Department of Materials

Imperial College

MEng Thesis

Machine learning of defects in crystals

Wonjun Choi

Supervisor: Prof Aron Walsh and Alex Ganose

Date of submission: 6th June 2022

Abstract

Metal oxides are widely used owing to its stability, low cost, abundance, and versatility. Oxygen vacancy formation energy plays a significant role in chemical and physical properties of metal oxides. Thus, accurately obtaining oxygen vacancy formation energy of various metal oxides is desired for both theoretical and experimental researchers. This project presents data-driven method for fast and accurately predicting oxygen vacancy formation energy of metal oxides. The dataset consisted 1665 metal oxides with structural and compositional diversity. Composition based feature vector (CBFV) was used to uniquely describe metal oxides. Gradient boosting regressor exhibits the best performance for accurate prediction with RMSE value of 0.465 eV. The regression model have tendency to overestimate low energy materials and underestimate high energy materials. Gradient boosting classifier successfully classify metal oxides with low (< 3 eV) and high oxygen vacancy formation energy with accuracy of 0.96. The recall value for low energy is 0.57 due to highly imbalanced dataset. Since CBFVs cannot distinguish structural difference between polymorphs, the best accuracy of the model is limited by the difference in energies between multiple inequivalent oxygen vacancy formation energies.

Contents

Acknowledgements

First and foremost, I wish to thank my supervisor, Professor Aron Walsh, for his guidance and help, and he provide related literature and detailed descriptions regarding experimental procedures, but also helped me to analyse the results. I wish to thank Alex Ganose, providing material dataset for the experiment. Also thanks to Saajan Shah and Braian Lew for the collaboration of the work that resulted a good proportion of the work in this project. Lastly, I would like to take the opportunity to thank my friends who supported and motivated me to complete the thesis.

1 Aims and Context

Metal oxides have been a popular material in scientific field owing to its stability, low cost, abundance and versatility [1]. Metal oxides can be tailored to meet its need, which is widely used in the fuel cells [2], superconductors [3], super capacitors [4], sensors [5], catalysts [6] and thermal insulators [7]. Point defects are ubiquitous in metal oxides, which affect their physical, chemical, electronic and optical properties [8]. One of the most prevalent point defects in metal oxide is oxygen vacancy, which was first introduced by Tompkins *et al.* as a substance in surface chemistry [9]. After then, many researches have found out that oxygen vacancies exists in many metal oxides, and the level of defects affects material's structure and thus their physical and chemical properties [10, 11]. Enormous studies have contended that the existence of oxygen vacancies have significant influence in the production of efficient materials. For instance, oxygen vacancies in photo-catalysts modify their band structures by either up-shifting valence band (VB) or down-shifting conduction band (CB) allowing electron excitation by visible light with narrowed band gap [12]. Additionally, oxygen vacancies improve photo or electocatalytic performance by tuning electron transfer, adsorption and activation of molecules, and catalysis kinetics [8]. Hence, accurately assessing the effect of oxygen vacancies in metal oxide is inevitable in the development of advanced novel materials. One of the methods to characterise vacancy in a solid material is obtaining vacancy formation energy(E_f). The vacancy formation energy is defined as the energy required to form a vacancy [13]. Oxygen vacancy formation energy plays a significant role in physically and chemically analysing metal oxides. Thus, obtaining the formation energy with high accuracy is desired for researchers.

For current industries, it has always been a great challenge in the material choices. In the relatively recent times, material selection only relied on the people's experiences and intuitions, and tremendous experiment need to be done to support a designer's evaluation and adopt new materials into practice with prior experimental results [14, 15]. Despite the discovery of novel materials is the key milestone of high technology, synthesising and evaluating requires a welter of experiments while only few of them are successful. Although few successful experiments leaded to a material development, but it still rely on trial and error, which is both "costly and time-consuming" [16].

The alternative approach was the use of first principal ab initio methods. Classically, a Schrödinger wave mechanics has been widely used as a theoretical tool for understanding elementary particle phenomena in a wide range of areas, and developing numerous advanced devices based on quantum effects. However, there is a problem that the Schrödinger equation cannot be solved accurately for complex molecular or material systems of two or more atoms. In 1965, Kohn and Sham have proposed a Kohn-Sham equation, which is for directly obtaining the exact electron density in the random space, r, for the ground state of the electron [17]. This new methodology based on Kohn-Sham equation is called density functional theory (DFT), and it had unparalleled impact on solving complex problems in chemistry [18]. The DFT allows accurate electron density and accurate prediction of all physical properties, hence the interest and demand has resulted in tremendous popularity. Recently, linear methods have been developed to enable the calculation of the first principle for large systems containing thousands of atoms [19]. However, the DFT results are strongly depends on the functional of metal oxides as the real functional is unknown [20]. Moreover, DFT requires very high computational cost. In other words, DFT results are very system dependent, and difficult to derive solutions for diverse materials.

To overcome the limitation of DFT, many researches are adopting data-driven analysis methods for high throughput screening, allowing automatic testing of various materials by the use of machine learning models. Machine learning is widely used for predicting band gap and electronic properties [21–23]. However, there are still many studies for developing accurate machine learning models and descriptors for diverse metal oxides, which highly contribute to the accuracy of prediction.

This project, therefore, focuses on developing the best machine learning model to fast and accurately predict oxygen vacancy formation energy of various metal oxides by hyperparameter tuning and feature engineering.

2 Literature Review

2.1 Project Summary

In recent materials industry, identifying novel materials using computational science (i.e. first principle, molecular dynamics, monte-carlo simulation) has grown to prominence. According to the IDC Market Research White Paper in 2005, modelling and simulation technology has improved efficiency of research, thus reducing research cost and time and empowered the development of materials. Material Genome Initiative (MGI) project, in particular, has increased interest in computational material science. This project served as a momentum in bringing new paradigm for the area of material design. The importance of database (DB) has been emphasised in material industry, and efforts to build the material database around research groups have begun since then. Research groups that built DB then started to graft Artificial Intelligence (AI) to the area of material design.

Recent studies of material design have an explosive interest in AI. Why material researchers are so excited to AI? This is closely related to time it takes for prediction. For instance, material prediction takes several months with classic calculation (i.e. chemical reaction path with first principle) whereas AI can reduce the time to several minutes. AI can provide such rapid calculation because it looks for the correlation between material information (composition, structure) and property to predict properties of novel materials. Here, principles such as quantum mechanics, classical mechanics and statistical mechanics are not used at all. Hence AI-based material design technology can provide a new paradigm to studies of material development. Figure 1 introduces materials AI technology as 4^{th} paradigm technology of material development. It is classified into a 1^{st} paradigm that relies on trial and error, a 2^{nd} paradigm based on thermodynamics, a 3^{rd} paradigm based on computational science and a 4^{th} paradigm based on AI.

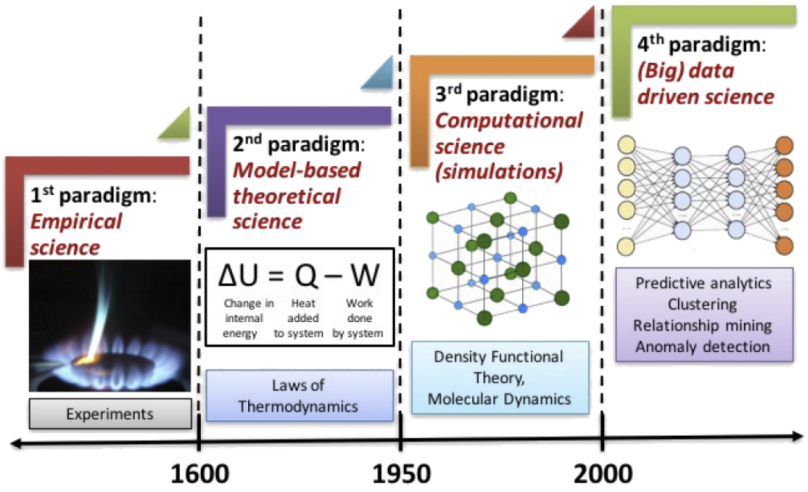


Figure 1: 4^{th} paradigm technology of material development: AI-based technology [24]

Currently, AI is being used in various fields of material design. In particular, research is actively underway in the field of developing new theoretical methodologies to overcome the shortcomings of existing computational science technology; the field of predicting physical properties from material information and reverse prediction technology that outputs material information with desiring physical properties. Following document therefore discusses critical review on current state of research regarding AI in the field of material design.

Table 1: Types of machine learning algorithms and their applications [28]

Machine Learning	Algorithm	Applications
Supervised Learning	Regularized least squares Support vector machines Kernel ridge regression Neural networks Decision trees Genetic programming	Predict processing- structure property relationships; Develop model Hamiltonians; Predict crystal structures; Classify crystal structures; identify descriptors
Unsupervised Learning	K-means clustering Mean shift theory Markov random fields Hierarchical cluster analysis Principal component analysis Cross-correlation	Analyze composition spreads from combinatorial experiments; Analyze micrographs; Identify descriptors; Noise reduction in data sets

2.1.1 Introduction of AI in Materials

In order to use AI in actual material research, three components are required: a material DB, a machine learning (ML) algorithm and a feature or descriptor [25].

Machine learning algorithms are divided into supervised learning and unsupervised learning depending on the form of the DB and the purpose to be analyzed [26]. In the case of supervised learning, a model between input and output information is developed by learning information on both input and output data. Thereafter, the model is used to predict the output value for a given new input value. In general, it is a method of predicting physical properties from material information (composition, atomic structure, etc.) during material design and development. For instance, in the field of electronic material development, supervised learning methods can be used for research such as predicting electronic structure properties such as band gap [21, 22], elastic constants [21, 23], Debye temperature [21] or predicting adsorption energy of reactants from material composition or atomic structure information.

Meanwhile, unsupervised learning can be efficiently used in classification fields such as specific statistical patterns or clustering of input value data without specifying output value. As an example, there is a research that outputs a phase diagram by determining the similarity of crystal structure of the material [27]. Table 1 summarizes the classification of these machine learning methods and their application fields.

In the field of material machine learning, input data is usually called a feature or descriptor, and output is called a property. For instance, if we want to obtain adsorption energy from its atomic structure and electronic structure, atomic and electronic structure becomes a feature/descriptor. Thus it is important to determine feature when using machine learning methods. Characteristics are clearly determined by the physical properties, but determining features depends heavily on the subjective judgement of the researcher. When machine learning an adsorption energy of materials, for example, higher prediction accuracy can be increased when using defect type, density, electronic structure for features. When selecting features like this, domain knowledge must be utilized. Selecting features using such domain knowledge is called feature engineering, which is considered the most important step in using machine learning. Typical examples of feature extraction methods will be discussed in section 2.2.1.

2.1.2 Material Database

In order to use AI for material development, a large-scale material DB must be premised. The reason why AI research is very active in areas such as social science and the financial sector is that it is easy to construct database. The field of bio-research has quickly became aware of the importance of big data, hence DB construction has been systematic for a long time. In comparison, the construction of DB started relatively late in the field of materials, but data collection process is at a fairly fast pace over the past decade.

Material structure database Collecting structural information of all possible materials must be prioritized in constructing material informatics DB. This includes not only materials already found or used, but also all materials that can be made virtually. The largest structural information databases are Inorganic Crystal Structure Database (ICSD) [29], Cambridge Structural Database (CSD) [30] and Chem Spider [31, 32].

Material property database Various material property information databases (i.e. thermodynamic properties electrical/magnetic/mechanical properties) are calculated based on material structure databases such as ICSD, CSD, Chem Spider and many more. The largest databases for this are Open Quantum Materials Database (OQMD) [33], Materials Project [34] and Citrination Materials-Data Analytics Platform [32, 35].

2.1.3 A Study on Materials using AI

AI research has been very active worldwide since 2015 using DBs discussed earlier. When AI prediction of the correlation between material structure and physical properties is highly efficient, it can be a stepping stone to the new paradigm of material development methods. This is because it is possible to overcome the excising trial error based Edisonian approach, and to select and search high performance materials using AI.

The development of an AI model is largely composed of three stages [36]. The first step is sample construction, which corresponds to the DB construction discussed earlier. Now, DB is constructed with data calculated by computational science method to increase data consistency. However in the future, it is desirable to collect a larger database from experimental results. Unnecessary or inaccurate data may also be collected while this process, and hence pre-processing of data must be done. In the sample construction, determining features as input values for machine learning model is carried out in company with data collection. Determining features requires researchers' intuition and domain knowledge. Recently, research is actively underway to automate feature creation and optimize it in the direction of maximizing the accuracy of machine learning.

The second step is the development of a machine learning model. When input and output values are defined, a model can be developed through various machine learning algorithms. However, classification can be done without defined values through unsupervised learning. The final step is the verification of the model. In this step, the accuracy of the trained model is verified. Here, cross-validation method is used in order to prevent high accuracy from occurring by chance. If the model is accurate above the standard, materials properties can be predicted at a very high speed by expanding to unknown materials that are not in the DB.

Prediction of bulk properties Over the past three years, machine learning has quickly dissolved into the area of material property prediction. Prof. Jefferey C. Grossman's team of MIT materials developed a Crystal Graph Convolutional Neural Network (CGCNN) AI algorithm, specialized in predicting bulk-state properties of materials [37]. In order for a machine to learn the correlation between the crystal structure of the materials and their physical properties, encoding the structure information of the material is the first step. They approached with a crystal graph that encodes both atomic information and bonding interactions between atoms, and successfully produced an algorithm (fig.2).

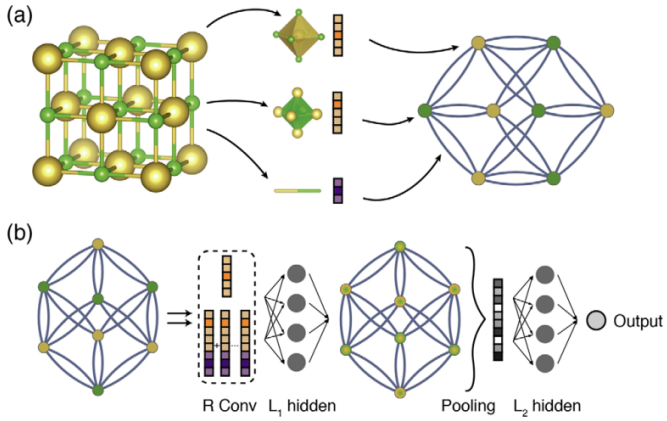


Figure 2: (a) Graphic representation of crystal structure information of a material, (b) A schematic diagram for predicting bulk properties through CNN of crystal graph [37]

The architecture was inspired by architectures for computer vision [38], natural language processing (NLP) [39], molecular fingerprinting [40] and general graph-structured data [41, 42]. In this research, the entire material information (about 70,000 materials) in the previously introduced Materials Project was learned, and thermodynamic, electrical and mechanical properties, which are the most basic intrinsic properties of the material, were mainly selected. This includes, bulk formation energy, absolute energy, band gap, fermi energy, bulk moduli, shear moduli, Poisson ratio and many more. The material property prediction accuracy was very good overall. The researchers showed that machine learning could become a major technology in material property prediction as errors from machine learning was significantly lower than from experimental results (table 2). Here, it is worth noting that CGCNN can predict the material properties to the level of first principle calculation using information that can be easily obtained from the periodic table.

Table 2: Accuracy of CGCNN prediction on material properties ($MAE = mean\ absolute\ error$)

Property	# of train data	Unit	$\text{MAE}_{\text{model}}$	MAE_{DFT}
Formation Energy	28046	eV/atom	0.039	0.081-0.136 [43]
Absolute Energy	28046	eV/atom	0.072	-
Band gap	16458	eV	0.388	0.6 [44]
Fermi Energy	28046	eV	0.363	-
Bulk moduli	2041	log(GPa)	0.054	0.050 [45]
Shear moduli	2041	log(GPa)	0.087	0.069 [45]
Poisson ratio	2041	-	0.030	-

CGCNN algorithm can be used very efficiently to predict material properties, but is limited to bulk structures. However other materials such as catalyst materials requires a technology capable of handling surface structures. Kim's research team thus developed Slab Graph Convolutional Neural Network (SGCNN) based on CGCNN but can be applied to the surface structure (fig.3) [46].

The main idea of SGCNN is considering a surface structure of material in a Slab form, and the slab structure can be divided into bulk and surface areas. Here, the atomic structure and element information in the bulk region are expressed with a crystal graph, and the atomic structure and element in the surface region are expressed with a separate surface graph in a similar manner to CGCNN. When expressing the surface graph, the chemical reaction-related molecules/atoms adsorbed on the catalyst surface are also

considered. Then, combining the encoded vectors from CGCNN and SGCNN enables the expression of the surface sreucruer adsorbed with molecules/atoms. Relating encoded vectors of surface and its adsorption energy produces a model that predicts adsorption energy of molecules/atoms on the catalyst surface. In fact, Kim's team has proposed six new catalysts [46], and they are currently applying the algorithm to the design of alloy catalysts for fuel cells.

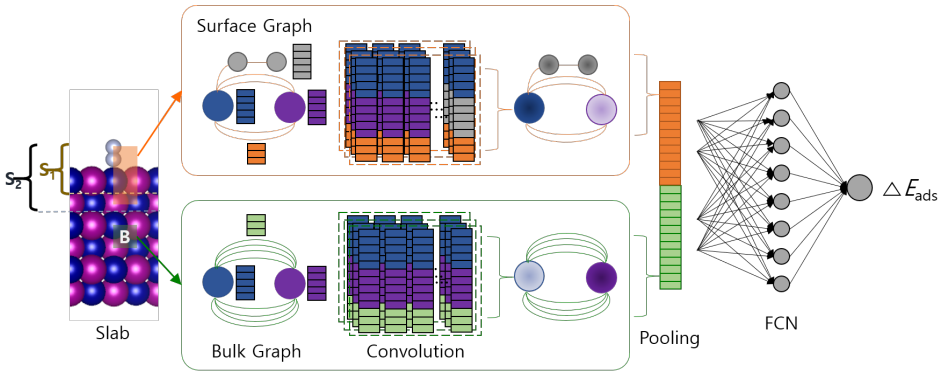


Figure 3: AI technology for catalyst material design [46]

Crystal structure classification Angelo Ziletti and Matthias Scheffler in Fritz-Haber Laboratory in Germany developed a machine learning model to learn the correlation between the x-ray diffraction pattern and the crystal structure of materials [47]. It is expected to have a great impact on industry, as it has the advantage of being able to predict the structure of materials from diffraction pattern for unknown materials. Since AI algorithms have been very specialized and developed in image analysis/processing (i.e. face analysis) before material industry, current deep learning algorithms based on convolutional neural network are specialized in image analysis and classification. The team understood and exploited the same algorithm to distinguish crystal structures by imaging diffraction patterns. The developed algorithm is also applicable to materials with many defects, hence will be inevitable for industries in the future.

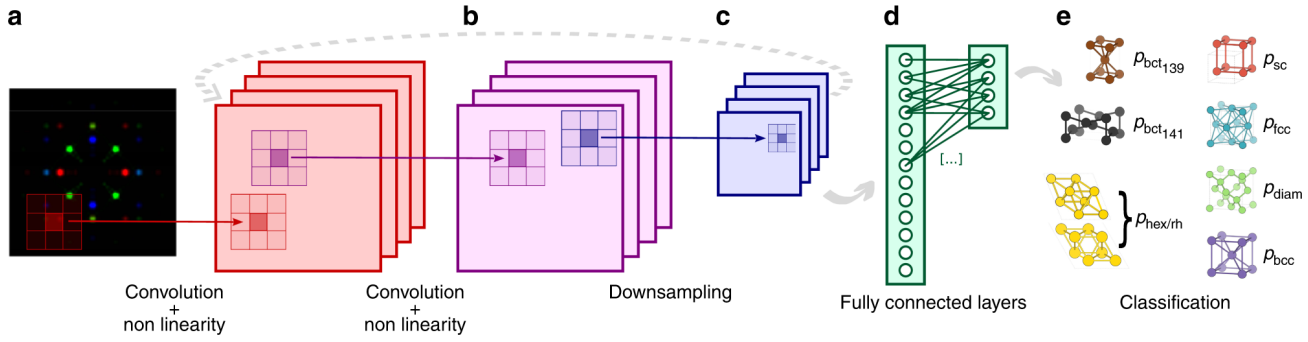


Figure 4: (a) Diffraction pattern of unknown materials,
(b-d) Convolution, pooling, fully-connected layer, (e) Crystal structure differentiation [47]

Catalysts for carbon dioxide reduction A research team led by Hongliang Xin introduced a machine learning method in developing alloy catalysts for CO_2 reduction [48]. The adsorption energy on the catalyst surface is widely known as a chemically reactive descriptor. However, the adsorption energy requires a considerable amount of time and computer resource with expensive quantum calculation. The neural network algorithm successfully learned the correlation between electronegativity, d-band properties and carbon monoxide adsorption energy. As a result, the adsorption energy of new catalyst materials that have never been learned was predicted very accurately with mean squared error (MSE) of 0.13eV. The field of catalyst has difficulties in finding novel catalysts as experiments and calculations need to be done for all possible reaction type, reaction path and materials. The introduction of statistical machine learning methods is expected to overcome these limitations and further accelerate the development of new catalyst materials. The machine learning method has been applied to catalyst development research to promote various reactions over the past four years, and some of the cases are summarized in table 3 below.

Table 3: Machine learning application to catalyst development research

Team	Algorithm	Input variable	Chemical reaction
Jinnouchi et al. [49]	Local Similarity Kernel	Surface structure (local structural similarity)	Direct NO decomposition
Tran et al. [50]	TPOT Regression	Structural information, atomic number, electronegativity, coordination number	Electrocatalysts for CO2 reduction
Gasper et al. [51]	Gradient Boosting Regression	D-band properties, coordination number	Electrocatalysts for CO2 reduction
O'Connor et al. [52]	LASSO	Oxide formation energy, oxygen vacancy formation energy	Single-atom catalyst
Han et al. [46]	CGCNN	Elemental properties of materials in periodic table	Ammonia generation

Ultraincompressible and superhard materials The research team of professor Jakoah Begoch used machine learing to explore novel material with super hardess mechanical properties [53]. The bulk moduli and shear moduli are the reference physical properties for evaluating the strength of materials. The team successfully developed a machine learning model that can accurately predict the bulk and shear moduli by sampling 2,572 materials. The machine learning model was based on a Support Vector Machine Regression (SVR) algorithm [54]. Descriptors for the model were 34 distinct compositional variables describing the elemental properties such as position on the periodic table, electronic structure and physical properties. Then, the model was applied to a total of 118,287 chemical compounds to select materials with a bulk modulus of higher than 300GPa and a shear modulus higher than 150GPa. The selected materials are then tested experimentally, and two types of carbide materials, $ReWC_2$ and $Mo_{0.9}W_{1.1}BC$ were discovered. It is virtually impossible to conduct about 120,000 experiments in real life through the Edisonian approach, whereas machine learning algorithms can be an alternative method with low error. This research is one of the good examples of material development acceleration.

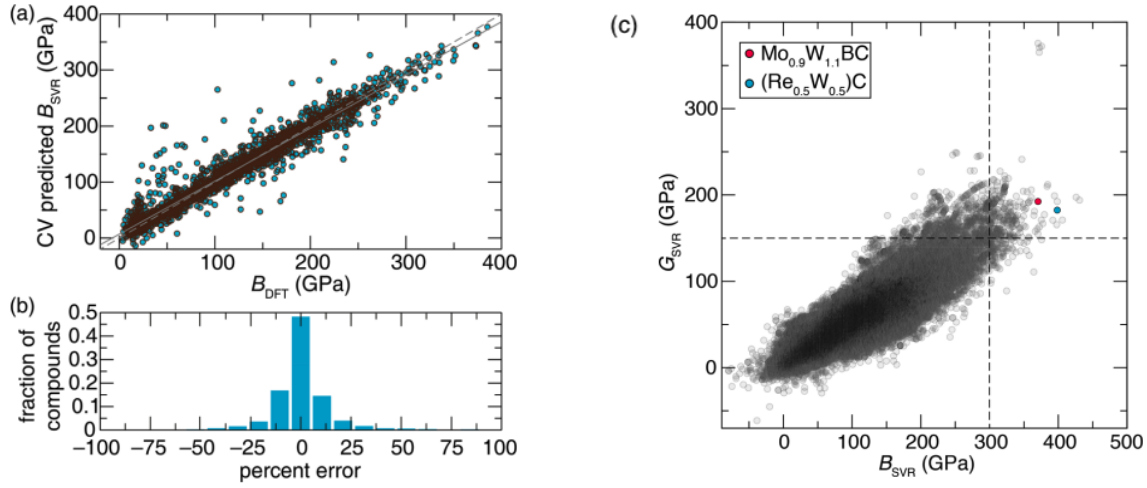


Figure 5: (a) Machine learning results (b) The accuracy of the model (c) Discovering materials from 120,000 possible materials [53]

Amorphous metal (glass former) Aperva Mahta and Christopher Wolverton introduced machine learning to discover new amorphous metals [55]. Amorphous metals have high mechanical strength and brittleness compared to highly crystalline metals. In many cases, amorphized metals are found in ternary or quaternary metals, but there are extensive amount of element combinations and compositions which hinders the research progress. The researchers generated a machine learning model that predicts whether a material can have amorphous forming ability in a wide variety of compositions of ternary metals (A-B-C). They used same features and similar data set, and used a single model for the entire data set rather using a separate models used by Ward et al. [56]. The random forest algorithm [57] was used by assessing 10-fold cross validation. The predicted materials were verified through experimental verification to confirm that the actual glass formers could actually be made, and the researchers foresee that it has the effect of improving the speed of discovering novel metallic glass materials in the research field by about 200 times.

Point defect energy of MPEA Multi-principal element alloys (MPEAs) are metal alloys consisting many principal elements that are distributed randomly in the crystal. Manzoor et al. developed a model that can predict point defect energies of ternary, quaternary and quinary MPEAs by using database of their constituent binary alloys [58]. They focused on Ni-Fe-Cr-Co-Cu alloy. The machine learning data, so called descriptors, in predicting vacancy formation energy were constructed by removing one of the atoms to create cation vacancy, recording the position of first and second neighbouring atoms from the vacancy, and their type and orientation (fig 6). In predicting vacancy migration energy, the distance moved by the atoms in relaxation and distance between vacancy and migrated atoms and their first neighbouring atoms were also added. The support vector regression (SVR) algorithm was used to train the data. They successfully predicted the vacancy migration energy and vacancy formation energy from binary alloy data, resulting root mean square error (RSME) of 0.07 0.09. It is yet challenging but there are many ongoing researches on machine learning of defects [59, 60]. Machine learning is becoming inevitable in material engineering, and it is expected to not only reduce computational cost, but also serve as a stepping stone for a new field of materials.

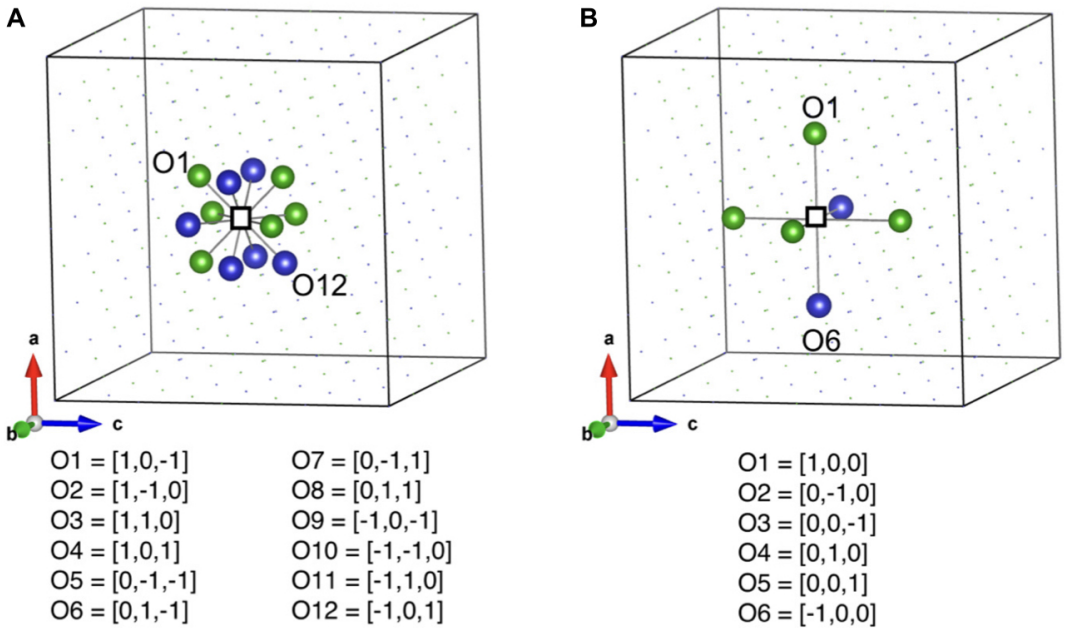


Figure 6: Schematic representation of descriptors for MPEAs: first and second neighbouring atoms [58]

2.2 Descriptors and Algorithms

2.2.1 Material Descriptors

The databases have been created with the results of density functional theory (DFT) simulations. The way compounds are represented from DFT results plays an important role in regulating the performance of a machine learning technique. As mentioned earlier, this is called a 'descriptor'. Since descriptors are selected based on the experts' knowledge via trial and error process, the predictive performance is heavily dependent on the quality and the size of the data of target characteristics.

In materials machine learning, a set of descriptors should satisfy followings: unique, continuous and differentiable, invariant to transformations such as translations, rotations, and nuclear permutations, machine-readable, and computationally cheap [61–64]. Descriptors are very similar to human's fingerprints or face. In the case of facial or fingerprint recognition, each person has a unique fingerprint or face and they must not depend on the location. There is no 'best' descriptors or machine learning models as referring to the 'No Free Lunch Theorem (NFL)' [65]. Therefore, there are multiple combinations of descriptors and machine learning models, and it is user's responsibility to optimise by choosing the best combination.

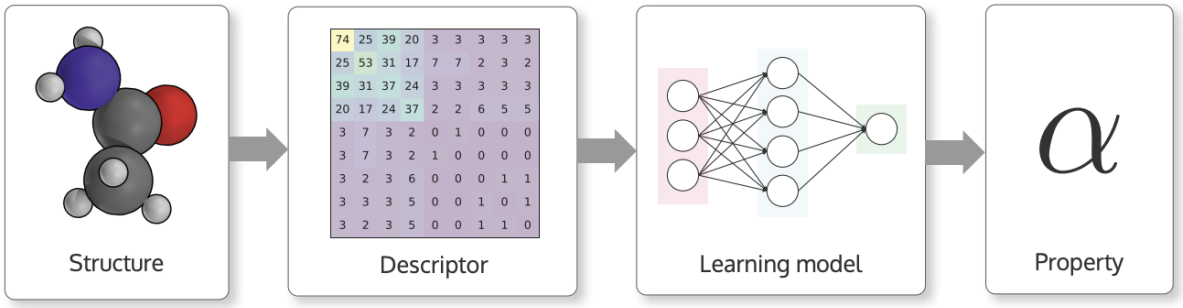


Figure 7: Typical steps in materials machine learning [66]

Compound descriptors The most popular descriptors are classified into three main groups. Physical properties from the library, physical properties calculated from DFT, and properties of elements and structure of a compound. The compound descriptors are usually based on DFT calculations such as volume, band gap, cohesive energy, elastic constants, dielectric constants, etc [67]. The DB based on the DFT calculation is limited as it is based on first principle calculation. This may be solved by discovering a machine learning model to discover properties of new materials. Research on discovering bulk properties of compounds was done since early 2010s, and can be found in the literature [68–70].

The simplest way to represent a compound as a descriptor is by using a binary digits. If there is m number of compunds to be trained, the descriptors would be prepared with m-dimensional arrays, where each array contains binary digit representing the presence of elements as shown below (table 4). Togo el al. used this descriptor to analyze the distributions of phonon lifetimes in Beillouin zones for compounds [71].

Table 4: Example of binary elemental descriptors [67]

	H	Li	Be	B	C	N	O	F	...
LiH	1	1	0	0	0	0	0	0	...
LiF	0	1	0	0	0	0	0	1	...
BeO	0	0	1	0	0	0	1	0	...
BN	0	0	0	1	0	1	0	0	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱

The another way to represent a compound as a descriptor is to combine structural and elemental representations. However, it is not as easy as mixing two descriptors, thus it is very important to consider both forms as a compound descriptors. Compounds are described by element types and neighbour environments that are affected by other atoms. Considering elemental descriptor as $N_{x,ele}$ and structural descriptor as $N_{x,st}$, each atom can be described by $N_x = N_{x,ele} + N_{x,st}$. Thus, a compound ξ can be described as a matrix, $\mathbf{X}^{\xi}$, with (N_a^{ξ}, N_x) dimensions, where N_a^{ξ} is a number of atoms in a unit cell of a compound ξ. An example of compound matrix can be described as,

$$\mathbf{X}^{\xi} = \begin{bmatrix} x_1^{(\xi,1)} & x_2^{(\xi,1)} & \dots & x_{N_x}^{(\xi,1)} \\ x_1^{(\xi,2)} & x_2^{(\xi,2)} & \dots & x_{N_x}^{(\xi,2)} \\ \vdots & x\vdots & \ddots & \vdots \\ x_1^{(\xi,N_a^{\xi})} & x_2^{(\xi,N_a^{\xi})} & \dots & x_{N_x}^{(\xi,N_a^{\xi})} \end{bmatrix} \tag{1}$$

where $x_n^{(\xi,i)}$ is n^{th} representation of atom i in compound ξ.

Structural descriptors The famous structural descriptor is developed by Himanen's research team, called DScribe [66]. DScribe is a python package that featurizes materials, a numerical fingerprints, from atomic structure data. DScribe provides different featurizing methods; coulomb matrix, sine matrix, ewald sum matrix, atom-centred symmetry functions (ACSF), smooth overlap of atomic positions (SOAP), many-body tensor representation (MBTR), local many-body tensor representation (LMBTR) and valle-oganov descriptor. Amongst all methods, SOAP encodes regions of atomic geometries by using a local expansion of a gaussian smeared atomic density with orthonormal functions based on spherical harmonics and radial basis functions. The output of SOAP is a partial power spectrum vector $\boldsymbol{p}$,

$$\boldsymbol{p}(x)_{b_1 b_2 l}^{\alpha\beta} = \pi \sqrt{\frac{8}{2l+1}} \sum_m (c_{b_1 lm}^{\alpha})^* c_{b_2 lm}^{\beta} \tag{2}$$

where b_1 and b_2 are indices for the different radial basis functions up to b_{max}, l is the angular degree of the spherical harmonics up to l_{max} and α and β are atomic species. Thus, the normalized polynomial kernel of the partial powers spectrum between two atomic environments is,

$$\mathbf{K}^{SOAP}(\vec{p}, \vec{p}') = \left(\frac{\vec{p} \cdot \vec{p}'}{\sqrt{\vec{p} \cdot \vec{p}} \vec{p}' \cdot \vec{p}'} \right)^{\xi} \tag{3}$$

Next, coulomb matrix descriptor is a global descriptor that mimics the electrostatic interaction between nuclei [72]. It is calculated by,

$$M_{ij}^{Coulomb} = \begin{cases} 0.5 Z_i^{2.4} & \text{for } i = j \\ \frac{Z_i Z_j}{R_{ij}} & \text{for } i \neq j \end{cases} \tag{4}$$

As can be seen in the equation 4, the diagonal element is a polynomial fit of the atomic energies to the nuclear charge Z_i. The other elements in the matrix represent the coulomb repulsion between atom i and j. It is hence useful to compare molecules and intuitive method to represent a molecule.

The MBTR is also a great descriptor, which encodes the structure by using a distribution of different structural motifs [61]. Therefore, it is easy to visualise input structure of a molecule. MBTR generates a geometry function, g_k, which transforms k atoms into a single scalar value. In addition, MBTR showed the lowest mean absolute error (MAE) in predicting energy of small organic molecules amongst coloumb matrix, bag of bonds, bonding angular machine learning, SOAP and MBTR [61].

Elemental descriptors The elemental descriptors are simply a representation of intrinsic quantities of elements. This includes physical, chemical, electrical and many more properties which also captures essential information about compounds. Such descriptors are hence used in building machine learning models and provides high prediction accuracy as shown in many researches [71, 73, 74].

Atomic descriptors It is always fascinating to transfer concepts from other fields. In the field of natural language processing (NLP), the simplest way to encode a word as a vector is using a one-hot vector [75]. However, the dimension of the vector increases by the number of words, and most of the vector elements are 0, which is inefficient use of memory. Thus Skip-gram and Word2Vec method were developed to construct database with dense vectors [76].

The main idea of Skip-gram is to predict surrounding words when a target word is given. A research team of Antunes came up with the idea that if machine can learn the relationship between words in a sentence, the relationship of atoms can also be learned from compounds [77]. They were inspired from Skip-gram model, thus named their model as SkipAtom; The crystal structure of materials are represented as a graph representing local connectivity as shown in figure 8(b). The result of SkipAtom is a set of vectors with dimension of d as shown in figure 8(a). These vectors can now show the magnitude of relationship with other atoms as the dot product is not zero.

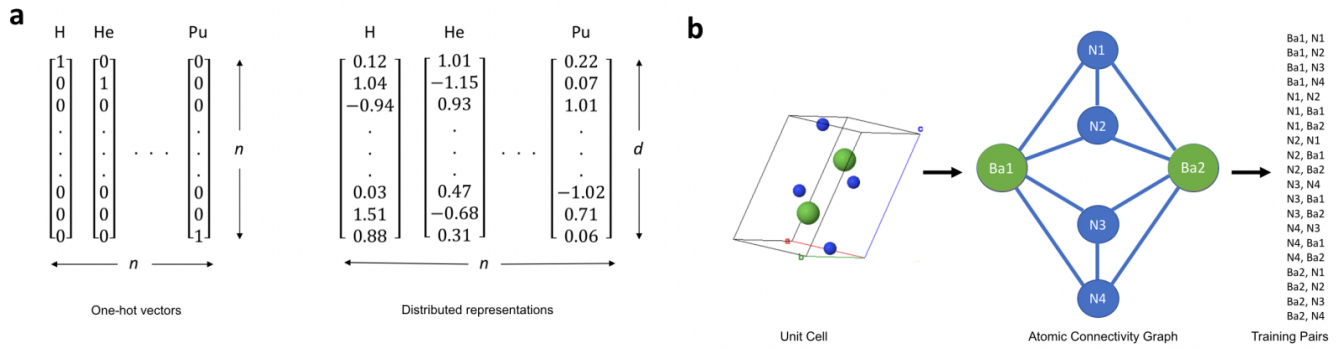


Figure 8: (a) Illustration of one-hot vector (sparse representation) and distributed (dense) representations (b)schematic representation of SkipAtom method [77]

Similarly, Word2Vec method was also used in material science for unsupervised word embeddings from materials science literature [78]. Research team of Tshitoyan has utilized NLP technique to generate information-dense vectors using 3.3 million material science literature. The main idea of Word2Vec is that similar words appear in the similar context, thus this can be applied as similar material will have similar properties. This information-dense word embedding support analogies. For instance, NiFe:ferromagnetic = IrMn:x, can be expressed by looking for nearest word neighbour by solving cosine similarity dot product. In this case, 'antiferromagnetic' will appear. Schematic representation of how this method can predict a material is shown below (fig 9).

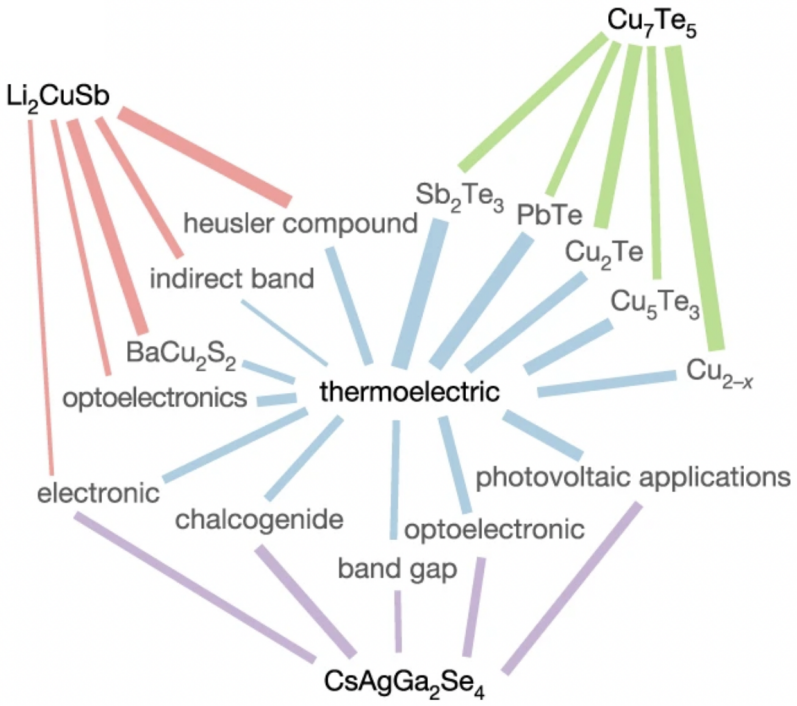


Figure 9: Illustration of how context words are predicted. Wider edges contribute more in prediction due to cosine similarity, and three materials (Li_2CuSb, $CsAgGa_2Se_4$ and Cu_7Te_5) are predicted [78]

2.2.2 Algorithms

There are many supervised/unsupervised machine learning frameworks for materials property prediction from descriptors. Matbench provides benchmark comparison data of machine learning algorithms for materials science [79].

2.3 Research Plan

Many state-of-the-art researches on materials machine learning were successful in predicting bulk properties, such as band gap, crystal structure and many more mentioned above. However, machine learning is not yet perfect to predict quantum properties, due to strong interaction between electrons which first-principle calculation cannot compute perfectly. Even more challenging calculation is studying the effects of defects in crystal to material properties due to vast number of cases in the defect search space [80]. However, research on this can provide a further insight to the researches on the effect of defects on materials such as Aron Walsh and Alex Zunger's research [81].

This project focuses on building models that describes and predicts the properties of point defects such as charged vacancies and interstitials. The performance of models are recorded by using different representation methods mentioned in section 2.2.

3 Methods

3.1 Data Preparation

The dataset was taken from the work by Yu Kumagai *et al.* [82]. The target metal oxides were retrieved from Material Project Database (MPD), and its application was done by PYMATGEN, which is a open-source python library for materials analysis [83, 84]. All materials were selected and preprocessed under several conditions: (i) stable metal oxide against its competing phases, (ii) band gap over 0.3eV, (iii) non-magnetic, and (iv) 30 or less atoms in their primitive cell. The oxygen vacancy formation energies of metal oxide supercells were then evaluated by the equation below [82].

$$E_f[V_O^q] = \{E[V_O^q] + E_{corr}[V_O^q]\} - E_p + \mu_O + q(\epsilon_{VBF} + \Delta\epsilon_f) \tag{5}$$

where E is total energy of a supercell in charge state of q, μ_O is chemical potential of oxygen, $\Delta\epsilon_f$ is the Fermi level to the valence band maximum (VBM), E_p and E_{corr} are total energy of pristine supercell and correction term respectively. With the equation, formation energies and the relaxed defect structures were calculated using density functional theory (DFT) from first principles.

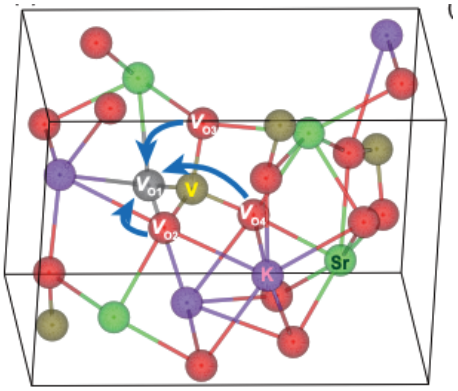


Figure 10: 4 different inequivalent oxygen sites of $KSrVO_4$, surrounding V ion. The blue arrows represent vacancy migration [82]

The dataset contains 1665 metal oxide structures, with 848 unique compositions. This is because some materials have more than one symmetry inequivalent oxygen sites. The metal oxides are comprised of 4 kinds of alkali metal elements, 5 kinds of alkali earth metal elements, 21 transition metal elements, 8 kinds of other metal elements, and Ce, which is a lanthanoid element. Besides B, Si, As and O, the compositional diversity of the oxides is pertinent for machine learning. In terms of structure, the dataset is comprised of 53 binary, 912 ternary, 673 quaternary, and 27 quinary metal oxide compounds. According to the crystal structure, the dataset is composed of 586 monoclinic, 455 orthorhombic, 215 trigonal, 194 tetragonal, 77 hexagonal, 70 triclinic, and 68 cubic structures. As mentioned, all materials are non-magnetic compounds.

3.2 Descriptors

Nowadays, machine learning techniques are utilised in everyday life. In the field of materials science, machine learning is used to extract meaningful patterns to predict material properties for advanced materials. The main factor that affects the performance of machine learning is the representation of existing materials, which is called features or descriptors. The way of choosing descriptors is trial and error, as it depends on the target, dataset, and its quality [67]. The descriptors must satisfy the following: easy to compute, low dimension and unique [85]. In this project, elemental representation - also known as compound based representation - will be focused and compared with different representations such as structural representation and graph based representation in section 4.4.

To develop a meaningful descriptors, CBFV (Composition-based fearture vector) - a python library for material featurisation - was used [86]. This technique allows to create a feature vector in the absence of structural information. A CBFV is one of the common ways to featurise chemical compounds for machine learning, and is generated by combining its compound's constituent element properties [87]. Element properties may include atomic number, atomic mass, ionisation energy, electronegativity, thermal conductivity, atomic radius, and many more descriptive statistics of element properties. The 6 different featurisation scheme were used in this project: Jarvis, Magpie, mat2vec, Oliynyk, Onehot and random-200.

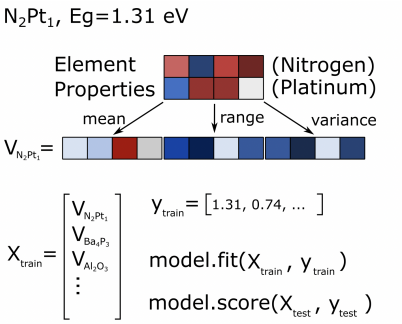


Figure 11: Construction of CBFV of N_2Pt_1 [88]

Jarvis, Magpie, and Oliynyk are all based on the elemental property databases. Each technique has unique spreadsheet that contains more than 100 encoded chemical properties on a per element basis, which is then combined in different ways. Figure 11 illustrates the construction of CBFV of N_2Pt_1. However, mat2vec uses completely different method to other techniques. The data is prepared by machine learning chemical domain knowledge through word embedding with NLP technique [78]. One-hot encoding technique contains no information about elemental properties, but it is a long vector size of total number of elements in periodic table. If an element is present in the compound, increment the number of element in the place for the element, and if it is not present, 0 is inserted. An illustration of one-hot encoding is shown in figure 8 (a). Lastly, a random-200 method contains a vector for each element, and each vector is encoded with 200 random numbers.

3.3 Feature Engineering

To avoid "curse of dimensionallity" and have accurate machine learning models, it is important to select features with high correlations with the target properties, and have low correlation with each features [89]. In other words, features need to embrace essential information with minimum redundancy. In machine learning, if the number of feature increases, the dimension of input data increases exponentially, leading to a sparse data density - with high data dimension, the distance between data also increases. The machine learning models hence become complex and the risk of overfitting rises, and it is likely to have high correlation between individual feature.

Principal component analysis (PCA), recursive feature elimination with cross-validation (RFECV), and correlation reduction methods were employed to reduce the dimensions of features. PCA is a method to minimise the loss of information while reducing dimensions. It is done by finding the axis in which covariance transforms the shape of data and the axis orthogonal to it. Recursive feature elimination (RFE) is, in some respects, the simplest method of feature selection. It is a process where removing the feature with lowest feature importance until the desired number of features is reached. In this way, features with a higher feature importance that corresponds to the desired number of features become the final feature selection result. RFE has shortcomings in selecting the number of features, but RFECV allows to select the number of features to be automatically selected with the highest performance. Lastly, the correlation reduction method removes highly correlated features, verifying the independence of features. The Pearson coefficient (r) is used to check the correlation, and is calculated through following equation:

$$r = \frac{\Sigma_i^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\Sigma_i^n (X_i - \overline{X})^2}\sqrt{\Sigma_i^n (Y_i - \overline{Y})^2}} \tag{6}$$

The equation is simply calculating cosine similarity of normalised vectors, X and Y. The correlation coefficient has a value of $-1 \leq r \leq 1$, and if $r = 0$, it does not mean they are irrelevant, but are not in a linear correlation. Here, $|r| < 0.8$ indicates that there is a strong correlation between the features, whilst a value of $|r| > 0.8$ indicates the features can be identified as independent [90]. PCA and RFECV were done with scikit-learn, a open source python library for machine learning and data analysis [91].

3.4 ML Model Development

The choice of algorithms is as important as data to develop models with high accuracy. When choosing a model, there is no perfect solution or approach that is suitable for every problem. There are several factors to consider, and in most cases, an important criteria is performance. Factors that needs to be considered are as following: define problem (supervised, unsupervised, etc.), data size, interpretability vs accuracy, linearity, and data dimension (feature size). To develop an efficient machine learning model with a consideration of given dataset, a random forest decision tree algorithm was employed.

A random forest algorithm uses a decision tree model, which is a model of tree structure used for predicting or classifying target variables according to the relationship or scale between explanatory variables. In other words, it is a supervised learning-based methodology that classifies or predicts target variables by entering observations of explanatory variables into the model. The main reasons for using the decision tree model are: tt has the advantage of identifying which explanatory variable is the most important factor in predicting the target variable or solving the classification problem, and furthermore, knowing the detailed criteria for each explanatory variable based on which scale it was predicted or classified. It is also easy to interpret the results, as the separation criteria is clear and intuitive, and has freedom of assumptions required by statistical models such as normality independence and equivariance, and allows navigation of linear or non-linear relationships by considering interactions between variables.

To prevent overfitting by classifying data with chosen features, random forest creates more than 100 decision trees by randomly sampling features to be classified. This process is called bootstrapping. Decision trees are trained on each of the bootstrapped dataset independently, and aggregate the result trees to obtain an averaged result. The bootstrapping and aggregating (bagging) process improves the model accuracy by avoiding overfitting issues that a decision tree model had before.

$$MAE = \frac{1}{N}\sum_{i=1}^{n} |E_{V,i} - \hat{E}_{V,i}| \tag{7}$$

$$MSE = \frac{1}{N}\sum_{i=1}^{n} (E_{V,i} - \hat{E}_{V,i})^2 \tag{8}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{n} (E_{V,i} - \hat{E}_{V,i})^2} \tag{9}$$

$$r2 = 1 - \frac{MSE}{Var(E_{V,i})} \tag{10}$$

When evaluating a regression model, it is common to use indicators using the difference between the measured value and the predicted value. Four indicators, mean average error (MAE), mean squared error (MSE), root mean squared error (RMSE), and r2 score are used to evaluate the performance of models, and defined above. N is the number of samples in the test set, $E_{V,i}$ and $\hat{E}_{V,i}$ is the i^{th} target and predicted values from the developed model respectively.

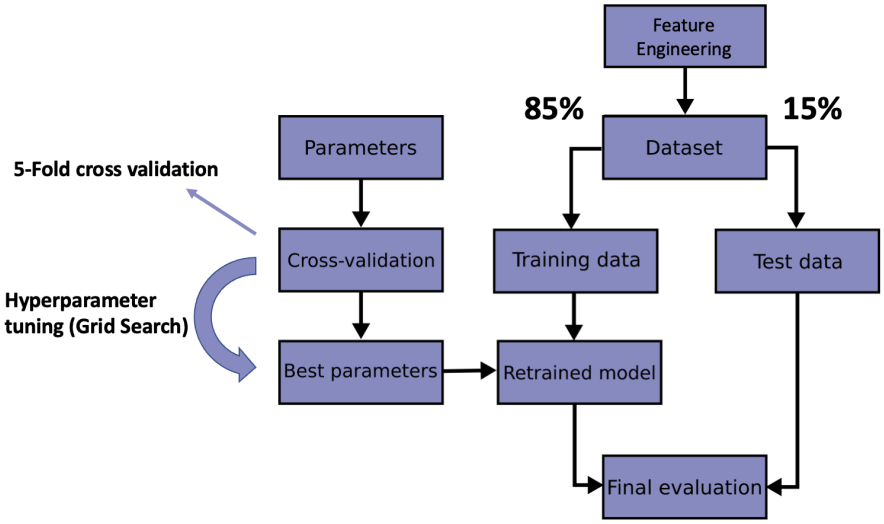


Figure 12: Schematic illustration of basic, high-level workflow of a machine learning project [91]

The basic workflow of machine learning is illustrated in figure 12. Once a data is selected, the task is then chosen. The dataset is explored and investigated before the feature engineering process, which is transforming the data provided to create new columns in the dataset to aid the model. The model is then trained on a portion of the dataset (85%) and evaluated on the remainder of the dataset (15%). If initial performance is good, then properties of the model called hyperparameters can be tuned to improve results. Otherwise, a different model is tested and the process begins again.

Most of the basic machine learning model is trained with a train set and the model is then evaluated with a test set. However, evaluating and modifying the performance of the model through a fixed test set will only show a high performance on the test set. In other words, the model is over-fitted, and predictions on the new dataset will be unreliable. Here, cross validation is one of the solutions, which separates dataset into a training set and a validation set, and then evaluate the model with the validation set. In the process of hyperparameter tuning, 5-fold cross validation technique was used - 80% of training set 20% of validation set, repeated 5 times. Cross validation has strong advantage in preventing underfitting due to small dataset, and the accuracy can thus be improved. Moreover, all dataset is used for the model evaluation and hence prevents data bias. However, if the number of iteration is high, the model training and evaluation will take considerable time. The process of hyperparameter tuning, hence, is a process of compromisation between computational time and accuracy.

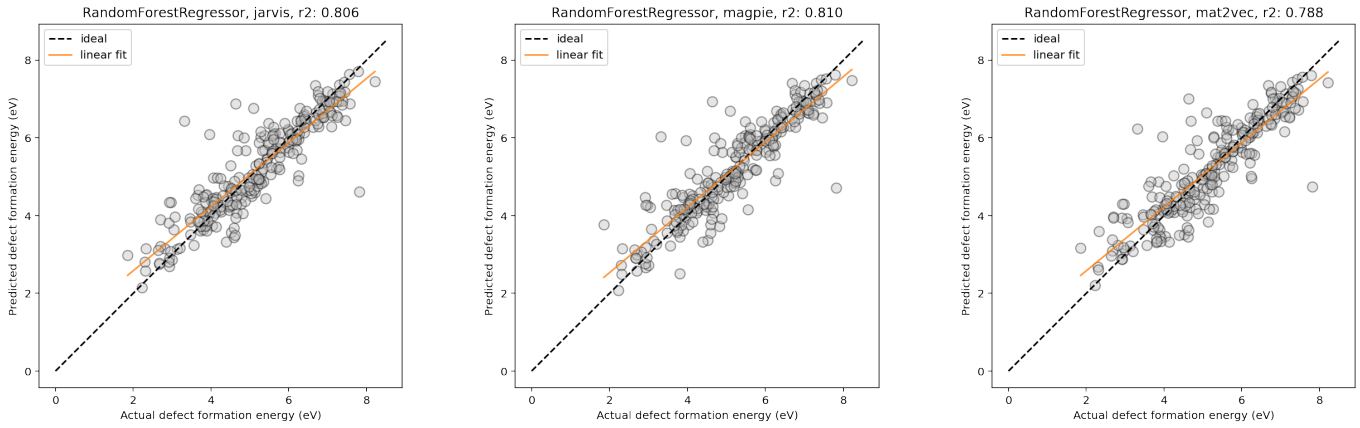
4 Results and Discussion

The models were trained with random forest algorithm as well as gradient boosting algorithm. It is based upon using an ensemble of decision trees, similar to a random forest algorithm. It uses custom trees that are built iteratively to predict the difference between an initial guess and the actual values. It differs from random forests as well in that trees that don't improve predictions can be pruned and removed. Trees are built to incrementally improve prediction performance as this small step approach has been empirically shown to generalise better. The outputs from the trees are multiplied by the learning rate, eta, to reduce the sensitivity of the algorithm to single tree outputs.

4.1 Initial Model Training

Table 5: Evaluation metrics of initial model training

Featurising Scheme	Model	r2 score	MAE	RMSE
Jarvis	Random Forest	0.806	0.401	0.592
	Gradient Boosting	0.801	0.420	0.599
Magpie	Random Forest	0.810	0.395	0.586
	Gradient Boosting	0.776	0.471	0.636
mat2vec	Random Forest	0.788	0.422	0.619
	Gradient Boosting	0.793	0.435	0.611
Oliynyk	Random Forest	0.807	0.388	0.590
	Gradient Boosting	0.792	0.445	0.613
Onehot	Random Forest	0.778	0.418	0.633
	Gradient Boosting	0.659	0.599	0.785
random200	Random Forest	0.783	0.432	0.626
	Gradient Boosting	0.733	0.500	0.694



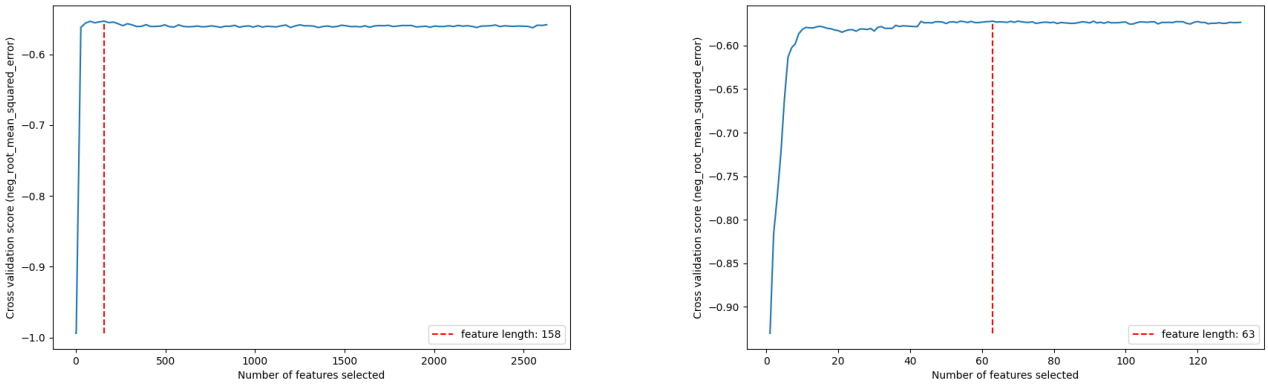
(a) Jarvis **(b)** Magpie **(c)** Oliynyk

Figure 13: Plots of DFT experimental data of oxygen vacancy formation energy and regression predicted values through random forest regressor on the various descriptors. Orange lines are fitted line of predicted formation energies, and black dotted line represents ideal result

The unprocessed CBFVs were initially trained with random forest regressor and gradient boosting regressor. The dimension of Jarvis, Magpie, mat2vec, Oliynyk, onehot and random200 features were 2628, 132, 1200, 264, 714 and 1200 respectively. The results are shown in the table 5. Also the linear regression of experimental data and predicted data on the testing data is illustrated in figure 13. Here, Jarvis, Magpie and Oilynyk performed well with random forest model in predicting the formation energy. The r2 score depicts that the predicted values of formation energy through gradient boosting regression is more predictive than the random forest regression except mat2vec. In terms of errors, Jarvis, Magpie and Oilynyk descriptors obtained the lowest RMSE in the range of 0.590 to 0.599 eV. Referring to the figure 13, the general trend of prediction can be noticed with few outlier data points. The most frequent outliers for all models were $Y_2Pt_2O_7$ in the low energy region, and $Y_2Sn_2O_7$ in the high energy region. Both metal oxides have two inequivalent oxygen vacancy sites. The initial models performed reasonably in prediction, but the error is large compared to the range of the dataset (1-9 eV).

4.2 Feature Engineering

Recursive Feature Elimination with Cross-Validation (RFECV) To increase training accuracy, the recursive feature reduction with cross-validation was employed. By sequentially removing the least important feature and evaluate at each step, the best performing features were selected. Due to the computation time, this method was done with random forest regression model, to compare with the initial results.



(a) Jarvis **(b)** Magpie

Figure 14: Plots of recursive feature elimination on Jarvis and Magpie. Red lines illustrates the optimised feature length

Figure 14 illustrates feature reduction and corresponding cross validation results (RMSE) on Jarvis and Magpie. The best length for the descriptors are shown in table 6. All features had a dramatic decrease in the feature dimension, but some features had little change in performance regardless on the feature size as shown in figure 14a.

Comparing with the initial results, most of the features had enhanced performance except Magpie and Oliynyk. The recursive feature reduction method does not seem to be a good method for the performance improvement. However, still Jarvis and Magpie have the best performance in predicting the oxygen vacancy formation energy with RMSE of 0.586 and 0.597 respectively.

Table 6: Evaluation metrics on reduced feature with RFECV

Featurising Scheme	Feature Length	r2	MAE	RMSE
Jarvis	158	0.810	0.392	0.586
Magpie	63	0.803	0.394	0.597
mat2vec	84	0.792	0.420	0.613
Oliynyk	64	0.801	0.396	0.600
Onehot	259	0.765	0.421	0.652
random_200	192	0.773	0.433	0.640

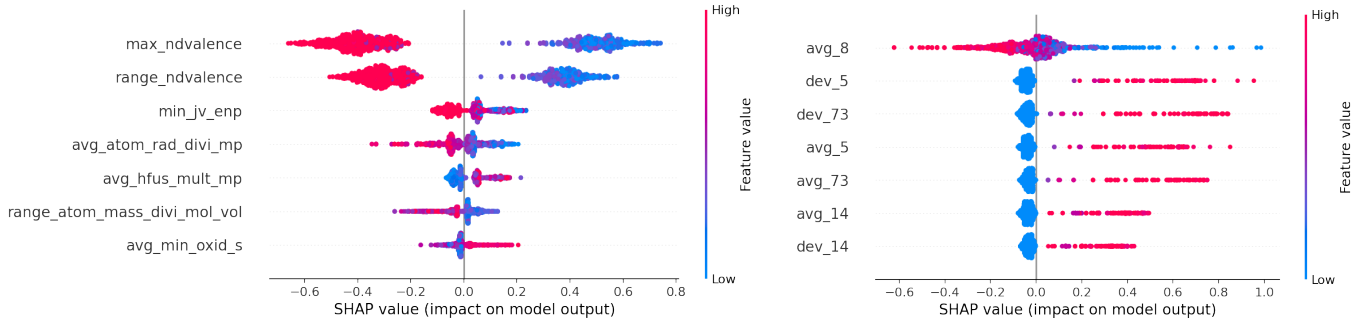
Principal Component Analysis (PCA) PCA is usually to summarise linear dataset to a lower dimension. However, if the given dataset is not linear, another method needs to be introduced. The Kernal PCA (KPCA) is an extension of PCA using a 'Kernel' function. The basic idea of KPCA is to project dataset into a higher dimension to separate, or classify, them. There are no clear performance metrics for selecting good kernels and hyperparameters, as KPCA is an unsupervised learning. However, dimensional reduction is often used as a preprocessing step in supervised learning, so grid navigation can be used to select the best performing kernel and hyperparameters for a given problem.

Table 7: Evaluation metrics of models trained with KPCA

Featurising Scheme	Model	r2 score	MAE	RMSE
Jarvis	Random Forest	0.755	0.462	0.666
	Gradient Boosting	0.724	0.524	0.707
Magpie	Random Forest	0.745	0.473	0.678
	Gradient Boosting	0.718	0.515	0.714
mat2vec	Random Forest	0.725	0.486	0.705
	Gradient Boosting	0.742	0.512	0.682
Oliynyk	Random Forest	0.750	0.460	0.673
	Gradient Boosting	0.759	0.481	0.660
Onehot	Random Forest	0.731	0.484	0.697
	Gradient Boosting	0.709	0.532	0.725
random200	Random Forest	0.704	0.508	0.731
	Gradient Boosting	0.718	0.523	0.714

The dimension of all dataset were reduced to 200 features. Due to the computational cost, the hyperparameter tuning for KPCA was not done. Since the reduced features were very sparse in values, they were standardized with standard scaler by removing the mean and scaling to unit variance. Amongst different types of kernel - polynomial, Radial Basis Function (RBF), sigmoid, and cosine - polynomial kernel was used as it had the best performance. Referring to the table 7, the performance of all models were poor. The r2 score decreased to 0.704 - 0.755 and RMSE increased to 0.650 - 0.731. The result suggests that the hyperparameter tuning and number of features to be selected are necessary for KPCA, in order to have reliable results.

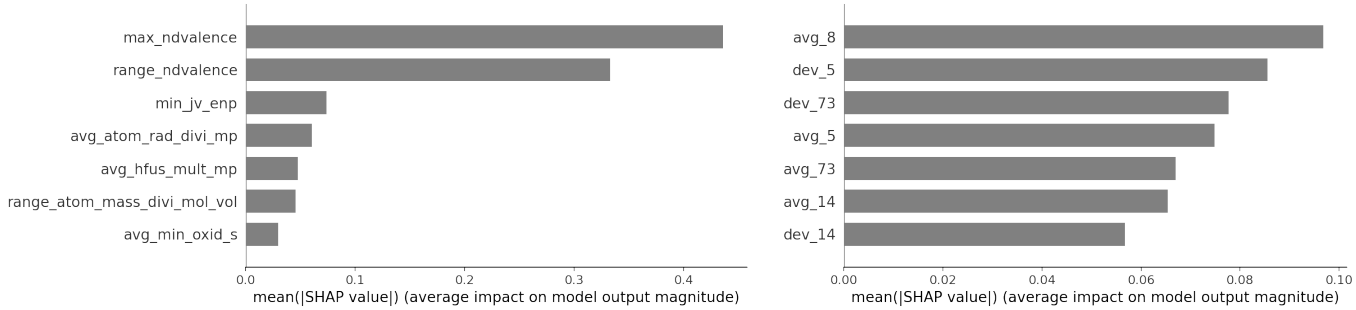
Correlation Reduction (Pearson Correlation) In the process of feature reduction, it is important to inspect the feature importance, and how they contribute to the prediction result. Here, the Shapley value was used, which calculates the importance of a variable compared to the case where the model has or does not have a specific variable dependence. However, the order in which the model inspects variables may affect the calculation, thus the variables are calculated in every possible combinations for equal consideration.



(a) Jarvis **(b)** onehot

Figure 15: Distributions of the SHAP values of compositional features and elemental fractions on the model output. The color represents the feature value (red high, blue low), and x-axis represents its contribution to the prediction of oxygen vacancy formation energy. Here only the top 7 features with the highest sum of absolute SHAP values are shown

Figure 15 shows the impacts of top 7 compositional structural features with the shap values on the model output. As can be seen, number of valence d-orbitals (ndvalence), energy per atom of an element from JARVIS-DFT (jv_enp), atom radius (atom_rad), enthalpy of fusion (hfus_mult_mp), atom mass per molar volume (atom_mass_divi_mol_vol), and oxidation state (oxid_s) were the most impactful properties with respect to the oxygen vacancy formation energy [92]. Most features of Jarvis shows distinct contribution on the oxygen vacancy formation energy, as negative and positive SHAP values represent their negative and positive contributions. For instance, by referring to the Jarvis plot (fig. 15a), the smaller maximum number of valence d-orbitals tends to have higher oxygen vacancy formation energy. On the other hand, most of the features are contributing positively on the predicted values with higher values. The difference in the contribution on the model (fig. 16) may affect the accuracy, thus feature reduction based on the correlation is essential.



(a) Jarvis **(b)** onehot

Figure 16: Illustration of the average impact of the feature by taking the absolute value of each feature's shape value. The greater the impact, the greater the relationship (not causality) with the target value

Jarvis + Magpie Since Jarvis and Magpie obtained the best performance with any method, it is likely to be effective in producing high accuracy model by combining two features. In other words, high performance regardless of the model depicts both feature may encapsulate essential compositional information. As mentioned in section 3.3, it is crucial to extract features that have Pearson correlation less than |0.8|. Figure 17 below illustrates a Pearson coefficient heatmap of concatenated descriptor. As can be seen, there are features that are highly correlated, which hinders the model performance. The dimension of the descriptor was reduced from 2760 to 517, meaning 2243 features were highly correlated and impeded the performance.

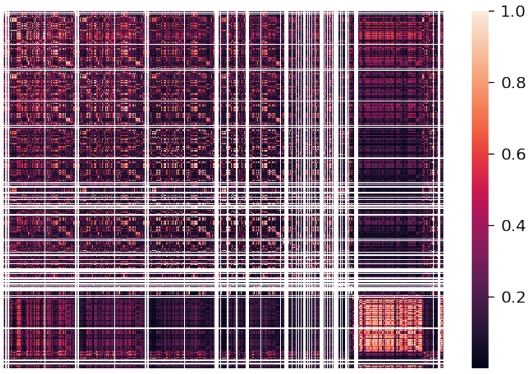


Figure 17: Pearson coefficient heatmap of concatenated Jarvis and Magpie descriptors

To further enhance the model performance, the hyperparameters of models were tuned with grid search with 5 fold cross validation. The grid search is a method to find the best parameter by trying all possible combinations of the parameters of interest. The best parameters for random forest regressor and gradient boosting regressor are listed below. The RMSE of random forest had the best performance with RMSE of 0.566 eV.

Table 8: Evaluation metrics of concatenated Jarvis and Magpie

Featurising Scheme	Model	r2	MAE	RMSE
Jarvis + Magpie	Random Forest	0.823	0.399	0.566
	Gradient Boosting	0.805	0.396	0.593

```
max_depth: 12,
min_samples_leaf: 8,
min_samples_split: 16,
n_estimator: 100
```

(a) Random Forest

```
learning_rate: 0.01,
max_depth: 4,
n_estimators: 1000,
random_state: 1,
subsample: 0.75
```

(b) Gradient Boosting

Figure 18: Selected parameters through grid search for different models

Smooth Overlap of Atomic Position (SOAP) Composition feature itself cannot connote all essential information of chemical compounds. To obtain more accurate and reliable model, the structural featurising scheme has been added. The Smooth Overlap of Atomic Position (SOAP) featuriser was used for the structural feature, which maps the local environment around sites very accurately. Since the atomic positions are points, where the number of basis functions required needs to be reduced for lower computational cost, hence the atomic positions are converted to atomic densities through Gaussian smoothing. Gaussian smoothing would remove the ability to differentiate between individual elements, thus SOAP is calculated for each present element and these values are concatenated. The smeared densities are decomposed using spherical harmonics in real space (for computational reasons) and a chosen orthogonal basis set and integrated [93].

The SOAP features are the partial power spectrum vector, $\mathbf{p(r)}$, where:

$$p(\mathbf{r})_{nn'l}^{Z_1Z_2} = \pi\sqrt{\frac{8}{2l+1}}\sum_m c_{nlm}^{Z_1}(\mathbf{r})^* c_{n'lm}^{Z_2}(\mathbf{r}) \tag{11}$$

where n and n' are indices of different radial basis function, l is the angular degree of spherical harmonics, and Z_1 and Z_2 are atomic species [94]. The coefficients, c_{nlm}^Z are defined as:

$$c_{nlm}^Z(\mathbf{r}) = \iiint_{R^3} dV\, g_n(\mathbf{r}) Y_{lm}(\theta,\phi) p^Z(\mathbf{r}) \tag{12}$$

where $p^Z(\mathbf{r})$ is Gaussian smoothed atomic density for atoms with atomic number Z, $Y_{lm}(\theta,\phi)$ is the real spherical harmonics, and $g_n(\mathbf{r})$ is the radial basis function [95].

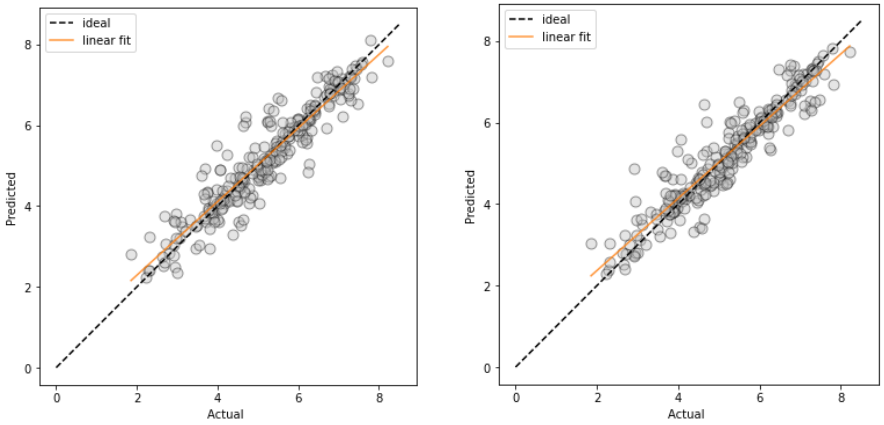
SOAP is site-specific since it is a local environment descriptor, and thus a summariser needs to be used. A fingerprint is used for this purpose to reduce the feature size through using the mean and standard deviation of the attributes, rather concatenating all sites. In this project, delta SOAP is used to emphasise the differences between the host feature vector and the defect feature vector.

Different SOAP descriptors are generated, and preprocessed in two different ways: correlation reduction (corr_reduced) and delta vector after correlation reduction (corr_reduced_delta). The dimension of correlation reduced SOAP descriptor and delta descriptor were 770 and 3281 respectively.

Referring to the scatter plot (fig. 19), and the corresponding RMSE values (table. 9), the model trained with Jarvis, Magpie, and correlation reduced delta vector of structural descriptor has the best performance with RMSE value of 0.465 eV. The model seems to overestimate the lower energies and underestimates the higher energies with a few larger outliers on the higher energy region.

Table 9: Evaluation metrics of concatenated Jarvis, Magpie, and SOAP descriptors

Featurising Scheme	Model	r2	MAE	RMSE
corr_reduced	Random Forest	0.810	0.385	0.586
	Gradient Boosting	0.878	0.331	0.470
corr_reduced_delta	Random Forest	0.862	0.368	0.533
	Gradient Boosting	0.880	0.321	0.465



(a) corr_reduced **(b)** corr_reduced_delta

Figure 19: Plots of DFT experimental data of oxygen vacancy formation energy and regression predicted values through gradient boosting regressor on different descriptors

4.3 Classification Model - Confusion Matrix

It is important to accurately predict the defect formation energy, but is also important to classify materials with high and low energies. The defects will not form easily if defects are high in formation energy. Therefore, the classification model could be used to identify which defects are important for further study - additional DFT calculations or to look out for in experiment.

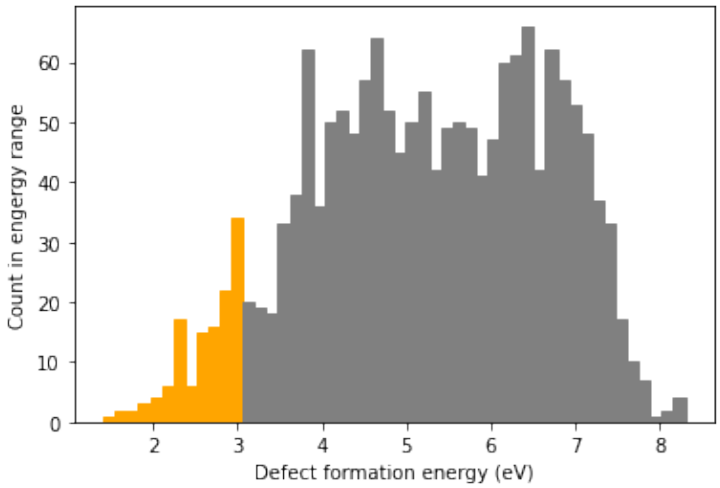


Figure 20: Histogram of count in energy range for the dataset. Orange bars represent lower energy materials and grey bars represent high energy materials

The gradient boosting classification was done with the dataset by 5-fold cross validation. Low formation energy was classified as lower 10^{th} percentile, which is below 3 eV and vice versa.

The evaluation metrics for classification models are different to regression models. The commonly used evaluation metrics for classification model are accuracy, precision, recall, and F1 score, which are mainly based on confusion matrix. Confusion matrix shows how accurately predict predicted the actual observed value. The confusion matrix summarises four situations by frequency: 1. prediction is true and it is actually true, 2. prediction is true, but it is actually false, 3. prediction is false, but it is actually true, and 4. prediction is false and it is actually false. By comparing the frequency of each cell with each other, one can examine how large the cell with wrong prediction is and how much the cell is treated properly. In order to perform well, of course, the frequency of cells with wrong guess should be as low as possible and the frequency of properly processed cells should be as high as possible.

It should be noted that the evaluator must have information of actual values in advance to compare with the predicted values. For instance, an exam cannot be graded without a mark scheme. Although it is often known as binary logic, which is treated in the form of positive or negative, the confusion matrix does not appear only in the form of a 2 by 2 matrix. It is also possible to represent the results of a multi-class classification with three or more classes as a confusion matrix. For instance, if a person is asked to guess Koreans, Japanese, or Chinese by showing pictures of the faces of Northeast Asians, the confusion matrix at this time appears in the form of a 3 by 3 matrix. In this case, it can be said that the cells, which were only marked as 'negative' in the existing 2 by 2 matrix, are subdivided.

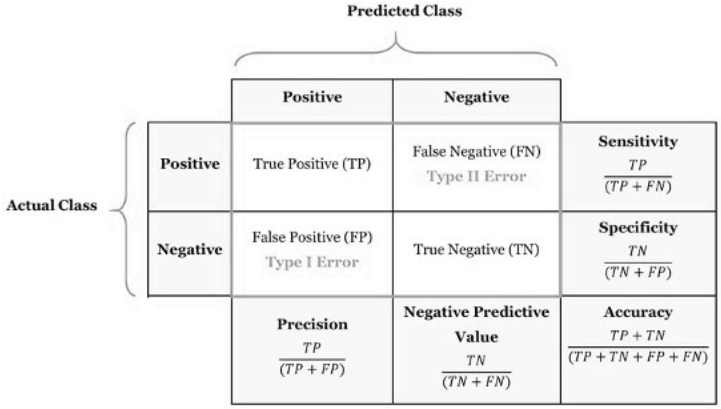


Figure 21: A confusion matrix with evaluation metrics [96]

Accuracy Accuracy is the probability that describes how the prediction fits the actual data. The sum of entire prediction result is in the denominator, and the frequency of successful prediction, whether it is true positive or true negative, is then evaluated as a value between 0 and 1. High accuracy means that predictions are often right, and prediction algorithms with high accuracy are highly likely to be used. On the other hand, the reversed accuracy (1 - Accuracy), is the proportion of false predictions among all predictions, and this value, which calculates false positives and false negatives, is called an error rate. From the perspective of data, the higher the accuracy, the lower the bias of the data, hence it can be evaluated as utilisable data. Since it is in conflict with the precision, most data analyses aim to compromise the two appropriately.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{13}$$

Precision Precision is the probability of actual value being positive when the prediction result is positive. Here, precision solely focuses in the case where the prediction results are positive - the true and false positives are added in the denominator - while numerator only contains the true positives. The high precision signifies that positive predictions are mostly right, and hence prediction algorithms with high precision are recognized to have high stability. However, precision does not provide information on how much to trust the prediction results when they are negative. Precision is used to calculate the F1 value below. Similar to accuracy, the reverse precision (1 - Precision) is the probability of obtaining positive predictions while the actual values are negative, and is called a negative predictive value. From the perspective of data, the higher the precision,

the lower the variance of the data, and thus, it can be evaluated as data with high stability.

$$Precision = \frac{TP}{(TP + FP)} \tag{14}$$

Recall (Sensitivity, True Positive Rate (TPR) Recall is the probability of having positive actual values and positive prediction results. The high sensitivity indicates high rate of obtaining positive predictions with actual positive values. However, sensitivity does not provide information on how predictions are made when the actual value is negative. The sensitivity is also used to calculate the F1 value below, and is also used to calculate the ROC curve.

$$Recall = \frac{TP}{(TP + FN)} \tag{15}$$

In the contrary, true negative rate (TNR) is calculated with equation below:

$$Specificity = \frac{TN}{(TN + FP)} \tag{16}$$

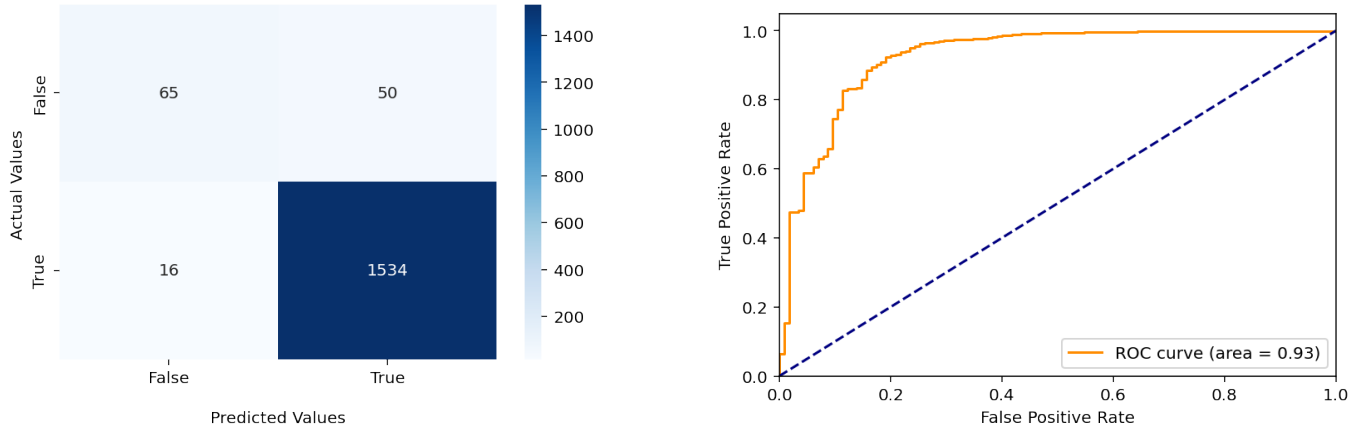
F1 score F1 score is an evaluation metric that utilizes precision and recall rate (sensitivity). F1 score is obtained by multiplying precision (P) and recall rate (R), and then multiplied by 2 to be weighted. The result of the calculation appears to be a value between 0 and 1. If the frequency of the three cells is the same with $TP = FP = FN$, the F1 value will be 0.5. As long as the TP cell does not change, the F1 value does not reflect the frequency difference between FP and FN if their sum is consistent. On the other hand, if TP is constant, the greater the sum of the frequencies of FP and FN, the greater the F1 score.

$$F_1 = \frac{2PR}{P + R} \tag{17}$$

ROC curve The Receiver Operating Characteristic (ROC) curve is a graph of FPR and TPR on the x and y axes, respectively. The ROC curve represents the change in FPR and TPR with the change in the threshold of the model, and is a curve connecting (0,0) and (1,1). The ROC curve is uesd to determine which model performs well. In other words, ROC curve is one of the evaluation criteria to select the best model that exhibits high sensitivity and specificity. Skewed ROC curve to the upper left indicates high performance. Area Under the Curve (AUC) is a numerical criterion in performance evaluation, and high performing models have the value closer to 1.

Table 10: Evaluation metrics gradient boosting classification model

	Precision	Recall	F1 score
Low energy	0.80	0.57	0.66
High energy	0.97	0.99	0.98
Accuracy			0.96



(a) Confusion matrix

(b) ROC-AUC curve

Figure 22: (a) Confusion matrix for classification of low and high formation energies (low: False, high: True), (b) ROC-AUC curve of classification model

The classification of low and high energy materials were machine-learnt with Jarvis + Magpie + corr_reduced_delta descriptor through gradient boosting classifier. All descriptors were correlation reduced, and depth and number of trees were set to 5 and 1500 respectively.

The classification results can be seen in figure 22a, which shows the numbers of TN, FN, FP and TPs. 65 out of 115 low energies were classified correctly. Moreover, there were 16 out of 1550 of the high energies that were falsely classified as high energies (16 false positives). The precision and recall were then calculated and the curve was plotted in figure 22b. The recall values for low and high energies were 0.57 and 0.99 respectively. The high difference in recall values can be observed due to the imbalanced dataset. In other words, the number of low and high energy materials were 115 and 1550 respectively. Since classical machine learning models assume a balanced class distribution, the performance of model is poor in classifying low energy materials. Classifying imbalanced data is a fundamental challenge for machine learning because of the skewed class distribution. Nevertheless, the final model was prominent with the accuracy of 0.96.

4.4 Discussion

The best model for regression, so far, was gradient boosting regressor with concatenated Jarvis, Magpie, and corr_reduced_delta descriptors. Random forest and gradient boosting algorithm is a recursive binary splitting procedure to obtain homogeneous (zero variance) or near-homogeneous terminal nodes [97]. The main purpose of the algorithms is to ensure improved homogeneity of two daughter nodes than their parent node. The random bootstrapping method, where the number of randomly chosen features are smaller than the descriptor size, decorrelates trees in the algorithm, thus reduce variance. In splitting trees into two daughter nodes, the splitting rule plays an important role in the performance of the model. Basic idea of splitting is based on the recursively 'removing impurities' [98].

The scikit-learn package, gradient boosting regressor utilises 'Friedman MSE' criteria to remove impurity of the current node and minimise it in trees. Friedman MSE impurity criterion uses following equation to improve purity (improvement = $i^2(\mathbf{R}_l, \mathbf{R}_r)$:

$$i^2(\mathbf{R}_l, \mathbf{R}_r) = \frac{w_l w_r}{w_l + w_r}(\overline{y}_l - \overline{y}_r)^2 \tag{18}$$

where $\overline{y}_l$ and $\overline{y}_r$ are left and right daughter response means respectively, w_l and w_r are corresponding sums of the weights and $\mathbf{R}_l, \mathbf{R}_r$ are split two sub regions.

$$w_l = \sum_{i \in \mathbf{R}_i} w_l(x_i) \tag{19}$$

Here, weight of left daughter node (w_l) is the sum of the possibilities where in the given region l, every x_i belongs to the certain k-class, which are the dependent variables in the regression.

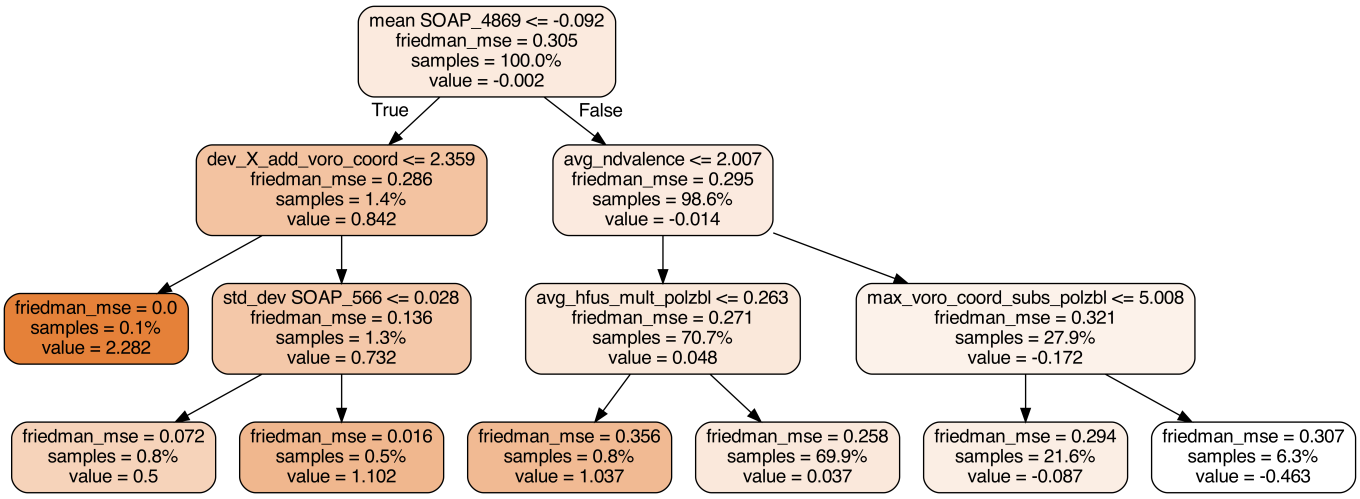


Figure 23: Visualisation of first decision tree (out of 1500) in random forest regressor

The final model consists 1500 trees, and the first tree is illustrated above. Each tree employs all features from bootstrap process, and each node produces daughter nodes with respect to lowest Friedman MSE value. The node on the left in the second depth (root node has depth of 0) have zero Friedman MSE, indicating a homogeneous node. Thus it terminates splitting and provides a value for 0.1% of samples. As can be noted, the right node in the first depth splits with the number of valence d-orbitals (ndvalence), where the value below 2.007 negatively affects to the value of the oxygen vacancy formation energy. Referring to the figure 24 below, the lower the ndvalence, the lower the oxygen vacancy formation energy can be observed.

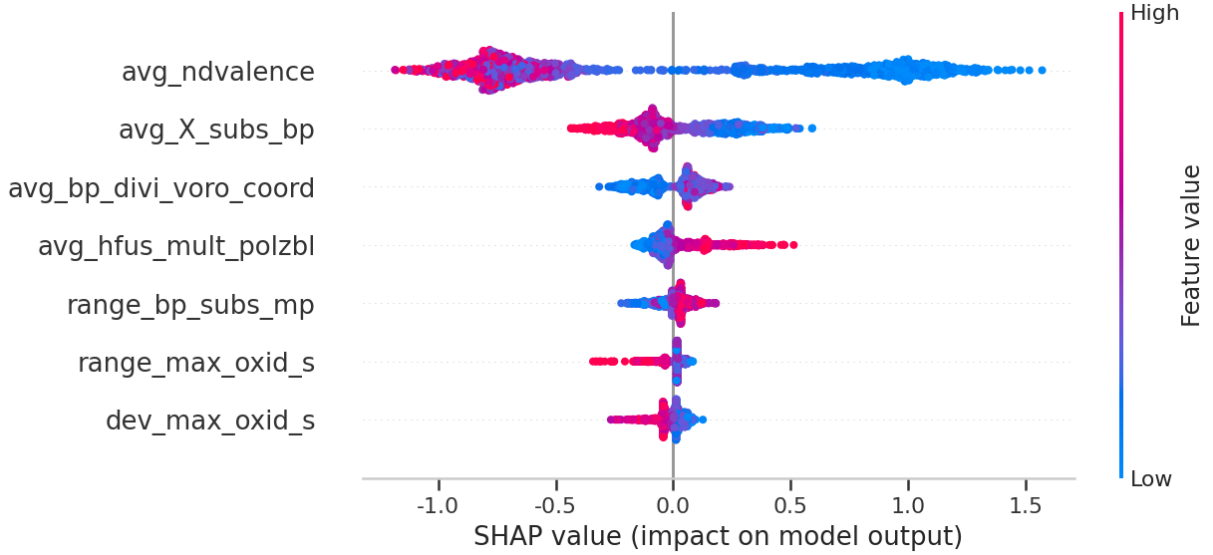
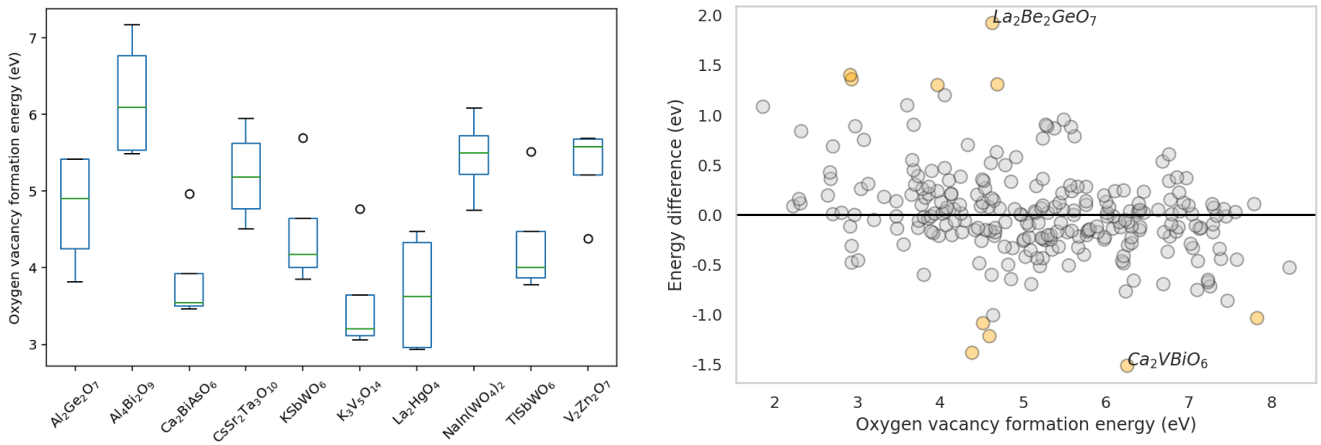


Figure 24: Distribution of the SHAP values of features in concatenated Jarvis, Magpie, and corr_reduced_delta on the gradient boosting regressor output. (red high, blue low). Top 7 features with the highest feature importance is illustrated

From the observation for all models, the model tends to overestimates the lower energy materials and underestimates the higher energy metal oxides. Figure 25b illustrates a scatter plot of the actual defect formation energy against the difference between actual and predicted energies with top overestimation and underestimation is labeled. The highest overestimated metal oxides are $LaAuO_3$, $K_3Nb_3(BO_6)_2$, $La_2Be_2GeO_7$, La_2HgO_4, and $BaBiBO_4$ which have 2, 3, 3, 4, and 3 inequivalent oxygen sites respectively. In contrary, the highest underestimated metal oxides are Ba_4AgAuO_6, Ca_2VBiO_6, $Ca(AsO_3)_2$, $Y_2Sn_2O_7$, and As_2PbO_6 which have 2, 2, 1, 2, and 2 inequivalent oxygen sites respectively. Since compositional descriptors only encapsulates the information of its chemical formula, the best accuracy of the model is limited by the difference in energies between multiple inequivalent oxygen vacancy formation energies. The original dataset contains 1612 metal oxides that are not binary metal oxides, thus the error of the predicted values can be explained. The box plot below illustrates the huge difference in oxygen vacancy formation energy of metal oxides with 4 inequivalent sites, which is also related to the prediction error. RMSE value of 0.46 eV can be considered as a reasonable result, considering the range of the data sample, which is from 1 to 9 eV. However, the targeted value of RMSE was 0.1 eV as for the model to be a reliable and highly accurate model, therefore it will be a further work optimising and improving the models.



(a) Box plot

(b) Scatter plot

Figure 25: (a) Box plot of the formation energies of the top 10 largest spread structures with 4 inequivalent sites, (b) Scatter plot of the actual defect formation energy on the x-axis against the difference between actual and predicted energies on the y axis. Orange dots represents top 5 overestimation and underestimation with labeled chemical formulas

The performance of the classification model was reasonable in identifying materials with high oxygen vacancy formation energy. However, the model had low recall value with 0.57, signifying only more than half of the materials with low energy are successfully classified. This problem has arouse due to highly skewed dataset. The imbalance in the data was recognized prior to machine learning, but machine learning without preprocessing caused a severe difference in the prediction, which differed sigificantly from the actual value. The gradient boosting model itself solved imbalance problem through setting a parameter - class_weight = 'balanced' - but the biased result still occurred as the imbalance distribution difference was 1:9.

There are three main solutions for the imbalanced dataset: under-sampling, over-sampling, and hybrid sampling. Under-sampling - so called down-sampling - refers to the process of trimming a high distributed data to match number of samples in total. This method can leave meaningful data, but there is a risk of losing some information. Since the number of metal oxides in the dataset is 1665, under-sampling will not improve the model performance as the number of samples to be learnt is insufficient. In contrast, over-sampling, also known as up-sampling, is a sampling method that matches the values of classes with low distributions to classes with high distributions. Over-sampling can prevent the loss of information in the perspective of high distribution class, but the model can be overfitted due to the redundant data on the low distribution class. Applying the same logic, over-sampling also cannot enhance the model performance due to data scarcity.

One of the solutions can be collecting more oxygen vacancy formation energy data through DFT. More accurate and abundant dataset will improve the performance, and will also affect its recall value. However, the endemic problem of composition based features may limit the model performance as it cannot describe the energy difference due to inequivalent sites.

The potential shortcomings of composition based feature vectors (CBFV) is the way of combining the pre-generated element vectors to form a descriptor. The element vectors are scaled to its fractional abundance, signifying each constituent element are dictating their chemical signal to the material property. In real life situation, a chemical compound may have huge difference in material properties to combination of its constituent elements material properties. For instance, this problem is significant in the case of material doping: the dopants are present in the very small portion of the compound, but alters its electrical, thermal and many more properties. The chemical signal from the dopant element will not be significant with respect to host materials by using CBFVs. Thus it will be a task for many researchers to find the best and accurate way to model material properties behaviour for high throughput prediction.

There are alternatives to CBFV, such as Crystal Graph based Convolutional Neural Network (CGCNN) and SOAP. Through gradient boosting regressor, CGCNN and SOAP had RMSE values of 0.500 eV and 0.484 eV respectively.

5 Conclusions

Oxygen vacancy formation energy of metal oxides plays significant role in material properties. However, accurate prediction of formation energies remains challenge for researchers and industries for the development of novel and efficient materials. This project, thus, focuses on developing efficient machine learning model for prediction of oxygen vacancy formation energy. The 1665 metal oxides with significant compositional and structural diversity were prepared, and represented with compositional based feature vectors to characterise metal oxides. Different feature engineering methods were utilised to remove features with high correlation and to prevent overfitting of the models. SHAP value was used for identifying important features, and number of valence d-orbitals, boiling point, melting point, Voronoi coordination number of an elemental-crystal structure, enthalpy, polarisability, and oxidation state features were screened as features with high importance. Three model performance metrics - r2, MAE, and RMSE - were used for model evaluation. The gradient boosting regressor with concatenated Jarvis, Magpie and SOAP descriptors exhibited the best performance. The classification model to classify metal oxides with low (< 3 eV) and high oxygen vacancy formation energy. Gradient boosting classification was the best performing model with accuracy of 0.96. The model had tendency to overestimate the low energy materials and underestimate the high energy materials due to the limitation of compositional based feature vectors. The way to accurately distinguish polymorphs needs to be further researched.

Code availability

The python based codes are available at: `https://github.com/Junibuni/Evo_prediction_ml`

References

[1] Xiang Gao, Guanyu Liu, Ye Zhu, Peter Kreider, Alicia Bayon, Thomas Gengenbach, Teng Lu, Yun Liu, Jim Hinkley, Wojciech Lipiński, et al. Earth-abundant transition metal oxides with extraordinary reversible oxygen exchange capacity for efficient thermochemical synthesis of solar fuels. *Nano Energy*, 50:347–358, 2018.

[2] Xiao Hua, Phoebe K Allan, Chen Gong, Philip A Chater, Ella M Schmidt, Harry S Geddes, Alex W Robertson, Peter G Bruce, and Andrew L Goodwin. Non-equilibrium metal oxides via reconversion chemistry in lithium-ion batteries. *Nature communications*, 12(1):1–11, 2021.

[3] Lu Li, Christoph Richter, Jochen Mannhart, and RC Ashoori. Coexistence of magnetic order and two-dimensional superconductivity at laalo3/srtio3 interfaces. *Nature physics*, 7(10):762–766, 2011.

[4] Rahul R Salunkhe, Yusuf V Kaneti, and Yusuke Yamauchi. Metal–organic framework-derived nanoporous metal oxides toward supercapacitor applications: progress and prospects. *ACS nano*, 11(6):5293–5308, 2017.

[5] N Barsan, D Koziej, and U Weimar. Metal oxide-based gas sensor research: How to? *Sensors and Actuators B: Chemical*, 121(1):18–35, 2007.

[6] MA Banares and IE Wachs. Molecular structures of supported metal oxide catalysts under different environments. *Journal of Raman Spectroscopy*, 33(5):359–380, 2002.

[7] RJ Ayen and PA Iacobucci. Metal oxide aerogel preparation by supercritical extraction. *Reviews in Chemical Engineering*, 5(1-4):157–198, 1988.

[8] Guoxin Zhuang, Yawen Chen, Zanyong Zhuang, Yan Yu, and Jiaguo Yu. Oxygen vacancies in metal oxides: recent progress towards advanced catalyst design. *Science China Materials*, 63(11):2089–2118, 2020.

[9] Frederick Clifford Tompkins. Superficial chemistry and solid imperfections. *Nature*, 186(4718):3–6, 1960.

[10] H Sawada and K Kawakami. Electronic structure of oxygen vacancy in ta 2 o 5. *Journal of applied physics*, 86(2):956–959, 1999.

[11] William C Chueh, Christoph Falter, Mandy Abbott, Danien Scipio, Philipp Furler, Sossina M Haile, and Aldo Steinfeld. High-flux solar-driven thermochemical dissociation of co2 and h2o using nonstoichiometric ceria. *Science*, 330(6012):1797–1801, 2010.

[12] Myeongjin Kim, Byeongyong Lee, Hyun Ju, Jin Young Kim, Jooheon Kim, and Seung Woo Lee. Oxygen-vacancy-introduced basno3- δ photoanodes with tunable band structures for efficient solar-driven water splitting. *Advanced Materials*, 31(33):1903316, 2019.

[13] Keonwook Kang and Wei Cai. Vacancy formation energy, 2005.

[14] KW Johnson, PM Langdon, and MF Ashby. Grouping materials and processes for the designer: an application of cluster analysis. *Materials & design*, 23(1):1–10, 2002.

[15] IEH Van Kesteren. Product designers' information needs in materials selection. *Materials & Design*, 29(1):133–145, 2008.

[16] Mark E Eberhart and Dennis P Clougherty. Looking for design in materials design. *Nature materials*, 3(10):659–661, 2004.

[17] Walter Kohn and Lu Jeu Sham. Self-consistent equations including exchange and correlation effects. *Physical review*, 140(4A):A1133, 1965.

[18] Aron J Cohen, Paula Mori-Sánchez, and Weitao Yang. Challenges for density functional theory. *Chemical reviews*, 112(1):289–320, 2012.

[19] José M Soler, Emilio Artacho, Julian D Gale, Alberto García, Javier Junquera, Pablo Ordejón, and Daniel Sánchez-Portal. The siesta method for ab initio order-n materials simulation. *Journal of Physics: Condensed Matter*, 14(11):2745, 2002.

[20] Zhongyu Wan, Quan-De Wang, Dongchang Liu, and Jinhu Liang. Data-driven machine learning model for the prediction of oxygen vacancy formation energy of metal oxide materials. *Physical Chemistry Chemical Physics*, 23(29):15675–15684, 2021.

[21] Olexandr Isayev, Corey Oses, Cormac Toher, Eric Gossett, Stefano Curtarolo, and Alexander Tropsha. Universal fragment descriptors for predicting properties of inorganic crystals. *Nature communications*, 8(1):1–12, 2017.

[22] Ghanshyam Pilania, Arun Mannodi-Kanakkithodi, BP Uberuaga, Rampi Ramprasad, JE Gubernatis, and Turab Lookman. Machine learning bandgaps of double perovskites. *Scientific reports*, 6(1):1–10, 2016.

[23] Maarten De Jong, Wei Chen, Randy Notestine, Kristin Persson, Gerbrand Ceder, Anubhav Jain, Mark Asta, and Anthony Gamst. A statistical learning framework for materials science: application to elastic moduli of k-nary inorganic polycrystalline compounds. *Scientific reports*, 6(1):1–11, 2016.

[24] Ankit Agrawal and Alok Choudhary. Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science. *Apl Materials*, 4(5):053208, 2016.

[25] Christopher M Bishop. Pattern recognition. *Machine learning*, 128(9), 2006.

[26] Jonathan Schmidt, Mário RG Marques, Silvana Botti, and Miguel AL Marques. Recent advances and applications of machine learning in solid-state materials science. *npj Computational Materials*, 5(1):1–36, 2019.

[27] CJ Long, J Hattrick-Simpers, Makoto Murakami, RC Srivastava, Ichiro Takeuchi, Vicky L Karen, and X Li. Rapid structural mapping of ternary metallic alloy systems using the combinatorial approach and cluster analysis. *Review of Scientific Instruments*, 78(7):072217, 2007.

[28] Tim Mueller, Aaron Gilad Kusne, and Rampi Ramprasad. Machine learning in materials science: Recent progress and emerging applications. *Reviews in Computational Chemistry*, 29:186–273, 2016.

[29] Rudolf Allmann and Roland Hinek. The introduction of structure types into the inorganic crystal structure database icsd. *Acta Crystallographica Section A: Foundations of Crystallography*, 63(5):412–417, 2007.

[30] Colin R Groom, Ian J Bruno, Matthew P Lightfoot, and Suzanna C Ward. The cambridge structural database. *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials*, 72(2):171–179, 2016.

[31] Harry E Pence and Antony Williams. Chemspider: an online chemical information resource, 2010.

[32] Joanne Hill, Gregory Mulholland, Kristin Persson, Ram Seshadri, Chris Wolverton, and Bryce Meredig. Materials science with large-scale data and informatics: Unlocking new opportunities. *Mrs Bulletin*, 41(5):399–409, 2016.

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

[34] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin a. Persson. The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, 2013.

[35] Jordan O'Mara, Bryce Meredig, and Kyle Michel. Materials data infrastructure: a case study of the citrination platform to examine data import, storage, and access. *Jom*, 68(8):2031–2034, 2016.

[36] Yue Liu, Tianlu Zhao, Wangwei Ju, and Siqi Shi. Materials discovery and design using machine learning. *Journal of Materiomics*, 3(3):159–177, 2017.

[37] Tian Xie and Jeffrey C Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical review letters*, 120(14):145301, 2018.

[38] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[39] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167, 2008.

[40] David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alan Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[41] Mikael Henaff, Joan Bruna, and Yann LeCun. Deep convolutional networks on graph-structured data. *arXiv preprint arXiv:1506.05163*, 2015.

[42] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1263–1272. PMLR, 06–11 Aug 2017.

[43] 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(1):1–15, 2015.

[44] Anubhav Jain, Geoffroy Hautier, Charles J Moore, Shyue Ping Ong, Christopher C Fischer, Tim Mueller, Kristin A Persson, and Gerbrand Ceder. A high-throughput infrastructure for density functional theory calculations. *Computational Materials Science*, 50(8):2295–2310, 2011.

[45] Maarten De Jong, Wei Chen, Thomas Angsten, Anubhav Jain, Randy Notestine, Anthony Gamst, Marcel Sluiter, Chaitanya Krishna Ande, Sybrand Van Der Zwaag, Jose J Plata, et al. Charting the complete elastic properties of inorganic crystalline compounds. *Scientific data*, 2(1):1–13, 2015.

[46] Myungjoon Kim, Byung Chul Yeo, Sang Soo Han, and Donghun Kim. Slab graph convolutional neural network for discovery of n2 electroreduction catalysts, 2019.

[47] Angelo Ziletti, Devinder Kumar, Matthias Scheffler, and Luca M Ghiringhelli. Insightful classification of crystal structures using deep learning. *Nature communications*, 9(1):1–10, 2018.

[48] Xianfeng Ma, Zheng Li, Luke EK Achenie, and Hongliang Xin. Machine-learning-augmented chemisorption model for co2 electroreduction catalyst screening. *The journal of physical chemistry letters*, 6(18):3528–3533, 2015.

[49] Ryosuke Jinnouchi and Ryoji Asahi. Predicting catalytic activity of nanoparticles by a dft-aided machine-learning algorithm. *The journal of physical chemistry letters*, 8(17):4279–4283, 2017.

[50] Kevin Tran and Zachary W Ulissi. Active learning across intermetallics to guide discovery of electrocatalysts for co 2 reduction and h 2 evolution. *Nature Catalysis*, 1(9):696–703, 2018.

[51] Raymond Gasper, Hongbo Shi, and Ashwin Ramasubramaniam. Adsorption of co on low-energy, low-symmetry pt nanoparticles: energy decomposition analysis and prediction via machine-learning models. *The Journal of Physical Chemistry C*, 121(10):5612–5619, 2017.

[52] Nolan J O'Connor, ASM Jonayat, Michael J Janik, and Thomas P Senftle. Interaction trends between single metal atoms and oxide supports identified with density functional theory and statistical learning. *Nature Catalysis*, 1(7):531–539, 2018.

[53] Aria Mansouri Tehrani, Anton O Oliynyk, Marcus Parry, Zeshan Rizvi, Samantha Couper, Feng Lin, Lowell Miyagi, Taylor D Sparks, and Jakoah Brgoch. Machine learning directed search for ultraincompressible, superhard materials. *Journal of the American Chemical Society*, 140(31):9844–9853, 2018.

[54] Harris Drucker, Chris JC Burges, Linda Kaufman, Alex Smola, Vladimir Vapnik, et al. Support vector regression machines. *Advances in neural information processing systems*, 9:155–161, 1997.

[55] 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.

[56] 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.

[57] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[58] Anus Manzoor, Gaurav Arora, Bryant Jerome, Nathan Linton, Bailey Norman, and Dilpuneet Singh Aidhy. Machine learning based methodology to predict point defect energies in multi-principal element alloys. *Frontiers in Materials*, 8:129, 2021.

[59] Falko Ziebert and Igor S Aranson. Computational approaches to substrate-based cell motility. *npj Computational Materials*, 2(1):1–16, 2016.

[60] Vinit Sharma, Pankaj Kumar, Pratibha Dev, and Ghanshyam Pilania. Machine learning substitutional defect formation energies in abo3 perovskites. *Journal of Applied Physics*, 128(3):034902, 2020.

[61] Haoyan Huo and Matthias Rupp. Unified representation of molecules and crystals for machine learning, 2018.

[62] Christopher R Collins, Geoffrey J Gordon, O Anatole von Lilienfeld, and David J Yaron. Constant size molecular descriptors for use with machine learning. *arXiv preprint arXiv:1701.06649*, 2017.

[63] Felix Faber, Alexander Lindmaa, O Anatole von Lilienfeld, and Rickard Armiento. Crystal structure representations for machine learning models of formation energies. *International Journal of Quantum Chemistry*, 115(16):1094–1101, 2015.

[64] Matthias Rupp. Machine learning for quantum mechanics in a nutshell. *International Journal of Quantum Chemistry*, 115(16):1058–1073, 2015.

[65] D.H. Wolpert and W.G. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, 1997.

[66] Lauri Himanen, Marc O.J. Jäger, Eiaki V. Morooka, Filippo Federici Canova, Yashasvi S. Ranawat, David Z. Gao, Patrick Rinke, and Adam S. Foster. Dscribe: Library of descriptors for machine learning in materials science. *Computer Physics Communications*, 247:106949, Feb 2020.

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

[68] Atsuto Seko, Tomoya Maekawa, Koji Tsuda, and Isao Tanaka. Machine learning with systematic density-functional theory calculations: Application to melting temperatures of single-and binary-component solids. *Physical Review B*, 89(5):054303, 2014.

[69] Koji Fujimura, Atsuto Seko, Yukinori Koyama, Akihide Kuwabara, Ippei Kishida, Kazuki Shitara, Craig AJ Fisher, Hiroki Moriwake, and Isao Tanaka. Accelerated materials design of lithium superionic conductors based on first-principles calculations and machine learning algorithms. *Advanced Energy Materials*, 3(8):980–985, 2013.

[70] Kazuaki Toyoura, Daisuke Hirano, Atsuto Seko, Motoki Shiga, Akihide Kuwabara, Masayuki Karasuyama, Kazuki Shitara, and Ichiro Takeuchi. Machine-learning-based selective sampling procedure for identifying the low-energy region in a potential energy surface: A case study on proton conduction in oxides. *Physical Review B*, 93(5):054112, 2016.

[71] Atsushi Togo, Laurent Chaput, and Isao Tanaka. Distributions of phonon lifetimes in brillouin zones. *Physical review B*, 91(9):094306, 2015.

[72] Matthias Rupp, Alexandre Tkatchenko, Klaus-Robert Müller, and O Anatole Von Lilienfeld. Fast and accurate modeling of molecular atomization energies with machine learning. *Physical review letters*, 108(5):058301, 2012.

[73] Atsuto Seko, Hiroyuki Hayashi, Keita Nakayama, Akira Takahashi, and Isao Tanaka. Representation of compounds for machine-learning prediction of physical properties. *Physical Review B*, 95(14):144110, 2017.

[74] Atsuto Seko, Hiroyuki Hayashi, Hisashi Kashima, and Isao Tanaka. Matrix-and tensor-based recommender systems for the discovery of currently unknown inorganic compounds. *Physical Review Materials*, 2(1):013805, 2018.

[75] Bin Wang, Angela Wang, Fenxiao Chen, Yuncheng Wang, and C-C Jay Kuo. Evaluating word embedding models: Methods and experimental results. *APSIPA transactions on signal and information processing*, 8, 2019.

[76] David Meyer. How exactly does word2vec work? *Uoregon. Edu, Brocade. Com*, pages 1–18, 2016.

[77] Luis M Antunes, Ricardo Grau-Crespo, and Keith T Butler. Distributed representations of atoms and materials for machine learning. *arXiv preprint arXiv:2107.14664*, 2021.

[78] Vahe Tshitoyan, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A Persson, Gerbrand Ceder, and Anubhav Jain. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, 571(7763):95–98, 2019.

[79] Alexander Dunn, Qi Wang, Alex Ganose, Daniel Dopp, and Anubhav Jain. Benchmarking materials property prediction methods: the matbench test set and automatminer reference algorithm. *npj Computational Materials*, 6(1):1–10, 2020.

[80] Nathan C Frey, Deji Akinwande, Deep Jariwala, and Vivek B Shenoy. Machine learning-enabled design of point defects in 2d materials for quantum and neuromorphic information processing. *ACS nano*, 14(10):13406–13417, 2020.

[81] Aron Walsh and Alex Zunger. Instilling defect tolerance in new compounds. *Nature materials*, 16(10):964–967, 2017.

[82] Yu Kumagai, Naoki Tsunoda, Akira Takahashi, and Fumiyasu Oba. Insights into oxygen vacancies from high-throughput first-principles calculations. *Physical Review Materials*, 5(12):123803, 2021.

[83] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL materials*, 1(1):011002, 2013.

[84] Shyue Ping Ong, Shreyas Cholia, Anubhav Jain, Miriam Brafman, Dan Gunter, Gerbrand Ceder, and Kristin A Persson. The materials application programming interface (api): A simple, flexible and efficient api for materials data based on representational state transfer (rest) principles. *Computational Materials Science*, 97:209–215, 2015.

[85] Luca M. Ghiringhelli, Jan Vybiral, Sergey V. Levchenko, Claudia Draxl, and Matthias Scheffler. Big data of materials science: Critical role of the descriptor. *Phys. Rev. Lett.*, 114:105503, Mar 2015.

[86] Steven K Kauwe, Jake Graser, Antonio Vazquez, and Taylor D Sparks. Machine learning prediction of heat capacity for solid inorganics. *Integrating Materials and Manufacturing Innovation*, 7(2):43–51, 2018.

[87] Anthony Yu-Tung Wang, Steven K Kauwe, Ryan J Murdock, and Taylor D Sparks. Compositionally restricted attention-based network for materials property predictions. *Npj Computational Materials*, 7(1):1–10, 2021.

[88] Ryan J Murdock, Steven K Kauwe, Anthony Yu-Tung Wang, and Taylor D Sparks. Is domain knowledge necessary for machine learning materials properties? *Integrating Materials and Manufacturing Innovation*, 9(3):221–227, 2020.

[89] Richard Bellman. Dynamic programming. *Science*, 153(3731):34–37, 1966.

[90] Mavuto M Mukaka. A guide to appropriate use of correlation coefficient in medical research. *Malawi medical journal*, 24(3):69–71, 2012.

[91] 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.

[92] Kamal Choudhary, Kevin F Garrity, Andrew CE Reid, Brian DeCost, Adam J Biacchi, Angela R Hight Walker, Zachary Trautt, Jason Hattrick-Simpers, A Gilad Kusne, Andrea Centrone, et al. The joint automated repository for various integrated simulations (jarvis) for data-driven materials design. *npj Computational Materials*, 6(1):1–13, 2020.

[93] Lauri Himanen, Marc O. J. Jäger, Eiaki V. Morooka, Filippo Federici Canova, Yashasvi S. Ranawat, David Z. Gao, Patrick Rinke, and Adam S. Foster. DScribe: Library of descriptors for machine learning in materials science. *Computer Physics Communications*, 247:106949, 2020.

[94] Sandip De, Albert P Bartók, Gábor Csányi, and Michele Ceriotti. Comparing molecules and solids across structural and alchemical space. *Physical Chemistry Chemical Physics*, 18(20):13754–13769, 2016.

[95] Marc OJ Jäger, Eiaki V Morooka, Filippo Federici Canova, Lauri Himanen, and Adam S Foster. Machine learning hydrogen adsorption on nanoclusters through structural descriptors. *npj Computational Materials*, 4(1):1–8, 2018.

[96] Data Science. Confusion matrix, Jan 1970.

[97] Hemant Ishwaran. The effect of splitting on random forests. *Machine learning*, 99(1):75–118, 2015.

[98] Fernando Berzal, Juan-Carlos Cubero, Fernando Cuenca, and Marıa J Martın-Bautista. On the quest for easy-to-understand splitting rules. *Data & Knowledge Engineering*, 44(1):31–48, 2003.