



**NANYANG  
TECHNOLOGICAL  
UNIVERSITY**  

---

**SINGAPORE**

**Nanyang Technological University  
CZ4041 Machine Learning Project  
Group 64**

Name	Matriculation Number	Contribution
Mak Qin Xiang	U2022775F	Code, Report, Slides and Video
Jared Chan	U2022954B	Code, Report, Slides and Video
Lim Zhi Qing	U2021897L	Code, Report, Slides and Video

# Content Page

<b>Problem Statement</b>	<b>3</b>
<b>Exploratory Data Analysis</b>	<b>3</b>
Descriptive Univariate Analysis	4
Pearson Correlation	4
Distance Correlation	5
Model-based Feature Analysis	7
Multivariate feature analysis	7
Lattice Vectors vs. Formation Energy & Band Gap Energy	7
Elemental Percentages vs. Formation Energy & Band Gap Energy	8
Number of Total Atoms vs. Formation Energy & Band Gap Energy	9
Space Group vs. Formation & Band Gap Energy	11
Geometry Information	11
Additional Features	13
Combining Space Group and Number of Total Atoms	13
XenonPy	13
<b>Solutions</b>	<b>15</b>
Linear Regression	15
Gradient boosted trees	16
Stacked Models	16
Feedforward Neural Networks	17
Recurrent Neural Networks & Convolutional Neural Networks	17
Composite Neural Network	18
XGBoost using Eigenspectrum	18
<b>Issues Faced</b>	<b>20</b>
<b>Public Leaderboard Score</b>	<b>21</b>
<b>References</b>	<b>21</b>

# Problem Statement

The aim of this challenge involves predicting the formation energy and band gap energy of specific categories of transparent conductors. The task involves investigating the chemical compositions of materials as well as their structural geometry and geometric properties. Suitable data-driven models need to be explored to accommodate the data provided. Compositional data, as well as geometric data, need to be pre-processed before it is suitable for data-driven models. The given dataset should also be augmented to improve model performance.

The scoring metric, **Root Mean Squared Logarithmic Error (RMLSE)** measures the performance of a predictive model by calculating the average logarithmic difference between predicted and actual values. By applying a logarithmic scale, **RMSLE** reduces the impact of large errors or outliers, making it useful for datasets containing wide-ranging or skewed data, ensuring that errors across the entire range of values are more uniformly considered in the evaluation of the model's performance. The formula for **RMSLE** on a single column is shown below.

$$\sqrt{\frac{1}{n} \sum_{i=1}^n (\log(p_i + 1) - \log(a_i + 1))^2}$$

## Exploratory Data Analysis

This segment delves into exploratory data analysis to explore and investigate the provided dataset. The training dataset comprises 2400 entries, each with an **ID** and 13 features. Four of these features represent the material's elemental composition.

- **Percent Aluminium (Al), Gallium (Ga), and Indium (In)** describe the percentage composition of each element in the material, varying between the range of **0** to **1**.
- **Number of Total Atoms** describes the total number of atoms in the material, and varies between the range of **10** to **80**.

The remaining features pertain to the geometry of a multi-atom grid, the fundamental structure of the material.

- **Space Group** is a categorical variable describing the arrangement and symmetry of atoms within a crystal lattice structure. Space groups with higher numbers describe higher symmetry.
- **Lattice Vectors 1, 2, 3 (ang)** describe the arrangement and distances in a lattice structure, measured in units of **angstroms (Å, 1 Å = 10<sup>-10</sup> m)**. **Lattice vector 1** varies from **12 Å** to **24 Å**, **lattice vector 2** varies from **2.9 Å** to **6.7 Å** and **lattice vector 3** from **5.7** to **6.9 Å** [1].
- **Lattice Angles Alpha, Beta, Gamma (degree)** describe the angles between lattice vectors, measured in degrees. The **lattice angles alpha** and **gamma** are effectively **90 degrees**, while **lattice angle beta** varies between **103.7°** to **106.2°** [1].

There are two target variables.

- **Formation Energy (eV/atom)** refers to the energy change that occurs during the formation of the material, measured in **electron volts per atom (eV/atom)**.
- **Band Gap Energy (eV/atom)** refers to the energy required to promote an electron from the highest energy level in the valence band to the lowest energy level in the conduction band within a material's electronic structure, measured in **electron volts per atom (eV/atom)**.

## Descriptive Univariate Analysis

### Pearson Correlation

The Pearson correlation coefficient, denoted as  $r$ , measures the magnitude and direction of the linear relationship between two continuous variables. This correlation coefficient ranges from -1 to +1, where a value of +1 indicates a perfect positive linear relationship, -1 denotes a perfect negative linear relationship, and 0 signifies no linear relationship. The formula for Pearson's correlation coefficient involves computing the covariance between two variables and dividing it by the product of their standard deviations. This coefficient assumes a linear relationship between the variables.

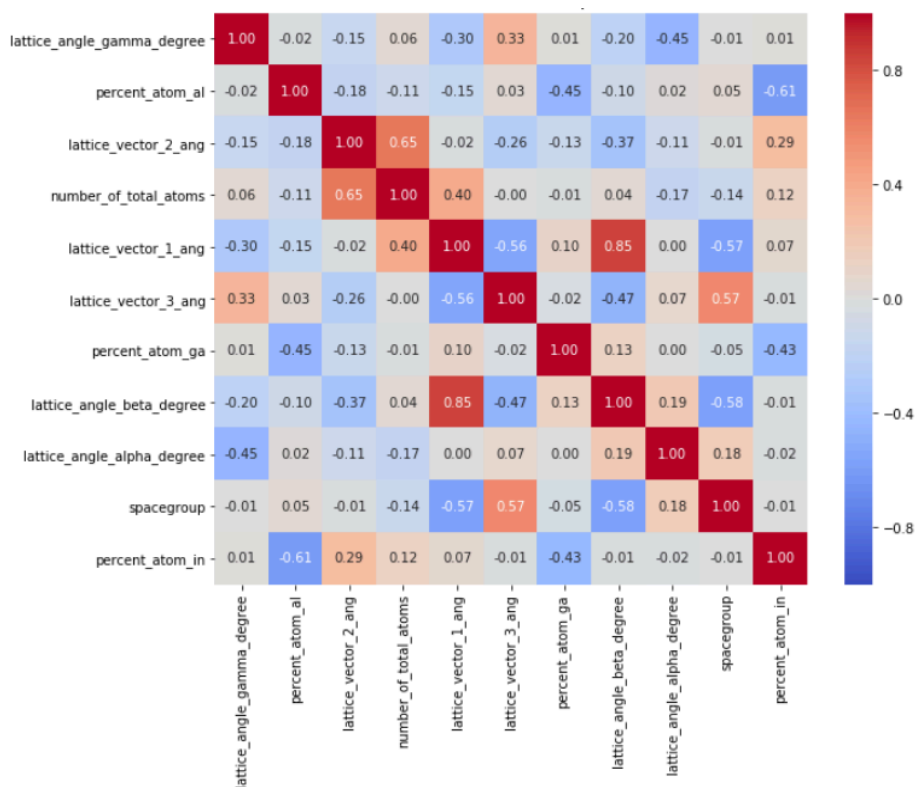


Figure 1. Pearson Correlation Heatmap of Features

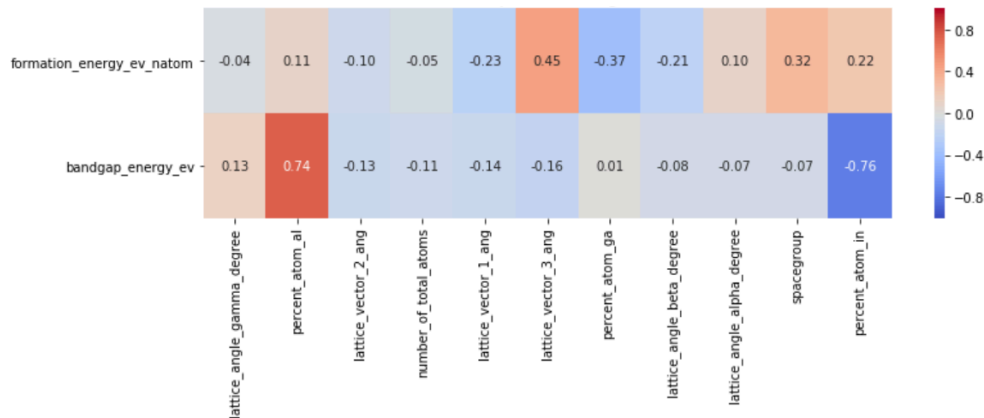


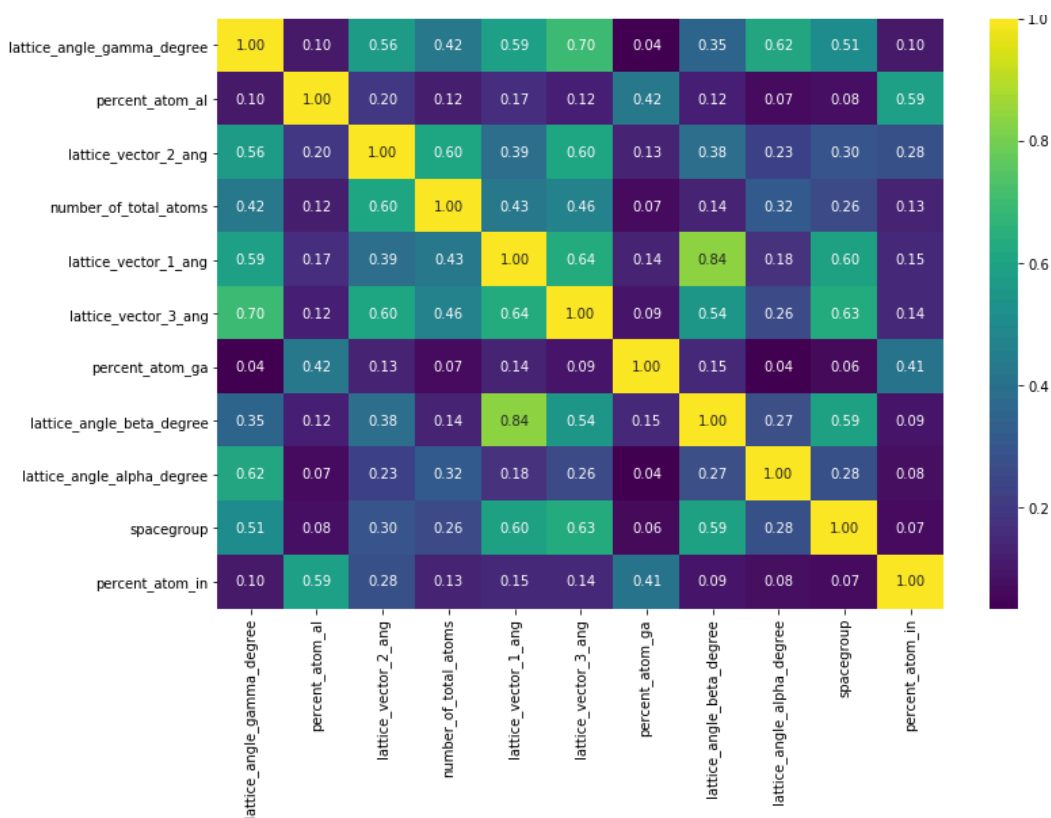
Figure 2. Pearson Correlation Heatmap between Features and Targets

As shown in **figure 1**, **lattice vector 1** exhibits a strong correlation with **lattice angle beta**, suggesting a potential opportunity to derive a new feature through their combination. However, this correlation is expected as **lattice angle beta** describes the angle between **lattice vectors 1** and **3**.

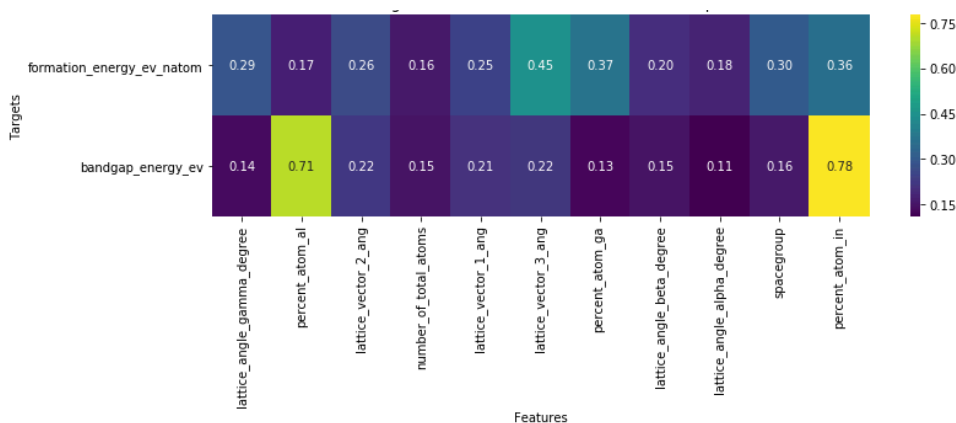
As shown in **figure 2**, The correlation between **bandgap energy** and **percent AI** is notably high. **Formation energy** also demonstrates a correlation with **lattice vector 3**, while its correlation with other features is comparatively weaker.

## Distance Correlation

Distance correlation is a statistical measure that captures both linear and nonlinear relationships between two variables. Unlike Pearson's correlation, it considers a broader range of associations and is useful for detecting nonlinear dependencies. The measure ranges from 0, indicating no dependence, to 1, indicating perfect dependence, and is suitable for evaluating the similarity of distances between pairs of observations within the joint space of two variables.



**Figure 3.** Distance Correlation Heatmap between Features



**Figure 4.** Distance Correlation Heatmap between Features and Targets

As shown in **figure 4**, **band gap energy** shows a high correlation with both **percent Al** and **percent In**, as similarly indicated by the Pearson correlation coefficient. **Formation Energy** exhibits correlations with **lattice vector 3**, **percent In**, and **percent Ga**.

## Model-based Feature Analysis

Analysis of feature importance follows a methodology similar to Pearson's correlation coefficient, where the standardised regression coefficient in a predictive linear regression model is equivalent. In this approach, different predictive models are built to assess the effectiveness of each individual feature in predicting the target variable. The findings are obtained through Random Forest Regressors, using one feature at a time to predict each target. This process incorporates cross-validation and utilises a shallow tree structure to prevent overfitting.

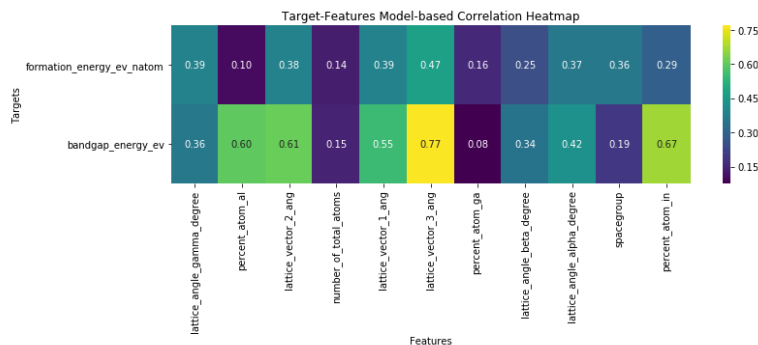


Figure 5. Target Features Model based Correlation heatmap

As shown in figure 5, **lattice vector 3** and **percent indium** show highly positive correlation with **band gap energy**, suggesting that these features may be valuable in predicting **band gap energy**.

## Multivariate feature analysis

### Lattice Vectors vs. Formation Energy & Band Gap Energy

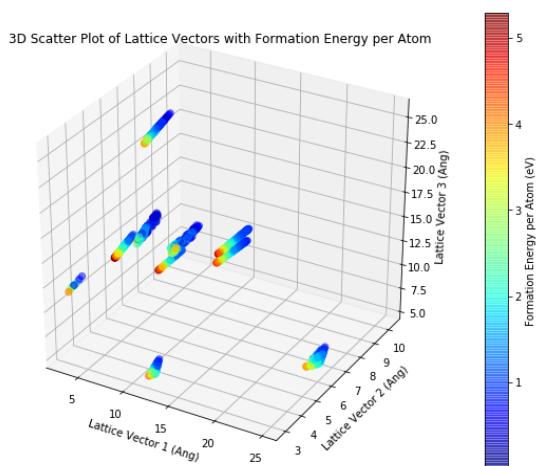
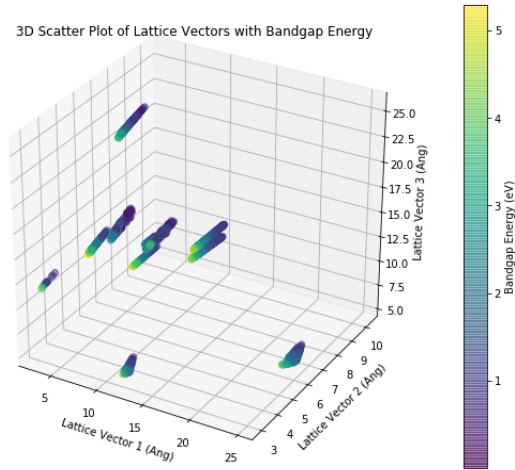


Figure 6. Scatter Plot of Lattice Vectors with Formation Energy

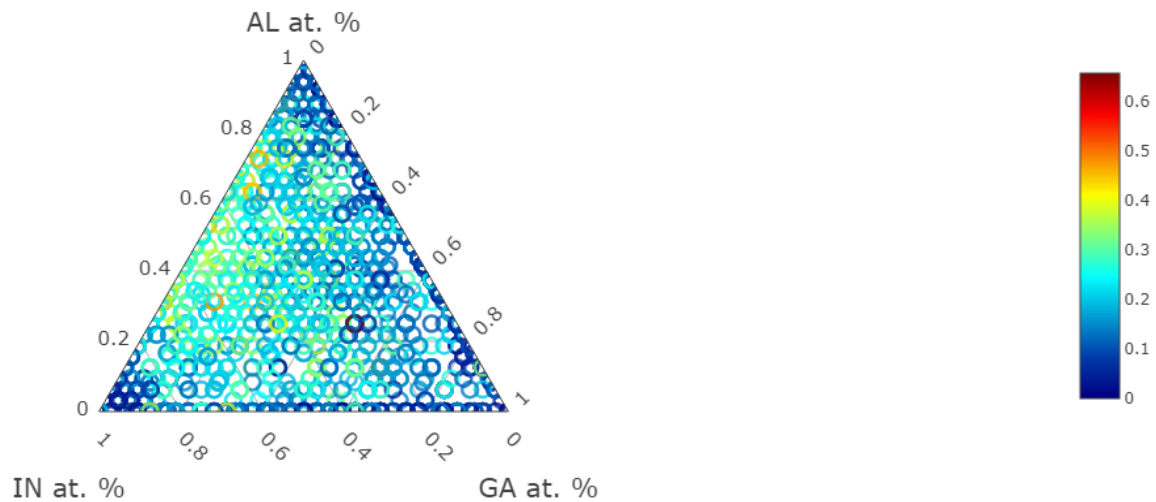


**Figure 7.** Scatter Plot of Lattice Vectors with Band Gap Energy

As shown in **figures 6 and 7**, there is a discernible correlation between the change in **formation energy** and **band gap energy** and variations in **lattice vector 2**, but not for **lattice vectors 1 and 3**.

## Elemental Percentages vs. Formation Energy & Band Gap Energy

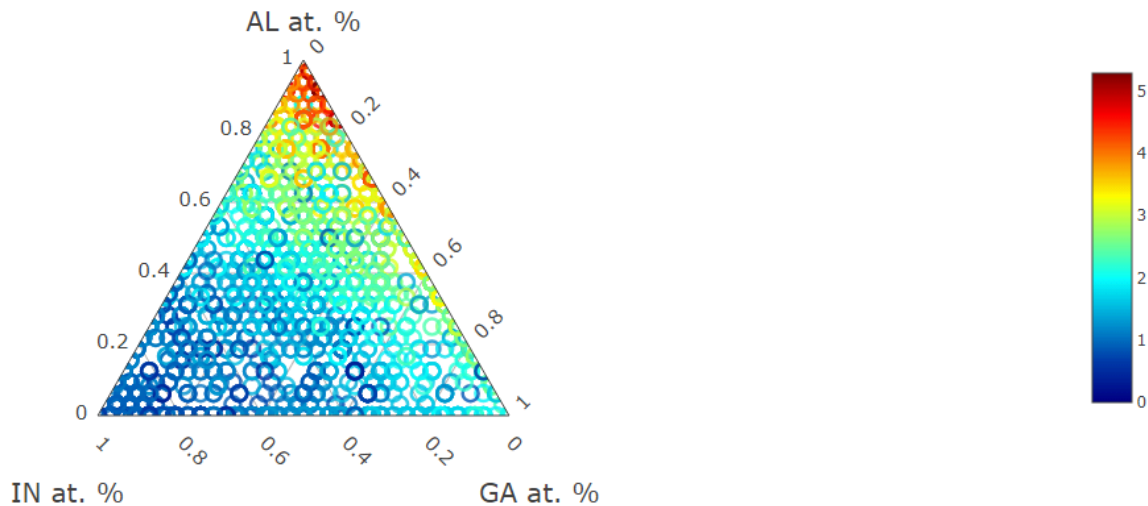
A ternary plot is employed to analyse the interplay between **percent Al**, **percent In**, and **percent Ga** concerning their impact on **formation energy**. The plot serves to visualise chemical compositions by portraying the three variables within a 2D graph. Ternary plots find prevalent usage in the domain of physical chemistry for their adeptness in illustrating relationships among multiple components within a compositional framework.



**Figure 8.** Ternary Plot of Formation Energy at different (AL, IN, GA) Atomic Percentages



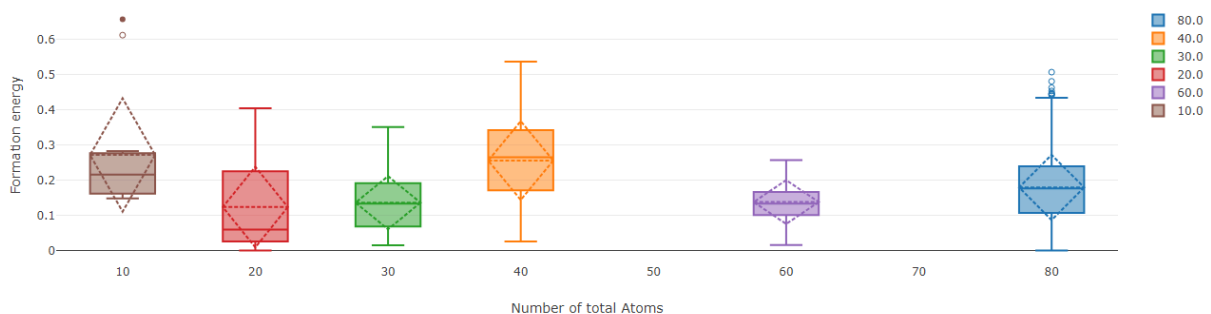
As shown in **figure 8**, **formation energy** tends to be low when the **percent In** or the **percent Al** are at their extremes. There is a gradual increase in **formation energy** as the **percent Ga** decreases.



**Figure 9.** Ternary Plot of Band Gap Energy at different (AL, IN, GA) Atomic Percentages

As shown in **figure 9**, **band gap energy** decreases as the **percent In** increases. **Band gap energy** is high when the **percent Al** is high. **Band gap energy** does not seem to be affected by the **percent Ga**.

## Number of Total Atoms vs. Formation Energy & Band Gap Energy



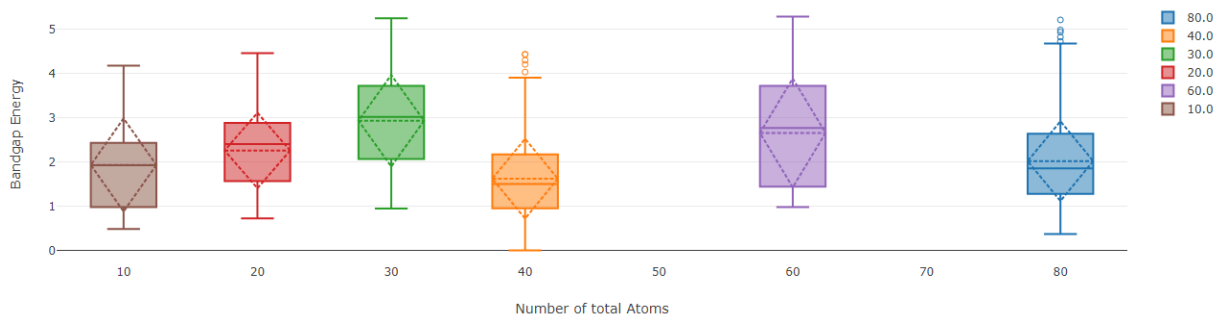
**Figure 10.** Box Plot of Total Atoms against Formation Energy

192	0.2351
660	0.2825
984	0.1856
1235	0.6121
1618	0.2744
1720	0.1491
1750	0.2456
1830	0.1549
1983	0.6572
2090	0.1483
2094	0.1642
2152	0.2043
2371	0.2159

Name: formation\_energy\_ev\_natom, dtype: float64

**Figure 11.** Formation Energy of Materials with 10 Atoms

As shown in **figure 10**, there are significant outliers where the **number of total atoms** is **10**. As shown in **figure 11**, the entries with indices **1235**, and **1983** appear to be outliers, and it would be advantageous to eliminate them from the training set. Materials with the **number of total atoms** at **10** and **60** display low variance, suggesting that the **formation energy** for materials with these **numbers of total atoms** fall within a smaller range.



**Figure 12.** Box Plot Total Atoms against Band Gap Energy

Analysis of **figure 12** does not reveal any distinct pattern between the **number of total atoms** and shows that there is a notable variability in **band gap energy** across different **numbers of total atoms**.

## Space Group vs. Formation & Band Gap Energy

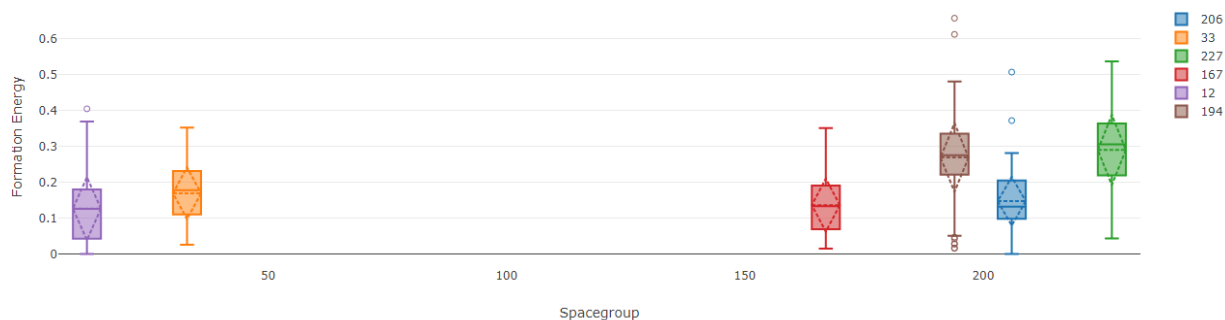


Figure 13. Box Plot of Space Group against Formation Energy

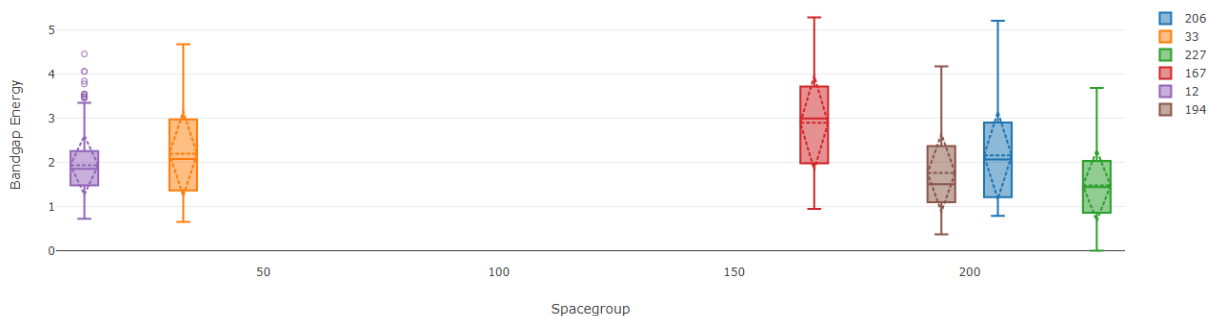
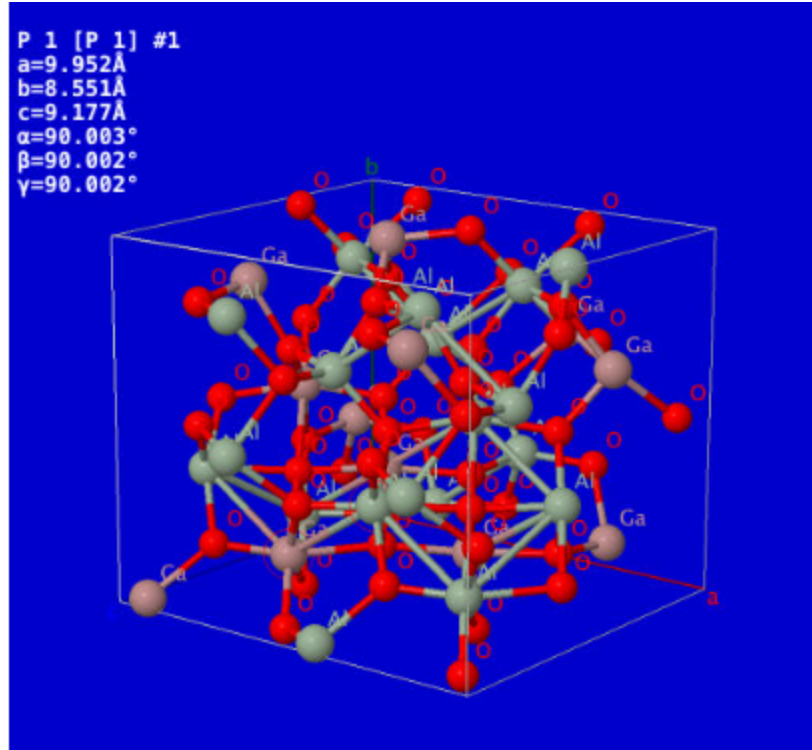


Figure 14. Box Plot of Space Group against Band Gap Energy

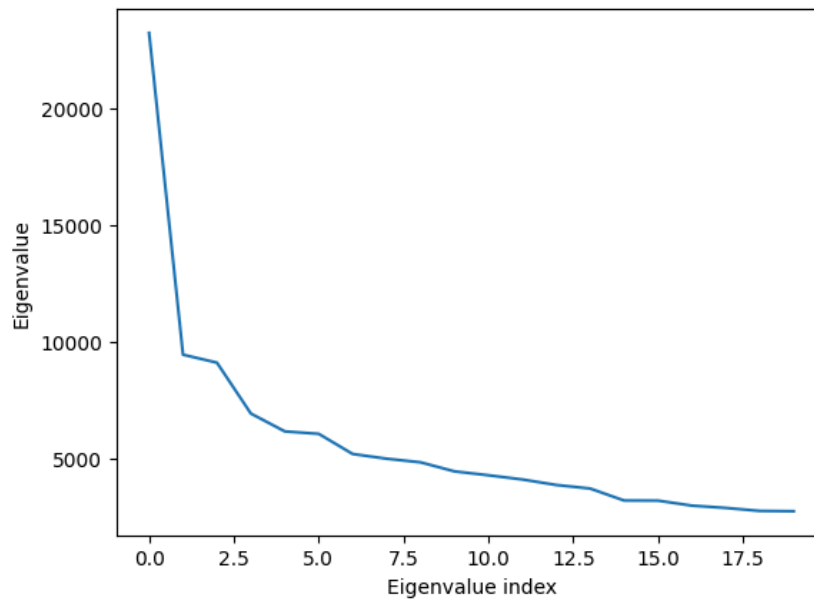
As shown in figure 13 and 14, there is no clear relationship between **space group** and **formation** and **band gap energy**, as energies for most **space groups** overlap.

## Geometry Information

Geometry information is extracted from the provided **.xyz** files. Lattice vectors, atom identities and atom coordinates were obtained. Lattice vectors are a set of vectors that define the periodicity and symmetry of a crystal lattice, which is a three-dimensional array of points denoting the positions of atoms, ions, or molecules that make up the crystal.



**Figure 15.** Example of a xyz file



**Figure 16.** Example of an Eigenspectrum

A sine matrix is constructed from the geometrical data for each material, in accordance with the research done by Faber et. al [2]. Eigenvalues are calculated and sorted from the sine matrix data, creating an eigenspectrum [3]. Standard scaling followed by the Principal Component Analysis (PCA) technique is applied to improve performance as

regions with relatively higher variance should have more significance in predicting the target variables.

## Additional Features

### Combining Space Group and Number of Total Atoms

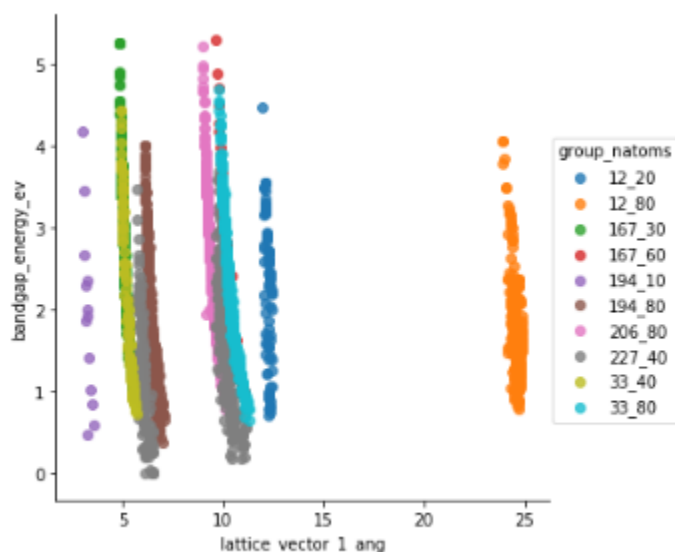


Figure 17. Example of an Eigenspectrum

The combination of the space group and number of total atoms is investigated. As shown in **figure 17** it is observed that the **lattice vector** and **lattice angle** features exhibit distinct variations based on the combination of **space group** and **number of total atoms**. To capture this relationship effectively, a list of categorical features is created by combining the **space group** and **number of total atoms** through the method of one-hot encoding.

## XenonPy

Using XenonPy, additional features are generated regarding chemical and structural properties.

As reported by Tien et al [4], the orbital field matrix seems to be highly correlated to the formation energies of crystalline materials. However, the XenonPy library is unable to generate the orbital field matrix due to unknown issues.

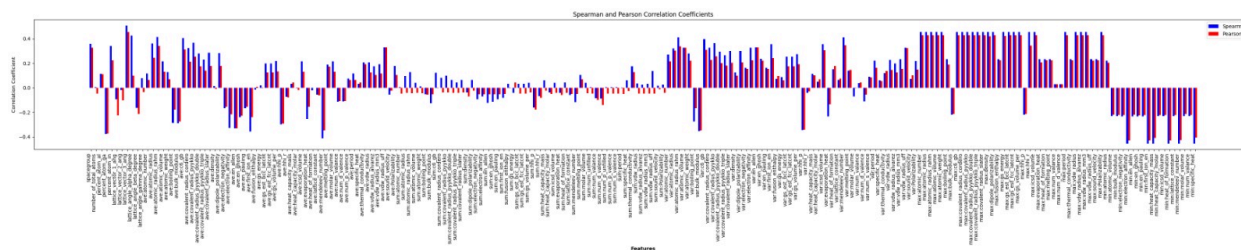
The Radial Distribution Function is generated, which is represented as a series of regular spikes corresponding to the atomic positions within the material lattice, giving insight into the frequency of distances between particles.

The mean, maximum, minimum, sum and standard deviation of the following features are generated, as shown in **figure 18**.

atomic\_number, atomic\_radius, atomic\_radius\_rahm, atomic\_volume, atomic\_weight, boiling\_point, ave:bulk\_modulus, ave:c6\_gb, covalent\_radius\_cordero, covalent\_radius\_pyykko, covalent\_radius\_pyykko\_double, covalent\_radius\_pyykko\_triple, covalent\_radius\_slater, density, dipole\_polarizability, electron\_negativity, electron\_affinity, en\_allen, en\_ghosh, en\_pauling, first\_ion\_en, fusion\_enthalpy, gs\_bandgap, gs\_energy, gs\_est\_bcc\_latcnt, gs\_est\_fcc\_latcnt, gs\_mag\_moment, gs\_volume\_per, hhi\_p, hhi\_r, heat\_capacity\_mass, heat\_capacity\_molar, icsd\_volume, evaporation\_heat, heat\_of\_formation, lattice\_constant, mendelevy\_number, melting\_point, molar\_volume, num\_unfilled, num\_valance, num\_d\_unfilled, num\_d\_valence, num\_f\_unfilled, num\_f\_valence, num\_p\_unfilled, num\_p\_valence, num\_s\_unfilled, num\_s\_valence, period, specific\_heat, thermal\_conductivity, vdw\_radius, vdw\_radius\_alvarez, vdw\_radius\_mm3, vdw\_radius\_uff, sound\_velocity, polarizability

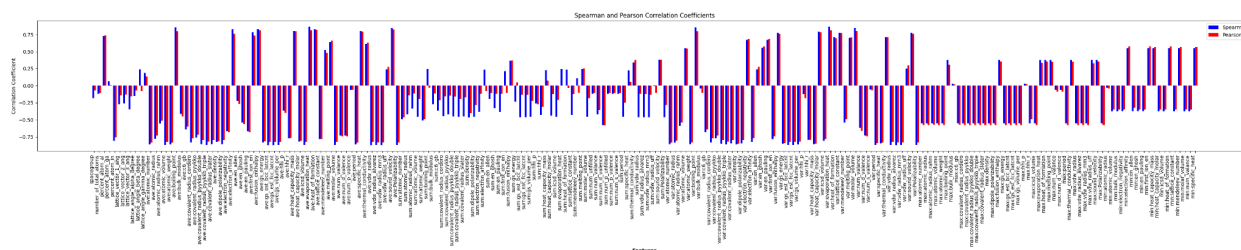
**Figure 18.** Generated Features

To enhance model fitting, features with a wide range of values are identified, specifically those where the maximum value exceeds twenty times the minimum. These features undergo logarithmic transformation to reduce skewness. Additionally, features that displayed constant values across all samples are deemed non-informative and subsequently eliminated to improve the model's performance. This results in a more streamlined dataset, conducive to effective modelling.



**Figure 19.** Correlation of Formation Energy and All Features

As shown in **figure 19**, the blue and red bars represent respectively the pearson and spearman correlations between the generated chemical properties and **formation energy**. All features display correlation between the range of **-0.6 to +0.6**, indicating the absence of strong correlation with **formation energy**. The magnitude of the spearman correlation coefficient is relatively higher compared to the spearman correlation coefficient, suggesting that a non-linear model will show better performance.



**Figure 20.** Correlation of Band Gap Energy and All Features

As shown in **figure 20**, the blue and red bars represent respectively the pearson and spearman correlations between the generated chemical properties and **band gap energy**. Several features display correlation outside the range of **-0.7 to +0.7**, indicating that **band gap energy** is well represented by these features. The magnitude of the spearman correlation coefficient is relatively higher compared to the spearman correlation coefficient, suggesting that a non-linear model will show better performance.

## Solutions

Several machine learning methods were explored and experimented on. In this section, the methods used and the results obtained are explained. A **80:20** train test split with shuffling is used for all neural networks, along with the Adam optimiser, and early stopping with a patience of **10**. Standard scaling was applied to all features.

## Linear Regression

Features and Models Used	Score
Linear regression with the given features and one hot encoding	0.08514
Above with PCA of eigenspectrum of sine matrix	0.06744
Above with best L2 regularisation	0.06714
Above but with best L1 regularisation	0.21225
With the additional generated features, with filtering on correlation	0.12747

**Table 1.** Public Leaderboard Score for Linear Regression

A simple linear regression with standard scaling is first attempted. For this simple model, a public score of **0.08514** is achieved, which is unsatisfactory. Adding the eigenspectrum data, improves the score to **0.06744**. L2 regularisation is then added,

which with the best alpha, improves the score to **0.06714**. However, the addition of L1 regularisation, even with weak regularisation, negatively impacts the score, bringing it to **0.21225**. Thus, L2 and L1 regularisation is not combined. The addition of additional features, despite the high correlation with band gap energy, causes the linear regression model to perform poorly with a score of **0.12747**.

## Gradient boosted trees

	Max depth 3	Max depth 5	Max depth 7
All features except Radial Basis Function	0.05665	0.05612	0.06065
All Features	0.05973	0.05872	0.06126

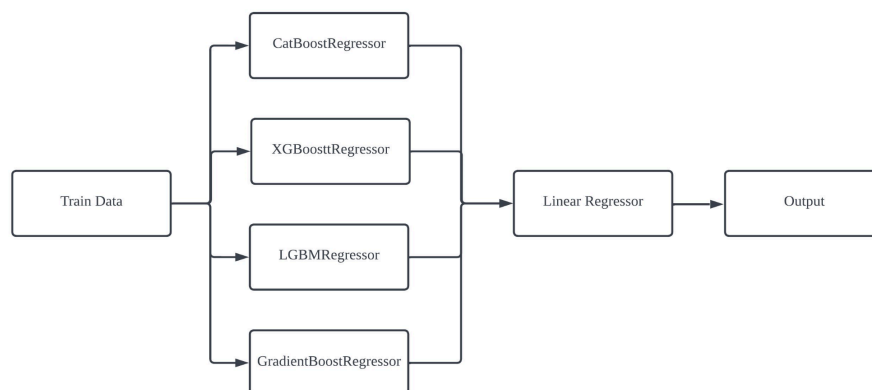
**Table 2.** Public leaderboard score for Gradient Boosted Trees

	Max depth 3	Max depth 5	Max depth 7
All Features	0.974	0.993	0.999

**Table 3.** In-sample  $R^2$  score

Adding the radial distribution function, seems to worsen the gradient boosted tree model. The best score was achieved with a max depth of 5 as it's complex enough to model the target variable, unlike max depth of 3 as shown in the higher  $R^2$  score in **Table 2**. and it's able to generalise better than the max depth of 7.

## Stacked Models



**Figure 21.** Stacked Model Architecture



For these stacked models, a combination of the principal components for the elements, oxygen, aluminium, indium, and gallium are used. These are obtained from PCA on the composition data, geometry data and additional generated chemical data. Columns with only a single unique value are dropped.

CatBoost, XGBoost, LGBM, and GradientBoost are used as individual learners. CatBoost is adept at managing categorical features, while XGBoost excels at capturing complex relationships. LGBM is lightweight and efficient, suitable for large datasets, and GradientBoost is effective in handling non-linear patterns. Together, this ensemble is well-equipped to navigate the complexities of the given data. Linear Regression is then used as the meta-learner. With these stacked models, a public score of **0.05596** is obtained, which is a slight improvement from **0.05612**.

## Feedforward Neural Networks

A shallow feedforward neural network is tested with 1 hidden layer trained on the composition data. However, the complexity of the problem can not be learned by a shallow network, as evidenced by the in-sample loss remaining constant. Adding more dense layers and varying the width of each layer helps the model learn better, improving the public score from **0.16432** to **0.10378**. Batch normalisation and dropout is also added, improving the public score to **0.09557** based on the given features.

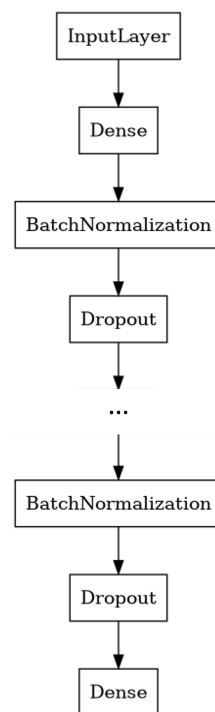


Figure 22. Feedforward Model Architecture

## Recurrent Neural Networks & Convolutional Neural Networks

An LSTM network is designed and trained on the eigenspectrum data. This model has much better performance on both training and validation data than the feedforward network and produces much better results on the test data, with a public score of **0.07456**. However, adding more LSTM layers or dense layers does not help to improve model performance. Model performance seems to be optimal at 128 LSTM cells with **0.01** validation loss. The network architecture is repeated with GRU layers and 1-dimensional convolutional layers substituting LSTM layers, but this does not produce any significant performance improvement, with the public score staying within the range of **0.074** to **0.082**. A 2-dimensional convolutional neural network is also designed and trained on the sine matrix data but does not display good performance, with a public score of **0.09842**. We also tried an Orbital graph convolutional neural network [7], which adds an encoder-decoder network before the convolution layers. Unfortunately it did not learn anything and predicted the same values for all samples.

## Composite Neural Network

The deep feedforward network is combined with the RNN to train on both the composition data and the eigenspectrum data. The composite model produces better results than the individual feedforward and RNN models, with a public score of **0.06544**. The RNN portion is tested with both bidirectional LSTM and GRU layers. It is observed that the composite model implemented with LSTM layers performs better.

A composite model is also tested separately on the added atom properties data and the radial distribution function data. The features for atom properties are reduced by first fitting a gradient boosted decision tree and taking the features that had relative importance above the mean. The performance is however unsatisfactory, with a public score of **0.2104** when trained on all features and **0.077** when trained on the filtered features, including the radial distribution function.

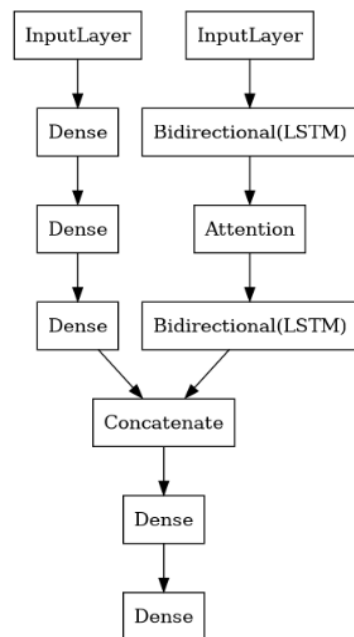
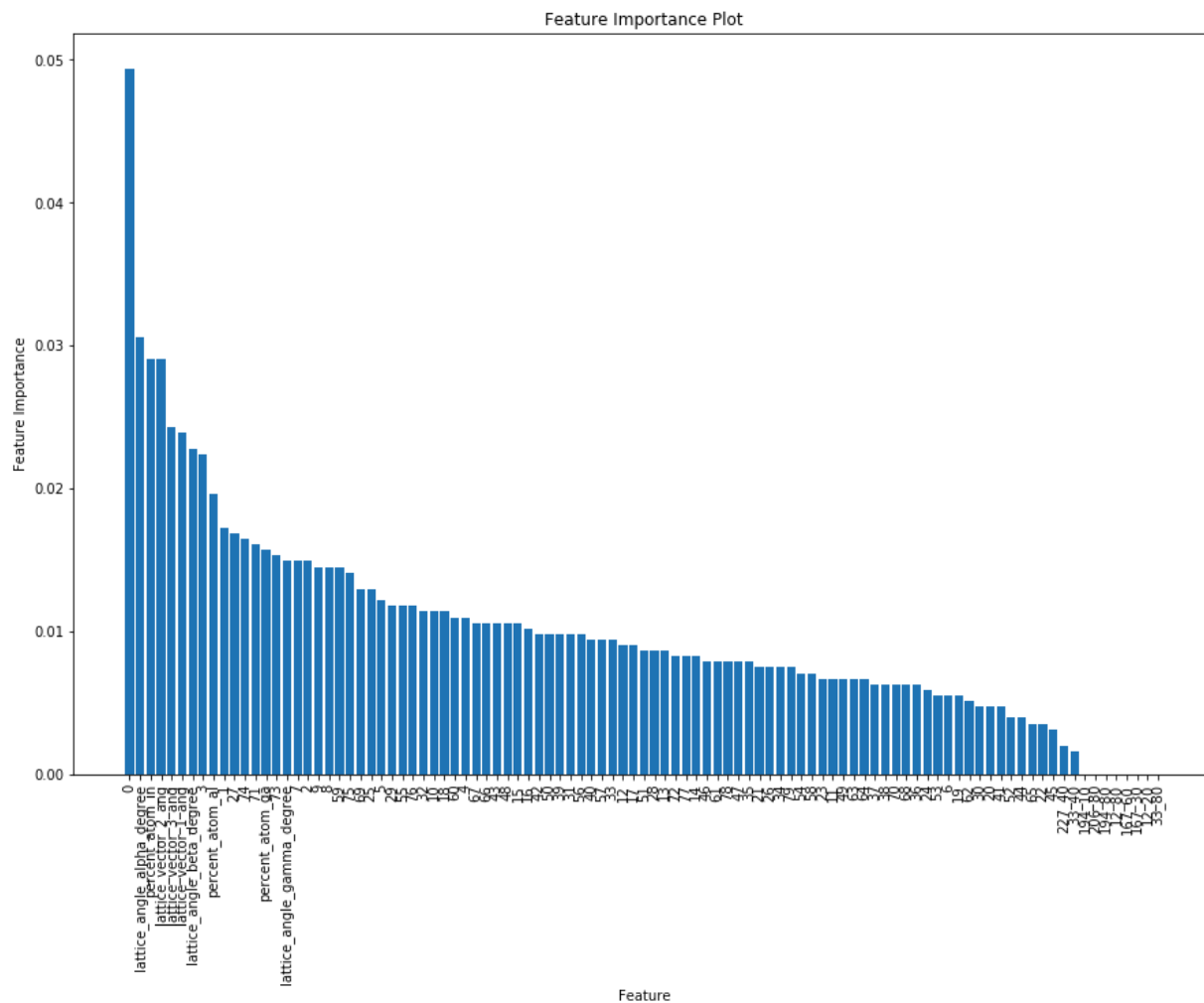


Figure 23. Composite Model Architecture

An implementation of the composite model architecture is shown. Several hyperparameter permutations were experimented on, including the addition of L1 and L2 regularisation, varying dropout rates, batch normalisation, adding attention layers, varying numbers of neurons, varying numbers of layers, and varying activation functions, including Leaky ReLU and ELU. It is observed that both Leaky ReLU and ELU had a negative impact on model performance, and the use of ReLU shows the best performance on the validation set.

## XGBoost using Eigenspectrum

XGBoost is implemented and trained on the composition data and eigenspectrum. Using 5 fold cross validation, a best max depth of **4** is observed for formation energy, with **190** estimators, and the best max depth of **4** is observed for band gap energy with **80** estimators. This model produced our best results.




**Figure 24.** Feature Importance for Formation Energy

As shown in **figure 24**, the first principal component is the most significant feature, capturing the regions of the eigenspectrum with the highest variance. This pivotal feature is succeeded in importance by **percent In** and the **lattice vector** dimensions, which are critical in predicting **formation energy**.



# Public Leaderboard Score

Our top score on the public leaderboard is 0.04806, achieved with a single XGBoost model. Attempts to combine and average other submissions have not surpassed this performance, indicating the single model's superiority in our current approach.

Competition Notebook	Run	Private Score	Public Score
 <a href="#">Nomad2018 Predicting Transparent Con...</a>	237.4s	0.06676	0.04806

## References

- [1] <https://www.kaggle.com/code/kemuel/python-exploration-with-domain-knowledge>
- [2] <https://arxiv.org/abs/1503.07406>
- [3] <https://www.kaggle.com/code/asatoonishi/using-sine-matrix>
- [4] <https://arxiv.org/abs/1705.01043>
- [5] <https://www.kaggle.com/code/leo1988/exploratory-data-analysis-using-plotly>
- [6] <https://www.kaggle.com/code/srserver85/boosting-stacking-and-bayes-searching>
- [7] M. Karamad *et al.*, "Orbital graph convolutional neural network for material property prediction," *Physical Review Materials*, vol. 4, no. 9, 2020.  
doi:10.1103/physrevmaterials.4.093801