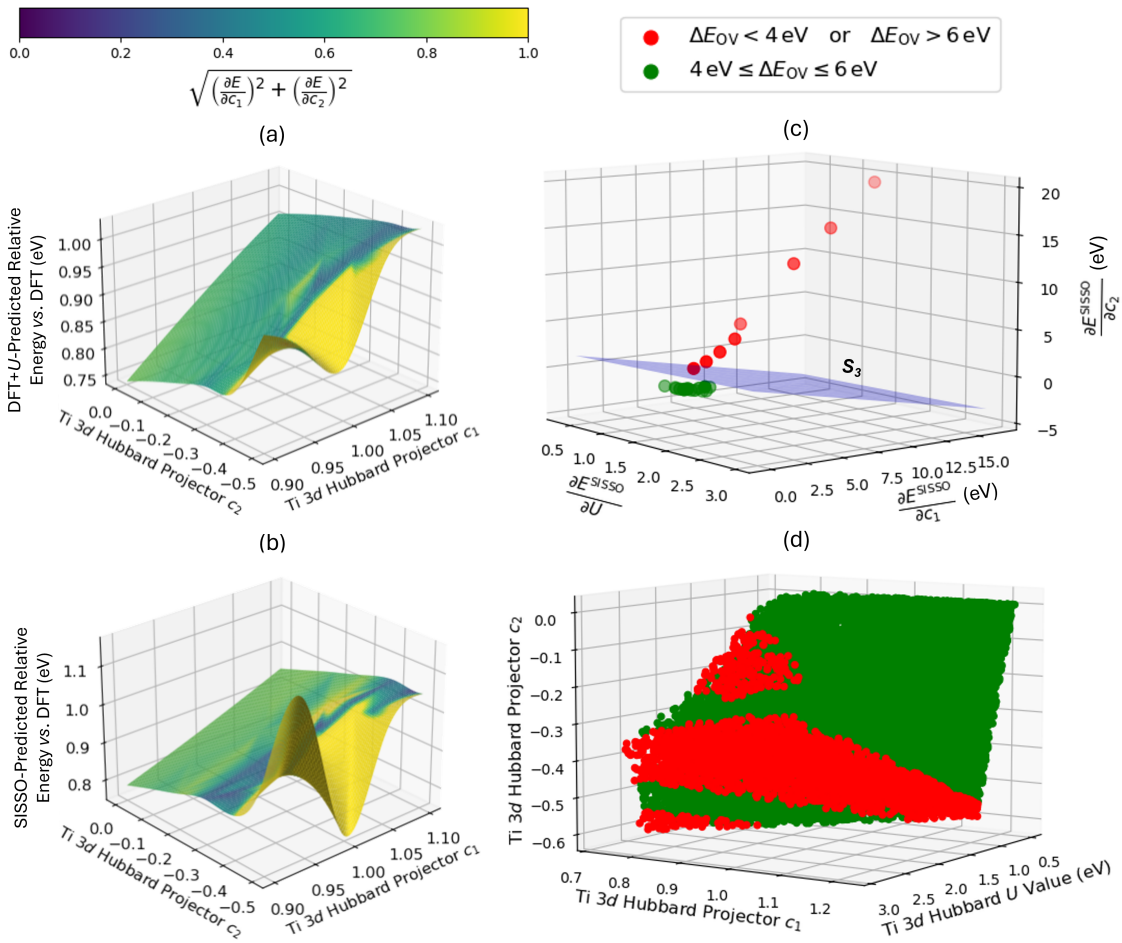


Figure 4.8: Interpolated surface plots of the (a) DFT+*U*-predicted total energy (E) and (b) SISSO-predicted total energy (E^{SISSO}), both normalised using the DFT-predicted total energy for anatase TiO_2, plotted as a function of c_1 and c_2 with U = 0.5 eV. Each surface plot is coloured according to the gradient norm of the partial derivatives of the relative total energy with respect to c_1 and c_2. (c) The linear boundary S_3, that classifies the defect energies of the converged bulk oxygen vacancy calculations in Figure 4.6, separates "physical" (green markers for 4 eV ≤ ΔE_{OV} ≤ 6 eV) and "unphysical" (red markers for ΔE_{OV} < 4 eV or ΔE_{OV} > 6 eV) oxygen vacancy formation energies, using the partial derivatives of E^{SISSO} with respect to U, c_1 and c_2. (d) The same plot as Figure 4.7(b), with markers coloured according to the satisfaction (green) or violation (red) of the SVM-constraint derived from S_3 in Equation (4.24).

The discontinuities are believed to be the root cause for the unphysical defect energies using U = 2.749 eV, c_1 = 0.758 and c_2 = -0.354. Therefore, an additional constraint was constructed to filter out Hubbard parameters in Figure 4.7(b) that are predicted to similarly lead to unphysical defect

energies. The constraint depends on a SISSO-computed correlation for E in terms of U, c_1, and c_2, *i.e.*, E^{SISSO} which is detailed in Section 4.2.2. E^{SISSO} reasonably captures the derivative discontinuities in Figure 4.8(a) (calculated using DFT+U), as shown in Figure 4.8(b) (calculated using E^{SISSO}). Using the partial derivatives of E^{SISSO} with respect to U, c_1 and c_2 as three primary features, the linear SVM S_3 classifies whether a particular combination of U, c_1, and c_2, for all converged defect calculations in Figure 4.6, leads to "physical" (4 eV $\leq \Delta E_{\text{OV}} \leq$ 6 eV) or "unphysical" ($\Delta E_{\text{OV}} <$ 4 eV or $\Delta E_{\text{OV}} >$ 6 eV) oxygen vacancy formation energies, as illustrated in Figure 4.8(c). The constraint is defined below, with the linear equation corresponding to S_3 detailed in Section 4.2.2:

$$-1 < S_3 < 1.56 \tag{4.24}$$

Filtering out the data points in Figure 4.7(b) that do not satisfy Equation (4.24), as illustrated in Figure 4.8(d), yields the optimal Hubbard parameters U = 2.575 eV, c_1 = 0.752 and c_2 = -0.486. The approximate similarity is noted of our optimised Ti 3d Hubbard projector with that determined by Jakob and Oberhofer for bulk rutile TiO_2, defined by c_1 = 0.828 and c_2 = -0.561, using a first-principles generalised LR-cDFT method in FHI-aims. [57]

Defect Energies and Computational Cost *vs.* Hybrid-DFT

Simulations of stoichiometric bulk anatase and rutile TiO_2, as listed in Table 4.9, show DFT+U simultaneously predicts both target properties, E_{bg} and V_0, with superior accuracy than DFT, relative to experimental references. The hybrid-DFT-predicted V_0 is closest to the experimental value, although hybrid-DFT significantly overestimates E_{bg}. DFT+U predicts a smaller ΔE_{Form} than DFT, hybrid-DFT and experiment, although all computational methods maintain a similar relative difference between ΔE_{Form} for anatase and rutile TiO_2, with $\Delta E_{\text{Form}}^{\text{Anatase}} - \Delta E_{\text{Form}}^{\text{Rutile}}$ values of -0.02 eV, -0.01 eV and -0.01 eV for DFT, DFT+U and hybrid-DFT, respectively. The DFT+U-predicted covalency of TiO_2 is also suppressed compared to hybrid-DFT, as evidenced by the reduced total Ti 3d orbital occupancy, *i.e.*, Tr[$\mathbf{n}$(Ti 3d)], whilst the DFT+U-predicted total O 2p orbital occupancy, given by Tr[$\mathbf{n}$(O 2p)], remains close to that of hybrid-DFT.

Table 4.9: Geometric, electronic and energetic properties of bulk anatase and rutile TiO_2, predicted using DFT (mBEEF exchange-correlation functional), DFT+U (mBEEF exchange-correlation functional, U = 2.575 eV, c_1 = 0.752 and c_2 = -0.486) and hybrid-DFT (PBE0 exchange-correlation functional), presented alongside experimental references: band gap (E_{bg}, eV), unit cell equilibrium volume (V_0, Å^3), formation energy (ΔE_{Form}, eV/atom), Ti 3d occupation matrix trace (Tr[$\mathbf{n}$(Ti 3d)]), O 2p occupation matrix trace (Tr[$\mathbf{n}$(O 2p)])

TiO_2 Polymorph	Method	E_{bg} (eV)	V_0 (Å^3)	ΔE_{Form} (eV/Atom)	Tr[$\mathbf{n}$(Ti 3d)]	Tr[$\mathbf{n}$(O 2p)]
Anatase	DFT	2.54	137.26	-3.28	1.18	4.91
Anatase	DFT+U	2.75	137.09	-3.02	0.69	4.64
Anatase	Hybrid-DFT	4.29	135.49	-3.21	1.25	4.66
Anatase	Experiment	3.20 [30]	136.28 [31]	-3.24 [51]	N/A	N/A
Rutile	DFT	2.21	63.50	-3.26	1.17	4.95
Rutile	DFT+U	2.42	62.51	-3.01	0.68	4.63
Rutile	Hybrid-DFT	4.05	62.39	-3.20	1.23	4.69
Rutile	Experiment	3.00 [30]	62.44 [31]	-3.26 [51]	N/A	N/A

Next, self-consistent DFT+U simulations were performed for anatase and rutile TiO_2 containing an oxygen vacancy and the substititutional dopants Nb, W, Au, Pd, Pt, Co and Mn. Table 4.10 summarises the success or failure of these simulations, both with or without a refined Ti 3d Hubbard projector and with or without the Hubbard parameters satisfying the error correction constraint derived from S_3 in Equation (4.24).

Table 4.10: The effect of Hubbard U value and projector modification on the numerical stability of point defect calculations in TiO_2. Ticks (crosses) correspond to successful convergence (calculation termination) of self-consistent calculations using the OMR method. The satisfaction of constraints derived from the SVM boundaries S_1, S_2 and/or S_3, given by Equations (4.23) and (4.24), affects the predicted defect energies corresponding to each set of Hubbard parameters, as shown in Figure 4.9(a).

					Point Defect							
Polymorph	U (eV)	c_1	c_2	Satisfied Constraints	$V_O^\times$	$Nb_{Ti}^\times$	$W_{Ti}^\times$	$Au_{Ti}^\times$	$Pd_{Ti}^\times$	$Pt_{Ti}^\times$	$Co_{Ti}^\times$	$Mn_{Ti}^\times$
Anatase	2.749	1	0	None	✗	✗	✗	✗	✗	✗	✗	✗
Anatase	2.575	1	0	None	✗	✗	✗	✗	✓	✓	✓	✓
Anatase	2.749	0.758	-0.354	S_1, S_2	✓	✓	✓	✓	✓	✓	✓	✓
Anatase	2.575	0.752	-0.486	S_1, S_2, S_3	✓	✓	✓	✓	✓	✓	✓	✓
Rutile	2.749	1	0	None	✗	✗	✗	✓	✓	✓	✓	✓
Rutile	2.575	1	0	None	✗	✗	✗	✓	✓	✓	✓	✓
Rutile	2.749	0.758	-0.354	S_1, S_2	✓	✓	✓	✓	✓	✓	✓	✓
Rutile	2.575	0.752	-0.486	S_1, S_2, S_3	✓	✓	✓	✓	✓	✓	✓	✓

Comparing the rows in Table 4.10 that use the same Ti 3d Hubbard U value but a different Ti 3d Hubbard projector shows that refining the Hubbard projector has a direct effect on enabling self-consistent defect simulations. All calculation failures were due to unphysical ground states causing the termination of calculations due to excessive polaron localisation as seen in Figure 4.5(e). Refining the Ti 3d Hubbard parameters further from U = 2.749 eV, c_1 = 0.758 and c_2 = -0.354 to U = 2.575 eV, c_1 = 0.752 and c_2 = -0.486 using the constraint derived from the SVM boundary S_3, gave physical defect energies for all studied defects in anatase and rutile TiO_2, as illustrated in Figure 4.9(a). Comparing the DFT-, DFT+U- (U = 2.575 eV, c_1 = 0.752 and c_2 = -0.486) and hybrid-DFT predicted defect energies, the DFT+U-predicted ΔE_{OV} in anatase (5.42 eV) and rutile (5.48 eV) are both much closer to the corresponding values computed using hybrid-DFT (5.29 and 5.76 eV respectively) than those computed using DFT (4.44 and 4.43 eV respectively). Without the application of a Hubbard correction to the dopant atom d orbitals, there was no obvious trend in the raw DFT+U-predicted substitutional defect energies, compared to hybrid-DFT. However, when normalised with respect to ΔE_{OV} in the same TiO_2 polymorph, the DFT+U-predicted values of ΔE_{Defect} were closer to the corresponding hybrid-DFT-predicted values compared to DFT, as illustrated in Figure 4.9(b), which shows the relative defect energies for Pt- and Pd-doped anatase and rutile TiO_2 computed using DFT, DFT+U and hybrid-DFT (single point calculations using the DFT+U geometry). The blue columns in Figure 4.9(c) plot the cost of the hybrid-DFT single-point calculation in core-hours per atom, relative to the cost of the DFT+U geometry optimisation calculation. For the four systems presented in Figure 4.9(b), DFT+U with geometry optimisation is between 79-189× faster than the hybrid-DFT single point calculations.

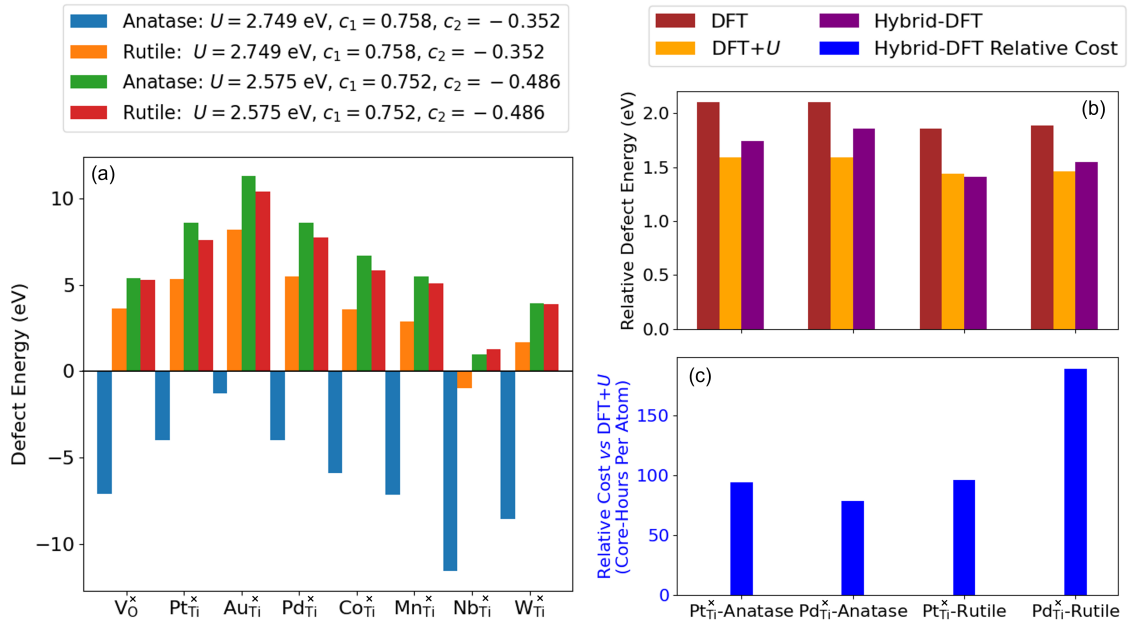


Figure 4.9: (a) DFT+*U*-predicted values of ΔE_{OV} and ΔE_{Defect} using a refined set of Hubbard parameters that satisfy the constraints derived from the SVM boundaries S_1 and S_2, calculated using the mBEEF exchange-correlation functional, U = 2.749 eV, c_1 = 0.758 and c_2 = -0.354 *vs.* Hubbard parameters that satisfy the constraints derived from the SVM constraints S_1, S_2 and S_3, calculated using the mBEEF exchange-correlation functional, U = 2.575 eV, c_1 = 0.752 and c_2 = -0.486. (b) ΔE_{Defect} relative to ΔE_{OV} for Pt- and Pd-doped anatase and rutile TiO_2, calculated using geometry optimisation calculations using DFT and DFT+*U* (mBEEF exchange-correlation functional, U = 2.575 eV, c_1 = 0.752 and c_2 = -0.486) and single point calculations using hybrid-DFT (PBE0 exchange-correlation functional, using the DFT+*U*-optimised geometry). (c) The cost of the hybrid-DFT single point calculations relative to the DFT+*U* geometry optimisation calculations in core-hours per atom.

Electron Polarons in Anatase vs. Rutile TiO_2

DFT+*U* using $U = 2.575$ eV, $c_1 = 0.752$ and $c_2 = -0.486$ can be used to simulate experimentally observed differences in electron polaron formation in anatase *vs.* rutile TiO_2, demonstrating the impact of our parameterisation work. For example, DFT+*U* predicts the polymorph-dependent formation of small Holstein electron polarons in W-doped TiO_2, characterised by strong electron-lattice interactions resulting in a localised state within the TiO_2 band gap. [2, 58] This is shown in Figure 4.10, where the formation of a localised W $5d_{z^2}$ state within the rutile TiO_2 band gap is increasingly formed by incrementing the level of theory from DFT in Figure 4.10(a), to DFT+*U* in Figure 4.10(b) and hybrid-DFT in Figure 4.10(c).

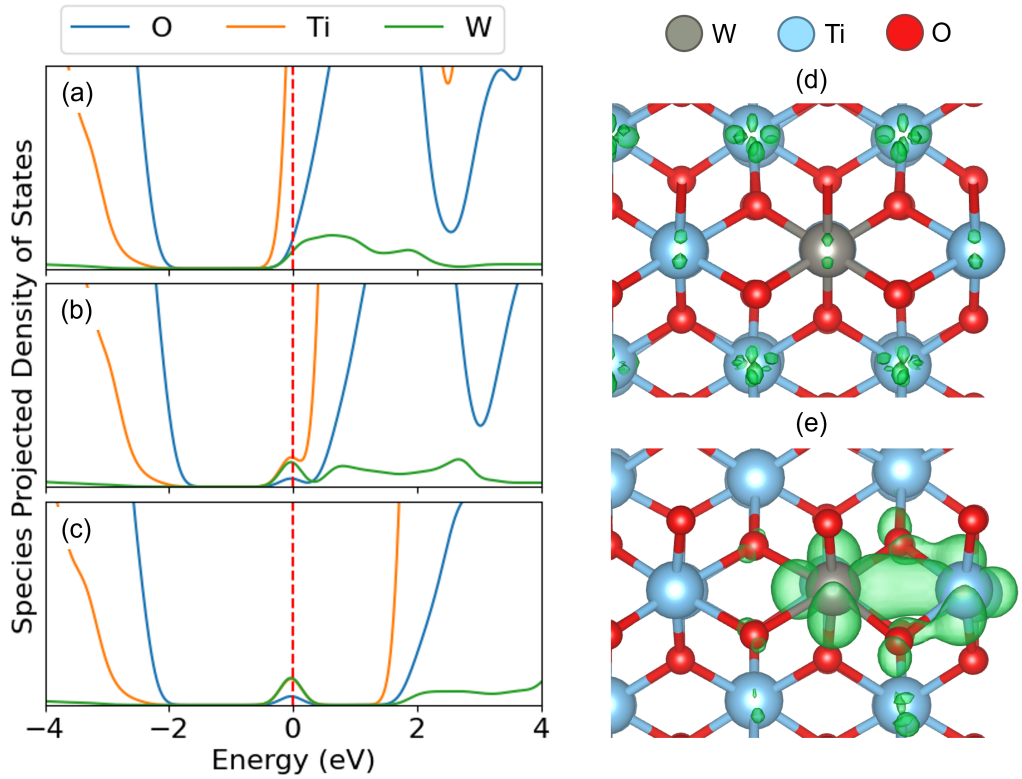


Figure 4.10: The elemental species projected density of states for W-doped rutile TiO_2, calculated using (a) DFT (mBEEF exchange-correlation functional), (b) DFT+U (mBEEF exchange-correlation functional, U = 2.575 eV, c_1 = 0.752 and c_2 = -0.486) and (c) hybrid-DFT (PBE0 exchange-correlation functional single point calculation using the DFT+U optimised geometry). The Fermi level is denoted by the red dashed line. The corresponding charge density isosurfaces for the highest occupied molecular orbital (HOMO), at the 0.025 eÅ^{-3} level, are shown for (d) DFT and (e) DFT+U.

The degree of defect state localisation in W-doped rutile TiO_2, given by the energy gap between the localised state and the TiO_2 conduction band minimum, is predicted as 0.46 eV using DFT+U and 1.76 eV using hybrid-DFT. As illustrated by the charge density isosurfaces of the highest occupied molecular orbital (HOMO) predicted by DFT in Figure 4.10(d) and DFT+U in Figure 4.10(e), DFT predicts both W 5d and Ti 3d states to be delocalised, *i.e.*, predicting W^{6+} formation, whilst DFT+U predicts a hybridised defect state of W 5d and Ti 3d character, indicating the predicted formation of W^{5+}. Similarly, DFT+U simulations of Nb-doped rutile TiO_2 predict a Nb 4d signature at the Fermi level that is 5.5× larger than that in Nb-doped anatase TiO_2, supporting previous work using EPR spectroscopy to characterise the formation of paramagnetic Nb^{4+} and W^{5+} in doped rutile TiO_2, which is recoverable using self-consistent DFT+U as opposed to standalone DFT or constrained DFT+U using OMC. [2]

More generally, the DFT+U simulations reveal clear differences in the electronic structures of substitutionally doped anatase and rutile TiO_2, particularly in the occupancies of the dopant atom d orbitals, as illustrated in Figure 4.11(a). In rutile TiO_2, there are greater occupancies of the dopant atom e_g orbitals (m = 0 and 2) [25] than in anatase, indicating a structural and electronic environment that favours the filling of orbitals aligned *along* the metal-oxygen bonds. Conversely, substitutionally doped anatase TiO_2 shows increased occupancies of the dopant atom t_{2g} orbitals (m = -2, -1 and 1)

[25] reflecting a local electronic environment that favours the filling of orbitals aligned *between* the metal-oxygen bonds.

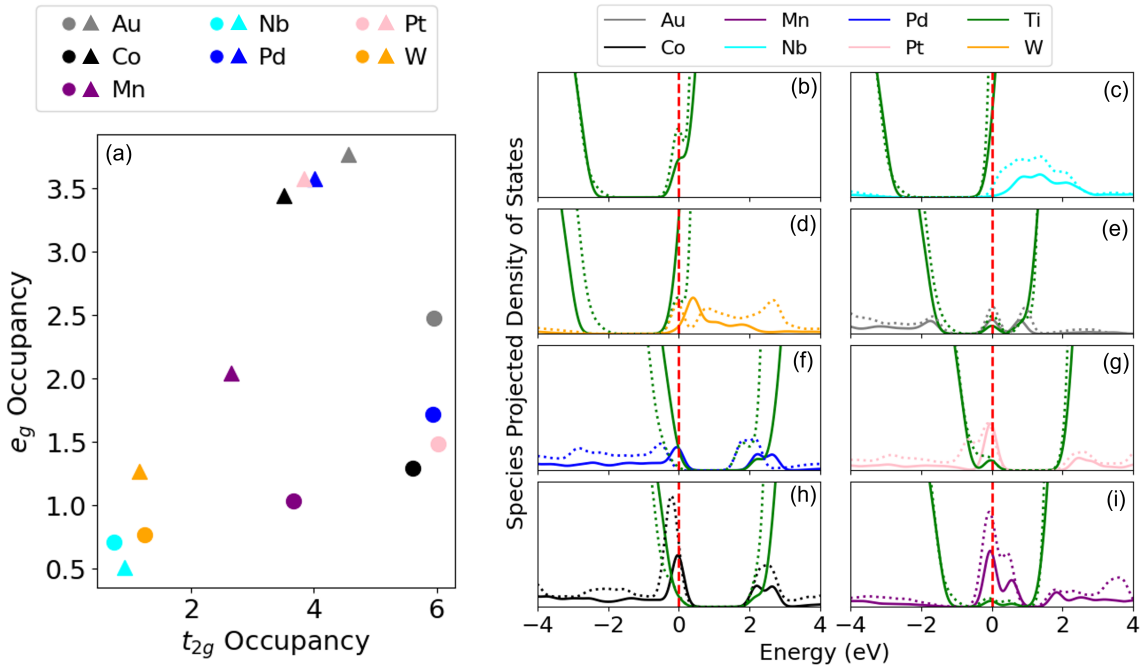


Figure 4.11: The DFT+*U*-predicted (a) occupancies of the dopant atom t_{2g} and e_g orbitals, in the 3*d*, 4*d* or 5*d* subshell, for all extrinsic defects in anatase (circles) and rutile (triangles), calculated using the mBEEF exchange-correlation functional, U = 2.575 eV, c_1 = 0.752 and c_2 = -0.486. The corresponding elemental species projected density of states for (b) $V_O^{\times}$, (c) $Nb_{Ti}^{\times}$, (d) $W_{Ti}^{\times}$, (e) $Au_{Ti}^{\times}$, (f) $Pd_{Ti}^{\times}$, (g) $Pt_{Ti}^{\times}$, (h) $Co_{Ti}^{\times}$, (i) $Mn_{Ti}^{\times}$ in bulk anatase (solid lines) and rutile (dashed lines) TiO_2, are normalised with respect to the different defect concentrations in the anatase and rutile simulation supercells. The Fermi level is denoted by the red dashed line.

The formation of localised *vs.* delocalised defects states is directly affected by the filling of the e_g and t_{2g} orbitals, respectively, leading to polymorph-sensitivity in the electrical conductivity and chemical reactivity of TiO_2-based materials. [2] Localised defect states within the TiO_2 band gap (or present at the valance or conduction band edges), as observed for Nb- and W-doped rutile in Figures 4.11(c) and (d), respectively, and both TiO_2 polymorphs containing an oxygen vacancy and Au, Pd, Pt, Co and Mn dopants in Figures 4.11(b), (e), (f) (g), (h) and (i), respectively, give rise to electronic conduction *via* thermally activated polaron hopping (conductivity increases with temperature) and provide readily available sites for activating reactant molecules. [59, 60] Conversely, delocalised defect states located deeper in the valence or conduction bands, as observed for Nb- and W-doped anatase in Figures 4.11(c) and (d), respectively, give rise to improved electronic conductivity according to a band-like model and can both reduce the degrading recombination of electrons and holes in solar cells and lower the rate determining step of the oxygen evolution reaction on rutile TiO_2 surfaces. [61, 62]

4.3.3 Optimising Hubbard U Values and Projectors from First-Principles

Motivation and Method Reconfiguration

Following the success of the semi-empirical machine learning approach in mitigating numerical instability during simulations of defects in TiO_2 (Section 4.3.2), we now turn to the development of a fully first-principles strategy for optimising Hubbard U values and projectors. Simultaneous optimisation of Hubbard U values and projectors is particularly attractive as a step towards the long-term goal of formulating DFT+U as a functional of the electron density and orbitals, with all corrective Hubbard parameters determined self-consistently, rather than being treated as external parameters. However, the reformulation of the semi-empirical machine learning approach is contingent on two major modifications. Firstly, it is necessary to redefine the semi-empirical cost function (J^{SE}) by removing all experimental reference data, but in a manner that will ultimately yield similar results leading to the numerically stable self-consistent defect calculations in Section 4.3.2. Secondly, the termination and SCF non-convergence of defect calculations observed for TiO_2 were also observed for a broad range of TMOs and REOs; therefore, it is necessary to generalise the SVM constraints in Section 4.3.2 in order to filter out Hubbard parameters that lead to unstable defect calculations for systems beyond TiO_2.

To investigate how best to construct a first-principles cost function (J^{FP}), the DFT+U-predicted Ti 3d and O 2p orbital occupancies in anatase and rutile TiO_2, using a refined and atomic Ti 3d Hubbard projector, were compared with the occupancies calculated using hybrid-DFT, as illustrated in Figure 4.12. Figures 4.12(a) and (c) show that DFT+U using an atomic Ti 3d Hubbard projector predicts Ti 3d orbital occupancies closer to those calculated using hybrid-DFT, although this is expected as the hybrid-DFT-calculated occupancies are derived from the atomic Ti 3d Hubbard projector and so cannot be compared in a like-for-like manner to those derived from a modified Hubbard projector (detailed in Section 2.3.2). Figures 4.12(b) and (d) show that DFT+U using the refined Ti 3d Hubbard projector predicts O 2p orbital occupancies closer to those predict using hybrid-DFT, with errors < 5% for all values of m, which is a like-for-like comparison as the O 2p orbital occupancies are derived from the atomic O 2p Hubbard projector across all methods (detailed in Section 2.3.2). The first-principles cost function was therefore defined as the average error in the DFT+U-predicted O 2p orbital occupancies *vs.* hybrid-DFT (defined in Section 4.2.3), which is obtained from a single reference calculation of a geometry optimised unit cell. To investigate the requirement for SVM constraints in a first-principles approach for optimising Hubbard U values and projectors, the dataset for classifying the validity of bulk oxygen vacancy calculations in anatase TiO_2 (Figure 4.12(e)) was considered again in terms of percentage errors of $\mathrm{Tr}[\mathbf{n}(\mathrm{Ti}\ 3d)]$ and $\mathrm{Tr}[\mathbf{n}(\mathrm{O}\ 2p)]$ *vs.* hybrid-DFT (Figure 4.12(f)). Figure 4.12(f) shows that minimising J^{FP} towards zero is necessary but not a sufficient condition to ensure the numerical stability of point defect calculations in anatase TiO_2, highlighting the requirement of satisfying SVM-derived constraints as well as minimising J^{FP} for accurate Hubbard parameter optimisation.

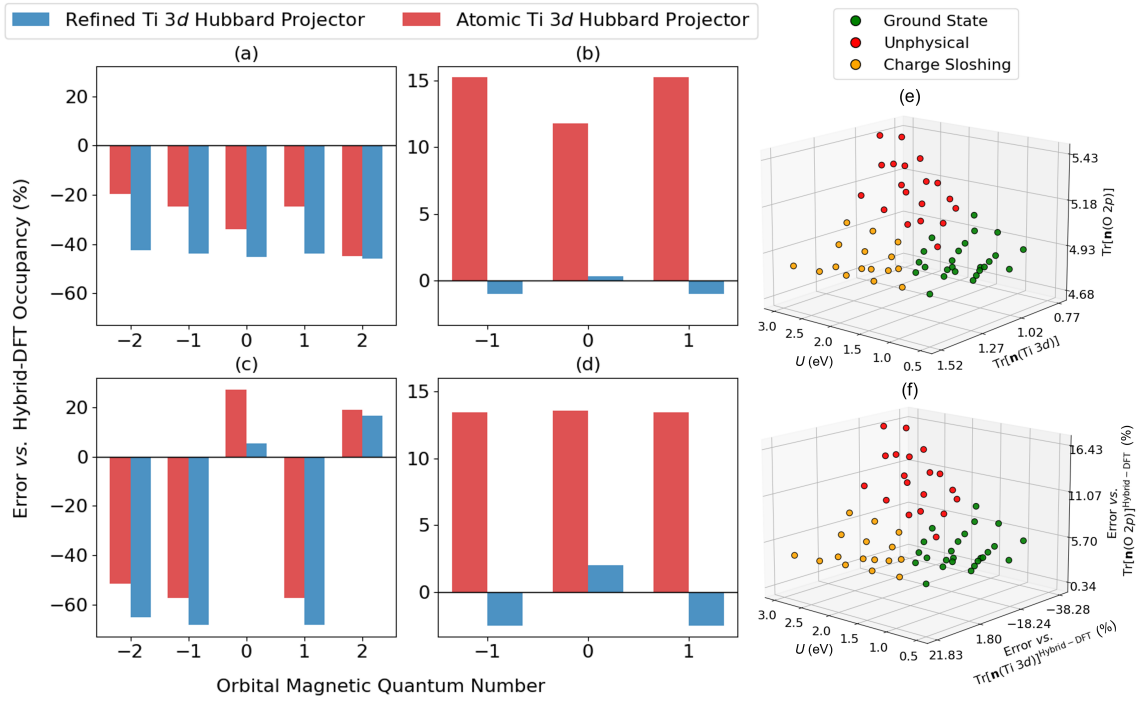


Figure 4.12: Comparing the percentage errors of the DFT+*U*-predicted Ti 3*d* and O 2*p* orbital occupancies in anatase and rutile TiO_2 *vs.* hybrid-DFT (PBE0 exchange-correlation functional); (a) Ti 3*d* in anatase TiO_2, (b) O 2*p* in anatase TiO_2, (c) Ti 3*d* in rutile TiO_2 and (d) O 2*p* in rutile TiO_2. Blue bars correspond to DFT+*U* using the mBEEF exchange-correlation functional, U = 2.575, c_1 = 0.752 and c_2 = -0.486. Red bars correspond to DFT+*U* using the mBEEF exchange-correlation functional, U = 2.575, c_1 = 1 and c_2 = 0. The same outcomes of bulk oxygen vacancy calculations plotted in Figure 4.6 are plotted in (e) and (f), where (e) is plotted in terms of the raw U, $\mathrm{Tr}[\mathbf{n}(\mathrm{Ti}\ 3d)]$ and $\mathrm{Tr}[\mathbf{n}(\mathrm{O}\ 2p)]$ values, whilst (f) is plotted in terms of U and the percentage errors of $\mathrm{Tr}[\mathbf{n}(\mathrm{Ti}\ 3d)]$ and $\mathrm{Tr}[\mathbf{n}(\mathrm{O}\ 2p)]$ *vs.* hybrid-DFT.

In the following sections, we outline the generalisation of both the regression and classification methods adopted in Section 4.3.2, extending them towards the development of a single unified model that can automatically parameterise Hubbard U values and projectors from material-dependent inputs, *i.e.*, in the spirit of a foundation model for DFT+*U* parameterisation. This serves as a test of how well the insights gained from TiO_2 transfer across chemical space, with the ultimate goal of enabling accurate, self-consistent simulations of defects and polarons across a broad range of TMOs and REOs. We note that such an approach could equally be formulated as an iterative active learning scheme for systems that are difficult to incorporate within a single unified model, as outlined in Section 6.2.1.

Generalised Symbolic Regression

Estimating the DFT+*U*-predicted O 2*p* occupancies was achieved using three HI-SISSO correlations, which were constructed using the expanded primary feature set of Hubbard parameters, basis set parameters, DFT predicted orbital occupancies and atomic material descriptors (listed in the Section 4.2.3), before searching the landscape of $J^{\mathrm{FP}}_{\mathrm{Predicted}}$ for the ten materials listed in Table 2.1, using the corresponding material-dependent descriptors and hybrid-DFT reference orbital occupancies. The

outputs of the linear search were families of solutions for each material (Figure 4.13), of which 94 combinations of U, c_1 and c_2 were validated using DFT+U to compare the accuracy of the $J^{\text{FP}}_{\text{Predicted}}$ values (using the HI-SISSO correlations) versus $J^{\text{FP}}_{\text{Validated}}$ values (from DFT+U calculations), achieving an average MAE across all materials of 5.02% in Figure 4.14(a).

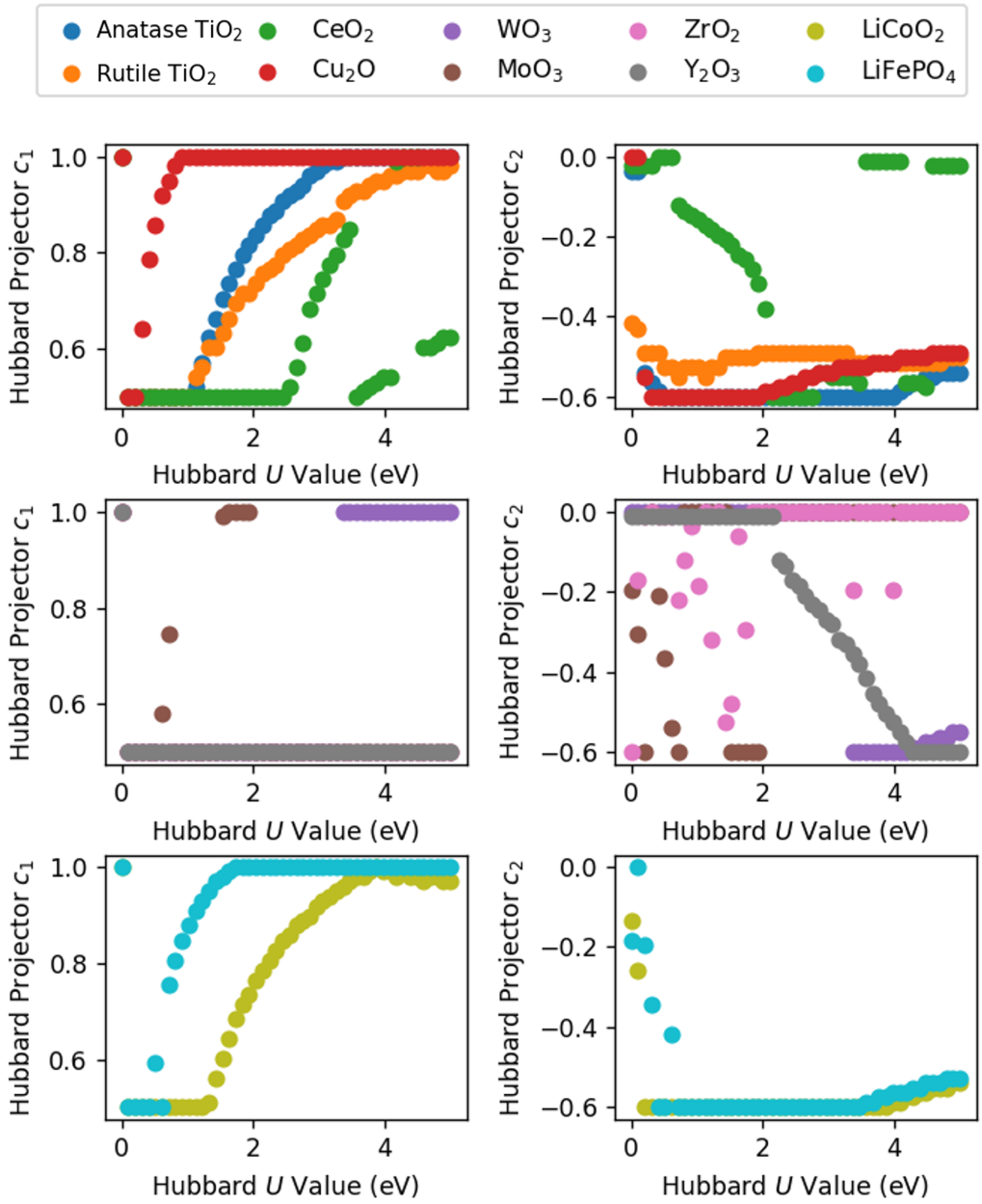


Figure 4.13: Computed families of solutions following a linear search of the landscape of $J^{\text{FP}}_{\text{Predicted}}$, producing optimised Hubbard projectors for a given Hubbard U value, to minimise $J^{\text{FP}}_{\text{Predicted}}$.

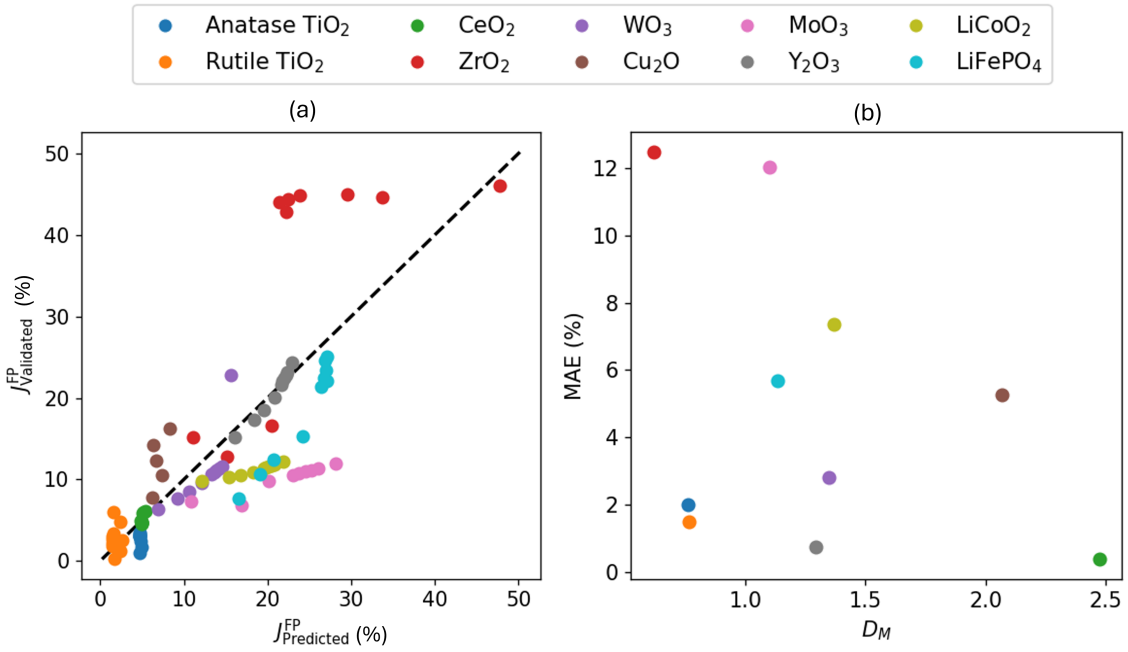


Figure 4.14: (a) Parity plot of the HI-SISSO-predicted ($J^{FP}_{Predicted}$) and DFT+U-validated ($J^{FP}_{Validated}$) cost function across ten TMOs and REOs using the generalised approach. (b) Comparison of the MAE of $J^{FP}_{Predicted}$ for each material in (a) *vs*. the corresponding Mahalanobis distance (D_M), averaged over all tested combinations of Hubbard parameters, to visualise the dependence of the accuracy of the generalised approach on the training set size for each material.

To investigate the relationship between the residuals in Figure 4.14(a) and the size of the training set per material, the MAE was compared with D_M (both averaged per material) to quantify whether Hubbard projector optimisation is interpolative or extrapolative. As illustrated in Figure 4.14(b), the generalised approach is more interpolative for anatase TiO_2, rutile TiO_2 and ZrO_2 which have the three smallest values of D_M. Large values of D_M, as seen for CeO_2, Cu_2O and $LiCoO_2$, indicate the generalised approach is comparatively extrapolative for these materials. As there is no correlation between the MAE and D_M, the larger residuals in Figure 4.14(a) for ZrO_2 and MoO_3 do not appear to be related to the training set size and are likely due to the choice of primary features used in HI-SISSO, such as the lack of structure-dependent features, requiring further study to enhance the method transferability.

Generalised Symbolic Classification

Given that the minimisation of J^{FP} is necessary but not sufficient for numerically stable point defect calculations in TiO_2, as illustrated in Figure 4.12(f), the constraints on the DFT+U-predicted covalency established in Figure 4.6 needed to be generalised across materials. Generalised constraints were determined using the primary features U and J^{FP}, as well as two additional material-dependent descriptors of covalency, including the average error in the DFT+U-predicted metal d or f orbital occupancies compared to hybrid-DFT (E^{Metal}) and the ratio of the traces of the metal d or f and O $2p$ occupation matrices predicted using hybrid-DFT (R^{Hybrid}), as defined in Section 4.2.3. With the primary features U, J^{FP}, E^{Metal} and R^{Hybrid}, generalised classification was performed to predict the

same outcomes shown in Figure 4.6 for bulk oxygen vacancy calculations, but across TiO_2, CeO_2, ZrO_2, MoO_3, WO_3 and Cu_2O, as illustrated in Figure 4.15(a) and (b), where the unitless SVM boundaries S_4 and S_5 are defined in Section 4.2.3 and the constant $\alpha = 1\ \mathrm{eV}^{-1}$ is introduced to ensure dimensional consistency. S_4 and S_5 are then combined to form a constraint on the generalised Hubbard parameter space, to ensure the numerical stability of point defect calculations using the OMR method:

$$S_4 \geq 0 \quad \wedge \quad S_5 \geq 0 \tag{4.25}$$

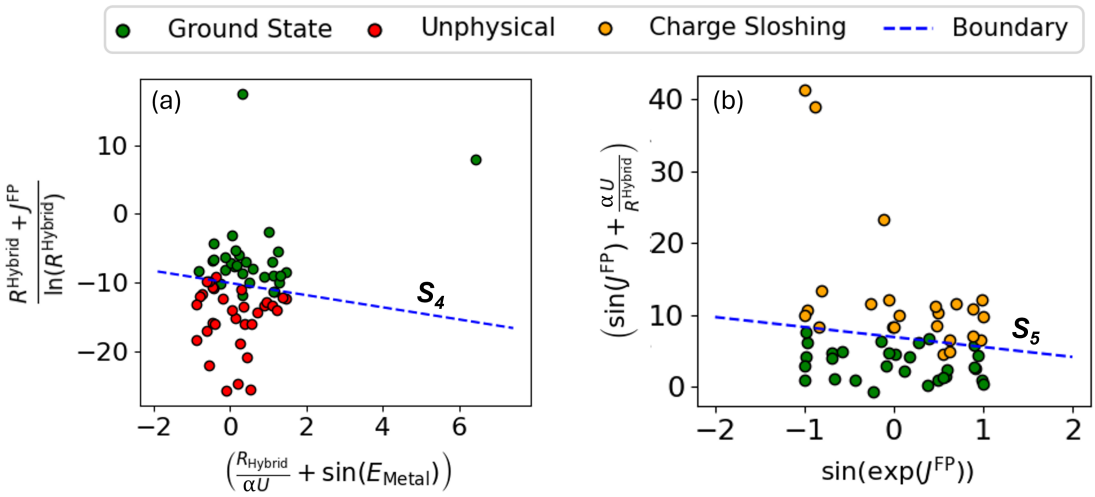


Figure 4.15: The linear boundaries (a) S_4 and (b) S_5 that classify the numerical stability of DFT+U simulations of a bulk oxygen vacancy in TiO_2, CeO_2, ZrO_2, MoO_3, WO_3 and Cu_2O, separating regions in the DFT+U-computed feature space that lead to successful convergence, termination due to an unphysical ground state and charge sloshing that prevents SCF convergence.

Electron and Hole Polarons in $LiCo_{1-x}Mg_xO_{2-x}$

The generalised regression and classification models were then integrated to perform a one-shot screening of the Hubbard parameter space for a previously unseen material, to test the accuracy and numerical stability of DFT+U simulations of the stoichiometric and defective electronic structure. We chose the common Li-ion battery cathode material $LiCoO_2$, which is the focus of wide-ranging studies to enhance its electrochemical properties, such as electrical conductivity and long-term cycling stability, *via* the introduction of low-valence ions resulting in charge compensation from the formation of point defects. [63–65] For example, substitution of Co^{3+} in $LiCoO_2$ with Mg^{2+} is proposed to result in the formation of Co^{4+} holes, that exist delocalised [66–68] and are attributed to the 100× enhanced electrical conductivity of $LiCo_{1-x}Mg_xO_2$ compared to pristine $LiCoO_2$. [69] Other reports suggest that charge compensation and enhancements in electrical conductivity occur *via* the formation of oxygen vacancies, whilst Mg doping does not cause a change in the oxidation state of Co atoms. [66, 70].

Literature reported DFT+U studies, in a planewave basis, have yet to provide an unambiguous rationale for the increased electrical conductivity of Mg-doped $LiCoO_2$, with significant sensitivity of the DFT+U-predicted localisation of Co^{4+} holes with respect to the choice of Hubbard parameters.

For example, DFT+U using a Co 3d Hubbard U value determined from first-principles using LR-cDFT (between 4.91 eV to 5.62 eV) predicts deep Co 3d defect states within the $LiCoO_2$ band gap, which is inconsistent with the experimentally observed high electrical conductivity of $LiCo_{1-x}Mg_xO_2$, whilst DFT+U using smaller Hubbard U values predicts shallow defect states, supporting the experimental observations. [15, 71] Whilst no explanation for the sensitivity of DFT+U simulations of $LiCo_{1-x}Mg_xO_2$ was given, Hoang and Johannes used DFT+U to simulate the self-trapping of Ni^{4+} holes in $LiNiO_2$, where the DFT+U-predicted instability of Ni^{4+} was attributed to the predicted O 2p character of the $LiNiO_2$ valence band edge, highlighting the importance of considering band edge effects when choosing appropriate Hubbard parameters. [72]

To test the DFT+U-predicted localisation of hole polarons in $LiCo_{1-x}Mg_xO_2$ in a NAO framework, the one-shot approach was first applied to determine an appropriate Co 3d Hubbard U value and projector for stoichiometric $LiCoO_2$. The parameters were determined using a linear search of the HI-SISSO-predicted cost function ($J^{\text{FP}}_{\text{Predicted}}$) for U between 0 eV and 5 eV, c_1 between 0.5 and 1 and c_2 between 0 and -0.6, before all Hubbard parameters that violate Equation 4.25 were discarded. The output of this initial screening is a region of the Hubbard parameter space (U, c_1 and c_2) that minimises $J^{\text{FP}}_{\text{Predicted}}$ and satisfies Equation 4.25 but contains multiple possible solutions. We therefore narrow this region down to three candidate Hubbard parameters for validation using K-means clustering of a reduced subset of the screened Hubbard parameter space that corresponds to the lowest 10 % of $J^{\text{FP}}_{\text{Predicted}}$ (illustrated by the hatched region in Figure 4.16).

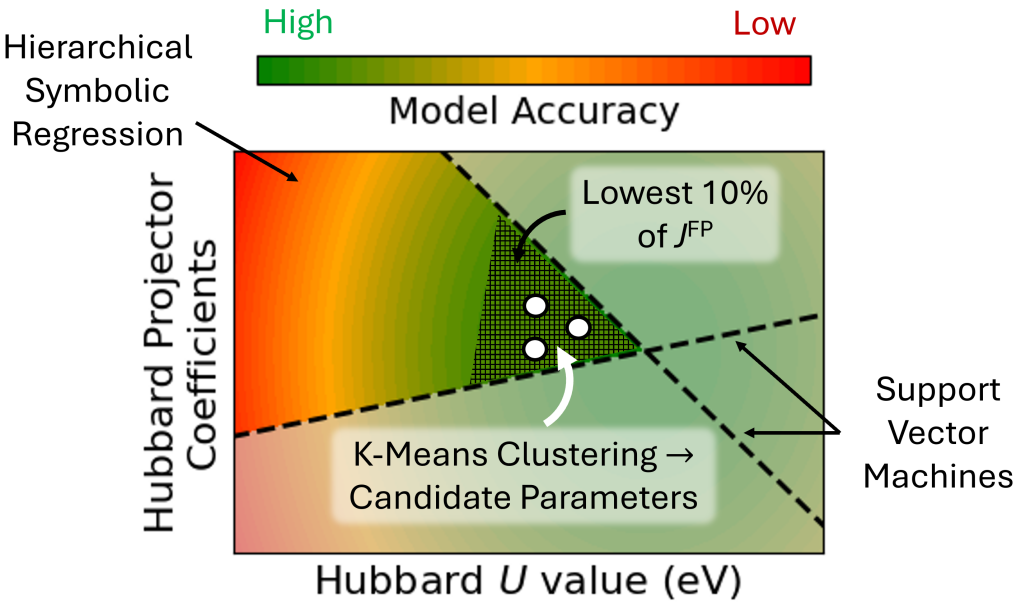


Figure 4.16: Integrated one-shot approach for simultaneously optimising Hubbard U values and projectors from first-principles. The landscape of the first-principles cost function ($J^{\text{FP}}_{\text{Predicted}}$) is predicted using hierarchical symbolic regression (Section 4.3.3) and unsuitable Hubbard parameter values are excluded using support vector machines (Section 4.3.3). The remaining region of the Hubbard parameter space is reduced to three candidate parameters using K-means clustering of a reduced subset of the screened Hubbard parameter space that corresponds to the lowest 10 % of $J^{\text{FP}}_{\text{Predicted}}$.

From the three candidate Hubbard parameters, DFT+U using U = 3.342 eV, c_1 = 0.792 and c_2

= -0.506, which is henceforth referred to as "refined-DFT+U", provided the best accuracy in the predicted E_{bg} (4.18 eV), V_0 (97.38 Å^3) and ΔE_{Form} (-2.40 eV/atom) relative to experimental references (Tables 4.11 and 4.12).

Table 4.11: Geometric and energetic properties of bulk $LiCoO_2$, predicted using DFT+U (mBEEF exchange-correlation functional) and hybrid-DFT (PBE0 exchange-correlation functional), presented alongside experiment: band gap (E_{bg}, eV), unit cell equilibrium volume (V_0, Å^3) and formation energy (ΔE_{Form}, eV/atom). Rows 1–3 (4–6) correspond to DFT+U with a refined (atomic) Co 3d Hubbard projector.

Method	U (eV)	c_1	c_2	E_{bg} (eV)	V_0 (Å^3)	ΔE_{Form} (eV/atom)
DFT+U	3.342	0.792	-0.506	4.18	97.38	-2.40
DFT+U	4.074	0.818	-0.485	4.67	97.58	-2.75
DFT+U	4.731	0.834	-0.470	4.97	97.75	-3.09
DFT+U	3.342	1	0	3.93	92.26	-4.67
DFT+U	4.074	1	0	4.05	90.49	-5.29
DFT+U	4.731	1	0	4.05	88.55	-5.87
Hybrid-DFT	N/A	N/A	N/A	5.02	95.01	-2.13
Experiment	N/A	N/A	N/A	2.70 [73]	96.48 [74, 75]	-1.76 [76]

Table 4.12: Electronic properties of bulk $LiCoO_2$, predicted using DFT+U (mBEEF exchange-correlation functional) and hybrid-DFT (PBE0 exchange-correlation functional), presented alongside experiment: Co 3d occupation matrix trace ($\mathrm{Tr}[\mathbf{n}(\mathrm{Co}\ 3d)]$), O 2$p$ occupation matrix trace ($\mathrm{Tr}[\mathbf{n}(\mathrm{O}\ 2p)]$) and valence band (VB) edge character. Rows 1–3 (4–6) correspond to DFT+U with a refined (atomic) Co 3d Hubbard projector.

Method	U (eV)	c_1	c_2	$\mathrm{Tr}[n(\mathrm{Co}\ 3d)]$	$\mathrm{Tr}[n(\mathrm{O}\ 2p)]$	VB Edge Character
DFT+U	3.342	0.792	-0.506	4.48	5.12	Co 3d and O 2p
DFT+U	4.074	0.818	-0.485	4.79	5.16	Co 3d and O 2p
DFT+U	4.731	0.834	-0.470	4.99	5.19	O 2p
DFT+U	3.342	1	0	7.33	5.09	O 2p
DFT+U	4.074	1	0	7.39	5.12	O 2p
DFT+U	4.731	1	0	7.43	5.15	O 2p
Hybrid-DFT	N/A	N/A	N/A	7.08	4.76	Co 3d and O 2p
Experiment	N/A	N/A	N/A	N/A	N/A	Co 3d and O 2p [77]

Comparing the DFT+U-predicted electronic structure of stoichiometric $LiCoO_2$ using refined-DFT+U and U = 3.342 eV, c_1 = 1 and c_2 = 0 (*i.e.*, the same Co 3d Hubbard U value with an atomic projector, which is henceforth referred to as "atomic-DFT+U"), projector refinement was found to be essential to predict the mixed Co 3d and O 2p valence band edge character of $LiCoO_2$ in Figure 4.17(a), which is also predicted using hybrid-DFT and reported experimentally. [77] Conversely, DFT+U using atomic-DFT+U, or a much larger Hubbard U value than in Figure 4.17(a), both predict the $LiCoO_2$ valence band edge to be dominated by O 2p states, as illustrated in Figure 4.17(b).

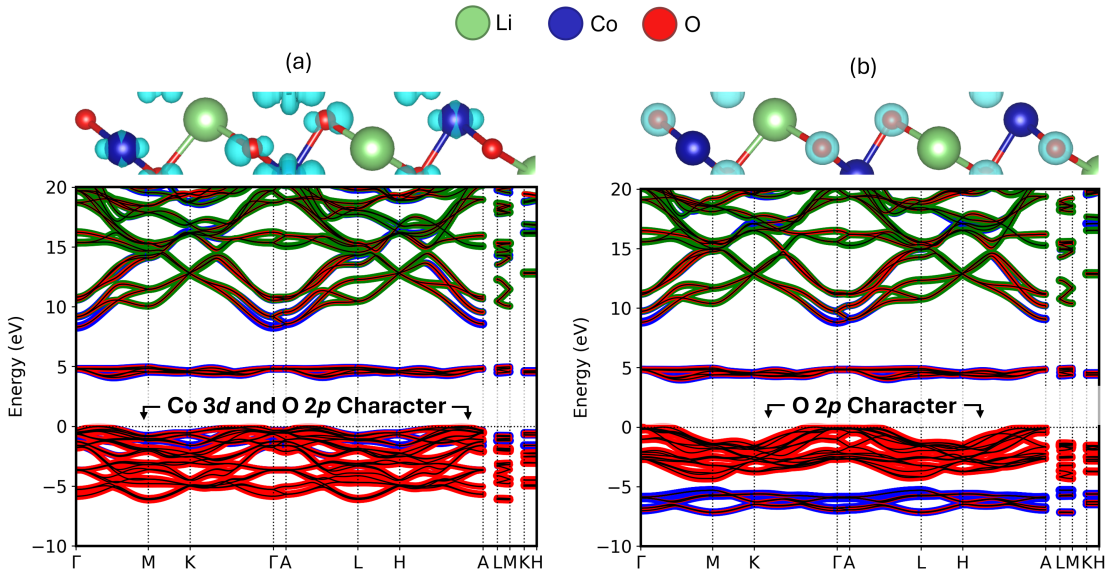


Figure 4.17: Charge density isosurface at the 0.05 eÅ^{-3} level for the eigenstate corresponding to the HOMO and the corresponding Mulliken-projected band structure for stoichiometric $LiCoO_2$ along the high-symmetry **k**-point path Γ–M–K–Γ–A–L–H–A–L–M–K–H, calculated using (a) DFT+U with the mBEEF exchange-correlation functional, U = 3.342 eV, c_1 = 0.792 and c_2 = -0.506 and (b) DFT+U with the mBEEF exchange-correlation functional, U = 3.342 eV, c_1 = 1 and c_2 = 0. Marker sizes and colours in the band structure plots correspond to the relative contribution for that species to the band. The valence band edge character is either (a) a mixture of Co 3d and O 2p states or (b) dominated by O 2p states. The band structure plots are centred with respect to the Fermi level.

The defective $LiCoO_2$ bulk containing $Mg_{Co}^{\times}$ was simulated using both refined-DFT+U and atomic-DFT+U, using the OMR method with the $3d_{z^2}$ orbital occupancy of the Co atom nearest the Mg dopant reduced by 1 to initialise Co^{4+}. In both simulations, SCF convergence and geometry optimisation completed successfully without termination or charge sloshing. Both simulations also predict the formation of shallow defect states at the top of the valence band in Figure 4.18(b), corresponding to a delocalised hole polaron that exists *hybridised* between Co 3d and O 2p states using refined-DFT+U in Figure 4.18(d), compared to *non-hybridised* across only O atoms using atomic-DFT+U in Figure 4.18(e). These differences in the DFT+U-predicted character of the hole polaron also directly affect their stability, with ΔE_{Defect} = 1.48 eV using refined-DFT+U and ΔE_{Defect} = 10.89 eV using atomic-DFT+U. Given that compositions with a 1:1 ratio of Co and Mg, *i.e.*, $LiCo_{0.5}Mg_{0.5}O_2$, have been synthesised experimentally, [78] DFT+U with the default atomic Co 3d Hubbard projector therefore incorrectly predicts the total insolubility of Mg in $LiCoO_2$.

Both refined- and atomic-DFT+U were further used to simulate a bulk oxygen vacancy in Mg-doped $LiCoO_2$ using the OMR method. Here, the nearest O atom to the Mg dopant was removed and the two nearest Co atoms to the vacancy were initialised as Co^{2+} by increasing their $3d_{yz}$ orbital occupancies by 1. With the atomic Co 3d Hubbard projector, the simulation did not converge due to charge sloshing, *i.e.*, oscillations in the charge density), similar to TiO_2 in Section 4.3.1. Conversely, DFT+U using the refined Co 3d Hubbard projector successfully converged without termination or charge sloshing, predicting a deep defect state of $3d_{x^2-y^2}$ character and an associated oxygen vacancy formation energy of ΔE_{OV} = 3.61 eV. These results suggest that both Mg-doping and oxygen vacancy

formation could enhance the electrical conductivity of $LiCoO_2$ *via* the transport of delocalised hole and localised electron polarons, respectively, whilst the generalised approach for optimising the Co 3*d* Hubbard *U* value and projector is robust with respect to numerical instability that is often observed when using the default atomic Hubbard projector. These results give promise to the development of the first-principles approach by extending the size and diversity of the training set used for generalised regression and classification.

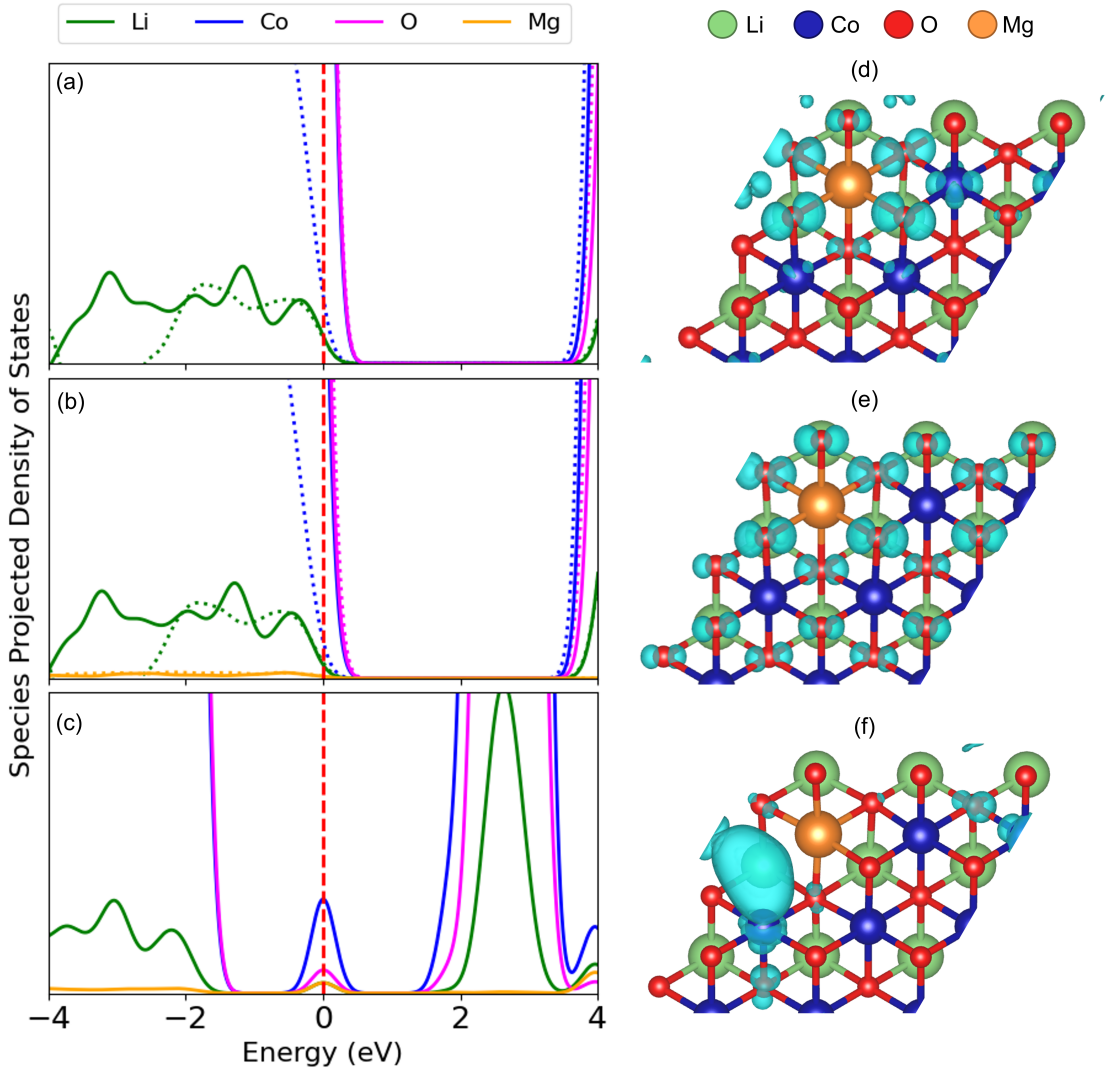


Figure 4.18: Elemental species projected density of states for (a) stoichiometric $LiCoO_2$, (b) defective $LiCoO_2$ containing $Mg_{Co}^{\times}$ and (c) defective $LiCoO_2$ containing both $Mg_{Co}^{\times}$ and $V_O^{\times}$, calculated using DFT+*U* with the mBEEF exchange-correlation functional, *U* = 3.342 eV, c_1 = 0.792 and c_2 = -0.506 (refined-DFT+*U*, solid lines) and DFT+*U* with the mBEEF exchange-correlation functional, *U* = 3.342 eV, c_1 = 1 and c_2 = 0 (atomic-DFT+*U*, dotted lines). All plots are relative to the Fermi level (red dashed line). The corresponding charge density isosurfaces at the 0.025 eÅ^{-3} level for the eigenstate corresponding to the HOMO are shown for defective $LiCoO_2$ containing $Mg_{Co}^{\times}$, calculated using (d) refined-DFT+*U* and (e) atomic-DFT+*U*, as well as (f) defective $LiCoO_2$ containing both $Mg_{Co}^{\times}$ and $V_O^{\times}$, calculated using refined-DFT+*U*.

4.4 Conclusions

To navigate the numerical instability of self-consistent DFT+U simulations of TMOs and REOs in a NAO framework, including the prediction of erroneous metallic ground states and the termination or non-convergence of point defect calculations, it is essential to carefully define an appropriate Hubbard projector. Simultaneous optimisation of the Hubbard U value and projector has been demonstrated semi-empirically using SR and SVMs, before using BO to minimise the errors of target properties relative to experimental references. The Ti 3d Hubbard U value and projector have been semi-empirically refined to enable self-consistent DFT+U simulations of intrinsic and extrinsic defects in both anatase and rutile TiO_2. The outcome is DFT+U simulations with comparable accuracy to hybrid-DFT in terms of the relative stabilities of point defects and the formation of localised Holstein polarons, but at orders of magnitude lower cost. The DFT+U-predicted occupation matrices reproduce the hybrid-DFT O 2p occupation matrix, which can be an effective cost function for a first-principles strategy for Hubbard U value and projector optimisation.

The semi-empirical approach was therefore defined as a first-principles approach and generalised across materials using hierarchical SR to screen the Hubbard parameter space using empirical correlations to learn the DFT+U potential energy surface in terms of orbital occupancies. Predictions of metal d or f orbital and O 2p orbital occupancies were made in terms of Hubbard parameters, basis set parameters, DFT-predicted orbital occupancies and atomic material descriptors. The first-principles approach enables the development of a generalised workflow for the one-shot computation of Hubbard U values and projectors for different materials that requires no successive DFT+U calculations, as in active learning schemes. The method transferability is demonstrated for 10 prototypical TMOs and REOs (anatase and rutile TiO_2, Cu_2O, MoO_3, WO_3, Y_2O_3, ZrO_2, CeO_2, $LiCoO_2$ and $LiFePO_4$), which each require one reference DFT and hybrid-DFT calculation as inputs, whilst generating families of solutions for each material, *i.e.*, optimised Hubbard projectors for a given Hubbard U value. Upon validating a subset of these solutions, a MAE of 5.02% for the DFT+U-predicted O 2p orbital occupancies was achieved, with demonstrated accuracy for materials unseen from model training ($LiCoO_2$ and $LiFePO_4$).

Predicting the numerical stability of point defect calculations can also be generalised across materials using symbolic classification, using Hubbard U values and material-dependent descriptors of covalency, enabling the determination of Hubbard U values and projectors that are robust against numerical instability. The validity of Hubbard U values and projectors determined from first-principles has been investigated for the self-consistent simulation of Mg-doped and oxygen deficient $LiCoO_2$, where refining the Co 3d Hubbard projector enables the numerically stable simulation of experimentally reported charge compensation mechanisms driving the material's high electrical conductivity. The same results were not possible using an atomic Co 3d Hubbard projector and did not require any prior testing of suitable Co 3d Hubbard U values or projectors, which gives promise for the development of a foundational tool for simultaneously determining multiple Hubbard parameters in a NAO framework and beyond. The work demonstrates how advanced machine learning algorithms can aid the development of inexpensive and transferable workflows for DFT+U parameterisation, achieving extrapolative accuracy beyond the limits of small training sets, for more accurate and efficient simulations of complex catalytic materials.

References

(1) A. Chaudhari, K. Agrawal and A. J. Logsdail, Machine learning generalised DFT+*U* projectors in a numerical atom-centred orbital framework, *Digit. Discov.* 2025, **4** 3701–3727.

(2) A. Chaudhari, A. J. Logsdail and A. Folli, Polymorph-Induced Reducibility and Electron Trapping Energetics of Nb and W Dopants in TiO_2, *J. Phys. Chem. C* 2025, **129** 15453–15461.

(3) O. Y. Long, G. Sai Gautam and E. A. Carter, Evaluating optimal *U* for 3*d* transition-metal oxides within the SCAN+*U* framework, *Phys. Rev. Mater.* 2020, **4** 045401.

(4) L. Wang, T. Maxisch and G. Ceder, Oxidation energies of transition metal oxides within the GGA + *U* framework, *Phys. Rev. B* 2006, **73** 195107.

(5) M. Cococcioni and S. de Gironcoli, Linear response approach to the calculation of the effective interaction parameters in the LDA + *U* method, *Phys. Rev. B* 2005, **71** 035105.

(6) I. Timrov, N. Marzari and M. Cococcioni, Hubbard parameters from density-functional perturbation theory, *Phys. Rev. B* 2018, **98** 085127.

(7) M. Springer and F. Aryasetiawan, Frequency-dependent screened interaction in Ni within the random-phase approximation, *Phys. Rev. B* 1998, **57** 4364–4368.

(8) N. J. Mosey, P. Liao and E. A. Carter, Rotationally invariant *ab initio* evaluation of Coulomb and exchange parameters for DFT+*U* calculations, *J. Chem. Phys.* 2008, **129** 014103.

(9) L. A. Agapito, S. Curtarolo and M. Buongiorno Nardelli, Reformulation of DFT + *U* as a Pseudohybrid Hubbard Density Functional for Accelerated Materials Discovery, *Phys. Rev. X* 2015, **5** 011006.

(10) G. C. Moore, M. K. Horton, E. Linscott, A. M. Ganose, M. Siron, D. D. O'Regan and K. A. Persson, High-throughput determination of Hubbard *U* and Hund *J* values for transition metal oxides via the linear response formalism, *Phys. Rev. Mater.* 2024, **8** 014409.

(11) K. Yu and E. A. Carter, Communication: Comparing *ab initio* methods of obtaining effective *U* parameters for closed-shell materials, *J. Chem. Phys.* 2014, **140** 121105.

(12) E. Macke, I. Timrov, N. Marzari and L. C. Ciacchi, Orbital-Resolved DFT+*U* for Molecules and Solids, *J. Chem. Theory Comput.* 2024, **20** 4824–4843.

(13) I. Timrov, N. Marzari and M. Cococcioni, Self-consistent Hubbard parameters from density-functional perturbation theory in the ultrasoft and projector-augmented wave formulations, *Phys. Rev. B* 2021, **103** 045141.

(14) C. Ricca, I. Timrov, M. Cococcioni, N. Marzari and U. Aschauer, Self-consistent site-dependent DFT+*U* study of stoichiometric and defective $SrMnO_3$, *Phys. Rev. B* 2019, **99** 094102.

(15) J. A. Santana, J. Kim, P. R. C. Kent and F. A. Reboredo, Successes and failures of Hubbard-corrected density functional theory: The case of Mg doped $LiCoO_2$, *J. Chem. Phys.* 2014, **141** 164706.

(16) M. Yu, S. Yang, C. Wu and N. Marom, Machine learning the Hubbard U parameter in DFT+ U using Bayesian optimization, *npj Comput. Mater.* 2020, **6** 180.

(17) W. Yu, Z. Zhang, X. Wan, H. Guo, Q. Gui, Y. Peng, Y. Li, W. Fu, D. Lu, Y. Ye and Y. Guo, Active Learning the High-Dimensional Transferable Hubbard U and V Parameters in the DFT+U+V$Scheme$, *J. Chem. Theory Comput.* 2023, **19** 6425–6433.

(18) P. Tavadze, R. Boucher, G. Avendaño-Franco, K. X. Kocan, S. Singh, V. Dovale-Farelo, W. Ibarra-Hernández, M. B. Johnson, D. S. Mebane and A. H. Romero, Exploring DFT+U parameter space with a Bayesian calibration assisted by Markov chain Monte Carlo sampling, *npj Comput. Mater.* 2021, **7** 182.

(19) G. Cai, Z. Cao, F. Xie, H. Jia, W. Liu, Y. Wang, F. Liu, X. Ren, S. Meng and M. Liu, Predicting structure-dependent Hubbard U parameters via machine learning, *Mater. Futures* 2024, **3** 025601.

(20) M. Uhrin, A. Zadoks, L. Binci, N. Marzari and I. Timrov, Machine learning Hubbard parameters with equivariant neural networks, *npj Comput. Mater.* 2025, **11** 19.

(21) L. Foppa, T. A. R. Purcell, S. V. Levchenko, M. Scheffler and L. M. Ghiringhelli, Hierarchical Symbolic Regression for Identifying Key Physical Parameters Correlated with Bulk Properties of Perovskites, *Phys. Rev. Lett.* 2022, **129** 055301.

(22) V. Blum, R. Gehrke, F. Hanke, P. Havu, V. Havu, X. Ren, K. Reuter and M. Scheffler, *Ab initio* molecular simulations with numeric atom-centered orbitals, *Comput. Phys. Commun.* 2009, **180** 2175–2196.

(23) F. Birch, Finite Elastic Strain of Cubic Crystals, *Phys. Rev.* 1947, **71** 809–824.

(24) M. Kick, K. Reuter and H. Oberhofer, Intricacies of DFT+U, Not Only in a Numeric Atom Centered Orbital Framework, *J. Chem. Theory Comput.* 2019, **15** 1705–1718.

(25) J. P. Allen and G. W. Watson, Occupation matrix control of d- and f-electron localisations using DFT+U, *Phys. Chem. Chem. Phys.* 2014, **16** 21016–21031.

(26) B. Dorado, B. Amadon, M. Freyss and M. Bertolus, DFT + U calculations of the ground state and metastable states of uranium dioxide, *Phys. Rev. B* 2009, **79** 235125.

(27) R. Ouyang, S. Curtarolo, E. Ahmetcik, M. Scheffler and L. M. Ghiringhelli, SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates, *Phys. Rev. Mater.* 2018, **2** 083802.

(28) T. A. R. Purcell, M. Scheffler, C. Carbogno and L. M. Ghiringhelli, SISSO++: A C++ Implementation of the Sure-Independence Screening and Sparsifying Operator Approach, *JOSS* 2022, **7** 3960.

(29) T. A. R. Purcell, M. Scheffler and L. M. Ghiringhelli, Recent advances in the SISSO method and their implementation in the SISSO++ code, *J. Chem. Phys.* 2023, **159** 114110.

(30) Y. Zhang, J. W. Furness, B. Xiao and J. Sun, Subtlety of TiO_2 phase stability: Reliability of the density functional theory predictions and persistence of the self-interaction error, *J. Chem. Phys.* 2019, **150** 014105.

(31) T. Arlt, M. Bermejo, M. A. Blanco, L. Gerward, J. Z. Jiang, J. Staun Olsen and J. M. Recio, High-pressure polymorphs of anatase TiO_2, *Phys. Rev. B* 2000, **61** 14414–14419.

(32) M. Setvin, C. Franchini, X. Hao, M. Schmid, A. Janotti, M. Kaltak, C. G. Van de Walle, G. Kresse and U. Diebold, Direct View at Excess Electrons in TiO_2 Rutile and Anatase, *Phys. Rev. Lett.* 2014, **113** 086402.

(33) F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, Scikit-learn: Machine Learning in Python, *JMLR* 2011, **12** 2825–2830.

(34) The GPyOpt authors, *GPyOpt: A Bayesian Optimization framework in Python*, `http://github.com/SheffieldML/GPyOpt`, Accessed June 2024, 2016.

(35) D. R. Jones, M. Schonlau and W. J. Welch, Efficient Global Optimization of Expensive Black-Box Functions, *J. Glob. Optim.* 1998, **13** 455–492.

(36) R. J. B. M. D. Mckay and W. J. Conover, A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output From a Computer Code, *Technometrics* 2000, **42** 55–61.

(37) S. Dalton and D. Richardson, *pyDOE: The experimental design package for Python*, `https://pythonhosted.org/pyDOE/`, Accessed June 2024, 2014.

(38) R. Ramakrishnan, P. O. Dral, M. Rupp and O. A. von Lilienfeld, Big Data Meets Quantum Chemistry Approximations: The -Machine Learning Approach, *J. Chem. Theory. Comput.* 2015, **11** 2087–2096.

(39) D.-H. Seo, A. Urban and G. Ceder, Calibrating transition-metal energy levels and oxygen bands in first-principles calculations: Accurate prediction of redox potentials and charge transfer in lithium transition-metal oxides, *Phys. Rev. B* 2015, **92** 115118.

(40) P. Mahalanobis, On the Generalised Distance in Statistics, *Sankhya A* 1936, **80** 1–7.

(41) V. I. Anisimov, F. Aryasetiawan and A. I. Lichtenstein, First-principles calculations of the electronic structure and spectra of strongly correlated systems: the LDA+*U* method, *J. Phys. Condens. Matter.* 1997, **9** 767–808.

(42) B. Himmetoglu, A. Floris, S. de Gironcoli and M. Cococcioni, Hubbard-corrected DFT energy functionals: The LDA+*U* description of correlated systems, *Int. J. Quantum Chem.* 2014, **114** 14–49.

(43) T. Schäfer, N. Daelman and N. López, Cerium Oxides without *U*: The Role of Many-Electron Correlation, *J. Phys. Chem. Lett.* 2021, **12** 6277–6283.

(44) R. song Li, D. qiang Xin, S. qi Huang, Z. jian Wang, L. Huang and X. hua Zhou, A full potential all-electron calculation on electronic structure of NiO, *Chin. J. Phys.* 2018, **56** 2829–2836.

(45) G. Gebreyesus, L. Bastonero, M. Kotiuga, N. Marzari and I. Timrov, Understanding the role of Hubbard corrections in the rhombohedral phase of $BaTiO_3$, *Phys. Rev. B* 2023, **108** 235171.

(46) T.-c. Liu, D. Gaines, H. Kim, A. Salgado-Casanova, S. B. Torrisi and C. Wolverton, Anomalous reversal of stability in Mo-containing oxides: A difficult case exhibiting sensitivity to DFT + U and distortion, *Phys. Rev. Mater.* 2025, **9** 055402.

(47) S. Phoka, P. Laokul, E. Swatsitang, V. Promarak, S. Seraphin and S. Maensiri, Synthesis, structural and optical properties of CeO_2 nanoparticles synthesized by a simple polyvinyl pyrrolidone (PVP) solution route, *Mater. Chem. Phys.* 2009, **115** 423–428.

(48) F. B. Baker, E. J. Huber, C. E. Holley and N. Krikorian, Enthalpies of formation of cerium dioxide, cerium sesquicarbide, and cerium dicarbide, *J. Chem. Thermodyn.* 1971, **3** 77–83.

(49) M. Kick, C. Grosu, M. Schuderer, C. Scheurer and H. Oberhofer, Mobile Small Polarons Qualitatively Explain Conductivity in Lithium Titanium Oxide Battery Electrodes, *J. Phys. Chem. Lett.* 2020, **11** 2535–2540.

(50) M. Kick, C. Scheurer and H. Oberhofer, Formation and stability of small polarons at the lithium-terminated $Li_4Ti_5O_{12}$ (LTO) (111) surface, *J. Chem. Phys.* 2020, **153** 144701.

(51) M. Arrigoni and G. K. H. Madsen, A comparative first-principles investigation on the defect chemistry of TiO_2 anatase, *J. Chem. Phys.* 2020, **152** 044110.

(52) J. Aarons, M. Sarwar, D. Thompsett and C.-K. Skylaris, Perspective: Methods for large-scale density functional calculations on metallic systems, *J. Chem. Phys.* 2016, **145** 220901.

(53) T. D. Pham and N. A. Deskins, Efficient Method for Modeling Polarons Using Electronic Structure Methods, *J. Chem. Theory Comput* 2020, **16** 5264–5278.

(54) R. Zhang, A. Chutia, A. A. Sokol, D. Chadwick and C. R. A. Catlow, A computational investigation of the adsorption of small copper clusters on the CeO_2(110) surface, *Phys. Chem. Chem. Phys.* 2021, **23** 19329–19342.

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

(56) K. Warda, E. Macke, I. Timrov, L. C. Ciacchi and P. M. Kowalski, Getting the manifold right: The crucial role of orbital resolution in DFT+ U for mixed df electron compounds, *arXiv preprint: 2508.16435* 2025.

(57) K. Jakob and H. Oberhofer, "Self-Consistency in the Hubbard-Corrected DFT+U Method", Master's thesis, Faculty of Chemistry, Technical University of Munich, 2021.

(58) M. Reticcioli, U. Diebold and C. Franchini, Modeling polarons in density functional theory: lessons learned from TiO_2, *J. Condens. Matter Phys.* 2022, **34** 204006.

(59) C. Spreafico and J. VandeVondele, The nature of excess electrons in anatase and rutile from hybrid DFT and RPA, *Phys. Chem. Chem. Phys.* 2014, **16** 26144–26152.

(60) C. Bigi, Z. Tang, G. M. Pierantozzi, P. Orgiani, P. K. Das, J. Fujii, I. Vobornik, T. Pincelli, A. Troglia, T.-L. Lee, R. Ciancio, G. Drazic, A. Verdini, A. Regoutz, P. D. C. King, D. Biswas, G. Rossi, G. Panaccione and A. Selloni, Distinct behavior of localized and delocalized carriers in anatase TiO_2 (001) during reaction with O_2, *Phys. Rev. Mater.* 2020, **4** 025801.

(61) P. Nandi, S. Shin, H. Park, Y. In, U. Amornkitbamrung, H. J. Jeong, S. J. Kwon and H. Shin, Large and Small Polarons in Highly Efficient and Stable Organic-Inorganic Lead Halide Perovskite Solar Cells: A Review, *Sol. RRL* 2024, **8** 2400364.

(62) P. Gono, J. Wiktor, F. Ambrosio and A. Pasquarello, Surface Polarons Reducing Overpotentials in the Oxygen Evolution Reaction, *ACS Catal.* 2018, **8** 5847–5851.

(63) Y. Huang, Y. Zhu, H. Fu, M. Ou, C. Hu, S. Yu, Z. Hu, C.-T. Chen, G. Jiang, H. Gu, H. Lin, W. Luo and Y. Huang, Mg-Pillared $LiCoO_2$: Towards Stable Cycling at 4.6V, *Angew. Chem. Int. Ed.* 2021, **60** 4682–4688.

(64) J. Xia, N. Zhang, D. Yi, F. Lu, Y. Yang, X. Wang and Y. Wang, Stabilizing 4.6 V $LiCoO_2$ via Er and Mg Trace Doping at Li-Site and Co-Site Respectively, *Small* 2024, **20** 2311578.

(65) M. Mladenov, R. Stoyanova, E. Zhecheva and S. Vassilev, Effect of Mg doping and MgO-surface modification on the cycling stability of $LiCoO_2$ electrodes, *Electrochem. Commun.* 2001, **3** 410–416.

(66) S. Levasseur, M. Ménétrier and C. Delmas, On the Dual Effect of Mg Doping in $LiCoO_2$ and $Li_1{+}\delta CoO_2$: Structural, Electronic Properties, and 7Li MAS NMR Studies, *Chem. Mater.* 2002, **14** 3584–3590.

(67) S. Shi, C. Ouyang, M. Lei and W. Tang, Effect of Mg-doping on the structural and electronic properties of $LiCoO_2$: A first-principles investigation, *J. Power Sources.* 2007, **171** 908–912.

(68) X. G. Xu, C. Li, J. X. Li, U. Kolb, F. Wu and G. Chen, Electronic Structure of Li(Co, Mg)O_2 Studied by Electron Energy-Loss Spectrometry and First-Principles Calculation, *J. Phys. Chem. B.* 2003, **107** 11648–11651.

(69) H. Tukamoto and A. R. West, Electronic Conductivity of $LiCoO_2$ and Its Enhancement by Magnesium Doping, *J. Electrochem. Soc.* 1997, **144** 3164.

(70) W. Luo, X. Li and J. R. Dahn, Synthesis and Characterization of Mg Substituted $LiCoO_2$, *J. Electrochem. Soc.* 2010, **157** A782.

(71) F. Zhou, M. Cococcioni, C. A. Marianetti, D. Morgan and G. Ceder, First-principles prediction of redox potentials in transition-metal compounds with LDA+U, *Phys. Rev. B* 2004, **70** 235121.

(72) K. Hoang and M. D. Johannes, Defect chemistry in layered transition-metal oxides from screened hybrid density functional calculations, *J. Mater. Chem. A* 2014, **2** 5224–5235.

(73) J. van Elp, J. L. Wieland, H. Eskes, P. Kuiper, G. A. Sawatzky, F. M. F. de Groot and T. S. Turner, Electronic structure of CoO, Li-doped CoO, and $LiCoO_2$, *Phys. Rev. B* 1991, **44** 6090–6103.

(74) X. Wang, I. Loa, K. Kunc, K. Syassen and M. Amboage, Effect of pressure on the structural properties and Raman modes of $LiCoO_2$, *Phys. Rev. B* 2005, **72** 224102.

(75) J. Akimoto, Y. Gotoh and Y. Oosawa, Synthesis and Structure Refinement of $LiCoO_2$ Single Crystals, *J. Solid State Chem.* 1998, **141** 298–302.

(76) M. Wang and A. Navrotsky, Enthalpy of formation of $LiNiO_2$, $LiCoO_2$ and their solid solution, $LiNi_{1x}Co_xO_2$, *Solid State Ion.* 2004, **166** 167–173.

(77) L. Dahéron, H. Martinez, R. Dedryvère, I. Baraille, M. Ménétrier, C. Denage, C. Delmas and D. Gonbeau, Surface Properties of $LiCoO_2$ Investigated by XPS Analyses and Theoretical Calculations, *J. Phys. Chem. C.* 2009, **113** 5843–5852.

(78) R. Sathiyamoorthi, P. Shakkthivel, R. Gangadharan and T. Vasudevan, Layered $LiCo_{1-x}Mg_xO_2$ (x = 0.0, 0.1, 0.2, 0.3 and 0.5) cathode materials for lithium-ion rechargeable batteries, *Ionics* 2007, **13** 25–33.

Chapter 5

Ab Initio Insights into Support-Induced Sulfur Resistance of Ni-Based Reforming Catalysts

This chapter is based on the published work *Ab initio insights into support-induced sulfur resistance of Ni-based reforming catalysts* in *Catalysis Science & Technology*, which is co-authored by Dr Pavel Stishenko (Cardiff University, CU), Akash Hiregange (CU), Dr Christopher Hawkins (Johnson Matthey, JM), Dr Misbah Sarwar (JM), Dr Stephen Poulston (JM) and Dr Andrew Logsdail (CU). [1] The experimental testing and characterisation was assisted by Dr Andrew Steele (JM), Dr Gregory Goodlet (JM), Dr Riho Green (JM) and Jason Raymond (JM).

This work builds upon Chapters 3 and 4, which introduced the challenges and solutions associated with modelling defects in transition metal and rare-earth metal oxides with experimental accuracy. In this chapter, these concepts are extended to examine how defect formation in the bulk phase of metal oxide supports influences the poisoning and regeneration of supported Ni nanoparticle catalysts. I performed the electronic structure calculations and Dr Pavel Stishenko performed the Monte Carlo simulations. Akash Hiregange and I fine-tuned and inferenced the machine learned interatomic potential. Dr Christopher Hawkins and Dr Andrew Steele performed the experimental testing. Dr Gregory Goodlet, Dr Riho Green and Jason Raymond from the Advanced Characterisation Department at Johnson Matthey Technology Centre performed the SEM, XPS and ICP analysis, respectively. All input/output files for electronic structure calculations, Monte Carlo sampling and MACE fine-tuning are available open-source in the GitHub repository *https://github.com/amitmc1/GCMC-Adlayers* and as a supplementary dataset on Figshare at the DOI: *https://doi.org/10.6084/m9.figshare.29562377*.

5.1 Introduction

Methane steam reforming (MSR) is an established industrial process that produces $\sim$ 95% of the global H_2 supply [2] *via* the conversion of natural gas (primarily CH_4, with smaller amounts of higher hydrocarbons) to syngas (mixtures of CO, CO_2 and H_2), at high temperature and pressure, in the presence of a catalyst. The commercial Ni-based catalysts are highly susceptible to sulfur poisoning by impurities in the feedstock, *e.g.*, H_2S, SO_2, H_2SO_4 and/or COS, and therefore an expensive feed desulfurisation process is necessary to achieve sub-ppm sulfur concentrations. [3] The additional

cost and complexity of feed desulfurisation also limits the development of biogas reforming processes for scalable H_2 production from renewable feedstocks, *e.g.*, using solid oxide fuel cells [4] or *via* combined steam and dry reforming. [5] Understanding the factors that affect the catalyst sulfur tolerance is essential to enable the direct use of sulfur-containing feedstocks; a challenge that is particularly important for Ni-based catalysts as they are more economically viable than those based on platinum group metals (PGMs).

A number of strategies have been considered to enhance the sulfur tolerance of Ni-based catalysts, such as alloying with PGMs, including Au, Cu, Mn, Pd, Pt and Rh. [6] Alloys are widely reported in the literature and are proposed to enhance the catalyst sulfur tolerance *via* different mechanisms, *e.g.*, promoting sulfur scavenging by secondary metallic active phases, [7] promoting sulfur oxidation and desorption at high temperatures [8, 9] and suppressing the dissociative adsorption of feedstock poisons like H_2S. [10] The optimisation of metal oxide supports is another effective strategy to enhance the sulfur tolerance of supported Ni nanoparticles during catalytic reforming reactions, with the mechanism widely hypothesised to involve oxygen buffering from reducible supports like CeO_2 and Y_2O_3. [11, 12] In these materials, lattice oxygen is proposed to migrate from the support to the Ni active phase under reducing conditions at high temperatures, resulting in the oxidation and desorption of catalyst poisons *e.g.*, $C \rightarrow CO_2$ [13–17] and $S \rightarrow SO_2$. [18–21] Similarly, a number of established chemical and electrochemical regeneration methods have been shown to restore the activity of poisoned Ni catalysts by modulating the transfer of oxygen to the poisoned Ni active sites. Chemical regeneration of sulfur-poisoned Ni catalysts can be achieved using exposure in steam, H_2 and/or O_2 depending on the degree of sulfur poisoning. [22, 23] Electrochemical regeneration can also be used to control the O^{2-} spillover from both aqueous environments; and Y_2O_3-stabilised ZrO_2 (YSZ) supports, towards sulfur poisoned Pt and Ni species, enabling catalyst oxidative regeneration using a negative electrode potential. [24–26]

Ab initio computational modelling methods, such as density functional theory (DFT), provide an atomic-level insight into the surface chemistry of sulfur-poisoned Ni nanoparticles. Atomic sulfur is often used to represent H_2S poisoning at low/medium surface coverage (θ_S) due to the predicted dissociative adsorption of $H_2S \rightarrow S$ on Ni(111), which does not cause surface reconstruction or sulfur penetration into the Ni bulk as observed at high θ_S. [27–30] DFT studies of oxygen-mediated sulfur removal from Ni(111) show that both atomic O and molecular O_2 (which adsorbs dissociatively) can lead to the sequential oxidation of $S \rightarrow SO \rightarrow SO_2$, which then desorbs at high temperatures. [31, 32] These studies were limited to idealised adlayer representations of S, with θ_S = 0.25 ML and 0.5 ML, and do not account for variations in configurational entropy at intermediate coverages; therefore, whether the formation of SO_2 is thermodynamically or kinetically driven at experimentally relevant surface coverages remains unresolved. To move beyond idealised models of surface adsorption, grand canonical Monte Carlo (GCMC) sampling can be used to sample the large configurational space of adsorption complexes on a lattice model of the surface, which would otherwise be computationally infeasible to sample with DFT alone (summarised in Section 2.5). Machine learned interatomic potentials offer a computationally tractable means to validate the predictions from GCMC, by capturing off-lattice effects such as many-body lateral interactions from DFT datasets, before being used to perform classical geometry optimisation calculations (summarised in Section 2.5).

Accurate simulations of poisoning and reactivity of Ni-based MSR catalysts are also very challenging to realise due to the interplay between oxygen buffering (causing catalyst regeneration) and phase transformations of the metal oxide support (causing catalyst deactivation). For example, Ni/γ-Al_2O_3 catalysts can undergo progressive Ni substitution for Al, resulting in the *in situ* transformation of Ni/γ-Al_2O_3 to spinel-type $NiAl_2O_4$. [33] Conflicting reports exist for the utility of Ni-based spinel-type oxides and whether they deactivate Ni-based catalysts [34] or enhance catalytic activity [35–40] and tolerance to S and C poisons [41] due to the facile formation of oxygen vacancies. Accurate predictions of the energetics of oxygen vacancy formation and substitutional doping for these support materials are non-trivial using DFT, particularly for reducible transition metal oxides (TMOs) *e.g.*, TiO_2, and rare-earth metal oxides (REOs) *e.g.*, CeO_2, which are experimentally reported to exhibit favourable oxygen buffering capacities. [42, 43] As discussed in Section 2.3.2, the SIE of local and semi-local DFT, results in erroneous defect formation energies in TMOs and REOs; [44–46] therefore, it is necessary to use "*beyond-DFT*" methods such as DFT+U to combat the SIE. As discussed in Chapters 3 and 4, the determination of appropriate simulation parameters, including the Hubbard U value and projector, is non-trivial for simulating defects in TMOs and REOs with accuracy that matches experimental observations, and care is therefore necessary in application. [47, 48]

In this chapter, a combined computational and experimental approach is adopted to investigate the enhanced sulfur tolerance of Ni nanoparticles on reducible metal oxide supports, with the aim of establishing strategies for future catalyst optimisation. The thermodynamic driving force for oxygen-mediated sulfur removal from Ni(111) is investigated using grand canonical Monte Carlo (GCMC) sampling of a DFT-parameterised lattice model, thereby providing insights into the regenerative effects of support oxygen buffering. The GCMC-predicted adlayers enable the prediction of the surface coverage and composition of competitively adsorbed S and O atoms as a function of temperature and the chemical potentials of S and O across an extended Ni(111) surface. The GCMC-predicted adlayers are validated using geometry optimisation simulations with a fine-tuned MACE MLIP to reveal entropic contributions and limitations to catalyst regeneration at experimentally relevant surface coverages. Simulations of the surface chemistry of Ni(111) are complemented by DFT+U predictions of the energetics of bulk defect formation (oxygen vacancies and Ni substitution) in prototypical metal oxide support materials, providing insights into the proclivity for oxygen release and phase transformation during catalytic reactions. The computational modelling is correlated with experimental characterisation (TPD-MS, XPS, ICP) and MSR activity testing for H_2S-poisoned Ni nanoparticle catalysts to rationalise the experimentally observed differences in the catalyst sulfur tolerance. The work demonstrates the integration of *ab initio* computational modelling, statistical sampling and machine learning to construct more realistic models of complex catalytic materials, which further complement experimental characterisation to inform future strategies for catalyst rational design.

5.2 Methodology

5.2.1 Electronic Structure Calculations

DFT

All electronic structure calculations were performed using the Fritz-Haber Institute *ab initio* materials simulation (FHI-aims) software package, [49] which uses an all electron numerical atom-centred orbital (NAO) basis set, interfaced with the Python-based Atomic Simulation Environment (ASE). [50] Periodic boundary conditions were applied using converged **k**-point spacing with the standard light basis set (2020), with equivalent accuracy to the TZVP Gaussian-type orbital basis set, [51] as decided after benchmarking of the bulk Ni vacancy formation energy ($\Delta E_{\text{Ni Vac}}$) in a $3\times3\times3$ supercell, relative to an experimental reference (1.79 eV), as detailed in Figure A.8. [52] Relativistic effects were accounted for using the zeroth order regular approximation (ZORA) [49] as a scalar correction. The system charge and spin were set to zero, given the reported quenching of Ni(111) surface magnetic moments following oxygen adsorption [53] and the temperatures of MSR far exceeding the Curie temperature of Ni (631 K), only below which long-range magnetic order is observed. [54] The mBEEF meta-GGA exchange-correlation functional was used, [55, 56] as defined in Libxc, [57] providing the best accuracy compared to other local and semi-local functionals (detailed in Table A.3), which was determined by comparing the Euclidean norm of the percentage errors of the DFT-predicted $\Delta E_{\text{Ni Vac}}$, V_0 and cohesive energy (ΔE_{Coh}), relative to experimental references (1.79 eV, [52] 43.61421 Å^3 [58] and 4.48 eV/atom, [59] respectively). Meta-GGAs are further reported to provide desirable accuracy for modelling sulfur adsorption complexes on transition metal surfaces with accuracy that matches experimental observations. [60] Dispersion corrections were not explicitly included as sulfur and oxygen bind strongly to Ni(111) through short-range chemisorption, which are well described by the mBEEF density functional. [55] For such systems, long-range van der Waals interactions provide only minor contributions to adsorption energies, whilst any van der Waals correction may also be detrimental to the representation of the support material; therefore, no further dispersion corrections are included.

Self-consistent field (SCF) optimisation of the electronic structure was achieved using a convergence criteria of 1×10^{-6} eV for the change in total energy, 1×10^{-4} eV for the change in the sum of eigenvalues and 1×10^{-6} e a_0^{-3} for the change in charge density. Unit cell equilibrium volumes (V_0) were calculated by fitting to the Birch-Murnaghan equation of state using ASE. [61] Geometry optimisation was performed using the quasi-Newton BFGS algorithm [62–65] with a force convergence criteria of 0.01 eV/Å. The pristine Ni(111) surface was modelled using a six layer symmetric periodic slab, of which the bottom three layers were frozen to mimic the system bulk, which is in line with computational literature studying the adsorption of catalyst poisons on Ni(111). [66] Our periodic slab model yields a surface energy γ^{surf} of 1.73 Jm^{-2}, which is in reasonable agreement with experimental references (1.94 Jm^{-2}). [67] γ^{surf} is defined as: [68]

$$\gamma^{\text{surf}} = \gamma^{\text{cleave}} + \gamma^{\text{relax}} = \frac{E^{\text{Unrelaxed}}_{\text{Ni slab}} - N^{\text{Form}} \times E_{\text{Ni bulk}}}{2\times A} + \frac{E^{\text{Relaxed}}_{\text{Ni slab}} - E^{\text{Unrelaxed}}_{\text{Ni slab}}}{A} \tag{5.1}$$

where $E^{\text{Relaxed}}_{\text{Ni slab}}$ ($E^{\text{Unrelaxed}}_{\text{Ni slab}}$) denotes the energy of the geometry optimised (initial) Ni slab, $E_{\text{Ni bulk}}$

denotes the energy of the geometry optimised Ni bulk, N^{Form} denotes the number of formula units in the slab and A denotes the slab surface area. A 20 Å vacuum gap was used in the direction perpendicular to the surface to eliminate artificial interactions between periodic images. A dipole correction was applied to compensate for the inhomogeneous electric field arising from surface adsorption. Adsorption energies were calculated as:

$$\Delta E_{\text{Ads}} = E_{[\text{Ni(111)} + \text{Ads}]} - E_{\text{Ni(111)}} + \mu_{\text{Ads}} \tag{5.2}$$

where the chemical potential of the adsorbed species (μ_{Ads}) was calculated using the energies of isolated atomic S, atomic O, molecular SO and molecular SO_2.

DFT+U and Defect Calculations

All DFT+U calculations were performed with FHI-aims, using the on-site definition of the occupation matrix and the Fully Localised Limit (FLL) double counting correction.[69] A Hubbard correction was applied to treat the Coulomb self-interaction of Ti 3d orbital electrons in tetragonal rutile TiO_2 and Ce 4f orbital electrons in cubic CeO_2. No Hubbard correction was applied for the Ni dopants or for γ-Al_2O_3. Hubbard U values for Ti 3d and Ce 4f orbital electrons were chosen as $U^{\text{Ti } 3d} = 2.575$ eV and $U^{\text{Ce } 4f} = 2.653$ eV, which are both valid with a refined atomic-like Hubbard projector, as defined in Table 5.1.

Table 5.1: Parameterised Hubbard U values (eV), projector coefficients c_1 and c_2 and supercell sizes used in Chapter 5

Support Material	Corrected Orbital	Hubbard U (eV)	c_1	c_2	Supercell Size
γ-Al_2O_3	N/A	N/A	N/A	N/A	1×1×3
TiO_2	Ti 3d	2.575	0.752	-0.486	2×2×5
CeO_2	Ce 4f	2.653	0.561	-0.600	2×2×3

The DFT- and DFT+U-predicted band gap (E_{bg}), V_0 and formation energy (ΔE_{Form}) of each material are compared with experimental references in Table 5.2, where ΔE_{Form} is calculated using the energies of bulk Ti (in the hexagonal close packed, HCP, crystal structure), bulk Al and Ce (both in the cubic crystal structure) and an isolated O_2 molecule:

$$\Delta E_{\text{Form}} = E_{\text{MO}_i} - E_{\text{M}} - i \times E_{\text{O}_2} \tag{5.3}$$

where i denotes the stoichiometric coefficient for oxygen. Hubbard U values and projectors were simultaneously determined using the machine learning-based workflow in Chapter 4, with the target of reproducing the bulk material covalency as calculated using hybrid-DFT, which results in numerically stable self-consistent simulations of point defects. [48]

Table 5.2: DFT- and DFT+U-predicted geometric, electronic and energetic properties of bulk γ-Al_2O_3, rutile TiO_2 and CeO_2 versus experimental references. The Hubbard parameters for Ti 3d orbitals are U = 2.575 eV, c_1 = 0.752 and c_2 = -0.486, whilst those for Ce 4f orbitals are U = 2.653 eV, c_1 = 0.561 and c_2 = -0.600. No Hubbard correction is applied for γ-Al_2O_3 or Ni in this work.

Material	Method	E_{bg} (eV)	V_0 (Å^3)	ΔE_{Form} (eV/Atom)
γ-Al_2O_3	DFT	5.66	371.77	-3.22
γ-Al_2O_3	Experimental	7.20 [70]	371.12 [71]	-3.43 [72]
TiO_2	DFT+U	2.47	62.86	-3.00
TiO_2	Experimental	3.00 [73]	62.44 [74]	-3.26 [75]
CeO_2	DFT+U	2.38	159.95	-3.73
CeO_2	Experimental	3.20 [76]	158.43 [77]	-3.77 [78]

Defect calculations in γ-Al_2O_3, TiO_2 and CeO_2 were performed using the supercell sizes listed in Table 5.1, with suitable sizes to ensure a consistent defect concentration across the three systems whilst also accurately representing the dilute limit. Defect energies (ΔE_{Defect}) following substitution of a host metal atom (Al in γ-Al_2O_3, Ti in TiO_2 and Ce in CeO_2) with a Ni atom were calculated as:

$$\Delta E_{\text{Defect}} = E_{\text{Defective Bulk}} + \mu_{\text{Host}} - E_{\text{Stoichiometric Bulk}} - \mu_{\text{Dopant}} \tag{5.4}$$

where the chemical potentials μ_{Host} and μ_{Dopant} were calculated using the energy of bulk Ti (hexagonal close packed) as well as Al, Ce and Ni (all cubic). Oxygen vacancy formation energies (ΔE_{OV}) were calculated as:

$$\Delta E_{\text{OV}} = E_{\text{Defective Bulk}} + \mu_{\text{O}} - E_{\text{Stoichiometric Bulk}} \tag{5.5}$$

where the chemical potential μ_O was calculated using half the energy of an isolated O_2 molecule. Defect calculations in TiO_2 and CeO_2 were performed using the "*occupation matrix release*" (OMR) method to initialise Ti^{3+} and Ce^{3+} polarons at nearest neighbour atoms to the defect, before the DFT+U-predicted total energy (E^{DFT+U}) is pre-converged using fixed orbital occupancies until $\Delta E^{DFT+U} \leq 0.001$ eV and then all orbital occupancies are calculated self-consistently. [69]

5.2.2 Monte Carlo Sampling

All lattice modelling and Monte Carlo sampling was performed using the the Surface Science Modeling and Simulation Toolkit (SuSMoST) software package, [79] considering adsorption complexes of S, O, SO and their pairs, and the occupation of hollow HCP and hollow FCC active sites on Ni(111) motivated by our results in Sections 5.3.1 and 5.3.2. Full DFT geometry optimisation was performed for 70 symmetrically inequivalent pairs of adsorption complexes on either a 10×10 or 7×7 Ni(111) surface supercell within a 10 Å or 5 Å radial cutoff, respectively, as explained further in Section 5.3.2, before calculating the energy of lateral interactions, $\Delta E_{Lateral}$, using:

$$\Delta E^{s_1,s_2}_{\text{lateral}} = E^{s_1,s_2}_{x\text{-}x\,\text{Pair}} - E_{\text{Ni(111)}} - \left(E^{s_1}_x + E^{s_2}_x\right) \tag{5.6}$$

where $E_{Ni(111)}$ is the energy of the pristine surface, $E^{s_1,s_2}_{x\text{-}x\,\text{Pair}}$ is the energy of a pair of adsorbates x at sites s_1 and s_2 for $x \in \{S, O\}$ and $s_1, s_2 \in \{\text{Hollow HCP}, \text{Hollow FCC}\}$, $E^{s_1}_x$ is the energy of a single

adsorbate x occupying site s_1 and $E_x^{s_2}$ is the energy of a single adsorbate x occupying site s_2. 35 adsorption complexes consisting of pairs of S-S, O-O and S-O atoms, with $|\Delta E_{\text{Lateral}}| \geq 0.04$ eV, were chosen for parameterising a pairwise Hamiltonian ($\mathscr{H}$) for subsequent GCMC sampling, based on the generalised lattice-gas model of adsorption monolayers by Akimenko *et al.*, [80] using:

$$\mathscr{H} = \sum_{i \in L} \Delta E_{\text{Ads}}(\sigma_i) + \sum_{i,j \in L} \Delta E_{\text{lateral}}(\sigma_i, \sigma_j, \mathbf{r}_{ij}) \tag{5.7}$$

where L is a set of lattice sites, σ_i is an adsorption complex at site i, $\Delta E_{\text{Ads}}(\sigma_i)$ is the adsorption energy of the adsorption complex at site i in the zero coverage limit and $\Delta E_{\text{lateral}}(\sigma_i, \sigma_j, \mathbf{r}_{ij})$ is the energy of lateral interactions between adsorption complexes at sites i and j, given the radius-vector ($\mathbf{r}_{ij}$) between the two sites. Geometry optimisation of S-O pairs with a short interatomic separation of 1.45 Å, corresponding to adsorption at neighbouring hollow HCP and hollow FCC active sites, resulted in atomic diffusion to other active sites, therefore these adsorption complexes were disregarded for subsequent GCMC sampling. Similarly, molecularly adsorbed SO was predicted to be less stable than individually adsorbed S and O atoms at low surface coverage, and therefore was not included in the GCMC sampling (see Section 5.3.2).

GCMC sampling was performed on a hexagonal lattice of 30×30 centers with periodic boundary conditions, which was large enough to avoid finite size effects. Each Monte Carlo step involved 30×30 attempted moves, *i.e.*, one attempt for each active site per step to change the state of the adsorbed layer through adsorption, desorption and surface diffusion of atomic S and O. The acceptance or rejection of a new configuration of the model adsorbed layer of S and O was determined using the Metropolis algorithm, [81] where a new configuration is accepted if the total energy ($\mathscr{H}$) is less than that of the previous configuration (*i.e.*, $\Delta\mathscr{H} \leq 0$ eV) or, if $\Delta\mathscr{H} > 0$ eV, the new configuration is accepted with the probability $\min\left\{1, \exp\left(-\frac{\Delta\mathscr{H}}{RT}\right)\right\}$. One million Monte Carlo steps were used to reach thermodynamic equilibrium and then the same number of steps were used to calculate ensemble averages. The parallel tempering algorithm was used to improve convergence to equilibrium and calculate the temperature dependence of the predicted adlayer coverage and composition, while also accounting for variations in configurational entropy. [82] The following temperatures were used for parallel tempering replicas: 300, 400, 600, 800, 1000, 1200, 1500 and 1700 K. Each simulation was performed with varying relative chemical potentials (μ^{R}) of sulfur ($\mu^{\text{R}}_{\text{S}}$) and oxygen ($\mu^{\text{R}}_{\text{O}}$) between -1 and 1 eV. $\mu^{\text{R}}_{\text{S}}$ ($\mu^{\text{R}}_{\text{O}}$) = 1 eV corresponds to the adsorption energies of a single S (O) atom on Ni(111) in the zero coverage limit, before geometry relaxation. Negative values of μ^{R} correspond to surfaces that are less likely to adsorb atoms in the zero coverage limit, whilst positive values of μ^{R} correspond to surfaces that are more likely to adsorb atoms in the zero coverage limit. We note that non-zero adsorbate coverages are still possible for both positive and negative values of μ^{R} after geometry relaxation, due to entropic effects or attractive lateral interactions. To enable direct comparison with experiment, the relative chemical potentials used for GCMC sampling were mapped to gas phase partial pressures, corresponding to reservoirs of O_2 and H_2S, using ideal gas thermodynamics at the same temperature and a standard-state pressure of 1 bar:

$$\mu^{\text{R}}_{\text{S}}(T, p) = \Delta E^{\text{S}}_{\text{Ads}} + \left[G_{\text{H}_2\text{S}}(T, p) - E_{\text{H}_2\text{S}}\right] - \left[G_{\text{H}_2}(T, p) - E_{\text{H}_2}\right] \tag{5.8}$$

$$\mu_{\mathrm{O}}^{\mathrm{R}}(T,p) = \Delta E_{\mathrm{Ads}}^{\mathrm{O}} + \frac{1}{2}\left[G_{\mathrm{O_2}}(T,p) - E_{\mathrm{O_2}}\right] \tag{5.9}$$

where $\Delta E_{\mathrm{Ads}}^{\mathrm{S}}$ ($\Delta E_{\mathrm{Ads}}^{\mathrm{O}}$) are the DFT-computed adsorption energies for a S (O) atom on Ni(111) in the zero-coverage limit, $G_{\mathrm{H_2S}}$, $G_{\mathrm{H_2}}$ and $G_{\mathrm{O_2}}$ are the Gibbs free energies of the isolated H_2S, H_2 and O_2 molecules, respectively, obtained from ideal gas thermochemistry using ASE, and $E_{\mathrm{H_2S}}$, $E_{\mathrm{H_2}}$ and $E_{\mathrm{O_2}}$ are the DFT-computed energies of the isolated H_2S, H_2 and O_2 molecules, respectively.

5.2.3 Many-Body Tensor Representations

To quantify the differences in the GCMC-predicted spatial distribution of adsorbed S and O on Ni(111), the GCMC-predicted adlayers were encoded into structural fingerprints using many-body tensor representations (MBTRs), [83] with the *DScribe* Python library. [84, 85] Two-body MBTRs were used to encode pairwise interatomic distances between adsorbed S and O atoms as a smooth density distribution over a continuous grid, defined over the range of 0 to 10 Å, with a Gaussian broadening parameter set to 0.1. An exponential weighting function was applied with a decay scale of 0.5, as well as a threshold of 10^{-3}, which acts as a cutoff for discarding small Gaussian contributions and therefore emphasise closer atomic interactions. No normalisation was applied to preserve the raw spatial distributions. The smooth density distribution was discretised into five equally spaced bins, yielding five two-body MBTR descriptors (D_i), before being reduced to a one-dimensional descriptor using principal component analysis (PCA) with the *Scikit-learn* Python library. [86] The principal component output from PCA ($\mathrm{PC^{MBTR}}$) is defined as:

$$\mathrm{PC^{MBTR}} = (0.4852 \times D_2) + (0.5003 \times D_3) + (0.5070 \times D_4) + (0.5071 \times D_5) \tag{5.10}$$

where D_i denotes the discretised MBTR descriptors, with $D_1 = 0$ as short-range S-O interactions are excluded from the GCMC sampling and therefore not present in the resulting adlayers. $\mathrm{PC^{MBTR}}$ therefore captures the most significant trends in the spatial disorder of co-adsorbed S and O.

5.2.4 Interatomic Potential Training and Inferencing

The GCMC predictions were validated using geometry optimisation calculations with a MACE (version 0.3.10) MLIP, [87] providing a computationally efficient means to relax the high-coverage GCMC-predicted adlayers on the 30×30 Ni(111) surface (~5800 atoms, surface area ~50 $\mathrm{nm^2}$). The MACE MLIP was trained using the diverse dataset of 5921 DFT-optimised structures collected in the work, including isolated atoms and molecules (S, O, SO, SO_2 and SO_3), Ni(111) periodic slab models of different thicknesses and adsorption complexes involving S, O, SO and SO_2 at both low and high surface coverage on Ni(111). Training was performed using multihead replay fine-tuning of the off-the-shelf MACE-MPA-0 (medium) foundation model, [88] trained on approximately 146,000 unique materials in the Material Project Trajectory (MPTrj) dataset [89, 90] and 3.2 million unique materials in a subset of the Alexandria dataset. [91] No dispersion correction was used and the model precision was set to *float64*. A randomly selected 4737 structures (80%) were used for model training, with the remaining 1184 structures (20%) used for validation. The Adam optimiser [92] was used to minimise a cost function comprised of an equally weighted average of energy and force errors, with the learning rate set to 0.01. The MACE model consists of two message-passing layers and employs

a radial cutoff for learning interatomic interactions of 6 Å, resulting in a total receptive field of 12 Å, which is greater than the distance when lateral interactions between surface adsorbed pairs of S-S, O-O and S-O atoms decay to zero at low surface coverage, as computed using DFT. Fine-tuning was performed for 24 epochs, to balance cost and accuracy due to plateauing of the energy and force errors (Figures 5.1(b) and (c), respectively). The fine-tuned model gave a training (validation) root mean squared error (RMSE) of 14.4 (14.2) meV/atom in total energies and 16.3 (17.2) meV/Å in atomic forces. When inferenced on the full dataset, the pre-trained foundation model gave a RMSE of 1.43×10^{10} meV in total energies and 10.7 eV/Å in maximum atomic forces, which were reduced by > 99 % upon fine-tuning the model as shown in the parity plots in Figures 5.1(d) and (e).

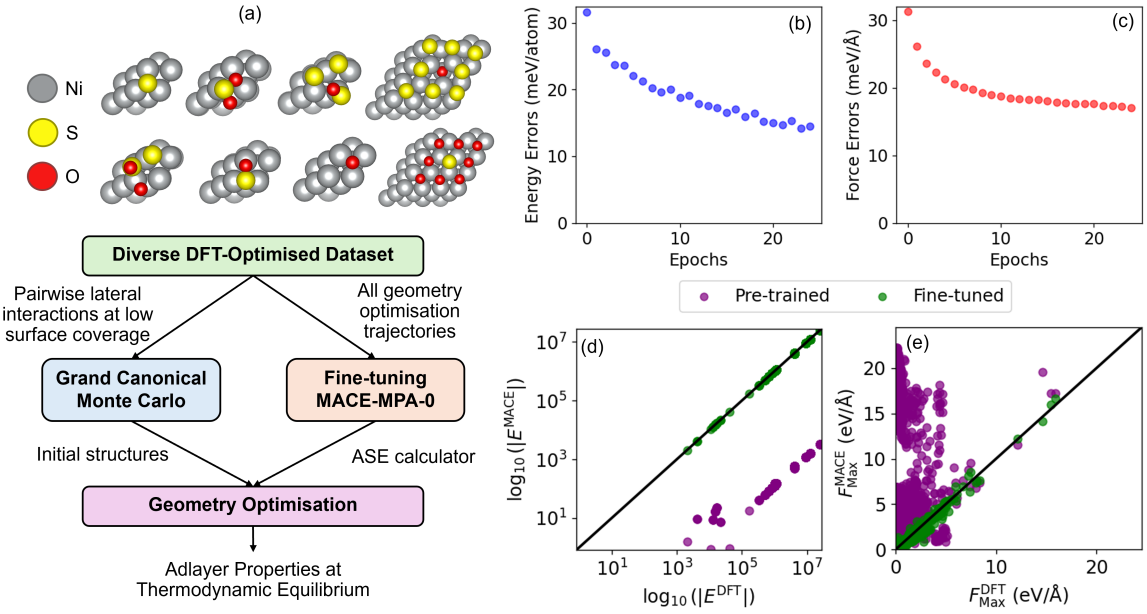


Figure 5.1: (a) Overview of the use of grand canonical Monte Carlo (GCMC) sampling and a fine-tuned MACE machine learned interatomic potential for studying the co-adsorption of S and O atoms on Ni(111) at thermodynamic equilibrium. The MACE model is fine-tuned from the MACE-MPA-0 pre-trained foundation model for 24 epochs, which results in a reduction in the (b) energy and (c) force errors until both start to plateau. When inferenced on the full dataset of DFT-optimised structures, the fine-tuned model yields a reduction in the RMSE in total energies and maximum atomic forces of > 99 % *vs.* the pre-trained foundation model, as shown in the parity plots for (d) total energies and (e) maximum atomic forces.

The fine-tuned MACE model was then used as the ASE calculator to run geometry optimisation calculations using the BFGS algorithm [62–65] with a force convergence criteria of 0.01 eV/Å. Six GCMC-predicted adlayers of differing coverages and intermixing of adsorbed S and O were validated using MACE: for $\mu_{\mathrm{S}}^{\mathrm{R}}$ = -1 eV, $\mu_{\mathrm{O}}^{\mathrm{R}}$ = -1 eV, -0.7 eV and -0.5 eV, and T = 600 K and 1200 K. The accuracy of the GCMC-predicted adlayers were validated by computing the root mean squared deviation (RMSD) of the S and O atomic positions (x and y co-ordinates) between the initial GCMC-predicted adlayers and the final MACE-optimised adlayers:

$$\mathrm{RMSD}_i = \sqrt{(x_i^{\mathrm{MACE}} - x_i^{\mathrm{GCMC}})^2 + (y_i^{\mathrm{MACE}} - y_i^{\mathrm{GCMC}})^2} \tag{5.11}$$

where x_i^{GCMC} and y_i^{GCMC} are the x and y coordinates of atom i (either S or O) in the initial GCMC-predicted adlayer and x_i^{MACE} and y_i^{MACE} are the corresponding coordinates in the final MACE-optimised adlayer.

5.2.5 Experimental Characterisation

The following catalyst preparation, sulfur poisoning and methane steam reforming activity testing was carried out by Dr Christopher Hawkins and Dr Andrew Steele from Johnson Matthey. The SEM, XPS and ICP analysis were conducted by Dr Gregory Goodlet, Dr Riho Green and Jason Raymond, respectively, from the Advanced Characterisation Department at Johnson Matthey Technology Centre.

To investigate how support oxygen buffering affects the sulfur tolerance of the Ni catalyst, we select three model supports spanning a range of reducibilities. γ-Al_2O_3 is chosen as a high surface area, structurally robust support material with negligible oxygen buffering behaviour. [93] Rutile TiO_2 is chosen as a moderately reducible support material, which can form oxygen vacancies and facilitate mild oxygen buffering at high temperatures. [43] CeO_2 is chosen as the prototypical support material for strong oxygen buffering under catalytic reaction conditions due to the ease of switching between the Ce^{3+} and Ce^{4+} oxidation states, and low oxygen vacancy formation energy. [42, 93]

The three supported catalysts of 10 wt % NiO on γ-Al_2O_3 (commercial, surface area= 140 m^2/g), rutile TiO_2 (commercial, surface area= 20 m^2/g) and CeO_2 (commercial, surface area= 20 m^2/g) were synthesised using the standard *incipient wetness impregnation* method, where the support materials were first impregnated with a Ni nitrate precursor solution, then dried and calcined at 773 K for 2 hours to obtain the final catalyst samples. [94] The catalysts were pelletised to a size of 250-355 μm and activated in a tube furnace, in a mixture of 10 % H_2 in N_2 at 923 K for 10 hours. Scanning electron microscopy (SEM) was used to visualise the morphology of the prepared catalysts using a Zeiss ultra 55 Field emission electron microscope equipped with in-lens secondary electron and backscattered detectors. X-ray diffraction (XRD) was performed using a Bruker D8 Advance Davinci design unit to measure the NiO crystallite size in the prepared catalysts.

A 1 g portion of each catalyst was then saturated with H_2S at room temperature for 18 hours in a fixed bed reactor, using a feed gas of 100 ppm of H_2S in a mixture of 2.5 % H_2 in N_2, with a relative humidity of 50 % and a flowrate of 500 ml/min. The total sulfur content following room temperature saturation was quantified using inductively coupled plasma (ICP) analysis. As the focus of this work is to investigate the thermodynamic driving force for sulfur removal and catalyst regeneration, rather than the kinetics of sulfur adsorption under operating reaction conditions, the room temperature sulfur loading protocol provides a consistent baseline from which we assess the temperature-dependent catalyst regeneration behaviour. We note that the measured sulfur content for each catalyst is expected to be a high (upper bound) estimate, with reduced adsorption at higher temperatures. The surface speciation of the H_2S-poisoned catalysts, with a measurement depth of 5-10 nm, was analysed using X-ray photoelectron spectroscopy (XPS). Temperature programmed desorption-mass spectrometry (TPD-MS), using a Micromeritics Autochem II Chemisorption analyser linked with a MKS Cirrus

2 mass spectrometer, was used to track the desorption of H_2O, SO and SO_2 from the H_2S-poisoned catalysts under a fixed temperature ramp of 10 K/min from room temperature to 1223 K in N_2.

MSR activity testing was carried out in a low-pressure rig designed to flow dry gas mixtures of N_2, CH_4 (and higher hydrocarbons) and H_2 for catalyst pre-reduction. The dry gas composition used was 68.4 % CH_4 and 3.6 % C_2H_6, with a balance of N_2. The dry gas mixture is then combined with steam (following prior heating and evaporation in an oven) forming a reaction gas mixture that is flowed through a packed catalyst bed, contained in a quartz tube, within a furnace that is electrically heated up to 1223 K. The MSR activity for each H_2S-poisoned catalyst was evaluated at steady state, at temperatures of 873, 973 and 1073 K, under regulated outlet backpressures of 100, 120 and 150 mbar, respectively. During the reaction, the dry gas is combined with steam resulting in a steam to carbon ratio of 3:1, with a total gas flowrate of 200 ml/min. The quartz tube (diameter 0.4 cm) was loaded to a 1.5 cm bed length, equating to 0.097 g (0.094 cm^3) of catalyst and 0.155 g (0.094 cm^3) of SiC inert dilutant. We note that the studied support materials are chosen as model systems to investigate the key principles driving the catalyst sulfur tolerance, but are not immediately compatible with existing industrial MSR processes due to differences in the catalyst form (*i.e.*, pellets *vs.* powders) and thermal instability at very high temperatures over long timescales.

5.3 Results and Discussion

5.3.1 Atomic and Molecular Adsorption on Ni(111)

To ascertain the number of non-equivalent adsorption sites on Ni(111), atomic S and O were adsorbed at the four initial positions illustrated in Figure 5.2(a), including hollow HCP, hollow FCC, atop and bridge sites. Geometry optimisation of atomic S adsorbed at both atop and bridge sites resulted in S diffusion to the hollow HCP site, whilst atomic O adsorbed at atop and bridge sites diffused to hollow HCP and hollow FCC sites, respectively. The hollow HCP sites in Figures 5.2(b) and (d) and the hollow FCC sites in Figures 5.2(c) and (e) were therefore determined to be the relevant non-equivalent sites for adsorption. Both atomic S and O strongly chemisorb on the Ni(111) surface and display an energetical preference for adsorption at hollow FCC sites, by 0.05 eV for S and 0.23 eV for O. The trends in adsorption energies and site preferences are in agreement with computational literature detailed in Table 5.3, although the absolute values of adsorption energies are found to vary slightly with the choice of exchange-correlation functional, as GGAs from the literature tend to underbind, [95] and the choice of Ni(111) surface model parameters. [30, 96–98] The adsorption of molecular SO was also considered, with both S and O directly bonded to the surface. At both hollow HCP and FCC sites, S-bound SO was calculated to be more energetically stable by 2.35 eV and 2.10 eV, respectively. Finally, SO_2 adsorption was tested at the four initial positions in Figure 5.2(a), from which the non-equivalent adsorption sites were atop and bridge sites in Figures 5.2(h) and (i), respectively. SO_2 is calculated to be most stable when S occupies the bridge site of Ni(111), as is reported experimentally, [99] with the same preferential stability as reported in the DFT study of Liu *et al*. [98] All calculated adsorption energies are reported in Table 5.3.

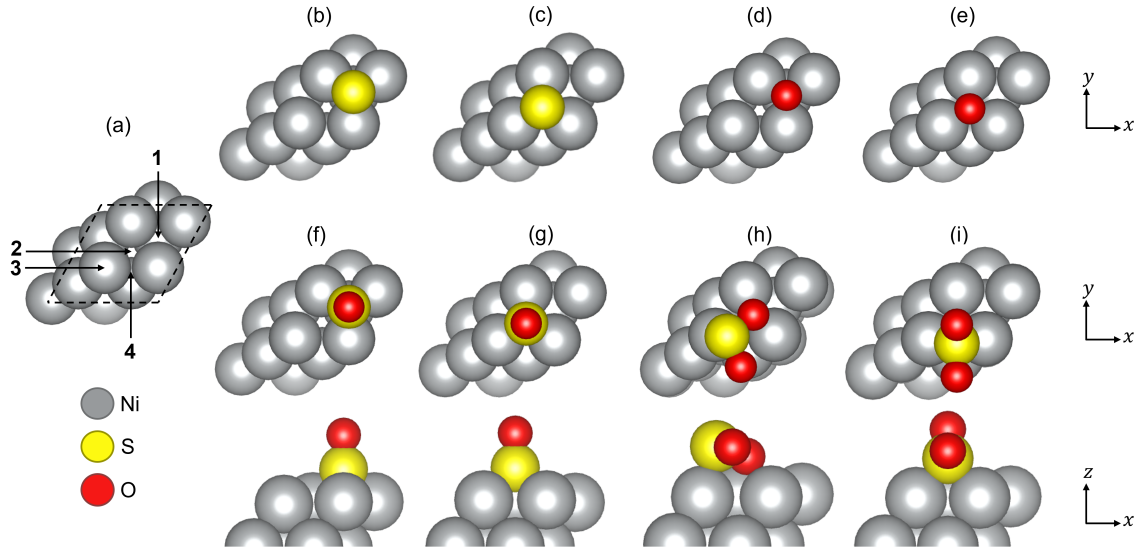


Figure 5.2: (a) The four studied adsorption sites on the Ni(111) surface, with the unit cell boundaries denoted in the black dashed lines, including (1) hollow HCP, (2) hollow FCC, (3) atop and (4) bridge. (b)-(i) The most stable single atom (S and O) and molecular (SO and SO_2) adsorption complexes on a 1×1 Ni(111) surface, calculated using DFT with the mBEEF exchange-correlation functional, where (b) and (c) correspond to S adsorption, (d) and (e) correspond to O adsorption, (f) and (g) correspond to SO adsorption and (h) and (i) correspond to SO_2 adsorption. (a)-(i) are top down views of the Ni(111) surface and the bottom row is a side view for adsorption complexes (f)-(i). The corresponding adsorption energies for the adsorption complexes (b)-(i) are listed in Table 5.3

Table 5.3: Adsorption energies (ΔE_{Ads}, eV) for atomic S, atomic O, molecular SO (for S binding to surface and OS for O binding to the surface) and molecular SO_2 on a 1×1 Ni(111) surface, calculated using DFT with the mBEEF exchange-correlation functional. The active sites are (1) hollow HCP, (2) hollow FCC, (3) atop and (4) bridge, as illustrated in Figure 5.2. Available literature comparisons are included with the corresponding exchange-correlation functional in brackets. Our results match the relative stabilities of adsorption complexes between active sites, but differences in absolute adsorption energies *vs*. the available literature are noted due to the use of GGA exchange-correlation functionals and different Ni(111) surface parameters, *e.g.*, number of layers and supercell dimensions.

Adsorbate	Active Site	ΔE_{Ads} (eV)	Literature
S	1	-7.09	-5.07 (PW91), -4.57 (RPBE), -5.56, -5.16 (PBE) [96–98]
S	2	-7.14	-5.12 (PW91), -4.62, -4.69 (RPBE), -5.62, -5.21 (PBE) [30, 96–98]
O	1	-6.25	-5.02 (PW91), -4.42, -5.76 (PBE), -5.91 (PBE) [96–98]
O	2	-6.48	-5.13 (PW91), -4.52 (RPBE), -5.87, -6.02 (PBE) [96–98]
SO (OS)	1	-4.62 (-2.27)	N/A
SO (OS)	2	-4.64 (-2.54)	-2.08 (PW91) [100]
SO_2	3	-2.00	-1.03 (PBE) [98]
SO_2	4	-2.08	-1.08 (PBE) [98]

5.3.2 Pairwise and Many-Body Lateral Interactions on Ni(111)

The four non-equivalent adsorption complexes of atomic S and O in Figures 5.2(b)-(e), were used to construct new adsorption complexes of S-S, O-O and S-O pairs at low surface coverage on a 10

× 10 Ni(111) surface (for S-S and O-O pairs) and a 7 × 7 Ni(111) surface for S-O pairs (to reduce computational cost at no detriment to accuracy). Following geometry optimisation, the energies of adsorbed single atoms and pairs were then used to compute lateral energies (E_{lateral}, defined in Section 5.2.2, Equation (5.6)), which are plotted in Figures 5.3(a)-(c) for pairs of S-S, O-O and S-O, respectively. Lateral interactions are repulsive for all pairs in Figures 5.3(a)-(c), indicating that the O-mediated removal of adsorbed S occurs at high surface coverage and would require a large supply of O atoms to the surface to overcome the repulsive lateral interactions between adsorbed S and O, *e.g.*, from a reducible metal oxide support with a large oxygen buffering capacity or using a high partial pressure of O_2 gas during experimental catalyst regeneration. All adsorption complexes corresponding to $|E_{\text{Lateral}}| \geq 0.04$ eV, *i.e.*, green markers in Figures 5.3(a)-(c), were used to parameterise the pairwise Hamiltonian ($\mathscr{H}$, defined in Section 5.2.2, Equation 5.7) for GCMC sampling. Geometry optimisation of S-O pairs at low surface coverage reveals the instability of short-range interactions of ≤ 1.45 Å between adjacent hollow HCP and hollow FCC sites, which results in atomic diffusion to neighbouring sites in Figures 5.3(d) and (e). Short-range S-O interactions were therefore not included in the GCMC sampling by assigning $E_{\text{lateral}} = \infty$ eV within the lattice model for both initial configurations in Figures 5.3(d) and (e).

The validity of excluding short-range S-O interactions from the GCMC sampling was investigated further by considering the effects of the S and O surface coverages on the energetics of S oxidation to SO. The geometry optimisations in Figures 5.3(d) and (e) were repeated on a smaller 1 × 1 Ni(111) surface in Figures 5.3(f) and (g), respectively, corresponding to a higher surface coverage, before evaluating the relative stability ($\Delta E_{\text{Relative}}$) of an adsorbed SO molecule at the most stable hollow-FCC site *vs.* atomic S and O, using:

$$\Delta E_{\text{Relative}} = E^{n\times n}_{\text{SO/Ni(111)}} - E^{n\times n}_{\text{S,O/Ni(111)}} \tag{5.12}$$

where $E^{n\times n}_{\text{SO/Ni(111)}}$ is the energy of a geometry optimised SO molecule adsorbed at a hollow-FCC site on an $n \times n$ Ni(111) surface and $E^{n\times n}_{\text{S,O/Ni(111)}}$ is the energy of a geometry optimised pair of S and O atoms adsorbed at an initial interatomic separation of 1.45 Å on an $n \times n$ Ni(111) surface.

Comparing the relative energies in Figures 5.3(d)-(g), there is a significant site-dependence in the energetic feasibility of S oxidation to SO, where relaxation of S adsorbed at hollow-FCC sites and O adsorbed at hollow-HCP sites dramatically reduces $\Delta E_{\text{Relative}}$ compared to relaxation of S adsorbed at hollow-HCP sites and O adsorbed at hollow-FCC sites. This observation is consistent with the spin-polarised DFT study of Das and Saida, who calculated $\Delta E_{\text{Relative}} = 0.41$ eV for S adsorbed at a hollow-FCC site and O adsorbed at a hollow-HCP site and $\Delta E_{\text{Relative}} = 2.98$ eV for both atoms adsorbed at hollow-FCC sites, on a 2 × 2 Ni(111) surface. [100] Our results further show a strong coverage-dependence for the feasibility of S oxidation, as shown by the reduction in $\Delta E_{\text{Relative}}$ from 0.57 eV to 0.01 eV by increasing the surface coverage from Figure 5.3(e) to Figure 5.3(g). The pairwise GCMC Hamiltonian, which excludes short-range S-O interactions that are energetically unfavourable at low surface coverage, is concluded to be valid for simulated adlayers with low θ_S and θ_O, only shown as the lighter regions in the GCMC-predicted isotherms in Figures 5.4(a) and (b), as well as regions of low intermixing between S and O shown as the lighter regions in Figure 5.4(c). In these regions, strong adsorbate interactions with the Ni(111) surface exceed any attractive lateral

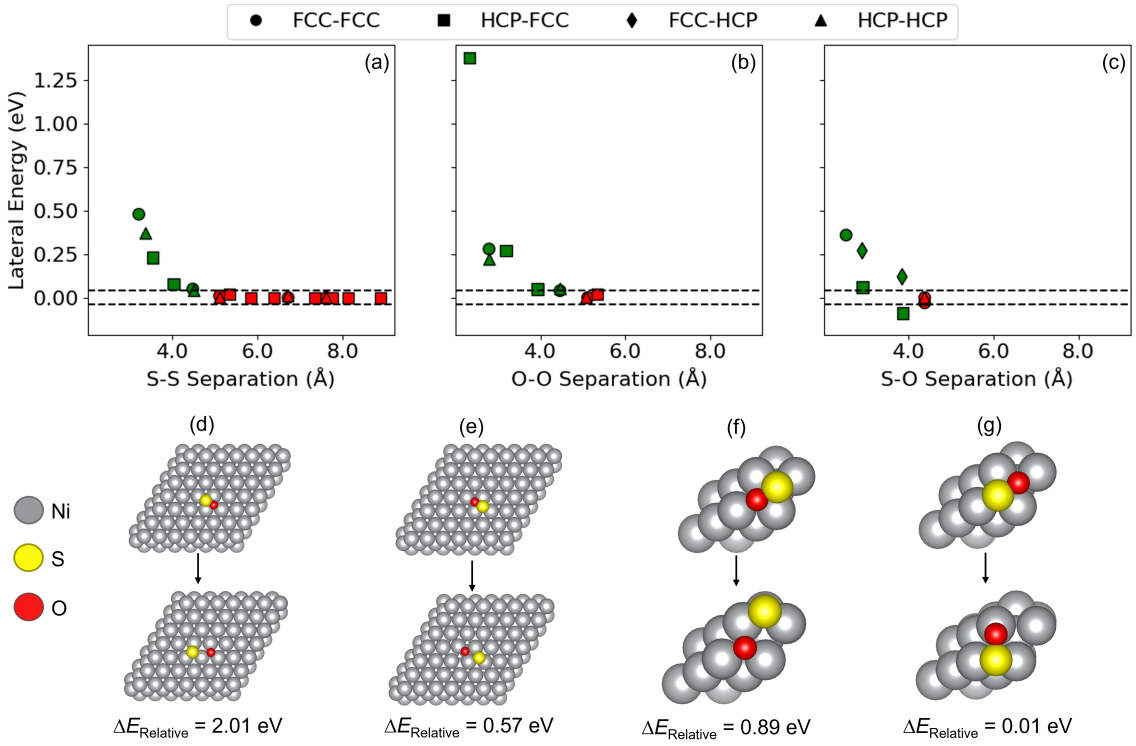


Figure 5.3: Lateral energies between adsorbed (a) S-S, (b) O-O and (c) S-O atomic pairs, at low surface coverage on Ni(111), calculated using DFT with the mBEEF exchange-correlation functional. Green (red) markers correspond to adsorption complexes that are included (not included) in the pairwise GCMC Hamiltonian. The marker shape corresponds to the type of active site occupied by each atom in the pairs. The initial (top row) and final optimised geometries (bottom row) for DFT relaxations of short-range S-O interactions, where S occupies a hollow-HCP site and O occupies a hollow-FCC site in (d) and (f), whilst S occupies a hollow-FCC site and O occupies a hollow-HCP site in (e) and (g). Adsorption complexes (d) and (e) correspond to low surface coverage on a 7×7 Ni(111) surface, whilst complexes (f) and (g) correspond to high surface coverage on a 1×1 Ni(111) surface. The relative energy for each adsorption complex (d)-(g), calculated using Equation (5.12), is listed underneath each subfigure.

interactions between adsorbed S and O as may be required for the formation of oxidised sulfur species. Under sulfur-rich conditions ($\mu_S^R \rightarrow -1$ eV), the GCMC-predicted isotherm in Figure 5.4(a) predicts a large sulfur coverage of up to 0.45 ML that is thermodynamically stable even at extremely low H_2S feed concentrations in a H_2S/H_2 mixture, on the order of parts per million. This reflects the strong chemisorption of atomic S to Ni(111) relative to the weak thermodynamic driving force for desorption into H_2S. In contrast, Figure 5.4(b) shows that co-adsorbed oxygen can reduce sulfur coverages on Ni(111) *via* site competition under sufficiently oxygen-rich conditions ($\mu_O^R \rightarrow -1$ eV); although this does not occur under any realistic oxygen partial pressures at 600 K. These results suggest that a high temperature is essential for oxygen-assisted catalyst regeneration *via* site competition between co-adsorbed S and O.

To investigate the entropic contributions to catalyst regeneration *via* oxidation of S $\rightarrow$ SO, six GCMC-predicted adlayers were validated for μ_S^R = -1 eV, μ_O^R = -1 eV, -0.7 eV and -0.5 eV, and T = 600 K and 1200 K, using geometry optimisation simulations with the fine-tuned MACE model (trained

on both low coverage and high coverage DFT relaxations). The mean and standard deviation of the RMSD of adsorbate atomic displacements is shown in Figure 5.4(d), where the MACE relaxation trajectories do not lead to S oxidation. In all cases in Figure 5.4(d), the differences in the GCMC-predicted and MACE-optimised adlayer structures are driven by surface diffusion of some adsorbed S atoms to nearest neighbour sites without any S oxidation to SO or SO_2, whilst the RMSD in atomic positions is consistently lower for adsorbed O than S (Figure 5.5).

The results suggest that combinations of μ_S^R and μ_O^R that lead to higher coverages and intermixing of S and O, illustrated by the dark blue regions in Figure 5.4(c), create conditions that are necessary but not sufficient alone for SO formation and that thermal activation is essential for SO formation irrespective of the degree of S and O co-adsorption. As a result, the use of metal oxide support materials with a large oxygen buffering capacity can aid the regeneration of S-poisoned catalysts at high temperature, where the formation and desorption of SO and SO_2 is feasible. However, tuning the support oxygen buffering capacity is unlikely to improve the sulfur tolerance of low temperature catalysts, which requires modification of the Ni catalyst to reduce the high affinity of S, O, SO and SO_2. These findings are consistent with the kinetic modelling of S oxidation on Ni(111) by Galea *et al.*, who combined DFT simulations with TPD experiments to investigate the removal of adsorbed S atoms using gas-phase O_2. [32] Their TPD results showed no SO_2 formation at temperatures below 600 K for surfaces with low S coverage, indicating that direct oxidation of S atoms is not thermally accessible at these conditions. Instead, S removal was only observed above 600 K and at sufficiently high O_2 exposures, to facilitate O-assisted S diffusion and oxidation. Their DFT calculations similarly demonstrated a high activation barrier (>1 eV) for SO formation from isolated S and O atoms on Ni(111).

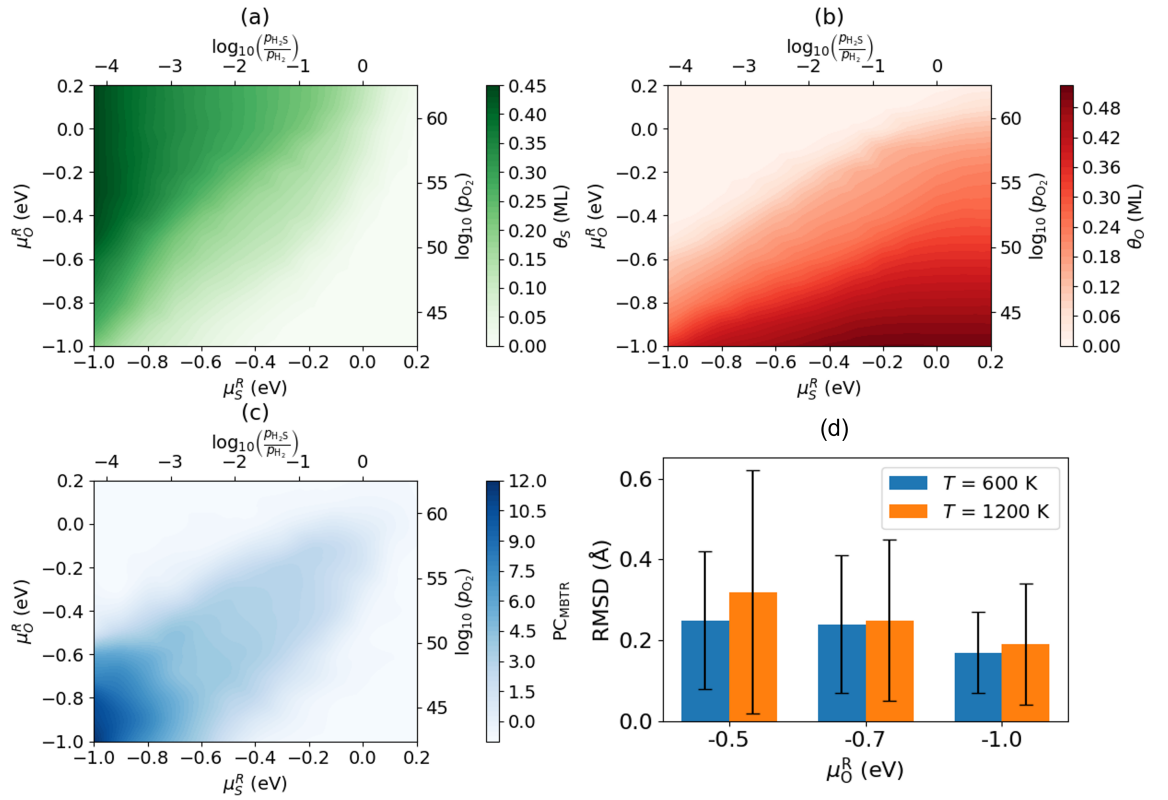


Figure 5.4: GCMC-predicted surface coverages of (a) S and (b) O at 600 K for relative chemical potentials of S (μ_S^R) and O (μ_O^R) ranging between -1 eV and 0.2 eV, as defined in Section 2.2. (c) The principal component derived from two-body many-body tensor representations (PC^{MBTR}, discussed in the SI Section S2), which encodes the pairwise interatomic distances between adsorbed S and O atoms across 10 GCMC-predicted adlayers for 441 combinations of μ_S^R and μ_O^R at 600 K. The secondary axes in (a), (b) and (c) show the equivalent gas phase thermodynamic control variables corresponding to the relative chemical potentials, including the ratio of partial pressures (p) of H_2S to H_2 (for a fixed p_{H_2} = 1 bar) and the partial pressure of O_2, which were obtained from ideal gas thermodynamics at the same temperature and a standard-state pressure of 1 bar. (d) The root-mean-square deviation (RMSD) in S and O x and y atomic co-ordinates, between GCMC-predicted and MACE-reoptimised adlayers. Bars represent the mean RMSD for each μ_O^R value at T = 600 K and 1200 K. Error bars represent the standard deviation of the RMSD. All bars correspond to μ_S^R = -1 eV, thereby testing the validity of adlayers with varied intermixing of adsorbed S and O atoms on Ni(111), which increases for larger values of μ_O^R.

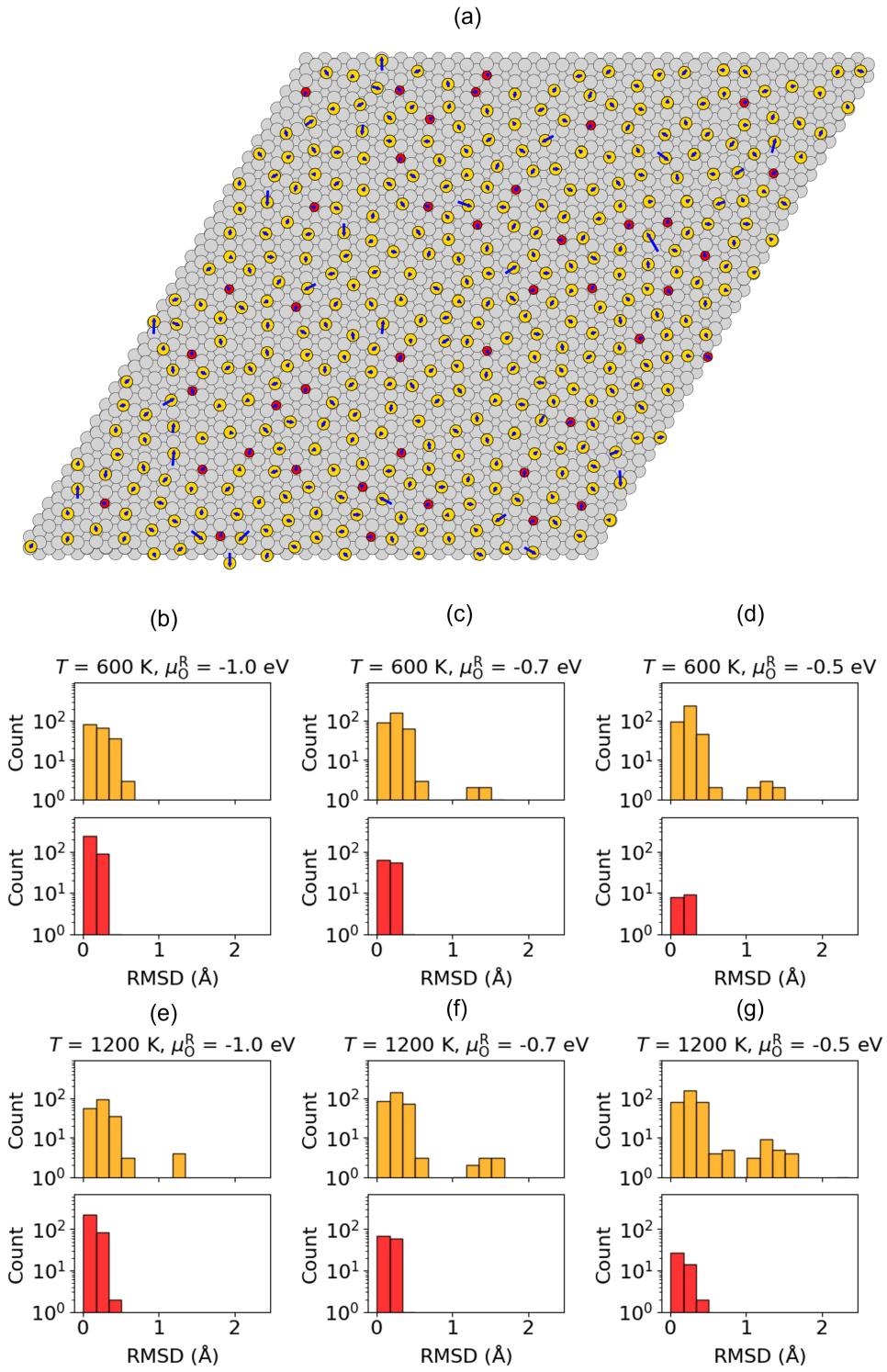


Figure 5.5: (a) The MACE-reoptimised structure for the GCMC-predicted adlayer for μ_S^R = -1 eV, μ_O^R = -0.5 eV and T = 1200 K, which corresponds to the largest RMSD in Figure 5.4(d). The arrows indicate the direction and magnitude of atomic S and O displacements from the initial GCMC-predicted atomic positions. (b)-(e) Histograms of RMSD of S (yellow bars) and O (red bars) atoms between the GCMC-predicted and MACE-reoptimised adlayers for all six validated adlayers in Figure 5.4(d).

5.3.3 Reversible *vs*. Irreversible Catalyst Deactivation

The results in Section 5.3.2 can be used to rationalise the outcomes of experimental MSR activity testing for fresh and H_2S-poisoned Ni nanoparticle catalysts in Figure 5.6, which shows methane conversion as a function of the reaction temperature.

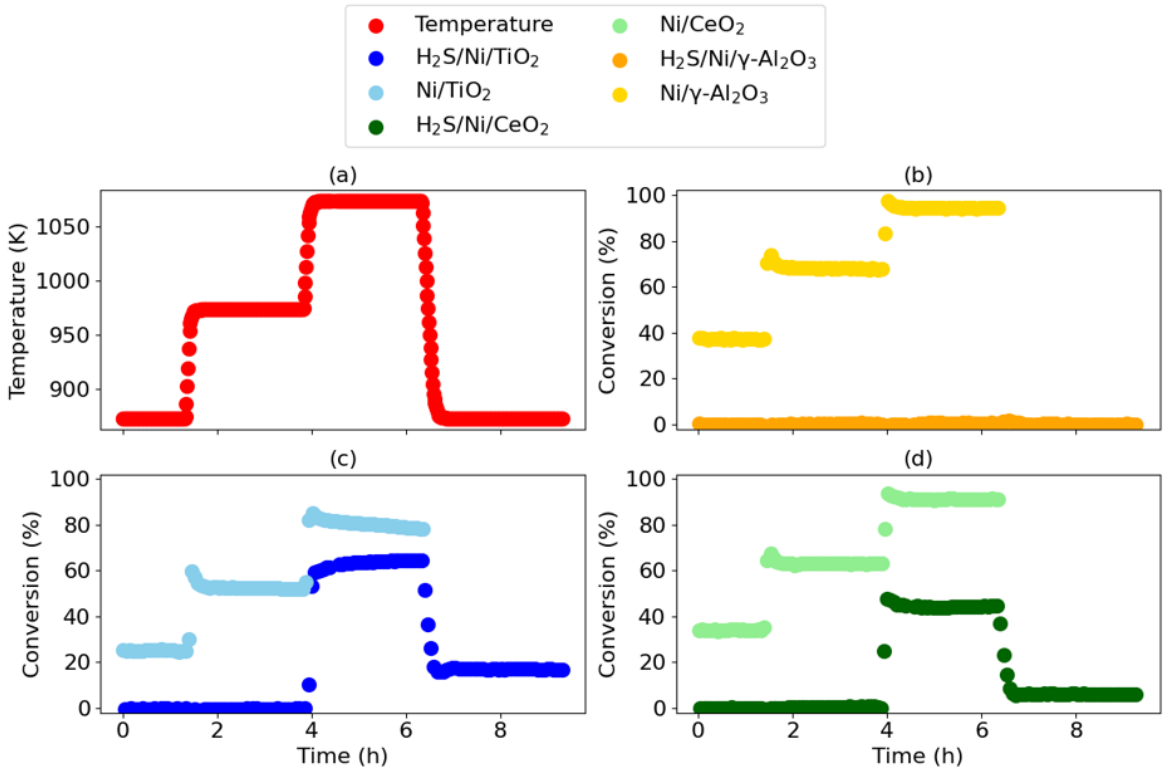


Figure 5.6: (a) Temperature profile for MSR activity testing of fresh and H_2S-poisoned Ni catalysts supported on (b) γ-Al_2O_3, (c) TiO_2 and (d) CeO_2. The reduction in temperature from 1073 K to 873 K after t = 6 hours was only performed for the H_2S-poisoned catalysts. All fresh catalysts were subject to an additional pre-reduction in H_2 at 923 K, prior to t = 0 hours. *These results were collected by Dr Christopher Hawkins and Dr Andrew Steele.*

For both H_2S-poisoned Ni/TiO_2 and H_2S-poisoned Ni/CeO_2, catalyst regeneration and partial restoration of activity (to $\sim$ 80 % and $\sim$ 50 % of that of fresh Ni/TiO_2 and Ni/CeO_2, respectively) is achieved upon increasing the temperature beyond 973 K. Although H_2S-poisoned Ni/TiO_2 is restored to the highest absolute value of catalytic activity in Figure 5.6(a), ICP analysis indicates a total uptake of H_2S during room temperature saturation of 0.11 weight percentage of sulfur ($\%_{S\ wt}$), which is an order of magnitude lower than that of Ni/γ-Al_2O_3 (2.14 $\%_{S\ wt}$) and Ni/CeO_2 (2.53 $\%_{S\ wt}$). The reduced sulfur loading on Ni/TiO_2 likely stems from the reduced dispersion of Ni in the experimentally prepared catalyst, as evident by the SEM imaging in Figure 5.7, which is consistent with the much larger XRD-determined NiO crystallite size of 17.9 nm on TiO_2 *vs.* 12.1 nm on CeO_2. As a result, Figure 5.6(a) shows that H_2S-poisoned Ni/CeO_2 is restored to a substantially greater catalytic activity than H_2S-poisoned Ni/TiO_2, relative to its sulfur-content, which is in line with our DFT+U calculated oxygen vacancy formation energies of 3.44 eV for CeO_2 and 5.35 eV for TiO_2, *i.e.*, oxygen from the CeO_2 lattice facilitates S oxidation. Both values are much lower than the DFT-calculated oxygen vacancy formation energy of 7.00 eV for γ-Al_2O_3, indicating support oxygen buffering may drive the

enhanced sulfur resistance of Ni/CeO_2, although not in a manner to reduce the temperature required for catalyst regeneration, as discussed in Section 5.3.2.

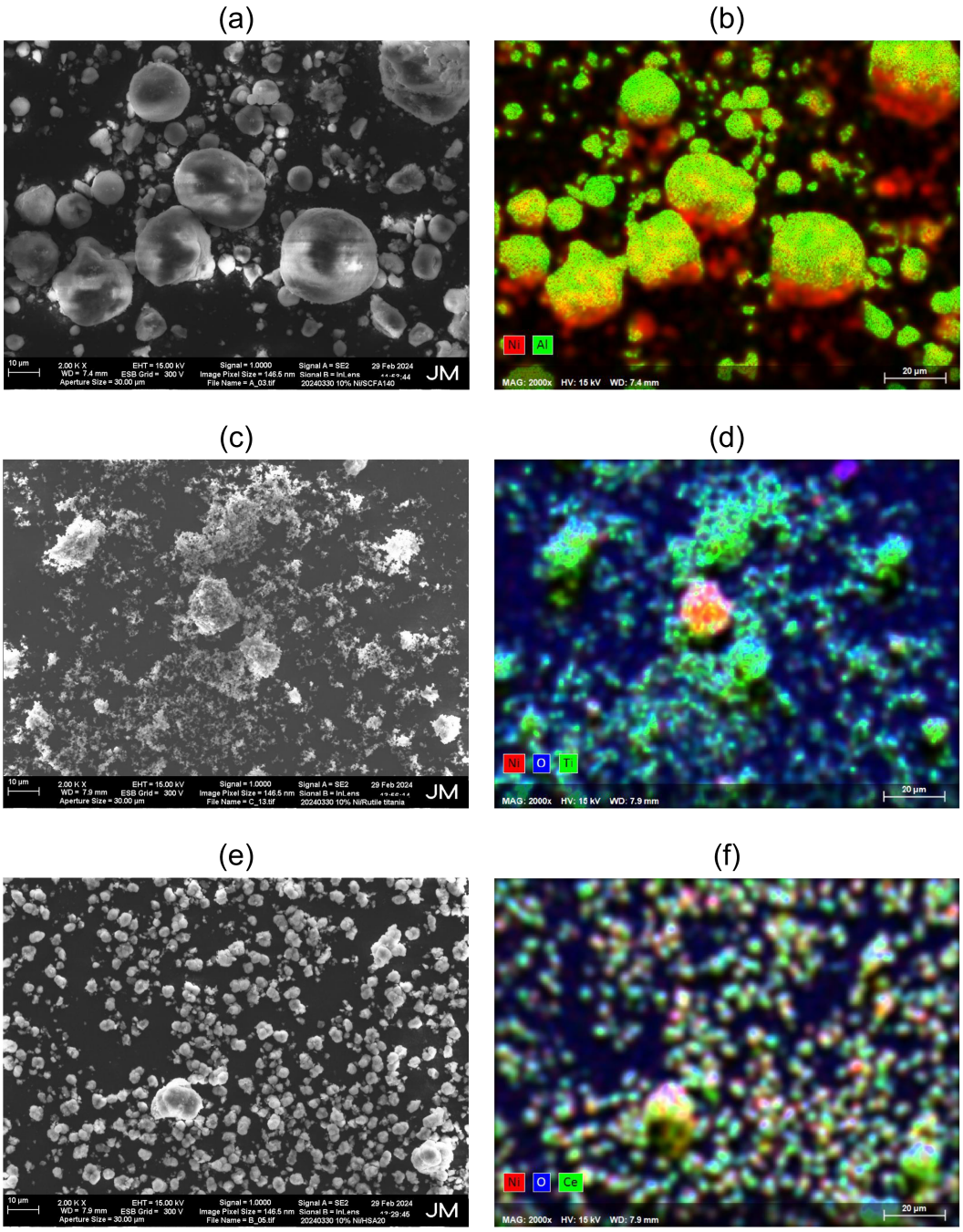


Figure 5.7: Scanning electron microscopy images of the microstructure of the prepared (a) $Ni/\gamma\text{-}Al_2O_3$, (c) Ni/TiO_2 and (e) Ni/CeO_2 catalysts. The corresponding elemental mapping of Ni (red), O (blue) and either (b) Al, (d) Ti or (f) Ce (green) shows the variation in the Ni dispersion amongst the prepared catalysts, which is significantly lower for Ni/TiO_2. *These images were collected by Dr Gregory Goodlet.*

The H_2S-poisoned $Ni/\gamma\text{-}Al_2O_3$ catalyst was found to deactivate irreversibly in Figure 5.6(b), with no restoration of catalytic activity upon increasing temperature. Given the measured activity of

the fresh Ni/γ-Al_2O_3 catalyst, which is subject to a pre-reduction in H_2 at 923 K, the irreversible deactivation of H_2S-poisoned Ni/γ-Al_2O_3 is likely due to the variation in the Ni oxidation state with respect to the reducibility of the reaction environment. The observed irreversible catalyst deactivation is consistent with the experimentally reported *in situ* transformation of Ni/γ-Al_2O_3 to spinel-type $NiAl_2O_4$, *i.e.*, switching the Ni oxidation state from Ni^0 in Ni^{2+} on the surface and in the bulk, which is inactive for MSR. [101–103] The suppression of Ni^0 when Ni/γ-Al_2O_3 is exposed to oxidising atmospheres, *e.g.*, when exposed to air in ambient conditions before characterisation, is further supported by the Ni $2p_{3/2}$ XPS spectra in Figure 5.8(a), where the Ni surface speciation on the different supports is distinctly different at ~ 853 eV, which corresponds to Ni^0, whilst being similar at ~ 856 eV, which corresponds to Ni^{2+}. [104] Given that the relative intensity of the peak at ~ 853 eV is lowest for H_2S-poisoned Ni/γ-Al_2O_3, this suggests that γ-Al_2O_3 suppresses the formation of Ni^0 in oxidising conditions.

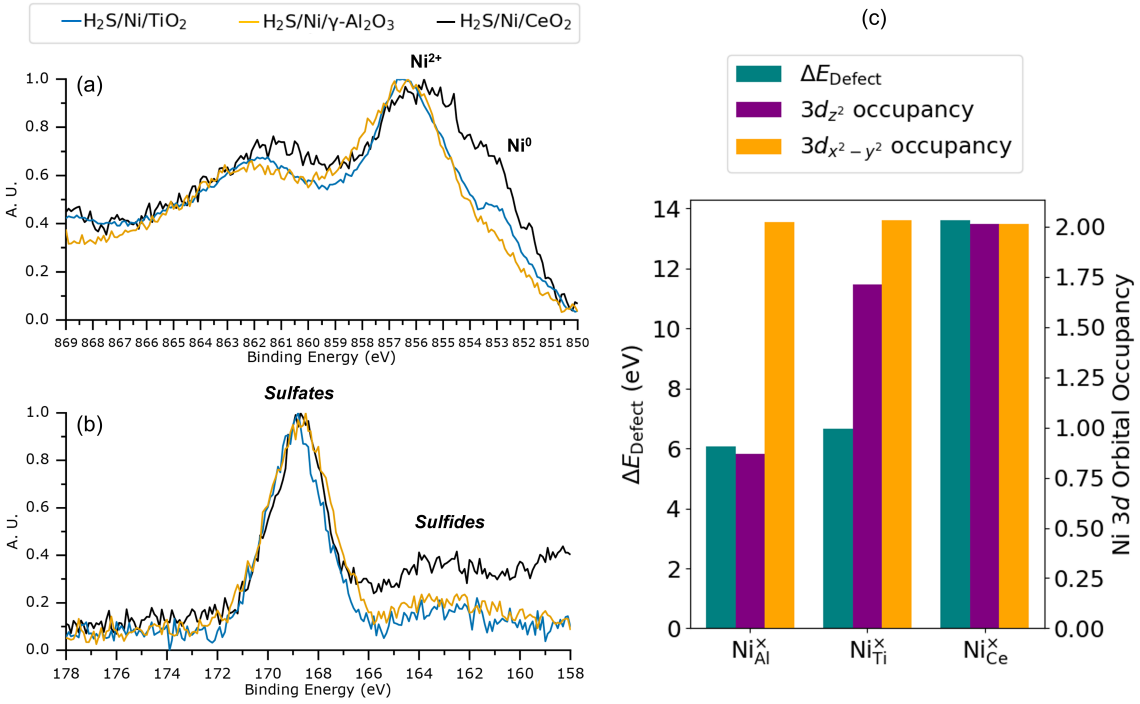


Figure 5.8: Normalised XPS spectra for (a) Ni $2p_{3/2}$ and (b) S $2p$ for the three H_2S-poisoned Ni catalysts following room temperature saturation with H_2S (before MSR activity testing). (c) Substitutional defect energies for $Ni^{\times}_{Al}$ in bulk γ-Al_2O_3 (DFT), $Ni^{\times}_{Ti}$ in bulk TiO_2 (DFT+U) and $Ni^{\times}_{Ce}$ in bulk CeO_2 (DFT+U), calculated using the mBEEF exchange-correlation functional and Hubbard parameters detailed in the Section 5.2.1. The defect energies are plotted alongside the corresponding occupancies of the Ni $3d$ e_g orbitals, including both $3d_{z^2}$ and $3d_{x^2\text{-}y^2}$ orbitals. Large differences between $3d_{z^2}$ and $3d_{x^2\text{-}y^2}$ orbital occupancies are reportedly characteristic of systems with stabilising Jahn-Teller distortions. [105, 106] *The XPS spectra were collected by Dr Riho Green.*

To investigate the driving force for irreversible catalyst deactivation further, the energetics of substitutional defect formation in the support materials were calculated using DFT and DFT+U, as outlined in Section 5.2.1. As shown in Figure 5.8(c), the substitutional defect energy for $Ni^{\times}_{Al}$ in γ-Al_2O_3 is calculated as 6.08 eV, which is lower than $Ni^{\times}_{Ti}$ in TiO_2 (6.67 eV) and $Ni^{\times}_{Ce}$ in CeO_2 (13.61 eV), supporting a hypothesis that the deactivating phase transformation is more favourable for

Ni/γ-Al_2O_3, whereas Ni/TiO_2 and Ni/CeO_2 are more resistant to forming bulk solid solutions. Figure 5.8(c) further shows that the increasing defect energies from $Ni^{\times}_{Al}$ to $Ni^{\times}_{Ce}$ correlate inversely with the polarisation of the Ni 3*d* e_g orbitals, comprised of the $3d_{z^2}$ and $3d_{x^2\text{-}y^2}$ orbitals that align along the metal-oxygen bonds, [107] which is characteristic of complex oxides containing divalent ions such as Ni^{2+} resulting in stabilisation *via* Jahn-Teller distortions that break the system symmetry. [105, 106] These results indicate an energetic favourability for the initial stages of phase transformation in γ-Al_2O_3, in agreement with the DFT+*U*-parameterised Monte Carlo study of Elias *et al*., who concluded the $NiAl_2O_4$ can be more stable than NiO and γ-Al_2O_3 in Ni-rich conditions at high temperatures. [33] The predicted insolubility of Ni in CeO_2 is in contrast with literature-reported defect energies of $\sim$ 2-3 eV using DFT+*U* in a planewave basis. [108, 109] Whilst the two sets of results are not directly comparable due to differences in the employed Hubbard projectors, our results align with the results in Chapters 3 and 4 that shows self-consistent DFT+*U* in a NAO framework can successfully rationalise experimentally observed defect chemistry in TMOs, *e.g.*, the varying oxidation states of Nb and W dopants in different TiO_2 polymorphs [47, 48] and the energetics of Mg doping in $LiCoO_2$, [48] the results for which can vary ambiguously in the plane-wave DFT+*U* literature.[110–113] The large defect energy for $Ni^{\times}_{Ce}$ is confirmed as not an artifact of our chosen DFT+*U* parameters by repetition of the calculation using standalone DFT, which yields a defect formation energy of 13.81 eV.

5.3.4 Sulfur Speciation and the Role of Water

To gain further insights into the mechanisms that drive sulfur removal from the H_2S-poisoned catalysts, TPD-MS was performed in N_2 to track the signals for H_2O, SO and SO_2, which correspond to measurements from mass spectrometry (Figure 5.9). For H_2S-poisoned Ni/CeO_2, sulfur removal occurs partially in a low temperature regime (between 423-573 K) and also a high temperature regime (beyond 973 K), which can be attributed to lattice and surface oxygen, respectively, based on the thermogravimetric analysis of Zhu *et al*., who studied pure and Ni-doped CeO_2 nanorods showing surface oxygen release between 423-593 K and lattice oxygen release between 593-1073 K.[114] Liu *et al*. similarly used TPD-MS to investigate SO_2 release from H_2S-poisoned CeO_2, concluding that peaks between 473-673 K corresponded to the formation of SO_2 which could react with lattice oxygen above 673 K to form $Ce(SO_4)_2$ and then this decomposes back to SO_2 at 873 K. [115] The role of oxygen in facilitating sulfur removal was further supported by observations that SO_2 TPD-MS signals were greatest when the catalyst was pretreated in O_2, compared to inert Ar or reducing H_2. [115] Figures 5.9(b) and (c) show a greater TPD-MS signal for SO and SO_2 release from H_2S-poisoned Ni/γ-Al_2O_3 at low temperatures than H_2S-poisoned Ni/CeO_2. We attribute this difference to the increased formation of surface $Ni_xAl_{1-x}O_2$ solid solutions, based on our calculated bulk defect formation energies in Section 5.3.3 and the H_2 temperature programmed reduction (TPR) study of Shan *et al*., which correlated the bimodal distribution at low temperatures in Figures 5.9(b) and (c) to the existence of both Ni^0 and Ni^{2+} on the catalyst surface. [116]

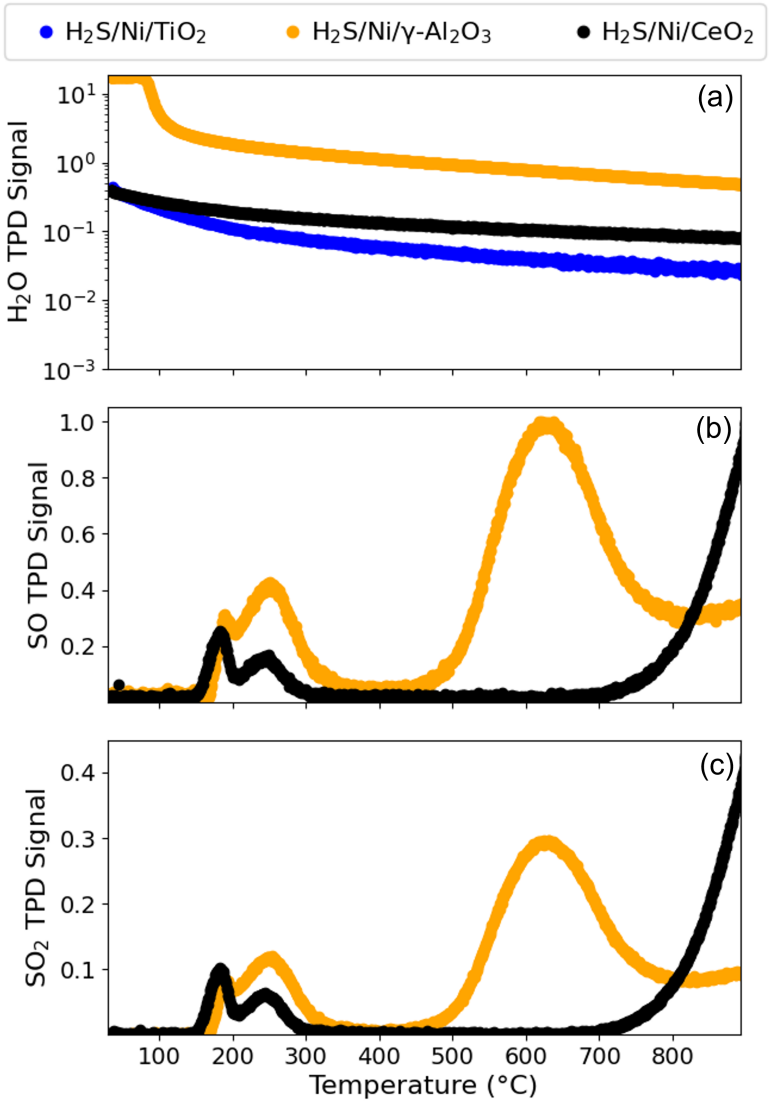


Figure 5.9: Temperature-programmed-desorption-mass spectrometry (TPS-MS) spectra obtained using a fixed temperature ramp of 10 K/min from room temperature to 1223 K in N_2 for (a) H_2O (mass = 18 g/mol) release from H_2S-poisoned γ-Al_2O_3, TiO_2 and CeO_2, (b) SO (mass = 48 g/mol) release from H_2S-poisoned γ-Al_2O_3 and CeO_2, and (c) SO_2 (mass = 64 g/mol) release from H_2S-poisoned γ-Al_2O_3 and CeO_2. The TPD-MS spectra for SO and SO_2 release from H_2S-poisoned Ni/TiO_2 were negligible (due to the lower H_2S loading as discussed in Section 5.3.3) and therefore are not shown. TPD-MS signals for H_2S (mass = 34 g/mol) release from all catalysts were negligible, indicating H_2S desorption and/or dissociation before analysis. These catalysts were not subject to a pre-reduction in in H_2 at 923 K, as discussed for the fresh catalysts in Section 5.2.5. *These spectra were collected by Dr Christopher Hawkins and Dr Andrew Steele.*

To rationalise the differences between the high temperature SO and SO_2 desorption behaviour from Ni/γ-Al_2O_3 and Ni/CeO_2 in Figures 5.9(b) and (c), the S 2*p* XPS spectra in Figure 5.8(b) is considered, where sulfates and sulfides (NiS) were identified as the peaks at ~ 169 eV and ~ 162 eV, respectively. Around 85 % of all sulfur species in the three H_2S-poisoned catalysts were quantified to be sulfates using curve fitting of the S 2*p* XPS spectra in Figure 5.8(b). The temperature-dependent oxidation (reduction) of SO_2 to (from) sulfates is hypothesised to drive the differences in the TPD-MS

spectra of Ni/γ-Al_2O_3 and Ni/CeO_2 in Figures 5.9(b) and (c). The hypothesis is supported by the study of Hamzehlouyan *et al.*, who combined TPD and diffuse reflectance infrared Fourier transform spectroscopy (DRIFTS) to investigate SO_2 release from SO_2-poisoned Pt/Al_2O_3 catalysts, concluding that SO_2-TPD peaks at ~ 509 K and ~ 947 K correspond to the desorption of molecularly adsorbed SO_2 and the dissociation of aluminium sulfate, respectively. [117] Furthermore, Smirnov *et al.* used temperature-resolved XPS to show that water vapour inhibits SO_2 oxidation to sulfates on an Al_2O_3 thin film but enhances sulfate formation on a CeO_2 thin film, due to a Ce^{3+} redox-mediated mechanism of SO_2 oxidation. [118] Together with our TPD-MS results in Figure 5.9(a), which show orders of magnitude greater water adsorption on Ni/γ-Al_2O_3 than Ni/CeO_2 due to the 7× greater surface area, the findings of Hamzehlouyan *et al.* and Smirnov *et al.* support the hypothesis that SO and SO_2 desorb at lower temperatures from Ni/γ-Al_2O_3 as water vapour inhibits the formation and retention of thermally stable sulfates.

5.4 Conclusions

Understanding the atomic level mechanisms that govern the sulfur tolerance of Ni-based catalysts is essential for designing next-generation catalysts for industrial H_2 production *via* MSR and low-temperature processes from renewable feedstocks. In this study, a combined computational and experimental approach is adopted to investigate the enhanced sulfur tolerance of Ni nanoparticles on reducible metal oxide supports, with the aim of uncovering strategies for future catalyst optimisation. Combining DFT, GCMC and a fine-tuned MACE MLIP, a high oxygen chemical potential provided *via* support oxygen buffering is shown to be insufficient for the removal of adsorbed S from Ni(111), with additional thermal activation being essential. The results support experimental MSR activity tests showing that the catalytic activity of Ni supported on reducible CeO_2 can be readily restored from a poisoned state at high temperatures, compared to Ni supported on less reducible TiO_2 and γ-Al_2O_3. The results are further validated using DFT+U computed oxygen vacancy formation energies for the bulk support materials, which show the ease of oxygen vacancy formation in the order CeO_2 > TiO_2 > γ-Al_2O_3. The MSR activity testing also indicates the critical role of phase transformations into catalytically inactive phases, which is widely reported to occur for Ni/γ-Al_2O_3, and that agrees with our DFT+U computed defect energies for substitutional Ni doping, which indicate the initial stages of bulk phase transformation are more favourable in the order γ-Al_2O_3 > TiO_2 > CeO_2. TPD-MS and XPS highlight the critical role of water in the formation of thermally stable sulfate species that can increase the temperatures required for catalyst regeneration.

Overall, the combined computational and experimental investigation points to three critical aspects for the rational design of metal oxide support materials for sulfur tolerant catalysts: (1) the feasibility of bulk oxygen vacancy formation in the support; (2) the resistance of the bulk support to phase transformations into catalytically inactive solid solutions; and (3) the support- and temperature-dependent surface chemistry of SO_2 to sulfates. The integration of *ab initio* computational modelling, statistical sampling and machine learning further demonstrates the importance of advanced workflows for studying complex catalytic materials in a manner that faithfully bridges theory and experiment.

References

(1) A. Chaudhari, P. Stishenko, A. Hiregange, C. Hawkins, M. Sarwar, S. Poulston and A. J. Logsdail, *Ab initio* insights into support-induced sulfur resistance of Ni-based reforming catalysts, *Catal. Sci. Technol.* 2026, –.

(2) J. Moore, J. Durham, A. Eijk, E. Karakas, R. Kurz, J. Lesak, M. McBain, P. McCalley, L. Moroz, Z. Mohamed, B. Pettinato, G. Phillippi, H. Watanabe and B. Williams, *Chapter 9 - Compressors and expanders*, ed. K. Brun and T. Allison, Elsevier, 2022, pp. 333–424.

(3) W. S. Jablonski, S. M. Villano and A. M. Dean, A comparison of H_2S, SO_2, and COS poisoning on Ni/YSZ and Ni/K_2O-$CaAl_2O_4$ during methane steam and dry reforming, *Appl. Catal. A: Gen.* 2015, **502** 399–409.

(4) B. Hua, M. Li, Y.-F. Sun, Y.-Q. Zhang, N. Yan, J. Chen, J. Li, T. Etsell, P. Sarkar and J.-L. Luo, Biogas to syngas: flexible on-cell micro-reformer and NiSn bimetallic nanoparticle implanted solid oxide fuel cells for efficient energy conversion, *J. Mater. Chem. A* 2016, **4** 4603–4609.

(5) N. Schiaroli, M. Volanti, A. Crimaldi, F. Passarini, A. Vaccari, G. Fornasari, S. Copelli, F. Florit and C. Lucarelli, Biogas to Syngas through the Combined Steam/Dry Reforming Process: An Environmental Impact Assessment, *Energ. Fuel.* 2021, **35** 4224–4236.

(6) J. A. Rodriguez and J. Hrbek, Interaction of sulfur with well-defined metal and oxide surfaces: unraveling the mysteries behind catalyst poisoning and desulfurization, *Acc. Chem. Res.* 1999, **32** 719–728.

(7) C. Brady, J. Pan and B. Xu, Sulfur resilient nickel based catalysts for steam reforming of jet fuel, *Catal. Sci. Technol.* 2020, **10** 8429–8436.

(8) C. Xie, Y. Chen, Y. Li, X. Wang and C. Song, Sulfur poisoning of CeO_2–Al_2O_3-supported mono- and bi-metallic Ni and Rh catalysts in steam reforming of liquid hydrocarbons at low and high temperatures, *Appl. Catal. A: Gen.* 2010, **390** 210–218.

(9) S. L. Lakhapatri and M. A. Abraham, Deactivation due to sulfur poisoning and carbon deposition on Rh-Ni/Al_2O_3 catalyst during steam reforming of sulfur-doped n-hexadecane, *Appl. Catal. A: Gen.* 2009, **364** 113–121.

(10) A. Cho, B. Hwang and J. W. Han, Development of Ni-based alloy catalysts to improve the sulfur poisoning resistance of Ni/YSZ anodes in SOFCs, *Catal. Sci. Technol.* 2020, **10** 4544–4552.

(11) J. Xu, J. Harmer, G. Li, T. Chapman, P. Collier, S. Longworth and S. C. Tsang, Size dependent oxygen buffering capacity of ceria nanocrystals, *Chem. Commun.* 2010, **46** 1887–1889.

(12) G. Sun, K. Hidajat, X. Wu and S. Kawi, A crucial role of surface oxygen mobility on nanocrystalline Y_2O_3 support for oxidative steam reforming of ethanol to hydrogen over Ni/Y_2O_3 catalysts, *Appl. Catal. B: Environ.* 2008, **81** 303–312.

(13) U. Oemar, K. Hidajat and S. Kawi, Pd–Ni catalyst over spherical nanostructured Y_2O_3 support for oxy-CO_2 reforming of methane: Role of surface oxygen mobility, *Int. J. Hydrogen Energy.* 2015, **40** 12227–12238.

(14) Z. Li and K. Sibudjing, Facile Synthesis of Multi-Ni-Core@Ni Phyllosilicate@CeO_2 Shell Hollow Spheres with High Oxygen Vacancy Concentration for Dry Reforming of CH_4, *ChemCatChem* 2018, **10** 2994–3001.

(15) D. Guo, Y. Lu, Y. Ruan, Y. Zhao, Y. Zhao, S. Wang and X. Ma, Effects of extrinsic defects originating from the interfacial reaction of CeO_{2-x}-nickel silicate on catalytic performance in methane dry reforming, *Appl. Catal. B: Environ.* 2020, **277** 119278.

(16) L. Pino, C. Italiano, A. Vita, M. Laganà and V. Recupero, $Ce_{0.70}La_{0.20}Ni_{0.10}O_{2-}$ catalyst for methane dry reforming: Influence of reduction temperature on the catalytic activity and stability, *Appl. Catal. B: Environ.* 2017, **218** 779–792.

(17) H. Wang, X. Dong, T. Zhao, H. Yu and M. Li, Dry reforming of methane over bimetallic Ni-Co catalyst prepared from $La(Co_xNi_{1-x})_{0.5}Fe_{0.5}O_3$ perovskite precursor: Catalytic activity and coking resistance, *Appl. Catal. B: Environ.* 2019, **245** 302–313.

(18) G.-R. Hong, K.-J. Kim, S.-Y. Ahn, B.-J. Kim, H.-R. Park, Y.-L. Lee, S. S. Lee, Y. Jeon and H.-S. Roh, Sulfur-Resistant CeO_2-Supported Pt Catalyst for Waste-to-Hydrogen: Effect of Catalyst Synthesis Method, *Catalysts* 2022, **12** 1670.

(19) Y.-L. Lee, K.-J. Kim, G.-R. Hong, S.-Y. Ahn, B.-J. Kim, H.-R. Park, S.-J. Yun, J. W. Bae, B.-H. Jeon and H.-S. Roh, Sulfur-Tolerant Pt/CeO_2 Catalyst with Enhanced Oxygen Storage Capacity by Controlling the Pt Content for the Waste-to-Hydrogen Processes, *ACS Sustain. Chem. Eng.* 2021, **9** 15287–15293.

(20) S. d. S. Eduardo, J. P. Mendonça, P. N. Romano, J. M. A. R. de Almeida, G. Machado and M. A. S. Garcia, Tailoring Ceria-Based Nanocatalysts for Enhanced Performance in Steam Reforming Processes: Exploring Fundamentals and Morphological Modulations, *Hydrogen* 2023, **4** 493–522.

(21) M. A. Ocsachoque, J. I. Eugenio Russman, B. Irigoyen, D. Gazzoli and M. G. González, Experimental and theoretical study about sulfur deactivation of Ni/CeO_2 and Rh/CeO_2 catalysts, *Mater. Chem. Phys.* 2016, **172** 69–76.

(22) L. Li, C. Howard, D. L. King, M. Gerber, R. Dagle and D. Stevens, Regeneration of Sulfur Deactivated Ni-Based Biomass Syngas Cleaning Catalysts, *Ind. Eng. Chem. Res.* 2010, **49** 10144–10148.

(23) P. Wachter, C. Gaber, J. Raic, M. Demuth and C. Hochenauer, Experimental investigation on H_2S and SO_2 sulphur poisoning and regeneration of a commercially available Ni-catalyst during methane tri-reforming, *Int. J. Hydrogen Energy.* 2021, **46** 3437–3452.

(24) A. de Lucas-Consuegra, A. Caravaca, P. Martínez, J. Endrino, F. Dorado and J. Valverde, Development of a new electrochemical catalyst with an electrochemically assisted regeneration ability for H_2 production at low temperatures, *J. Catal.* 2010, **274** 251–258.

(25) S. Zha, Z. Cheng and M. Liu, Sulfur Poisoning and Regeneration of Ni-Based Anodes in Solid Oxide Fuel Cells, *J. Electrochem. Soc.* 2006, **154** B201.

(26) T. Morooka, T. Shishido, R. Devivaraprasad, G. Elumalai, M. Aoki, T. Shirasawa, T. Nakanishi, A. Ishikawa, T. Kondo and T. Masuda, Potential-Dependent and Face Orientation-Dependent Electrochemical Oxidative Desorption Behavior of Sulfur Species Adsorbed on Platinum Single-Crystal Surfaces, *J. Phys. Chem. C.* 2024, **128** 16426–16436.

(27) W. Chen, T. Li, L. Peng, G. Shen, Z. Jiang, B. Huang and H. Zuo, First-principles study of H_2S adsorption and dissociation on the Ni(111) and Cl-covered Ni(111) surfaces, *Comput. Theor. Chem.* 2024, **1231** 114443.

(28) M. Zhang, Z. Fu and Y. Yu, Adsorption and decomposition of H_2S on the Ni(111) and Ni(211) surfaces: A first-principles density functional study, *Appl. Surf. Sci.* 2019, **473** 657–667.

(29) J.-H. Wang and M. Liu, Computational study of sulfur–nickel interactions: A new S–Ni phase diagram, *Electrochem. Commun.* 2007, **9** 2212–2217.

(30) C. R. Bernard Rodríguez and J. A. Santana, Adsorption and diffusion of sulfur on the (111), (100), (110), and (211) surfaces of FCC metals: Density functional theory calculations, *J. Chem. Phys.* 2018, **149** 204701.

(31) C.-H. Yeh and J.-J. Ho, A First-Principle Calculation of Sulfur Oxidation on Metallic Ni(111) and Pt(111), and Bimetallic Ni@Pt(111) and Pt@Ni(111) Surfaces, *ChemPhysChem* 2012, **13** 3194–3203.

(32) N. M. Galea, J. M. Lo and T. Ziegler, A DFT study on the removal of adsorbed sulfur from a nickel(111) surface: Reducing anode poisoning, *J. Catal.* 2009, **263** 380–389.

(33) I. Elias, A. Soon, J. Huang, B. S. Haynes and A. Montoya, Atomic order, electronic structure and thermodynamic stability of nickel aluminate, *Phys. Chem. Chem. Phys.* 2019, **21** 25952–25961.

(34) P. Littlewood, S. Liu, E. Weitz, T. J. Marks and P. C. Stair, Ni-alumina dry reforming catalysts: Atomic layer deposition and the issue of Ni aluminate, *Catal. Today* 2020, **343** 18–25.

(35) F. F. Tao, J. J. Shan, L. Nguyen, Z. Wang, S. Zhang, L. Zhang, Z. Wu, W. Huang, S. Zeng and P. Hu, Understanding complete oxidation of methane on spinel oxides at a molecular level, *Nat. Commun.* 2015, **6** 7798.

(36) L. Yu, M. Song, P. T. Williams and Y. Wei, Alumina-Supported Spinel $NiAl_2O_4$ as a Catalyst for Re-forming Pyrolysis Gas, *Ind. Eng. Chem. Res.* 2019, **58** 11770–11778.

(37) F. Chen, C. Wu, L. Dong, A. Vassallo, P. T. Williams and J. Huang, Characteristics and catalytic properties of $Ni/CaAlO_x$ catalyst for hydrogen-enriched syngas production from pyrolysis-steam reforming of biomass sawdust, *Appl. Catal. B: Environ.* 2016, **183** 168–175.

(38) D. Li, Y. Li, X. Liu, Y. Guo, C.-W. Pao, J.-L. Chen, Y. Hu and Y. Wang, $NiAl_2O_4$ Spinel Supported Pt Catalyst: High Performance and Origin in Aqueous-Phase Reforming of Methanol, *ACS Catal.* 2019, **9** 9671–9682.

(39) A. Jamsaz, N. Pham-Ngoc, M. Wang, D. H. Jeong and E. W. Shin, Favorable formation of needle-shaped $NiAl_2O_4$ phase over macroporous $Ni/Ce_xZr_{1-x}O_2$–Al_2O_3 catalysts in one-pot preparation and coke-resistant catalytic performance in dry reforming of methane, *Chem. Eng. J.* 2024, **500** 156932.

(40) S. Li, J. Li, Z. He, Y. Sheng and W. Liu, Superior catalytic combustion of methane over Pd supported on oxygen vacancy-rich $NiAl_2O_4$, *Catal. Sci. Technol.* 2024, **14** 5864–5873.

(41) S. T. Misture, K. M. McDevitt, K. C. Glass, D. D. Edwards, J. Y. Howe, K. D. Rector, H. He and S. C. Vogel, Sulfur-resistant and regenerable Ni/Co spinel-based catalysts for methane dry reforming, *Catal. Sci. Technol.* 2015, **5** 4565–4574.

(42) Z.-K. Han, W. Liu and Y. Gao, Advancing the Understanding of Oxygen Vacancies in Ceria: Insights into Their Formation, Behavior, and Catalytic Roles, *JACS Au* 2025, **5** 1549–1569.

(43) T. Zhang, P. Zheng, J. Gao, X. Liu, Y. Ji, J. Tian, Y. Zou, Z. Sun, Q. Hu, G. Chen et al., Simultaneously activating molecular oxygen and surface lattice oxygen on Pt/TiO2 for low-temperature CO oxidation, *Nat. Commun.* 2024, **15** 6827.

(44) N. L. Nguyen, N. Colonna, A. Ferretti and N. Marzari, Koopmans-Compliant Spectral Functionals for Extended Systems, *Phys. Rev. X* 2018, **8** 021051.

(45) J. P. Perdew and A. Zunger, Self-interaction correction to density-functional approximations for many-electron systems, *Phys. Rev. B* 1981, **23** 5048–5079.

(46) M. Reticcioli, U. Diebold and C. Franchini, Modeling polarons in density functional theory: lessons learned from TiO_2, *J. Condens. Matter Phys.* 2022, **34** 204006.

(47) A. Chaudhari, A. J. Logsdail and A. Folli, Polymorph-Induced Reducibility and Electron Trapping Energetics of Nb and W Dopants in TiO_2, *J. Phys. Chem. C* 2025, **129** 15453–15461.

(48) A. Chaudhari, K. Agrawal and A. J. Logsdail, Machine learning generalised DFT+U projectors in a numerical atom-centred orbital framework, *Digit. Discov.* 2025, **4** 3701–3727.

(49) V. Blum, R. Gehrke, F. Hanke, P. Havu, V. Havu, X. Ren, K. Reuter and M. Scheffler, *Ab initio* molecular simulations with numeric atom-centered orbitals, *Comput. Phys. Commun.* 2009, **180** 2175–2196.

(50) A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, E. D. Hermes, P. C. Jennings, P. B. Jensen, J. Kermode, J. R. Kitchin, E. L. Kolsbjerg, J. Kubal, K. Kaasbjerg, S. Lysgaard, J. B. Maronsson, T. Maxson, T. Olsen, L. Pastewka, A. Peterson, C. Rostgaard, J. Schiøtz, O. Schütt, M. Strange, K. S. Thygesen, T. Vegge, L. Vilhelmsen, M. Walter, Z. Zeng and K. W. Jacobsen, The atomic simulation environment—a Python library for working with atoms, *J. Phys. Condens. Matter.* 2017, **29** 273002.

(51) O. Lamiel-Garcia, K. C. Ko, J. Y. Lee, S. T. Bromley and F. Illas, When Anatase Nanoparticles Become Bulklike: Properties of Realistic TiO_2 Nanoparticles in the 1–6 nm Size Range from All Electron Relativistic Density Functional Theory Based Calculations, *J. Chem. Theory Comput.* 2017, **13** 1785–1793.

(52) B. Medasani, M. Haranczyk, A. Canning and M. Asta, Vacancy formation energies in metals: A comparison of Meta-GGA with LDA and GGA exchange–correlation functionals, *Comput. Mater. Sci.* 2015, **101** 96–107.

(53) F. Passek and M. Donath, Magnetic surface state becomes nonmagnetic by oxygen adsorption, *Phys. Rev. Lett.* 1993, **71** 2122–2125.

(54) B Legendre and M Sghaier, Curie temperature of nickel, *J. Therm. Anal. Calorim.* 2011, **105** 141–143.

(55) J. Wellendorff, K. T. Lundgaard, K. W. Jacobsen and T. Bligaard, mBEEF: An accurate semi-local Bayesian error estimation density functional, *J. Chem. Phys.* 2014, **140** 144107.

(56) J. P. Perdew, A. Ruzsinszky, G. I. Csonka, O. A. Vydrov, G. E. Scuseria, L. A. Constantin, X. Zhou and K. Burke, Restoring the Density-Gradient Expansion for Exchange in Solids and Surfaces, *Phys. Rev. Lett.* 2008, **100** 136406.

(57) S. Lehtola, C. Steigemann, M. J. Oliveira and M. A. Marques, Recent developments in libxc — A comprehensive library of functionals for density functional theory, *SoftwareX* 2018, **7** 1–5.

(58) M. Asadikiya, V. Drozd, S. Yang and Y. Zhong, Enthalpies and elastic properties of Ni-Co binary system by *ab initio* calculations and an energy comparison with the CALPHAD approach, *Mater. Today Commun.* 2020, **23** 100905.

(59) P. Janthon, S. A. Luo, S. M. Kozlov, F. Viñes, J. Limtrakul, D. G. Truhlar and F. Illas, Bulk Properties of Transition Metals: A Challenge for the Design of Universal Density Functionals, *J. Chem. Theory Comput.* 2014, **10** 3832–3839.

(60) P. M. Spurgeon, D.-J. Liu, J. Oh, J. W. Evans and P. A. Thiel, Identification of an AgS_2 Complex on Ag(110), *Sci. Rep.* 2019, **9** 19842.

(61) F. Birch, Finite Elastic Strain of Cubic Crystals, *Phys. Rev.* 1947, **71** 809–824.

(62) C. G. Broyden, The Convergence of a Class of Double-rank Minimization Algorithms 1. General Considerations, *IMA J. Appl.* 1970, **6** 76–90.

(63) R. Fletcher, A new approach to variable metric algorithms, *Comput. J.* 1970, **13** 317–322.

(64) D. F. Shanno, Conditioning of Quasi-Newton Methods for Function Minimization, *Math. Comput.* 1970, **24** 647–656.

(65) D. Goldfarb, A Family of Variable-Metric Methods Derived by Variational Means, *Math. Comput.* 1970, **24** 23–26.

(66) S. S. Yadavalli, G. Jones and M. Stamatakis, DFT benchmark studies on representative species and poisons of methane steam reforming on Ni (111), *Phys. Chem. Chem. Phys.* 2021, **23** 15601–15612.

(67) H. Meltzman, D. Chatain, D. Avizemer, T. M. Besmann and W. D. Kaplan, The equilibrium crystal shape of nickel, *Acta Mater.* 2011, **59** 3473–3483.

(68) J. C. Boettger, Nonconvergence of surface energies obtained from thin-film calculations, *Phys. Rev. B* 1994, **49** 16798–16800.

(69) M. Kick, K. Reuter and H. Oberhofer, Intricacies of DFT+U, Not Only in a Numeric Atom Centered Orbital Framework, *J. Chem. Theory Comput.* 2019, **15** 1705–1718.

(70) R. H. French, Electronic Band Structure of Al_2O_3, with Comparison to Alon and AIN, *J. Am. Ceram. Soc.* 1990, **73** 477–489.

(71) S. Wilson, The dehydration of boehmite, γ-AlOOH, to γ-Al_2O_3, *J. Solid State Chem.* 1979, **30** 247–255.

(72) S. Carstens, R. Meyer and D. Enke, Towards Macroporous α-Al_2O_3—Routes, Possibilities and Limitations, *Materials* 2020, **13** 1787.

(73) L. Kavan, M. Grätzel, S. E. Gilbert, C. Klemenz and H. J. Scheel, Electrochemical and Photoelectrochemical Investigation of Single-Crystal Anatase, *JACS* 1996, **118** 6716–6723.

(74) T. Arlt, M. Bermejo, M. A. Blanco, L. Gerward, J. Z. Jiang, J. Staun Olsen and J. M. Recio, High-pressure polymorphs of anatase TiO_2, *Phys. Rev. B* 2000, **61** 14414–14419.

(75) M. Arrigoni and G. K. H. Madsen, A comparative first-principles investigation on the defect chemistry of TiO_2 anatase, *J. Chem. Phys.* 2020, **152** 044110.

(76) S. Phoka, P. Laokul, E. Swatsitang, V. Promarak, S. Seraphin and S. Maensiri, Synthesis, structural and optical properties of CeO_2 nanoparticles synthesized by a simple polyvinyl pyrrolidone (PVP) solution route, *Mater. Chem. Phys.* 2009, **115** 423–428.

(77) L. Gerward, J. Staun Olsen, L. Petit, G. Vaitheeswaran, V. Kanchana and A. Svane, Bulk modulus of CeO_2 and PrO_2—An experimental and theoretical study, *J. Alloys Compd.* 2005, **400** 56–61.

(78) F. B. Baker, E. J. Huber, C. E. Holley and N. Krikorian, Enthalpies of formation of cerium dioxide, cerium sesquicarbide, and cerium dicarbide, *J. Chem. Thermodyn.* 1971, **3** 77–83.

(79) S. S. Akimenko, G. D. Anisimova, A. I. Fadeeva, V. F. Fefelov, V. A. Gorbunov, T. R. Kayumova, A. V. Myshlyavtsev, M. D. Myshlyavtseva and P. V. Stishenko, SuSMoST: Surf. Sci. Modeling and Simulation Toolkit, *J. Comput. Chem.* 2020, **41** 2084–2097.

(80) S. S. Akimenko, V. A. Gorbunov, A. V. Myshlyavtsev and P. V. Stishenko, Generalized lattice-gas model for adsorption of functional organic molecules in terms of pair directional interactions, *Phys. Rev. E* 2016, **93** 062804.

(81) D. Landau and K. Binder, *A guide to Monte Carlo simulations in statistical physics*, Cambridge University Press, 2021.

(82) D. J. Earl and M. W. Deem, Parallel tempering: Theory, applications, and new perspectives, *Phys. Chem. Chem. Phys.* 2005, **7** 3910–3916.

(83) H. Huo and M. Rupp, Unified representation of molecules and crystals for machine learning, *Mach. Learn.: Sci. Technol.* 2022, **3** 045017.

(84) L. Himanen, M. O. J. Jäger, E. V. Morooka, F. Federici Canova, Y. S. Ranawat, D. Z. Gao, P. Rinke and A. S. Foster, DScribe: Library of descriptors for machine learning in materials science, *Comput. Phys. Commun.* 2020, **247** 106949.

(85) J. Laakso, L. Himanen, H. Homm, E. V. Morooka, M. O. J. Jäger, M. Todorović and P. Rinke, Updates to the DScribe library: New descriptors and derivatives, *J. Chem. Phys.* 2023, **158** 234802.

(86) F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, Scikit-learn: Machine Learning in Python, *JMLR* 2011, **12** 2825–2830.

(87) I. Batatia, D. P. Kovacs, G. Simm, C. Ortner and G. Csányi, MACE: Higher order equivariant message passing neural networks for fast and accurate force fields, *Adv. Neural Inf. Process.* 2022, **35** 11423–11436.

(88) I. Batatia, P. Benner, Y. Chiang, A. M. Elena, D. P. Kovács, J. Riebesell, X. R. Advincula, M. Asta, M. Avaylon, W. J. Baldwin et al., A foundation model for atomistic materials chemistry, *arXiv preprint: 2401.00096* 2023.

(89) A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder and K. A. Persson, Commentary: The Materials Project: A materials genome approach to accelerating materials innovation, *APL Mater.* 2013, **1** 011002.

(90) B. Deng, *Materials Project Trajectory (MPtrj) Dataset*, 2023.

(91) M. M. Ghahremanpour, P. J. van Maaren and D. van der Spoel, The Alexandria library: a quantum-chemical database of molecular properties for force field development, *Sci. Data* 2018, **5** 180062.

(92) K. D. B. J. Adam et al., A method for stochastic optimization, *arXiv preprint: 1412.6980* 2014, **1412**.

(93) P. Promhuad, B. Sawatmongkhon, K. Theinnoi, T. Wongchang, N. Chollacoop, E. Sukjit, S. Tunmee and A. Tsolakis, Effect of metal oxides (CeO2, ZnO, TiO_2, and Al_2O_3) as the support for silver-supported catalysts on the catalytic oxidation of diesel particulate matter, *ACS Omega* 2024, **9** 19282–19294.

(94) J. R. Sietsma, A. Jos van Dillen, P. E. de Jongh and K. P. de Jong, in *Scientific Bases for the Preparation of Heterogeneous Catalysts*, ed. E. Gaigneaux, M. Devillers, D. De Vos, S. Hermans, P. Jacobs, J. Martens and P. Ruiz, Elsevier, 2006, vol. 162, pp. 95–102.

(95) M. A. Hefnawy, S. A. Fadlallah, R. M. El-Sherif and S. S. Medany, Systematic DFT studies of CO-Tolerance and CO oxidation on Cu-doped Ni surfaces, *J Mol Graph Model.* 2023, **118** 108343.

(96) Y. Bai, D. Kirvassilis, L. Xu and M. Mavrikakis, Atomic and molecular adsorption on Ni(111), *Surf. Sci.* 2019, **679** 240–253.

(97) V. Alexandrov, M. L. Sushko, D. K. Schreiber, S. M. Bruemmer and K. M. Rosso, Adsorption and diffusion of atomic oxygen and sulfur at pristine and doped Ni surfaces with implications for stress corrosion cracking, *Corros. Sci.* 2016, **113** 26–30.

(98) L. Liu, C. Zhang, W. Wang, G. Li and B. Zhu, Decomposition of SO_2 on Ni(111) Surface and the Effect of Metal Doping: A First-Principles Study, *Molecules* 2023, **28** 6739.

(99) T. Yokoyama, S. Terada, A. Imanishi, Y. Kitajima, N. Kosugi and T. Ohta, Adsorption of SO_2 on Ni(100) and Ni(111) surfaces, *J. Electron Spectrosc. Relat. Phenom.* 1996, **80** 161–164.

(100) N. K. Das and W. A. Saidi, Effects of Cr-doping on the adsorption and dissociation of S, SO, and SO_2 on Ni(111) surfaces, *J. Chem. Phys.* 2017, **146** 154701.

(101) J. Zieliński, Effect of alumina on the reduction of surface nickel oxide; morphology of the surfaces of the surfaces of Ni/Al_2O_3 catalysts, *J. Mol. Catal.* 1993, **83** 197–206.

(102) S. Maluf and E. Assaf, Ni catalysts with Mo promoter for methane steam reforming, *Fuel* 2009, **88** 1547–1553.

(103) Y. Guo, T. P. Tran, L. Zhou, Q. Zhang and H. Kameyama, Steam Methane Reforming Using an Anodic Alumina Supported Nickel Catalyst ($NiAl_2O_3$/Alloy): Analysis of Catalyst Deactivation, *J. Chem. Eng. Jpn.* 2007, **40** 1221–1228.

(104) D. Li, Q. Zhu, Z. Bao, L. Jin and H. Hu, New insight and countermeasure for sulfur poisoning on nickel-based catalysts during dry reforming of methane, *Fuel* 2024, **363** 131045.

(105) X. Wang, D. Santos-Carballal and N. H. de Leeuw, Cation doping and oxygen vacancies in the orthorhombic $FeNbO_4$ material for solid oxide fuel cell applications: A density functional theory study, *J. Chem. Phys.* 2024, **160** 154713.

(106) R. Prasad, R. Benedek and M. M. Thackeray, Dopant-induced stabilization of rhombohedral $LiMnO_2$ against Jahn-Teller distortion, *Phys. Rev. B* 2005, **71** 134111.

(107) J. P. Allen and G. W. Watson, Occupation matrix control of d- and f-electron localisations using DFT+U, *Phys. Chem. Chem. Phys.* 2014, **16** 21016–21031.

(108) Z. Chafi, N. Keghouche and C. Minot, DFT study of Ni–CeO_2 interaction: Adsorption and insertion, *Surf. Sci.* 2007, **601** 2323–2329.

(109) E. F. de Souza, C. A. Chagas, R. L. Manfro, M. M. V. M. Souza, R. Bicca de Alencastro and M. Schmal, Combined DFT and experimental study of the dispersion and interaction of copper species in Ni-CeO_2 nanosized solid solutions, *RSC Adv.* 2016, **6** 5057–5067.

(110) J. A. Santana, J. Kim, P. R. C. Kent and F. A. Reboredo, Successes and failures of Hubbard-corrected density functional theory: The case of Mg doped $LiCoO_2$, *J. Chem. Phys.* 2014, **141** 164706.

(111) K. K. Ghuman and C. V. Singh, A DFT+U study of (Rh, Nb)-codoped rutile TiO_2, *J. Phys.: Condens. Matter.* 2013, **25** 085501.

(112) B. J. Morgan, D. O. Scanlon and G. W. Watson, Small polarons in Nb- and Ta-doped rutile and anatase TiO_2, *J. Mater. Chem.* 2009, **19** 5175–5178.

(113) A. Raghav, K. Hongo, R. Maezono and E. Panda, Electronic structure and effective mass analysis of doped TiO_2 (anatase) systems using DFT+U, *Comput. Mater. Sci.* 2022, **214** 111714.

(114) Y. Zhu, W. Wang, G. Chen, H. Li, Y. Zhang, C. Liu, H. Wang, P. Cheng, C. Chen and G. Seong, Influence of Ni Doping on Oxygen Vacancy-Induced Changes in Structural and Chemical Properties of CeO_2 Nanorods, *Crystals* 2024, **14** 746.

(115) B. Liu, H. Xu and Z. Zhang, Temperature-Programmed Surface Reaction Study of Adsorption and Reaction of H_2S on Ceria, *Chin. J. Catal.* 2012, **33** 1631–1635.

(116) W. Shan, M. Luo, P. Ying, W. Shen and C. Li, Reduction property and catalytic activity of $Ce_{1x}Ni_xO_2$ mixed oxide catalysts for CH_4 oxidation, *Appl. Catal. A: Gen.* 2003, **246** 1–9.

(117) T. Hamzehlouyan, C. Sampara, J. Li, A. Kumar and W. Epling, Sulfur Poisoning of a Pt/Al_2O_3 Oxidation Catalyst: Understanding of SO_2, SO_3 and H_2SO_4 Impacts, *Top. Catal.* 2016, **59** 1028–1032.

(118) M. Smirnov, A. Kalinkin, A. Pashis, A. Sorokin, A. Noskov, V. Bukhtiyarov, K. Kharas and M. Rodkin, Comparative XPS study of Al_2O_3 and CeO_2 sulfation in reactions with SO_2, SO_2+O_2, SO_2+H_2O, and SO_2+O_2+H_2O, *Kinet. Catal.* 2003, **44** 575–583.

Chapter 6

Outlook

6.1 Conclusions

This thesis has investigated the application of advanced computational methods and workflows for modelling defect-driven phenomena in catalysis, with the aim of rationalising experimental observations, uncovering future catalyst optimisation strategies and demonstrating capabilities that extend far beyond the model systems selected. Achieving this has required addressing two of the most significant barriers in computational materials modelling: the Coulomb self-interaction error and the prohibitive computational cost of simulating large systems that are more representative of catalysts under reaction conditions.

In Chapter 3, we investigated the challenges of accurately modelling polarons in transition metal oxides, focusing on the model systems of Nb- and W-doped TiO_2 (NTO and WTO, respectively), which are promising materials for next-generation solar cells and photocatalysis. There is currently no explanation in the literature for the polymorph-sensitivity of polaron formation in doped anatase *vs*. rutile TiO_2, as characterised by Dr Andrea Folli using electron paramagnetic resonance (EPR) spectroscopy. EPR spectroscopy shows that charge compensation is highly sensitive to the TiO_2 polymorph, with Nb^{4+} and W^{5+} signals present in substitutionally doped rutile but not in doped anatase. The observations are rationalised by DFT+U calculations in an all electron numerical atom-centred orbital (NAO) framework, which required modification of the Ti 3d Hubbard projector to enable self-consistent resolution of the Ti 3d, Nb 4d and W 5d orbital occupancies. Self-consistent DFT+U predicts favourability of Nb^{4+} in rutile NTO through greater filling of the Nb 4d t_{2g} orbitals and reduced filling of the e_g orbitals compared to anatase NTO. Self-consistent DFT+U also predicts W^{5+} in rutile WTO through the formation of a localised mid-gap state of $5d_{z^2}$ character that is not formed in anatase WTO. Given the numerical instabilities and inaccuracies of self-consistent and constrained DFT+U calculations with an atomic Ti 3d Hubbard projector, respectively, it is essential to consider the effect of the Hubbard projector on DFT+U simulations in more detail.

Chapter 4 built upon the findings in Chapter 3 by developing machine learning-based workflows to guide the determination of Hubbard U values and projectors, to ensure the DFT+U method is robust across a wide range of simulations of defects in TMOs and REOs. Simultaneous optimisation of the Hubbard U value and projector has been demonstrated semi-empirically using symbolic regression and support vector machines, before using Bayesian optimisation to minimise the errors of target properties relative to experimental references. The Ti 3d Hubbard U value and projector have been semi-empirically refined to enable self-consistent DFT+U simulations of intrinsic and extrinsic defects

in both anatase and rutile TiO_2. The outcome is DFT+U simulations with comparable accuracy to hybrid-DFT in terms of the relative stabilities of point defects and the formation of localised Holstein polarons, whilst at orders of magnitude lower cost. The DFT+U-predicted occupation matrices reproduce the hybrid-DFT O 2p occupation matrix, which can be an effective cost function for a first-principles strategy for Hubbard U value and projector optimisation.

The semi-empirical approach was therefore redefined as a first-principles approach and generalised across materials using hierarchical symbolic regression to screen the Hubbard parameter space using empirical correlations to learn the DFT+U potential energy surface in terms of orbital occupancies. Predictions of metal d or f orbital and O 2p orbital occupancies were made in terms of Hubbard parameters, basis set parameters, DFT-predicted orbital occupancies and atomic material descriptors. The first-principles approach enables the development of a generalised workflow for the one-shot computation of Hubbard U values and projectors for different materials that requires no successive DFT+U calculations, as in active learning schemes. The method transferability is demonstrated for 10 prototypical TMOs and REOs (anatase and rutile TiO_2, Cu_2O, MoO_3, WO_3, Y_2O_3, ZrO_2, CeO_2, $LiCoO_2$ and $LiFePO_4$), which each require one reference DFT and hybrid-DFT calculation as inputs, whilst generating families of solutions for each material, *i.e.*, optimised Hubbard projectors for a given Hubbard U value. Upon validating a subset of these solutions, a MAE of 5.02% for the DFT+U-predicted O 2p orbital occupancies was achieved, with demonstrated accuracy for materials unseen from model training ($LiCoO_2$ and $LiFePO_4$).

Predicting the numerical stability of point defect calculations can also be generalised across materials using symbolic classification, using Hubbard U values and material-dependent descriptors of covalency, enabling the determination of Hubbard U values and projectors that are robust against numerical instability. The validity of Hubbard U values and projectors determined from first-principles has been investigated for the self-consistent simulation of Mg-doped and oxygen deficient $LiCoO_2$, where refining the Co 3d Hubbard projector enables the numerically stable simulation of experimentally reported charge compensation mechanisms driving the material's high electrical conductivity. The same results were not possible using an atomic Co 3d Hubbard projector and did not require any prior testing of suitable Co 3d Hubbard U values or projectors. Overall Chapter 4 demonstrates how supervised machine learning can be used to construct scalable and transferable solutions to the challenge of DFT+U parameterisation, showing promise for extension to schemes requiring multiple parameter optimisation as well as other electronic structure codes.

Chapter 5 broadened the scope of our investigations into metal oxide support effects towards industrially relevant H_2 production catalysts, with the aim of applying advanced computational workflows to investigate strategies for future catalyst optimisation. Chapter 5 adopted a combined computational and experimental approach to examine the atomic level mechanisms that govern the sulfur tolerance of supported Ni catalysts for MSR and low-temperature processes from renewable feedstocks. Combining DFT, grand canonical Monte Carlo (GCMC) sampling and a fine-tuned MACE machine learned interatomic potential (MLIP), we show that a high oxygen chemical potential provided *via* support oxygen buffering is not sufficient alone for the removal of adsorbed S from Ni(111), with thermal activation being essential. The results support experimental MSR activity tests showing that the catalytic activity of Ni supported on reducible CeO_2 can be readily restored from a poisoned state at high temperatures, compared to Ni supported on less reducible TiO_2 and γ-Al_2O_3. The results are

further validated using DFT+U computed oxygen vacancy formation energies for the bulk support materials, which show the ease of oxygen vacancy formation in the order CeO_2 > TiO_2 > γ-Al_2O_3. The MSR activity testing also indicates the critical role of phase transformations into catalytically inactive phases, which is widely reported to occur for Ni/γ-Al_2O_3, and that agrees with our DFT+U computed defect energies for substitutional Ni doping, which indicate the initial stages of bulk phase transformation are more favourable in the order γ-Al_2O_3 > TiO_2 > CeO_2. TPD-MS and XPS highlight the critical role of water in the formation of thermally stable sulfate species that can increase the temperatures required for catalyst regeneration.

Overall, the computational workflows and experimental characterisation points to three critical aspects for the rational design of metal oxide support materials for sulfur tolerant catalysts: (1) the feasibility of bulk oxygen vacancy formation in the support; (2) the resistance of the bulk support to phase transformations into catalytically inactive solid solutions; and (3) the support- and temperature-dependent surface chemistry of SO_2 to sulfates. The integration of *ab initio* computational modelling, statistical sampling and machine learning further demonstrates the importance of advanced workflows for studying complex catalytic materials in a manner that faithfully bridges theory and experiment.

6.2 Future Work

This thesis investigates a coherent strategy for advancing the accuracy of simulations of defect-driven catalytic phenomena, through mitigating the SIE with machine learning-parameterised DFT+U simulations and coupling DFT with statistical sampling and MLIPs for simulations beyond atomistic regimes. Looking forward, these approaches point to a future where physics-informed machine learning and multiscale modelling are seamlessly integrated to enable predictive and cost-effective simulations of complex catalytic materials under reaction conditions. Such developments will accelerate the rational design of catalysts by bridging the gap between atomistic modelling and experimental performance.

6.2.1 Fast, Accurate and Robust DFT+U Parameterisation

Chapter 4 demonstrated how supervised machine learning can be applied to simultaneously optimise Hubbard U values and projectors to improve the accuracy and numerical stability of DFT+U in a NAO framework; however, there remains clear avenues for further research that are not accounted for in the current work.

Extension to Challenging Open-Shell Systems

The methods and results presented generally hold well for closed-shell systems or those with ions in a low-spin state. In contrast, significant challenges were encountered when applying DFT+U to open-shell systems containing ions in a high spin state, *e.g.*, Co^{2+} in CoO, Mn^{4+} in MnO_2, Fe^{3+} in Fe_2O_3 and Cr^{3+} in Cr_2O_3. For these materials, DFT+U was found to consistently suffer from SCF non-convergence as a result of charge sloshing, *i.e.*, excessive oscillations in the charge density, which even occurred during single point calculations of small unit cells, irrespective of the chosen Hubbard U values and projectors. Upon analysing the DFT-predicted ground state electronic structures for

these materials, all cases are incorrectly predicted to be metallic with zero band gap; therefore, we hypothesise that the starting electronic structure from DFT makes SCF convergence for DFT+U very challenging in a NAO framework. After further testing, switching to spin polarisation in combination with meta-GGA exchange-correlation density functionals and carefully selected SCF mixing parameters has been found to improve the DFT-predicted electronic structure to restore a band gap (albeit which is underestimated compared to experiment). These changes make the convergence of DFT+U with refined Hubbard projectors feasible, opening the possibility of directly translating the work in Chapter 4 to defects in magnetic transition metal and rare-earth metal oxides. Whilst this remains an unresolved challenge at the moment, we note that the restoration of semiconducting or insulating ground states from an initial DFT-predicted metallic state is routinely achievable in planewave implementations of DFT+U, [1] therefore a more detailed investigation into exact differences between implementations may help.

Universal Models *vs*. Active Learning

Section 4.3.3 outlined the development of a single universal model for parameterising Hubbard U values and projectors across materials. This approach served as a test for how well the insights gained from TiO_2 transfer across chemical space, with the ultimate goal of enabling accurate, self-consistent simulations of defects and polarons across a broad range of TMOs and REOs. A valid alternative approach could be an iterative active learning scheme, where U, c_1 and c_2 are progressively refined so that the DFT+U-predicted electronic structure resembles that from hybrid-DFT, *e.g.*, *via* comparison of orbital occupation numbers or band structures. This active learning approach may work well for systems that are not accurately incorporated within a single universal model, but would require repeating for different materials and incur a potentially significant cost for the total number of DFT+U calculations required. Section 4.3.2 shows that Bayesian optimisation can efficiently sample a SISSO-derived semi-empirical cost function, therefore this could be repurposed to iteratively perform DFT+U geometry optimisation calculations for a small unit cell, sampling Hubbard parameters and targeting orbital occupancies calculated using hybrid-DFT. If the cost function landscape is very difficult to optimise with an iterative approach, both the universal and active learning schemes could be combined, *i.e.*, the outputs from the one-shot approach used as initial positions for active learning, which may improve convergence to good solutions whilst requiring fewer DFT+U iterations.

Compatibility with High-Throughput Simulations

The current approach outputs several combinations of Hubbard U values and projectors, which serve as optimised candidate parameters for validation with the aim of enabling self-consistent defect simulations. Further development into a high-throughput optimisation requires an improvement in accuracy of the generalised symbolic regression model on out-of-training data, which should be feasible since this work considers a very small training set of 197 sets of DFT+U-calculated orbital occupancies from geometry optimised unit cells, of which nearly half correspond to TiO_2. With a better extrapolative accuracy, the generalised approach could be used to output a single Hubbard parameter, rather than relying on K-means clustering to output several candidate parameters for further

validation. The reliance on small unit cells further reduces the computational overhead associated with generating corresponding hybrid-DFT reference data.

Choice of High-Level Reference and Extension to Other Codes

Chapter 4 uses hybrid-DFT as the high-level reference method to obtain target orbital occupancies for metal d or f and O $2p$ states. This choice was made in order to establish a consistent and computationally tractable benchmark against which machine learning–optimised U values and projectors could be compared. While this introduces a dependency on the choice of reference, it is entirely possible within our approach to select another level of theory depending on the target material and available computational resources. The approach could equally be extended to optimise against alternative physics-informed quantities, *e.g.*, minimising errors in band structure plots, enforcing piecewise linearity or maximising occupation matrix idempotency. In the case of magnetic systems, for which identifying the true ground state is complicated by the presence of metastable states in the potential energy surface, one could instead target experimentally observed magnetic orderings as the optimisation criteria.

The methods presented could also be extended to other electronic structure codes, for which there already exist several schemes that require simultaneous parameter optimisation, *e.g.*, DFT+U+V and orbital resolved DFT+U. [2, 3] In such approaches, rather than tuning linear expansion coefficients as in our NAO-based approach, one would instead adjust the definition of the Hubbard projector *via* the projector augmented wave (PAW) augmentation radius in a planewave basis, [4] or the muffin-tin radius in a linear augmented planewave (LAPW) basis, both of which have been reported to significantly influence the predicted geometric, energetic and electronic properties of complex oxides. [5, 6] Direct adaptation of the approach for other electronic structure codes will be essential, as Hubbard parameters are generally not transferable across codes with different basis sets and definitions of the Hubbard projector. As discussed by Kick *et al.*, DFT+U in a NAO framework typically requires U values ~1-2 eV smaller than other formalisms to achieve comparable charge localisation. [7] This observation supports the results of this work concerning Ti $3d$ Hubbard parameters for simulating TiO_2 (U = ~ 2.5-3 eV) *vs.* literature reported values computed in a real-space formalism (U = ~ 5-6 eV). [8]

6.2.2 Redox-Aware Machine Learned Interatomic Potentials

Robust schemes for DFT+U parameterisation from first-principles offer immense promise in materials discovery, where automated workflows can determine corrective parameters for vast numbers of new materials, before being used to train "*redox-aware*" MLIPs for scalable simulations of changing oxidation states in catalytic materials. [9, 10] DFT+U-parameterised MLIPs offer tremendous promise for large-scale simulations of defect-driven phenomena in heterogeneous catalysis, as is reported for the accurate modeling of reconstructions of non-stoichiometric TiO_2 surfaces and the resulting strong adsorption and geometry of adsorbed catalytic nanoparticles; [11–13] a strategy that could easily be applied to extend the work in Chapter 5, by investigating sulfur poisoning of supported Ni nanoparticles on metal oxide surfaces, to better account for critical effects at the metal-support interface that we did not consider.

While pre-trained foundation model MLIPs are emerging, their direct application to modelling defects in complex oxides is limited, as fine-tuning remains essential to capture charge transfer and localisation. Such fine-tuning is expensive as it requires high-fidelity datasets generated from expensive electronic structure calculations, with the accuracy of the resulting MLIP tightly bound to the quality and diversity of this training data. The machine learnining-based workflows for DFT+*U* parameterisation discussed in Chapter 5 offers a promising solution, by facilitating accurate and robust simulations of defects and polarons in transition metal and rare-earth metal oxides with almost equivalent cost as standalone DFT. This now enables the generation of much improved datasets for fine-tuning MLIPs, representing a key step towards accurate and scalable redox-aware MLIPs for simulating defect-rich systems on length scales that are totally intractable to simulate using DFT or beyond-DFT methods.

References

(1) N. E. Kirchner-Hall, W. Zhao, Y. Xiong, I. Timrov and I. Dabo, Extensive benchmarking of DFT+*U* calculations for predicting band gaps, *Appl. Sci.* 2021, **11** 2395.

(2) I. Timrov, F. Aquilante, M. Cococcioni and N. Marzari, Accurate Electronic Properties and Intercalation Voltages of Olivine-Type Li-Ion Cathode Materials from Extended Hubbard Functionals, *PRX Energy* 2022, **1** 033003.

(3) E. Macke, I. Timrov, N. Marzari and L. C. Ciacchi, Orbital-Resolved DFT+*U* for Molecules and Solids, *J. Chem. Theory Comput.* 2024, **20** 4824–4843.

(4) Z. Wang, C. Brock, A. Matt and K. H. Bevan, Implications of the DFT + *U* method on polaron properties in energy materials, *Phys. Rev. B* 2017, **96** 125150.

(5) Y.-C. Wang, Z.-H. Chen and H. Jiang, The local projection in the density functional theory plus U approach: A critical assessment, *The Journal of Chemical Physics* 2016, **144** 144106.

(6) K. Park, M. Raman, A.-J. Olatunbosun and J. Pohlmann, Revisiting DFT+*U* calculations of TiO_2 and the effect of the local-projection size, *AIP Adv.* 2024, **14** 065114.

(7) M. Kick, K. Reuter and H. Oberhofer, Intricacies of DFT+*U*, Not Only in a Numeric Atom Centered Orbital Framework, *J. Chem. Theory Comput.* 2019, **15** 1705–1718.

(8) S. Bhowmik, A. J. Medford and P. Suryanarayana, Real-space Hubbard-corrected density functional theory, *arXiv preprint: 2507.23612* 2025.

(9) L. Bastonero, C. Malica, E. Macke, M. Bercx, S. Huber, I. Timrov and N. Marzari, First-principles Hubbard parameters with automated and reproducible workflows, *Npj Comput. Mater.* 2025, **11** 183.

(10) C. Malica and N. Marzari, Teaching oxidation states to neural networks, *Npj Comput. Mater.* 2025, **11** 212.

(11) Y. Lee, X. Chen, S. M. Gericke, M. Li, D. N. Zakharov, A. R. Head, J. C. Yang and A. N. Alexandrova, Machine-Learning-Driven Exploration of Surface Reconstructions of Reduced Rutile TiO_2, *Angew. Chem.* 2025, e202501017.

(12) H. Ünal, E. Mete and Ş. Ellialtıoğlu, Surface energy and excess charge in (1×2)-reconstructed rutile TiO_2 (110) from DFT+ U calculations, *Phys. Rev. B* 2011, **84** 115407.

(13) V. Çelik, H. Ünal, E. Mete and Ş. Ellialtıoğlu, Theoretical analysis of small Pt particles on rutile TiO_2 (110) surfaces, *Phys. Rev. B* 2010, **82** 205113.

Appendix A

Appendix

A.1 Software Versions

The calculations in Chapters 3 and 4 were performed with FHI-aims versions 210618 and later, which were verified to yield consistent total energies and forces. After project completion, a minor error has been identified for the DFT+U implementation when used in combination with sparse matrix formats and hydrogenic basis functions. Specifically, an inconsistency has been identified between the occupation matrix and the corresponding Hamiltonian, due to an unintended double-counting of the Hubbard U correction on off-diagonal elements of the occupation matrix. To quantify the impact of this minor error, the self-consistent DFT+U simulations in Chapters 3 with the refined Hubbard projector were repeated with the corrected implementation, corresponding to the GitLab commit "a1ab632a0890b9a1c9373bbb1d75aa0f3faf4950", for both anatase and rutile NTO and WTO. The key differences in the localisation of polarons in anatase and rutile, as discussed in the density of states plots in Figure 3.3, are unchanged between the different code versions. The defect energies with the corrected implementation increase slightly for anatase NTO and WTO, by 0.08 eV and 0.17 eV, respectively; and decrease slightly for rutile NTO and WTO, by 0.23 eV and 0.20 eV, respectively. To quantify the impact of the minor error on the results presented in Chapter 4, self-consistent DFT+U simulations of bulk anatase TiO_2 with different definitions of the Ti $3d$ Hubbard projector were repeated with the corrected implementation, using systematically varied Hubbard parameters U (1 eV, 2 eV and 3 eV), c_1 (1 and 0.8) and c_2 (0, -0.3 and -0.5). Across all combinations of Hubbard parameters, differences in the ground state total energy between different implementations were typically less than 0.03 eV, with the majority of Hubbard parameters yielding differences below 1 meV. The largest observed difference was 0.151 eV, found at $U = 3$ eV, $c_1 = 0.8$ and $c_2 = -0.5$.

The quantitative differences noted for the different implementations do not affect the qualitative trends or conclusions as presented in Chapters 3 and 4, and are noted here for integrity purposes. The DFT+U calculations presented in Chapter 5 were performed with the corrected implementation.

A.2 DFT Parameterisation: Bulk TiO_2

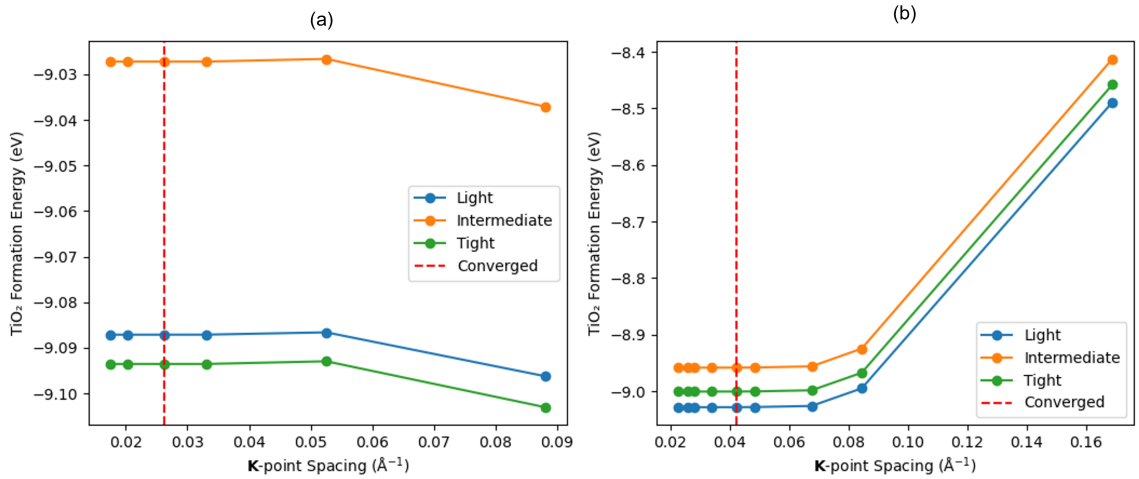


Figure A.1: Variation of the bulk TiO_2 formation energy with respect to the **k**-point spacing and basis set size (light, intermediate and tight), calculated using the PBE functional, for (a) anatase and (b) rutile. The red dashed line corresponds to the converged **k**-point spacing.

Table A.1: Comparison of DFT exchange correlation density functionals for predicting the formation energy, ΔE_{Form} (eV), band gap, E_{bg} (eV), unit cell equilibrium volume cell volume, V_0 (Å^3) for anatase and rutile TiO_2, as well as the CPU time per SCF cycle, t (s) for unit cell geometry optimisation.

Functional	ΔE_{Form}		E_{bg}		V_0		t	
	Rutile	Anatase	Rutile	Anatase	Rutile	Anatase	Rutile	Anatase
LDA	-3.42	-3.43	1.74	2.12	61.22	132.54	0.93	0.47
PBE	-3.01	-3.03	1.78	2.20	64.60	139.72	1.16	0.63
PBEsol	-3.12	-3.13	1.76	2.16	62.71	135.75	1.20	0.60
BLYP	-3.14	-3.17	1.76	2.21	66.26	143.21	1.23	0.63
M06-L	-3.12	-3.09	2.10	2.58	64.12	138.77	1.43	0.81
SCAN	-3.43	-3.42	2.08	2.59	63.08	136.41	1.48	0.78
rSCAN	-3.37	-3.37	2.11	2.59	62.95	136.18	1.51	0.79
mBEEF	-3.26	-3.28	2.21	2.65	63.50	137.26	1.69	0.94
B3LYP	-3.34	-3.36	3.56	3.93	63.98	138.30	108.05	32.69
PBE0	-3.20	-3.21	4.05	4.29	62.39	135.49	38.47	62.41

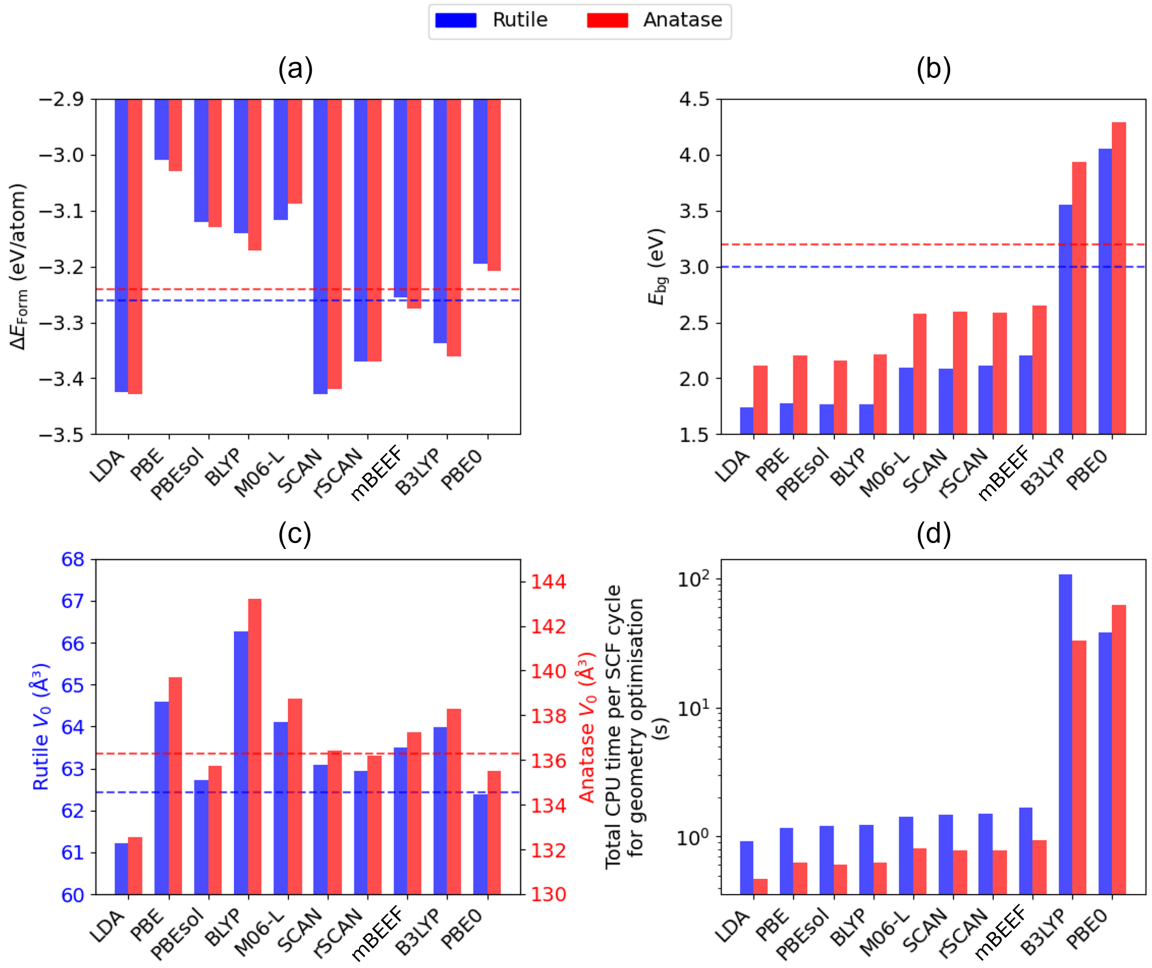


Figure A.2: Comparing the DFT-predicted (a) formation energy, (b) band gap, (c) unit cell equilibrium volume and (d) CPU time per SCF cycle for unit cell geometry optimisation for bulk anatase and rutile TiO_2 using 10 different exchange correlation density functionals. Experimental reference values are indicated by horizontal dashed lines [1–3]

A.3 Experimental Characterisation: Nb- and W-Doped TiO_2

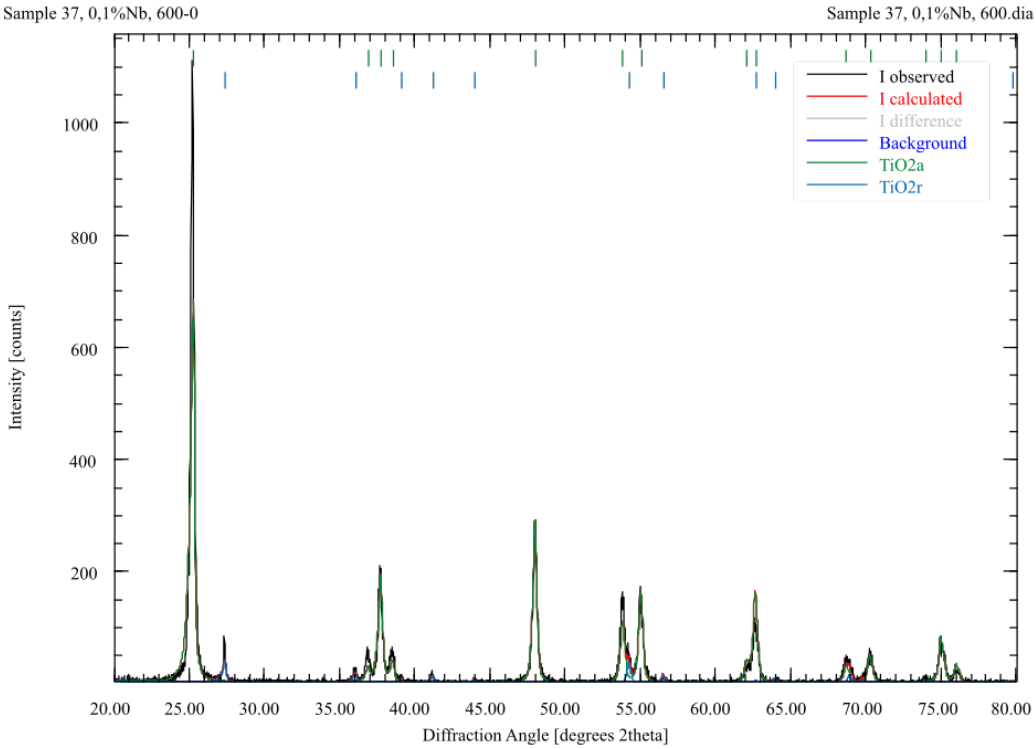


Figure A.3: Refined powder XRD pattern of $0.1\ \%_{at.}$ Nb doped TiO_2 (NTO-AR) indicating the presence of both anatase ($92\ \%$) and rutile ($8\ \%$) polymorphs. *This XRD data was collected by Dr Andrea Folli.*

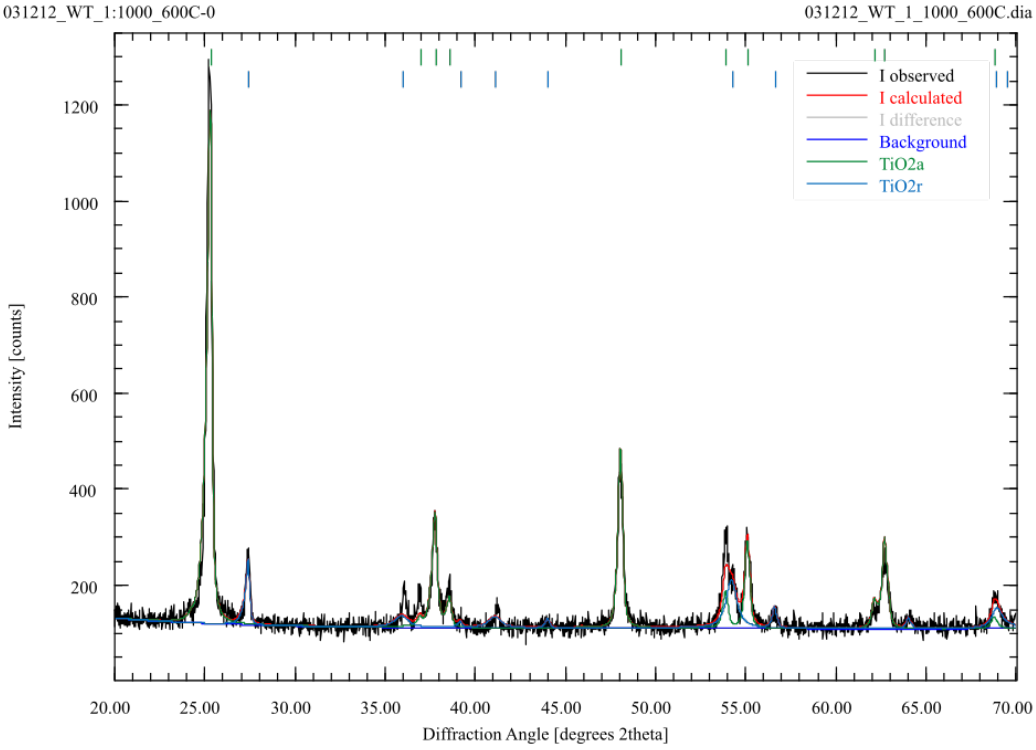


Figure A.4: Refined powder XRD pattern of $0.1\ \%_{at.}$ W doped TiO_2 (WTO-AR) indicating the presence of both anatase ($72\ \%$) and rutile ($28\ \%$) polymorphs. *This XRD data was collected by Dr Andrea Folli.*

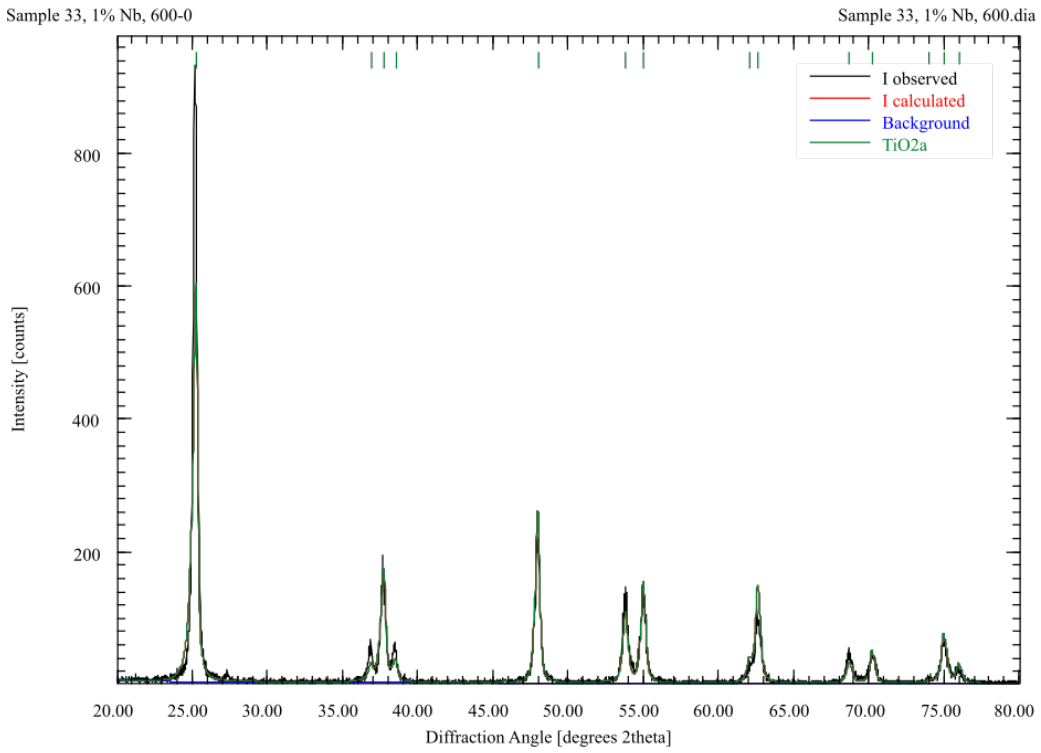


Figure A.5: Refined powder XRD pattern of $1.0\,\%_{\text{at.}}$ Nb doped TiO_2 (NTO-A) indicating the presence of anatase only polymorph. *This XRD data was collected by Dr Andrea Folli.*

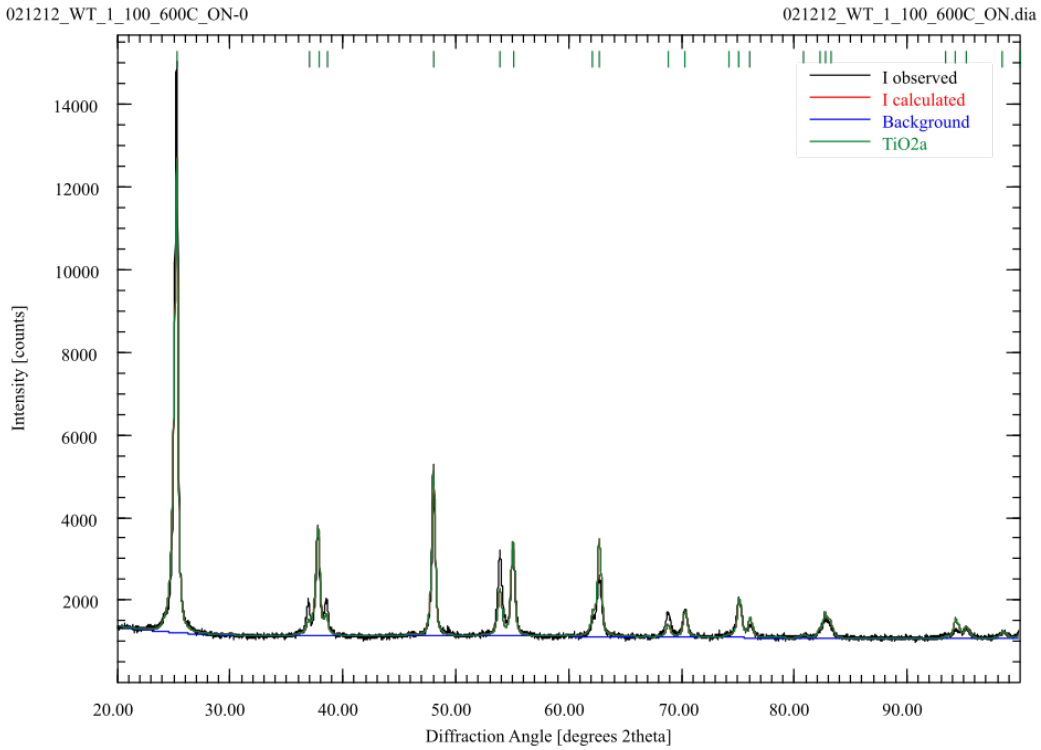


Figure A.6: Refined powder XRD pattern of $1.0\,\%_{\text{at.}}$ W doped TiO_2 (WTO-A) indicating the presence of anatase only polymorph. *This XRD data was collected by Dr Andrea Folli.*

The assignment of Nb^{4+} in Figure 3.2(a) can be explained as follows. Figures A.7(b) and A.7(c) show the simulated angular dependency of the single crystal EPR spectra at 4.2 K of Nb^{4+} centres in pure rutile NTO at X and Q band, respectively, computed using experimental values derived from a rutile single crystal. [4] The crystal structure of rutile in Figure A.7(a) is tetragonal (D_{4h}^{14}, space group P42/mnm) with two Ti per unit cell, which are equivalent except for a $\pi/2$ rotation about the c axis, as also evidenced by the orientation of the eigenframes of the magnetic tensors reported for the two inequivalent sites I and II in Figure A.7(a). The two Ti sites possess orthorhombic point symmetry D_{2h}. Details of the spin Hamiltonian used for the simulations can be found in Table A.2.

Table A.2: Spin Hamiltonian parameters of reduced dopant metal centres detected in Nb^{5+} and W^{6+} doped TiO_2. *This data was collected by Dr Andrea Folli.*

Reduced metal centre	TiO_2 polymorph	g_x	g_y	g_z	A_x *MHz*	A_y *MHz*	A_z *MHz*	Reference
Nb^{4+}	Rutile	1.970	1.985	1.941	n.d.	n.d.	n.d.	This work
Nb^{4+}	Rutile	1.973	1.981	1.948	5.0	23.8	7.0	[4]
W^{5+}	Rutile	1.594	1.473	1.443	277.5	122.4	191.1	This work
W^{5+}	Rutile	1.5944	1.4725	1.4431	277.3	122.3	191.0	[5]

Note: x, y, z are here along $[1\bar{1}0]$, $[110]$, $[001]$ respectively. $[001]$ corresponds to the rutile crystal c-axis (see also Figure A.7).

The g_x and g_y completely overlap at X band frequencies but they can be resolved at Q band frequencies. Hyperfine interaction of the $4d^1$ unpaired electron to the ^{93}Nb nucleus is expected ($I(^{93}Nb)= 9/2$, 100 % natural abundance) giving rise to 10 hyperfine lines. The A_y value is noticeably larger than A_x and A_z, with some resolved structure due to $\delta m_I = \pm 1$ transitions that are allowed *via* the electric quadrupole interaction when **B** is not aligned to the principal axes of the crystal. Hyperfine structure has been shown to vanish above 25 K on the single crystal, [4] with particularly the $\mathbf{B}//[001]$ and $\mathbf{B}//[1\bar{1}0]$ sets of lines coalescing into a single line already at 25 K. [4] The increasingly dispersive character of the signal above 25 K and the line narrowing observed upon coalescing of the hyperfine structure, was interpreted by Zimmermann [4] as a combination of two contributions, *i.e.*, thermal excitation of $4d^1$ donor electrons to the conduction band plus exchange scattering of $4d^1$ donor electrons with the conduction band. Both of these contributions give rise to electron hopping from different donor sites (*i.e.*, Nb^{4+}) *via* the conduction band (Anderson's model of random frequency modulation). The activation energy for the hopping between the donor level and the conduction band was found to be $E_a/k = 72$ K at temperatures below 40 K, while the same activation energy was much larger at temperatures up to 300 K, in agreement with more recent evidence suggesting that rutile NTO is resistive at room temperature. [6]

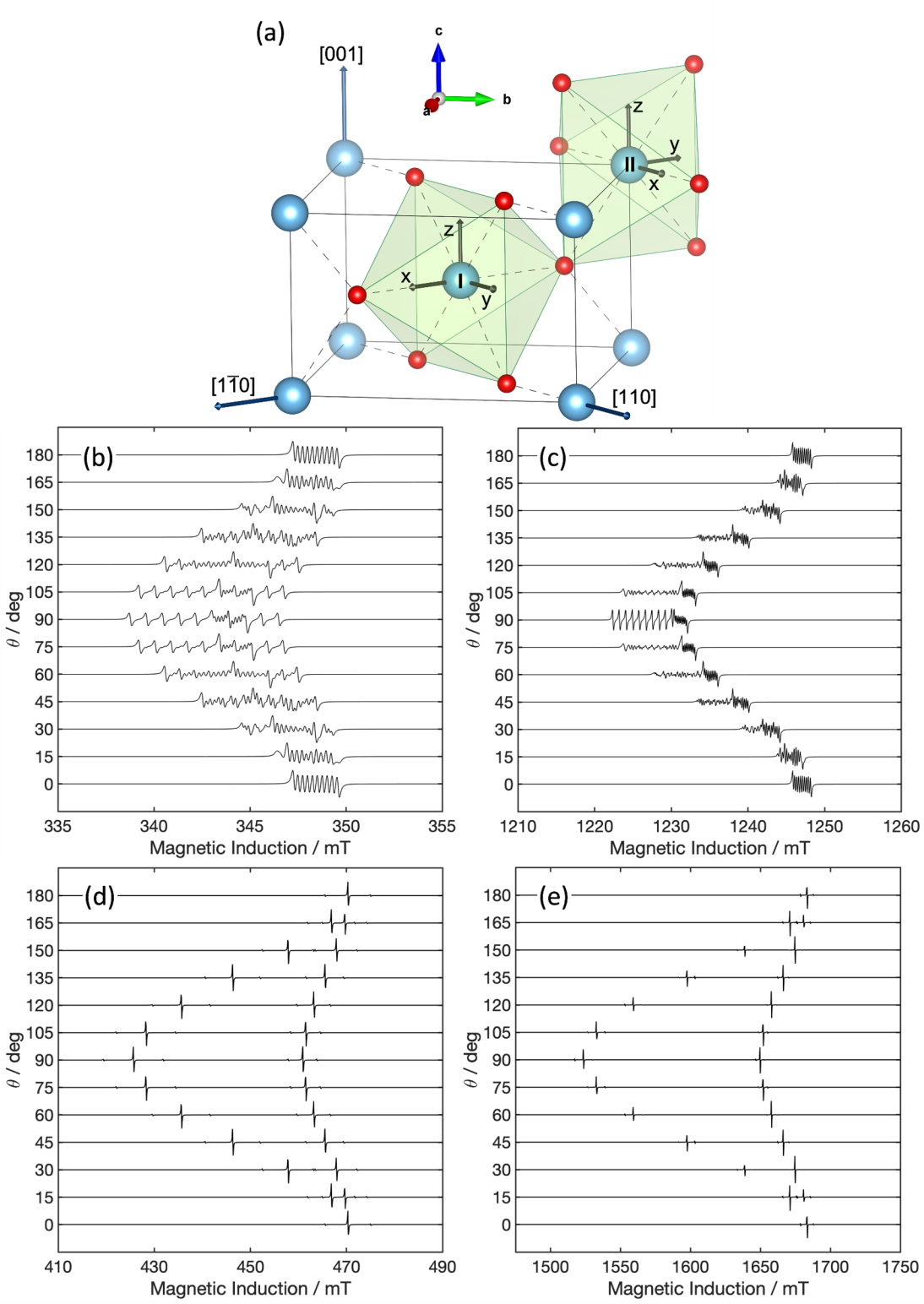


Figure A.7: (a) Tetragonal crystal structure of rutile TiO_2 with two inequivalent Ti atoms (I and II). (b)-(e) show the simulated angular dependency of the rutile single crystal EPR spectra at $4.2\,\mathrm{K}$ for Nb^{4+} centres at (b) X band and (c) Q band; and W^{5+} centres at (d) X band and (e) Q band. *These spectra were simulated by Dr Andrea Folli.*

A.4 DFT Parameterisation: Bulk Nickel

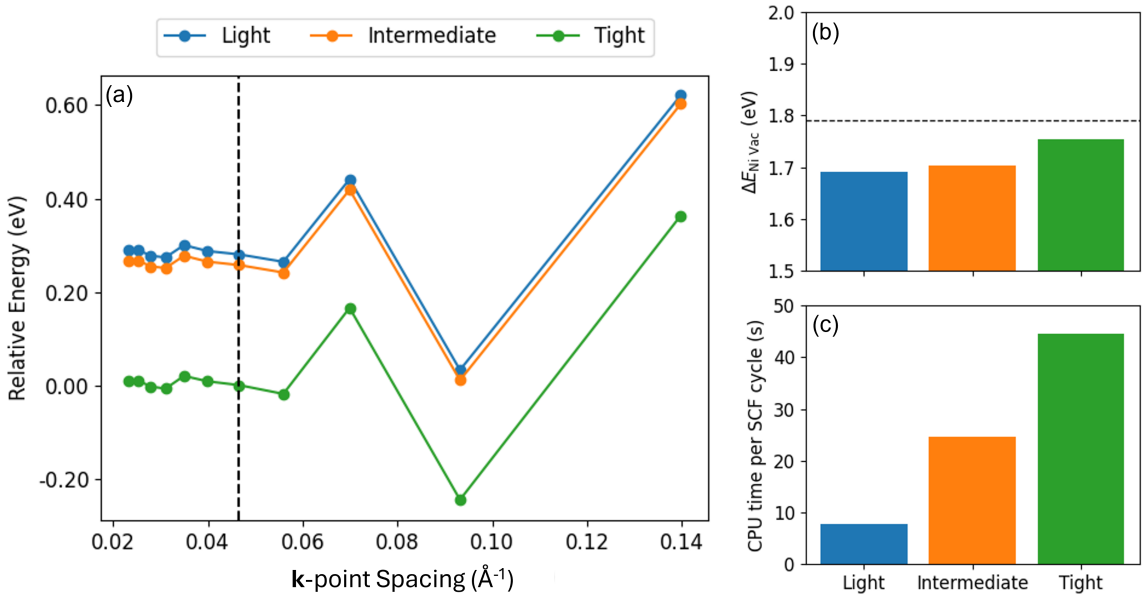


Figure A.8: Comparing the effect of the basis set size on the (a) convergence of the relative total energy (*vs.* the converged value with the tight basis set) with respect to the **k**-point spacing, (b) the bulk Ni vacancy formation energy in a $3 \times 3 \times 3$ supercell and (c) the CPU time per SCF cycle for the bulk Ni vacancy geometry optimisation simulation (all calculated using the PBE functional). The black dashed line in (a) denotes the converged **k**-point spacing and in (b) denotes the experimental defect energy. [7] The light basis set was determined to provide an adequate $\Delta E_{\text{Ni Vac}}$, whilst significantly reducing computational cost *vs.* intermediate or tight basis sets. The Ni vacancy formation energy was calculated using Equation A.1

$$\Delta E_{\text{Ni Vac}} = E_{\text{Defective Ni Bulk}} + E_{\text{Isolated Ni Atom}} - E_{\text{Stoichiometric Ni Bulk}} \tag{A.1}$$

Table A.3: DFT-calculated Ni vacancy formation energy (ΔE_{vac}), unit cell equilibrium volume (V_0), cohesive energy (ΔE_{Coh}) and average error in Equation (A.2) E^{Exc} for a range of exchange correlation density functionals. Rows are ordered from top to bottom for increasing values of E^{Exc}. Experimental reference values are included for comparison. [7–9]

Functional	ΔE_{vac} (eV)	V_0 (Å^3)	ΔE_{Coh} (eV)	E^{Exc} (%)
PBE	1.69	43.17	5.66	26.92
mBEEF	1.74	41.15	5.90	32.39
rSCAN	1.96	41.22	6.29	41.93
PBEsol	2.09	41.22	6.27	43.67
SCAN	1.97	40.84	6.46	45.66
BLYP	0.90	45.32	4.81	50.25
LDA	2.26	39.87	6.68	56.36
M06-L	2.88	42.08	6.55	76.44
Exp	1.79	43.61	4.48	N/A

$$E^{E_{\mathrm{xc}}} = \left\| \frac{100 \times \left(\Delta E_{\mathrm{Ni\ Vac}} - \Delta E^{\mathrm{Exp}}_{\mathrm{Ni\ Vac}}\right)}{\Delta E^{\mathrm{Exp}}_{\mathrm{Ni\ Vac}}} , \frac{100 \times \left(V_0 - V_0^{\mathrm{Exp}}\right)}{V_0^{\mathrm{Exp}}}, \frac{100 \times \left(\Delta E_{\mathrm{Coh}} - \Delta E^{\mathrm{Exp}}_{\mathrm{Coh}}\right)}{\Delta E^{\mathrm{Exp}}_{\mathrm{Coh}}} \right\| \tag{A.2}$$

where the cohesive energy is defined as:

$$\Delta E_{\mathrm{Coh}} = \frac{E_{\mathrm{Ni\ atom}} - E_{\mathrm{Ni\ bulk}}}{N} \tag{A.3}$$

where N is the number of atoms in the bulk.

Bibliography

(1) T. Arlt, M. Bermejo, M. A. Blanco, L. Gerward, J. Z. Jiang, J. Staun Olsen and J. M. Recio, High-pressure polymorphs of anatase TiO_2, *Phys. Rev. B* 2000, **61** 14414–14419.

(2) Y. Zhang, J. W. Furness, B. Xiao and J. Sun, Subtlety of TiO_2 phase stability: Reliability of the density functional theory predictions and persistence of the self-interaction error, *J. Chem. Phys.* 2019, **150** 014105.

(3) M. Arrigoni and G. K. H. Madsen, A comparative first-principles investigation on the defect chemistry of TiO_2 anatase, *J. Chem. Phys.* 2020, **152** 044110.

(4) P. H. Zimmermann, Temperature Dependence of the EPR Spectra of Niobium-Doped TiO_2, *Phys. Rev. B* 1973, **8** 3917–3927.

(5) T. T. Chang, Paramagnetic-Resonance Spectrum of W^{5+} in Rutile (TiO_2), *Phys. Rev.* 1966, **147** 264–267.

(6) S. X. Zhang, D. C. Kundaliya, W. Yu, S. Dhar, S. Y. Young, L. G. Salamanca-Riba, S. B. Ogale, R. D. Vispute and T. Venkatesan, Niobium doped TiO_2: Intrinsic transparent metallic anatase versus highly resistive rutile phase, *J. App. Phys.* 2007, **102** 013701.

(7) B. Medasani, M. Haranczyk, A. Canning and M. Asta, Vacancy formation energies in metals: A comparison of Meta-GGA with LDA and GGA exchange–correlation functionals, *Comput. Mater. Sci.* 2015, **101** 96–107.

(8) M. Asadikiya, V. Drozd, S. Yang and Y. Zhong, Enthalpies and elastic properties of Ni-Co binary system by *ab initio* calculations and an energy comparison with the CALPHAD approach, *Mater. Today Commun.* 2020, **23** 100905.

(9) P. Janthon, S. A. Luo, S. M. Kozlov, F. Viñes, J. Limtrakul, D. G. Truhlar and F. Illas, Bulk Properties of Transition Metals: A Challenge for the Design of Universal Density Functionals, *J. Chem. Theory Comput.* 2014, **10** 3832–3839.