University of Nottingham
UK | CHINA | MALAYSIA

Material Discovery and Modelling for Solid-State Hydrogen Storage and Fuel Cell Applications

Thesis submitted to the University of Nottingham for the degree of **Doctor of Philosophy, 30th September 2022.**

James K. M. Wakerley

14324172

Supervised by

Dr Sanliang Ling
Prof Gavin Walker
Prof David Grant
Dr Ming Li

Abstract

This thesis covers an attempt to construct a supervised machine learning model, for use in prediction of formation enthalpy values for novel metal hydride compositions. Further work, making use of static density functional theory calculations as well as *ab initio* and machine learning force field molecular dynamics simulations, to model oxygen transport in a La-Mg co-doped barium titanate system is also reported.

Utilising open-source, readily available repositories of previously calculated results, two gradient boosting regression models are developed; separately trained to qualitatively predict formation enthalpy data for metal hydrides, and for intermetallic alloys. Once developed, such predictions are compared to enthalpy values, calculated from first principles, for held-out samples from the original database, and known experimental values for select materials. A process is outlined for generating new ternary hydride compositions, previously unseen to the model, from which a select sample of promising predictions are subjected to crystal structure prediction processes. By introducing structural information, first principles calculations are used to determine formation enthalpies for comparison to predictions.

Intentionally trained using descriptors derived solely from chemical composition, without any dependence on crystal structure, the resultant

model ultimately struggles to generalise prediction of formation enthalpies across the diverse geometry space of hydride materials. The decades-long quest for reliable crystal structure prediction simply from chemical composition proves to be a challenge for effective model validation by calculation, given the range of hydride classes.

Oxygen transport through the prototypical perovskite system of barium titanate is studied to investigate the methodology of characterising oxide ion diffusion through the bulk of such a material. Inspired by unpublished experimental results, this system is then co-doped with lanthanum and magnesium, thus introducing titanium and oxygen vacancies, allowing for investigation of oxygen diffusion by means of dynamic simulation methods. This is performed using the relatively new method of on-the-fly machine learning force field molecular dynamics, an approach to the modelling of dynamical systems which, in theory, drastically reduces the time and computational cost of traditional methods based solely on *ab initio* molecular dynamics.

Approximations of low-energy transition paths for oxygen movement in the local vicinity of point defects suggest energetically favourable diffusion pathways introduced by lanthanum doping. Molecular dynamics simulation methods are used to construct oxygen self-diffusion trajectories, from which diffusion mechanics can be determined. These results suggest higher rates of diffusion events in magnesium-doped barium titanate than when co-doped with magnesium and lanthanum. Additionally, mobility in the co-doped system is shown to be influenced by geometry of lanthanum dopants.

Acknowledgements

I am beyond proud to have produced this work, and have enjoyed my time spent at the university whilst working towards this goal. Thank you to all that have been a part of this journey and have helped me along the way.

First and foremost, I would like to sincerely thank my supervisors for sharing their expertise and wisdom throughout the course of this project. Sanliang - thank you for introducing me to your world of computational chemistry; for showing me the ropes and steering me back in the right direction when needed. Gavin and Ming - thank you for your experimental insights and for tolerating the notion of calculations under ideal, yet unrealistic, conditions. David - thank you for being the voice of reason and for going above and beyond to help me out when times got tough.

Thank you to my fellow CDT colleagues; Dr. Marcus Adams, and (soon to officially be Dr.s) Jack Hart and James Felton, without whom the office would have been unbearably quiet - and potentially even productive. Additionally, thanks to the SusHy contingent that I have gotten to know in recent years, albeit some more virtually than others.

A huge thanks to all of the folks over in CompChem - my home away from home - who took me in as one of their own when I was new to the university. In particular, thank you to Josh Baptiste and Ellen Guest for introducing me to the gang, and also to Abi Miller, Ben Speake and Steve Skowron for the countless hours spent putting the world to rights at the JA.

Finally, I am truly grateful to all of my loved ones who have been with me though the good times and the bad. Thank you to Jess for being there throughout - for listening to my woes, and diligently nodding along as I explain my work for the umpteenth time - I couldn't have done this without your support. Thanks to my family - Mom, Dad and Elodie - for believing in me and for always being at the other end of the phone when needed.

This project is funded by the EPSRC Centre for Doctoral Training in Fuel Cells and their Fuels - Clean Power for the 21st Century, EPSRC Grant No. EP/L015749/1. I would also like to acknowledge the use of the Sulis supercomputer through the HPC Midlands+ Consortium and the ARCHER2 supercomputer through membership of the U.K.'s HPC Materials Chemistry Consortium, which are funded by EPSRC Grant Nos. EP/T022108/1 and EP/F067496/1, respectively.

Contents

Chapter 1

Introduction

1.1 Global climate concerns

With an increasing global impetus to reduce, and ultimately eliminate, carbon emissions in a technically challenging time scale, the need to shift the energy industry away from a dependence on fossil fuels is of great importance. From an engineering perspective, development of alternative energy technologies is the crux of this issue and, as such, there is interest in a solution, or combination of methods, that would be able to satisfy the needs of both developed and developing countries [1], and be practical enough to be adopted and implemented for a range of uses.

1.2 Potential of hydrogen as a fuel

It is proposed that a contributing factor to this necessary paradigm shift could be the increased utilisation hydrogen in many aspects of life. The most abundant element in the universe, only trace amounts of the gas are

found in the air on Earth but it is plentiful in the planets organic life and far-reaching oceans. Possessing one of the highest specific energy density values per mass compared to other fuels, the potential is clear to see [2]. However, to its detriment, low pressure in ambient conditions results in a low energy density per volume. Due to this, efforts are being made to develop storage solutions that would offer an energy-dense hydrogen store, whilst also taking into consideration the weight of such a system [3].

An approach for integrating hydrogen fuel into the current energy mix would be the introduction of fuel cells into pre-existing appliances as a clean and renewable energy carrier [4]. Fuel cells are electrochemical devices which generate direct current electricity by reacting hydrogen with oxygen, forming water as a sole by-product. These differ from batteries in that they are not a closed system, and require a continuous supply of reactants in order to maintain performance. Therefore, suitable storage and supply mechanisms are crucial aspects which can impact the performance of the system as a whole.

As with any technology, there is an ever-present objective to further improve performance and efficiency, but in such a developing market there is also the need to facilitate the uptake by means such as improving practicality and ease of conversion. A key point of interest with regards to storage is for mobile applications, such as light-duty fuel cell vehicles. To incentivise commercialisation, factors which include performance and driving experience are sought to be comparable to road vehicles that are currently available on the market or, with time, to even surpass them. Additionally, aspects such as range between refuelling stops, the ease of such a refuelling process, and maintenance and running costs, require consideration to encourage adoption. These matters are intrinsic to the architecture of the system and can be viewed as a material

challenge of components throughout a system's design. Nonetheless, the opportunity to mitigate the emissions from transport, which accounts for approximately 20% of global CO_2 emissions, exists as strong motivation for such development [5].

1.3 Technical challenges of transition

Technical performance targets, as set out by a partnership of the United States Department of Energy (DOE) and the United States Council for Automotive Research (USCAR), along with energy and utility companies and organisations, provide industry-level system objectives for the near-future and an ultimate commercially viable target [6]. As of 21-Jan-2022, these are:

Storage Parameter	**Units**	**2020**	**2025**	**Ultimate**
System Gravimetric Capacity				
Usable specific-energy from H_2 (net useful energy/max system energy)	kWh/kg (kg H_2/kg system)	1.5 (0.045)	1.8 (0.055)	2.2 (0.065)
System Volumetric Capacity				
Usable energy density from H_2 (net useful energy/max system volume)	kWh/L (kg H_2/L system)	1.0 (0.030)	1.3 (0.040)	1.7 (0.050)
Storage System Cost				
Storage system cost	\$/kWh net (\$/kg H_2)	10 (333)	9 (300)	8 (266)

Table 1.1: Light-duty fuel cell vehicle system storage targets, per the US Office of Energy Efficiency & Renewable Energy.

As can be seen, there is an emphasis on gravimetric and volumetric capacity as well as acknowledgement that the overall cost of the system is a key limiting factor as to adoption rates. It should be noted that these

targets are given for light-duty fuel cell vehicles such as a private passenger vehicle. With respect to other applications, different targets may apply on a situational basis.

Industrial vehicles, for example forklift trucks and excavators, make use of counterweights to balance heavy loads. There is the potential to utilise the weight of the fuel cell and hydrogen store for this purpose [7]. Elsewhere, efforts are being made to decarbonise the global freight network, with cargo ships, heavy goods vehicles, and trains all shown to be convertible to hydrogen power - again, being applications with a reduced dependence on system weight [8]. Similarly for stationary stores, the volumetric capacity and cost take priority.

Considering this, the multivariable optimisation of both gravimetric and volumetric capacity, along with system cost and performance, can be seen to have a range of solutions dependent on use case and tolerance, allowing for investigation of a broad selection of component materials. This thesis will focus on the storage materials gravimetric and volumetric capacities, as opposed to system capacity, which may include the weight and volume of components such as the physical tank and cooling apparatus, amongst others.

1.4 On-board hydrogen storage systems

The phase diagram of hydrogen as a function of temperature and pressure is presented in Figure 1.1. It shows three lines which represent condensation, freezing and sublimation, where hydrogen transitions between its solid, liquid and gas phases. Also specified are the triple point, boiling point and critical point. The first of these, the triple point, represents the required

temperature and pressure in which hydrogen can exist in all three phases simultaneously (13.8 K, 7.2 kPa). The boiling point represents the normal boiling point (NBP); the temperature at which the substance boils at atmospheric pressure, which is 20.3 K for hydrogen. The critical point, denotes the conditions for the coexistence of the liquid and vapour phases (33.145 K, 1.3 MPa) [9].

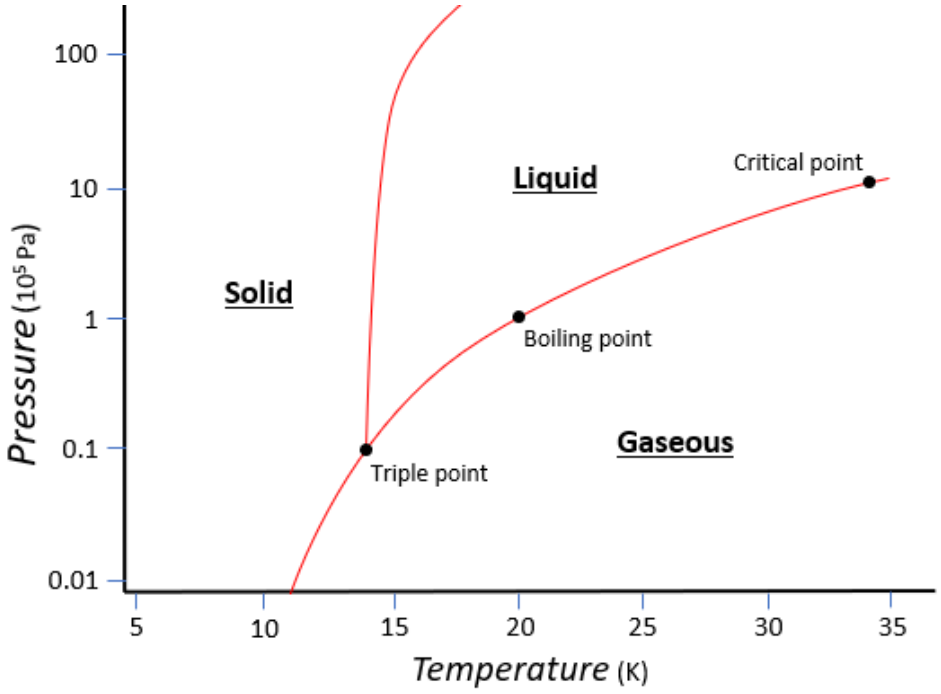


Figure 1.1: Sketch of the phase diagram for hydrogen as a function of temperature and pressure.

1.4.1 Physical-based methods

Physical storage methods refer to the notion that hydrogen can be contained by a material such that there are no strong chemical bonds between the hydrogen and the host compound. The most prevalent technologies of such class are compressed gaseous hydrogen (CGH2), and liquid hydrogen (LH2). Both of these methods tend to take the form of a physical tank system simply filled with the corresponding fuel.

The low energy density of hydrogen gas at ambient conditions lends itself to being stored as a compressed gas, in order to allow for a sufficient capacity energy store within a practical footprint. It is necessary for a light-duty vehicle to carry 5-6 kg of hydrogen to give a range of approximately 500 km. In order to maintain the available cabin space that might be expected from such a vehicle, it is required to store this gas in the region of 350-700 bar. The upper limits of viability for compression are dictated by the flattening of the mass energy density at these high pressures [10]. The inherent risks and energy costs of compressing and storing high pressure gas requires stringent safety solutions and development of lightweight storage tanks.

Liquid hydrogen utilises the vast increase in mass density when compared to the gaseous state. A comparable energy density of CGH2 systems at 700 bar can be achieved at $-253°C$ and 1 bar. Obvious practicability concerns revolve around the initial liquefaction and ability to maintain the store at these low temperatures. Hydrogen liquefaction is a very energy intensive process with 30% of the stored chemical energy consumed in doing so [11]. Heat transfer from external sources can lead to evaporation at which point any boil-off gas must be vented so as to maintain a certain temperature, resulting in a non-trivial loss of fuel unless these very low temperatures can be maintained. Ultimately, with the additional requirement of an effective cryogenic tank, the total cost of LH2 systems are at least comparable to that of CGH2 systems.

1.4.2 Materials-based methods

Alternatively, storage solutions that utilise interatomic interactions are often referred to as materials based methods. Be it by adsorption in porous or otherwise large surface area systems, making use of liquid

organic hydrogen carriers, or by creating chemical, interstitial, and complex hydrides; utilising chemical systems to store hydrogen has the potential for much higher volumetric densities than physical-based methods [12]. Whilst operating conditions for CGH2 and LH2 are largely reliant on the properties of hydrogen itself in the corresponding physical state, conditions for these chemical systems can vary significantly as a function of the chemical species involved, with the energy and/or temperature required for sorption and desorption reliant on the strength of the hydrogen interactions with the rest of the material.

1.5 Metal hydrides as a storage solution

In this work, the focus will be on metal hydride materials, which are usually interstitial or complex hydride species. These take the form of a solid-state fuel with higher hydrogen volumetric density than physical-based stores whilst operating at safer, more accessible temperatures and pressures. Usually consisting of a metallic crystalline host structure, hydrogen anions bond to these less electronegative elements to form a stable hydride species.

The transition from a metallic to hydride form is characterised by three stages of reaction. The first (α) entails hydrogen in a solid solution with the metal, and the final (β) is the complete hydride phase. The intermediary state ($\alpha + \beta$) is the phase of ongoing reaction, where both such states co-exist and transition from one to the other. On a pressure-composition (PCI) isotherm plot (see sketch in Figure 1.2), this transformation is represented by a flattening of the curve to a so-called plateau pressure. The system can exist exclusively in α or β phase for very low or high hydrogen to metal ratio (H/M), respectively. Another caveat is that with an increase

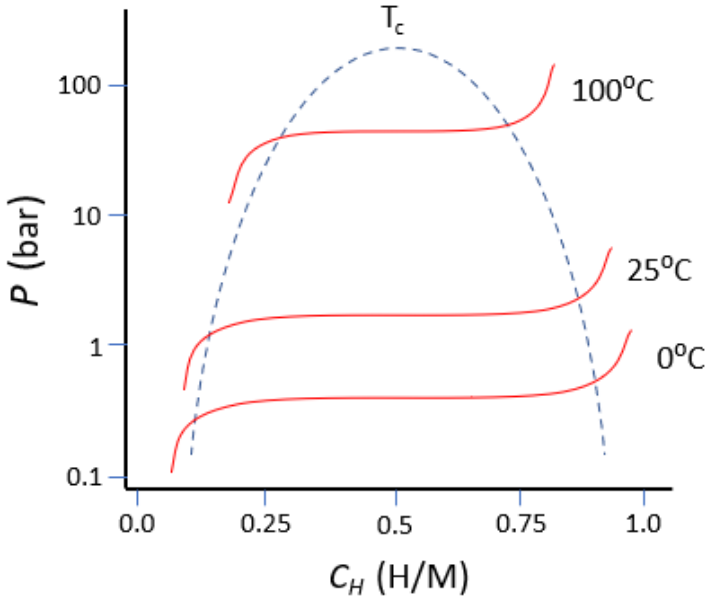


Figure 1.2: Sketch of a PCI plot for a range of temperatures, with relative hydrogen capacity against pressure for several temperatures.

in temperature the plateau pressure rises, narrowing the region of co-existence. It is ultimately eliminated if this exceeds the critical temperature (T_c). The parameterisation of such plots are composition specific. If the hydrogenation and dehydrogenation processes, which are exothermic and endothermic respectively, can be consecutively performed with minimal hysteresis or degradation, then a given system may be deemed practically cyclic.

The change in Gibbs free energy (J mol^{-1}) of a system at a given temperature (K) relates to the change in standard enthalpy ΔH (J mol^{-1}) and the change in standard entropy ΔS (J K^{-1} mol^{-1}), defined as:

$$\Delta G = \Delta H - T\Delta S. \tag{1.1}$$

The strength of the bond between hydrogen and metal is related to the enthalpy term, and the entropy term accounts for the transition from molecular to bound hydrogen, along with contributions from the metal

atoms. For many metal hydrides, the value of ΔS is approximated to be the standard entropy value of hydrogen (S_{300K} = 130.77 J K^{-1} $mol_{H_2}^{-1}$), corresponding approximately to contributions from the loss of degrees of freedom when gas phase hydrogen is absorbed into the metallic crystal [13, 14].

Whilst metal hydrides have been of consideration for decades [15], the search for an 'optimal' hydride as a storage solution is by no means trivial. Dependent on the application, operating requirements and conditions are situational. Important parameters include extrinsic factors, for example the practicality of a system in a given scenario as a function of volumetric and gravitation densities. Additionally, factors intrinsic to material choice such as uptake/discharge mechanics and kinetics, and hysteresis must be considered [16].

Low temperature hydrides are suggested to be cyclable storage solutions that operate at a reasonably low pressure, so as to reduce reliance on hydrogen compressors, whilst at, or just above, ambient temperature [17]. This would be useful in reducing the amount of external energy required to be input, improving the total system efficiency. For light-duty vehicles, this might allow for more efficient recycling of otherwise waste heat energy from fuel cell operation so as to assist the hydrogen supply mechanism. Properties sought after for these materials include being lightweight, with high gravitational and volumetric densities, to have a suitable plateau pressure at a near-ambient temperature, and to have sufficiently stable hydrogenation and dehydrogenation products.

1.6 Fuel cell systems

Solid oxide fuel cell (SOFC) technologies offer a means of sustainable, environmentally friendly energy production. Through the utilisation of fuels such as hydrogen or synthesis gas, this technology allows for conversion of chemical energy into electrical energy by use of an electrochemical device, with negligible emissions. The structure consists of the nominative solid oxide ceramic electrolyte that separates an anode and a cathode, which are in a fuel rich and oxygen rich environment respectively, as depicted in Figure 1.3. Catalytic membranes, selectively permeable with respect to oxygen or hydrogen, are used to construct a membrane electrode assembly on either side of the electrolyte.

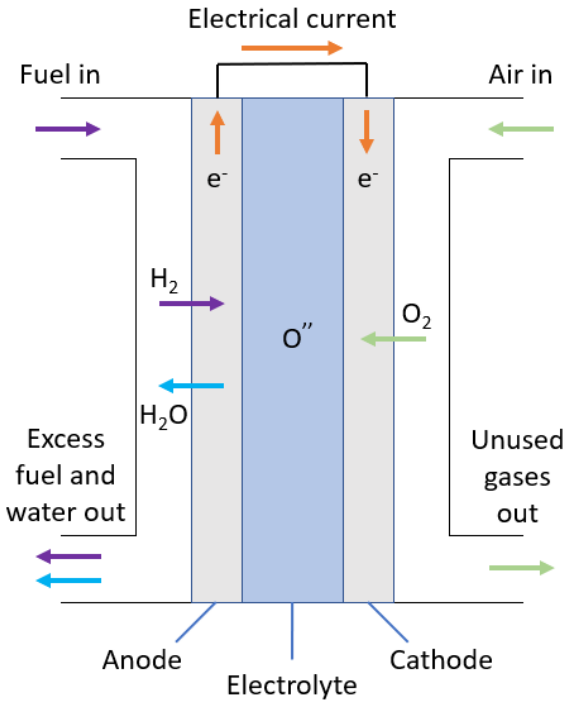


Figure 1.3: Schematic of a solid oxide fuel cell system.

Fuel is decomposed at the anode into protons and electrons, and an external circuit carries the free electrons to the cathode, where they are used for the reduction of oxygen. Oxygen ions move across the dense electrolyte,

combining with hydrogen ions at the anode side of the device to complete the circuit, producing water as a result [18].

1.6.1 Importance of oxygen transport for fuel cell systems

One of the rate-limiting factors for electrochemical activity in a SOFC system is the oxygen ionic transport mechanics of component materials. Electrolyte design requires a mechanically stable dense material with sufficient ionic conductivity to maintain good performance of the cell, whilst minimising the electrical conductivity to prevent leakage current or short circuiting [19]. SOFCs typically operate at temperatures of between 600°C and 1000°C, at which temperatures oxygen transport kinetics allow for a reasonable output. At temperatures towards this lower limit, common electrolyte materials display ionic transport resistances that drastically impact overall performance of the device [20].

An element of the oxygen ionic transport of an electrolyte, oxygen mobility, is a function of the movement of oxygen ions through the bulk of the material: self-diffusion. The nature of crystalline solid state materials allow for diffusion processes to be characterised by atomic hopping throughout the lattice. Macroscopic diffusion can be viewed as the collective effect of many such displacements throughout the crystal structure and is a function of the physical quantities of these microscopic hops. These include hop distance, jump rates, as well as geometric and correlation factors [21]. Lattice point defects, including substitutions, interstitials, or vacancies, act to influence these factors, and can combine to facilitate complex diffusion mechanisms.

1.7 Material discovery

A material challenge exists to find solid-state hydrogen storage materials that satisfy high density targets whilst also being thermodynamically suited to operate at the low temperatures congruent with on-board applications. Similarly, performance of fuel cell electrochemical devices is also subject to development of materials for components.

A poorly conceived notion of theoretical material science is that it is antagonistic to experimental work - vying for the same spot at the top of the hierarchy of modern science. This could not be more wrong. The reality is that simulations and calculations have become ingrained in the scientific process and a largely theoretical approach to a problem works in complement to experiment.

Experimental testing across a test sample space can be an arduous task, expensive both in terms of man-hours and the cost of resources required to effectively explore a large range of different chemical compositions due to the inherent trial-and-error process required. Upon consideration of more complicated chemical forms, we observe a combinatorial explosion of the chemical domain when accounting for all combinations of a sample space, e.g. chemical systems of the form A $\rightarrow$ A-B $\rightarrow$ A-B-C, for any elements A, B, C. The challenge in assessing the full domain of possibilities risks missing best-in-class performance, and incomplete knowledge of material behaviour [22]. Experimental exploration of these materials may also present practical issues. These processes involve many wet experiments, producing a large amount of harmful chemicals and waste, which would obviously be amplified by a huge testing operation.

Computational approaches allow for a mathematical analysis of large

combinatorial spaces, accelerating the screening of a complex chemical domain in a representative world. Whilst both methods have their own advantages and pitfalls, they are mutually complementary [23]. Accurate mathematical representations of the real world may highlight correlations and interactions, which might be excluded from current theory, facilitating further experimentation, whereas experiments have more potential to focus on new phenomena [24]. Given this, material discovery as a whole can be accelerated by supplementation of computational modelling with experimental work. Beyond considerations regarding in-lab safety and environmentally unfriendly by-products, efficient screening of potential combinations and further analyses at more accurate levels of theory increases the throughput of candidate material nomination.

1.7.1 Evolution of discovery approach

Classically, academia has always put a great emphasis on the notion of confidentiality. Intellectual property and unpublished results, *in situ* equipment and infrastructure - access to such resources have historically been reserved for those connected to the appropriate powers through one means or another. Over time, the academic world has gained access to developments in IT infrastructure, providing more effective means to share knowledge and practices, and we have seen behaviours change toward the sharing of work and collaboration. Scientists have opened up to the concept of accessibility of research being an important part of progress, appreciating the mutual benefits of assisting each other without *quid pro quo*, to the extent that data management plans are now commonplace, if not mandatory, for most funding agencies.

A milestone in this paradigm shift has been the development of the internet,

the global adoption of which has allowed for easy sharing of massive amounts of data at an archival level. Advancements in computing power has also spurred the generation of more data, bolstering the capabilities of computational science. This has facilitated evermore complex calculations at an expanding rate in line with developments in computing resources, whilst simultaneously seeing improvments in ease of use and access.

1.7.2 'Big data' approach

From both experimental and theoretical work, this wealth of information must be properly archived and available for access in a quick and efficient manner. This has led to the development of materials databases; collated banks of calculated properties accessible via the internet. These repositories have helped to revolutionise the exchange of information on a worldwide scale, providing this data in a quickly accessible and easy-to-use manner.

These databases may allow for more focused computational analysis of scientific trends, some of which may take a complex mathematical form, thus proving difficult for interpretation by human intelligence. In turn, statistical methods can be used to create predictive models to extend such patterns to new materials of interest. This open access approach to data collection also allows for more corroborative measures, assisting efforts in data curation and quality assurance. By combining data collection, AI-assisted modelling methods and computational simulations for material prototyping, testing and validation, development of an informatic inspired workflow for material science could lead to a much higher throughput of analysis than ever before [25].

1.8 Aims and objectives

1.8.1 Aims

This thesis includes work on materials that relate to two distinct areas of hydrogen research. The first of these being construction of a machine learning model to qualitatively predict the formation enthalpy of novel ternary hydride compositions, greatly narrowing down the compositional space to be considered, allowing for more focused further analyses with the scope of assessing feasibility as solid-state hydrogen storage systems. This involves construction of multiple models, to predict formation enthalpy of metal hydrides, and metal alloys. In addition to this, perovskite structures are modelled so as to study oxygen mobility through the crystal and consequential phenomena, and how this process could be used to analyse materials for fuel cell systems. Whilst not itself a candidate electrolyte material, a La-Mg doped barium titanate was used to qualitatively present the methodology for simulating such dynamic behaviour, where the chemistry of non-stoichiometric materials can be adjusted to either increase or suppress oxygen diffusion.

1.8.2 Objectives

1. Develop regression machine learning models to predict enthalpy of formation for metal hydrides, and metal alloys (Chapter 3)

2. Validate the hydride model's performance against known ground truth data, sourced from theory and experiment (Chapter 3-4)

3. Use these models to predict potential storage materials, and further analyse (Chapter 4)

4. Simulate oxygen mobility in the prototypical perovskite structure of barium titanate, before expanding the investigation to a system formed by the co-doping of barium titanate with lanthanum and magnesium (Chapter 5)

Chapter 2

Theory

2.1 Statistical learning

The theory of statistical learning concerns prediction and pattern recognition in order to deduce mathematical reasoning for data distribution or for data mapping. This is most commonly categorised as being either supervised or unsupervised learning. The former works by analysing input and output data and determining a mapping function for the process at hand in order to predict results when confronted with new inputs. Unsupervised learning, however, is a means of hypothesising patterns and correlations when only presented with input values, from which data structure is devised without pre-existing results for comparison.

The approach to the material discovery task presented in this work is to use a supervised machine learning method, fitted to data for materials whose ground truth values have been calculated through the means of density functional theory. Producing such a model should aid in identifying new materials to analyse to a more accurate theoretical degree such that they

might later be studied in realistic experimental conditions.

A key statistical tool to this task is regression analysis, a means of mathematically investigating the relationships between variables. By applying a set of statistical processes, one can ascertain the relative causal effect of several independent variables ('descriptors', 'predictors' or 'features') upon a given dependent variable ('criterion variable' or 'response'), the target value.

These techniques have seen use primarily for predictive means by comparing data points with known ground truth data. An important aspect of regression analysis, as alluded to earlier, is the ability to assess the effect that variation in the value of a model's descriptors has upon the criterion variable. This can help provide insight into the relationship between the two classes of data and can be investigated further by certain methods to facilitate estimation of relative dependencies.

In general, a regression model relates the criterion variable y to a function of the independent variables X and unknown parameters β (which may be vector or scalar values depending on the model used) through an equation of the form

$$H(\mathbf{X}) = f(\mathbf{X}, \boldsymbol{\beta}). \tag{2.1}$$

These parameters help to define the form of the function acting as our hypothesis. If the form of the function f is not known then one assumes an easily adjustable form, such as a linear combination i.e.

$$H(\mathbf{X}) = \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n \tag{2.2}$$

where $\mathbf{X} = (x_1, x_2, ...x_n)$ for any n descriptors.

The training process is intrinsically a mathematical optimisation problem.

In order to accomplish this, data must be represented numerically and suitable algorithms must be defined. In most supervised regression scenarios, a loss function is used to define a difference in prediction and known results, whilst an iterative process works to minimise such loss. For the sake of efficiency and to hone the ability to generalise, these optimisation algorithms are parameterised so as to scale the learning process. Called hyperparameters, these are often algorithm-specific and can be used to balance the speed and quality of model construction.

Different machine learning algorithms may have differing predictive performance on various datasets, even when considering consistent descriptors. As such, it is often required to test and benchmark a range of algorithms for a given dataset so as to identify the best-performing case. For a chosen algorithm, optimal performance is dependent on an optimised parameterisation of the fitting function. This is done by fine tuning hyperparameters.

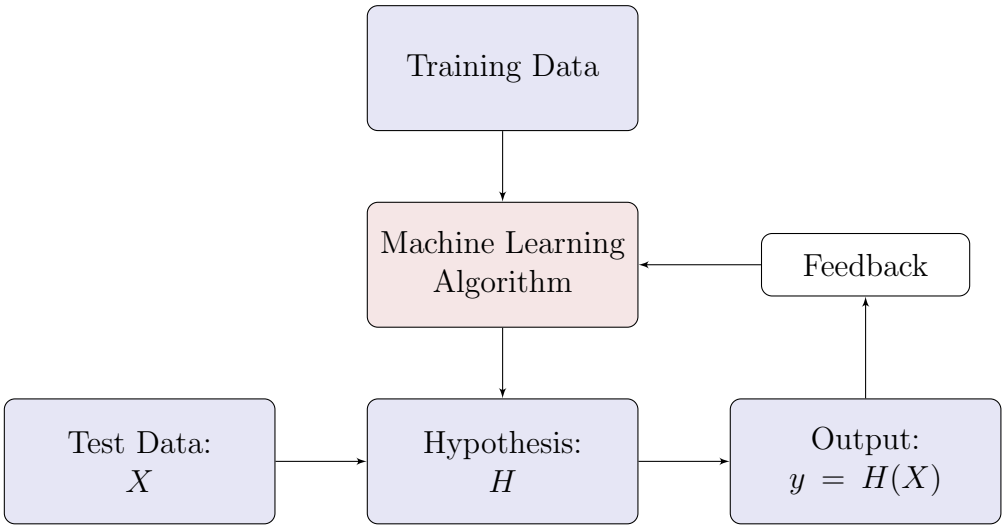


Figure 2.1: Sketch of the machine learning training process.

A generalised sketch of the construction of a supervised machine learning process is shown in Figure 2.1. The central column of the diagram depicts

the construction of a learner, be it a single learner, or a learner built from several which are combined together through a statistical process. This is trained on known data in order to formulate a mapping from independent to dependent variable, a function known as either a classifier or a regressor depending on its purpose. These two functions take an input and predict an output based on the determined hypothesis. A key aspect of the training process is the feedback, or validation, that can be ascertained by comparison of the known ground truth value for the test data inputs to the predicted value. This provides information as to a model's predictive performance on pseudo-'unseen data', allowing for further tuning of the model building process to more accuratly reflect the relationship in the data being investigated.

2.1.1 Regression methods

2.1.2 Decision trees

Decision trees, illustrated in Figure 2.2, are a very lightweight and intuitive learner. Starting with a source set of data, recursive partitioning occurs along the length of the tree, splitting data along branches at each node according to some known greedy (locally optimal) splitting rules. When used for regression, the hypothesis H, which estimates the mapping of descriptor data to the target function, is too large to feasibly search for exhaustively. As such, a heuristic search algorithm is used to find a hypothesis that minimises training error through the minimisation of a loss function.

At each node, the presented data is split into two branches according to some threshold, and an impurity value is calculated from a given function

for each possible split. A greedy choice of threshold and split point is selected so as to minimise the impurity before the process is repeated until a stopping rule is reached, such as a maximum tree depth or minimum number of instances at a node. The value of the final estimator is given as the mean value across all terminal nodes.

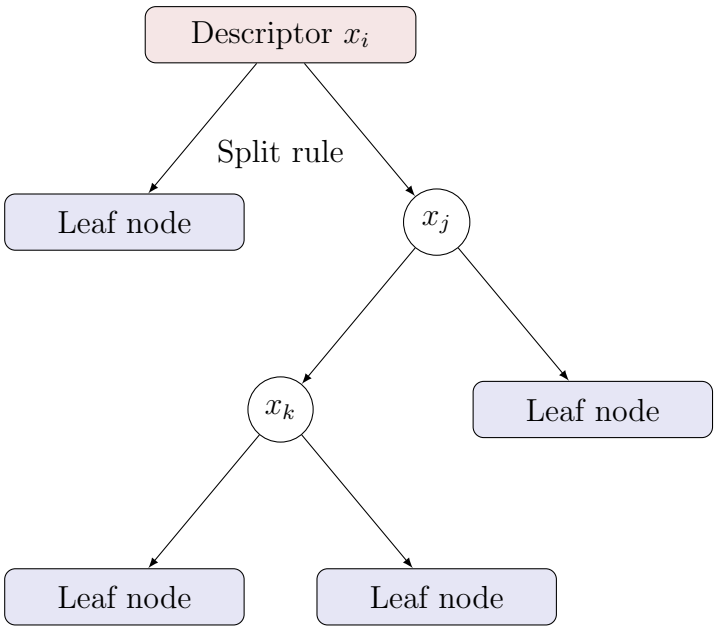


Figure 2.2: Sketch of a decision tree, denoting characteristic features.

2.2 Ensemble learning methods

One may also run several iterations of a base learner, either in series or in parallel, to construct what is called an ensemble. Whilst a single learner searches hypothesis space for a good prediction hypothesis, an ensemble combines multiple hypotheses in an attempt to reach a better result still. This technique can be used to bolster weaker learning algorithms or can be utilised by relatively fast algorithms, such as decision trees, to reinforce their predictions.

2.2.1 Bagging

The concept of bootstrap aggregating (referred to as 'bagging') was proposed as a way to improve the accuracy of both regression and classification machine learning algorithms by combining results from randomly sampled training sets taken from the input dataset [26]. Generating a number of training sets sampled with replacement from the original, known as a bootstrap sample, the algorithm is run on each of these sets before outputting the mean prediction. This method allows for the analysis of measures of accuracy to sample estimates, many of which are of interest including prediction error, bias and variance, and facilitates with the derivation of standard errors.

One application of this technique is that of tree bagging. Random bagged samples of the training set are taken and used to fit decision trees, following which predictions can be made at a specified point by averaging values given by the multitude of constructed trees. This helps to reduce the impact of correlated trees by making use of a wide range of training sets. Given this decrease in correlation, it can be seen that a tree bagging method reduces the models variance whilst not having a negative effect on the bias.

2.2.2 Random forests

Random forest regression (RFR) is an ensemble technique created using the random subspace method which is a feature bagging mechanic, training each learner on a randomly bootstrapped subset of the total descriptor space, and calculating the optimal branch (with respect to impurity) to take at each node from a random selection of the feature subspace [27]. This helps to decorrelate the trees further than simply tree bagging by allowing

more diversity in the descriptors being considered across different trees. As well as this, strong predictors that would otherwise be selected quite frequently while analysing a large number of trees have such dominance lessened by the feature bagging process, allowing a more rounded analysis of all descriptors.

2.2.3 Extremely randomised trees

The ExtraTrees regression (ETR) ensemble method is trained in a similar way to random forests but differs in that each tree is trained with the full training sample, and it introduces a randomisation of the top-down splitting process. Instead of taking the optimal splitting rule for each feature (based on aforementioned impurity calculations), this algorithm generates thresholds at random for each feature and the highest impurity value of these is selected as the splitting rule. This explicit randomisation of predictor and value at a cut-point, combined with averaging of the output over the ensemble, provides a greater reduction in variance as opposed to regimes with less randomisation, such as random forests. Additionally the random selection process as opposed to optimal split choice at each node is computationally preferable, something that may prove favourable upon dealing with very large systems.

2.2.4 Gradient boosting

A different approach to an ensemble system, gradient boosting regression (GBR) is a mechanism that sees the combination of weak learners to form a strong learner, with the aim of optimising an arbitrary loss function $L\left(y, F\left(x\right)\right)$ to approximate the criterion variable. By defining an upper

bound of iterations (N), each step starts with an imperfect model $\hat{F}_{n-1}(x)$, where $1 \leq n \leq N$, and at each stage a learner h is found such that it combines with $\hat{F}_{n-1}(x)$ to improve the overall estimator with respect to the loss function. An approximation of the final model is taken to be a weighted sum of these weak learners and a specified learning rate, a fixed step length which dictates the rate of change per iteration.

$$\hat{F}(x) = \sum_{N} \gamma_m h_m(x) + const. \tag{2.3}$$

Here it can be seen that the approximation iterates in a greedy manner so as to minimise the empirical risk relative to the immediately preceding step:

$$F_n(x) = F_{n-1}(x) + \arg\min_{h_n} \left[\sum_{i=1}^{k} L\left(y_i, F_{n-1}(x_i) + h_m(x_i)\right) \right] \tag{2.4}$$

From here, a steepest descent algorithm is used to numerically approximate the minimisation operator, which is given as the negative gradient local to $F_{n-1}(x)$.

Maintaining the approximation of linearity by use of a small step length γ, Equation 2.4 becomes:

$$F_n(x) = F_{n-1}(x) - \gamma \sum_{i=1}^{k} \nabla_{F_{n-1}} L\left(y_i, F_{n-1}(x_i)\right), \qquad \gamma > 0. \tag{2.5}$$

This methodology can be applied to decision trees in what is referred to as gradient tree boosting. Given a set of base learners of consistent size, a similar process takes place where one can fit a tree h_m to pseudo-residuals (gradients of the loss function). The depth of the trees can be adjusted to account for interactions between descriptors, with the price being the operation speed of the algorithm. It is suggested that a depth

ranging anywhere from 4 to 8 is sufficient to incorporate interdependencies of variables, with higher values providing negligible additional contribution [28].

2.2.5 Out-of-bag error and feature importance

Bagging data provides access to additional useful tools for assessing prediction error in resultant models. Out-of-bag (OOB) error is determined by calculating the error on a given data point $d_i = (x_i, y_i)$ using predictive learners trained without d_i in their bagged sample. This can be assessed for all i in the initial dataset. It is possible to utilise the out-of-bag error in order to rank the importance of the variables in the dataset [29]. Selecting OOB data corresponding to each constructed tree in the ensemble, these OOB data points are computed along each tree and an error value determined. By randomly permuting the features in this OOB set and repeating the process, another error value is obtained. Using the standard deviation of the differences as a normalisation factor, we can compare these scores, providing a new tool for interpretation of our data [30]. Descriptors with higher scores are those that have a relatively significant contribution to the model's predictive ability. Another method of evaluating feature importance is Gini Importance (GI), or Mean Decrease in Impurity (MDI). In this case, feature importance is given as the total decrease in node impurity, weighted by the proportion of samples reaching that node, averaged across all trees in the ensemble.

2.3 Caveats of statistical learning

Predictive ability of a machine learning model is developed through the process of quantitative learning of patterns from training sample data. The fidelity of a model is reliant on several factors. Those intrinsic to the machine learning algorithm include the selection of algorithm, selection of features, and parameterisation of the fitting process. For the training process, one must consider both the method used as well the training dataset used.

2.3.1 Input data sample size

Mathematical patterns, sought to be captured by a machine learning process, involves evaluation of the training data presented before the algorithm. Scarcity of data makes such a task more difficult, impacting the ability for a model to train, and ultimately affecting predictive performance for an unexplored test domain. Such model bias may be mitigated by means such as retraining with the inclusion of more data samples, consistent with any previously used data. In literature, some ML based investigations into thermodynamic properties of solids have reported predictive accuracy to scale monotonically with training dataset size, systematically reducing prediction error [31, 32].

2.3.2 Quality of input data

In order to ascertain correlations and relationships between the data, reliable input data is required. By the form of the hypothesis given in Equation 2.2, it can be seen to have an inherent risk of contamination

from noisy, incorrect or otherwise troublesome data. Issues may arise with example data that contradict other samples or require an increase in the algorithm complexity in order to accommodate. Such situations can drastically impact the model's ability to generalise, or may lead to overfitting.

In order to build a robust learner, training data should be curated to a certain extent, eliminating troublesome data points which may compromise understanding the fidelity of the mechanism under investigation. Data pruning methods include rudimentary outlier identification and elimination, as well as more involved iterative processes used to improve generalisation. However, pruning may lead to bias and so should be handled carefully.

2.3.3 Generalisation

A key concept in statistical learning is that of generalisation and how well a trained learner can perform when predicting results for unseen data. Ideally, a model would work well when presented with any new sample. However, there are reasons why this might not be the case.

Data diversity is important within the training sample. Deciphering the mathematical relations between variables presented during model construction may result that only new data samples similar to those included in the original dataset can be predicted sufficiently. The extent of this extrapolatory limitation is situational, but can be mitigated by ensuring the training sample contains enough consistent data across the sample space to be investigated.

2.3.4 Bias-variance tradeoff

A dilemma faced when using any supervised learning method is the notion of attempting to minimise mutually conflicting sources of model error: bias and variance.

Bias error is rooted in underfitting a model to training data. This causes an erroneous fitting to known data, incorrectly capturing relations between variables. High bias can often be detected by analysis of a suitable performance metric, identifying poor predictive ability on training data. Conversely, variance error occurs when a model is trained to be too sensitive to noise or random fluctuations in the sample data. By focusing upon these small changes within the data the model is said to be overfit, negatively impacting its ability to generalise which can lead to poor predictive performance on unseen data points. Cross-validation is useful for assessing overfitting by running multiple performance tests against random splits of the training sample.

This balance between high bias vs high variance, or overly simplified vs overly complex, is important in statistical learning and is susceptible to a variety of other factors, such as availability of training data, choice of descriptors, and reliability of source data.

2.4 Machine learning for material discovery

Machine learning is a useful tool for recognition of patterns and relationships in data. Such processes can be applied to material discovery to mathematically characterise underlying chemical mechanisms in an interpretable manner.

By systematically studying similar systems, a robust model can theoretically be developed to describe a class of materials through a computationally efficient process. This can then be further expanded upon using higher levels of theory to validate prediction. Having used ML on a broad selection of compositions, one can hone in on specific examples and more efficiently utilise expensive computational resources.

Further analyses include the use of density functional theory to directly model atomic structures. By introducing specific electronic and ionic considerations, structures can be relaxed into ground state configurations *via* geometry optimisation, from which material properties can be calculated.

2.5 Density functional theory

2.5.1 Approximation of the many-body Hamiltonian

In principle, a given system can be exactly described by its wavefunction. If modelled analytically, this would allow for exact understanding of how the system behaves. Alas, the Schrödinger equation for an N-body system cannot be solved with current understanding and as accurate an approximation as possible is required in order to depict an atomic scale environment.

By introducing the Born-Oppenheimer approximation, an assumption is made that the nuclear motion and the electronic motion in a given molecule can be separated. This allows the molecule to be described by the nuclear and electron positions [33]. An early attempt to simplify this complex system was by Hohenberg and Kohn [34]. The first Hohenberg-Kohn

theorem states that the external potential on a system is a unique functional of the electron density. The second theorem states that the functional that admits the ground state energy of the system gives the lowest energy *if and only if* the input density is the true ground state density. This spatially dependent electron density is a functional which is used to describe electronic behaviour in a many-body electronic system in only three spatial dimensions, rather than the 3N degrees of freedom given by an N-body electronic wavefunction.

The Hohenberg-Kohn theorems can thus be built upon to determine the ground state electron density. By starting with an explicit energy functional, $E[\rho(r)]$, and varying the spatially dependent electron density, $\rho(r)$, the energy can be minimised. This functional can be decomposed into a kinetic term, and terms for electron interaction with either nuclei or other electrons. The electron-electron interaction consists of Coulomb and exchange terms, with the former easily calculated as a system of repulsive terms. The nuclei-electron component can be found in a similar sense; attractive forces in this case. The remaining terms, kinetic energy and exchange, remain unknown and so must be modelled.

2.5.2 Kohn-Sham method

A formalism was established by Kohn and Sham to approximate the kinetic energy of a given system of N interacting electrons to the kinetic energy of a fictitious system of non-interacting electrons of the same spatial density [35]. Such a system admits a set of independent particle equations, molecular orbits (MOs), that collectively give the exact electron density.

$$\rho\left(r\right) = \sum_{i} \left|\phi_i\left(r\right)\right|^2 \tag{2.6}$$

Each electron equation has its own ground state energy functional as outlined by the Hohenberg-Kohn theorems and, as such, the system energy is a functional of these Kohn-Sham MOs. The total energy of the system is given by

$$E[\rho] = \int dr\nu\left(r\right)\rho\left(r\right) + \frac{1}{2}\iint drdr'\frac{\rho\left(r\right)\rho\left(r'\right)}{|r-r'|} + T\left[\rho\right] + E_{XC}\left[\rho\right] \quad (2.7)$$

$$= E_{ext}\left[\rho\right] + E_{coul}\left[\rho\right] + E_{kin}\left[\rho\right] + E_{XC}\left[\rho\right], \quad (2.8)$$

comprised of energy terms respectively corresponding to an external potential, classical Coulomb self-interaction of the electron density, kinetic energy of the particles, and many-body interactions between electrons collected into a term coined the exchange-correlation functional.

By use of the variational principle, the ground state energy can be found by minimising this energy functional with respect to the electron density. As the system has been shown to be composed of KS MOs, this means to minimise the energy with respect to such orbitals. Doing so admits the KS Hamiltonian, $\hat{h}_{KS}$, comprised of the kinetic and potential terms derived from Equation 2.8.

$$\hat{h}_{KS} = -\frac{1}{2}\nabla^2 + \nu_{eff}\left(r\right) \quad (2.9)$$

$$\nu_{eff}(r) = \nu_{ext}\left(r\right) + \nu_{coul}\left(r\right) + \nu_{XC}\left(r\right) \quad (2.10)$$

$$\hat{h}_{KS}\psi_i(r) = \varepsilon_i\psi_i\left(r\right), \quad (2.11)$$

where the effective potential, ν_{eff}, consists of external, Coulombic and exchange-correlation components.

Upon consideration of the effective potential given by Equation 2.10, it can be seen that the Coulomb and XC interaction terms are dependent on the orbitals sought from the calculation of the Kohn-Sham Schrödinger

equations. The common approach to tackle this quandary is an iterative method known as the Self-Consistent Field (SCF) procedure, as outlined in Figure 2.3. Following an initial supposition of an orbital set the electron density can be calculated and, from this, the Hamiltonian generated. This can then be used to calculate a new generation of orbitals in order to repeat the cycle. At each pass, the new energies calculated are compared to those of the previous generation, and the process is said to have converged should these values differ by less than some predefined threshold.

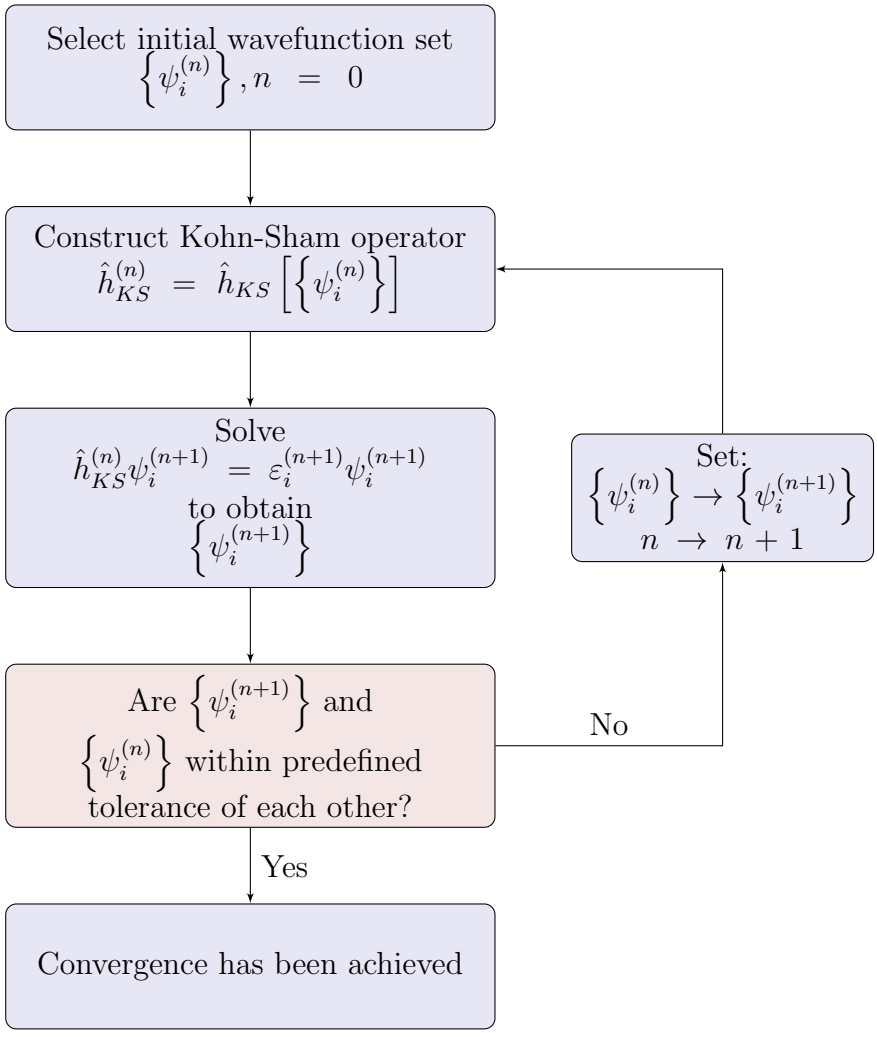


Figure 2.3: Workflow diagram of the Self-Consistent Field (SCF) cycle.

2.5.3 The exchange-correlation term

The exchange energy comes from the fermionic nature of electrons and the Pauli Exclusion Principle. This necessitates that electron wavefunctions be antisymmetric in order to accommodate the interchange of any two electrons in space. The correlation energy is less well defined, accounting for the dynamics of many electronic phenomena, including Fermi correlation and Coulomb correlation. The exact form of the E_{XC} term is not known but, by definition, is given to be the difference between the exact total energy and the other known quantities. As such, it is the only term in the Kohn-Sham energy functional that cannot be solved exactly and must be modelled. Whilst still a non-trivial feat, this is the crux of the Kohn-Sham approach to DFT.

The biblical notion of Jacob's Ladder is often invoked and used as analogy to the pursuit for perfect chemical accuracy. Each rung introduces new complexity by considering more exact terms in a trade off for higher computational cost. Taking approximations of $E_{XC}[\rho]$, of varying order of ρ and complexity, allows one to fine tune a calculation in order to optimise accuracy and efficiency.

Local-density approximations (LDA) of E_{XC} are functionals solely dependent on the electron density at a point in space, $\rho(r)$, approximated by the XC energy of electrons in a uniform electron gas (UEG) of the same density.

$$E_{XC}^{LDA}[\rho(r)] = \int dr \rho(r) \varepsilon_{XC}^{UEG}(\rho(r)) \tag{2.12}$$

Despite the apparent simplicity of this approach, LDA is an effective method when investigating systems with a slowly varying charge density such as bulk metals. However, this approach proves less useful in more

inhomogeneous systems, such as where ionic and covalent bonding is more prevalent. LDAs usually overbind molecules, underestimating bond lengths and the cell volume, resulting in unacceptable errors in geometry for more generalised applications.

Generalised gradient approximation (GGA) is an improvement on LDA, accounting for both the electron density at a point in space, $\rho(r)$, as well as the gradient, $\nabla\rho(r)$. This can now encapsulate information regarding the non-uniformity of the electron density at a given point by considering how charge density changes through space.

$$E_{XC}^{GGA}[\rho(r)] = \int dr\rho(r)\,\varepsilon_{XC}^{GGA}(\rho(r), |\nabla\rho(r)|) \tag{2.13}$$

Typically GGAs are more accurate than LDAs as to transition-state barriers and bond dissociation energy, but come with an appropriately increased computational cost. A wide range of such functionals exist, of varying construction. Non-empirical GGAs are the most widely applicable, built around the general rules of quantum mechanics, so as to satisfy as many exact conditions as possible without being fit to specific molecular properties. There also exists a scale of functionals which incorporate a range of empirically fitted parameters characteristic to certain chemical forms. These can incorporate more accurate depictions of certain dynamics, and can often do so more quickly than a non-empirical method, though at the cost of generalisability. The Perdew-Burke-Ernzerhof (PBE) functional is a popular choice due its time-tested applicability to a wide range of systems, yielding a reasonable accuracy in most practical cases [36].

The ladder extends further; with meta-GGA additionally incorporating $\nabla^2\rho$, hybrid DFT complementing previous tier approximations with a portion of exact exchange from Hartree-Fock theory, and random

phase approximation (RPA) involving exact exchange and partial exact correlation.

2.6 Plane-wave pseudopotential method

2.6.1 Plane waves basis sets and reciprocal space

A mathematical representation of the Kohn-Sham orbitals is required to convert the information encoded in the electronic wavefunction into algebraic equations for computational use. A choice of basis set should be able to encapsulate most of the dynamics of the wavefunction as a finite length linear combination of basis functions. A basis set can be localised which is to say that the functions are fitted to each atom or, especially of use in periodic systems, a set of plane waves which can span the system with the same periodicity. The choice of basis set impacts all further algorithms, and the complexity of the linear combination of functions is adjustable to tune for efficiency or accuracy in the final calculation.

Bloch's theorem states that solutions to the Schrödinger equation within a perfectly periodic potential can be given by a plane wave and a function that exhibits the periodicity of the potential,

$$\psi\left(r\right) = e^{ik\cdot r} u\left(r\right), \tag{2.14}$$

where $u\left(r\right)$ is a modulatory periodic function of the same period as the lattice, and $e^{ik\cdot r}$ is a plane wave characterised by the crystal momentum of the system. This theorem is true for any propagating particle in the lattice, independent of atomic positions, and has no dependence on the strength of the potential.

Such a wavefunction can be expanded to take the form of a three-dimensional Fourier series obeying the Born-von Karman periodic boundary conditions, ensuring the whole function reflects the translational symmetry of the lattice,

$$\psi(r) = \sum_{G} c_{Gk} e^{iG \cdot r}, \tag{2.15}$$

where c_{Gk} are complex Fourier coefficients, summed over all reciprocal lattice vectors. The sum is infinite in theory, however the coefficients c_{Gk} are inversely proportional to the squared norm of the vector G. This allows one to define an upper energy limit for the plane-waves to be used, represented by a maximum radius in reciprocal space, expressed as a cut-off energy,

$$E_{cut} = \frac{\hbar}{2m} |G|^2 . \tag{2.16}$$

Considering this set of independent valid wavefunctions, one for each possible k value, the electron density can be constructed by integrating the norm of all wavefunctions over k-space. By virtue of the periodicity of the lattice, all information regarding the repeating symmetry is encapsulated in the first Brillouin zone (BZ), centred around the gamma point. As a result of this, wavefunctions need only be considered at values of k within this BZ. The wavefunction evolves slowly as a function of k and so, to evaluate this in a computationally efficient manner, this integral is converted to a weighted sum over a specific set of k-points in the form of a density grid,

$$\psi(r) = \int_{\Omega_{BZ}} d^3k \, |\psi_k(r)|^2 \tag{2.17}$$

$$\approx \sum_{j} w_j \psi_{k_j}(r) . \tag{2.18}$$

2.6.2 Pseudopotentials and the Projector Augmented Wave method

The cut-off energy defined in Equation 2.16 provides an elegant parameterisation of the plane-waves to be used in a given calculation. Increasing the variable corresponds to an increase in the maximum frequency of plane-wave to be included, providing improved accuracy but at the price of further computational cost. With this in mind, the high kinetic energy, and thus high frequency plane-waves, of core electrons would prove expensive to be captured accurately, both in terms of time and computing resources. Compounded by the limited participation of these electrons in the reactions or chemical bonding often investigated by these methods, tools have been developed to incorporate core dynamics in a more efficient manner for calculations using non-localised basis sets. By encapsulating the Coulomb potential of core electrons into a smoother pseudopotential of a lower frequency, a smaller cut-off energy can be used. This pseudopotential is constructed to represent the screening effects of core electrons on the nuclear potential and presents this net core interaction to valence electrons.

A further evolution upon this approach is the projector augmented-wave (PAW) method [37]. Upon an assertion that the true all-electron wavefunction can be linearly transformed onto pseudised valence wavefunctions, this formalism defines an augmentation sphere centred around the nucleus of a given radius. At distances within this sphere, the all-electron wavefunction is transformed by projector functions so as to smooth the waves, whilst outside the sphere the pseudised valence wavefunctions are exact. By introducing an explicit linear transformation, physical quantities are similarly transformed, allowing for all-electron energies to be determined from the pseudised wavefunction.

First principles calculations presented in this work will make use of the Vienna Ab initio Simulation Package (VASP). Implementing the PAW method, VASP uses a proprietary scheme of real-space projectors, whilst providing a near-complete database of elemental PAW pseudopotentials, compatible for PBE calculations.

2.7 Utilisation of DFT

2.7.1 The potential energy surface and geometry optimisation

The potential energy of a system may be defined as a function of atomic coordinates, $E(\mathbf{x})$. This can be interpreted in the form of a hypersurface in multiple dimensions, which describes the potential energy of the system at each point along its surface for different atomic arrangements. This potential energy surface, or energy landscape, can yield structural information and corresponding energies for many systems in an intuitive, mathematical representation which can then be interpreted in terms of physical meaning.

Points of interest on the potential energy surface are often stationary points. Minima represent a net inter-atomic force of zero, suggesting stability of a given structure, whilst saddle points may be investigated to yield a transition state. Standalone maxima describe unstable states, which are often overshadowed in importance. There may be many stable states due to several minima on the energy surface, but the state with the lowest energy is represented by the global minimum.

In essence, locating these minimum points can be considered as a purely

mathematical optimisation problem seeking an arrangement such that the derivative of $E(\mathbf{x})$ with respect to the position vector, $\mathbf{x}$, is zero and the second derivatives are all positive. A range of optimisation algorithms may be used to try to minimise the forces at play by using $E(\mathbf{x})$, as well as its first and second positional derivatives. However, in many cases the full Hessian matrix may prove unjustifiably expensive to compute. Many minimisation algorithms work to only locate a local minimum for the given starting point. In order to search for a global minimum the algorithm must be able to span the hypersurface, analysing many starting points and optimising in turn or traversing the surface by other statistical means.

When investigating properties of a solid, it is imperative that the system being simulated is in its optimised ground state geometry as many processes and interactions are dependent on lattice parameters and exact positioning. Such calculations on non-equilibrated structures can lead to inaccurate results. Another consideration is that structures, and thus calculated values, will vary based on the level of accuracy of the calculation. Different functionals and parameterisation of the DFT code used may result in differences in results, which in turn may differ from empirical results.

2.7.2 Crystal structure prediction

When an initial structure or energy landscape is not known, sampling of multiple possible geometries is required in order to search for the minimum-energy atomic arrangement for a given composition. A concept known as crystal structure prediction (CSP), this process involves combinatorial sampling methods and knowledge of chemical interactions between constituent atoms to generate a range of chemically feasible structures, before calculating energies and outputting any resultant stable

or metastable structures. CSP has uses in the study of both organic and inorganic materials; systematically analysing, without human bias, multiple crystal packing possibilities that are potentially previously unseen by theory or experiment.

The Crystal structure AnaLYsis by Particle Swarm Optimization (CALYPSO) software [38, 39] uses the particle swarm optimization (PSO) algorithm and is an effective metaheuristic global optimisation method which starts from an initial guess of a set of geometries before traversing a parameterised energy landscape, repeatedly performing structural optimisations. Whilst ensuring specified minimal inter-atomic distances are maintained, multiple local minima are probed in order to map the potential energy surface. Genetic algorithms are used to maintain structural diversity, introducing randomly generated structures to each generation of the PSO algorithm in order to reduce the chance of stagnation in potential wells.

2.8 Molecular dynamics calculations

Whilst the aforementioned process of Kohn-Sham DFT calculations are performed for a static crystal, and so is independent of temperature related energy contributions and quantum fluctuations, it is often the case that such finite temperature dynamical trajectories are of interest. The notion of molecular dynamics (MD) is that of using atomic forces, determined from consecutive electronic structure calculations, to analyse the microscopic time evolution of a many-body system. This allows for analysis of time dependent processes such as transport properties, or energy and mass transfer.

In many cases, the inter-atomic or inter-molecular forces are pre-emptively parameterised in the form of a force field. These are empirical models tailored to specific individual, or specific classes of, materials and are tried-and-tested methods of encapsulating observed behaviours of certain systems.

An intrinsic limitation of this is that these force fields are fitted using previously observed or calculated results regarding specific systems and scenarios. For example, a force field parameterised to fit static observables may not accurately model dynamic properties, and *vice versa*. As such there is limited transferability of use for classical force fields. The bespoke nature of these interatomic potentials, and thus niche applicability, means that if no appropriate force field exists for the system at hand then the dynamics at play must be determined from first principles.

2.8.1 Ensembles and Ergodicity

The parameterisation of a realistic system to one suitable for computational calculation requires a degree of simplification. This can be done by the introduction of constraints, separating a system from the surrounding environment, whilst allowing for control of interactions between them, or by reducing an otherwise lengthy calculation required to exhaustively observe time evolution of a system into a large, yet manageable, number of discrete calculations representing possible states.

An ensemble is an exhaustive collection of states that are macroscopically identical but differ at the microscopic level (atomic positions, $\mathbf{r}$, and momenta, $\mathbf{p}$). These microstates are described by a selection of variables, most commonly; N particles, volume V, energy E, pressure p, temperature

T and chemical potential μ. By keeping some of these values constant, and relating the system to its external environment, a range of statistical ensembles may be defined such that each constituent microstate conforms to the macroscopic constraints of the system.

The process of molecular dynamics generates a trajectory where each point admits a set of 3-dimensional coordinates $\mathbf{r}^N$, as well as momenta $\mathbf{p}^N$, for the N particles in the system. By defining a $6N$-dimensional phase space, an observable property A for the ensemble as a whole, is given by the value of the property, weighted by the probability density, as calculated at a point in phase space:

$$\langle A \rangle_{ens} = \iint d\mathbf{p}^N d\mathbf{r}^N \, \rho\left(\mathbf{r}^N, \mathbf{p}^N\right) \, A\left(\mathbf{r}^N, \mathbf{p}^N\right). \tag{2.19}$$

An important tool in statistical physics is the ergodic hypothesis. The notion of ergodicity is a property of a mathematical system where by the entirety of a space that the system is said to exist in is traversed in a uniform but random manner. That is to say that a single trajectory will, over sufficient time evolution, sample the whole space. Equally, given a sufficiently large number of samples, a similar mapping can be obtained. The ergodic hypothesis assumes that over a reasonably long period of time, all available microstates are equiprobable and so the assumption is made to equate the time average and the ensemble average

$$\langle A \rangle_{ens} = \langle A \rangle_{time} = \lim_{\tau \to \infty} \frac{1}{\tau} \int_0^{\tau} dt \, A\left(\mathbf{r}^N(t), \mathbf{p}^N(t)\right). \tag{2.20}$$

The implication of this is that rather than observing many microstates of a given system, one can observe a time evolution of a single microstate over a sufficiently long time and obtain the same expectation value for an observable A.

2.8.2 Thermostats

Classical molecular dynamics are calculated with a constant number of particles, volume, and energy. Often denoted as the NVE ensemble in reference to these conserved quantities, this statistical ensemble is called the microcanonical ensemble. By modelling the system to be isolated from any external environment outside the simulation box, the conservation of energy ensures the total system energy is maintained constant.

More realistically, it is not feasible to completely control the energy of a system in an experimental scenario, whereas external factors such as pressure and temperature may be reasonably tightly regulated. An attempt to mimic conditions of an experiment requires an alternative ensemble method. Controlling the number of particles and volume, along with simulation temperature, the constant-temperature, constant-volume ensemble (also referred to as the NVT or canonical ensemble), is a better representation of experimental conditions, and works by introducing a heat exchange process with the external environment of the system without any transfer of matter. In order to model the energy transfer at the boundaries of an MD system it is modelled to be weakly coupled to a thermal reservoir. Thermostat algorithms are introduced as a modification to the classical MD calculation to facilitate the transfer of energy between the system and the reservoir in order to maintain a system temperature.

Many other ensembles exist, often characterised by conserved values of a selection of properties, as well as thermostats which will not be introduced here as they are not used in this work.

2.8.3 Nosé-Hoover thermostat

Outlined in Section 2.8.2, the conversion of a microcanonical ensemble to the more realistic canonical ensemble requires a numerical modification of the molecular dynamics approach so as to maintain the system temperature. One thermostating approach presented by Nosé and Hoover [40, 41] is to incorporate the notion of the heat bath into the system as an additional degree of freedom - a fictitious variable, ζ, with an associated coordinate, r, and effective mass, Q.

Introduced as an extension to the Newtonian equations of motion analogous to a friction term, this addition acts to keep the total kinetic energy constant:

$$\frac{\mathrm{d}v\left(t\right)}{\mathrm{d}t} = \frac{F\left(t\right)}{m} - \zeta v\left(t\right) \tag{2.21}$$

$$\frac{\mathrm{d}\zeta\left(t\right)}{\mathrm{d}t} = \frac{1}{Q}\left[\sum_{i} m_i v_i\left(t\right)^2 - (X-1)k_B T\right], \tag{2.22}$$

where Q determines the rate of temperature fluctuations, and X is the number of degrees of freedom.

A deterministic process, this approach does not impair correlated motions and thus is effective for describing kinetics and diffusion properties [42].

2.8.4 *Ab initio* molecular dynamics

As previously mentioned, classical molecular dynamics requires information regarding inter-atomic or inter-molecular forces at play in a given system. In many cases, this is specified in the form of a force field, constructed in a manner to be tailored to a given material, or class of materials, and for an explicit scenario. Otherwise, in situations without an *a posteriori*

description of the forces present, the method of *ab initio* molecular dynamics (AIMD) offers a process of calculating dynamical trajectories using forces obtained in a procedural manner as the simulation progresses.

Ab initio molecular dynamics, as implemented in VASP, generates dynamical trajectories by means of evaluating the time evolution of a system at a finite temperature *via* the numerical integration of classical equations of motion. By discretising these equations of motion in terms of a specified time step, Δt, the simple, yet effective integration scheme of Verlet is used to determine successive velocities and positions of particles:

$$v\left(t+\frac{\Delta t}{2}\right)=v\left(t-\frac{\Delta t}{2}\right)+\frac{f\left(r\left(t\right)\right)}{m}\Delta t+\mathcal{O}\left(\Delta t^{3}\right) \tag{2.23}$$

$$r\left(t+\frac{\Delta t}{2}\right)=r\left(t\right)+v\left(t+\frac{\Delta t}{2}\right)\Delta t+\mathcal{O}\left(\Delta t^{4}\right) \tag{2.24}$$

Here the forces at each velocity evaluation stage are evaluated from first principles, obtained from electronic structure calculations. These forces are used to update particle velocities and in turn positions, a process repeated until a desired simulation time is reached. For small Δt, higher order terms can be ignored.

Initial velocities are randomly generated which, when coupled with a thermostat, means the simulation requires an initial time period in order to equilibrate to a required temperature. Another consideration is the size of the time step; too large a step may not allow for correct sampling of high-frequency modes which may cause numerical issues such as problems with SCF convergence, however too small can result in inefficient sampling, which could be a waste of computing resources whilst not gathering enough useful statistics of time evolution.

2.8.5 'On-the-fly' machine learning force field molecular dynamics

The need to calculate forces for each step from first principles with AIMD means that the method can accrue a sizeable cost, both computationally and timewise. An emerging method offering an alternative approach is a molecular dynamics scheme utilising on-the-fly machine learning force field generation [43, 44, 45].

This method constructs a force field in the background of a molecular dynamics simulation, using the information computed from first principles. Trained using structure datasets, these consist of information regarding the Bravais lattice, and atomic positions, as well as energies, forces, and stresses obtained from *ab initio* calculation steps. From this data, local configurations around each atom can be probed for radial and angular relationships with neighbours, and the force field appropriately parameterised with such descriptors.

The algorithm for on-the-fly machine learning force field molecular dynamics (MLFF / MLFF MD) implemented in VASP works to adjudge at each MD step whether to calculate forces from first principles, or not. Through methodology akin to an AIMD step, doing so involves analysing new structures and gathering new data for the structure dataset, which can be used to further bolster the training of the force field in the background, hence 'on-the-fly' learning. Alternatively, the MD step can be assessed by using said force field, skipping a training step, but drastically reducing the time taken to evaluate the equations of motion. By interweaving the two options the dynamics of a system can theoretically be captured in a bespoke force field, which may be extracted for further use, whilst also making use of it throughout the training process to speed up the MD simulation.

2.8.6 Mean square displacement

The random movement of particles diffusing through a system can be compared to the mathematical notion of a random walk. By taking a series of steps in random directions throughout the available degrees of freedom, a naïve sum of displacement of particles would admit an average displacement of zero. In order to capture the gross movement of particles over time, the square of the displacement for each particle between each time step is averaged over the number of particles being considered to disclose the mean square displacement (MSD). The Einstein formula for MSD evaluated at a given time t_0, for N particles with n-dimensional coordinates x is given as

$$MSD = \left\langle \frac{1}{N} \sum_{i=1}^{N} |x_n - x_n(t_0)|^2 \right\rangle_{t_0}. \tag{2.25}$$

For n-dimensional Brownian motion over a given time, t, the MSD is related to a coefficient of diffusion D which can be shown to be

$$MSD = 2nDt. \tag{2.26}$$

From here it can be seen that in an isotropic, three-dimensional medium, a diffusion coefficient can be determined by dividing a calculated MSD value by $6t$.

When considering MSD calculations for individual particles, displacements corresponding to certain time intervals, or time lags, between positions allows for maximisation of samples garnered from the MD calculation. Defining a time interval τ, time lags are given as

$$\tau, 2\tau, 3\tau, ..., N\tau \tag{2.27}$$

where $N\tau \leq$ the total calculation time, T. Taking MSD as a function of time invervals, τ, results in data collected for each length interval regardless of position throughout the duration of the simulation,

$$MSD\left(\tau\right) = \left\langle \frac{1}{N}\sum_{i=1}^{N}\left|x_n\left(t_0+\tau\right)-x_n\left(t_0\right)\right|^2 \right\rangle_{t_0}. \tag{2.28}$$

For large T and $\tau \ll T$, one may assume ergodicity and evaluate an ensemble average MSD. Such an approach greatly increases statistical performance, providing many values from a single trajectory. A consequence of averaging over time lags is that the data has ballistic regions corresponding to small and large τ values, whilst statistically significant information is gleaned from the linear trend obtained across a truncated central region. By plotting MSD as a function of time-lags, the slope of such linear trend can be calculated and Equation 2.26 admits

$$D = \frac{1}{2n}\frac{MSD}{\tau}. \tag{2.29}$$

Chapter 3

Metal hydride predictive modelling

3.1 Introduction

As outlined in Section 1.5, the potential for metal hydride storage systems relies on the intrinsic material properties that govern performance, for example storage capacity, as well as how a use case might accommodate such a system's size and weight. With regards to a solid-state hydrogen storage material, the material weight of interest refers to the mass of constituent atoms, whilst the material gravimetric capacity corresponds to the relative mass of hydrogen to the other atoms involved. A crucial thermodynamic factor that governs the conditions under which such a storage material will absorb and desorb hydrogen is the dehydrogenation enthalpy, or enthalpy of formation, of the metal hydride species (ΔH_f). Intuitively, this is the measure of the change of enthalpy during the formation of a substance from its constituents, given a specified reaction pathway.

The operating conditions permissible for on-board chemical storage for a light-duty vehicle suggests desirable dehydrogenation enthalpy in the region of -10 to -60 kJ/mol$_{H_2}$ [46, 47]. Experimentally investigating this enthalpy value for a wide range of materials would require many material synthesis and characterisation processes, as well as numerous hydrogenation cycles of these hydrides.

The aim of this chapter is to construct a machine learning model using readily available data, independent of structural information, in order to predict the enthalpy of formation for a presented metal hydride composition. This would be useful for identifying novel hydride materials or for further analysing metal hydrides that may have already been synthesised, but have yet to have their hydrogen storage properties properly characterised. If of a reasonable predictive accuracy, this tool could provide a means of drastically narrowing the compositional space when searching for a compound with a target enthalpy value, whilst doing so at a relatively low expense.

3.2 Hydride storage materials literature

3.2.1 Miedema model

Miedema and others worked to devise a semi-empirical model, seeking to match known experimental results for enthalpy of formation [48, 49, 50, 51]. An assumption is made that the Wigner-Seitz atomic cells of metals, A and B, within a binary alloy structure are similar to the atomic cells of the corresponding pure metals. The hydrogenation mechanism is taken to be an interaction along the interface between atomic cells of A and B,

raising two main energetic contributions. The first is a contribution of negative value, representing the loss of atomic contact between A and B metals, which corresponds to the binary compound's formation enthalpy. The second results from an approximation that the contact surface for A-H and B-H is the same. It therefore follows that AB_nH_{x+y} can be given to be energetically equivalent to a mixture of AH_x and B_nH_y [52]. Often referred to as the 'rule of reversed stability', ΔH_f of a ternary AB_nH_{x+y} can thus be given as

$$\Delta H_f\left(AB_nH_{x+y}\right) = \Delta H_f\left(AH_x\right) + \Delta H_f\left(B_nH_y\right) - \Delta H_f\left(AB_n\right). \quad (3.1)$$

As the informal name implies, this equation suggests an inversely proportional relationship between the stability of a ternary hydride and a binary alloy consisting of the corresponding intermetallic species.

Further development of this concept introduces a modification based on empirical results for systems where hydrogenation does not break all bonds between A and B. More generally, and particularly for small n,

$$\Delta H_f\left(AB_nH_{x+y}\right) = \Delta H_f\left(AH_x\right) + \Delta H_f\left(B_nH_y\right) - (1-F)\Delta H_f\left(AB_n\right). \quad (3.2)$$

Here, F is dependent on the composition of the alloy and empirically derived values for F are often used, relating to the B element involved [13].

This model has been shown to work reasonably well for binary hydrides, however it has issues with predictions for alkali metal hydrides. For ternary hydrides, the model generally over predicts the enthalpy value, proposing higher stability than seen in experiment, through a non-systematic trend of deviations [13]. A predictive machine learning model trained on known stability information may have improved predictive accuracy, as well as

the potential to incorporate any underlying relationships that may be responsible for the irregular deviations seen here.

3.2.2 Machine learning material discovery for hydride materials

In recent years, general approaches to the task of material discovery have shifted from a more Edisonian approach of trial-and-error to a more systematic regime of utilising theoretical data to rationalise experimental choices (see Section 1.7.1). This is true in many branches of materials science and engineering, and has been demonstrated in the field of metal hydride systems for solid-state hydrogen storage.

A publication by Hattrick-Simpers et al. covers the development and implementation of a regression model to predict the enthalpy of hydrogenation of metal hydrides for high pressure compressors [53]. Data regarding a wide range of storage materials was used from the *Hydrogen Storage Materials Database*, a collaborative effort between the International Energy Agency (IEA) and the U.S. Department of Energy (DOE) to collate results from their funded research projects, non-DOE research, and computational models into a comprehensive repository [54]. Material classes including interstitial Laves phase material hydrides, complex hydrides, and solid solution interstitial hydrides are included in the mixed theoretical-experimental search space. After reducing this sample space to only reversible metal alloys, with explicitly reported formation enthalpy values, a random forest regression model is constructed using the Weka software platform to predict hydrogenation enthalpy for each corresponding intermetallic species. Following this, binary, ternary, and quaternary alloys are generated using the elements Ca, Al, Si, Fe, Mg, Na, Mn, Zn, Cr,

Mo and Ti, filtered by constraints such as an upper limit to the cost per kilogram for the alloy, confining chemistries to AB, AB_2, A_2B or AB_5 Laves phases, and predicted enthalpy values ranging from -18 kJ/mol_{H_2} to -30 kJ/mol_{H_2}. Using a genetic algorithm for structure and phase prediction, phase space for target compositions was sampled and DFT calculations used to verify predictions with limited success at validation.

Rahnama et al. reported an investigation involving two separate machine learning approaches concerning hydrides for storage applications, released as consecutive journal entries [55, 56]. The first involved testing a range of regression processes to predict hydrogen weight percentage. Training data is again collected from the *Hydrogen Storage Materials Database*, however only the entry with highest hydrogen concentration result for a given intermetallic is taken. The predictive performance of four regression models are compared, namely linear regression, neural network, Bayesian linear regression and boosted decision tree regression, as implemented in Microsoft Azure Machine Learning Studio. It was found that the best predictive performance corresponded to boosted decision tree regression, and feature importance analysis admitted the highest ranking descriptor to be material class, followed by temperature and then the heat of formation, whilst composition formula was shown to be an insignificant variable. The second entry by Rahnama et al. instead focused on developing a model to classify metal hydrides into materials classes based on the values of properties collected from the *Hydrogen Storage Materials Database*.

Work by Witman et al. rather uses an entirely experimental hydride database to construct an empirical based machine learning model to predict the natural logarithm of the equilibrium pressure of H_2 at ambient temperature ($\ln P^{\circ}_{eq}$) [57]. Using the *HydPARK* database, a repository of experimental metal hydride information and empirical

results constructed through collaboration between the International Energy Agency (IEA) and the U.S. Department of Energy (DOE) [58], data pruning removes compositions with incomplete data such that $\ln P^{\circ}_{eq}$ cannot be calculated, and duplicate entries for a given composition are represented by their median value. Building a gradient boosting regression model as implemented in scikit-learn [59], and using Magpie descriptors [60], the resultant model admits strong feature importance for the mean volume per atom in the ground state structure. This is further expanded on, using this structure-property relationship as justification for DFT analysis of similar materials, examining A-site substitutions in the $LaNi_5$ series.

3.3 Model training methodology

In principle, the construction and subsequent use of a machine learning model is procedural. As sketched in Figure 3.1, this process consists of three main regimes; data collection and processing, model construction and testing, and ultimate application.

Data collection: All models require training data from which to analyse any form of underlying mathematical pattern. This data may take many forms, and can be obtained from a range of sources. For regression applications in material science, this encompasses experimental results, theoretical calculations, or a combination thereof. In addition to this, descriptor variables must be chosen and the corresponding information collected or generated along with, in the case of supervised learning, data to be taken as the ground truth values of the target variable.

In order to obtain a robust model boasting a strong mathematical

representation of the thermodynamic mechanisms that govern the target variable, this data must be reliable, as covered in Section 2.3.2. Inconsistency in simulation or experimental methods between data points, or simply incorrect results, can contaminate the dataset and impact the predictive ability of the resultant model. Data must therefore be cleaned, pruning the training dataset of troublesome results.

An appropriate algorithm should be used to construct the model. Learners perform differently on different datasets and so a suitable choice of algorithm may require investigation of multiple candidates, whilst also considering optimisation of corresponding hyperparameters.

Model construction: With a finalised, cleaned dataset, the results are split into subsets so as to enable testing of the model's performance at a variety of stages throughout construction. The majority will be used for the actual training of the model, providing information of the chemistry of these systems as encoded by descriptor values, whilst a random sample should be held-out in order to test the predictive ability of the final model. A portion of training data is used during construction to evaluate performance at various stages during the training process, however cross-validation may be performed instead. Throughout this work, this is quantified by the mean absolute error (MAE) of prediction relative to ground truth values.

Once the model has been built according to the initial parameterisation of the estimator, the held-out test set can be used to assess accuracy of prediction when the model is presented with previously unseen compositions, which should be indicative of the model's ability to generalise to new data.

Application: After the model has been built, it can be used to predict the target variable for any composition. For this work, new compositions are generated by a heuristic process. Descriptor values are determined for each formulae and run against the trained model, outputting a predicted enthalpy value. From here, a variety of filtering criteria are established to reduce the sample space of these candidate materials.

Once reduced sufficiently, structure prediction processes are used to expand upon the information known about sampled compositions, allowing for further analysis by means of first principles calculations. By evaluating energies for the final hydride compound, as well as possible dehydrogenation product species, the accuracy of the enthalpy prediction for such materials can be verified.

Whilst Figure 3.1 depicts a generalised model construction process, the work presented in this chapter requires slight deviation to compensate for nuances in our investigation. To account for the limited size of the available training dataset, cross-validation is used as an alternative to an independent validation set. Additionally, the descriptor data is suitably formatted for input to the relevant algorithms, so as to bypass the pre-processing of data that may otherwise be required for non-explicitly numerical descriptor information.

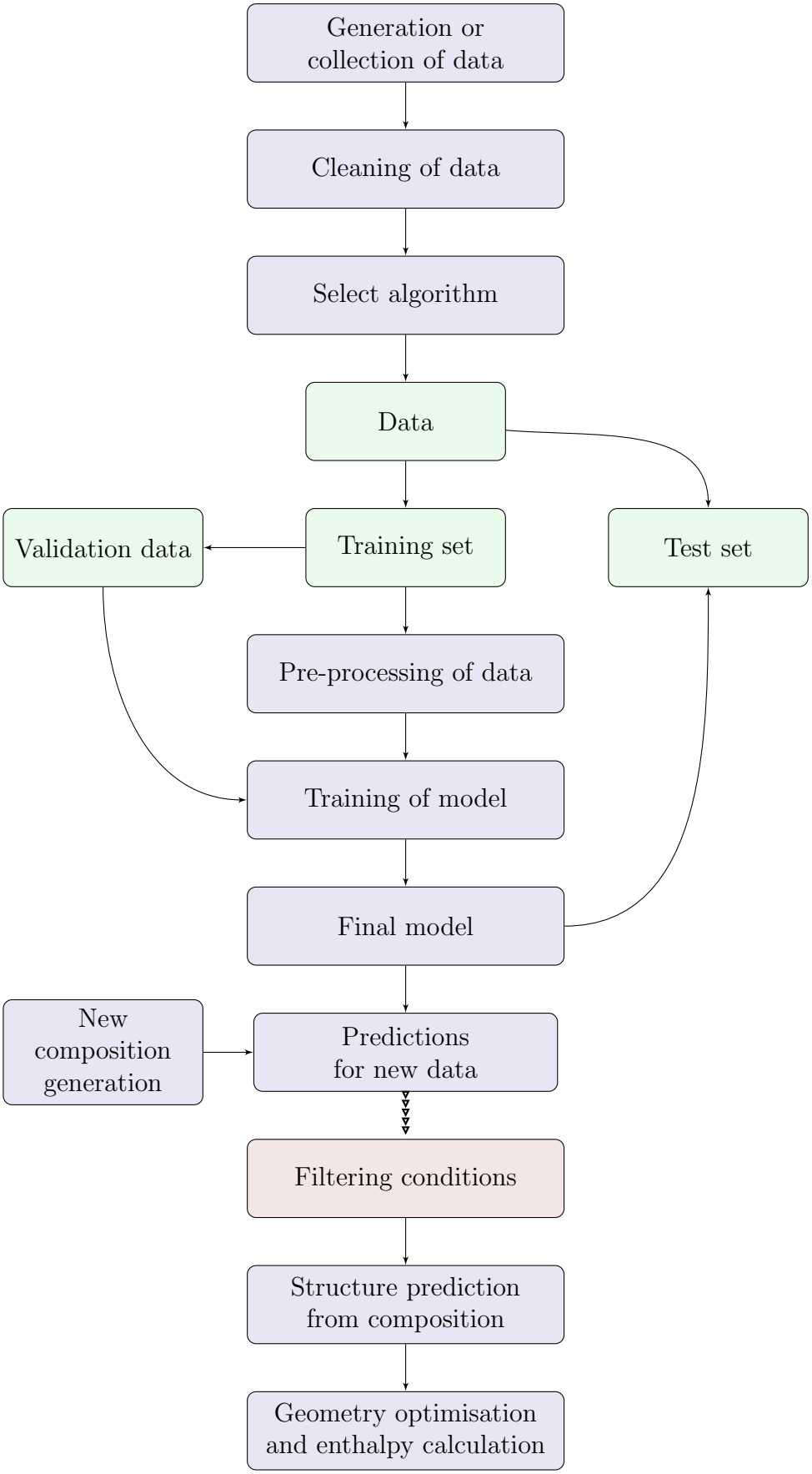


Figure 3.1: Workflow diagram for data collection, followed by model construction, testing, and application.

3.3.1 Cross-validation

Constructing a machine learning model involves parameterisation of the fitting function with respect to the set of training data presented. Testing performance, and thus these parameters, using the samples involved in training would bias the model to these results. This causes overfitting of the learner and can lead to poor predictive accuracy when presented with yet-unseen data. Common practice is to partition the input dataset into training, validation, and testing sets. From here, the test set is kept aside whilst the model learns from training data, evaluating and adjusting fitting parameters according to accuracy with relation to known validation results. Once optimised, final evaluation against the test sample space provides quantification of performance according to chosen metrics. When dealing with smaller sets of data, a three-way splitting of samples may have a notable impact on the model by limiting the amount of available data points for the learning process. Additionally, prediction issues may arise as a result of correlations between training and validation data.

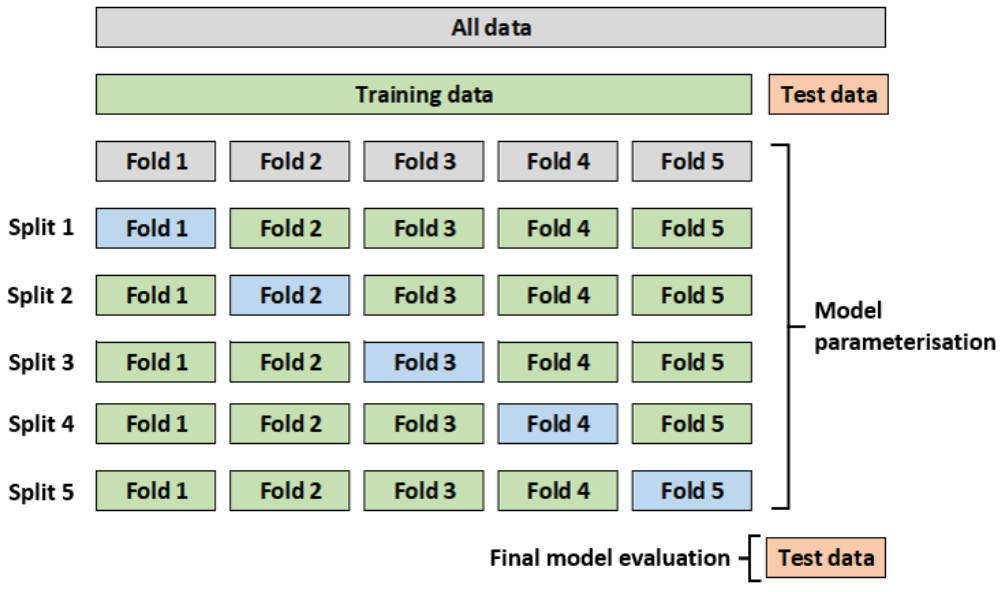


Figure 3.2: Schematic of a k-fold splitting of training data, in this case 5-fold.

An alternative method is by performing cross-validation (CV). For this, data need only be split into train and test sets, allowing what would have been the independent validation samples to be included in the learning process. A simple CV scheme is that of k-fold CV which intuits splitting the training set into k subsets. The model is constructed using $k-1$ subsets, whilst validated against the remaining one. This is repeated k times such that each fold is used to validate the rest of the data, after which parameters are averaged across all splits in the loop, as shown in Figure 3.2. This fitted model is then used against the separate test set to determine predictive performance.

3.3.2 Hyperparameter optimisation

The parameterisation of a predictive function, such as Equation 2.1, is determined by the machine learning algorithm being used. As explained in Section 2.1, such an algorithm itself can be controlled by hyperparameters in order to parameterise the optimisation procedure. Scikit-learn, the toolkit used in this work for construction of a variety of machine learning models, provides sets of default hyperparameters for each of its regression functions. However, in order to maximise predictive ability and build a more robust model, one should investigate methods of selecting optimal hyperparameters, given the dataset and method.

One method used early in this investigation is to exhaustively train models over a grid of specified hyperparameters. *GridSearchCV*, as implemented in scikit-learn, operates in a fairly self-explanatory manner; given a dictionary of hyperparameter names and possible values, a k-fold cross-validated grid-search over combinations of such possible values is performed and the optimal choice presented. This approach requires some intuition as to

the values certain hyperparameters may take, and it can be seen that the consideration of additional variables, values, or combinations of such, can greatly add to time complexity.

Another process used is that of a black box Bayesian optimisation of a given hyperparameter space, as implemented in scikit-optimize [61]. One defines target ML algorithm variables and an appropriate range for each, as well as an objective function to minimise, which in this work is set to be the MAE of a 5-fold GBR model trained on the input data. The function is approximated using a Gaussian process whilst optimising a cheap acquisition function - expected improvement, by default. This is performed by iteration over the antecedent distribution at each step for a specified number of calls.

3.4 Data representation

As outlined in Section 2.1, the machine learning process works to ascertain a relationship between a set of variables, fitting a function to evaluate a target property. By this definition, the process can be used to study materials where these values encode information about the system, being correlated with material properties and combining to portray macroscale or microscale mechanisms [62].

A mathematical representation of materials data, theoretical descriptors are based on symbolic representations of molecules. This can vary from low-dimensional data such as information regarding constituent atoms and relative stoichiometry, to topological information including structural features, distance and Coulomb matrices, and even further to spatially dependent descriptors, encoding atomic coordinates [63]. As the complexity

and scope increases, so does the complexity of respective calculations to obtain the descriptor values.

For the work presented in this thesis, machine learning models are constructed using descriptors which are independent of crystal structure information. By using a descriptor generation system based on constituent elements, stoichiometry, and atomic structure, the model can be used to predict results for unseen compositions for which the crystal structure is unknown. This is because the patterns and correlations interpreted by the ML training process is based on information implicitly encoded by the elements involved, and relative quantities.

Magpie, or Materials AGnostic Platform for Informatics and Exploration [60], uses elemental property databases and ratios of elements present, along with electronic and ionic attributes, to construct a set of 131 descriptors. The diverse range of attributes covered results in a widely applicable set of values to describe many different classes of materials [64, 65, 66]. By using Magpie to generate features for initial training compositions and fitting to ground truth property values, we can then repeat the process to featurise either test compositions or unseen data.

3.4.1 Data source - OQMD

The material-related ground truth enthalpy data and compositional information used to construct our machine learning models is acquired from the Open Quantum Materials Database (OQMD) [67, 68]. It contains plentiful data on relevant inorganic compounds, including thermodynamic and structural properties. This can be downloaded as a searchable SQL database. Calculations for OQMD entries are performed using VASP,

making use of the PBE GGA approximation of exchange and correlation referred to in Section 2.5.3. A range of relaxation schemes are used to initially test physical feasibility and to consider inclusion of magnetism, before an iterative process of refinement involving higher cut-off energies and denser k-point meshes are implemented.

With the latest version, as of writing, boasting approximately 300,000 structures [69], approximately 10% of these are obtained in partnership with the Inorganic Crystal Structure Database (ICSD) [70, 71]. The world's largest database for identified inorganic structures, ICSD is a collection of experimental and theoretical inorganic structures, collated and catalogued alongside their method of synthesis or calculation. This internal methodological inconsistency is a key motivation behind the OQMD approach of standardising the calculation parameterisation across all structures.

The remaining ∼90% of data is computed from prototypical structures generated for a range of Strukturbericht types [69]. A process known as crystal structure prediction by analogy, this includes many unary, binary, ternary and quaternary compositions fit to realistic symmetries and stoichiometries, similar to known examples. Resultant systems are then processed in the same systematic manner as existing data to provide approximation of convex hulls and relative stabilities. Whilst this approach greatly improves the sample size of materials data, the large proportion of data generated for newly generated compositions and configurations potentially poses issues for the integrity of statistical learning tools built upon this foundation.

Understanding the derivation of formation enthalpy as implemented for OQMD entries is of great importance for this work. As presented by Kirlin

et al. [68], the formation enthalpy is generally given as

$$\Delta H_f = E_{tot} - \sum_i \mu_i x_i, \tag{3.3}$$

where E_{tot} is the DFT total energy of a given compound, and μ and x are the chemical potential and quantity of an element i in the compound, respectively. The convention used consistently throughout the database is to equate the chemical potential of a given species to the DFT total energy calculated for the elemental ground state. Doing so assumes a reaction pathway of

$$\alpha A + \beta B + \gamma C \rightarrow A_\alpha B_\beta C_\gamma, \tag{3.4}$$

for any elements A, B, C. Implications of this will be further discussed.

3.4.2 Hydride data

Data from OQMD is downloadable as a MySQL database dump from https://www.oqmd.org/download/ and accessed by use of the *qmpy* Python backend [72]. Results for hydride compositions were collected by exhaustively searching for combinations of compositions of varying chemical complexity, comprised of metallic elements and hydrogen. Querying a *Composition* object for the *delta_e* entry value admits the lowest enthalpy of formation for that particular stoichiometry, considering the possibility of multiple structures for a given chemical formula. The work in this chapter utilises OQMD version 1.3.0 for hydride compositions and corresponding ground truth formation enthalpy values. Released in October 2019, data was collected for hydrides of three degrees of complexity; binary (A_xH_i), ternary ($A_xB_yH_i$), and quaternary ($A_xB_yC_zH_i$) compounds, for any metals A, B, C. The availability of information for each such class of composition

is shown in Table 3.1.

Binary	Ternary	Quaternary	Total
336	451	59	846

Table 3.1: Availability of results for each class of hydride composition collected from OQMD v1.3.0.

A range of hydride chemistries are included in this sample set, including those based on A2B, AB AB2 and AB5 alloys, as well as alanate structures, and other miscellaneous systems generated through the prototypical generation method.

3.4.3 Caveats of such data

It should be noted that there is scope for possible inaccuracies in the database results. It is possible that the prototype-based method for the generation of new compositions and structures simply outputs an incorrect, or unrealistic structure, or that it doesn't contain the most stable crystal structure for a given composition. There is also a chance of issues with the calculations performed on such structures. The SCF calculations may have been incorrect, or improperly converged, thus admitting an incorrect ground state structure from which material property values have been calculated. In addition to all of this, there may be correctly calculated results admitting extreme values that may skew the fitting of a model. A data pruning method for cleaning the dataset prior to model construction will be covered later in this chapter.

3.5 Comparative testing of ML algorithms

Supervised learning is an ever-developing field and algorithms have been designed for a multitude of use cases [73]. For the problem at hand, it is important to consider the performance of a selection of such methods, in order to find an optimal process for the final predictive model. A range of nonlinear processes, as well as standard linear regression, were selected in order to compare predictive performance for the nonlinear hydrogenation mechanism [74].

Multiple models were constructed using training data obtained from version 1.3.0 of the OQMD database. They were developed using the below methods as implemented in scikit-learn [59]. Each method had certain characteristic hyperparameters cross-validated *via* the GridSearchCV process across a suitable range and the predictive error recorded for the optimal combination.

- Linear regression.
  - Default hyperparameters.
- Kernel ridge regression.
  - '$alpha$' $= 10^{-n}$, for n $\in [0, 10]$
  - '$gamma$' $= 10^{-m}$, for m $\in [0, 10]$
  - All else, default hyperparameters.
- Lasso cross-validation.
  - 'n_alphas' $=$ n, for n $\in [1, 71]$
  - All else, default hyperparameters.

- Random forest regression.
  - 'max_depth' = n, for n $\in \{1, 2, 3, 4, 8, 10\}$
  - All else, default hyperparameters.
- ExtraTrees regression.
  - 'max_depth' = n, for n $\in \{1, 2, 3, 4, 8, 10\}$
  - All else, default hyperparameters.
- Gradient Boasting regression.
  - 'max_depth' = n, for n $\in \{1, 2, 3, 4, 8, 10\}$
  - '$learning\_rate$' = 10^{-m}, for m $\in [0, 10]$
  - All else, default hyperparameters.

Additionally, this entire process was calculated for a variety of train/test cuts, using 5-fold cross-validation, as well as across five separate seeds of initial data randomisation.

From the results shown in Table 3.2, it can be seen that the tree-based ensemble approaches performed best amongst the methods tested. There is similar performance from ExtraTrees Regression and Random Forest Regression, however Gradient Boosting Regression is best-in-class with the lowest test error across all test splits. GBR also appears to better capture the mechanics aimed to be represented by the fitting process, with a noticeably smaller train error than other methods. As such, it shall be used going forward.

Method	Test split proportion	Average test MAE, eV/atom	Average train MAE, eV/atom
GBR	0.33	0.2066	0.0368
	0.25	0.1949	0.0422
	0.20	0.1976	0.0427
	0.15	0.1920	0.0367
	0.10	0.1870	0.0398
ETR	0.33	0.2117	0.0595
	0.25	0.1968	0.0625
	0.20	0.1991	0.0622
	0.15	0.2064	0.0655
	0.10	0.2004	0.0662
RFR	0.33	0.2088	0.0588
	0.25	0.1987	0.0634
	0.20	0.2004	0.0623
	0.15	0.2075	0.0657
	0.10	0.2008	0.0664
Linear	0.33	0.3199	0.2402
	0.25	0.3131	0.2588
	0.20	0.2669	0.2300
	0.15	0.2718	0.2325
	0.10	0.2666	0.2390
KRR	0.33	0.2890	0.2250
	0.25	0.3131	0.2588
	0.20	0.2669	0.2300
	0.15	0.2718	0.2325
	0.10	0.2556	0.2334
LASSO	0.33	0.3175	0.2978
	0.25	0.3135	0.3016
	0.20	0.3141	0.3016
	0.15	0.3183	0.3016
	0.10	0.3036	0.3050

Table 3.2: Average test and train MAE values for the selection of algorithms, averaged over five instances of randomised test and train set allocations for each test split value.

3.6 Data cleaning process

3.6.1 Noise in data

By the motivation outlined in Section 2.3.2, for the benefit of the generalisability of the final model, it is important to consider the quality of data used in the model training process. This is accomplished by identifying and removing troublesome outliers from the base OQMD dataset.

3.6.2 Procedure

The original hydride data acquired from OQMD, as summarised in Table 3.1, is represented by the histogram plot in Figure 3.4. Divided into 35 bins according to the given ground truth enthalpy values, it can be seen to have a long upper tail representing a proportionally small number of highly unstable compounds. An initial gradient boosting regression model was built using this data, as implemented in scikit-learn and optimal hyperparameters generated with scikit-optimize.

Using 5-fold cross-validation across the whole dataset, each data point is presented on the prediction error plot Figure 3.3. With a mean absolute error of 114.4 meV/atom, the data points can be seen to be fairly spread out, with a degree of correlation but poorly represented by this model. Further to this, the line plot overlaid on top of the histogram in Figure 3.4 depicts the average MAE for data points with formation enthalpy corresponding to each bin. It is clear that the model has two main regimes of predictive performance; a consistent predictive error for more populated regions of enthalpy values, and a sparsely represented region of enthalpy data with much higher error values.

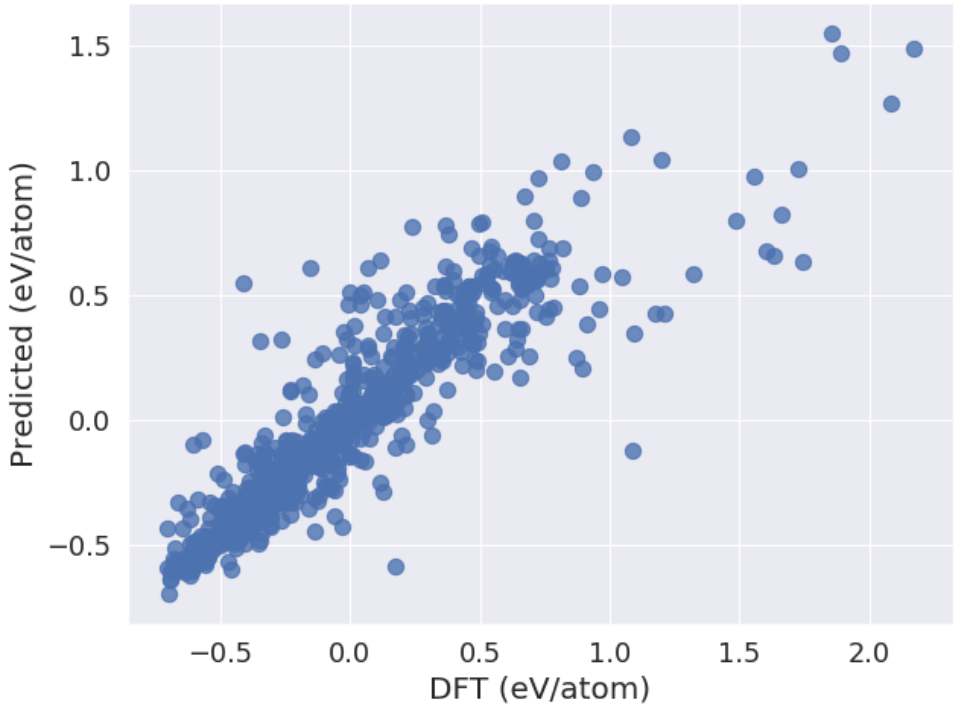


Figure 3.3: Prediction error plot for GBR model constructed using the base hydride data obtained from OQMD.

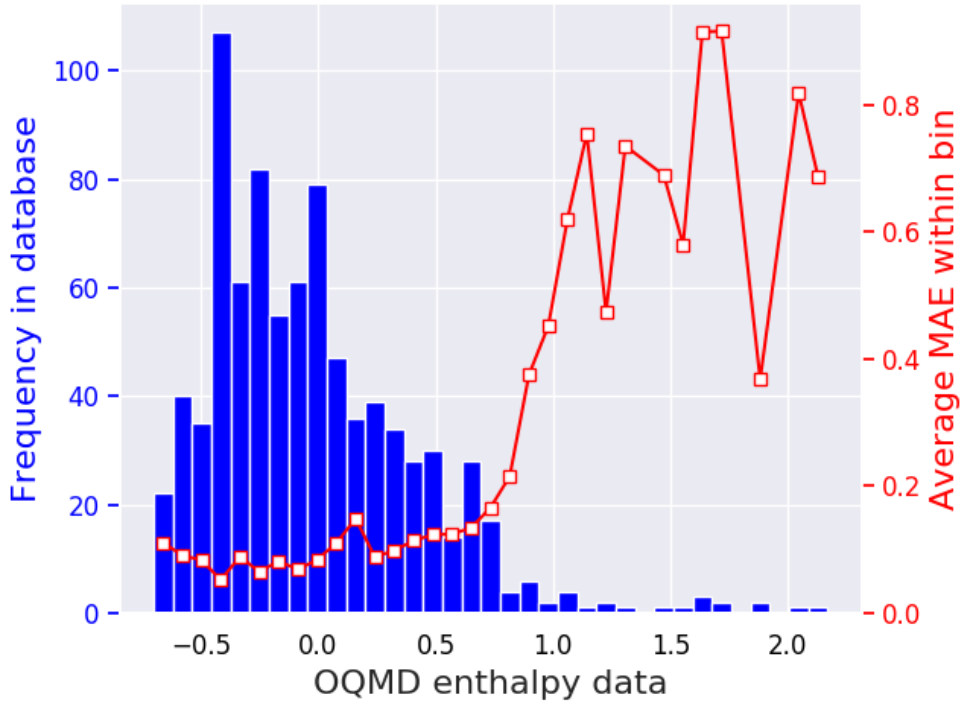


Figure 3.4: Histogram of formation enthalpy distribution for base hydride data obtained from OQMD, overlaid with MAE for data in each bin per the GBR model represented in Figure 3.3.

As suggested in Section 3.4.3, there may be several reasons for the existence of outlier data points amongst this theoretical data. In order to improve the quality of data, a three-stage pruning process was undertaken to remove outliers. The first stage consists of removing hydride compositions with unrealistic stoichiometry with regards to having a low hydrogen content. Compounds were filtered to remove any with a gravimetric density of less than 0.5 wt% of hydrogen, or any composition where hydrogen represents less than 25% of constituent atoms. This process not only helps to remove some rogue compounds that may have been generated as part of the prototype-based generation routine, but also removes those with notably poor hydrogen wt% that would be unfeasible as storage materials. Doing so further consolidates the compositional space to metal hydrides with non-trivial hydrogen content, the class of materials intended to be represented by such a model.

This reduces the sample size to 722 compositions, which were then used to generate a new GBR model, again with the same methodology of scikit-learn and scikit-optimize. Admitting a MAE = 102.5 meV/atom, Figure 3.5 suggests a stronger correlation between model output and known values, as shown by a reduction in the general spread of points, symbolising prediction error for data along an $x = y$ relationship. Remaining outlier results appear to follow an alternative linear relationship, and must be accounted for by either inclusion to the model, or removal from the training set.

The MAE plot in Figure 3.6 shows that this GBR model still presents relatively high errors in less represented enthalpy ranges, but appears to perform well elsewhere. Making use of this generalisation relationship, a new filtering procedure was developed. Taking the remaining materials data, two hundred independent GBR models were constructed using a 25%

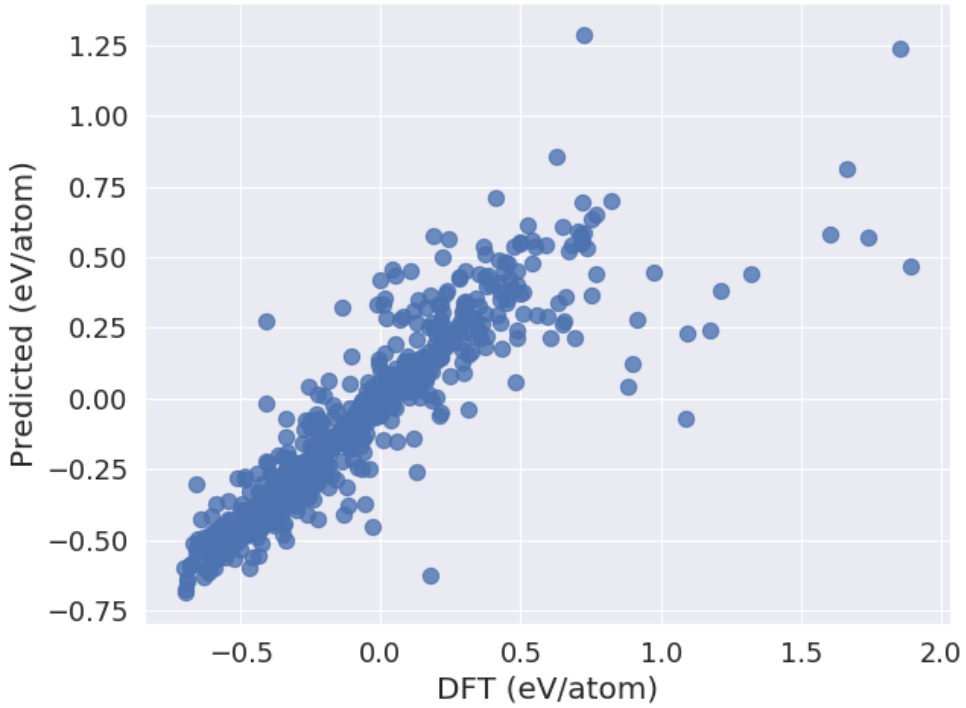


Figure 3.5: Prediction error plot for GBR model constructed using the remaining hydride data following the first filtering stage.

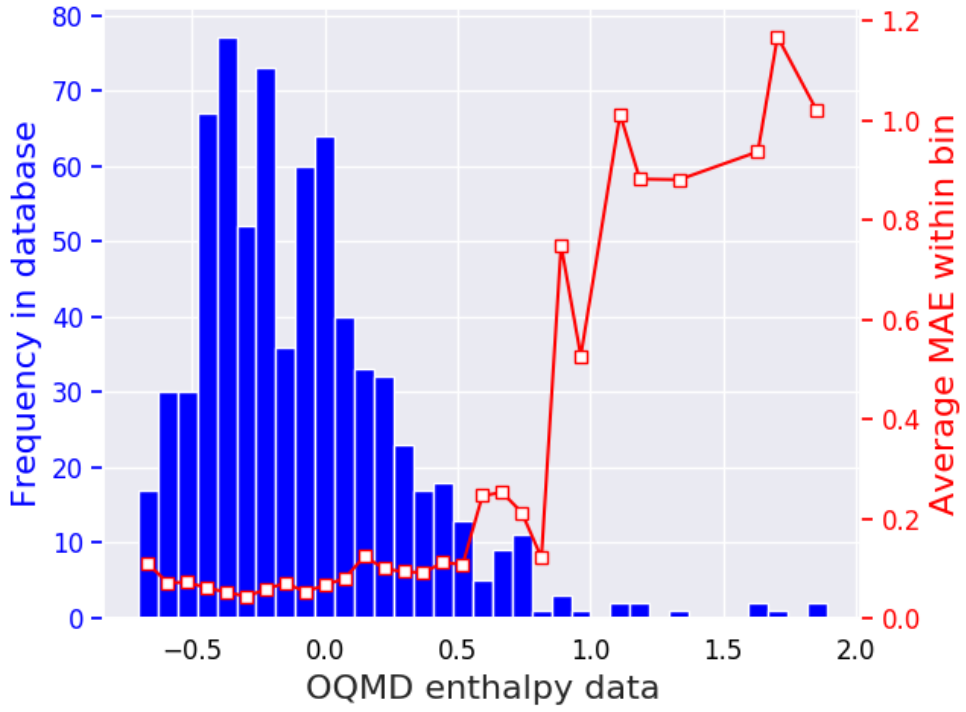


Figure 3.6: Histogram of formation enthalpy distribution for the remaining hydride data following the first filtering stage, overlaid with MAE for data in each bin per the GBR model represented in Figure 3.5.

test split, each using a different random seed to assign training/testing data. For each iteration, the test MAE for each material included in that model's test set is noted. After the two hundred models have been built and errors recorded, the MAE scores are averaged per composition, and the compounds with an error greater than 0.4 eV/atom are discarded from the dataset. This entire process was iterated until the maximum average MAE result was less than 0.4 eV/atom.

In order to produce results in a timely fashion, but at the risk of introducing further variance, hyperparameters were generated for a one-off build of a GBR model using the initial dataset. The hyperparameters defined to deviate from default values are given in Table 3.3, the first of which was chosen to ensure a good fit, whilst the others were optimised by use of scikit-optimize.

n_estimators	1000
max_depth	200
learning_rate	0.013510
max_features	90
min_samples_split	8
min_samples_leaf	1
loss	'lad'

Table 3.3: Hyperparameters chosen or generated for iterative data cleaning process.

This process reduces the composition set down to 694 compounds. A model was constructed using the resultant dataset from this method and predictive errors can be seen in Figure 3.7, showing a much more robust relationship with a reduced prediction error of 74.9 eV/atom. Seen here, as well as in Figure 3.8, the majority of samples possessing larger positive enthalpy values have been removed, consistent with the outliers observed in the previous stage of filtering. The binned MAE chart continues to suggest

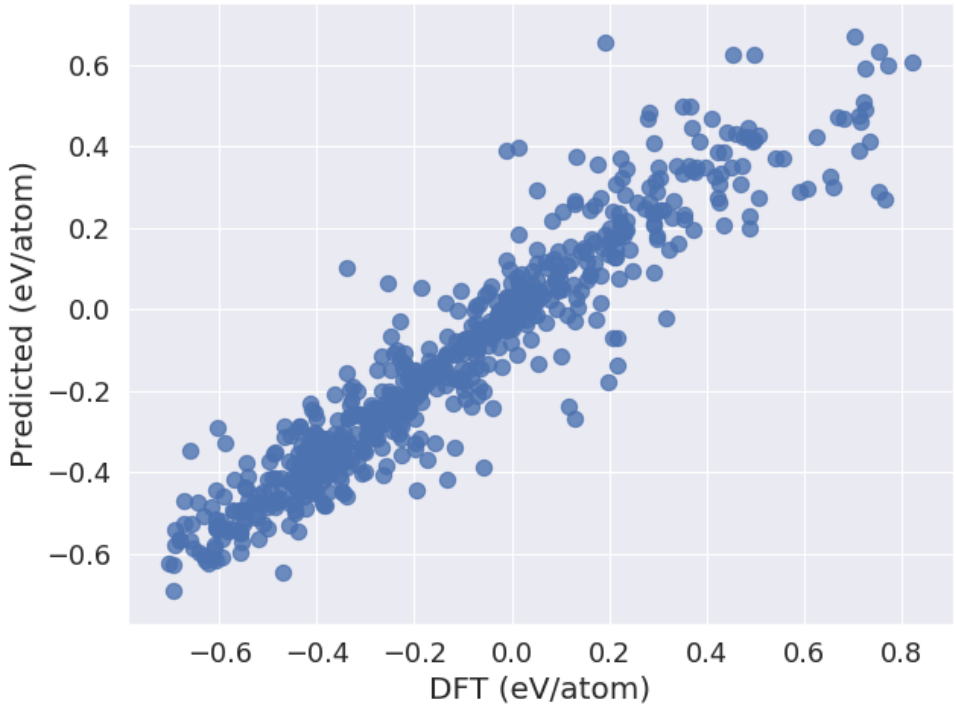


Figure 3.7: Prediction error plot for GBR model constructed using the remaining hydride data following the second filtering stage.

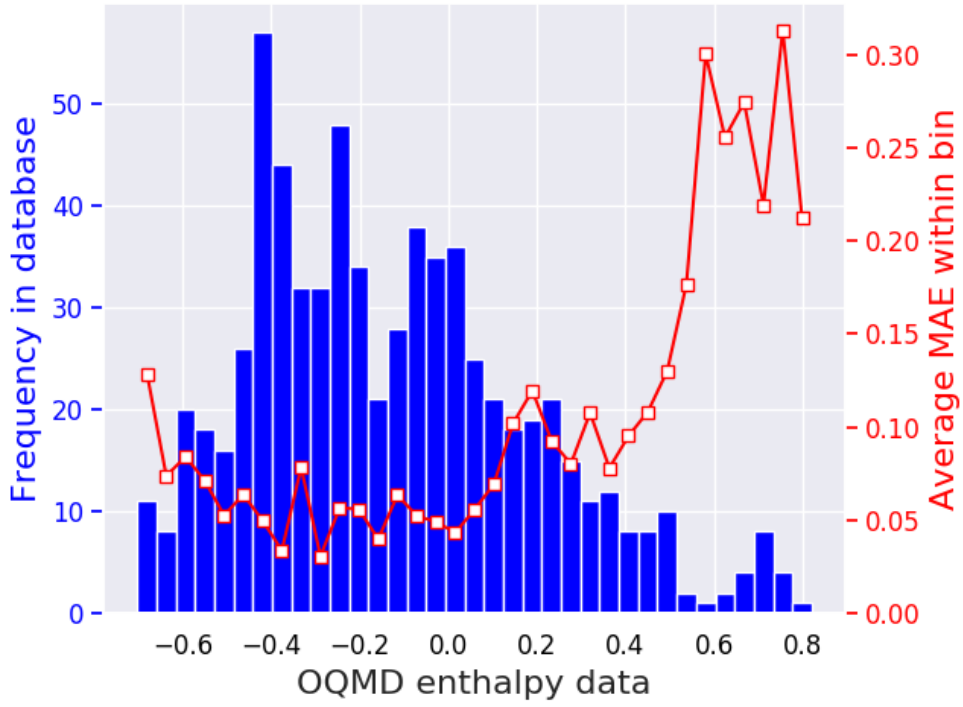


Figure 3.8: Histogram of formation enthalpy distribution for the remaining hydride data following the second filtering stage, overlaid with MAE for data in each bin per the GBR model represented in Figure 3.7.

a suitable ability to generalise to the bulk of the samples, yet still shows poor predictive performance for sparsely sampled enthalpy values.

Doing so consolidates the composition space used in the construction of the model to approximately ± 0.75 eV/atom, as per the histogram in Figure 3.8. However, the error for higher positive value ground truth enthalpy results is still comparatively large. Referring to the ultimate objective of this model, that being to predict formation enthalpy for stable metal hydrides, the positive values corresponding to unstable compounds are potentially superfluous to the model. A second cycle of the iterative GBR filtering process was conducted, using the same method and parameterisation as the previous step.

The error thresholds for the average MAE required for removal of a data point as well as the termination condition, were both reduced to a value of 0.2 eV/atom. The resultant filtered dataset, to be used going forward in the construction of the final model, consists of 623 compositions and admits a GBR model with a MAE of 58.7 meV/atom.

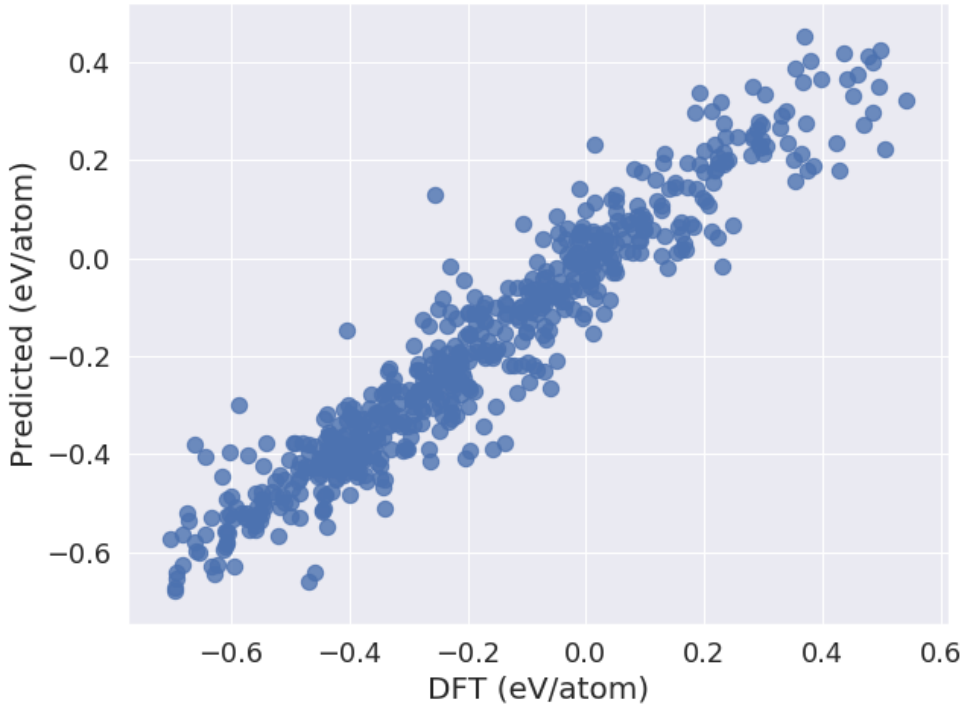


Figure 3.9: Prediction error plot for GBR model constructed using the remaining hydride data following the third filtering stage.

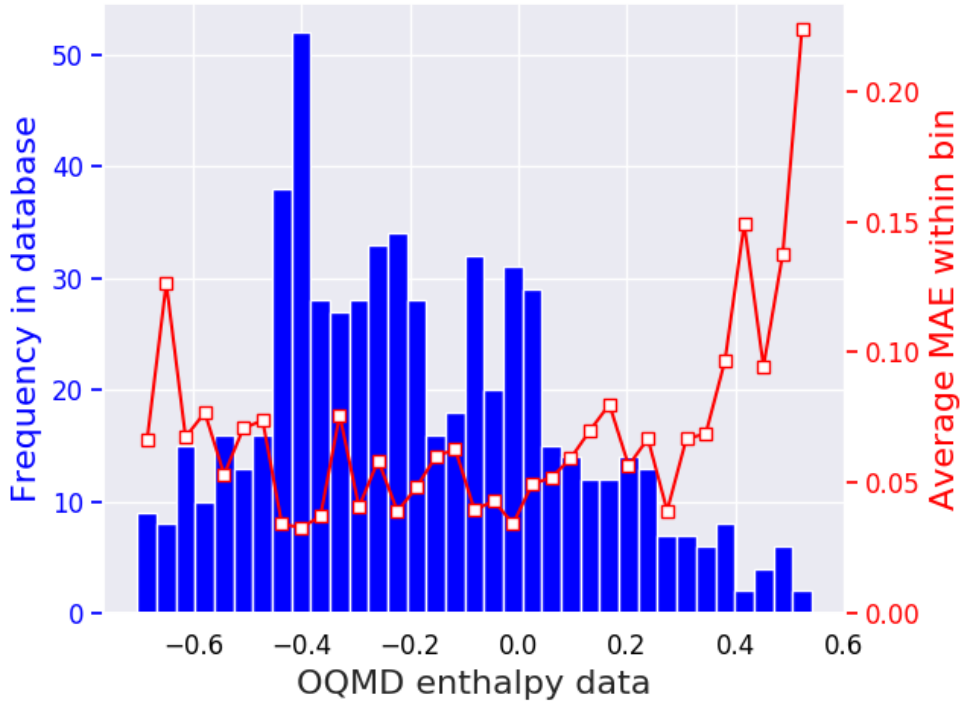


Figure 3.10: Histogram of formation enthalpy distribution for the remaining hydride data following the third filtering stage, overlaid with MAE for data in each bin per the GBR model represented in Figure 3.9.

3.7 Construction of final metal hydride predictive model

Having now consolidated the training data into a dataset well representative of the majority of the data obtained from OQMD, the final model can be constructed. A small sample of data points were taken from the training data and held aside to be used as a final test set to assess performance on unseen data for which there is ground truth data. Considering the already small dataset size, and concerns regarding performance as a result of this (see Section 2.3.1), the withheld test set consisted of six samples; two randomly selected compositions for each binary, ternary, or quaternary hydrides (Table 3.4).

Composition	**OQMD ΔH_f, eV/atom**
LaH_3	-0.573
La_4H_9	-0.073
$AlMgH_5$	-0.364
$CoSr_2H_6$	-0.622
$LaNiMg_2H_7$	-0.483
$LiAlK_2H_6$	-0.275

Table 3.4: Held-aside test compositions and corresponding ground truth enthalpy values, to seven decimal places.

As the models trained during the data cleaning process all used the same predefined hyperparameters, it was possible to further improve the predictive error admitted by the final pruning stage by reconsidering hyperparameter generation. Utilising the final cleaned dataset, a GBR model was built for the dataset without the held-aside validation data points, with newly generated optimised hyperparameters using scikit-optimize (see Table 3.5).

n_estimators	1000
max_depth	90
learning_rate	0.097294
max_features	56
min_samples_split	2
min_samples_leaf	25
loss	'lad'

Table 3.5: Hyperparameter values used in the GBR model constructed without held-aside validation data.

A 5-fold cross-validation was performed, as implemented in scikit-learn, resulting in a predictive model, with a MAE of 61.5 meV/atom (see Figures 3.11 & 3.12). This model will have limited use, implemented only in prediction of enthalpy values for these select validation compositions.

Composition	**OQMD ΔH_f, eV/atom**	**Predicted ΔH_f, eV/atom**	**Absolute Error, eV/atom**
LaH_3	-0.579	-0.573	0.006
La_4H_9	-0.616	-0.622	0.006
$AlMgH_5$	-0.058	-0.073	0.015
$CoSr_2H_6$	-0.475	-0.483	0.008
$LaMg_2NiH_7$	-0.333	-0.364	0.031
AlK_2LiH_6	-0.255	-0.275	0.020

Table 3.6: Compositions held out from training dataset used to validate the trained model, to seven decimal places.

As can be seen in Table 3.6, the predictive ability for these unseen data points is satisfactory. The largest absolute error value of 31 meV/atom, corresponding to the composition $LaMg_2NiH_7$, is less than the mean absolute error of 61.5 meV/atom admitted through the model construction. With this confidence in the model to predict enthalpy data similar to the DFT-based values sourced from OQMD, the model can be used further for predicting results for novel materials, previously unseen to the model.

As such, a final GBR model can be built using the full filtered dataset along

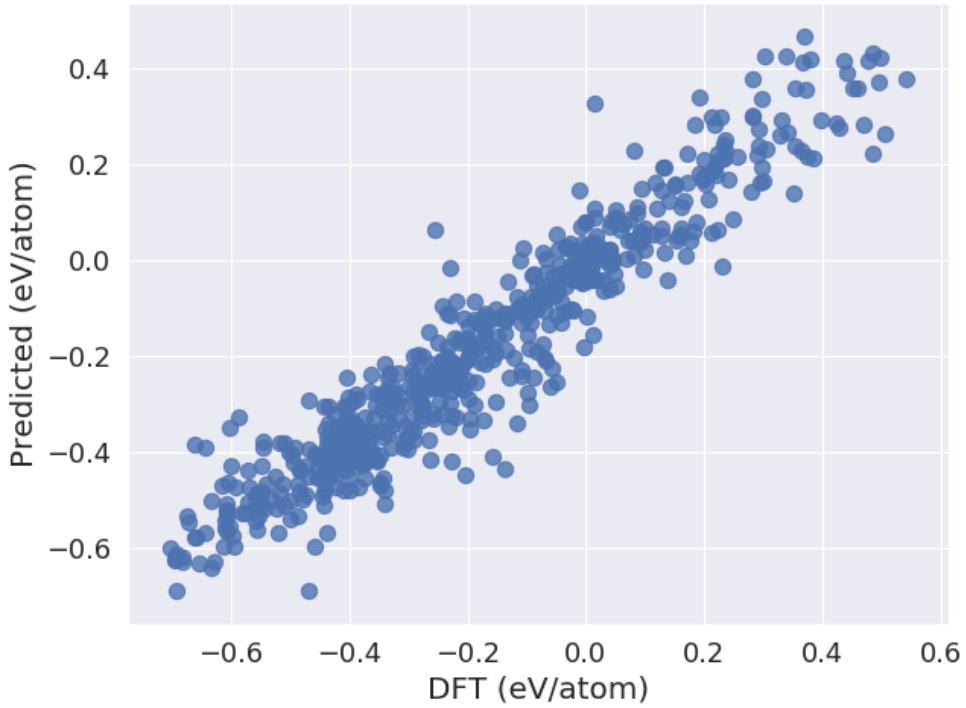


Figure 3.11: Prediction error plot for GBR model constructed using the final hydride dataset minus the held-aside validation set.

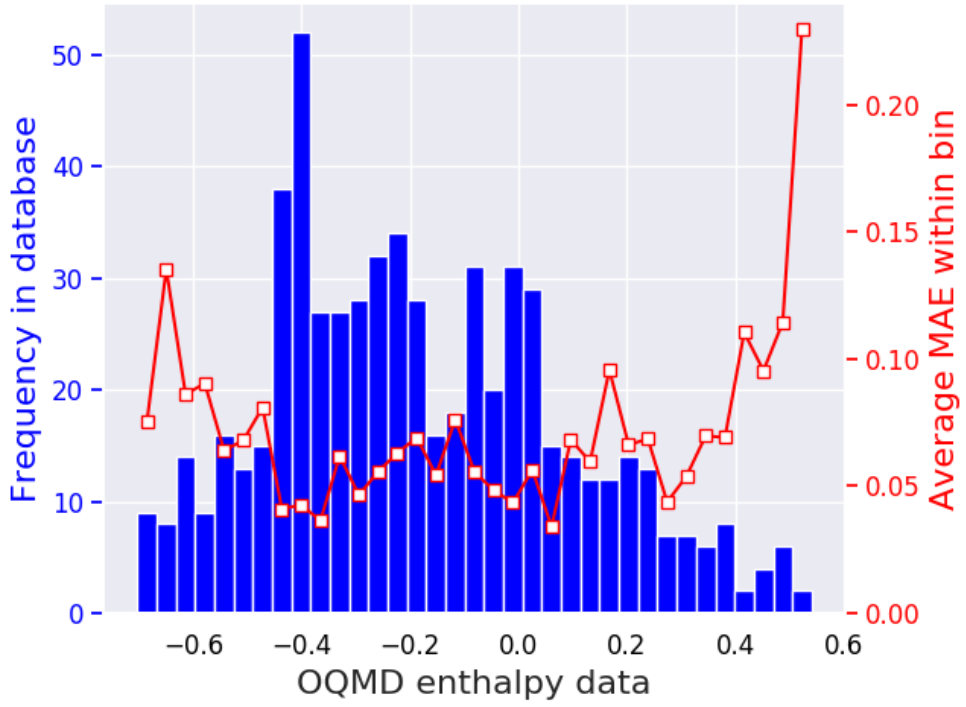


Figure 3.12: Histogram of formation enthalpy distribution for the final hydride dataset minus the held-aside validation set, overlaid with MAE for data in each bin per the GBR model represented in Figure 3.11.

with the held-aside data, which will be used in future prediction steps on data unseen to model construction. Admitting an ultimate predictive MAE of 57.9 meV/atom, this model is represented in Figures 3.13 & 3.14.

3.8 Construction of a binary alloy predictive model

3.8.1 Motivation

In Section 3.4.1, equations 3.3 & 3.4 outline the formalism used in the OQMD scheme for calculation of formation enthalpy values for a given species. From a bank of previously-calculated elemental ground-state energies, the convention is to follow a reaction pathway consisting of elemental crystals combining to form a species that is the sum of its parts. However this reaction mechanism, assumed *a priori*, is less realistic for the formation of metal hydrides. Whilst, by definition, appropriate for binary hydride compounds, empirical studies show alternative reactions to be more prevalent - for example; combining multiple hydride species, destabilising an initial hydride species by combining with another metal, or hydriding an alloy [75, 76].

A decision was made to investigate the more commonly occurring case of interstitial hydrides. To facilitate investigation of this reaction mechanism based on hydriding an alloy, more information is required regarding the initial intermetallic species. In particular, the stability of such an alloy is an important factor as to the ability to repeatedly hydrogenate and dehydrogenate. In order to check this, a secondary predictive model was constructed in a very similar manner to the hydride model, but this time

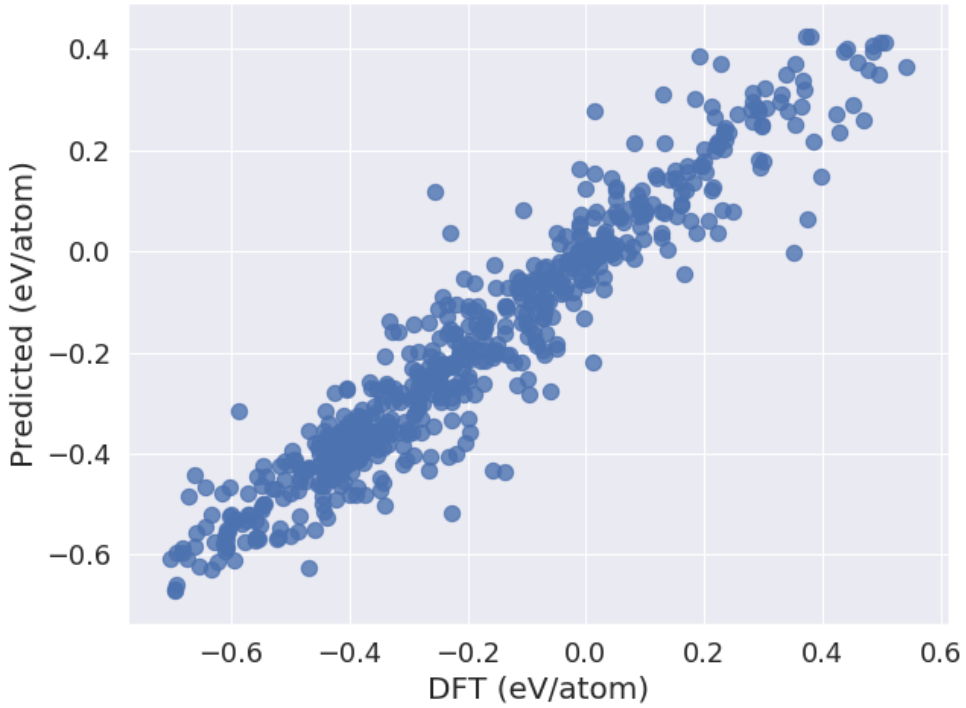


Figure 3.13: Prediction error plot for the final GBR model for hydride formation enthalpy prediction.

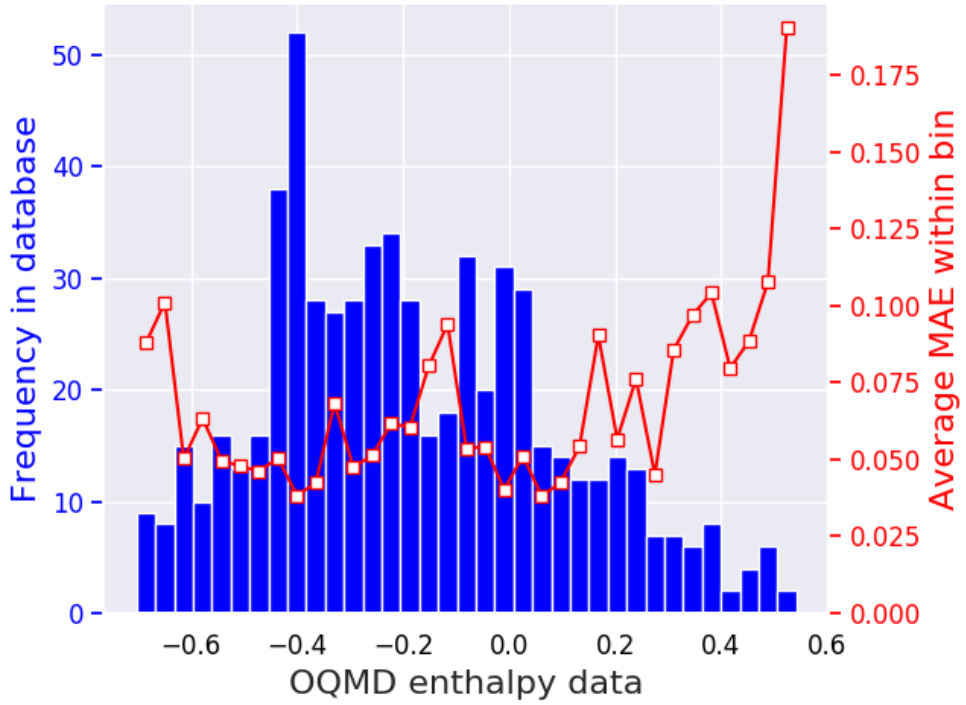


Figure 3.14: Histogram of formation enthalpy distribution for the final GBR model for hydride formation enthalpy prediction, overlaid with MAE for data in each bin per the GBR model represented in Figure 3.13.

using alloy data.

Considering the objective of this investigation is to search for novel ternary hydride compounds, the model will be used to predict enthalpy of formation for the initial binary alloys. To be useful as a cyclable hydrogen store, both the base alloy and the resultant hydrogenated species must be stable.

3.8.2 Data cleaning and model construction

Initial results were garnered from OQMD v1.3.0 in the same manner as outlined in Section 3.4.2 - taking the lowest formation enthalpy value for each binary (A_xB_y), ternary ($A_xB_yC_z$), and quaternary ($A_xB_yC_zD_q$) alloys, for any metals A, B, C, D. The availability of information for each such class of composition is shown in Table 3.7.

Binary	Ternary	Quaternary	Total
11,769	191,136	25,397	228,302

Table 3.7: Availability of results for each class of alloy composition collected from OQMD v1.3.0.

It is clear to see that the range of alloy compositions is a vast increase compared to the data handled when constructing the hydride predictive model. As alluded to in Sections 2.3.3 & 2.3.1, a larger unbiased training data set is usually beneficial to a model's performance. Being introduced to a much wider sample of a given population can greatly improve the ability to generalise to unseen data. Nonetheless, the dataset still requires some degree of pruning to remove any outliers and ensure it is still adequately representative of alloy results.

First, a rudimentary filter is created to remove any compositions with wildly

skewed stoichiometry. Removing data points for which

$$\frac{\text{maximum molar ratio}}{\text{total number of atoms}} \geq 0.9 \tag{3.5}$$

removed 212 outliers, leaving 228,090 alloy compounds. Considering the size of the training set and the iterative nature of the GBR filtering process used twice in Section 3.6, only the second stage of this method is implemented, with the MAE threshold set to 0.2 eV/atom. A GBR model was trained on the remaining data, as implemented in scikit-learn, and with hyperparameters optimised by use of scikit-optimize, given in Table 3.8. As can be seen with the prediction error results in Figures 3.15 & 3.16, predictive ability is impacted by a non-trivial number of extremely high ground truth values. Whether these data points are incongruous calculation values or not, they must be addressed in order to make the model practicable.

n_estimators	1000
max_depth	38
learning_rate	0.048372
max_features	32
min_samples_split	15
min_samples_leaf	33
loss	'lad'

Table 3.8: Hyperparameter values used in the GBR model constructed without held-aside validation data.

This results in a final pruned dataset of 216,255 alloy compositions which was used to construct a final GBR model with a MAE of 38.8 meV/atom (Figures 3.17 & 3.18).

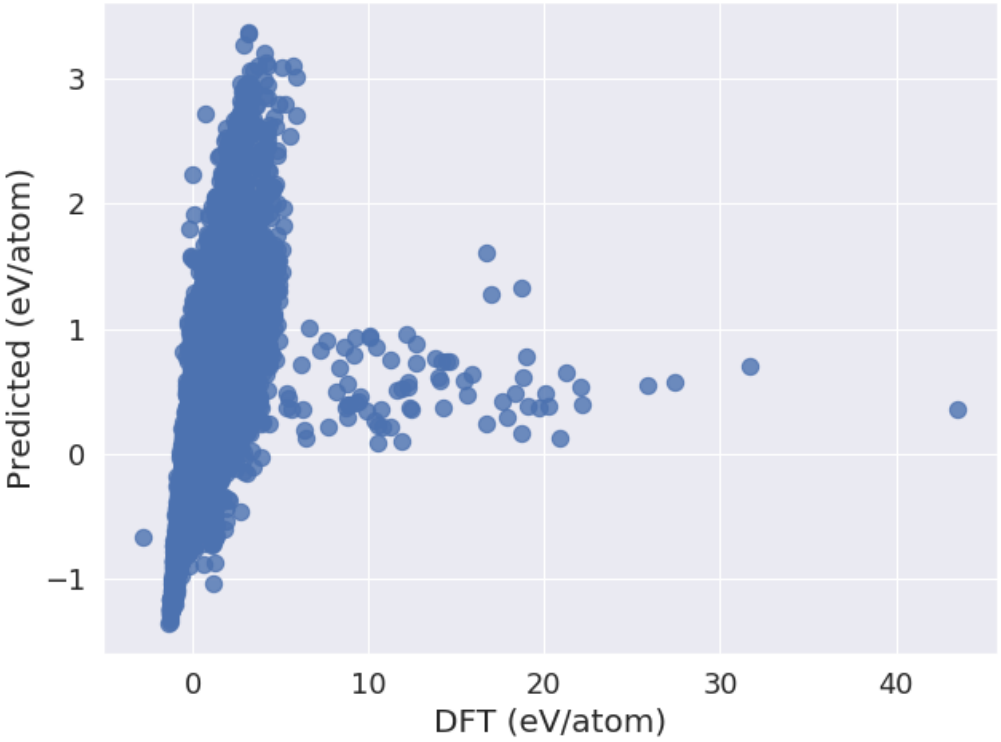


Figure 3.15: Prediction error plot for GBR model constructed using the base alloy data obtained from OQMD.

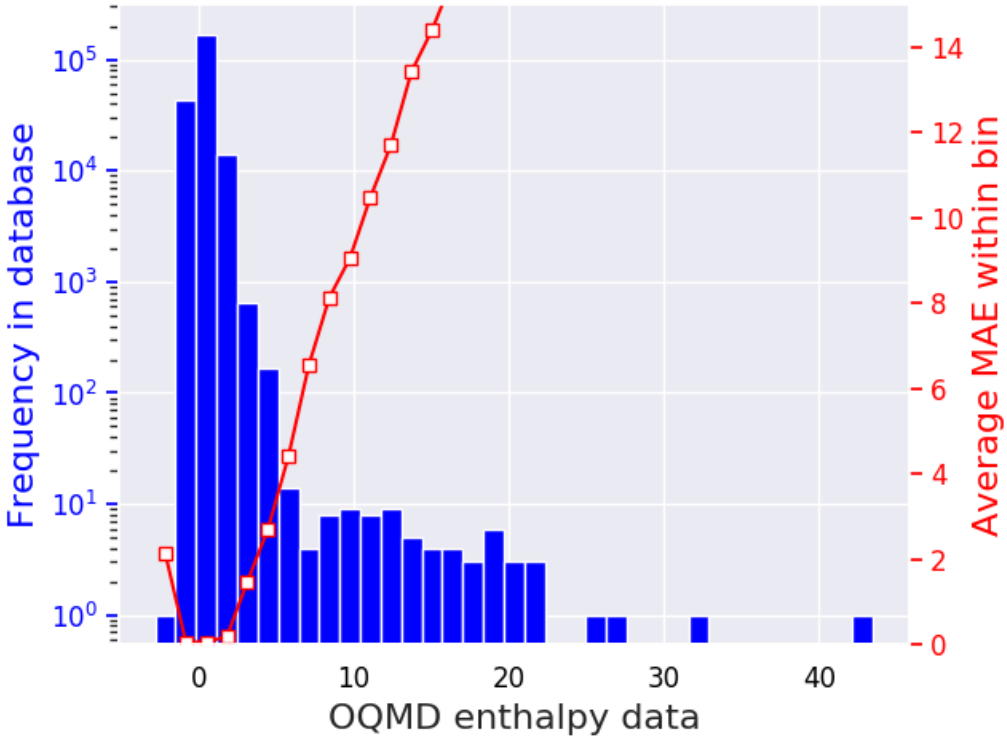


Figure 3.16: Histogram of formation enthalpy distribution for base alloy data obtained from OQMD, overlaid with MAE for data in each bin per the GBR model represented in Figure 3.15.

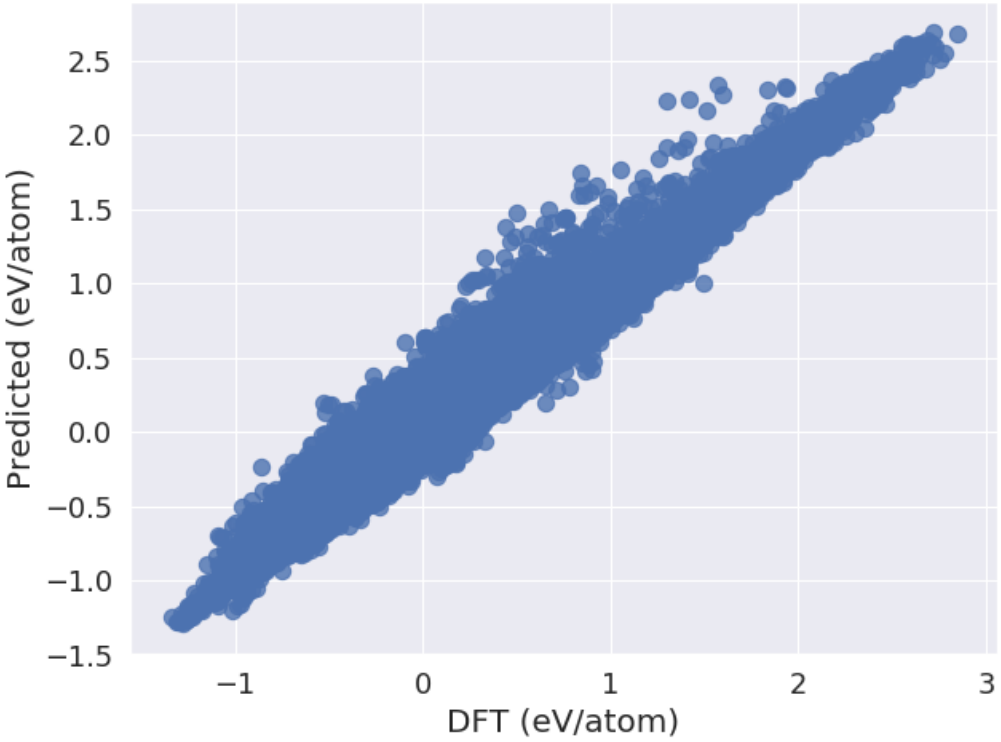


Figure 3.17: Prediction error plot for the final GBR model for alloy formation enthalpy prediction.

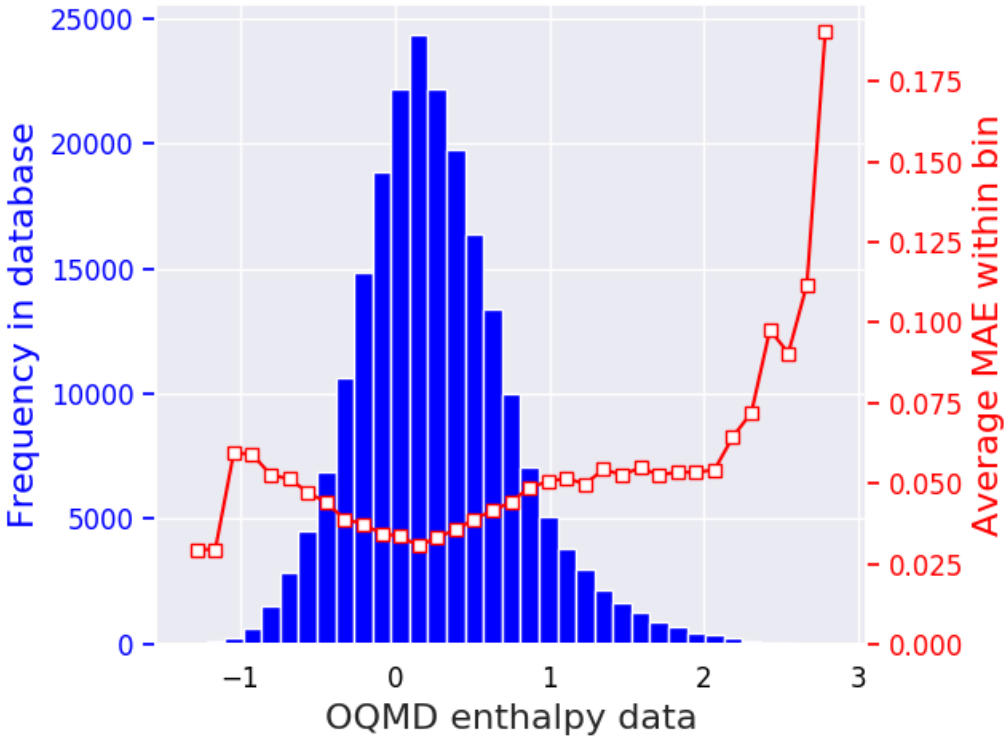


Figure 3.18: Histogram of formation enthalpy distribution for the final GBR model for alloy formation enthalpy prediction, overlaid with MAE for data in each bin per the GBR model represented in Figure 3.17.

3.9 Discussion

3.9.1 Choice of training data source

As alluded to in Section 3.4.1, the entries of OQMD are approximately 10% structures from ICSD, whilst the remainder consists of generated structures constructed by analogy, based upon prototypical structures obtained from ICSD. The sample space from this database is a medley of structures obtained from experimental reports - either fully characterised with specified coordinates, or with a known structure type such that coordinates and parameters can be inferred - as well as theoretical structures, collated from journal publications. These entries are then iterated upon according to certain chemistry rules to produce the remaining samples.

An appealing consequence of this is that OQMD offers a wider selection of structural entries than a solely experimental database would. As seen in the work reported by Witman et al. [57], the mostly-experimental HydPARK database admitted 570 hydride compositions with values for ΔH_f prior to data cleaning methods, whilst the approach taken in this work collected 846 compounds from OQMD. In addition, reported empirical results are dependent on the experimental processes used for synthesising these materials. Inherent inconsistency in methodology between entries introduces some intrinsic error within this collated data. Whilst experimental errors may be encoded in ICSD structures that are in turn included in OQMD, the standardised approach to calculations taken may work to mitigate such an error, with consistent DFT methods applied universally. Considering the dependence on the quantity and quality of training data for most ML methods (see Sections 2.3.1 & 2.3.2), this choice of database was intuited to be useful for this investigation.

Descriptor data was chosen to be solely composition dependent to investigate whether a model could be trained over data for a wide range of hydride material classes and geometries, with the aim of associating hydride chemistry with intrinsic thermodynamical properties, whilst providing chemical insight *via* the fitting process. Preliminary work used handmade archives of textbook data, but Magpie was soon settled on for ease of use due to the implementation in the Matminer *Featurizer* package [77].

3.9.2 Algorithm selection

Performance of a selection of machine learning algorithms over a range of train/test splits is presented in Section 3.5. The selection includes multiple ensemble-based methods which have been shown to be useful in regression of DFT-calculated energies and properties from geometry-only and composition-only descriptors [78, 79], as well as linear regression methods that have shown success [80]. Such a range of methods have been used for validation of algorithm performance, as in the work by Faber at al. [81].

It can be seen that the ensemble methods generally performed better, as quantified by the average MAE result across the five iterations of each train/test split for both training and testing. ETR and RFR perform similarly for both error values but GBR is shown to have a slightly lower test error across the board, as well as a much better fit during the cross-validation process during construction. Respective testing and training errors averaged 0.1956 eV/atom and 0.0396 eV/atom across the test splits examined, compared to the next best in class of 0.2029 eV/atom and 0.0632 eV/atom for ETR. Gradient boosting regression has been shown to be effective for the study of material properties, and has demonstrated

provable success in material discovery [82, 83]. These results and those from literature can be used to justify the choice to use gradient boosting regression as the machine learning algorithm of choice in this work.

It may be noted that the algorithm selection process and the data cleaning stage are presented in a different order to that in Figure 3.1. This is due to the test of algorithm choice having been previously performed using an older version of OQMD (v.1.2.0) which also justified the use of GBR, as was consequently used in the filtering processes to clean results from the updated database.

3.9.3 Feature importance

The use of a bagging-based algorithm allows for the determination of relative importance of features in the fitting function of a constructed ML model *via* the out-of-bag error (see Section 2.2.5). Relative importance of variables in the final hydride model were determined and the 20 highest values are presented in Figure 3.19. Due to the ever-present nature of hydrogen in all compositions by definition, statistical elemental features may be seen to describe the metal species involved. For example, hydrogen has a high electronegativity value and Mendeleev number, as well as the small atomic and covalent radii. Considering periodic trends, the statistical concepts of mean, minimum, range etc. of certain features can be interpreted as indicative of these properties in the intermetallic components.

The two most significant results are shown to be the minimum electronegativity value and the average deviation from the mean for the number of unfilled s valence electrons among elements, followed by

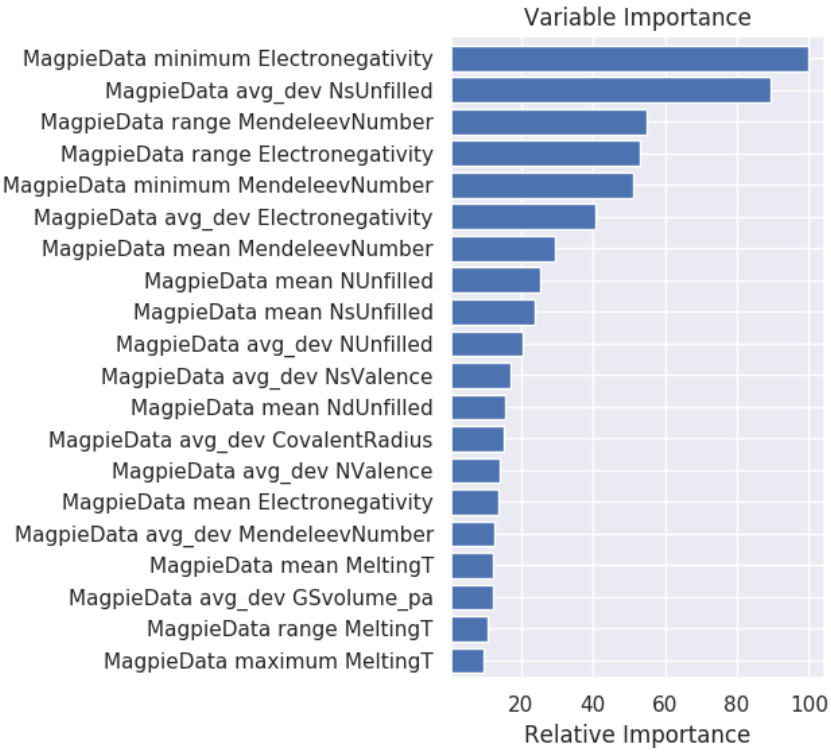


Figure 3.19: The top 20 results of variable importance for the final metal hydride model.

a selection of statistical variations on electronegativity and Mendeleev number.

Electronegativity is known to play a role in the thermodynamic stability of binary hydride species such that elements with low Pauling electronegativities can form very stable ionic hydrides (e.g. lithium) whilst those with electronegativity values closer to hydrogen can form stable covalent hydrides (e.g. carbon) [84]. This property is also correlated to atomic and ionic radii in metals, generally having an inverse relationship.

Deviation from the mean value of unfilled s valence electrons is a more nuanced variable. As the s-block is composed of the Group I and II elements, along with hydrogen and helium, one can deduce that such deviation refers to these certain metals. Hydride compounds of Group I metals tend to form saline, or salt-like, hydrides with ionic character that

increases down the group, whilst Group II metal hydrides can take the form of electron deficient covalent compounds, or ionic structures. It is possible that these distinct differences from simple interstitial hydride forms may be implicitly encoded within a Magpie descriptor.

Mendeleev number is defined as a simple enumeration process, traversing down each group of the periodic table in numerical order consecutively, with each element assigned an incremental number [85]. This metric can introduce trends as a function of group number. The descriptor of Mendeleev number range can impart information as to types of component metallic elements, from a coarse level of mostly alkali, alkaline earth, or transition metals, to a finer degree of traversal across transition metal, whilst also capturing information regarding trends down a given group, in terms of mass, volume, and various radii variables.

3.9.4 Alloy model

Whilst the main goal of this work was to develop a predictive model with regards to metal hydrides, the construction of a similar model trained on metal alloy data could allow for further interpretation of possible reaction pathways for hydrogenation or dehydrogenation. Alloy data was much more widely available from OQMD, and used to train a model that admitted a lower predictive error to the hydride model. With such abundant training data, the hyperparameters generated are tuned to mitigate overfitting, broadening the tree constraints of minimum samples per split and leaf, whilst reducing tree depth and learning rate to lower the chance of overtly biasing the fit.

During the generation of new hydride compositions, this model would

primarily be used qualitatively to predict whether a metal alloy is stable or not. With this considered, a test set and further validation was not performed for this model, whilst the cross-validation process was presumed to be sufficient in parameterising the model to our needs.

3.9.5 Qualitative prediction

The ML models generated here can be used to predict the formation enthalpy values for a range of compositions previously unseen to the model. It should be emphasised that the intention of this tool is to provide an initial means of screening a large combinatorial composition space. By first assessing samples independent of structure in a computationally efficient manner, such an approach would allow for focused investigation of highlighted candidate materials at higher levels of theory.

3.10 Conclusion

In this chapter, methodology for constructing a machine learning model and a process for comparative testing of performance for a range of algorithms is outlined. Of these, nonlinear ensemble methods were shown to perform best on average, and this was used to justify the selection of gradient boosting regression for the construction of models in this work.

The workflow of data acquisition and cleaning of this dataset was defined, filtering outlier data by implementing a threshold error for prediction, ensuring that the data retained is well represented by the model. This is performed for data regarding both metal hydrides and metal alloys, before the pruned data is used in the development of final production models,

shown to have mean absolute errors of 57.9 meV/atom and 38.8 meV/atom respectively through the cross-validation stage of model construction.

A separate hydride predictive model was built by excluding a selection of compositions sampled across the space of chemical complexity involved with the data, to allow for further validation of prediction using such a held-out test dataset. Initial comparison between prediction and ground truth data suggests good predictive ability to results at the standard of the database, but corroboration shall be tested further by use of first principles calculations.

Analysis of relative feature importance offers insight into the weighting of variables in the fitting process, which appears to encapsulate factors known to influence thermodynamical properties of metal hydrides such as the electronegativity of the intermetallic component and periodic table group trends relating to mass and radii.

Chapter 4

Density functional theory validation of machine learning predicted hydride systems

4.1 Introduction

Having constructed machine learning models for the prediction of formation enthalpy for both metal hydride compositions and intermetallic alloys, this chapter will cover validation of predictions by means of first principles calculations. Initially, calculations for the held-out validation dataset and known experimental storage systems will be used to determine predictive performance. Following this, new ternary compositions will be generated and subjected to a range of filtering stages to reduce the sample space. For select results, stable structures will be determined by crystal structure prediction methods and first principles calculations used to determine enthalpy of formation, from which predictions and calculated values can be compared.

4.2 DFT calculation process

Further use of the ML models initially requires proof of validation of prediction when compared to known theoretical results, after which they can be justified in use as a shortcut to study further results. In order to study these materials we use first principles calculations and the method of geometry optimisation to calculate ground state energies for both the ultimate metal hydride materials as well as any constituent species.

To ensure consistency with the OQMD regime, any comparison to their data will be calculated by means of constituent elemental ground state energies, as per Equation 3.4. The ground state energy for each simulation cell is then reduced to an energy per atom, and the formation enthalpy can be calculated as the energy difference between products and reactants, normalised per atom. For a compound $A_\alpha B_\beta H_\gamma$:

$$\Delta H_f = \frac{E\left(A_\alpha B_\beta H_\gamma\right) - \left[\alpha E\left(A\right) + \beta E\left(B\right) + \frac{\gamma}{2} E\left(H_2\right)\right]}{\alpha + \beta + \gamma}, \qquad (4.1)$$

where $E(X)$ is the DFT energy of a species X in a minimised energy structure.

4.2.1 DFT settings

Calculations are performed using the Vienna Ab initio Simulation Package (VASP), using a plane-wave basis set with an energy cutoff of 400 eV. Using a generalised gradient approximation (GGA) of exchange and correlation through the use of the Perdew-Burke-Ernzerhof (PBE) functional [36], the projector augmented wave method is used to solve the Kohn-Sham equations [37, 86]. For elemental metals or intermetallic alloys, a first-

order Methfessel-Paxton smearing scheme [87] is employed with a width of 0.2 eV, otherwise a Gaussian smearing process is used with width 0.05 eV. The energy convergence threshold to break the electronic self-consistency loop is set to 10^{-4} eV, whilst the break condition for the ionic relaxation loop is set to -5×10^{-2} eV/Å. The values for the k-point mesh are defined with a spacing of approximately 0.20 $Å^{-1}$.

These geometry optimisations were performed on the ARCHER2 UK National Supercomputing Service using the implemented standard VASP 5 software, assigned to a single node on the standard partition [88].

4.3 Validating model using theoretical calculations

In Section 3.7, a validation method was used to compare the predicted enthalpy from the model to that of the ground truth data collected from the OQMD database. Whilst suggesting relatively good alignment, further validation is required to suggest that the model works sufficiently as to predict results from first principles calculations.

By calculating the formation enthalpy for the compositions in Table 3.4 as per the procedure used for OQMD data generation, we may gain some insight into the effectiveness of the predictive model. The final hydride structure is directly taken from the database to facilitate energy calculations, and reference elemental DFT energies are also determined for each elemental species involved (see Table 4.1).

Elemental crystals are subjected to geometry optimisation and DFT energies normalised per atom, along with the final hydride structures.

Compound	Atoms in simulation cell	Total energy, eV	Energy per atom, eV/atom
H_2	2	-6.759	-3.379
Al	4	-16.337	-4.084
La	4	-21.178	-5.294
Ni	4	-23.506	-5.876
Co	4	-28.717	-7.179
K	4	-4.057	-1.014
Sr	4	-7.306	-1.827
Li	2	-4.136	-2.068
Mg	2	-3.577	-1.789
AlK_2LiH_6	20	-64.336	-3.217
$AlMgH_5$	28	-94.205	-3.365
$CoSr_2H_6$	36	-145.159	-4.032
LaH_3	16	-72.168	-4.511
La_4H_9	28	-120.920	-4.319
$LaMg_2NiH_7$	88	-345.031	-3.921

Table 4.1: DFT energies, to three decimal places.

Hydrogen is simulated as a dimer in a 10 $Å^3$ simulation box. Using Equation 4.1, some examples are shown in Equations 4.2-4.7 and presented in Table 4.2.

$$La + 1.5H_2 \rightarrow LaH_3 \qquad \Delta H_f = -0.653 \text{ eV/atom} \quad (4.2)$$

$$4La + 4.5H_2 \rightarrow La_4H_9 \qquad \Delta H_f = -0.634 \text{ eV/atom} \quad (4.3)$$

$$Al + Mg + 2.5H_2 \rightarrow AlMgH_5 \qquad \Delta H_f = -0.112 \text{ eV/atom} \quad (4.4)$$

$$Co + 2Sr + 3H_2 \rightarrow CoSr_2H_6 \qquad \Delta H_f = -0.576 \text{ eV/atom} \quad (4.5)$$

$$La + 2Mg + Ni + 3.5H_2 \rightarrow LaMg_2NiH_7 \qquad \Delta H_f = -0.430 \text{ eV/atom} \quad (4.6)$$

$$Al + 2K + Li + 3H_2 \rightarrow AlK_2LiH_6 \qquad \Delta H_f = -0.371 \text{ eV/atom} \quad (4.7)$$

These calculated enthalpy values are shown in Table 4.2, alongside the ground truth values from the source data, and the predicted value from the model trained on the finalised dataset, but excluding the withheld validation data points.

Ternary hydride	**Predicted ΔH_f, eV/atom**	**Calculated ΔH_f, eV/atom**	**Database ΔH_f, eV/atom**
LaH_3	-0.579	-0.652	-0.573
La_4H_9	-0.616	-0.633	-0.622
$AlMgH_5$	-0.058	-0.112	-0.073
$CoSr_2H_6$	-0.475	-0.576	-0.483
$LaMg_2NiH_7$	-0.333	-0.430	-0.364
AlK_2LiH_6	-0.255	-0.371	-0.275

Table 4.2: Comparison of enthalpy values for the hydride compositions given in Table 3.4.

For comparative purposes, data for known hydrogen storage materials was gathered from the *HydPARK* database, a repository of experimental metal hydride information and empirical results constructed through collaboration between the International Energy Agency (IEA) and the U.S. Department of Energy (DOE) [58].

Composition	**Frequency**	**Lowest enthalpy ternary hydride**
Mg_2Ni	36	Mg_2NiH_4
$NaAlH_4$	18	$NaAlH_4$
$LaNi_5$	18	$LaNi_5H_7$
Na_3AlH_6	11	Na_3AlH_6
$ZrCr_2$	11	$ZrCr_2H_3$

Table 4.3: Most frequently occurring 'Composition' samples obtained from the HydPARK database. Accessed 21-October-2020.

Whilst not possible to directly compare thermodynamic results calculated by first principles to those obtained by experiment (see Section 3.9.5), it is worthwhile to consider the predictive ability of the model versus both the OQMD value for formation enthalpy as well as that directly

obtained from *ab initio* calculations. To do this, a rudimentary search of the HydPARK database was performed to highlight the most commonly studied compositions. Taking the five lowest enthalpy ternary hydride compositions of these samples as per OQMD entries (see Table 4.3), first principles calculations are used to determine elemental ground state energies for all constituent elements as well as the final species, using structure files obtained from the ICSD.

It is worth noting that the only Zr-Cr-H species in the OQMD database, $ZrCr_2H_3$, is not presented as stable, having a positive theoretical formation enthalpy of 0.13 eV/atom. Results of Zr-Cr hydrides in the HydPARK database suggest hydrogen loading of H/M=1.8-2.1. This will be expanded upon in the discussion section later in this chapter.

Compound	**Atoms in simulation cell**	**Total energy, eV**	**Energy per atom, eV/atom**
H_2	2	-6.759	-3.379
Mg	2	-3.577	-1.789
Al	4	-16.337	-4.084
Cr	2	-19.878	-9.939
La	4	-21.178	-5.294
Na	2	-2.919	-1.459
Ni	4	-23.506	-5.876
Zr	2	-18.091	-9.045
$LaNi_5H_7$	26	-124.271	-4.780
Mg_2NiH_4	36	-123.727	-3.437
$NaAlH_4$	24	-80.184	-3.341
Na_3AlH_6	20	-62.255	-3.113
$ZrCr_2H_3$	24	-153.646	-6.402

Table 4.4: DFT energies, to three decimal places.

$$2Mg + Ni + 2H_2 \rightarrow Mg_2NiH_4 \qquad \Delta H_f = -0.294 \text{ eV/atom} \qquad (4.8)$$

$$Na + Al + 2H_2 \rightarrow NaAlH_4 \qquad \Delta H_f = -0.164 \text{ eV/atom} \qquad (4.9)$$

$$La + 5Ni + 3.5H_2 \rightarrow LaNi_5H_7 \qquad \Delta H_f = -0.293 \text{ eV/atom} \qquad (4.10)$$

$$3Na + Al + 3H_2 \rightarrow Na_3AlH_6 \qquad \Delta H_f = -0.239 \text{ eV/atom} \qquad (4.11)$$

$$Zr + 2Cr + 1.5H_2 \rightarrow ZrCr_2H_3 \qquad \Delta H_f = 0.108 \text{ eV/atom} \qquad (4.12)$$

Energies for the relevant species are presented in Table 4.4, which are then used in Equations 4.8-4.12 to determine formation enthalpies for these hydride compositions. Enthalpy data is presented for comparison in Table 4.5, with predictions computed by use of the finalised model containing all cleaned data.

Ternary hydride	**Predicted ΔH_f, eV/atom**	**Calculated ΔH_f, eV/atom**	**Database ΔH_f, eV/atom**
Mg_2NiH_4	-0.234	-0.294	-0.249
$NaAlH_4$	-0.149	-0.164	-0.149
$LaNi_5H_7$	-0.203	-0.293	-0.235
Na_3AlH_6	-0.230	-0.239	-0.206
$ZrCr_2H_3$	-0.217	0.108	0.130

Table 4.5: Comparison of enthalpy values for hydride compositions in Table 4.3.

4.4 Crystal structure prediction

Up until this point, first principles calculations have been performed using known crystal structures. Utilisating the ICSD catalogue of results, as well as OQMD's prototypical structures, that are ultimately derived from them, the starting structures from which to study these systems are readily

available. However, upon the generation of novel compositions without any prior structural insight, an alternative predictive methodology is required.

Following the methodology outlined in Section 3.4, this machine learning model was intentionally designed to be independent of explicit crystal structure information. In doing so, the model is trained to interpret the underlying chemistry encoded in the composition representation of a compound. Nonetheless, in order to verify and utilise the capability of this predictive tool, a conversion from mere chemical formulae to a crystal structure suitable for initiating geometry optimisation calculations is required.

4.4.1 CALYPSO

One process uses the CALYPSO method, as described in Section 2.7.2; a means of predicting energetically stable or metastable crystal structures for a target composition by directly interfacing with the VASP code. The global optimisation process works to find many structures across a potential energy surface. By providing a selection of INCAR files, consecutive geometry optimisation calculations are performed on found structures, gradually increasing in precision, until converged results are obtained from the finest of these processes.

In addition to these INCAR_* files (one for each optimisation stage, where * indexes the level of refinement), an appropriate POTCAR file to enable these calculations, and the executable file calypso.x, this process is parameterised using the file input.dat. Here the system is introduced, constraints such as minimal distances between atoms of each species are imposed, and settings for the PSO algorithm and the VASP calculations

can be adjusted.

An iterative process, this method involves a global or local optimisation algorithm with a large number of steps, as well as multiple geometry optimisation calculations for structures found along the way, resulting in a non-trivial cost of computational resources and time.

CALYPSO settings

Calculations are parameterised such that one formula unit is present per simulation cell. Minimal distances between each chemical species is set to be 0.1 Å less than the minimum distance between the given ions in the respective binary compound (rounded down to one decimal place). This is determined from the crystal structures of such compositions as presented in OQMD. A local PSO algorithm is used, with 60% of structures systematically constructed in this fashion. Twenty iterative steps are used, and each structure is optimised four times, over increasingly refined settings; with the energy convergence threshold monotonously reducing from 3×10^{-2} eV to 1×10^{-4} eV, the threshold for interatomic forces to be considered converged reducing from 4×10^{-2} eV/Å to 1×10^{-2} eV/Å, and a k-point sampling grid spacing of 0.25 $Å^{-1}$ initially used before being reduced to 0.25 $Å^{-1}$ for the final stage. Using the proprietary CALYPSO_ANALYSIS KIT script, results are ranked in terms of energies, and the lowest energy structure is taken.

4.4.2 Tetrahedral atomic structure search algorithm

A more efficient, yet more rudimentary, approach to constructing an initial crystal structure to be further refined is to use a degree of chemical

intuition. In a heuristic approach as developed by colleagues [89], should the crystal structure be known for the intermetallic component of a ternary hydride, the atomic structure of the alloy is explored and favourable hydrogen binding sites identified and systematically occupied.

For all atoms in the alloy, local geometry is determined as a function of atomic covalent radii and electronegativity. Tetrahedral arrangements of atoms are identified from which a central coordinate is identified. Additionally the tetrahedra are ranked by average electronegativity of their constituent atoms. For an AB2 alloy, it is known that hydrogen occupies these tetrahedral holes, energetically favourable for lower electronegativity. In other alloys, hydrogen may occupy octahedral holes instead.

A minimum distance, d_{min}, is defined and enforced between hydrogen instances - usually $d_{min} = 1.4$ Å. This hole-hole distance within the alloy structure can be taken to be shorter than the minimum H-H distance quoted in literature due to the volume expansion that occurs during hydrogenation. Hydrogen atoms are sequentially inserted at the central coordinate of the tetrahedral holes in order of ascending electronegativity, unless a new site falls within d_{min} of an existing H atom, in which case that site is skipped. This process shall henceforth be referred to as a 'tetra search'.

4.5 Generation of ternary compositions

Having constructed a machine learning model validated for the prediction of formation enthalpy values on known data, it can be presented with unseen compositions and offer insight into their thermodynamical properties. From here, materials suggesting promising results can be processed further. The

use of crystal structure prediction can expand upon this simple text-based form of a given composition, offering structural information to facilitate analysis at a higher level of theory.

With the combinatorial explosion of chemical space for each additional element introduced to the system, a decision is made to investigate novel ternary compositions. Given the limitations of the extrapolatory nature of machine learning (see Section 2.3.4), candidate materials are generated as a combination of binary hydrides collected from the OQMD database.

This combination method consists of concatenation and linear combination processes. By selecting the binary hydride composition with the lowest ΔH_f value corresponding to each metal (see Figure 4.1), linear combinations of two of these compounds are constructed with coefficients for each term varying from one to ten:

$$\alpha A_x H_y + \beta B_i H_j = A_{\alpha x} B_{\beta y} H_{\alpha y + \beta j} \qquad 1 \leq \alpha, \beta \leq 10, \tag{4.13}$$

for any two given binary hydrides, $A_x H_y$ and $B_i H_j$, in this set.

For the 67 metals considered, this results in 221,000 compositions. To remove any doubly counted compounds as a result of the linear combination process (for example, $Mg_2Li_2H_4$ as well as $MgLiH_2$), these are then filtered for only unique compositions by taking reduced formulae and extracting unique elements of the set.

4.5.1 Filtering of predicted compositions

Given the context of searching for realistically applicable hydrogen storage materials for on-board applications, remaining results are filtered by their

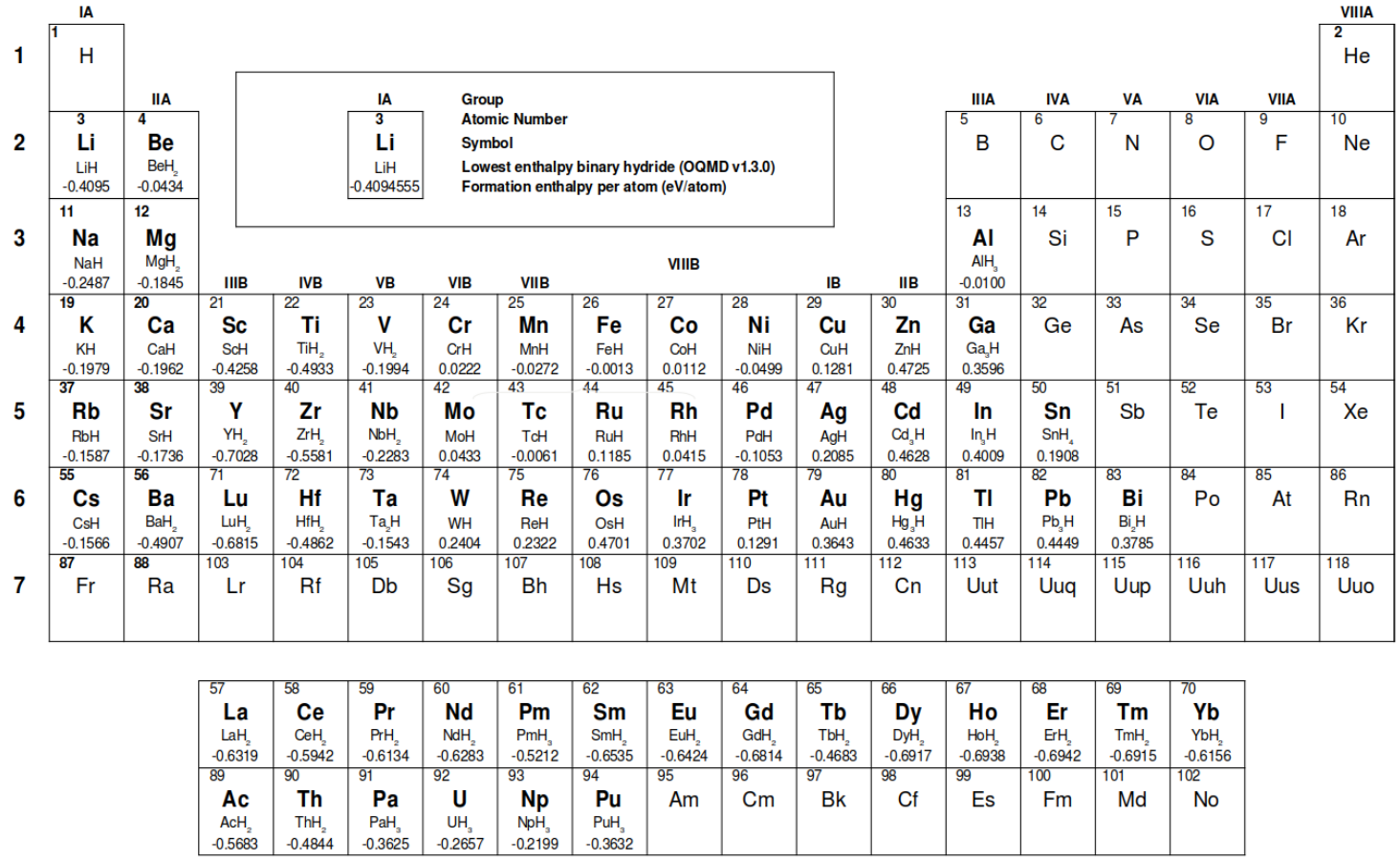


Figure 4.1: Graphical representation of the lowest enthalpy binary hydride composition for each metal, as obtained from the OQMD database.

gravimetric hydrogen content, removing those whose mass is less than 3% hydrogen. This is done by calculating the proportion of the material's total mass for which hydrogen is responsible, as a function of atomic mass for each constituent element, narrowing the field further to 8,440 samples. This lower limit was chosen so as to aim for storage materials with a reasonably high gravimetric capacity whilst also implicitly targeting less heavy, and often less expensive, intermetallic components.

A final constraint implemented is for the formation enthalpy to be defined between -50 and -30 kJ $\text{mol}_{\text{H}_2}^{-1}$, consistent with values suggested for on-board applications [46, 47]. Whilst convention is to regard standard entropy changes for hydrides as constant, intrinsically linked to the entropy of hydrogen gas (see Section 1.5), it has been reported that this may not be the case and that standard entropy changes for alloy-based metal hydrides are in the range of -100 to -150 J K^{-1} $\text{mol}_{\text{H}_2}^{-1}$ [90]. From Equation 1.1, for a system in equilibrium such that $\Delta G = 0$, it can be seen that these two factors directly relate to the temperature required for a plateau pressure of 1bar of H_2:

$$T(1 \text{ bar}) = \frac{\Delta H}{\Delta S}. \tag{4.14}$$

Whilst enthalpy values output by the predictive model are given in eV/atom, this can be readily converted to $\text{kJ/mol}_{\text{H}_2}$:

$$[kJ/mol_{\text{H}_2}] = [eV/atom] * 96.484934... * \frac{\text{total \# of atoms}}{\text{\# of hydrogen atoms / 2}} \tag{4.15}$$

Upon consideration of the range of entropy change values that might apply to the hydride samples in the data set, this chosen enthalpy range corresponds to a highest lower limit of ~27°C and a lowest upper limit of ~60°C, suitable for low-temperature storage solutions. This step admits a final pool of 727 candidate materials.

As alluded to in Section 3.8.1, the reaction pathway based on the combining of elemental crystals is a less realistic approach to hydride synthesis. We now also consider the mechanism of hydriding a metal alloy. A cyclable store of this form would require materials sufficiently stable such that the products for neither the hydrogenation nor dehydrogenation processes would spontaneously decompose. Nonetheless, the underlying alloy chemistry is sought to facilitate the generation of example hydride structures from just a composition representation.

The remaining dataset is processed and the intermetallic species extracted from the ternary hydride compositions. Using the second of the ML models - that trained on alloy data (see Section 3.8) - as well as the hydride model, formation enthalpies are predicted for both the ternary composition, and the alloy form admitted by the omission of the hydrogen component. Taking data entries for which both values are negative left 439 results.

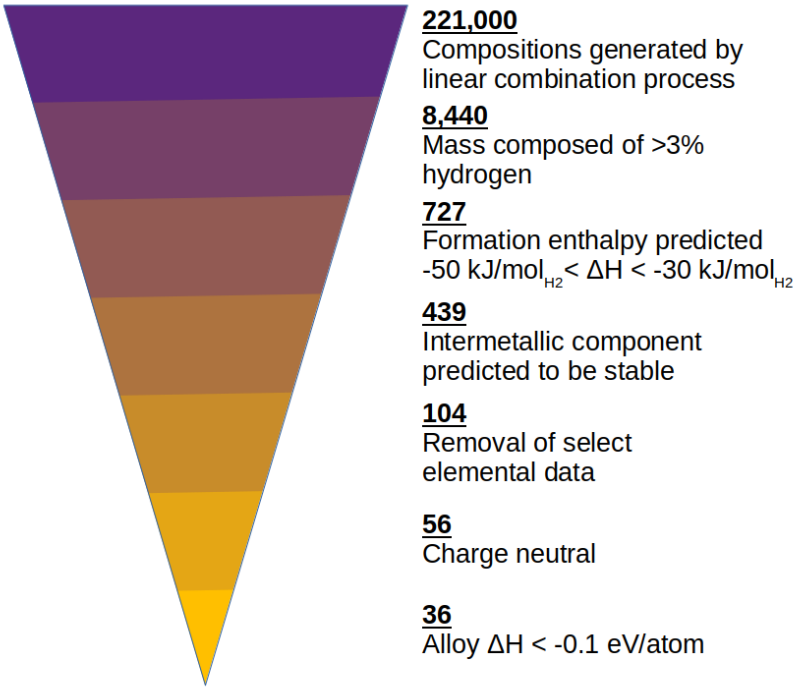


Figure 4.2: Sketch of filtering procedure for generated ternary hydride compositions.

A more involved approach was used for any further filtering of candidate compositions. Considering the remaining materials, constituent elements were assessed as to feasibility and removed as deemed necessary. Whilst all filtering stages up to this point have been rooted in intrinsic chemical or technical performance-based reasoning, certain extrinsic properties should also be considered. Rhodium, platinum, gold, silver, iridium, palladium, ruthenium, technetium, and scandium were all removed due to scarcity and/or wholesale cost which would make implementation as a practical hydrogen store unrealistic. Aluminium-containing compositions are removed on the basis that many ternary complex alanate hydrides are known to form through pathways involving the destabilising of a binary hydride with a second metal species, as opposed to *via* an alloy [91]. Beryllium is excluded due to its toxic effects and the risks involved with exposure [92].

Alloy chemistry	**Frequency**
Mg-Ni	29
Mg-Zn	18
Co-V	16
Ni-V	10
Mo-V	7
Ca-Sn	5
Sn-Ti	4
Mg-Sn	4
Na-Sn	3
Cd-Mg	3
Li-Sn	2
Fe-V	1
Cu-Mg	1
Bi-Mg	1

Table 4.6: Chemistry of the intermetallic component of remaining generated ternary hydrides following explicit element removal.

The outcome of this is 104 results, shared across a range of intermetallic chemistries, as presented in Table 4.6. From here, these compositions can either be taken in their exact form as given, or their alloy chemistry used to inspire a broader investigation into further possible hydride compositions of such a form. Of these results, only 56 are already in a charge neutral composition, split as per Table 4.7, suggesting that either approach may be used.

Alloy chemistry	**Frequency**
Mg-Ni	29
Co-V	16
Ni-V	10
Cu-Mg	1

Table 4.7: Alloy chemistries from Table 4.6 with charge neutral compositions.

This may be further refined by only considering alloys predicted to have a formation enthalpy of less than -0.1 eV/atom. Given the prediction error in this ML model, predictions of borderline metastable cases may in fact not be reliable. Thus, the sample space further reduces to the list of 36 compositions in Appendix B, which is summarised in Table 4.8.

Alloy chemistry	**Frequency**
Mg-Ni	20
Co-V	15
Ni-V	1

Table 4.8: Remaining charge neutral alloy chemistries such that ΔH_f(alloy) < -0.1 eV/atom.

4.6 Systems of interest

Calculation of DFT energies for materials known only by a compositional form requires an appropriate crystal structure geometry (see Section 4.4). For elemental crystals, alloys, or hydrides represented in the ICSD or OQMD repositories, these structures can be directly found. In many cases, alloy or hydride compositions may not have a known geometry in these databases. If there is structure for an intermetallic but not a corresponding hydride, a tetra search can be used to generate a starting point for further geometry optimisation. If neither are present, then a CALYPSO calculation is required for the alloy from which a tetra search can be performed.

4.6.1 Known alloy structures

Considering the systems presented in Table 4.8, OQMD possesses data for stable alloys of the form Mg_2Ni, $MgNi_2$, Co_3V, CoV_3, Ni_2V, Ni_3V, and NiV_3, as well as data for stable ternary hydrides Co_3VH, and Mg_2NiH_4. Of the generated compositions, two correspond to the above alloy stoichiometries: CoV_3H_7 and NiV_3H_7. These examples lend themselves to the use of the tetrahedral structure search algorithm for generation of a possible initial hydride structure. Other compounds will require more involved structure prediction of hydride form and/or alloy crystal.

Extracting the geometries of elemental reference structures, and dehydrogenation products across various other possible reaction pathways, DFT energies can be calculated. Product species consist of elemental crystals, the corresponding alloy to directly hydrogenate into the ternary species, and the most stable binary hydride constructed from constituent elements.

Exact alloys known for generated compositions

The alloy structures of CoV_3 and NiV_3 are taken from OQMD and a tetra search is used to insert hydrogen atoms into tetrahedral holes. The maximum capacity is seven H atoms, conveniently admitting the same composition as the predicted ternary hydride.

Compound	Atoms in simulation cell	Total energy, eV	Energy per atom, eV/atom
H_2	2	-6.759	-3.379
Co	4	-28.717	-7.179
Ni	4	-23.506	-5.877
V	2	-19.273	-9.637
CoV_3	8	-74.463	-9.308
NiV_3	8	-71.011	-8.876
VH_2	12	-68.700	-5.725
CoV_3H_7	22	-114.426	-5.201
NiV_3H_7	22	-111.558	-5.071

Table 4.9: DFT energies from relevant species using structures available in OQMD, to three decimal places.

$$\mathrm{Co} + 3\mathrm{V} \rightarrow \mathrm{CoV_3} \qquad \Delta H_f = -0.286 \text{ eV/atom} \tag{4.16}$$

$$\mathrm{V} + \mathrm{H_2} \rightarrow \mathrm{VH_2} \qquad \Delta H_f = -0.260 \text{ eV/atom} \tag{4.17}$$

$$\mathrm{Ni} + 3\mathrm{V} \rightarrow \mathrm{NiV_3} \qquad \Delta H_f = -0.180 \text{ eV/atom} \tag{4.18}$$

$$\mathrm{Co} + 3\mathrm{V} + 3.5\mathrm{H_2} \rightarrow \mathrm{CoV_3H_7} \qquad \Delta H_f = 0.230 \text{ eV/atom} \tag{4.19}$$

$$\mathrm{CoV_3} + 3.5\mathrm{H_2} \rightarrow \mathrm{CoV_3H_7} \qquad \Delta H_f = 0.334 \text{ eV/atom} \tag{4.20}$$

$$\mathrm{Co} + 3\mathrm{VH_2} + 0.5\mathrm{H_2} \rightarrow \mathrm{CoV_3H_7} \qquad \Delta H_f = 0.443 \text{ eV/atom} \tag{4.21}$$

$$\mathrm{Ni} + 3\mathrm{V} + 3.5\mathrm{H_2} \rightarrow \mathrm{NiV_3H_7} \qquad \Delta H_f = 0.242 \text{ eV/atom} \tag{4.22}$$

$$\mathrm{NiV_3} + 3.5\mathrm{H_2} \rightarrow \mathrm{NiV_3H_7} \qquad \Delta H_f = 0.307 \text{ eV/atom} \tag{4.23}$$

$$\mathrm{Ni} + 3\mathrm{VH_2} + 0.5\mathrm{H_2} \rightarrow \mathrm{NiV_3H_7} \qquad \Delta H_f = 0.455 \text{ eV/atom} \tag{4.24}$$

Composition	**Predicted ΔH_f, eV/atom**	**Calculated ΔH_f, eV/atom**
VH_2	-0.199	-0.260
CoV_3	-0.168	-0.286
NiV_3	-0.112	-0.180
CoV_3H_7	-0.132	0.230
NiV_3H_7	-0.161	0.242

Table 4.10: Comparison of enthalpy values for compositions in Table 4.9; predicted by respective alloy or hydride model and calculated from elemental DFT energies.

For the hydride structures, and possible dehydrogenation products, energies were calculated and are presented in Table 4.9. These values were used in Equations 4.16-4.24 to calculate corresponding formation enthalpies, which were then compared to predictions and are presented in Table 4.10.

This process and presentation format was repeated for the remaining intermetallic pairings.

Mg-Ni

The stable Mg-Ni alloy species in OQMD are Mg_2Ni and $MgNi_2$. A glitch in the tetra search failed to conserve stoichiometry of the base Mg_2Ni system on expansion for new tetrahedra. The process only correctly admitted results for $MgNi_2$, and as such a ternary composition of $MgNi_2H_{5.75}$.

Compound	**Atoms in simulation cell**	**Total energy, eV**	**Energy per atom, eV/atom**
Mg_2Ni	18	-60.307	-3.350
$MgNi_2$	24	-115.077	-4.795
MgH_2	6	-18.468	-3.078
$MgNi_2H_{5.75}$	70	-265.917	-4.542

Table 4.11: DFT energies from relevant species using structures available in OQMD, to three decimal places.

$$\mathrm{Mg} + \mathrm{H_2} \rightarrow \mathrm{MgH_2} \qquad \Delta H_f = -0.229 \text{ eV/atom} \quad (4.25)$$

$$\mathrm{Mg} + 2\mathrm{Ni} \rightarrow \mathrm{MgNi_2} \qquad \Delta H_f = -0.281 \text{ eV/atom} \quad (4.26)$$

$$2\mathrm{Mg} + \mathrm{Ni} \rightarrow \mathrm{Mg_2Ni} \qquad \Delta H_f = -0.199 \text{ eV/atom} \quad (4.27)$$

$$\mathrm{Mg} + 2\mathrm{Ni} + 2.875\mathrm{H_2} \rightarrow \mathrm{MgNi_2H_{5.75}} \qquad \Delta H_f = -0.031 \text{ eV/atom} \quad (4.28)$$

$$\mathrm{MgNi_2} + 2.875\mathrm{H_2} \rightarrow \mathrm{MgNi_2H_{5.75}} \qquad \Delta H_f = 0.066 \text{ eV/atom} \quad (4.29)$$

Composition	**Predicted ΔH_f, eV/atom**	**Calculated ΔH_f, eV/atom**
MgH_2	-0.146	-0.229
$MgNi_2$	-0.213	-0.281
Mg_2Ni	-0.118	-0.199
$MgNi_2H_{5.75}$	-0.111	-0.031

Table 4.12: Comparison of enthalpy values for compositions in Table 4.11; predicted by respective alloy or hydride model and calculated from elemental DFT energies.

Co-V

The only other stable Co-V alloy species in OQMD not yet covered is Co_3V.The tetra search admits a ternary composition of Co_3VH_6.

Compound	**Atoms in simulation cell**	**Total energy, eV**	**Energy per atom, eV/atom**
Co_3V	24	-196.790	-8.200
VH_2	12	-68.700	-5.725
Co_3VH_6	60	-272.492	-4.542

Table 4.13: DFT energies from relevant species using structures available in OQMD, to three decimal places.

$$3Co + V \rightarrow Co_3V \qquad \Delta H_f = -0.406 \text{ eV/atom} \tag{4.30}$$

$$3Co + V + 3H_2 \rightarrow Co_3VH_6 \qquad \Delta H_f = 0.603 \text{ eV/atom} \tag{4.31}$$

$$Co_3V + 3H_2 \rightarrow Co_3VH_6 \qquad \Delta H_f = 1.209 \text{ eV/atom} \tag{4.32}$$

$$3Co + VH_2 + 2H_2 \rightarrow Co_3VH_6 \qquad \Delta H_f = 0.681 \text{ eV/atom} \tag{4.33}$$

Composition	**Predicted ΔH_f, eV/atom**	**Calculated ΔH_f, eV/atom**
Co_3V	-0.406	-0.086
Co_3VH_6	0.015	0.603

Table 4.14: Comparison of enthalpy values for compositions in Table 4.13; predicted by respective alloy or hydride model and calculated from elemental DFT energies.

Cu-Mg

The stable Cu-Mg alloy species in OQMD are Cu_2Mg and $CuMg_2$. The tetra search admits ternary compositions of $CuMg_2H_5$ and Cu_2MgH_5.

Compound	Atoms in simulation cell	Total energy, eV	Energy per atom, eV/atom
$CuMg_2$	48	-131.043	-2.730
Cu_2Mg	24	-86.543	-3.606
MgH_2	6	-18.468	-3.078
$CuMg_2H_5$	128	-404.892	-3.163
Cu_2MgH_5	64	-210.439	-3.288

Table 4.15: DFT energies from relevant species using structures available in OQMD, to three decimal places.

$$Cu + 2Mg \rightarrow CuMg_2 \qquad \Delta H_f = -0.099 \text{ eV/atom} \quad (4.34)$$

$$2Cu + Mg \rightarrow Cu_2Mg \qquad \Delta H_f = -0.132 \text{ eV/atom} \quad (4.35)$$

$$Mg + H_2 \rightarrow MgH_2 \qquad \Delta H_f = -0.229 \text{ eV/atom} \quad (4.36)$$

$$Cu + 2Mg + 2.5H_2 \rightarrow CuMg_2H_5 \qquad \Delta H_f = -0.065 \text{ eV/atom} \quad (4.37)$$

$$CuMg_2 + 2.5H_2 \rightarrow CuMg_2H_5 \qquad \Delta H_f = -0.027 \text{ eV/atom} \quad (4.38)$$

$$Cu + 2MgH_2 + 0.5H_2 \rightarrow CuMg_2H_5 \qquad \Delta H_f = 0.107 \text{ eV/atom} \quad (4.39)$$

$$2Cu + Mg + 2.5H_2 \rightarrow Cu_2MgH_5 \qquad \Delta H_f = 0.127 \text{ eV/atom} \quad (4.40)$$

$$Cu_2Mg + 2.5H_2 \rightarrow Cu_2MgH_5 \qquad \Delta H_f = 0.176 \text{ eV/atom} \quad (4.41)$$

$$2Cu + 2Mg + 1.5H_2 \rightarrow Cu_2MgH_5 \qquad \Delta H_f = 0.212 \text{ eV/atom} \quad (4.42)$$

Composition	**Predicted ΔH_f, eV/atom**	**Calculated ΔH_f, eV/atom**
MgH_2	-0.146	-0.229
$CuMg_2$	-0.041	-0.099
Cu_2Mg	-0.091	-0.132
$CuMg_2H_5$	-0.060	-0.065
Cu_2MgH_5	-0.071	0.127

Table 4.16: Comparison of enthalpy values for compositions in Table 4.15; predicted by respective alloy or hydride model and calculated from elemental DFT energies.

Ca-Sn

The stable Ca-Sn alloy species in OQMD are CaSn, Ca_2Sn, and $CaSn_3$. The tetra search admits ternary compositions of Ca_2SnH_6 and $CaSn_3H_9$, as CaSn does not have many available tetrahedral structures to allow for a useful amount of hydrogen atom placements.

Compound	Atoms in simulation cell	Total energy, eV	Energy per atom, eV/atom
Ca_2Sn	12	-42.318	-3.527
CaH_2	12	-43.495	-3.625
CaSn	8	-30.918	-3.865
$CaSn_3$	4	-16.489	-4.122
Ca_2SnH_6	36	-117.881	-3.275
$CaSn_3H_9$	13	-41.242	-3.173

Table 4.17: DFT energies from relevant species using structures available in OQMD, to three decimal places.

$$2Ca + Sn \rightarrow Ca_2Sn \qquad \Delta H_f = -0.704 \text{ eV/atom} \tag{4.43}$$

$$Ca + Sn \rightarrow CaSn \qquad \Delta H_f = -0.715 \text{ eV/atom} \tag{4.44}$$

$$Ca + 3Sn \rightarrow CaSn_3 \qquad \Delta H_f = -0.481 \text{ eV/atom} \tag{4.45}$$

$$Ca + H_2 \rightarrow CaH_2 \qquad \Delta H_f = -0.649 \text{ eV/atom} \tag{4.46}$$

$$2Ca + Sn + 3H_2 \rightarrow Ca_2SnH_6 \qquad \Delta H_f = -0.081 \text{ eV/atom} \tag{4.47}$$

$$Ca_2Sn + 3H_2 \rightarrow Ca_2SnH_6 \qquad \Delta H_f = 0.154 \text{ eV/atom} \tag{4.48}$$

$$2CaH_2 + Sn + H_2 \rightarrow Ca_2SnH_6 \qquad \Delta H_f = 0.352 \text{ eV/atom} \tag{4.49}$$

$$Ca + Sn + 4.5H_2 \rightarrow CaSn_3H_9 \qquad \Delta H_f = 0.287 \text{ eV/atom} \tag{4.50}$$

$$CaH_2 + 3Sn + 3.5H_2 \rightarrow CaSn_3H_9 \qquad \Delta H_f = 0.437 \text{ eV/atom} \tag{4.51}$$

$$CaSn + 2Sn + 4.5H_2 \rightarrow CaSn_3H_9 \qquad \Delta H_f = 0.397 \text{ eV/atom} \tag{4.52}$$

$$CaSn3_2 + 4.5H_2 \rightarrow CaSn_3H_9 \qquad \Delta H_f = 0.435 \text{ eV/atom} \tag{4.53}$$

Composition	**Predicted ΔH_f, eV/atom**	**Calculated ΔH_f, eV/atom**
CaH_2	-0.541	-0.649
CaSn	-0.703	-0.715
Ca_2Sn	-0.605	-0.704
$CaSn_3$	-0.485	-0.481
Ca_2SnH_6	-0.404	-0.081
$CaSn_3H_9$	-0.073	0.287

Table 4.18: Comparison of enthalpy values for compositions in Table 4.17; predicted by respective alloy or hydride model and calculated from elemental DFT energies.

4.6.2 Generated alloy structures

Arbitrarily selecting compositions from the sample represented in Table 4.7, alloy structures are generated using CALYPSO, and interstitial hydrogen atoms inserted by means of the tetra search process. Relaxing these structures provides a low energy geometry from which elemental reference energies can be used to calculate an enthalpy of formation.

The data is presented in the same format as used in Section 4.6.1, with the addition of figures depicting the initial generated alloy structure, the alloy crystal with rudimentary interstitial occupation and the final relaxed geometry.

Mg-Ni

Compound	Atoms in simulation cell	Total energy, eV	Energy per atom, eV/atom
$Mg_{10}Ni_7$	34	-85.295	-2.509
$Mg_{10}Ni_7H_{10}$	54	-160.536	-2.973

Table 4.19: DFT energy for predicted species using the most energetically favourable alloy structure from CALYPSO, occupied with interstitial hydrogen by use of the tetra search process, to three decimal places.

$$10Mg + 7Ni \rightarrow Mg_{10}Ni_7 \qquad \Delta H_f = 0.963 \text{ eV/atom} \tag{4.54}$$

$$10Mg + 7Ni + 5H_2 \rightarrow Mg_{10}Ni_7H_{10} \qquad \Delta H_f = 0.089 \text{ eV/atom} \tag{4.55}$$

Composition	Predicted ΔH_f, eV/atom	Calculated ΔH_f, eV/atom
$Mg_{10}Ni_7$	-0.119	0.963
$Mg_{10}Ni_7H_{10}$	-0.230	0.089

Table 4.20: Comparison of enthalpy values for the composition in Table 4.19; predicted by hydride model and calculated from elemental DFT energies.

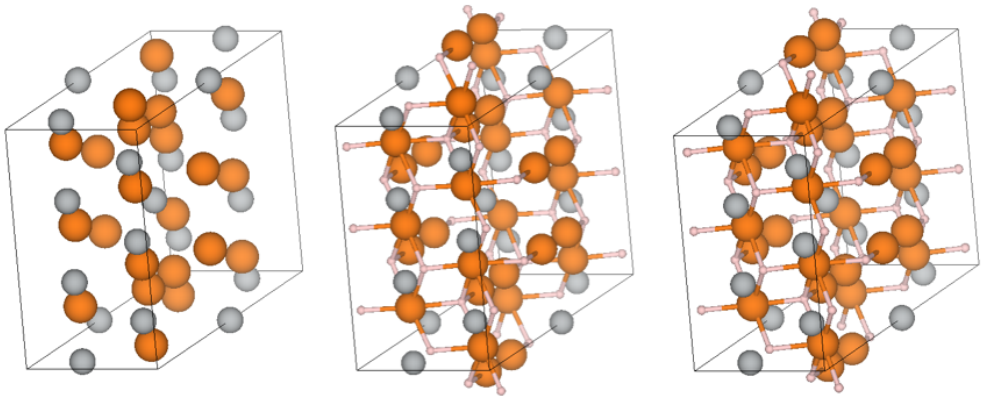


Figure 4.3: $Mg_{10}Ni_7H_{10}$: lowest energy generated alloy structure, occupied by interstitial hydrogen *via* tetra search and final relaxed hydride structure, respectively.

Co-V

Compound	Atoms in simulation cell	Total energy, eV	Energy per atom, eV/atom
CoV_2	6	-51.694	-8.616
CoV_2H_5	16	-88.056	-5.504

Table 4.21: DFT energy for predicted species using the most energetically favourable alloy structure from CALYPSO, occupied with interstitial hydrogen by use of the tetra search process, to three decimal places.

$$\mathrm{Co} + 2\mathrm{V} \rightarrow \mathrm{CoV_2} \qquad \Delta H_f = \quad 0.076 \text{ eV/atom} \tag{4.56}$$

$$\mathrm{Co} + 2\mathrm{V} + 2.5\mathrm{H_2} \rightarrow \mathrm{CoV_2H_5} \qquad \Delta H_f = -0.085 \text{ eV/atom} \tag{4.57}$$

Composition	Predicted ΔH_f, eV/atom	Calculated ΔH_f, eV/atom
CoV_2	-0.141	0.076
CoV_2H_5	-0.133	-0.085

Table 4.22: Comparison of enthalpy values for the composition in Table 4.21; predicted by hydride model and calculated from elemental DFT energies.

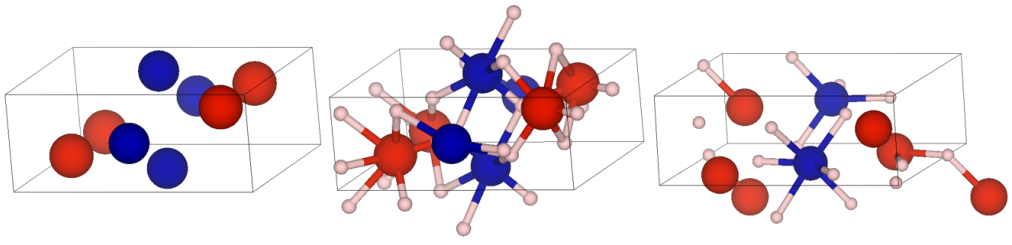


Figure 4.4: CoV_2H_5: lowest energy generated alloy structure, occupied by interstitial hydrogen *via* tetra search and final relaxed hydride structure, respectively.

Ca-Sn

Compound	Atoms in simulation cell	Total energy, eV	Energy per atom, eV/atom
Ca_5Sn_7	12	-45.943	-3.829
$Ca_5Sn_7H_7$	19	-70.032	-3.6859

Table 4.23: DFT energy for predicted species using the most energetically favourable alloy structure from CALYPSO, occupied with interstitial hydrogen by use of the tetra search process, to three decimal places.

$$5Ca + 7Sn \rightarrow Ca_5Sn_7 \qquad \Delta H_f = -0.515 \text{ eV/atom} \tag{4.58}$$

$$5Ca + 7Sn + 3.5H_2 \rightarrow Ca_5Sn_7H_7 \qquad \Delta H_f = -0.348 \text{ eV/atom} \tag{4.59}$$

Composition	Predicted ΔH_f, eV/atom	Calculated ΔH_f, eV/atom
Ca_5Sn_7	-0.653	-0.515
$Ca_5Sn_7H_7$	-0.396	-0.348

Table 4.24: Comparison of enthalpy values for the composition in Table 4.23; predicted by hydride model and calculated from elemental DFT energies.

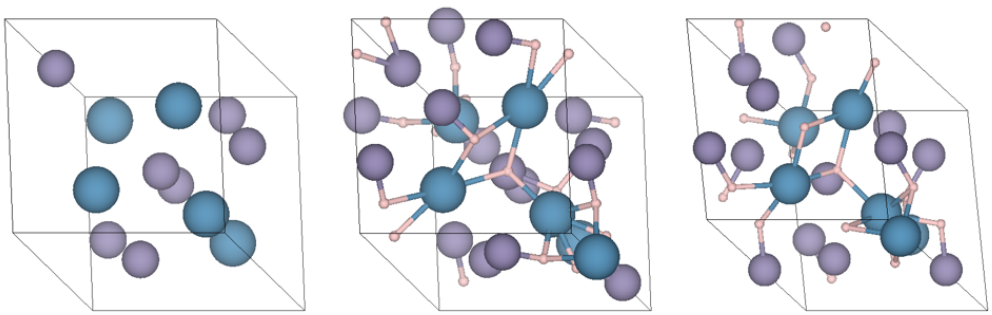


Figure 4.5: $Ca_5Sn_7H_7$: lowest energy generated alloy structure, occupied by interstitial hydrogen *via* tetra search and final relaxed hydride structure, respectively.

4.7 Discussion

4.7.1 DFT settings

It should be noted that the DFT calculations in this work are conducted to a somewhat lenient level of precision, as outlined in Section 4.2.1. Whilst production level calculations are usually parameterised with an energy cut-off of no less than 520 eV, lower k-point spacing, and often reduced loop convergence thresholds, an approach was taken to perform preliminary calculations so as to gauge approximate energy values from which to determine formation enthalpy values. Upon verification of predictions from the model, these calculations would be refined to more comprehensively assess the accuracy of prediction.

4.7.2 Validation to known compositions

Calculated enthalpy values for the compositions in the withheld test set suggest reasonable agreement between the predicted values and the ground truth data from the database, as shown in Table 4.25. Calculations were performed by the DFT process outlined in Section 4.2, using crystallographic information obtained directly from Composition entries in OQMD, whilst the machine learning model built with hydride data minus the validation data points was used. Predictive errors appear to be generally smaller than calculated errors, however this is potentially due to the precision of these first principles calculations, as mentioned above, and could well be further mitigated.

Ternary hydride	Prediction error, eV/atom	Calculation error, eV/atom
LaH_3	0.006	0.079
La_4H_9	0.006	0.011
$AlMgH_5$	0.015	0.039
$CoSr_2H_6$	0.008	0.093
$LaMg_2NiH_7$	0.031	0.066
AlK_2LiH_6	0.020	0.096

Table 4.25: Prediction and calculation absolute errors for data in Table 4.2, to three decimal places.

Of these samples, LaH_3, $AlMgH_5$, and AlK_2LiH_6 are structures generated from prototypical structures, whilst the others are from known ICSD entries. It is also worth noting that the two binary hydrides which were randomly selected are both La-H species and both admit reasonable prediction error, which may suggest good predictive ability for similar species of slightly varying stoichiometry. Overall this suggests good predictive ability, with the largest error of prediction in this test set being 31 meV/atom. The systematically lower calculated value to the database results is likely as a result of the elemental reference data used.

These results are useful, and suggest the model is suitably trained to the standard of structures and calculated values presented by the database that it was trained on. This, however, is not proof of ability to generalise. As mentioned in Section 3.4.1, the methodology used by the OQMD database is rooted in derivation of enthalpy of formation exclusively from elemental reference energies. In order to analyse whether this is a realistic approach to such a problem and, by extension, to test the accuracy and generalisability of the model, experimental enthalpy values were gathered from a mostly-empirical database for alternative validation. The top five most commonly occurring 'Composition' entries from the HydPARK database were subjected to the same calculation and prediction process

as the above test set. If this was a hydride species, the entry with the lowest enthalpy value in OQMD was used to obtain a structure for the calculations. If a binary alloy, the structure with the lowest enthalpy value for any corresponding ternary hydride was used instead. The final machine learning model constructed using the full pruned dataset was used for predictions.

Ternary hydride	**Prediction error, eV/atom**	**Calculation error, eV/atom**
Mg_2NiH_4	0.015	0.045
$NaAlH_4$	0	0.015
$LaNi_5H_7$	0.032	0.058
Na_3AlH_6	0.024	0.033
$ZrCr_2H_3$	0.347	0.022

Table 4.26: Prediction and calculation absolute errors for data in Table 4.5, to three decimal places.

Predictive and calculation errors relative to the database values are mostly comparable to that of the test set, with the exception of the prediction for $ZrCr_2H_3$ (see Table 4.26). Of these selections from the HydPARK database, several entries for the alloy $ZrCr_2$ exist with H/M ratios ranging from 1.8-2.1, yet the only Zr-Cr-H species in OQMD is $ZrCr_2H_3$, where H/M=1. Further still, the OQMD entry implies a theoretical enthalpy values of 0.13 eV/atom, suggesting this is an unstable hydride species and as such unlikely to be of use as a hydrogen store. This discrepancy might be explained by the structure generation process used by OQMD. The method of iterating upon prototypical geometries does not ensure complete coverage of a combinatorial space, and could lead to missing potential best-in-class results. It is also possible that not all experimental results are collected and/or converted correctly, as multiple entries for $ZrCr_2H_4$ can be found in ICSD. Another explanation, which may also be more widely

applicable to other entries in OQMD, is that the hydride structure stored may not be in the ground state, or hydrogen atoms not located at the optimal interstitial sites. If so, this could impact the model's performance if trained on incorrect or inaccurate data. These reasons demonstrate the need to sufficiently 'clean' the OQMD dataset before productive machine learning training can take place.

4.7.3 Ternary composition generation and filtering

When determining a method of generating new compositions, a systematic process is required so as to drastically narrow down the exhaustive combinatorial space across a range of stoichiometries. The method presented here, being to concatenate multiples of the most stable binary hydride result for each metal, is based on the presumption that the H/M ratio of such compositions should implicitly encode information regarding the oxidation states and electronegativity of the metallic component. If this were the case, the resultant combinations have a good likelihood of being charge neutral with respect to the most common metal oxidation states. Once reduced to a unique set, prior to any prediction process, these ternary compositions required filtering by means of empirically justified reasoning so as to reduce the space to a more manageable size.

For an on-board hydrogen store, weight is at a premium. Weight considerations of all aspects of system design are of great importance for such an application - from fuel cell stack, to fuel tank - as it can greatly impact the fuel consumption, and as such range, of the vehicle [93]. Whilst overall system weight is of importance, and the extent of cooling required is a function of the heat of reaction for the dehydrogenation process, this work focuses solely on material weight of these hydrides. A lower limit of 3%

hydrogen by mass, determined by the mass of hydrogen atoms as a fraction of the total mass of constituent atoms, is chosen as a preliminary filtering stage. This is towards the lower end of gravimetric capacity of known useful complex and metal hydride storage materials, whilst still being significantly below technical targets [94].

To determine a window for enthalpy values of interest the aforementioned range of entropy values for metal hydrides is used. By using Equation 4.14, it can be seen that the -50 to -30 kJ $mol_{H_2}^{-1}$ envelope corresponds to 300K - 500K (26.85°C - 226.85°C), and 200K - 333K (-73.15°C - 59.85°C), for dS equal to -100 and -150 J K^{-1} $mol_{H_2}^{-1}$ respectively. The mutually encompassed region of temperature for 1 bar hydrogen release is approximately 27°C to 60°C which corresponds to a range from a near-ambient value, up to a typical PEM fuel cell operating temperature.

At this stage, formation enthalpy predictions are made for the ternary hydrides and the corresponding dehydrogenated intermetallic species, from which only those ternary materials within the above enthalpy range, and with a stable alloy component, are taken. This was decided to facilitate several reaction pathways, being the combination of elemental reference species, direct hydrogenation of an alloy, or introducing a secondary metal and appropriate hydrogen to combine with a known binary hydride. Theoretically, each of these scenarios would thus correspond to stable product materials throughout the cycling process.

A degree of chemical intuition was then required to reject certain elements from the resultant set. With the target of effective, yet affordable, renewable technologies, expensive components should be avoided where possible. An example of this is ongoing research to substitute the pricey platinum-based materials used to catalyse fuel cell reactions with

a cheaper, more sustainable alternative. By this logic, it was decided to exclude precious metals such as those in the platinum group, as well as some similarly costly noble metals. Regardless of results with respect to predicted energies, such systems would simply be impractical for the use case at hand. As previously mentioned, beryllium is not considered due to health hazards. Technetium, whose isotopes are all radioactive, is excluded for a combination of cost and safety issues. Compositions with an aluminium component are an exceptional case. Group I and II elements can form a ternary complex hydride - a salt-like structure with anions of AlH_4^- tetrahedra, and cations of the secondary metal. Such materials have very high gravimetric densities of hydrogen storage but have proven difficult to cycle due to high kinetic barriers to hydrogenation and dehydrogenation [84]. Considering the interstitial nature of most other metal hydrides, a presumption was made that the machine learning model might struggle to generalise to these alanates.

Somewhat surprisingly given the construction process from known binary hydrides, only around half of the remaining candidates were charge neutral, suggesting that an alternative process of ternary composition generation, more stringent in cross-referencing stoichiometry to ionisation energies, may have been more appropriate. Of these, alloys were filtered to have a predicted formation enthalpy of less than -0.1 eV/atom in order to remove any predicted metastable values which may have fallen within the margin of error of the predictive model. Whilst this might be considered a large threshold value, the results admitted were identical to when considering a threshold of -0.05 eV/atom.

4.7.4 Crystal structure prediction methodology

The ability to predict a material's structure solely from its composition has been sought after for decades [95, 96]. Whilst chemical rules and stipulations can be defined and considered in such a process, no perfectly reliable process has been developed as of yet.

Similar workflows to that presented here have been reported for a range of material types: construction of a machine learning model from previously calculated and collated data sources, followed by property prediction for novel compositions coupled with structure prediction, and finally validation by means of first principles calculations. This includes work for oxides and nitrides, amongst others [82, 97]. However, such materials are very widely studied, and often have systematic geometries similar to prototypical structures. The wider availability of data facilitates better trained machine learning models, improving prediction capabilities, and if certain samples exist naturally, it is more likely that this information closely describes ground state structures. This more accurate data, and any similar geometries (e.g. perovskites), allows for better use of methods for structure prediction by analogy. In this case, metal alloys and hydrides have great diversity in atomic configuration and, as such, do not lend themselves to substitution *via* prototypical systems. Coupled with the limited training data available, this may limit the effectiveness of resultant predictive models.

CALYPSO is used to predict the intermetallic crystal structure, before interstitial hydrogen is inserted into tetrahedral holes. The parameterisation of these calculations searched for systems with one formula unit of the alloy per unit cell, which may be insufficient in finding the ground state structure by assuming a particularly high symmetry of

the species. Further, the tetra search process assumes the formation of an interstitial hydride. Whilst relatively commonplace, various chemistries exist for hydride compounds which may require consideration on a case-by-case basis.

4.7.5 Predictions for new compositions

In this section, alloy structures were first either sourced from crystallographic information available in OQMD if available, or predicted using CALYPSO if not. These were then occupied with interstitial hydrogen atoms after which geometry optimisation, consistently parameterised to the validation stages, was used to calculate enthalpy values with respect to elemental reference DFT energies.

Whilst the method outlined by the OQMD data calculation protocol is the reaction process that these machine learning models are trained upon (see Section 4.2), and as such will provide the mathematical rules for prediction, other possible dehydrogenation reaction pathways were also considered and used to determine formation enthalpy values for the ternary species. Due to the proposed generalisation of the machine learning models, it would also allow for prediction of these binary compounds from elemental reference energies. Ultimately, prediction will only be analysed for calculation from elemental DFT energies.

Of the ternary compositions which remain following the filtering process, two of them contain an intermetallic component corresponding to a stable alloy structure found in OQMD, as well as two entries for stable hydrides of these chemistries. Using the alloy crystallographic data, a tetra search was performed which admitted a maximum of seven tetrahedral holes

that could be occupied with interstitial hydrogen atoms. Conveniently, this directly coincides with exact generated compositions with these alloy stoichiometries. Alternatively, the broad intermetallic chemistries are searched for in the database and stable alloy structures collected, for which the tetra search method is used to provide a ternary hydride form to be further relaxed.

Composition	**Absolute prediction error, eV/atom**
VH_2	0.061
CoV_3	0.118
NiV_3	0.058
CoV_3H_7	0.362
NiV_3H_7	0.403
MgH_2	0.083
$MgNi_2$	0.068
Mg_2Ni	0.081
$MgNi_2H_{5.75}$	0.080
Co_3V	0.320
Co_3VH_6	0.588
MgH_2	0.083
$CuMg_2$	0.058
Cu_2Mg	0.041
$CuMg_2H_5$	0.005
Cu_2MgH_5	0.198
CaH_2	0.108
$CaSn$	0.012
Ca_2Sn	0.099
$CaSn_3$	0.004
Ca_2SnH_6	0.323
$CaSn_3H_9$	0.360

Table 4.27: Prediction absolute errors for data in Section 4.6.1, to three decimal places.

The absolute errors of prediction are presented in Table 4.27. In general, these error values are noticeably greater for ternary species compared to the binary compounds. This could perhaps be a function of the rudimentary insertion of hydrogen atoms into the metallic crystal; assuming an interstitial nature of the hydride, or failing to consider other geometries. Issues also arise from the uncertainty as to the true ground state of such species. Whilst the precision of these first principles calculations may have an impact on the accuracy of these values, as highlighted earlier, these error values are strikingly large. For example, the largest absolute error for a ternary species is that of Co_3VH_6, where 0.588 eV/atom corresponds to 189 kJ/mol_{H_2}.

Composition	**Absolute prediction error, eV/atom**
$Mg_{10}Ni_7$	1.082
$Mg_{10}Ni_7H_{10}$	0.319
CoV_2	0.217
CoV_2H_5	0.048
Ca_5Sn_7	0.138
$Ca_5Sn_7H_7$	0.048

Table 4.28: Prediction absolute errors for data in Section 4.6.2, to three decimal places.

A handful of compositions were selected from the set of generated ternary compounds and CALYPSO was used to generate a stable structure for the intermetallic component. The same process as above was then used to generate a hydride species, and DFT energies calculated so as to determine formation enthalpy data (see Table 4.28).

Albeit considered across only a small sample of compositions, it is evident that calculated enthalpy values are not all qualitatively consistent with

predicted values. Relatively large disparities can be seen in results for the alloy structures; issues which could carry over when introducing interstitial hydrogen. Some large error values were found, with errors as high as 1.082 eV/atom for the metal alloy model, and 0.319 eV/atom for the hydride model with the Mg-Ni systems. The other chemistries analysed offer more appropriate predictive errors for the hydride species, in line with the test error expected from construction, but admit large errors for the intermetallic species. This would suggest poor accuracy in prediction of stable alloy structures.

A clear limiting factor in accurate evaluation of predictive accuracy is the prediction of crystal structures for both the alloy and hydride species. Without a more reliable method of structure prediction, it is difficult to definitively judge the model's performance. Unfortunately, this is not a simple task. The approach in this work has been to assume an intermetallic form of hydride structure by occupying interstitial holes with no significant phase transition between the alloy and hydride state. It is likely that the OQMD data used to develop the model is based on cases where a phase transition occurs from the metallic species to the hydride form. As a result, the model may generalise poorly for compositions that would realistically take another geometry. A more robust crystal structure prediction method, or direct prediction of the hydride phase, could potentially provide more accurate results for validation. Heuristic processes for independent determination of alloy and hydride structures may prove more effective, but come with the associated increase in computational cost.

As previously discussed, machine learning models can only be expected to perform well for predictions similar to the data they are trained with. The predictive accuracy relative to the withheld test set proved to be reasonable, as with the compositions inspired by experimental results. That is to

say, it has been shown to work well with known structures and for the reaction pathway of combining elemental reference materials. Alas, it is when using this model on new data that we see poor performance. The most likely culprit is the crystal structure prediction method used. Success stories are aplenty for prediction software such as CALYPSO, USPEX [98], and AIRSS [99], but calculations require correct parameterisation. In this work, the choice of a single formula unit per cell is likely to be insufficient for the systems at hand, resulting in incorrect assumed symmetry and final geometries. Additionally, the assumption of an intermetallic hydride form without meaningful phase transition may also limit validation. Good prediction of structures is required to appropriately validate the predictions of the model for the pathway on which it was trained, and to comment on its ability to extrapolate to alternative reaction mechanisms.

4.8 Conclusion

In this chapter, the previously built predictive models, for determination of formation enthalpy from composition alone, are presented with new data previously unseen to the models. Model prediction is initially validated by comparison to calculated enthalpy values for the withheld test set and known experimental data for storage materials.

New ternary hydride data points were generated by a linear combination and concatenation process of binary hydride compositions obtained from the database, before being filtered in an iterative process to identify 36 hydrides with a suitable predicted enthalpy for on-board applications (between -50 and -30 kJ $\mathrm{mol}_{\mathrm{H_2}}^{-1}$), along with a reasonable gravimetric density of hydrogen (>3 wt%) and stable dehydrogenation products.

Of these results, those with known structures of their intermetallic component had crystallographic information collected from the database. Otherwise, sample compositions have the structure of their alloy component predicted by a heuristic crystal structure prediction algorithm. These structures are occupied with interstitial hydrogen and the systems of maximal occupancy are relaxed, allowing for calculation of enthalpy data.

Chapter 5

Oxygen transport in perovskite materials

5.1 Introduction

As mentioned in Section 1.6.1, the movement of oxygen ions through an electrolyte material is fundamental to the operation of a solid oxide fuel cell. The discovery of new cathode and electrolyte materials which cater for more improved oxygen transport dynamics is crucial in the development of more efficient, durable and financially viable fuel cell systems. The mechanics of oxygen self-diffusion can be analysed through a range of simulation techniques on a given candidate material.

The aim of this chapter is to investigate such oxygen transport dynamics in the prototypical perovskite system of barium titanate, before introducing the co-doping of lanthanum and magnesium on the A-site and B-site, respectively, with the intention of studying how such changes in stoichiometry and point defects might influence oxygen self-diffusion.

Static methods will be used to isolate and investigate energy barriers for localised diffusion around a single defect instance in an otherwise perfect barium titanate structure. These systems will then be extended to include multiple different point defects, from which molecular dynamics simulations, at a range of finite temperatures, will be used to qualitatively compare systems with varying arrangements.

5.2 Background

5.2.1 Technical challenge

In the world of semiconductors, fine tuning of composition can drastically affect electrical and magnetic properties. Doping and co-doping, as well as defect chemistry, are useful tools in this respect [100, 101]. Substitution with dopant species could cause changes in the crystal structure and microstructure, modifying ferroelectric and dielectric properties [102]. Anion transport can be influenced by such practices to the extent of having a significant contribution to electrical conductivity through a bulk material.

As explained in Section 1.6.1, oxide ion conductivity can be a rate limiting factor for fuel cell performance. Conversely, applications in the fields of capacitors and thermistors requires suppression of such mobility. Doped barium titanate has been shown to perform well in these applications, as well as for purposes such as photocatalysis [103, 104]. Understanding diffusion processes through these materials, as a function of defect concentration and distribution, would be a valuable tool for comparing ion conductivity between systems, and may help to develop processes to further suppress or enhance oxide conductivity for a range of

applications.

Investigations of diffusion in oxide ion conductors, even across a wide range of applications, ultimately revolve around the same methodologies - directly analysing the movement of oxide ions through a simulation cell of a bulk material. Aims of tuning such a system to minimise or maximise diffusion can then tailor a material for a particular use case. Low levels of non-stoichiometry in metal oxides can result from several processes, such as deliberate doping, natural impurities, or contamination during sample processing. The consequential changes in local geometry and coordination can drastically impact electrical and ionic conduction mechanisms, as well as introduce electrical inhomogeneity. By considering a prototypical perovskite structure, a series of point defects can be introduced and qualitatively compared so as to characterise their influence on localised oxygen diffusion.

5.2.2 Oxygen mobility mechanics

The potential energy surface can be used to conceptualise the movement of oxygen ions between two stable system configurations. The energetics of traversal between neighbouring minima of the potential energy surface is characterised by an intermediary saddle point. Representing a transition state configuration, the increase in the potential energy at this point, relative to the value of the surface corresponding to initial and final geometries, defines a potential barrier between the stable states. This barrier admits the activation energy required to traverse such a transition path. In the context of oxygen diffusion through a perovskite crystal, this barrier would dictate the minimum energy required for an oxygen ion to diffuse to a nearby vacant site. Intuitively, a lower activation barrier would

facilitate a higher rate of diffusion through the bulk material and thus increase ionic mobility.

In solids, activation energy for diffusion can be extracted from the Arrhenius equation by use of diffusion coefficient results calculated at a range of temperatures:

$$D = D_0 \exp\left(-\frac{E_A}{RT}\right), \tag{5.1}$$

where D is the diffusion coefficient determined for a temperature T, D_0 is the maximal diffusion coefficient (a pre-exponential factor, theoretically corresponding to infinite temperature), R is the universal gas constant, and E_A is the diffusion activation energy. By using diffusion coefficient results from MSD data, plotting reciprocal temperature against the natural log of the corresponding diffusion coefficients reveals the activation energy to be the slope multiplied by (-R).

5.2.3 Nudged elastic band

A method of investigating both the transition path and activation energy, as introduced in Section 5.2.2, by means of *ab initio* calculations, is to make use of the nudged elastic band (NEB) method. This process approximates the minimum energy path (MEP) between stable states *via* a transition state and evaluates the corresponding potential energy barrier. Fixed initial and final states, both optimised to stable geometries, are connected by an interpolated reaction path, as defined by the use, along which a set of equidistant images are defined. These transition images are connected to neighbours by a spring force which acts to maintain separation and prevent images from falling into the nearest minima.

All images are simultaneously optimised. The forces imparted on an image by the potential energy minimisation that are perpendicular to the band, in conjunction with the spring forces parallel to the band, are used to nudge the estimated transition path towards the true pathway [105]. This process is depicted in Figure 5.1. The number of NEB calculation cycles required is dependent on how close the initially defined path is to the low-energy path, and the cut-off thresholds specified for the DFT calculations.

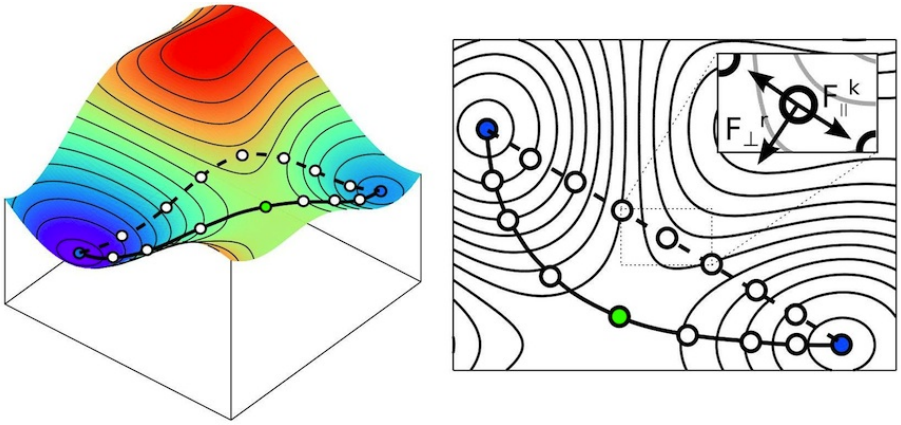


Figure 5.1: Nudged elastic band example between two minima [106].

A widely used modified version of NEB, not yet included in the latest version of VASP, is the climbing image nudged elastic band method [107]. Actively driving the image with the highest energy toward the saddle point, facilitated by allowing for variable spring constants along the length of the band to alter the spacing of images, one of the images will then converge to a geometry and energy near to that of the transition state.

5.3 Barium titanate systems

5.3.1 Unit cell structures

Room-temperature $BaTiO_3$ has a tetragonal perovskite phase consisting of TiO_6 octahedra, and Ba^{2+} on the A-site that is 12-fold coordinated to oxygen (see Figure 5.2). As temperature decreases, the orientation of the octahedral units adjusts through tilting and rotating to stabilise the structure, undergoing transitions through orthorhombic and rhombohedral phases [108, 109, 110, 111]. We shall be assuming a cubic-type perovskite structure ($Pm\bar{3}m$), for the sake of computational ease, to investigate oxygen vacancy formation and migration.

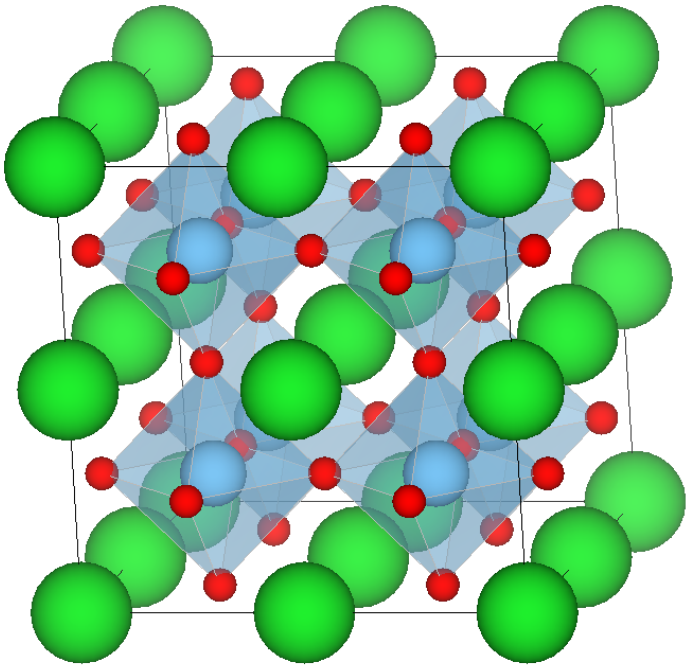


Figure 5.2: Barium titanate crystallographic structure. Ba (green), Ti (blue), O (red).

5.3.2 Defects of interest

Inspired by unpublished experimental results, a system of interest is developed by co-doping barium titanate with lanthanum and magnesium, where such point defects are used to tweak the base stoichiometry. Mg^{2+} is used as an isovalent acceptor dopant on the B-site, substituting directly for a Ti^{4+} ion, which is compensated by oxygen vacancy formation. An aliovalent La^{3+} species is a donor dopant assumed to occupy the A-site in place of a Ba^{2+} ion, whilst introducing a quarter of a titanium vacancy, $\frac{1}{4}V_{Ti}$, in recompense.

5.4 Static calculations

5.4.1 Determining a primitive transition path

Initially, the geometry of the lower-energy pathways for oxygen ion diffusion in $BaTiO_3$ were investigated. A simulation cell was created of 2×2×2 unit cells of barium titanate, and an oxygen vacancy introduced by simply deleting an oxygen atom. Now with the system $Ba_8Ti_8O_{23}$, NEB calculations were used to investigate and compare energy barriers related to the two shortest possible oxygen diffusion paths, determined intuitively from local coordination. The first of these involves movement around a neighbouring B-site atom, rounding a 'corner' of a (2 2 1) sample of the B-site sublattice, whilst the second sees traversal 'across' to an opposite side of such sublattice sample (see Figure 5.3). Calculations employ the static NEB method as implemented in VASP, and make use of the climbing image NEB functionality and NEB analysis tools of the VTST-Tools package [112]. The initial and ultimate structures are generated and relaxed by

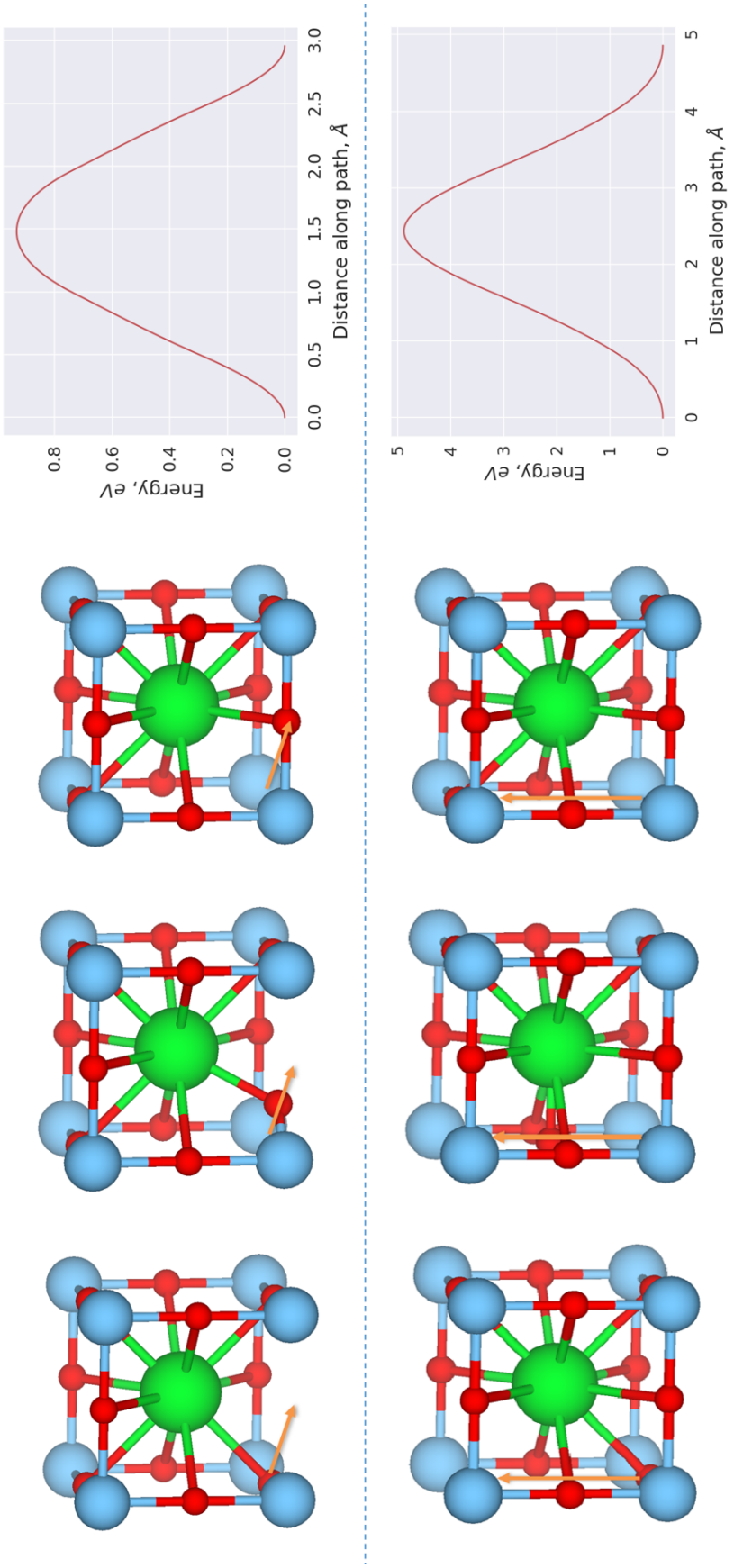


Figure 5.3: Sketch of the two shortest oxygen migration paths in $BaTiO_3$ and corresponding low-energy paths. Ba (green), Ti (blue), O (red).

Transition	Activation energy, eV	Migration path length, Å
Corner	0.93	2.93
Corner, with V_{Ba}	0.91	3.51
Across	4.87	4.86
Across, with V_{Ba}	4.76	5.16

Table 5.1: Activation energy barriers and migration path length for each of the transition path types, with and without the neighbouring V_{Ba}.

geometry optimisation, parameterised consistently to that seen in Section 4, before five intermediary images are generated by linear interpolation between these two coordinate sets. The spring constant, responsible for the nudging of the band, is kept at the default value of -5 eV/Å.

The low-energy pathways predicted for these two cases are also shown in Figure 5.3 with energy barriers of 0.93 eV and 4.87 eV. The process was repeated for a near-identical system, but now introducing a barium vacancy on the A-site adjacent to the oxygen diffusion process (the central ion in the sketches in Figure 5.3). As before, the vacancy was naïvely introduced by simply deleting the Ba^{2+} ion from the system and thus charge was not correctly compensated. Results are presented in Table 5.1.

5.4.2 NEB calculations for single defect instances

Each individual point defect was investigated as to how each influences the energy of the base barium titanate structure, and how they may impact the activation energy of local oxygen diffusion. Starting with a periodic simulation box consisting of 3×3×3 unit cells, a single instance of each defect was introduced, and a single oxygen vacancy at a coordinated site. A larger supercell was used in an attempt to reduce self-interaction between

the resultant distortions across the periodic boundary. With an aim to use NEB to study the transition path around the defect site - a 'corner' transition akin to that in Section 5.4.1 - a corresponding second cell was constructed with the oxygen vacancy manipulated to an adjacent oxygen site around the local TiO_6 octahedron, equidistant to the defect. Both initial and final position structures are relaxed and energies are presented in Table 5.2.

Climbing image nudged elastic band calculations are performed between these terminal states, with five images generated between them. Figure 5.4 presents schematics for each defect system, and their corresponding energy barrier plots. These migration barriers are further quantified in Table 5.3.

Defect	**Atoms in simulation cell**	**Total energy, eV**	**Energy per atom, eV/atom**
None	134	-1091.609	-8.146
Ba_vac	133	-1084.953	-8.158
La	134	-1096.739	-8.185
Mg	134	-1084.143	-8.091
Ti_vac	133	-1074.353	-8.078

Table 5.2: DFT energy for initial NEB structures, possessing a specified point defect and neighbouring oxygen vacancy.

Jump index	**Activation energy, eV**
a	1.20
b	1.13
c	2.03
d	1.13
e	0.75

Table 5.3: Activation energy barriers for each of the transition paths around respective defect types, corresponding to index values given in Figure 5.4.

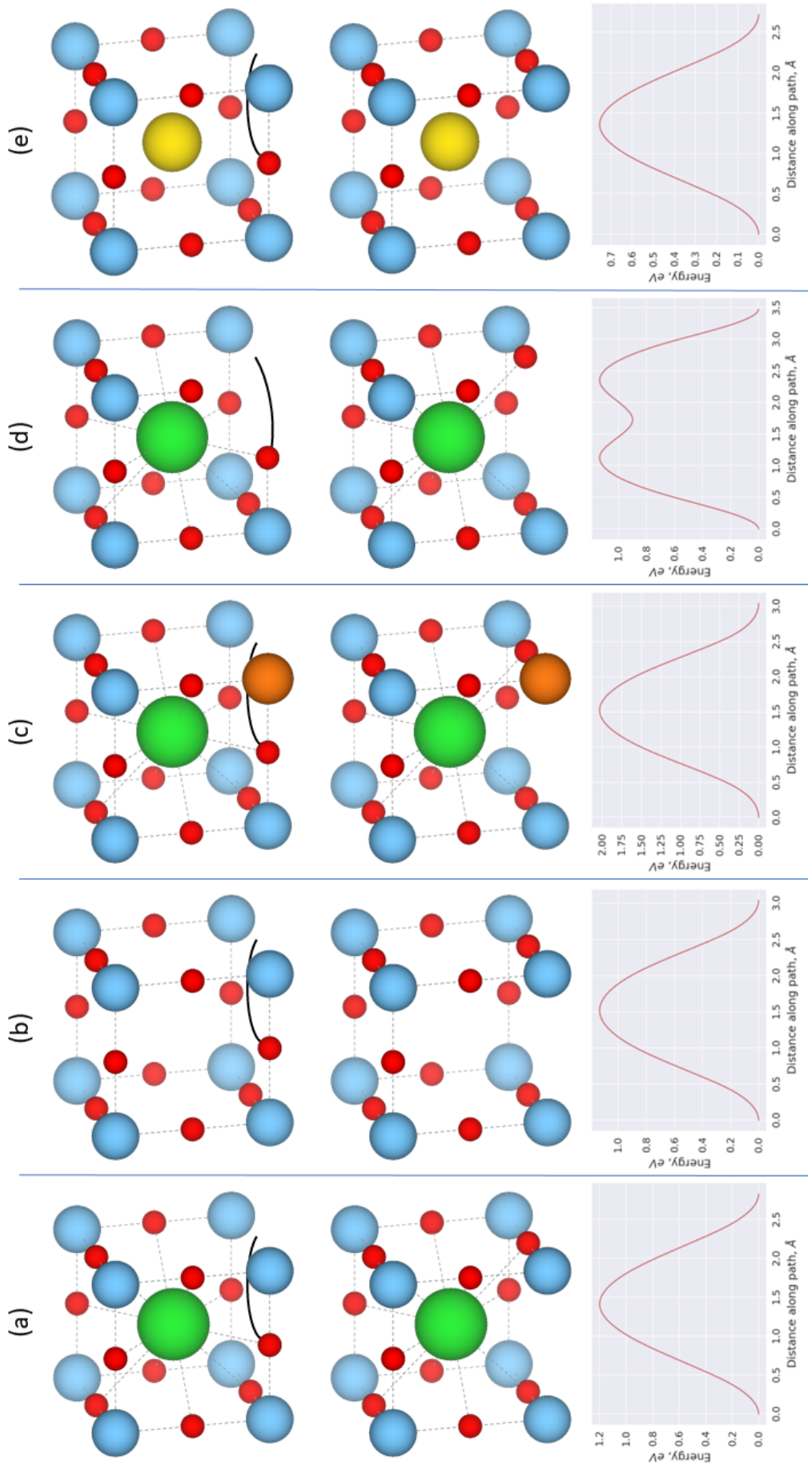


Figure 5.4: Schematic of minimum energy pathways for oxygen diffusion with a single oxygen vacancy in the presence of point defects, as predicted by first-principles calculations using the NEB method. (a) complete barium titanate cell, (b) adjacent to Ba vacancy, (c) adjacent to Mg dopant on B-site, (d) adjacent to Ti vacancy, (e) adjacent to La dopant on A-site. Ba (green), Ti (blue), O (red), Mg (orange), La (yellow).

It should be noted that only the energy of the relaxed initial state is given due to the inherent symmetry of the cell. Shown in Table 5.2, it can be seen that the energy of the system decreased upon the substitution of a lanthanum ion onto an A-site. The local defect structure around the oxygen vacancy migration path affects the migration barrier and trajectory, in turn offering favourable migration paths which, if extrapolated throughout a bulk sample, would likely have a noticeable effect on the material's conduction properties.

5.4.3 System energy dependant on lanthanum positioning

As per the doping mechanism being investigated, four lanthanum ions further introduce a titanium vacancy. Given this, there is interest as to how lanthanum ions positioned relative to such a vacancy site might impact the system energy. Possible positioning of four lanthanum ions in the immediate vicinity of the resultant B-site vacancy in a 3×3×3 supercell were exhaustively determined (see Figure 5.5) and DFT total energies calculated.

Whilst only a small difference in energy, admittedly within calculation tolerances, the lowest calculated system energy was for cell number 1. This suggested, albeit only slightly, an energetically favourable positioning of lanthanum ions local to the vacancy site in the form of a planar configuration (the full simulation cell is shown in Figure 5.6). Questions as to whether this would scale with further increases in doping and cell size, and any resultant impact on the isotropy of oxygen diffusion in the bulk material, are taken forward and considered with dynamic calculations.

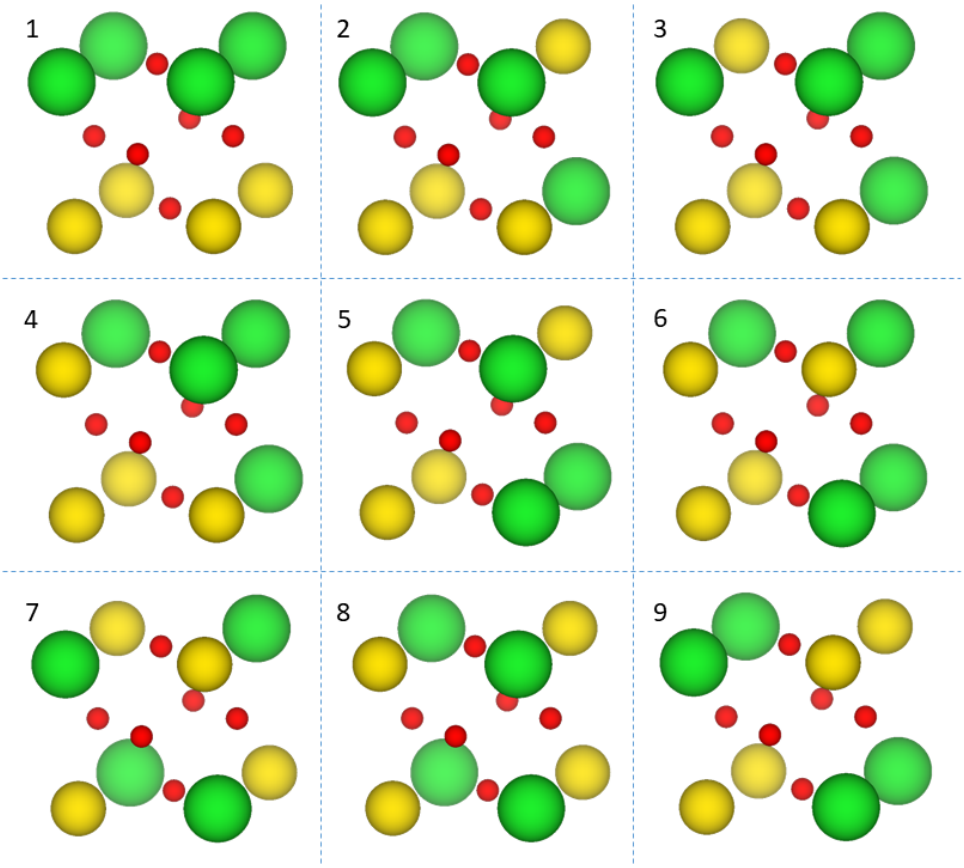


Figure 5.5: All possible geometries of four lanthanum ions substituted into a (2 2 2) sample of the A-site sublattice surrounding a B-site vacancy. Systems obtained by symmetry of these cases are energetically equivalent. Ba (green), O (red), La (yellow).

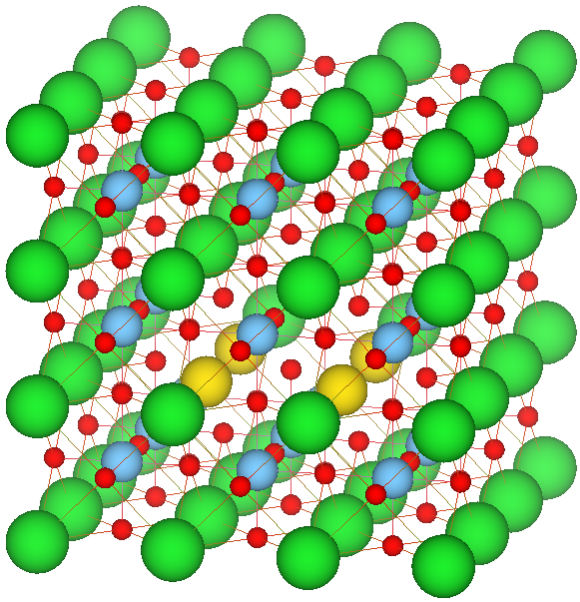


Figure 5.6: Planar lanthanum configuration of interest in an unrelaxed $3\times3\times3$ supercell. Ba (green), Ti (blue), O (red), La (yellow).

5.5 Molecular dynamics simulations at finite temperatures

The NEB calculations above reveal information regarding the transition path of oxygen, and the respective transition energy barriers, as it travels around each defect proposed for the larger system. This is presented in the form of spatially separated *ab initio* calculations along a presupposed trajectory, between which a path is interpolated. These calculations are categorised as a 'static' approach, and are calculated at 0 K. Whilst useful for understanding localised dynamics for a predetermined cell configuration, these calculations are less representative of the bulk structure with inhomogeneous defect distribution, and also have no consideration of temperature effects. Additionally, the computational cost of running numerous calculations at the required level of accuracy to develop these trajectories is a major limiting factor in continuing to evaluate more reasonable time evolution of these dynamics.

Molecular dynamics, as outlined in Section 2.8, introduces ion velocities and makes use of classical mechanical calculations discretised over time to construct dynamic particle trajectories by an iterative process. With this method, temperature dependent dynamics can be introduced to the system.

5.6 Dynamics calculations

5.6.1 Method

As aforementioned, it is possible to develop a more realistic understanding of oxygen mobility dynamics through the use of molecular dynamics. To expand upon the static calculations of these systems, the simulation cell was increased to a system of 4×4×4 unit cells of $Ba_3Ti_3O_9$, resulting in 320 total ions. Periodic boundary conditions were applied, and the NVT ensemble was considered by use of the Nosé-Hoover thermostat.

As the length of primitive lattice vectors in real-space are inversely proportional to those in reciprocal space, the use of a larger simulation cell means that fewer intersections are necessary for an equally spaced mesh. As a result of the size of the supercell used here, a k-point mesh of (1 1 1) is sufficient. This equates to consideration of only a single k-point, located at the gamma-point, in turn allowing the use of the gamma-point only VASP executable which can run up to 1.5 times faster than the standard version.

Point defects were introduced by removing or replacing appropriate ions from randomly selected sites. Total charge is conserved through combinations of various defects. The much larger simulation cell allows for implementation of a reasonable defect concentration, randomly distributed across the cell, when compared to the previous static calculations.

The molecular dynamics functionality implemented in the VASP code is used to perform a succession of calculations. Upon introducing the appropriate defects to an initial barium titanate (henceforth referred to as BTO) structure, a geometry optimisation calculation is required to relax the structure and accommodate the resultant changes in chemistry.

Initialisation of the molecular dynamics simulation involves introducing velocities to particles and equilibrating to a target temperature. This was done by means of a 1,000 step AIMD run. Once at temperature and the atoms set in motion, the output files are used to initiate a 50,000 step MLFF MD simulation. By extracting atomic coordinates from each step, dynamical trajectories for each particle can be constructed and MSD statistics can be analysed. If such statistics are deemed insufficient to characterise movement, the MLFF MD process is continued from its final state for a further 50,000 steps, when possible.

These molecular dynamics calculations were performed on the ARCHER2 UK National Supercomputing Service using the implemented VASP 6.3.0 gamma-point executable, assigned to 8 nodes on the highmem partition [88].

5.6.2 Initial supercell geometries

Starting from a 4×4×4 supercell of barium titanate, dopants are introduced with sufficient concentrations as to hopefully observe a reasonable degree of oxygen diffusion. Given the computation cost of a large timescale simulation, it is reasonable to exaggerate dopant levels in order to evaluate the qualitative impact that they might have on the overall dynamics of the system by increasing the likelihood of diffusion events occurring.

Taking inspiration from unpublished experimental results, the level of doping is chosen to be $La_{0.25}$ (16 La^{3+} on A-site), and $Mg_{0.125}$ (8 Mg^{2+} on B-site); thus introducing 8 V_O ($O_{0.958\dot{3}}$) and 12 V_{Ti} ($Ti_{0.8125}$)

Three simulation cells were created; one with magnesium doping but without lanthanum substitution (x=0, y=0.125, $Ti_{0.875}$), another with both

dopants present (x=0.25, y=0.125, $Ti_{0.8125}$), and a final cell with the same level of doping as the last, but with geometry inspired by the ground state energy results seen in Section 5.4.3 which suggested the possibility of an energetically favourable planar configuration of lanthanum.

- Mg-BTO: $Ba_{64}Ti_{56}Mg_8O_{184}$
  - Starting with an initial 4x4x4 BTO supercell; Mg atoms are randomly distributed onto B-sites, Ti vacancies are randomly introduced on B-sites, and O vacancies are randomly distributed.
- La-Mg-BTO: $Ba_{48}La_{16}Ti_{52}Mg_8O_{184}$
  - Using the Mg-BTO cell as an initial structure; La atoms are randomly substituted onto A-sites, and corresponding Ti vacancies onto B-sites.
- La-Mg-BTO_o: $Ba_{48}La_{16}Ti_{52}Mg_8O_{184}$
  - Using the Mg-BTO cell as an initial structure; identical Ti vacancy positioning to La-Mg-BTO cell, but with La introduced to form a (4 4 1) plane.

Diagrams of these systems are presented in Appendix D, visualised using the VESTA software package [113]. An example of the La-Mg-BTO system is shown in Figure 5.7.

It should be noted that these simulation cells represent idealised bulk crystalline structures, possessing only the point defects that have been outlined. Realistic oxide crystals are likely to possess extended defects within the microstructure of the material which will provide a variety of paths for diffusing species. These factors were not considered for this work.

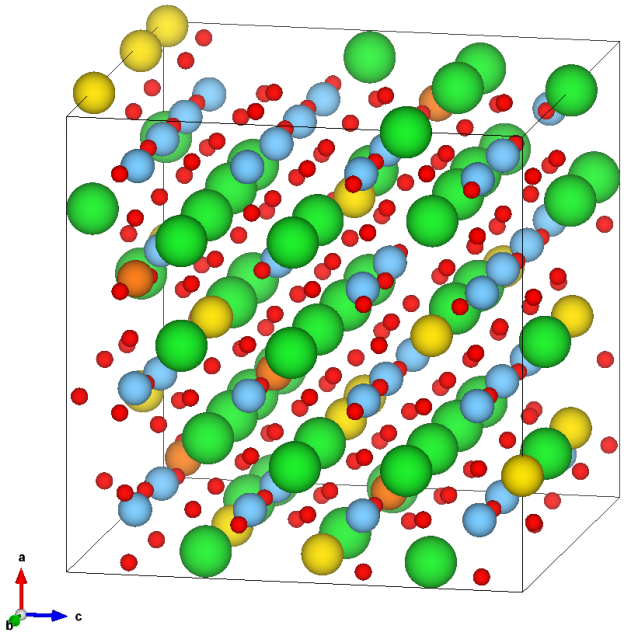


Figure 5.7: An off-axis view of the relaxed La-Mg-BTO system.

5.6.3 Calculations

The systems are initially equilibrated using AIMD as implemented in VASP, using the same energy cut-off value and convergence criteria for energy and ionic relaxation as outlined in Section 4.2.1. Calculated in a canonical NVT ensemble, a Nosé-Hoover thermostat is used to drive the system to, and maintain at, a target temperature which is parameterised by a Nosé mass of 1.0. In all cases, the equations of motion are solved using the Verlet scheme, discretised by a time step of 2 fs, which has been demonstrated to be suitable for capturing oxygen diffusion mechanics in oxides [114, 115]. This is calculated for 1000 steps, over 2 ps, to relax the lattice at the target temperature and establish ion velocities.

Having established ionic dynamism, the final states of these AIMD runs can be carried over and used to initiate MLFF MD simulations. This is run for

50,000 steps, utilising the on-the-fly process of replacing most *ab initio* calculation steps with machine learning trained force-field calculations, whilst updating the structure sample set at each AIMD stage.

The system predicts the energy, forces, and stress tensor, as well as their uncertainties, for each consecutive step using the MLFF process. By comparing to built-in error thresholds, the algorithm decides whether to continue using the current version of the force-field to calculate the above values, or to perform an *ab initio* step; adding the calculated structure to the sample set from which a new force-field is generated to use going forward. At each AIMD step, errors in energy, forces and stress of the force-field are compared to the calculated structure results in the sample set. An example of this for La-Mg-BTO is given in Appendix Table C.1 & C.2. The error in force for all calculations were found to be of the same order of magnitude as those presented in literature [44]. As these systems consist of four/five elements, the training process might be expected to take longer than for simpler systems, and/or yield larger errors in force field construction. Once sufficiently trained, characterising the system at hand, this force-field can be extracted and used in further calculations without the need for further AIMD steps.

After 50,000 steps, outputs are analysed to check that the simulation is running correctly, and is then continued for a further 50,000 steps. A simulation of this length should provide sufficient movement statistics to assess some degree of self-diffusion. In order to analyse the movement of oxygen atoms, ionic coordinates for each elemental species are isolated at each calculation step, allowing for the construction of trajectories for each individual atom. Visual representations of the dynamic paths for oxygen are presented in Appendix E, by means of the VMD software package [116].

5.7 Discussion

5.7.1 Static calculations

Research into the minimum energy path for oxygen ion migration in perovskite structures suggests a curved path around the adjacent B-site cation, which appears to corroborate the nudged elastic band calculations in Section 5.4.1, as well as results in literature [117, 118, 119, 120]. The activation energies for both potential paths are shown to be decreased by the introduction of an adjacent A-site vacancy. Oxide ion diffusion in pure barium titanate has been widely investigated experimentally, giving activation energies in the range of 0.5-1.28 eV [121, 122]. Previous DFT simulations have given migration energies of ≈0.89 eV (discussed in ref. [122]). Despite potential incomparability between experimental results and such rudimentary simulations presented here using NEB, the results are seen to be qualitatively similar to those in literature, which are within the margin of error for these calculations.

Table 5.3 presents the activation energies of NEB calculations for each individual defect case. The difference in energy and reaction coordinate between the systems in Section 5.4.1 and the 3×3×3 scaled equivalents could potentially be as a result of the strain effects across periodic boundary conditions. Further scaling of supercell size would be required to further mitigate any self-interaction between such diffusion events. In a similar result to the smaller cell, the energy barrier is slightly reduced in the presence of a neighbouring barium vacancy. The oxygen vacancy associated with the doped magnesium ion admits a significantly larger energy barrier of 2.03 eV, comparable to values seen in literature for larger systems [123]. A B-site vacancy presents a different geometry of low-energy path, mostly

following the rounded nature of other examples, but instead curves in the other direction closer to the B-site coordinate. This case also presents a local minima at the midpoint of the path suggesting a local potential well, which would be worth further investigation. Of all defects, the lowest energy barrier corresponded to a path in the vicinity of a lanthanum-substituted A-site. In theory, as this admits the most energetically favourable transition path, it could be implied that lanthanum doping of the base barium titanate material may increase oxygen ion movement through the bulk by facilitating diffusion events.

In Section 5.4.3, an investigation into how lanthanum can be positioned around a titanium vacancy was initially motivated by curiosity as to how a high density of these defects may impact the system energy and geometry. The results suggested a slightly lower energy for the (2 2 1) planar configuration, as presented in Figure 5.6. By only introducing asymmetry in one axis direction, it may well be the case that this scenario introduces less stress on the structure due to lattice distortion. This may be worth further investigation, extensively mapping lattice distortion and system energy as a function of doped ion positioning and density.

5.7.2 Dynamic calculations

Whilst the use of MLFF MD in this work has been justified to relieve some of the computational demand of exclusively AIMD calculation steps, there are some caveats to using this method. The force-field is only able to describe structures similar to those collected in its training set. There is a non-trivial concern of reliability if undertrained, as this would bias the model to only the dynamic events that may have been observed up to that stage.

It is also possible that resultant force-fields from such a process may have limited generalisability with regards to further application to similar systems with different geometries. The notion of a force-field is to provide information regarding properties such as interatomic bond lengths, bond angles and torsions and electrostatic interactions, to name but a few. The MLFF will only have captured these details for structures presented throughout training. Separate MLFFs were trained for each case for this reason, despite the identical stoichiometry of the five-species systems, in the event that the interactions local to the lanthanum sites differed for the random distribution or planar geometry.

Initially, this process is parameterised with a target temperature of 1500 K. In Appendices E.1-E.3, it can be seen that oxygen diffusion events are few and far between. The linear profile expected along the time-lag vs. MSD plot is not evident for the La-Mg-BTO system but seems more apparent in the La-Mg-BTO_o and Mg-BTO cases, following the nomenclature in Section 5.6.2. This is likely an issue with the training of the force-field failing to correctly describe oxygen movement. From the trajectory images for all cases, it can be seen that most diffusion events appear to be fairly localised to certain B-sites, and diffusion is not seen throughout the bulk of the simulation cell. Instead, most oxygen movement is naturally shown to be local oxygen vibration. A possible cause being that diffusion events at this temperature may be infrequent, and the training of the force-field rarely captures structures undergoing these dynamic processes. This causes the training to stagnate. Another option might be that the simulation simply has not run long enough to gather statistics on sufficient diffusion events.

The aim of these calculations was to gain qualitative information as to diffusion rates through these cells by use of these methods, rather than

quantitatively exact values. The calculations and training processes were repeated, with the target temperature raised to 2100 K. Increasing the temperature of simulation to observe less common events more frequently is a commonly used technique in molecular dynamics. Introducing more energy to these systems increases the velocities of ions, in theory increasing the chance of diffusion events occurring. With simulations run using the same methodology - equilibrated to the target temperature, and then used to train an MLFF for 100,000 steps - the trajectories presented in Appendices E.4-E.6 reveal a much greater amount of movement throughout the bulk of the cell.

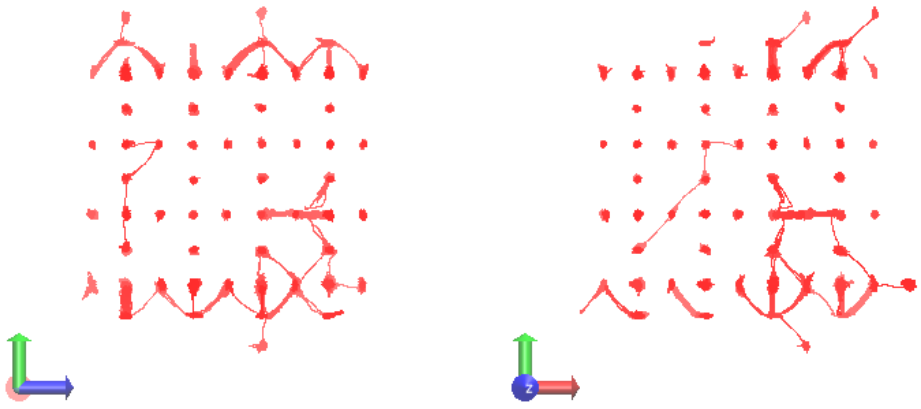
Determining and plotting the time-lag MSD for oxygen species in each cell reveals the characteristic linear regime expected from the data. Truncating the ballistic periods for both small and large time-lags, the diffusion coefficient is determined by calculating the slope using a linear fit between time steps of 25ps and 125ps. These values are presented in Table 5.4, along with the respective coefficients of determination.

	La-Mg-BTO	La-Mg-BTO_o	Mg-BTO
D	1.15e-6	1.29e-6	2.26e-6
R^2	0.997	0.997	0.999

Table 5.4: Coefficient of diffusion (D) and coefficient of determination (R^2) for each MLFF MD simulation at 2100 K, each to three significant figures.

The results suggest that the magnesium doped system facilitates more oxygen diffusion events compared to the co-doped system. These simulations also appear to show that the planar lanthanum substructure facilitates the movement of oxygen through the system more so than the randomly doped system. Whilst the La-Mg-BTO system admits little in the way of order, showing fairly unstructured diffusion paths as per the trajectories in Appendix E.4, the system with ordered lanthanum

atoms suggests oxygen diffusion predominantly throughout this plane. Figures 5.8a & 5.8b, looking along the axes of the lanthanum plane, reveal oxygen diffusion pathways to be mostly concentrated within the lanthanum substructure. This seems to agree with the above static calculations, suggesting such a pathway to be the most energetically favourable amongst the possibilities in this system. The Mg-BTO system admits more widespread oxygen diffusion throughout the simulation cell, with a diffusion coefficient almost double that of the systems co-doped with La. It has regions of localised diffusion around select B-site ions, which does not correspond to magnesium substitution sites. Quantitative comparison to literature results is difficult due to the exaggerated levels of doping in these simulation cells. Whilst such co-doped systems have not been previously simulated, MD results for a barium titanate system with 1% Mg doping have been reported, and suggests activation energies of approximately double that of undoped $BaTiO_3$ [123].



(a) (b)

Figure 5.8: Oxygen trajectories for the La-Mg-BTO_o simulation at 2100 K, viewed along the x-axis and z-axis, respectively.

For further analysis of the system with planar lanthanum geometry, oxygen species MSD data can be determined with respect to each Cartesian axis by simply considering movements projected onto each axis individually. The

determination of separate axial diffusion coefficients can then be used to comment on the isotropy of diffusion in the system. The time-lag vs. MSD plots for each axis are presented in Appendix F. As above, the slope of the linear fit between 25ps and 125ps is used to calculate D. However, it is apparent that the z-axis plot has a non-trivial deviation from a background linear trend between 60ps and 140ps. It is probable that this is due to limited simulation length, and that running the molecular dynamics calculations for a longer period of time would collect more statistics from diffusion events, averaging to the expected form. As an exception, the slope for the z-axis data is evaluated between 25ps and 150ps so as to circumvent this issue, and results are shown in Table 5.5.

	x-axis	y-axis	z-axis
D	2.57e-6	1.65e-6	1.66e-6
R^2	0.999	0.998	0.983

Table 5.5: Coefficient of diffusion (D) and coefficient of determination (R^2) in each axis in the La-Mg-BTO_o system, each to three significant figures.

By recalling the construction of the supercell, the ordered lanthanum positions form an x-z plane. With this, any consequential anisotropy of diffusion might be expected to be characterised by comparable diffusion rates in the x and z axes, with disparity to movement in the y-axis. Here, we instead observe near identical rates of movement in the y and z axes with approximately a 55% increase in movement in the x direction. Considering the geometry of the trajectories admitted for this system, this is perhaps counter-intuitive. It is possible that this simulation has not sufficiently sampled diffusion events and would need to be continued for longer to yield more conclusive statistics regarding the isotropy of diffusion.

Given the uncharacteristic diffusion dynamics admitted by the MLFF MD process at 1500 K, it was not possible to use these results to construct

an Arrhenius plot from MSD data and determine activation energies. This may be possible with a further round of simulations at a higher temperature than 2100 K. However, the training process for this latest set of simulations proved to be vastly memory-intensive, even with the available high-performance computing resources. Considering the size of the simulation cell, and the large amount of unique elements involved, the structure set developed throughout the force field training grew to be very large. Additionally, higher temperatures may require a reduced time step size, as ions move faster. Nonetheless, such memory demand issues pales in comparison to the resources saved compared to AIMD.

5.8 Conclusion

In this chapter, the general methodology for simulation of oxygen self-diffusion in a prototypical perovskite structure is described, from which the diffusion coefficient and activation energy can be determined, and how to approximate the minimum energy path between stable states *via* a transition state. These processes are applied to a nonstoichiometric co-doped barium titanate system inspired by unpublished experimental work.

The activation energy for an oxygen diffusion event near to each individual point defect in this system is calculated, using NEB to determine a low-energy path immediately adjacent to the corresponding A-site or B-site defect in a 3×3×3 supercell. Lanthanum substitution for a barium ion admitted the lowest energy barrier of 0.75 eV, whilst resultant energies are comparable to those found in literature for barium. Oxygen vacancy jumps involving magnesium-oxygen octahedra are calculated to have an activation barrier of 2.03 eV. A titanium, or B-site, vacancy suggested an

intermediary local minima which could be interesting to investigate further.

The positional dependence of lanthanum doping site on system energy is investigated to a limited degree, suggesting an energetically favourable planar configuration which is carried forward to be considered in dynamical simulations. Given the lower calculated energy barrier local to lanthanum, a more exhaustive investigation of lanthanum positioning as a function of system size and doping concentration, and how this might impact oxygen self-diffusion could be of interest.

The MLFF MD method implemented in VASP is used, following initial AIMD equilibration, to simulate the dynamics of oxygen movement through a 4×4×4 simulation cell at 1500 K and 2100 K. Expected diffusion behaviour was not seen at the lower of these temperatures, with analysis of trajectories suggesting the training process to have become biased toward local vibrational motion. Increasing the temperature, thus encouraging more frequent diffusion events, resulted in more characteristic self-diffusion profiles. Mean squared displacement statistics over a range of time-lags was used to calculate diffusion coefficients. Due to difficulties in reliable MLFF MD simulation at multiple temperatures, it was not possible to develop Arrhenius plots and as such calculate activation energy values.

Chapter 6

Conclusions

6.1 Machine learning for discovery of novel hydride storage materials

In an attempt to produce a machine learning model to predict formation enthalpy that would be generalisable to many metal hydride materials, a supervised gradient boosting regression learner was developed, trained on DFT calculated data. This model was validated to a withheld dataset and to a selection of known materials, which are commonly studied experimentally. To use the model to predict new storage materials suitable for operating conditions corresponding to on-board storage applications, a methodology for generating and filtering novel ternary hydride compositions is outlined. An attempt to validate enthalpy predictions for such systems by crystal structure prediction proved difficult as a result of the inherent challenge of such a task.

Determination of thermodynamic properties for these materials is ultimately reliant on sufficient knowledge of the ground state of the

system. Prediction of such quantities by use of a machine learning model requires confidence in the training procedure and the initial data used for development. Alternatively, direct calculation from first principles requires information regarding the crystallographic structure. This work sought to investigate a method to allow for prediction of such characteristics *via* a high-throughput process, reducing the need for exhaustive DFT calculations when sampling a composition space. The developed model proved useful in predicting formation enthalpy values for systems comparable to the training data, that being following a reaction pathway involving the combination of elemental reference crystals.

Difficulty in model validation arises through the generation of structures for calculation of energies. The diversity amongst metal hydride geometries, from interstitial Laves phase based structures to salt-like complex hydrides, proves a challenge for crystal structure prediction, as well as generalisation of a predictive model trained on such varied data. Similar methodological approaches in literature have been shown to find more success in determination of new structures by analogy when investigating material types with more similar, prototypical geometries. The large predictive errors are most likely due to the structure prediction procedure used, and as such should not be used as an indictment against the machine learning model without further analysis. Until this happens, the generalisability of the learner to alternative hydriding pathways cannot be accurately assessed.

6.2 Modelling how defects in barium titanate impact oxygen mobility

By outlining the procedure for determination of oxide ion diffusion in a prototypical perovskite structure, analysis of how the manipulation of system chemistry, in the form of point defects, can impact oxygen self-diffusion rates was presented. This work centres around computational analysis of a lanthanum-magnesium co-doped barium titanate system, inspired by as-yet unpublished experimental results. Activation energies were calculated for oxygen in the vicinity of single instances of each point defect, quantitatively suggesting the existence of energetically favourable diffusion pathways local to a lanthanum dopant. Further analysis is conducted by means of dynamics simulations, using machine learning force field molecular dynamics to alleviate the cost of many *ab initio* calculation steps, whilst considering a selection of doped barium titanate systems.

Machine learning force field molecular dynamics calculations at 1500 K appeared to struggle to provide sufficient statistics of diffusion events to reliably characterise mean squared displacement. These simulations suggest the development of a bias during training towards local vibrational motion, most apparent in the system with a random distribution of the co-doping species, where diffusion events are shown to be infrequent. Increasing the simulation temperature to 2100 K across these systems admits more distinctive self-diffusion behaviour.

A cell constructed with a planar substructure of the lanthanum dopant atoms is shown to have oxygen movement more concentrated to diffusion pathways throughout this plane, as suggested by the energetics of the previous static calculations. However, this apparent anisotropy does not

seem to be represented in MSD analysis along each axis of this system. The diffusion coefficient of a barium titanate system doped solely with magnesium is found to be 75-130% greater than that of the two co-doped systems.

6.3 Future avenues for this work

The intentionally structure-independent approach presented here is perhaps not suitable considering the range of hydride forms. Other machine learning approaches may well prove more effective by incorporating *a priori* information regarding local coordination. However, it should be noted that this would require knowledge of the crystal structure, which is only available for very limited compositions. Local fingerprint descriptor systems which can encode interactions between neighbouring atoms, such as Atom-Centred Symmetry Functions (ACSF) [124] or Smooth Overlap of Atomic Positions (SOAP) [125], used alongside an appropriate algorithm, may result in more accurate and/or more generalised prediction across the chemical environment space of hydride materials.

Developments in reliable and affordable crystal structure prediction methods will allow for more stringent validation practices of similar machine learning models to those presented here. In addition to this, it would facilitate high-throughput calculations for the determination of properties of predicted species. An alternative approach for construction of a predictive model could be to isolate known hydride geometries within training datasets and develop separate models trained on each class. Whilst perhaps more cumbersome to construct, this may prove more effective in generalising to new samples should a structure type be explicitly known or

reasonably intuited. An obvious caveat to this is the scarcity of training data currently available, and as such is a compromise that one cannot afford to make at this time.

As for the perovskite systems studied, it appears that the distribution of lanthanum dopant ions objectively influenced the diffusion pathways between the molecular dynamics simulations of the two co-doped systems. This could be interesting to probe further as to how concentration of doping and distribution of such ions through the lattice might influence system energy and diffusion dynamics.

Inspection of the axial MSD results within the ordered planar lanthanum substructure suggested the potential for further analysis as to possible anisotropic diffusion behaviour. A longer simulation, possibly also at an increased temperature, would likely exhibit more diffusion events, providing more data points to characterise this phenomena. Additionally, running a new set of simulations for all systems at other simulation temperatures would allow for construction of an Arrhenius plot to determine activation energies for diffusion through the cells.

Whilst not resulting in a provably generalisable machine learning model, nor a new best-in-class novel electrolyte material, the findings in this work contribute in the wider sense of the scientific process. By attempting to tackle a technical challenge through a particular method, insight can be gained into alternative, and potentially more effective, approaches to such a task, whilst potentially uncovering new lines of investigation worth further attention. Nonetheless, this has proven to be a worthwhile endeavour.

Bibliography

[1] Nicole Vandaele and Wendell Porter. Renewable energy in developing and developed nations: Outlooks to 2040. *Journal of Undergraduate Research*, 15(3):1–7, 2015.

[2] Seth Dunn. Hydrogen futures: toward a sustainable energy system. *International journal of hydrogen energy*, 27(3):235–264, 2002.

[3] Louis Schlapbach and Andreas Züttel. Hydrogen-storage materials for mobile applications. In *Materials for sustainable energy: a collection of peer-reviewed research and review articles from nature publishing group*, pages 265–270. World Scientific, 2011.

[4] Zainul Abdin, Ali Zafaranloo, Ahmad Rafiee, Walter Mérida, Wojciech Lipiński, and Kaveh R Khalilpour. Hydrogen as an energy vector. *Renewable and sustainable energy reviews*, 120:109620, 2020.

[5] Data Explorer — Climate Watch. https://www.climatewatchdata.org/data-explorer/. Accessed: 2022-09-20.

[6] DOE Technical Targets for Onboard Hydrogen Storage for Light-Duty Vehicles. https://www.energy.gov/eere/fuelcells/doe-technical-targets-onboard-hydrogen-storage-light-duty-vehicles. Accessed: 2022-01-21.

[7] William G Houf, GH Evans, IW Ekoto, EG Merilo, and MA Groethe. Hydrogen fuel-cell forklift vehicle releases in enclosed spaces. *International Journal of Hydrogen Energy*, 38(19):8179–8189, 2013.

[8] Rail, Aviation, and Maritime Metrics. https://www.hydrogen.energy.gov/pdfs/review21/ta034_ahluwalia_2021_o.pdf. Accessed: 2022-09-28.

[9] Vladimir Molkov. Fundamentals of hydrogen safety engineering. *Bookboon. com ISBN*, pages 978–87, 2012.

[10] Ulrich Eberle, Michael Felderhoff, and Ferdi Schueth. Chemical and physical solutions for hydrogen storage. *Angewandte Chemie International Edition*, 48(36):6608–6630, 2009.

[11] Rittmar Von Helmolt and Ulrich Eberle. Fuel cell vehicles: Status 2007. *Journal of Power Sources*, 165(2):833–843, 2007.

[12] Andreas Züttel, Arndt Remhof, Andreas Borgschulte, and Oliver Friedrichs. Hydrogen: the future energy carrier. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 368(1923):3329–3342, 2010.

[13] Anders Andreasen. Predicting formation enthalpies of metal hydrides. *Ris National Laboratory Report*, 2004.

[14] Darren P Broom. *Hydrogen storage materials: the characterisation of their storage properties*, volume 1. Springer, 2011.

[15] J HN Van Vucht, FAr Kuijpers, and H CAM Bruning. Reversible room-temperature absorption of large quantities of hydrogen by intermetallic compounds. *Philips Res. Rep. 25: 133-40 (Apr 1970).*, 1970.

[16] Mykhaylo V Lototskyy, Ivan Tolj, Lydia Pickering, Cordellia Sita, Frano Barbir, and Volodymyr Yartys. The use of metal hydrides in fuel cell applications. *Progress in Natural Science: Materials International*, 27(1):3–20, 2017.

[17] Poojan Modi and Kondo-Francois Aguey-Zinsou. Room temperature metal hydrides for stationary and heat storage applications: A review. *Frontiers in Energy Research*, 9:616115, 2021.

[18] Brian CH Steele and Angelika Heinzel. Materials for fuel-cell technologies. In *Materials for sustainable energy: a collection of peer-reviewed research and review articles from Nature Publishing Group*, pages 224–231. World Scientific, 2011.

[19] Saddam Hussain and Li Yangping. Review of solid oxide fuel cell materials: Cathode, anode, and electrolyte. *Energy Transitions*, 4(2):113–126, 2020.

[20] Zhan Gao, Liliana V Mogni, Elizabeth C Miller, Justin G Railsback, and Scott A Barnett. A perspective on low-temperature solid oxide fuel cells. *Energy & Environmental Science*, 9(5):1602–1644, 2016.

[21] Helmut Mehrer. *Diffusion in solids: fundamentals, methods, materials, diffusion-controlled processes*, volume 155. Springer Science & Business Media, 2007.

[22] C Collins, MS Dyer, MJ Pitcher, GFS Whitehead, M Zanella, P Mandal, JB Claridge, GR Darling, and MJ Rosseinsky. Accelerated discovery of two crystal structure types in a complex inorganic phase field. *Nature*, 546(7657):280, 2017.

[23] Emily A. Carter. Challenges in modeling materials properties without experimental input. *Science*, 321(5890):800–803, 2008.

[24] Mary S Morgan. Experiments versus models: New phenomena, inference and surprise. *Journal of Economic Methodology*, 12(2):317–329, 2005.

[25] Rajan Jose and Seeram Ramakrishna. Materials 4.0: Materials big data enabled materials discovery. *Applied Materials Today*, 10:127–132, 2018.

[26] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, Aug 1996.

[27] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, Oct 2001.

[28] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics)*. Springer, 2016.

[29] Leo Breiman. Out-of-bag estimation, 1996.

[30] Ruoqing Zhu, Donglin Zeng, and Michael R Kosorok. Reinforcement learning trees. *Journal of the American Statistical Association*, 110(512):1770–1784, 2015.

[31] Felix A Faber, Alexander Lindmaa, O Anatole Von Lilienfeld, and Rickard Armiento. Machine learning energies of 2 million elpasolite (ABC_2D_6) crystals. *Physical review letters*, 117(13):135502, 2016.

[32] Jonathan Schmidt, Jingming Shi, Pedro Borlido, Liming Chen, Silvana Botti, and Miguel AL Marques. Predicting the thermodynamic stability of solids combining density functional theory and machine learning. *Chemistry of Materials*, 29(12):5090–5103, 2017.

[33] M. Born and R. Oppenheimer. Zur Quantentheorie der Molekeln. *Annalen der Physik*, 389:457–484, 1927.

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

[35] W. Kohn and L. J. Sham. Self-Consistent Equations Including Exchange and Correlation Effects. *Physical Review*, 140:1133–1138, November 1965.

[36] John P Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized gradient approximation made simple. *Physical review letters*, 77(18):3865, 1996.

[37] Peter E Blöchl. Projector augmented-wave method. *Physical review B*, 50(24):17953, 1994.

[38] Yanchao Wang, Jian Lv, Li Zhu, and Yanming Ma. Crystal structure prediction via particle-swarm optimization. *Physical Review B*, 82(9):094116, 2010.

[39] Yanchao Wang, Jian Lv, Li Zhu, and Yanming Ma. CALYPSO: A method for crystal structure prediction. *Computer Physics Communications*, 183(10):2063–2070, 2012.

[40] Nosé Shuichi. A unified formulation of the constant temperature molecular dynamics methods. *The Journal of chemical physics*, 81(1):511–519, 1984.

[41] William G Hoover. Canonical dynamics: Equilibrium phase-space distributions. *Physical review A*, 31(3):1695, 1985.

[42] Nosé Shuichi. Constant temperature molecular dynamics methods. *Progress of Theoretical Physics Supplement*, 103:1–46, 1991.

[43] Ryosuke Jinnouchi, Jonathan Lahnsteiner, Ferenc Karsai, Georg Kresse, and Menno Bokdam. Phase transitions of hybrid perovskites simulated by machine-learning force fields trained on the fly with Bayesian inference. *Physical review letters*, 122(22):225701, 2019.

[44] Ryosuke Jinnouchi, Ferenc Karsai, and Georg Kresse. On-the-fly machine learning force field generation: Application to melting points. *Physical Review B*, 100(1):014105, 2019.

[45] Ryosuke Jinnouchi, Ferenc Karsai, Carla Verdi, Ryoji Asahi, and Georg Kresse. Descriptors representing two-and three-body atomic distributions and their effects on the accuracy of machine-learned inter-atomic potentials. *The Journal of Chemical Physics*, 152(23):234102, 2020.

[46] Vincent Berube, Gregg Radtke, Mildred Dresselhaus, and Gang Chen. Size effects on the hydrogen storage properties of nanostructured metal hydrides: A review. *International Journal of Energy Research*, 31(6-7):637–663, 2007.

[47] Gary Sandrock. A panoramic overview of hydrogen storage alloys from a gas reaction point of view. *Journal of alloys and compounds*, 293:877–888, 1999.

[48] AR Miedema, PF De Chatel, and FR De Boer. Cohesion in alloys—fundamentals of a semi-empirical model. *Physica B+ c*, 100(1):1–28, 1980.

[49] Frank R De Boer, W Mattens, R Boom, AR Miedema, and AK Niessen. Cohesion in metals. Transition metal alloys. 1988.

[50] AR Miedema. The electronegativity parameter for transition metals:

heat of formation and charge transfer in alloys. *Journal of the less common metals*, 32(1):117–136, 1973.

[51] KHJ Buschow, PCP Bouten, and AR Miedema. Hydrides formed from intermetallic compounds of two transition metals: a special class of ternary alloys. *Reports on progress in physics*, 45(9):937, 1982.

[52] Katsuhiko Hirose. *Handbook of hydrogen storage: new materials for future energy storage*. John Wiley & Sons, 2010.

[53] Jason R Hattrick-Simpers, Kamal Choudhary, and Claudio Corgnale. A simple constrained machine learning model for predicting high-pressure-hydrogen-compressor materials. *Molecular Systems Design & Engineering*, 3(3):509–517, 2018.

[54] Hydrogen Storage Materials Database. http://hydrogenmaterialssearch.govtools.us/. Accessed: 2022-08-29.

[55] Alireza Rahnama, Guilherme Zepon, and Seetharaman Sridhar. Machine learning based prediction of metal hydrides for hydrogen storage, part I: Prediction of hydrogen weight percent. *International Journal of Hydrogen Energy*, 44(14):7337–7344, 2019.

[56] Alireza Rahnama, Guilherme Zepon, and Seetharaman Sridhar. Machine learning based prediction of metal hydrides for hydrogen storage, part II: Prediction of material class. *International Journal of Hydrogen Energy*, 44(14):7345–7353, 2019.

[57] Matthew Witman, Sanliang Ling, David M Grant, Gavin S Walker, Sapan Agarwal, Vitalie Stavila, and Mark D Allendorf. Extracting an empirical intermetallic hydride design principle from limited data via

interpretable machine learning. *The Journal of Physical Chemistry Letters*, 11(1):40–47, 2019.

[58] G Sandrock and G Thomas. The IEA/DOE/SNL on-line hydride databases. *Applied Physics A*, 72(2):153–155, 2001.

[59] 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. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[60] Logan Ward, Ankit Agrawal, Alok Choudhary, and Christopher Wolverton. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Materials*, 2(1):1–7, 2016.

[61] scikit-optimize: Sequential model-based optimization in Python. https://scikit-optimize.github.io/stable/.
Accessed: 2022-09-03.

[62] Atsuto Seko, Atsushi Togo, and Isao Tanaka. Descriptors for machine learning of materials data. In *Nanoinformatics*, pages 3–23. Springer, Singapore, 2018.

[63] Machine learning descriptors for molecules. https://chemintelligence.com/blog/machine-learning-descriptors-molecules.
Accessed: 2022-07-21.

[64] Valentin Stanev, Corey Oses, A Gilad Kusne, Efrain Rodriguez, Johnpierre Paglione, Stefano Curtarolo, and Ichiro Takeuchi.

Machine learning modeling of superconducting critical temperature. *npj Computational Materials*, 4(1):1–14, 2018.

[65] Fang Ren, Logan Ward, Travis Williams, Kevin J Laws, Christopher Wolverton, Jason Hattrick-Simpers, and Apurva Mehta. Accelerated discovery of metallic glasses through iteration of machine learning and high-throughput experiments. *Science advances*, 4(4):eaaq1566, 2018.

[66] Logan Ward, Ruoqian Liu, Amar Krishna, Vinay I Hegde, Ankit Agrawal, Alok Choudhary, and Chris Wolverton. Including crystal structure attributes in machine learning models of formation energies via Voronoi tessellations. *Physical Review B*, 96(2):024104, 2017.

[67] James E. Saal, Scott Kirklin, Muratahan Aykol, Bryce Meredig, and C. Wolverton. Materials design and discovery with high-throughput density functional theory: The open quantum materials database (oqmd). *JOM*, 65(11):1501–1509, Nov 2013.

[68] Scott Kirklin, James E Saal, Bryce Meredig, Alex Thompson, Jeff W Doak, Muratahan Aykol, Stephan Rühl, and Chris Wolverton. The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. *npj Computational Materials*, 1:15010, 2015.

[69] OQMD - Structure Sources.
https://oqmd.org/documentation/structures.
Accessed: 2022-07-26.

[70] Guenter Bergerhoff, R Hundt, R Sievers, and ID Brown. The inorganic crystal structure data base. *Journal of chemical information and computer sciences*, 23(2):66–69, 1983.

[71] Alec Belsky, Mariette Hellenbrandt, Vicky Lynn Karen, and Peter Luksch. New developments in the Inorganic Crystal Structure Database (ICSD): accessibility in support of materials research and design. *Acta Crystallographica Section B: Structural Science*, 58(3):364–369, 2002.

[72] qmpy - qmpy v1.2.0 documentation. https://static.oqmd.org/static/docs/index.html. Accessed: 2022-08-09.

[73] 1. Supervised learning - scikit-learn 1.1.1 documentation. https://scikit-learn.org/stable/supervised_learning.html. Accessed: 2022-07-28.

[74] Zhang Yang, Yang Fu-sheng, Zhao Feng-qi, and Xu Si-yu. Interaction Mechanism between Metal Hydrides and Energetic Compounds: an Extensive Literature Survey. *FirePhysChem*, 2021.

[75] Kandavel Manickam, Priyen Mistry, Gavin Walker, David Grant, Craig E Buckley, Terry D Humphries, Mark Paskevicius, Torben Jensen, Rene Albert, Kateryna Peinecke, et al. Future perspectives of thermal energy storage with metal hydrides. *International Journal of Hydrogen Energy*, 44(15):7738–7745, 2019.

[76] John J Vajo, Florian Mertens, Channing C Ahn, Robert C Bowman Jr, and Brent Fultz. Altering hydrogen storage properties by hydride destabilization through alloy formation: LiH and MgH_2 destabilized with Si. *The Journal of Physical Chemistry B*, 108(37):13977–13983, 2004.

[77] Logan Ward, Alexander Dunn, Alireza Faghaninia, Nils ER Zimmermann, Saurabh Bajaj, Qi Wang, Joseph Montoya, Jiming

Chen, Kyle Bystrom, Maxwell Dylla, et al. Matminer: An open source toolkit for materials data mining. *Computational Materials Science*, 152:60–69, 2018.

[78] Jack D Evans and François-Xavier Coudert. Predicting the mechanical properties of zeolite frameworks by machine learning. *Chemistry of Materials*, 29(18):7833–7839, 2017.

[79] Takashi Toyao, Keisuke Suzuki, Shoma Kikuchi, Satoru Takakusagi, Ken-ichi Shimizu, and Ichigaku Takigawa. Toward effective utilization of methane: machine learning prediction of adsorption energies on metal alloys. *The Journal of Physical Chemistry C*, 122(15):8315–8326, 2018.

[80] Ann M Deml, Ryan O'Hayre, Chris Wolverton, and Vladan Stevanović. Predicting density functional theory total energies and enthalpies of formation of metal-nonmetal compounds by linear regression. *Physical Review B*, 93(8):085142, 2016.

[81] Felix A Faber, Luke Hutchison, Bing Huang, Justin Gilmer, Samuel S Schoenholz, George E Dahl, Oriol Vinyals, Steven Kearnes, Patrick F Riley, and O Anatole Von Lilienfeld. Prediction errors of molecular machine learning models lower than hybrid DFT error. *Journal of chemical theory and computation*, 13(11):5255–5264, 2017.

[82] Daniel W Davies, Keith T Butler, and Aron Walsh. Data-driven discovery of photoactive quaternary oxides using first-principles machine learning. *Chemistry of Materials*, 31(18):7221–7230, 2019.

[83] Jino Im, Seongwon Lee, Tae-Wook Ko, Hyun Woo Kim, YunKyong Hyon, and Hyunju Chang. Identifying Pb-free perovskites for solar cells by machine learning. *npj Computational Materials*, 5(1):1–8, 2019.

[84] Gavin Walker. *Solid-state hydrogen storage: materials and chemistry*. Elsevier, 2008.

[85] P Villars, K Cenzual, J Daams, Y Chen, and S Iwata. Data-driven atomic environment prediction for binaries using the Mendeleev number: Part 1. Composition AB. *Journal of alloys and compounds*, 367(1-2):167–175, 2004.

[86] D Hobbs, G Kresse, and J Hafner. Fully unconstrained noncollinear magnetism within the projector augmented-wave method. *Physical Review B*, 62(17):11556, 2000.

[87] MPAT Methfessel and AT Paxton. High-precision sampling for brillouin-zone integration in metals. *Physical Review B*, 40(6):3616, 1989.

[88] ARCHER2.
https://www.archer2.ac.uk/.
Accessed: 2022-09-28.

[89] Matthew Witman, Gustav Ek, Sanliang Ling, Jeffery Chames, Sapan Agarwal, Justin Wong, Mark D Allendorf, Martin Sahlberg, and Vitalie Stavila. Data-driven discovery and synthesis of high entropy alloy hydrides with targeted thermodynamic stability. *Chemistry of Materials*, 33(11):4067–4076, 2021.

[90] Yoshitsugu Kojima and Masakuni Yamaguchi. Investigation on hydrogen dissociation pressure, heat of formation and strain energy of metal hydrides. *Journal of Alloys and Compounds*, 840:155686, 2020.

[91] Chiara Milanese, Sebastiano Garroni, Fabiana Gennari, Amedeo Marini, Thomas Klassen, Martin Dornheim, and Claudio Pistidda.

Solid state hydrogen storage in alanates and alanate-based compounds: A review. *Metals*, 8(8):567, 2018.

[92] Tammy P Taylor, Mei Ding, Deborah S Ehler, Trudi M Foreman, John P Kaszuba, and Nancy N Sauer. Beryllium in the environment: a review. *Journal of Environmental Science and Health, Part A*, 38(2):439–469, 2003.

[93] Kyuhyun Sim, Ram Vijayagopal, Namdoo Kim, and Aymeric Rousseau. Optimization of component sizing for a fuel cell-powered truck to minimize ownership cost. *Energies*, 12(6):1125, 2019.

[94] Materials-Based Hydrogen Storage — Department of Energy. https://www.energy.gov/eere/fuelcells/materials-based-hydrogen-storage.
Accessed: 2022-09-20.

[95] AI Kitaigorodskii. Organic Crystal Chemistry. *Izd. Akad. Nauk SSSR, Moscow*, page 15, 1955.

[96] A Io Kitaigorodskii. Molecular crystals, 1971.

[97] Wenhao Sun, Christopher J Bartel, Elisabetta Arca, Sage R Bauers, Bethany Matthews, Bernardo Orvañanos, Bor-Rong Chen, Michael F Toney, Laura T Schelhas, William Tumas, et al. A map of the inorganic ternary metal nitrides. *Nature materials*, 18(7):732–739, 2019.

[98] Colin W Glass, Artem R Oganov, and Nikolaus Hansen. Uspex—evolutionary crystal structure prediction. *Computer physics communications*, 175(11-12):713–720, 2006.

[99] Chris J Pickard and RJ Needs. Ab initio random structure searching. *Journal of Physics: Condensed Matter*, 23(5):053201, 2011.

[100] Jingzhao Zhang, Kinfai Tse, Manhoi Wong, Yiou Zhang, and Junyi Zhu. A brief review of co-doping. *Frontiers of Physics*, 11(6):1–21, 2016.

[101] Ingrid Denk, Wolfram Münch, and Joachim Maier. Partial conductivities in $SrTiO_3$: bulk polarization experiments, oxygen concentration cell measurements, and defect-chemical modeling. *Journal of the American Ceramic Society*, 78(12):3265–3272, 1995.

[102] MM Vijatović Petrović, JD Bobić, T Ramoška, J Banys, and Biljana D Stojanović. Electrical properties of lanthanum doped barium titanate ceramics. *Materials characterization*, 62(10):1000–1006, 2011.

[103] Mirjana Vijatović, Jelena Bobić, and Biljana D Stojanović. History and challenges of barium titanate: Part II. *Science of Sintering*, 40(3):235–244, 2008.

[104] Pushkar Kanhere and Zhong Chen. A review on visible light active perovskite-based photocatalysts. *Molecules*, 19(12):19995–20022, 2014.

[105] Hannes Jónsson, Greg Mills, and Karsten W Jacobsen. Nudged elastic band method for finding minimum energy paths of transitions. 1998.

[106] Nudged Elastic Band - RheoMan. https://umet.univ-lille.fr/Projets/RheoMan/en/to-learn-more-about/nudged-elastic-band.php.html. Accessed: 2022-09-27.

[107] Graeme Henkelman, Blas P Uberuaga, and Hannes Jónsson. A climbing image nudged elastic band method for finding saddle points

and minimum energy paths. *The Journal of chemical physics*, 113(22):9901–9904, 2000.

[108] GH Kwei, AC Lawson, SJL Billinge, and SW Cheong. Structures of the ferroelectric phases of barium titanate. *The Journal of Physical Chemistry*, 97(10):2368–2377, 1993.

[109] Boštjan Zalar, Valentin V Laguta, and Robert Blinc. NMR evidence for the coexistence of order-disorder and displacive components in barium titanate. *Physical review letters*, 90(3):037601, 2003.

[110] Manuel I Marqués. First-principles study of instantaneous and averaged local potential in $BaTiO_3$. *Physical Review B*, 71(17):174116, 2005.

[111] Qingsong Zhang, Tahir Cagin, and William A Goddard. The ferroelectric and cubic phases in $BaTiO_3$ ferroelectrics are also antiferroelectric. *Proceedings of the National Academy of Sciences*, 103(40):14695–14700, 2006.

[112] VTSTTools 3.1 - The Transition State Tools implementation for VASP can be obtained from.
http://theory.cm.utexas.edu/vtsttools/.
Accessed: 2022-08-18.

[113] Koichi Momma and Fujio Izumi. VESTA: a three-dimensional visualization system for electronic and structural analysis. *Journal of Applied crystallography*, 41(3):653–658, 2008.

[114] Michael William Donald Cooper, RW Grimes, Michael E Fitzpatrick, and Alexander Chroneos. Modeling oxygen self-diffusion in UO2 under pressure. *Solid State Ionics*, 282:26–30, 2015.

[115] SR G Christopoulos, A Kordatos, Michael William D Cooper, Michael E Fitzpatrick, and A Chroneos. Activation volumes of oxygen self-diffusion in fluorite structured oxides. *Materials Research Express*, 3(10):105504, 2016.

[116] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD: visual molecular dynamics. *Journal of molecular graphics*, 14(1):33–38, 1996.

[117] M Cherry, M Saiful Islam, and CRA Catlow. Oxygen ion migration in perovskite-type oxides. *Journal of Solid State Chemistry*, 118(1):125–132, 1995.

[118] AV Petrov, SC Parker, and Armin Reller. Computer simulation of the oxygen mobility in $CaMnO_{3-x}$. *Phase Transitions*, 55(1-4):229–244, 1995.

[119] Scott M Woodley, Julian D Gale, Peter D Battle, and C Richard A Catlow. Oxygen ion migration in orthorhombic $LaMnO_{3-\delta}$. *The Journal of chemical physics*, 119(18):9737–9744, 2003.

[120] M Saiful Islam. Ionic transport in ABO3 perovskite oxides: a computer modelling tour. *Journal of Materials chemistry*, 10(4):1027–1038, 2000.

[121] J Maier, G Schwitzgebel, and H-J Hagemann. Electrochemical investigations of conductivity and chemical diffusion in pure and doped cubic $SrTiO_3$ and $BaTiO_3$. *Journal of Solid State Chemistry*, 58(1):1–13, 1985.

[122] Markus Kessel, Roger A De Souza, and Manfred Martin. Oxygen diffusion in single crystal barium titanate. *Physical Chemistry Chemical Physics*, 17(19):12587–12597, 2015.

[123] Wolfgang Preis. Molecular dynamics simulations of oxygen diffusion in barium titanate doped with Mg and Ca. *Journal of Solid State Chemistry*, page 123290, 2022.

[124] Jörg Behler. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *The Journal of chemical physics*, 134(7):074106, 2011.

[125] Albert P Bartók, Risi Kondor, and Gábor Csányi. On representing chemical environments. *Physical Review B*, 87(18):184115, 2013.

[126] kgrid: Calculate the required k-point density from the input geometry for periodic quantum chemistry calculations.
https://github.com/WMD-group/kgrid.
Accessed: 2022-08-18.

Appendices

Appendix A

VASP input files

Essential input files for a VASP calculation are POSCAR, POTCAR, and INCAR. Some further control of calculations is provided with an additional KPOINTS file.

POSCAR: This is the file used for system geometry information. Easily converted to from other crystallographic data representations, such as commonplace CIF files, this is the input for lattice geometry, ionic positions and, optionally, starting velocities. The starting geometry for calculations, it shares a filetype with the CONTCAR output file, enabling easy continuation of calculations and analysis of final structures.

POTCAR: Pseudopotentials for each atomic species in the target compound, as provided by the VASP PAW_PBE file repository, must be concatenated into this file. Ensuring they are in the same order and the elements are declared in the POSCAR file, this will facilitate the plane augmented wave calculation method used in VASP (see Section 2.6.2).

INCAR: The main input file for VASP, the INCAR file is used to parameterise the calculation process. A wide array of tags can be called and corresponding values given will allow for tuning of the entire process, including algorithm selection and settings. For many, the correct parameterisation of this file is essential for an accurate depiction of the physics involved (e.g. magnetism). There are many important tags for particular use cases, which will be introduced in this work when relevant.

KPOINTS: More intuitively named, this optional file can be used to specify the mesh density of the k-point grid used to sample the Brillouin zone (see Section 2.6.1). A number of subdivisions can

be specified along each reciprocal lattice vector. For larger periodic systems, a reduced number of k-points may be used which can help to ease the taxing calculations of ever-larger systems (as mentioned in Section 5.6.1). In most cases in this work, a gamma centred k-point grid is generated for a corresponding POSCAR by means of the *kgrid* python package [126].

Appendix B

Final set of non-metastable filtered generated ternary compositions in Section 4.5

a	b	c	d	e	f
Co1 H5 V2	-38.245	3.04	Co1 V2	-0.141	62.50
Co3 H17 V7	-38.366	3.11	Co3 V7	-0.156	62.96
Co4 H22 V9	-38.413	3.10	Co4 V9	-0.148	62.86
Co2 H16 V7	-39.789	3.29	Co2 V7	-0.154	64.00
Co3 H23 V10	-39.864	3.27	Co3 V10	-0.156	63.89
Co1 H7 V3	-40.137	3.22	Co1 V3	-0.168	63.64
Co1 H9 V4	-40.780	3.34	Co1 V4	-0.151	64.29
Co3 H19 V8	-40.803	3.17	Co3 V8	-0.139	63.33
Co2 H12 V5	-41.300	3.14	Co2 V5	-0.158	63.16
H25 Mg9 Ni7	-41.743	3.85	Mg9 Ni7	-0.114	60.98
Co1 H11 V5	-42.629	3.41	Co1 V5	-0.125	64.71
Co2 H20 V9	-42.987	3.38	Co2 V9	-0.137	64.52
H19 Mg7 Ni5	-43.653	3.97	Mg7 Ni5	-0.112	61.29
Co1 H13 V6	-43.699	3.47	Co1 V6	-0.116	65.00
Co1 H15 V7	-44.110	3.51	Co1 V7	-0.117	65.22
Co1 H17 V8	-44.130	3.54	Co1 V8	-0.111	65.38
Co1 H19 V9	-44.185	3.57	Co1 V9	-0.106	65.52
H28 Mg9 Ni10	-44.628	3.38	Mg9 Ni10	-0.252	59.57
H3 Mg1 Ni1	-44.681	3.52	Mg1 Ni1	-0.168	60.00
H14 Mg5 Ni4	-45.143	3.81	Mg5 Ni4	-0.123	60.87
H17 Mg6 Ni5	-45.301	3.75	Mg6 Ni5	-0.140	60.71
H27 Mg10 Ni7	-45.327	4.00	Mg10 Ni7	-0.119	61.36
H20 Mg7 Ni6	-45.606	3.72	Mg7 Ni6	-0.140	60.61
H23 Mg8 Ni7	-45.708	3.69	Mg8 Ni7	-0.149	60.53
H25 Mg8 Ni9	-46.452	3.37	Mg8 Ni9	-0.251	59.52
H26 Mg9 Ni8	-46.700	3.67	Mg9 Ni8	-0.148	60.47
H29 Mg10 Ni9	-46.851	3.65	Mg10 Ni9	-0.146	60.42
H13 Mg4 Ni5	-47.633	3.25	Mg4 Ni5	-0.233	59.09
H16 Mg5 Ni6	-47.852	3.29	Mg5 Ni6	-0.252	59.26
H22 Mg9 Ni4	-47.992	4.66	Mg9 Ni4	-0.122	62.86
H17 Mg7 Ni3	-48.109	4.72	Mg7 Ni3	-0.127	62.96
H19 Mg6 Ni7	-48.126	3.33	Mg6 Ni7	-0.253	59.38
H22 Mg7 Ni8	-48.501	3.35	Mg7 Ni8	-0.254	59.46
H7 Ni1 V3	-48.799	3.23	Ni1 V3	-0.111	63.64
H23 Mg7 Ni9	-49.671	3.21	Mg7 Ni9	-0.240	58.97
H10 Mg3 Ni4	-49.816	3.17	Mg3 Ni4	-0.238	58.82

Table B.1: (a) Ternary composition, (b) Predicted hydride ΔH_f, kJ/mol$_{H_2}$, (c) Hydrogen wt%, (d) Intermetallic component of hydride, (e) Predicted alloy ΔH_f, eV/atom, (f) H/M ratio

Appendix C

MLFF error log for each *ab initio* calculated step

nstep	rmse_energy, eV/atom	rmse_force, eV/Å	rmse_stress, kB
1	4.73169189E-02	3.14446610E-01	1.77266159E+02
2	1.37872159E-04	2.18034148E-01	7.93165204E-02
3	2.25198327E-04	1.69173053E-01	7.85855691E-02
4	2.81093562E-04	1.56099813E-01	1.13375647E-01
5	2.74662032E-04	1.39031700E-01	1.60515729E-01
6	2.54792294E-04	1.25941339E-01	1.66408840E-01
7	2.42860424E-04	1.18297887E-01	1.68840157E-01
8	2.75115184E-04	1.10916523E-01	1.61964420E-01
9	3.02245071E-04	1.09519022E-01	1.58200538E-01
10	3.12625338E-04	1.08970021E-01	1.44062372E-01
11	2.91008889E-04	1.08037064E-01	1.35012779E-01
21	8.80999242E-04	1.28452311E-01	2.57049575E-01
40	8.25760019E-04	1.41961356E-01	4.06355956E-01
47	8.54961325E-04	1.47289428E-01	3.91209924E-01
61	1.12160397E-03	1.54837219E-01	4.62571380E-01
71	9.38580781E-04	1.59019400E-01	5.63018262E-01
93	1.04425249E-03	1.63619651E-01	6.29525416E-01
102	1.12114725E-03	1.66678387E-01	7.05369289E-01
152	1.27852124E-03	1.75728133E-01	7.70140788E-01
202	1.12231284E-03	1.80186738E-01	8.71223128E-01
261	1.39298205E-03	1.83680294E-01	9.22509983E-01
324	1.39155711E-03	1.84820036E-01	9.72894079E-01
404	1.42183905E-03	1.86443856E-01	9.96348685E-01
656	1.50161533E-03	1.89804412E-01	1.03809682E+00
849	1.54840413E-03	1.90873433E-01	1.04875258E+00
1070	1.54482945E-03	1.92707916E-01	1.11519571E+00
1408	1.51137816E-03	1.92741448E-01	1.10143455E+00
1950	1.51296327E-03	1.93406152E-01	1.10312126E+00
2232	1.49047395E-03	1.93680083E-01	1.11608396E+00
3255	1.49047785E-03	1.94723147E-01	1.12772442E+00
3547	1.44072264E-03	1.95557927E-01	1.15962545E+00
5286	1.52377707E-03	1.95937932E-01	1.14839875E+00
5937	1.57701849E-03	1.95503335E-01	1.13909981E+00
7468	1.56183191E-03	1.95581368E-01	1.13491865E+00
12571	1.55527814E-03	1.95490484E-01	1.13431880E+00
17501	1.57342339E-03	1.95340998E-01	1.12323740E+00
20778	1.56438543E-03	1.96280071E-01	1.12587030E+00
22193	1.57716618E-03	1.95625620E-01	1.11940999E+00
30565	1.57094541E-03	1.95189770E-01	1.11921366E+00
40270	1.54914127E-03	1.94824309E-01	1.11686875E+00
50000	1.50705631E-03	1.95012580E-01	1.11382422E+00

Table C.1: Error log for first round of MLFF for La-Mg-BTO at 1500K.

nstep	rmse_energy, eV/atom	rmse_force, eV/Å	rmse_stress, kB
0	1.50705658E-03	1.95012550E-01	1.11382071E+00
1	1.52589293E-03	1.88656424E-01	1.07754306E+00
2	1.52697885E-03	1.85467396E-01	1.06194167E+00
3	1.52385938E-03	1.84034496E-01	1.05517727E+00
4	1.52397665E-03	1.83066625E-01	1.04587314E+00
5	1.52062205E-03	1.82005883E-01	1.03448070E+00
6	1.51016703E-03	1.81013266E-01	1.02533324E+00
7	1.49666943E-03	1.79887688E-01	1.01561553E+00
8	1.46015424E-03	1.78873983E-01	1.00567805E+00
9	1.43634192E-03	1.77865618E-01	9.95133616E-01
10	1.42549891E-03	1.77138793E-01	9.93294624E-01
31	1.41823121E-03	1.77059425E-01	9.89971930E-01
81	1.44743593E-03	1.77322644E-01	9.88979487E-01
141	1.42053002E-03	1.77716588E-01	9.91773121E-01
191	1.43733562E-03	1.77722331E-01	9.97854174E-01
241	1.42530230E-03	1.77692683E-01	1.00755784E+00
296	1.39005944E-03	1.77579757E-01	1.00876095E+00
372	1.37532908E-03	1.77524093E-01	1.01228600E+00
646	1.35955201E-03	1.77946322E-01	1.02075959E+00
784	1.36211131E-03	1.77936763E-01	1.02546237E+00
1193	1.35755999E-03	1.77688822E-01	1.02365805E+00
1555	1.34237633E-03	1.77542266E-01	1.01788241E+00
2191	1.35497896E-03	1.77735633E-01	1.02263265E+00
2682	1.36868091E-03	1.77918735E-01	1.03037482E+00
3490	1.36503010E-03	1.78093442E-01	1.03314768E+00
5884	1.36528884E-03	1.77760308E-01	1.03244406E+00
11879	1.38029529E-03	1.77874340E-01	1.04059158E+00
13247	1.37365458E-03	1.77957232E-01	1.04431868E+00
14215	1.36188634E-03	1.77973204E-01	1.04162864E+00
20578	1.35644301E-03	1.77950724E-01	1.03988896E+00
29212	1.35157586E-03	1.77781529E-01	1.03842597E+00
32094	1.34565245E-03	1.77474202E-01	1.03970096E+00
37533	1.36006178E-03	1.77423052E-01	1.04144157E+00
41761	1.34584699E-03	1.77529268E-01	1.04480398E+00
49914	1.34911407E-03	1.77645336E-01	1.04621018E+00

Table C.2: Error log for second round of MLFF for La-Mg-BTO at 1500K.

Appendix D

Supercell structures for molecular dynamics calculations in Section 5.6

Starting structures used for molecular dynamics calculations following initial geometry optimisation, visualised using VESTA [113]. Ba (green), Ti (blue), O (red), Mg (orange), La (yellow).

D.1 Mg-BTO: $Ba_{64}Ti_{56}Mg_8O_{184}$

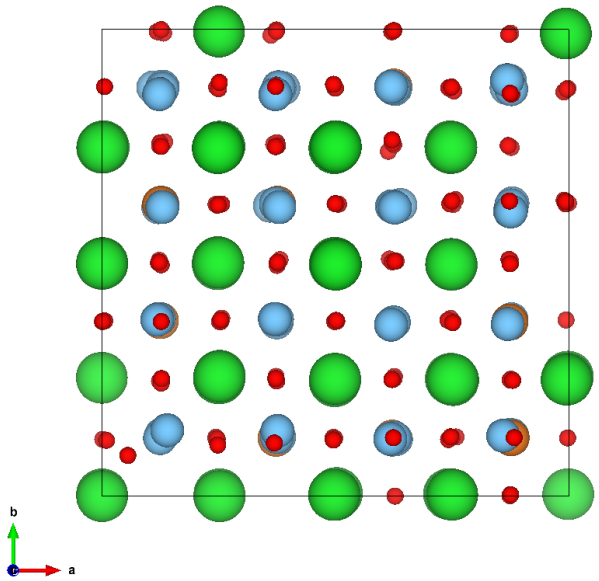


Figure D.1: Mg-BTO: a-b plane.

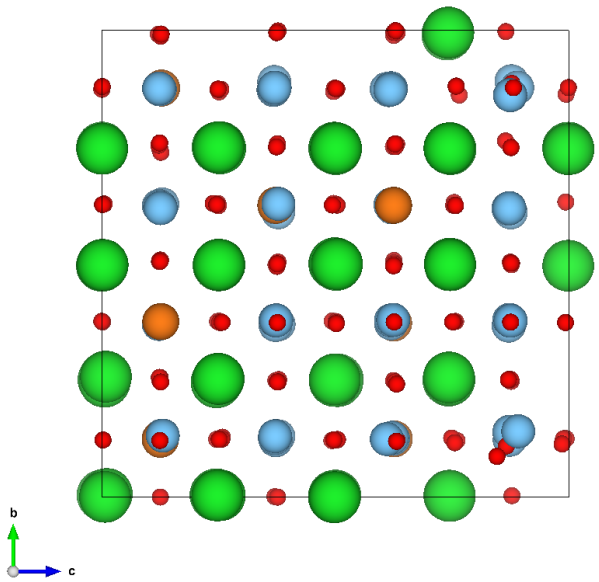


Figure D.2: Mg-BTO: b-c plane.

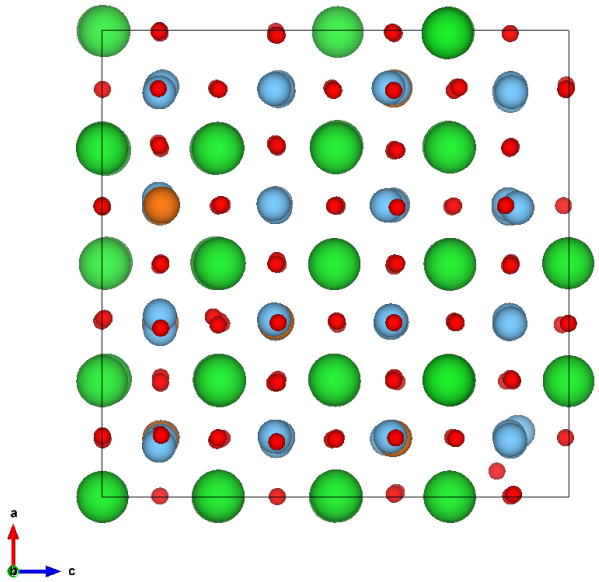


Figure D.3: Mg-BTO: a-c plane.

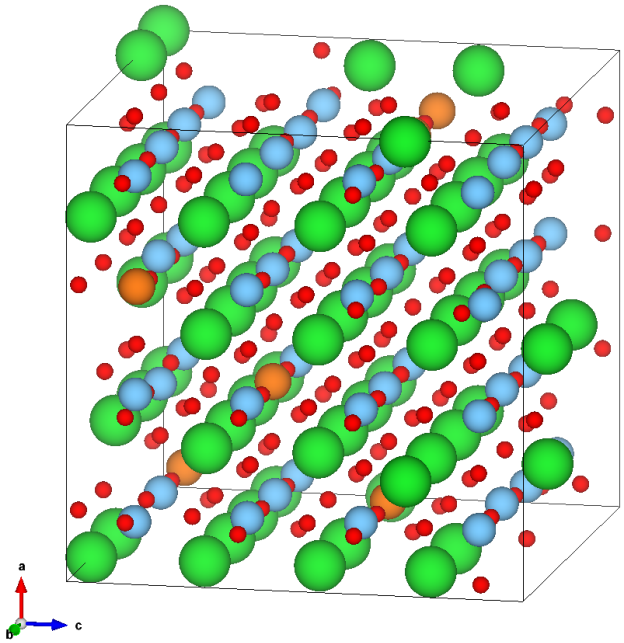


Figure D.4: Mg-BTO: off-axis.

D.2 La-Mg-BTO: $Ba_{48}La_{16}Ti_{52}Mg_8O_{184}$

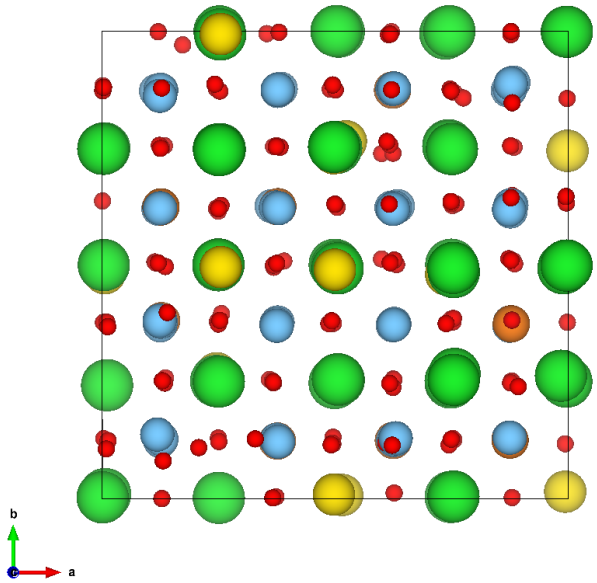


Figure D.5: La-Mg-BTO: a-b plane.

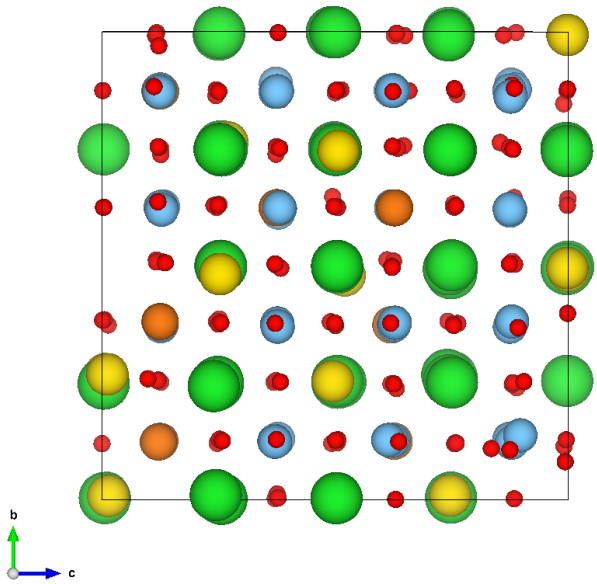


Figure D.6: La-Mg-BTO: b-c plane.

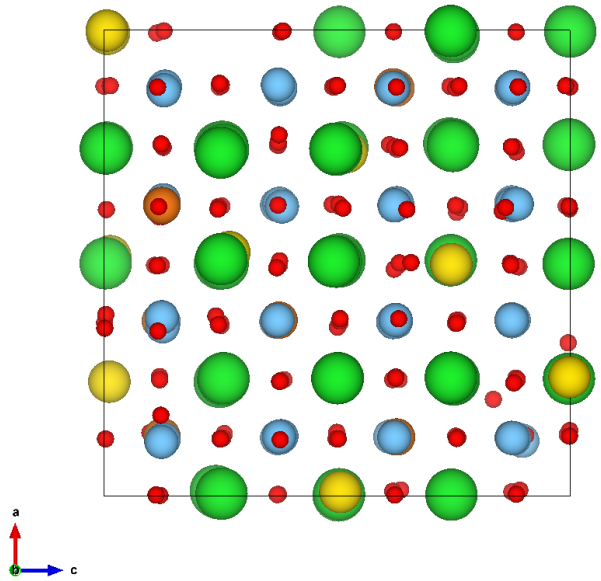


Figure D.7: La-Mg-BTO: a-c plane.

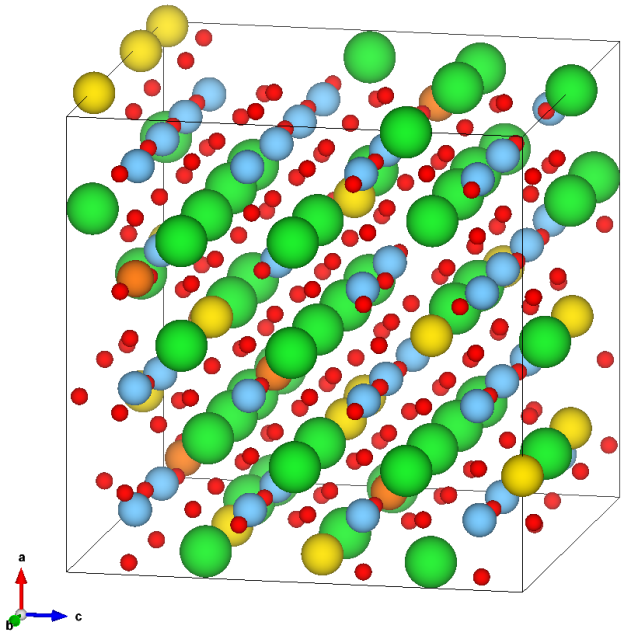


Figure D.8: La-Mg-BTO: off-axis.

D.3 La-Mg-BTO_o: $Ba_{48}La_{16}Ti_{52}Mg_8O_{184}$

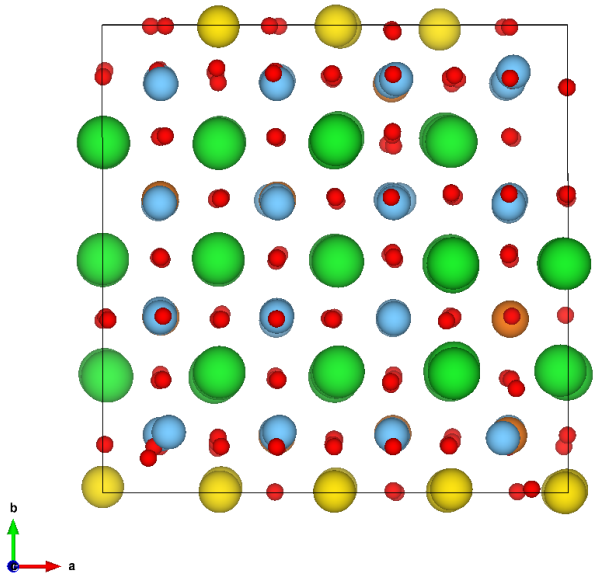


Figure D.9: La-Mg-BTO_o: a-b plane.

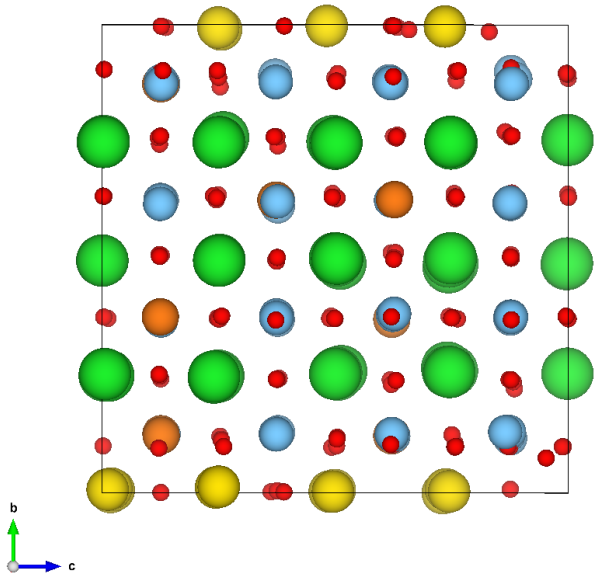


Figure D.10: La-Mg-BTO_o: b-c plane.

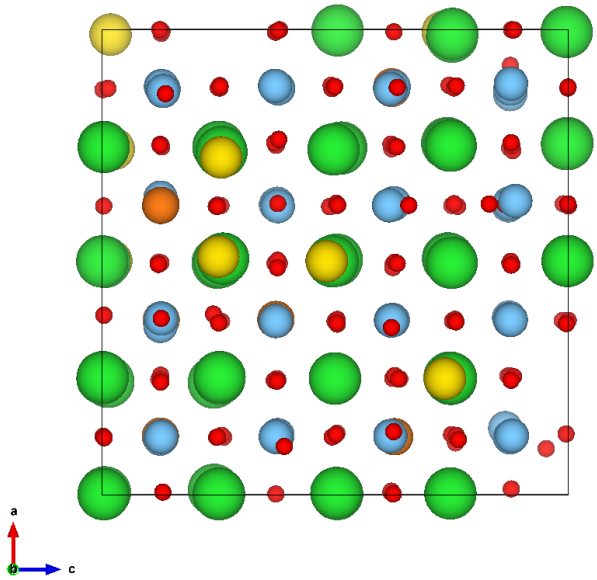


Figure D.11: La-Mg-BTO_o: a-c plane.

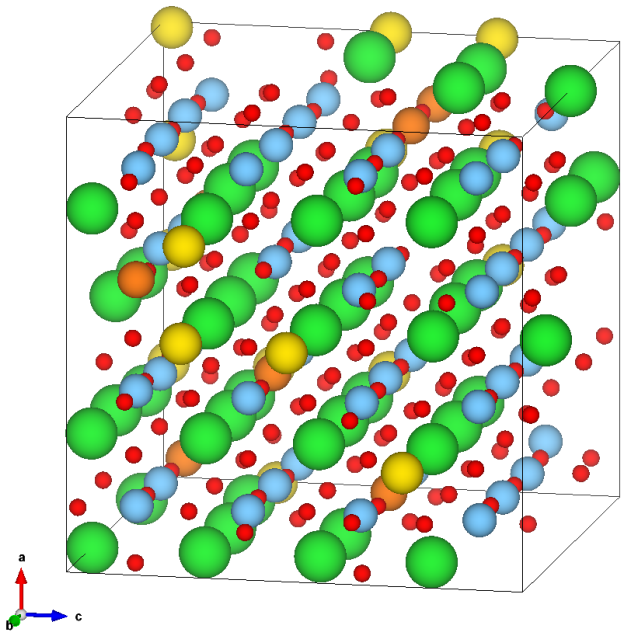


Figure D.12: La-Mg-BTO_o: off-axis.

Appendix E

Oxygen trajectories for MLFF simulations in Section 5.6

Trajectories of oxygen atoms for each system at a specified temperature - visualised using VMD, with a 'trajectory smoothing window size' of 50 [116]. Each section shows the given system along each axis, as well as a plot of time-lags vs. mean squared displacement.

E.1 La-Mg-BTO at 1500K

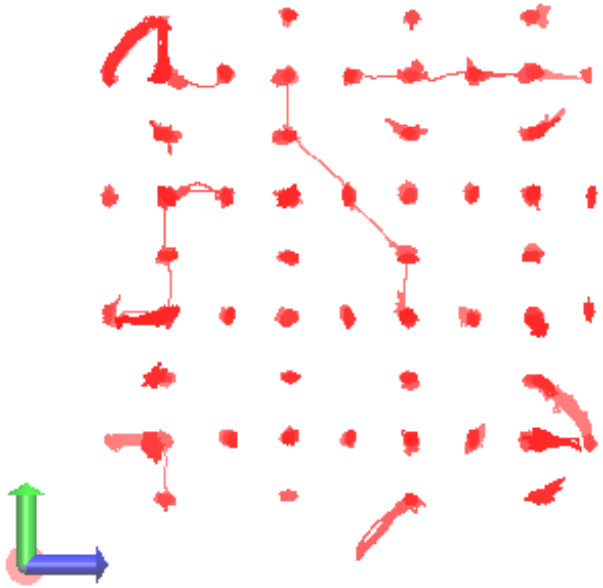


Figure E.1: Oxygen trajectories, viewed along the x-axis.

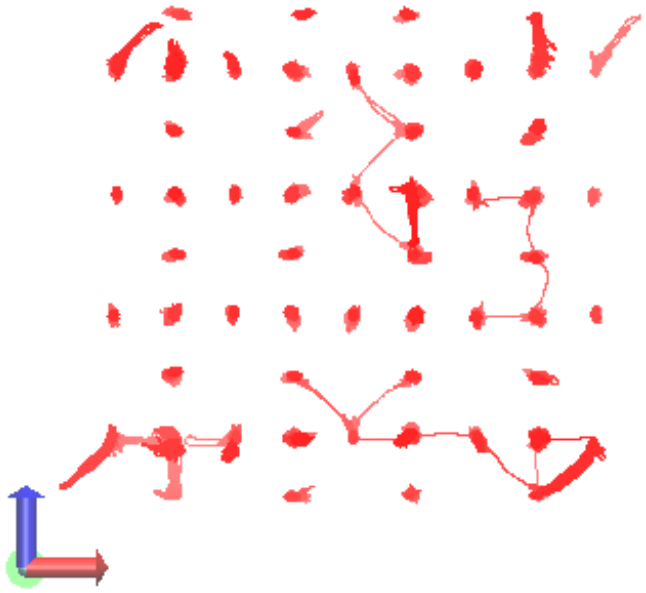


Figure E.2: Oxygen trajectories, viewed along the y-axis.

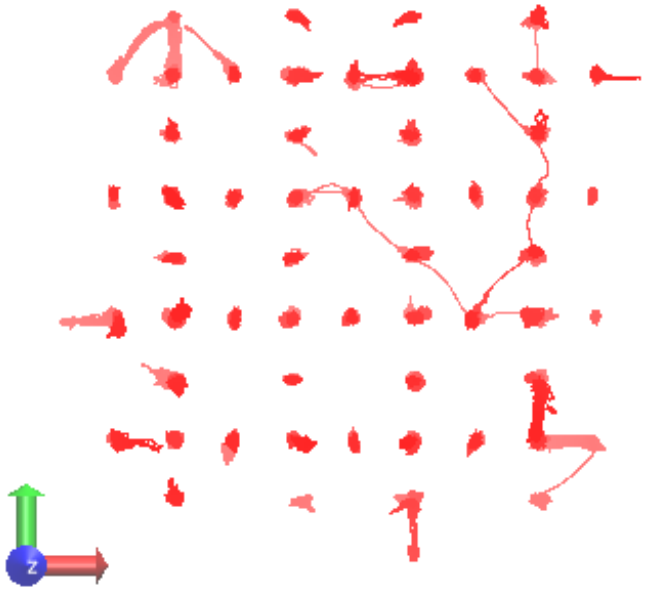


Figure E.3: Oxygen trajectories, viewed along the z-axis.

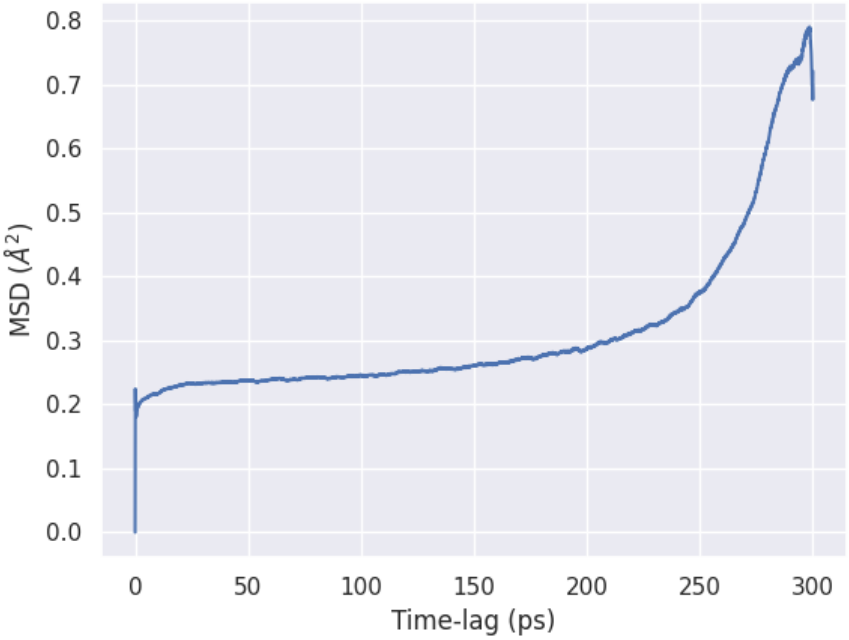


Figure E.4: Time-lag vs. MSD plot.

E.2 La-Mg-BTO_o at 1500K

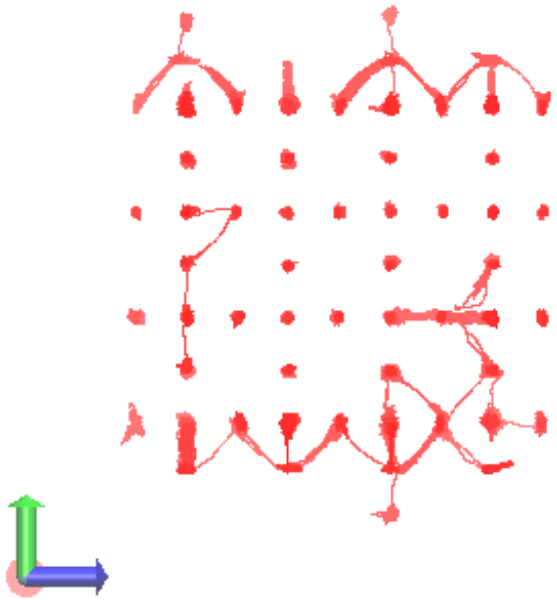


Figure E.5: Oxygen trajectories, viewed along the x-axis.

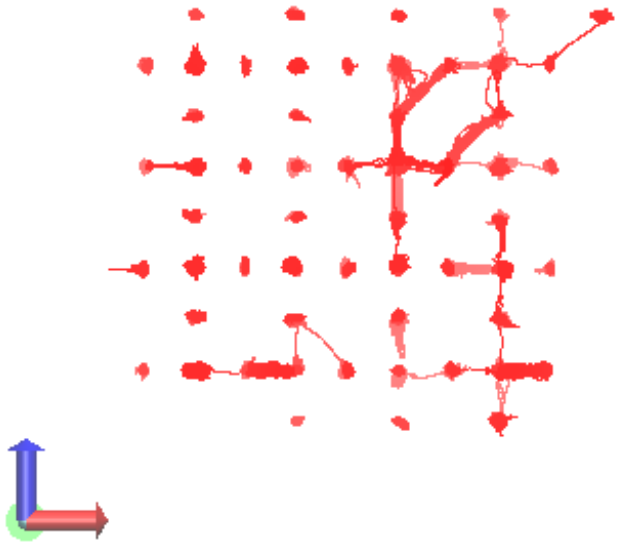


Figure E.6: Oxygen trajectories, viewed along the y-axis.

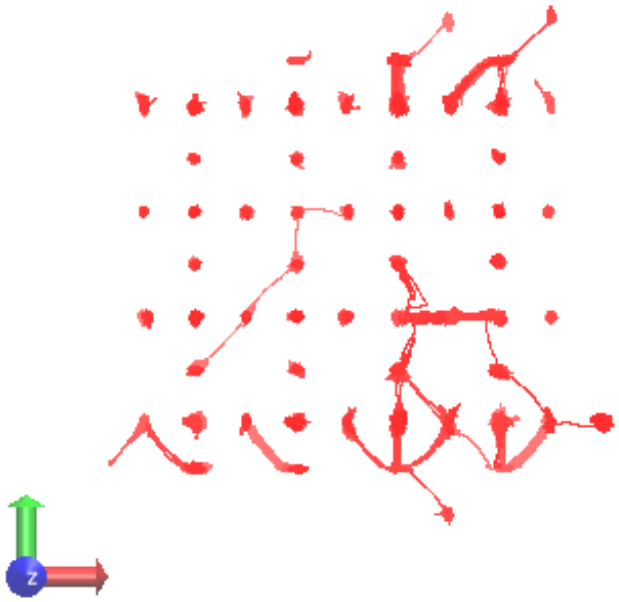


Figure E.7: Oxygen trajectories, viewed along the z-axis.

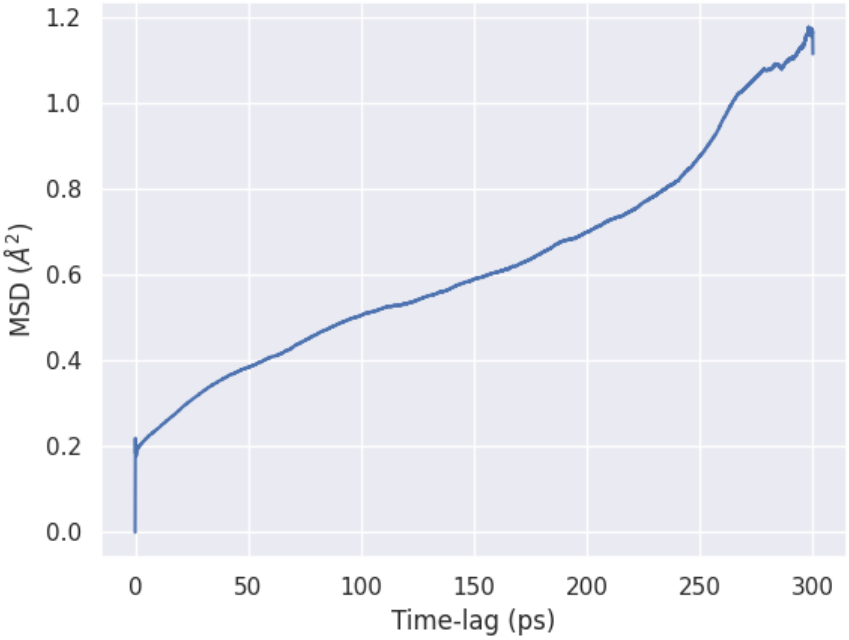


Figure E.8: Time-lag vs. MSD plot.

E.3 Mg-BTO at 1500K

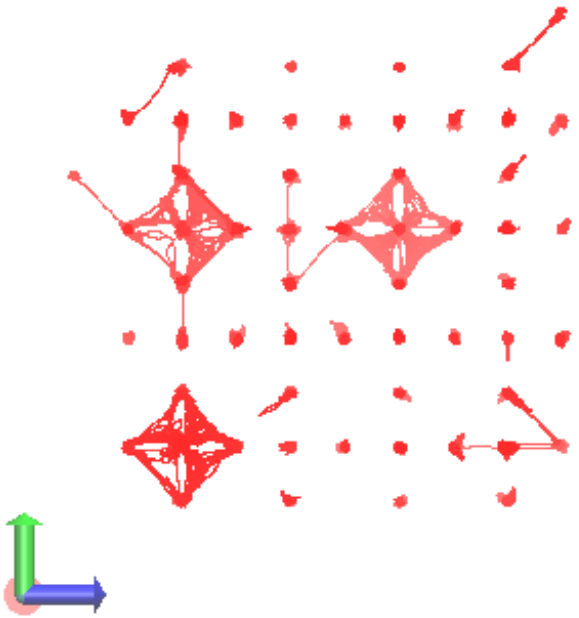


Figure E.9: Oxygen trajectories, viewed along the x-axis.

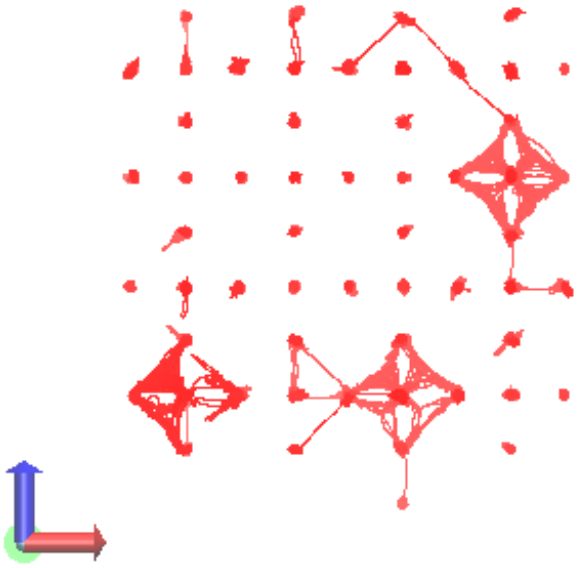


Figure E.10: Oxygen trajectories, viewed along the y-axis.

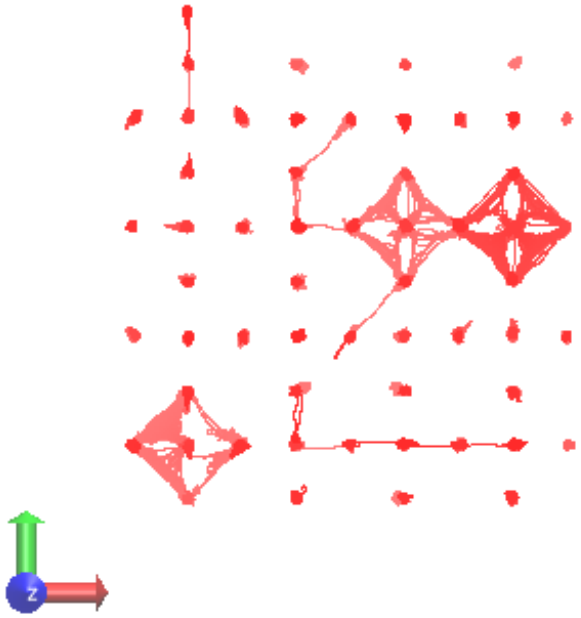


Figure E.11: Oxygen trajectories, viewed along the z-axis.

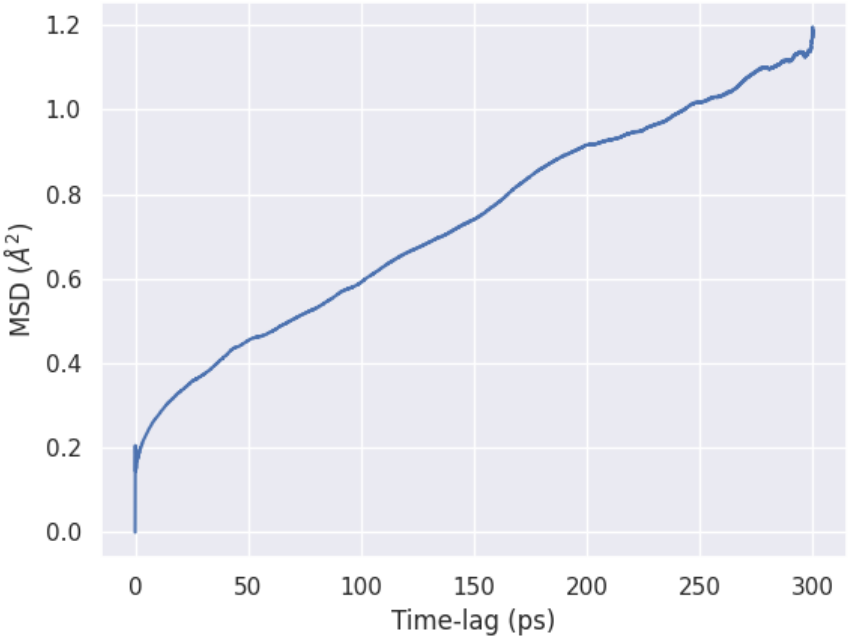


Figure E.12: Time-lag vs. MSD plot.

E.4 La-Mg-BTO at 2100K

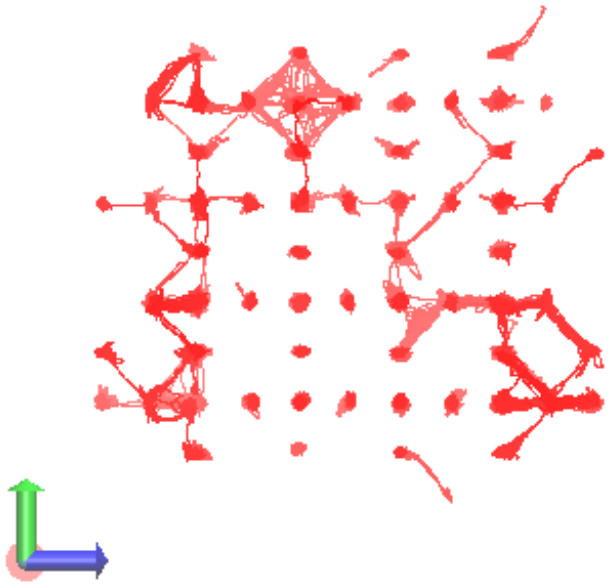


Figure E.13: Oxygen trajectories, viewed along the x-axis.

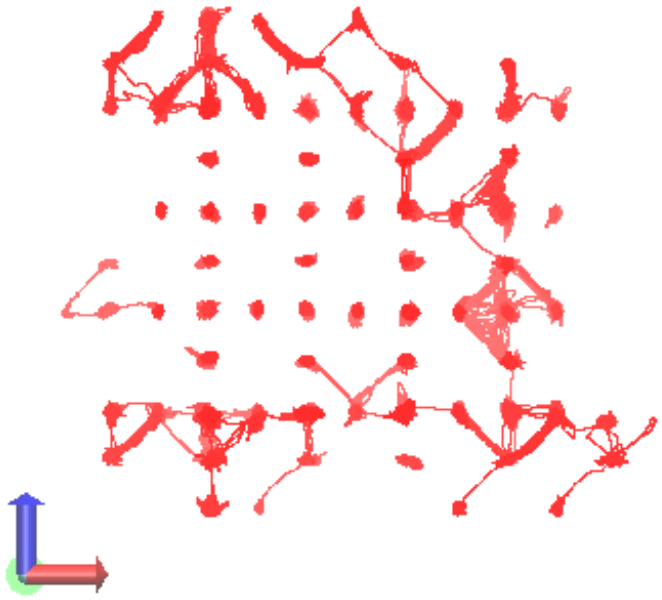


Figure E.14: Oxygen trajectories, viewed along the y-axis.

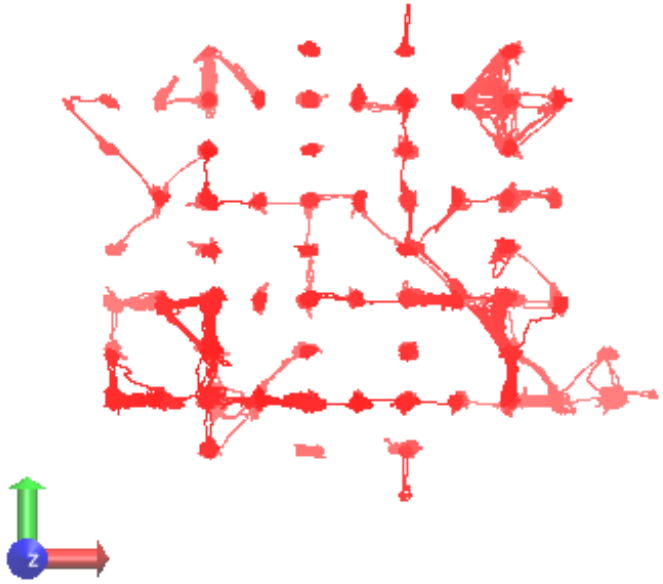


Figure E.15: Oxygen trajectories, viewed along the z-axis.

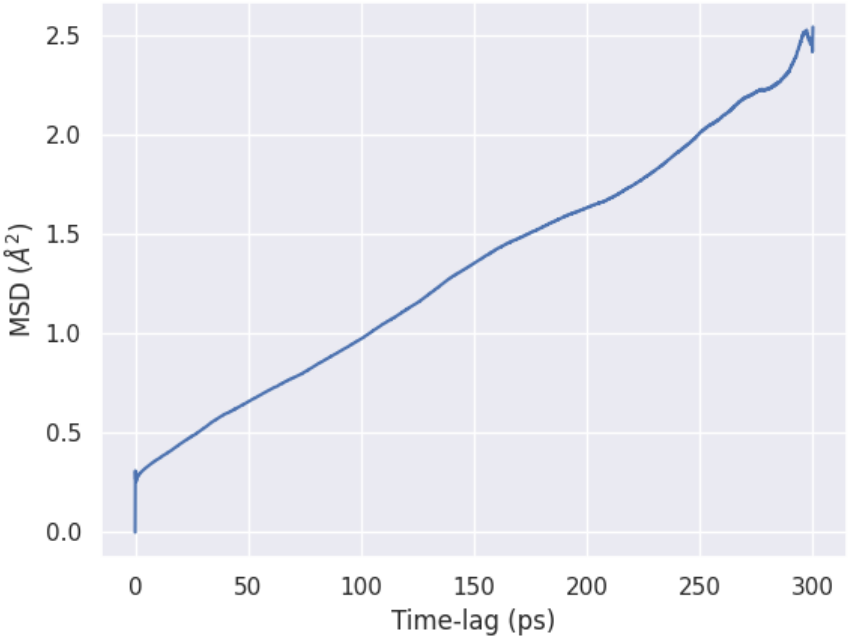


Figure E.16: Time-lag vs. MSD plot.

E.5 La-Mg-BTO_o at 2100K

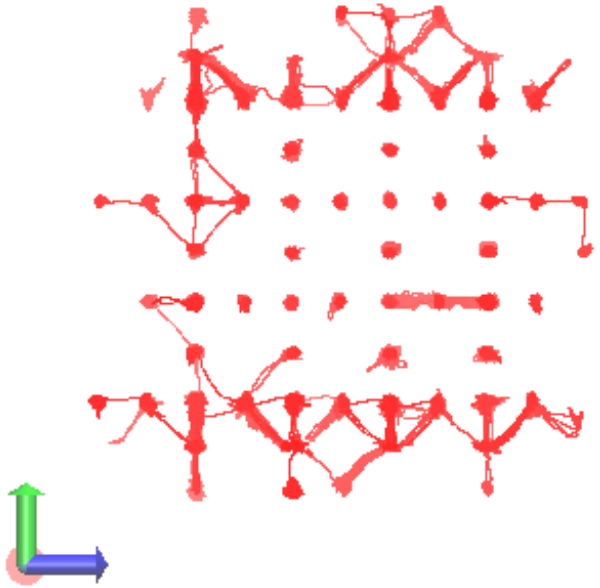


Figure E.17: Oxygen trajectories, viewed along the x-axis.

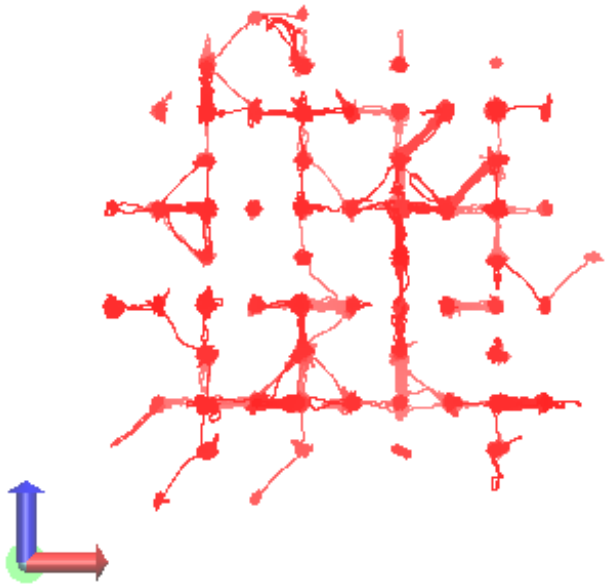


Figure E.18: Oxygen trajectories, viewed along the y-axis.

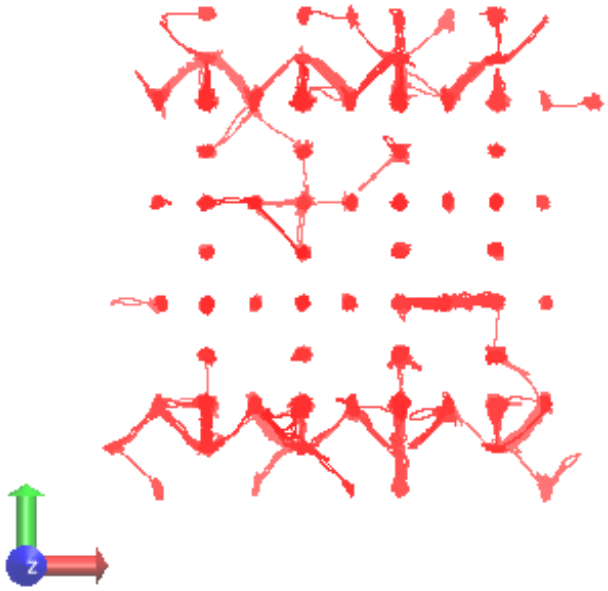


Figure E.19: Oxygen trajectories, viewed along the z-axis.

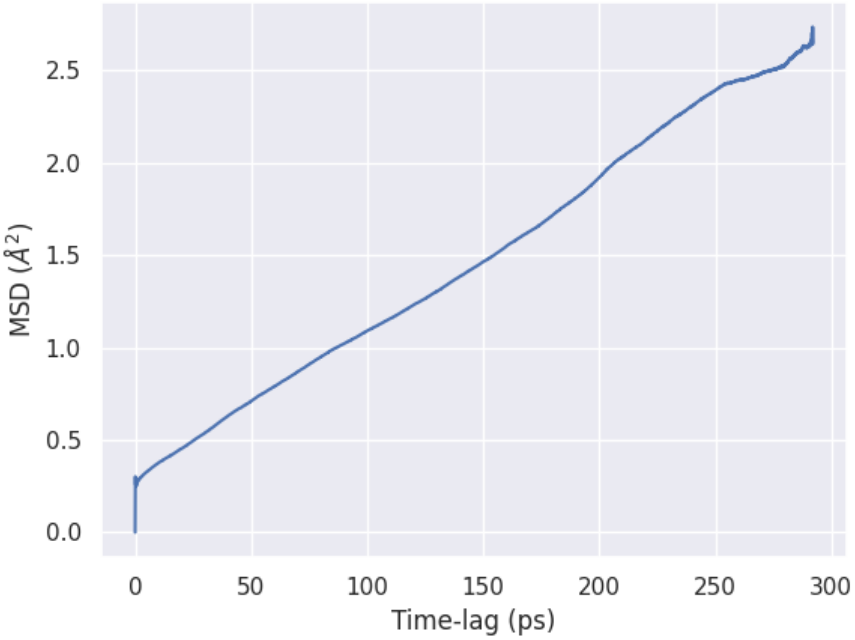


Figure E.20: Time-lag vs. MSD plot.

E.6 Mg-BTO at 2100K

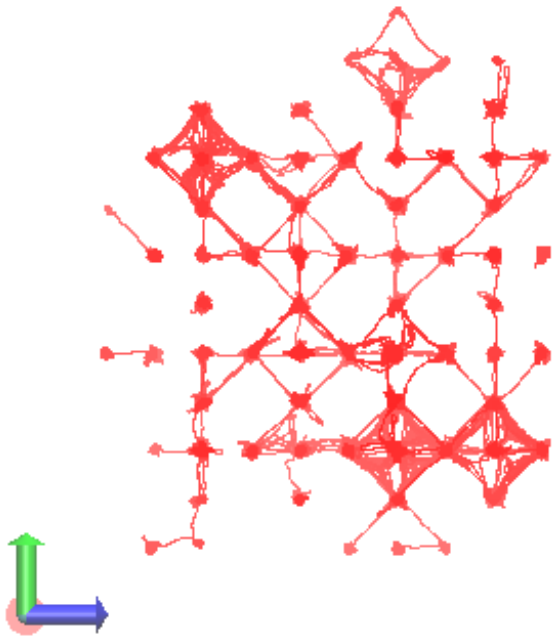


Figure E.21: Oxygen trajectories, viewed along the x-axis.

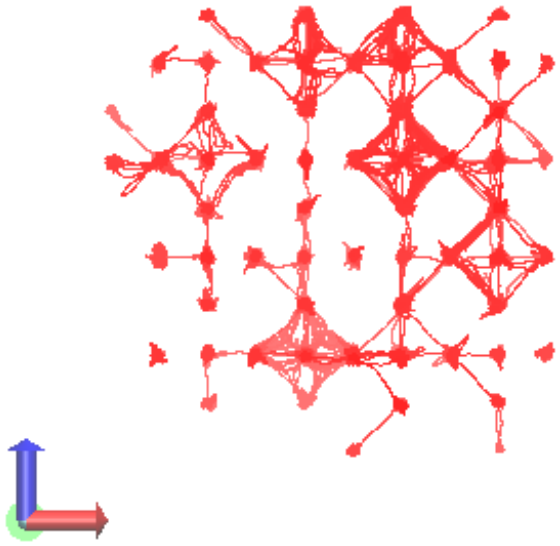


Figure E.22: Oxygen trajectories, viewed along the y-axis.

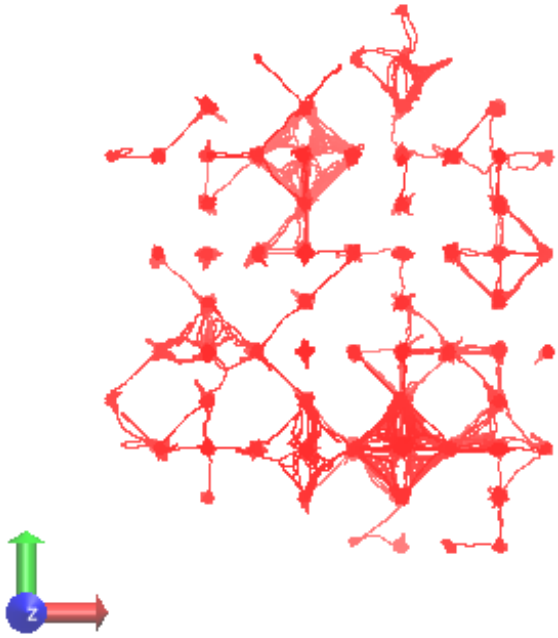


Figure E.23: Oxygen trajectories, viewed along the z-axis.

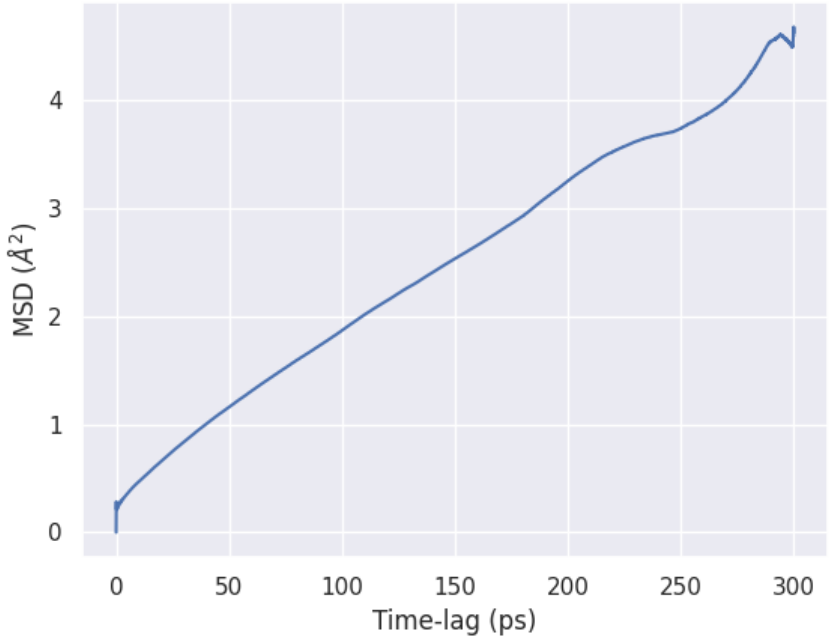


Figure E.24: Time-lag vs. MSD plot.

Appendix F

Axial MSD plots for oxygen diffusion in the La-Mg-BTO_o system at 2100K, as shown in Section 5.6

F.1 La-Mg-BTO_o at 2100K

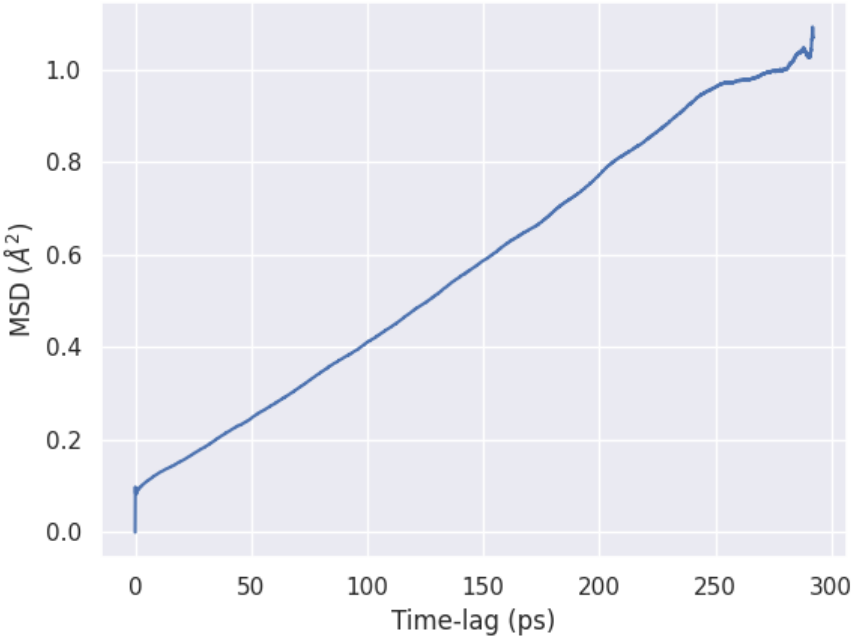


Figure F.1: MSD vs time-lag plot in the x-axis.

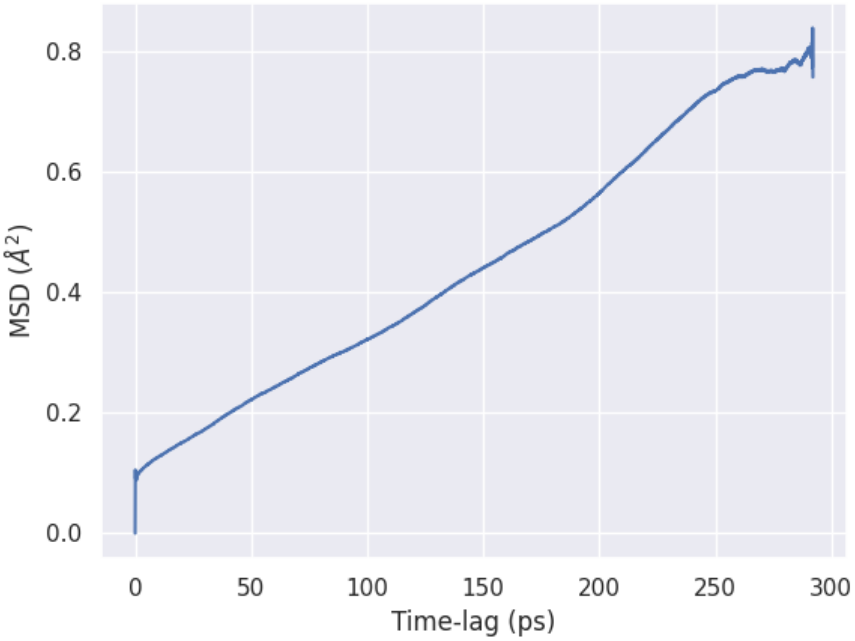


Figure F.2: MSD vs time-lag plot in the y-axis.

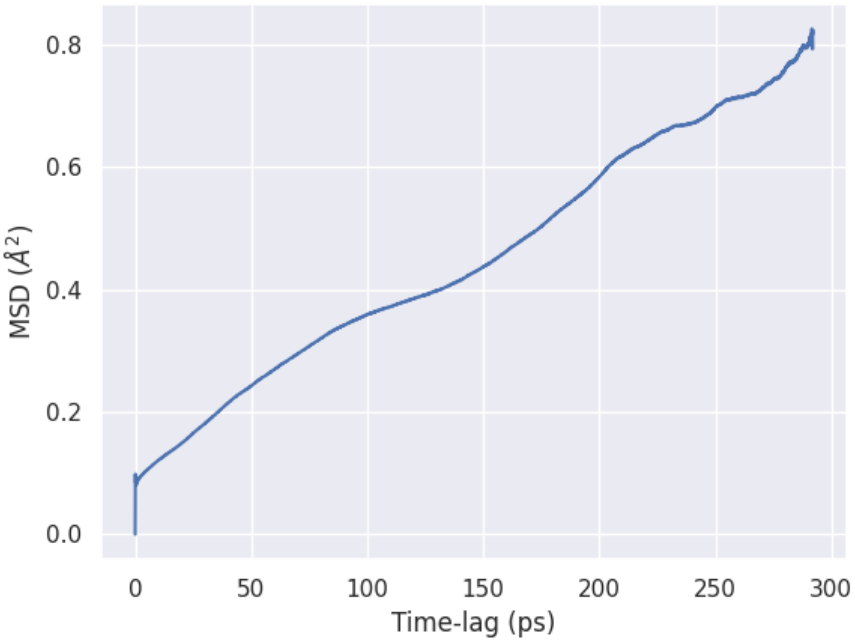


Figure F.3: MSD vs time-lag plot in the z-axis.