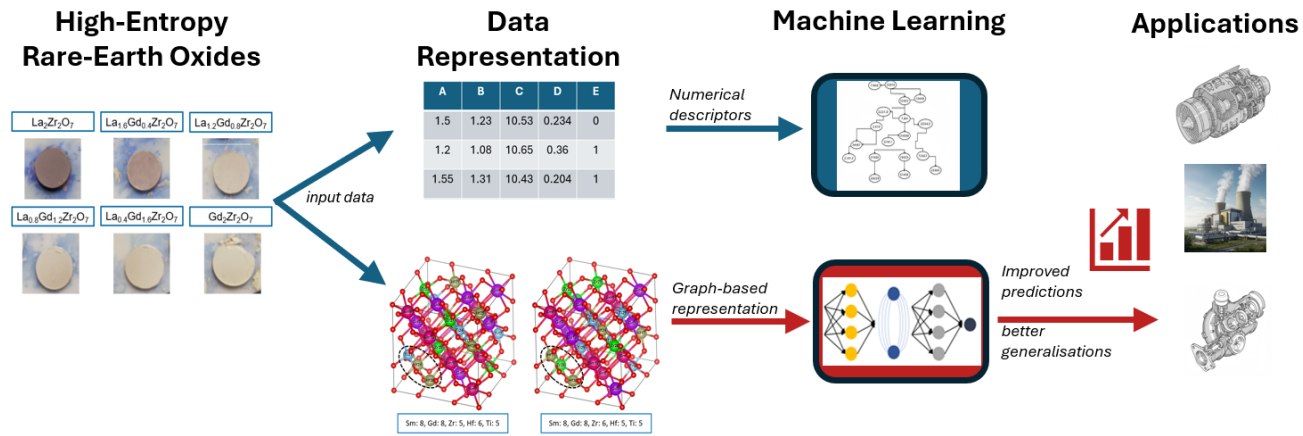


Graphical Abstract

A Methodological Study on Data Representation for Machine Learning Modelling of Thermal Conductivity of Rare-Earth Oxides

Amiya Chowdhury, Acacio Rincón Romero, Eduardo Aguilar-Bejarano, Halar Memon, Graziela Figueredo, Tanvir Hussain



## Highlights

### **A Methodological Study on Data Representation for Machine Learning Modelling of Thermal Conductivity of Rare-Earth Oxides**

Amiya Chowdhury, Acacio Rincón Romero, Eduardo Aguilar-Bejarano, Halar Memon, Graziela Figueredo, Tanvir Hussain

- Use of graph neural networks to model and predict thermal conductivity for high-entropy rare-earth oxides
- Comparison of graph data representation of compositions with handcrafted descriptors found in the literature
- Experiments show improved performance and generalisation with graph representations, demonstrating its suitability for modelling the problem

# A Methodological Study on Data Representation for Machine Learning Modelling of Thermal Conductivity of Rare-Earth Oxides<sup>\*</sup>

Amiya Chowdhury<sup>a,\*,1</sup>, Acacio Rincón Romero<sup>b,2</sup>, Eduardo Aguilar-Bejarano<sup>c,3</sup>, Halar Memon<sup>d,4</sup>, Graziela Figueredo<sup>e,5</sup> and Tanvir Hussain<sup>f,6</sup>

<sup>a</sup>Centre for Excellence in Coatings and Surface Engineering, University of Nottingham, University Park, Nottingham, NG7 2RD, UK

<sup>a</sup>Centre for Excellence in Coatings and Surface Engineering, University of Nottingham, University Park, Nottingham, NG7 2RD, UK

<sup>b</sup>School of Chemistry, University of Nottingham, Jubilee Campus, Nottingham, NG8 1BB, , UK

<sup>a</sup>Centre for Excellence in Coatings and Surface Engineering, University of Nottingham, University Park, Nottingham, NG7 2RD, UK

<sup>c</sup>Centre for Health Informatics, University of Nottingham, University Park, Nottingham, NG7 2RD, UK

<sup>a</sup>Centre for Excellence in Coatings and Surface Engineering, University of Nottingham, University Park, Nottingham, NG7 2RD, UK

## ARTICLE INFO

### Keywords:

graph neural networks, thermal conductivity, data representation, rare-earth oxides, machine learning

## Abstract

Quantitative structure-activity relationship (QSAR) modelling is widely employed in materials science to predict properties of interest and extract useful descriptors for measured properties. In thermal barrier coatings (TBC), QSAR can significantly shorten the experimental discovery cycle, which can take years. Although machine learning methods are commonly employed for QSAR, their performance depends on the data quality and how instances are represented. Traditional, hand-crafted descriptors based on known material properties are limited to represent materials that share the same basic crystal structure, limited the size of the dataset. By contrast, graph neural networks offer a more expressive representation, encoding atomic positions and bonds in the crystal lattice. In this study, we compare Random Forest (RF) and Gaussian Process (GP) models trained on hand-crafted descriptors from the literature with graph-based representations for high-entropy, rare-earth pyrochlore oxides using the Crystal Graph Convolutional Neural Network (CGCNN). Two different types of augmentation methods are also explored to account for the limited data size, one of which is only applicable to graph-based representations. Our findings show that the CGCNN model substantially outperforms the RF and GP models, underscoring the potential of graph-based representations for enhanced QSAR modelling in TBC research.

## 1. Introduction


Operating temperatures for nickel based superalloys are between 800-1000°C (Mouritz (2012)). Turbine inlet temperatures in gas-turbine engines can exceed 1500°C (Centrich, Shehab, Sydor, Mackley, John and Harrison (2014)). Thermal barrier coatings (TBC) provide temperature reductions on the nickel-based superalloy of 100-200°C (Darolia (2013)), allowing turbine engines to operate effectively at higher temperatures. Higher engine operating temperatures lead to increased thermodynamic efficiency and lower CO<sub>2</sub> emissions. Ytria stabilised zirconia has been the industry standard TBC since the 1970s due to its combination of thermal and mechanical properties (Stecura (1978); Cao, Vassen and Stöver (2004); Vaßen, Bakan, Mack and Guillon (2022)). However, this composition is limited to a temperature of 1200°C due to the formation of tetragonal and cubic phases above this temperature, leading to undesired expansion of about 3 to 5 % (Cao et al. (2004);


Liu, Shi, Geng, Wang, Xu and Chen (2022); Pakseresht, Sharifianjazi, Esmaeilkhani, Bazli, Reisi Nafchi, Bazli and Kirubakaran (2022)). It is also susceptible to chemical attack by calcia-magnesia-alumino-silicate (CMAS). To develop a replacement that outperforms YSZ as a coating, the new desirable material would have lower thermal conductivity, comparable fracture toughness (1 – 2 MPam<sup>1/2</sup>), thermal expansion coefficient (~ 10–11×10<sup>-10</sup> K<sup>-1</sup>)(Vaßen, Jarligo, Steinke, Mack and Stöver (2010)) and thermal cyclic life-span, higher resistance to CMAS and better phase stability at its operating temperatures (~ 1200–1300 °C)(Bakan and Vaßen (2017); Vaßen et al. (2022); Mehboob, Liu, Xu, Hussain, Mehboob and Tahir (2020)).

Due to the existing limitations of YSZ coatings, there has been an increasing interest in the discovery of high-performance alternatives (Vaßen et al. (2022)). Pyrochlore structures are currently the most investigated group (Bakan and Vaßen (2017); Vassen, Cao, Tietz, Basu and Stöver (2004); Fergus (2014); Wu, Wei, Padture, Klemens, Gell, García, Miranzo and Osendi (2004)). This is partly due to their lower thermal conductivity, higher thermal stability and lower oxygen diffusivity (Vassen et al. (2004); Zhu, Meng, Zhang, Li, Xu, Reece and Gao (2021a)) compared to YSZ. However, these compositions have a much lower thermal expansion coefficient and mechanical toughness compared to YSZ, making them unsuitable for replacement. The pyrochlore structure can be modified by cation substitution, forming high-entropy or multi-component ceramics

\*

\*Corresponding author

 amiya.chowdhury@nottingham.ac.uk (A. Chowdhury); (A.R. Romero); (E. Aguilar-Bejarano); (H. Memon); (G. Figueredo); tanvir.hussain@nottingham.ac.uk (T. Hussain)

 <https://orcid.org/0009-0004-2537-4962> (A. Chowdhury); (A.R. Romero); (E. Aguilar-Bejarano); (H. Memon); G.Figueredo@nottingham.ac.uk (G. Figueredo); (T. Hussain)

ORCID(s):

1

(HECs and MCCs). The cubic pyrochlore structure has the general formula  $A_2^{3+}B_2^{4+}O_6O'$ , where  $A^{3+}$  is typically a rare-earth element and  $B^{4+}$  is usually a transition metal. By including multiple elements in different combinations in the  $A$  and  $B$  sub-lattices, both the thermal and mechanical properties of the material can be manipulated (Wan, Pan, Xu, Qin, Wang, Qu and Fang (2006); Ren, Wan, Zhao, Yang and Pan (2015); Liu, Zhao and Jiang (2014)). The introduction of elements with different ionic radii, mass and electro-negativities, provides the potential to tailor the thermal and mechanical properties of the material, allowing for an ensemble of properties that would not otherwise be achievable with single elements. Additionally, distortion in the lattice caused by different  $Re - O$  bond lengths has been shown to reduce thermal conductivity (Wright, Wang, Ko, Chung, Chen and Luo (2019); Tsai and Yeh (2014)). The ability to tailor both the mechanical and thermal properties of these compositions makes it a worthwhile search space for potential new TBC materials.

Due to the large number of combinations this allows, experimentally measuring the properties of these potential TBC materials would be too costly and time-consuming. Computational methods using first-principles calculations, such as density functional theory (DFT) take time to set up and run (Seko, Togo and Tanaka (2018)). And third order problems, such as thermal conductivity are computationally expensive and often limited to crystals with a small number of atoms in their unit cell (Choi, Lee, Kim, Moon, Jeong and Han (2022); Luo, Li, Yuan, Liu and Fang (2023)). As an alternative to overcome laborious, resource intensive processes, machine learning (ML) has been increasingly used in materials science, for property and micro-structure predictions, and composition generation (Liu, Zhao, Ju and Shi (2017)). Conventional machine learning methods used include random forest, support vector machines and shallow neural networks. They often take as input hand-crafted and/or non-problem specific descriptors to represent information about the input data points representing materials. These descriptors are hypothesised to be relevant, that is, to have some degree of correlation to the target variable (Ward, Agrawal, Choudhary and Wolverton (2016); Jha, Ward, Paul, Liao, Choudhary, Wolverton and Agrawal (2018)). The definition and selection of these descriptors generally requires domain knowledge— although often generic chemistry descriptors are also used — and an understanding of the machine learning algorithms (Ward et al. (2016); Jha et al. (2018); Guyon and Elisseeff (2003)). This is an important step, as relevant descriptors can help understand materials structure-property relationships.

Yang *et al.* conducted an investigation of the atomic parameters that mostly affect the prediction of thermal and mechanical properties of pyrochlores Yang, Zhu, Sheng, Nian, Li, Song, Lu, Yang and Liu (2018)). Parameters include the average atomic mass of  $A$  and  $B$  cations ( $M_A$  and  $M_B$ ), period ( $P_A$  and  $P_B$ ) in the periodic table, average cationic radius ( $R_A$  and  $R_B$ ), cationic radius ratio ( $R_A/R_B$ ),

electronegativity ( $EN_A$  and  $EN_B$ ), density ( $\rho$ ) and lattice constant ( $\alpha$ ). Miyazaki, Tamura, Mikami, Watanabe, Ide, Ozkendir and Nishino (2021) developed a ML model to predict the lattice thermal conductivity of half-heusler compounds using a two stage system, that first predicted lattice parameters using ML and then used the predicted lattice parameter as a descriptor for training the thermal conductivity prediction model. The descriptors for the lattice parameter prediction model included the average atomic radius and average atomic mass of each site in the lattice ( $r_1, r_2, r_3, m_1, m_2, m_3$ ). The thermal conductivity model used the previously predicted lattice parameter and 54 other descriptors. These 54 descriptors were various functions of  $r_1, r_2, r_3, m_1, m_2, m_3$  describing crystallographic properties specific to half-heusler compounds. Using the Wrapper method (Kohavi and John (1997)), these 55 descriptors were then reduced to just 4 in order to optimise the training of the ML model. Their database was obtained from DFT calculated values of lattice parameter and thermal conductivity of 143 materials. The predictive models employed were multiple linear regression and boosted decision tree regression, with the latter achieving better results in both stages. For thermal conductivity, the boosted decision tree model performed best with an error of  $\pm 4\%$  and  $R^2$  of 0.84. The final trained model predicted thermal conductivity near instantaneously with a an accuracy of  $\pm 4\%$ , while DFT calculations for the same compound took up to 72 h.

Despite the advantages over numerical methods, ML are restricted by their training data and how the data is represented — that is, the quality of the descriptors defined. ML is expected to learn from data sets that provide a good representation of the search space to achieve good predictive power. Furthermore, trained models have been shown to be inaccurate when predicting properties of compounds with values of the target property that are significantly different to those represented in the training set. A general model trained on various different types of compounds and materials would be ideal, but is difficult to implement with algorithms like Random Forest. Crystal structures are not well represented for modelling when using solely descriptors such as their chemical formula or simple crystal structure information. The position of atoms in the lattice and bonding between atoms in different sites, for instance, have significant effects on thermal transport through the material. Thermal conductivity in crystals, is governed by phonon (or quantised lattice vibration modes) scattering within the lattice (Pala, Abbas, Rockstuhl, Menzel, Mühligh, Lederer, Brown, Bright, Curley, Koh and et al. (2012)). Phonon scattering is dictated by the strengths and positions of the various bonds present throughout the lattice structure (Liu, Wang, Zhou, Liao and Li (2007)). A representation that captures this information will likely lead to models with more predictive power. In the current literature, most ML models investigating crystal properties use simple descriptors (composition and crystal structure information), being limited to the same type of materials (Luo et al. (2023); Wright et al. (2019)).

Graph Neural Networks (GNN) capture structural information as a graph composed of nodes and edges. Edges represent connections between nodes; they also contain edge features information, such as the distance between two nodes. GNNs are commonly used in pharmaceutical research to represent 2D molecules (Carracedo-Reboredo, Liñares-Blanco, Rodríguez-Fernández, Cedrón, Novoa, Carballal, Maojo, Pazos and Fernandez-Lozano (2021)) for QSAR modelling and materials discovery (Zhang, Chen, Zhong, Wang, Jiang, Zhang, Jiang, Zheng and Li (2022); Jiang, Wu, Hsieh, Chen, Liao, Wang, Shen, Cao, Wu and Hou (2021)). While various methods exist to represent organic molecules as graphs, like SMILES, they are not suitable for representation of 3D, periodic crystal structures. Xie and Grossman, 2018, represented crystals as a graph structure allowing the use of over 3000 data-points to train various property prediction models using their custom algorithm, the Crystal Graph Convolutional Neural Network (CGCNN) (Xie and Grossman (2018)). They convert crystallographic information files (CIF) to a 2D graph used as the input for CGCNN. Zhu, He, Gong, Xie, Gorai, Nielsch and Grossman (2021b) used CGCNN to successfully predict the thermal conductivity of all known rare earth chalcogenides in the Inorganic Crystal Structure Database (ICSD), for applications in thermo-electric power generation.

For an improved screening of potential materials for TBC's, different representation methods for high-entropy pyrochlore compositions are explored in this study. We evaluate hand-crafted descriptors and a graph-based representation of the lattice structure. The hand-crafted representations tested use compositional descriptors and crystallographic descriptors. Compositional descriptors to create a predictive model allow us to study the effect of different atomic species on the thermal conductivity, making it useful for material design. Crystallographic descriptors, however, offer a more standardised representation of materials, potentially reducing model bias due to the distribution of species in the training dataset. Graph-based representation combines the compositional and crystallographic hand-crafted properties and allows for additional material information to be included. Using interpretative ML, each representation can offer different insights into those properties to be selected in next generation rational designs. This study focuses on the accuracy of the models trained with each approach to determine their potential usefulness both as a predictive model for high-throughput screening, and an interpretative model that can be used to study the science of thermal conductivity. Materials with a pyrochlore structure are the primary focus due to their potential for low thermal conductivity and capability to form high-entropy compositions which provides a sufficiently large search-space to test the capabilities of machine learning. Thermal conductivity is selected as the prediction target as without a sufficiently low thermal conductivity the material would not be able to function as a thermal insulator (Vaßen et al. (2022)). For the ML models, the hand-crafted descriptors are used to train

Random Forest Regressor (Pedregosa, Varoquaux, Gramfort, Michel, Thirion, Grisel, Blondel, Prettenhofer, Weiss, Dubourg, Vanderplas, Passos, Cournapeau, Brucher, Perrot and Duchesnay (2011)) and Gaussian Process Regressor (Pedregosa et al. (2011)), while the graph based representation was used to train the CGCNN. Training data for the model was gathered from literature.

It should be noted that most of the CIF files used in the training data are not generated by experimental measurements or DFT. DFT calculated CIF files of the multi-component/high-entropy compositions would likely have more accurate information regarding atomic coordinates and bond lengths. This avenue would, however, require large amounts of time and computational resources and was, thus, not explored for this paper.

## 2. Methodology

### 2.1. Database

Thermally sprayed coatings have varying micro-structures that depend on the processing conditions and compositions. To simplify the selection of descriptors, only data from pellets were considered. While DFT generated data was considered, there is a lack of sufficient calculated thermal conductivity data on rare-earth pyrochlores, especially high-entropy compositions. To ensure consistency, the data was collected from a single source. The data selected for our investigation was obtained from Wright, Wang, Ko, Chung, Chen and Luo (2020). The authors studied thermal and mechanical properties of 22 primarily single-phase, rare-earth pyrochlores with multiple components in both *A* and *B* positions. The *A* position was composed of rare-earths in various combinations from single element to up to 7 elements. The *B* position was composed of *Zr*, *Sn*, *Hf* and *Ti* in various combinations from single element to up to 4 elements.

Densities of samples were measured using the Archimedes method, following the ASTM C373-18 Standard. Relative density of the samples are around 95% on average. Thermal conductivity was calculated from thermal diffusivity which was measured at room temperature. Thermal diffusivity was measured using LFA 467 *HT Hyperflash* (NETZSCH, Germany), laser flash analysis. Specific heat was calculated using Neumann-Kopp (Suresh, Seenivasan, Krishnaiah and Srirama Murti (1997); Leitner, Voňka, Sedmidubský and Svoboda (2010)) rule and used to calculate the thermal conductivity of the pellets.

### 2.2. Hand-Crafted Descriptors

Yang et al. used least absolute shrinkage and selection operator (LASSO) to investigate atomic/crystallographic parameters affecting the prediction of thermal and mechanical properties of pyrochlores (Yang et al. (2018)). Parameters include atomic mass of *A* and *B* cations ( $M_A$  and  $M_B$ ), period ( $P_A$  and  $P_B$ ) in the periodic table, ionic radius ( $R_A$  and  $R_B$ ), cationic radius ratio ( $R_A/R_B$ ), electronegativity ( $EN_A$  and  $EN_B$ ), density ( $\rho$ ) and lattice constant ( $\alpha$ ). Compositional parameters, represented by atomic fractions of



Descriptors	Description
$LaA - YbA$	Atomic fraction of elements in the $A$ cation site, from lanthanum to ytterbium, normalised to 1.
$ZrB, HfB, SnB, CeB,$ $TiB$	Atomic fraction of elements in the $B$ cation site, normalised to 1.
$RA$	Effective ionic radii of $A$ cations
$RB$	Effective ionic radii of $B$ cations
$MA$	Effective atomic mass of $A$ cations
$MB$	Effective atomic mass of $B$ cations
$RA/RB$	Cationic ratio
$P$	Presence or absence of pyrochlore phase represented by 1 or 0. This was based on the $RA/RB$ criteria for pyrochlore.
$a$	Theoretical lattice constant calculated using (Mouta, Silva and Paschoal (2013)).
<i>Entropy</i>	Lattice configurational Entropy

**Table 1**

Brief description of the 26 descriptors used to describe the compositions in the model

each element present in the composition, were considered alongside crystallographic properties, such as the effective mass and radii of each cation, as well as the lattice parameter. Table 1 lists all descriptors considered for our model.

Theoretical lattice constant ( $a$ ) was calculated using Equation 1 from Mouta (Mouta et al. (2013)) using the average ionic radii of the cations.

$$a_2 = \frac{8}{3^{1/2}} \left[ 1.43373(R_A + R_O) - 0.42931 \frac{(R_A + R_B)^2}{R_B + R_O} \right] \quad (1)$$

Where  $R_A$ ,  $R_B$  and  $R_O$  are the ionic radii of the  $A$  cation,  $B$  cation and Oxygen, respectively. This was used as a quick method to obtain the lattice parameters of multi-component composition data obtained from literature.

Stability of the pyrochlore structure is dependant on the ratio of ionic radii of  $A$  and  $B$  cations ( $R_A/R_B$ ). For lanthanide zirconates ( $Ln_2Zr_2O_7$ ), when  $R_A/R_B \leq 1.46$  (Fuentes, Montemayor, Maczka, Lang, Ewing and Amador (2018)) a fluorite phase is formed instead. A simple phase descriptor ( $P$ ) is used to differentiate between the two phases. The remaining descriptors are applicable to both pyrochlores and fluorites. This is possible because the database used consists mainly of single-phase compositions.

The configurational entropy refers to the portion of the entropy of a system related to the precise positions of individual particles that make up the system. This parameter accounts for the number and proportion of each distinct atomic species present in the  $A$  and  $B$  sub-lattices of the pyrochlore lattice. To account for multiple sub-lattices in complex crystal structures, (Dippo and Vecchio (2021)) proposed a new entropy metric shown in Equation 2.

$$EM = \frac{S_{SL/mol\ atoms}^{config}}{R} * L \quad (2)$$

Where  $R$  is the ideal gas constant,  $L$  is the total number of sub-lattices and  $S_{SL/mol\ atoms}$  is the configurational entropy calculated using the sub-lattice model shown in Equation 3.

$$S_{config}^{SL} = \frac{-R \sum_S \sum_i a^S X_i^S \ln(X_i^S)}{\sum_S a^S} \quad (3)$$

Where  $a^S$  is the number of sites on the  $S$  sub-lattice and  $X_i$  is the fraction of each individual species  $i$  in the sub-lattice. For the  $A_2B_2O_6O'$ ,  $\sum a_S$  is 11 and  $L$  is 5 when considering the oxygen vacancy as an additional species.

Given that all the crystallographic parameters are directly calculated from the atomic fractions, it would be redundant to consider them together. A correlation matrix (figure ??) further supports this argument. Due to that, the atomic fractions and the atomic/crystallographic parameters are used as two separate methods. The crystallographic parameters are a more common representation of the problem than the atomic fractions. However, the effect of composition on thermal conductivity is also of interest, as it would be more intuitive from a materials design perspective.

### 2.2.1. Machine Learning Algorithms for hand-crafted descriptors

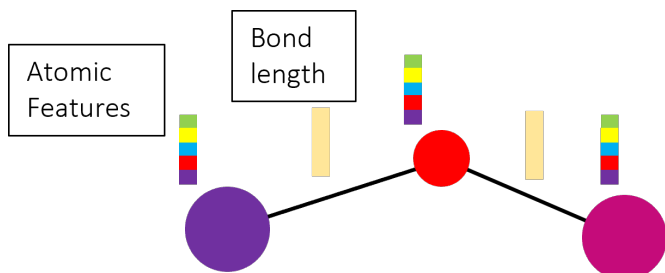
Random Forest (RF) (Pedregosa et al. (2011)) and Gaussian Process (GP) regressors (Pedregosa et al. (2011)) were selected to create predictive models for our data. They were implemented using the *scikit-learn* package (Pedregosa et al. (2011)) in *Python* 3.9.12. Hyper-parameters were optimised using *GridSearchCV*.

The RF was set up with 1000 decision trees ( $n\_estimators$ ) and a value of 2 was used for  $min\_sample\_leaf$  based on the results using *GridSearchCV*. Random state of the model was set to 42. All other values were kept at default for the package.

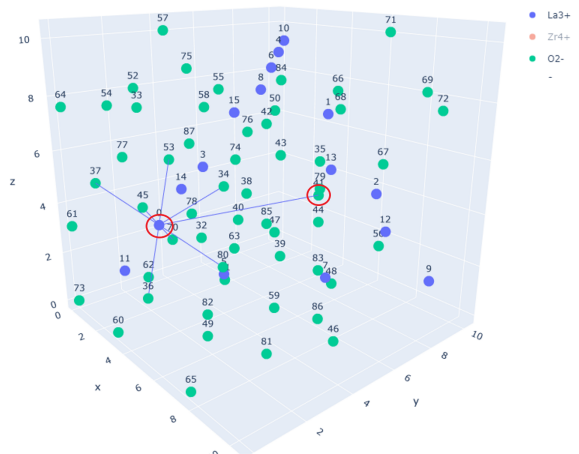
The GP used a combination of *ConstantKernel* and *Radial Basis Function* (RBF) for the kernel. For *ConstantKernel*, the constant was set to 1.0 and the bounds were set to "fixed". For *RBF*, length scale was set to 1.0 and the bounds were also set to "fixed".

### 2.3. Graph-based representation

The CGCNN takes CIF files representing crystal structures as inputs (Xie and Grossman (2018)). The CIF is



**Figure 1:** Structure of the graph used to represent each composition. The nodes represent each atom in the unit cell, containing 9 atomic descriptors and the edges represent bonds between two atoms and contain information about the bond length.



**Figure 2:** Virtual bond between a lanthanum atom (atom 0) and an oxygen atom (atom 41), marked by red circles

converted to a graphical representation, with nodes carrying atomic species data and edges denoting inter-atomic bonding with their corresponding distance. A simplified representation is shown in Figure 1.

The nodes contain detailed atomic information about the atomic species represented by the node. Bonded atoms are defined as being located within a radius of  $3\text{\AA}$  around each atom. Bonds are formed starting with the shortest bond length up to a maximum of 12 possible bonds. The CIF file only contains information about the unit cell; bonded atoms outside the cell are not represented. Due to the symmetry of crystal lattices, this limitation can be overcome by forming a virtual bond with another atom in the equivalent position across the line of symmetry. This allows the periodicity of the crystal to be represented in the graph. Figure 2 shows the virtual bond between a lanthanum atom (atom 0) and an oxygen atom (atom 41), across the cell. The inter-atomic distance between the atoms is found by subtracting the actual distance from the lattice parameter.

The CGCNN is commonly used with symmetrised CIF files that represent the *occupancy* or probability of an atom

species to exist in a particular site within the unit cell, rather than the actual coordinate of each individual atom. For single-element pyrochlores, the occupancy of each species in its specific sub-lattice is 1.0. In high-entropy or multi-component pyrochlores, the species have partial occupancies, with values smaller than 1.0. The CGCNN is unable to handle partial occupancies from the conventionally used symmetrised CIF files. The materials project (Jain, Ong, Hautier, Chen, Richards, Dacek, Cholia, Gunter, Skinner, Ceder and et al. (2013)) possesses an alternative, *computed* CIF file, which contain no symmetry information, but contains the actual Cartesian coordinates of each atom present in the lattice.

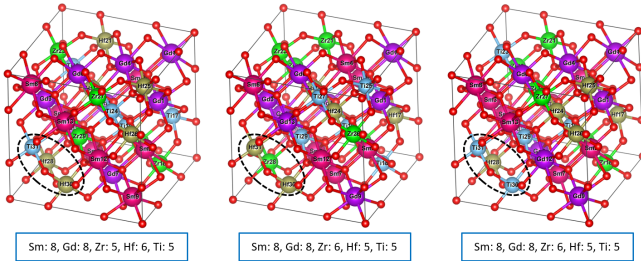
### 2.3.1. Generating Crystallographic Information Files

While CIF files for the single-element rare earth compositions are easily found in databases such as the Inorganic Crystal Structure Database (ICS) and the Materials Project (Jain et al. (2013)), there are very few available for multi-component pyrochlores. Either experimental measurements or DFT calculations are required to generate the lattice constants of different crystals. This method is very time-consuming to generate the CIFs for the all data being used in this project. CIF files for multi-component compositions were generated artificially by using an existing pyrochlore CIF file as a template, replacing the atomic species present at each coordinate and changing the lattice parameters. CIF files using the conventional standard, i.e., with the distribution of species represented by sub-lattice occupancies rather than cartesian coordinates, cannot be used to generate the lattice graphs needed for the GNN, specifically in the case of HECs and MCCs. This is because HECs and MCCs contain partial occupancies, i.e., occupancy values that are less than one. As the graph consists of nodes and edges representing the actual position and connections between atoms, it cannot deal with partial occupancies. To account for this, the non-symmetrised format for CIFs are used to generate the lattice graphs. The non-symmetrised formats always use the *P1* space group and contain real positions for each atom in the lattice, rather than the partial occupancies. For compositions with a fractional number of atoms, i.e., 5.333 atoms of *Zr*, *Hf* and *Ti*, the values are rounded by randomly sampling the species from the list for the A/B sub-lattices and rounding it up or down as needed to keep the total number of atoms at 16 (the maximum number of atoms in the A and B sub-lattices) and then iterating until all species in the sub-lattice have discrete numbers. Furthermore, the positions of each atom within the appropriate sub-lattice is randomised to avoid clustering of the same species. Detailed steps of the method are shown in Table 2. Figure 3 shows the visualisation of the unit cell of  $(Sm_{1/3}Gd_{1/3}Eu_{1/3})_2(Zr_{1/2}Hf_{1/2})_2O_7$ , generated by our method. In this case, the europium atoms in the 16d site (A sub-lattice), are replaced with samarium, gadolinium and europium according to their proportions.

The template file used for generating the CIFs is  $Eu_2Zr_2O_7$ , obtained from the Materials Project (Project (2020)). Lattice

CIF generation algorithm	
1:	Read template CIF
2:	Read database csv file
3:	Iterate over each entry in database
4:	Create new file for each entry and copy template data into the new file
5:	Replace lattice parameter, cell volume, composition on appropriate lines using info from database
6:	Calculate number of atoms in A and B sub-lattices using atomic fractions of species in database entry
7:	Round the number of atoms (for HECs and MCCs) ensuring the total number of atoms is 16 for each sub-lattice
8:	Create two separate lists of all atoms in A and B sub-lattices and shuffle the order of atoms in each list to randomise their positions within each sub-lattice
9:	Go through the A/B sub-lattice coordinates in the new file, replacing each species present in that coordinate with the species from the aforementioned list

**Table 2**  
Method for generating CIF for high-entropy and multi-component ceramics



**Figure 3:** Three iterations of  $(Sm_{1/3}Gd_{1/3}Eu_{1/3})_2(Zr_{1/2}Hf_{1/2})_2O_7$  with slightly differing proportions of each species (as shown below each image) and positions of atoms within the same sub-lattice. Here the changes are only in the *B* sub-lattice as the number of *Sm* and *Gd* atoms are exactly 8.

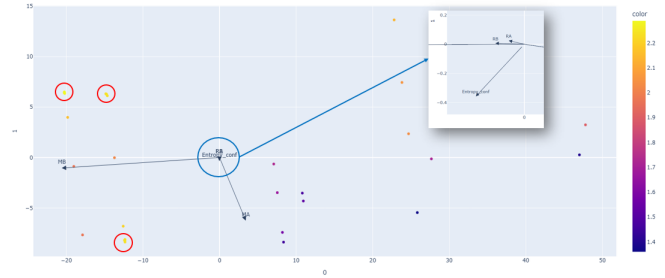
parameters and cell volumes were obtained using Equation 1.

## 2.4. Data Augmentation

Data augmentation uses existing data to artificially create new instances for training ML. This is particularly useful for small data-sets, where the augmented data can improve both size and data distribution. Two augmentation techniques were applied to the data, one for the handcrafted descriptors and another for the graph-based descriptors. The objective was to further test the efficacy of the input data representation for more points, other than those provided in the original data investigated.

Thermal conductivity ( $W/mK$ )	Relevancy
1.36	1
1.5	1
1.7	0.8
1.97	0.6
2.1	0.8
2.145	1
2.29	1

**Table 3**  
Custom relevancy matrix for SMOGN



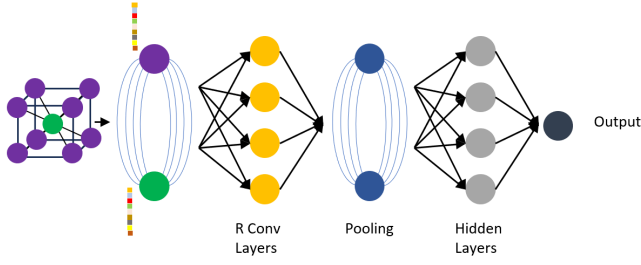
**Figure 4:** PCA plot of SMOGN augmented data-set

### 2.4.1. Handcrafted Descriptors - SMOGN

For the handcrafted descriptors, a data balancing algorithm, namely SMOGN (Kunz (2020)) was applied to augment the dataset. SMOGN combines over-sampling of minority cases with Gaussian noise to generate synthetic data points. It starts by identifying regions where the target variable is under represented (minority cases). Subsequently, it generates synthetic data points in these regions. Synthetic data are created by interpolating nearby instances within the minority region. Gaussian noise is then added to the synthetic instances to ensure that the new points are not simply linear combinations of existing data. This helps to improve generalisation and avoid potential biases in the model. Areas with more scarcity of available instances are identified via a relevancy value between 1 and 0 (1 being higher relevancy). The default function for identifying relevant regions for over-sampling assumes a bell curve distribution of values in the data-set, attributing higher relevancy to values on either end of the distribution. To prioritise increasing the number of data points available, we define the relevancy points as follows (Table 3):

It should be noted, that SMOGN is mainly designed to balance an existing data-set and does not simply increase the number of data points by an arbitrary amount. The variable chosen for optimisation was thermal conductivity. A total of 13 new, synthetic data points were generated with SMOGN. Of those 13, 4 were found to be duplicates. Overall, 9 new data points were generated through SMOGN, increasing the database size from 21 to 30. The principal component analysis plot (PCA), of the SMOGN augmented data-set is shown in Figure 4.





**Figure 5:** CGCNN model architecture. CIF file is converted to a graph consisting of nodes and edges representing atoms and their bonds, respectively. The nodes are embedded with vectors containing information representing the atomic species which are then aggregated in the pooling layer to represent the entire crystal

The figure shows a correlation between high  $M_B$  and low  $M_A$  with high thermal conductivity. The data points added by SMOGN are clustered within the areas denoted by red circles.

The relevancy matrix was designed to generate the most number of data points, which is likely why the distribution of TC for the augmented data-set is still skewed and even slightly worse than the original. The majority of the generated data points had a TC value of approximately  $2.2 \text{ W/mK}$ . A cursory validation of the generated data points was done by using a simple script to match potential compositions of  $A_2B_2O_7$  structure to the descriptors. The thermodynamic stability of these compositions were not validated theoretically or experimentally as this is outside the scope of this paper.

#### 2.4.2. Graph-based Data - Lattice perturbation

Multi-component crystal lattices can contain random variations in the positions and amount of different species, within their respective sub-lattice (Aidhy (2024)). These variations were used to generate copies of each composition in the database, tripling the size of the initial training set. Figure 3 shows three versions of the same composition  $((Sm_{1/3}Gd_{1/3}Eu_{1/3})_2(Zr_{1/2}Hf_{1/2})_2O_7)$  with slight differences in the position and proportion of each species. The difference in proportion arises due to fractional occupancy.

### 2.5. CGCNN Model

The CGCNN architecture consists of 3 convolutional layers, 2 hidden layers and 1 pooling layer. Each convolution layer, learns the information from neighbouring atoms to generate a new feature vector for each atom in the graph. Following that, the pooling layer then uses the information from the convolutional layers to generate a vector representing the entire crystal.

The mean squared error is used as the loss function during training. Table 4 shows the hyperparameters used for the CGCNN model.

It should be noted that the bond length is not included as a descriptor for the GNN. Calculating the bond lengths of different  $Re - O$  pairs in HECs and MCCs is a complicated

Hyper-Parameter	Value
learning rate	0.01
batch-size	64
Optimiser	SGD

**Table 4**  
Effect of element distribution on prediction

process that would require DFT modelling for each composition. This is outside the scope of the current work and the decision was made to remove bond lengths as a descriptor entirely.

### 2.6. Model Evaluation

. Mean Squared Error (Equation 5) is used for the CGCNN during the training process. Ultimately, mean absolute error in Equation 4 is used to evaluate all the models as it is more intuitive than the MSE.

$$MAE = \frac{1}{N} \sum_{i=1}^n |y_i - y_i^*| \quad (4)$$

$$MSE = \frac{\sum_{i=1}^N (y_i - y_i^*)^2}{N} \quad (5)$$

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - y_i^*)^2}{\sum_{i=1}^N (y_i - \bar{y})^2} \quad (6)$$

Where,  $y_i$  is the actual measured value,  $y_i^*$  is the predicted value and  $\bar{y}$  is the mean of the actual values. Leave-one-out cross validation (LOOCV) is used to evaluate all models due to the limited size of the data-set. The MAE from each fold is then averaged and used to compare the different models as it is a more intuitive metric. The  $R^2$  was calculated by combining the results of all folds, for each model.

## 3. Results

A summary of the model error metrics derived from LOOCV is listed in Table 5, while Table 7 lists the performance of the models trained using the augmented datasets (both SMOGN and Lattice perturbation).

Figures 6-11 show the predicted vs. actual thermal conductivity values for each composition, using the different models. The Gaussian Process model trained with compositional descriptors shows the worst performance in all metrics. The negative  $R^2$  value indicates that the model was unable to learn any meaningful relationships within the data. The model performance seems to be strongly affected by a single data point ( $Sm_2Zr_2O_7$ ). As seen in Figure 6, this composition has a significantly higher error than the average (55 % vs the average of 11 %). The predicted thermal conductivity is significantly lower than

Model	MAE		$R^2$
	Mean	Std. Dev	Mean
(1) GP (compositional)	0.209	0.245	-1.255
(2) RF (compositional)	0.117	0.084	0.776
(3) RF (crystallographic)	0.112	0.089	0.781
(4) CGCNN	<b>0.029</b>	<b>0.043</b>	<b>0.97</b>

**Table 5**

Leave-One-Out-Cross validation results for models trained on the original dataset

Model	RF (compositional)	RF (crystallographic)	CGCNN
GP (compositional)	0.288	0.128	
RF (compositional)	-	0.973	
RF (crystallographic)	-	-	

**Table 6**

Statistical significance test to study differences between models using Wilcox Signed Rank Test

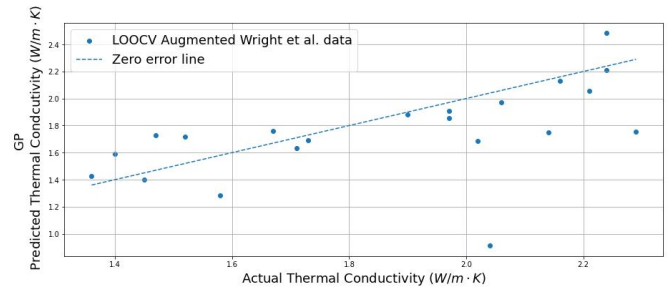
Model	MAE		$R^2$
	Mean	Std. Dev	Mean
(5) RF(crystallographic,SMOBN)	0.061	0.072	0.908
(6) CGCNN (augmented)	<b>0.024</b>	<b>0.042</b>	<b>0.975</b>
(7) CGCNN (SMOBN)	0.047	0.081	0.872

**Table 7**

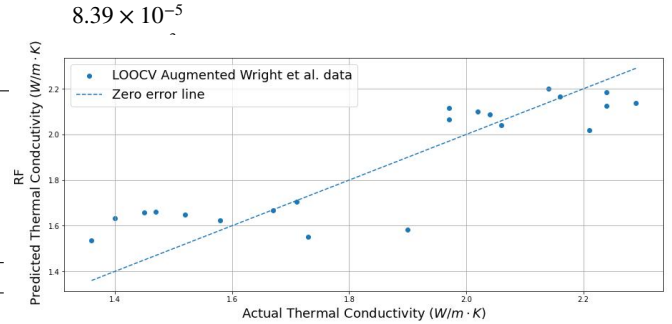
Leave-One-Out-Cross validation results for models trained on augmented data

the minimum in the training data-set. This point could be considered an outlier in the sense that it is one of only two single-element compositions in the training data, the other being lanthanum zirconate ( $La_2Zr_2O_7$ ). RF trained with compositional descriptors shows improvement over the GP model. Interestingly, the differences between the RF models using compositional and crystallographic descriptors seem to be minimal. Both models also show a similar trend in the error values for each composition as can be seen in figures 6 and 7. This implies that both sets of descriptors are similarly informative in regards to thermal conductivity. The Random Forest model trained on the SMOBN augmented data-set shows a significant improvement over the regular data-set. However, as can be seen from figure 9, the error trend for each composition appears to be similar. Particularly, in both models, the composition  $(Sm_{1/2}Gd_{1/2})_2(Zr_{1/3}Hf_{1/3}Ti_{1/3})_2O_7$  exhibit a higher error compared to other data points in the region. The augmented model appears to work better for thermal conductivity between 2.0 and 2.4  $W/mK$ . The  $R^2$  is also significantly improved, suggesting better generalisability compared to model 3.

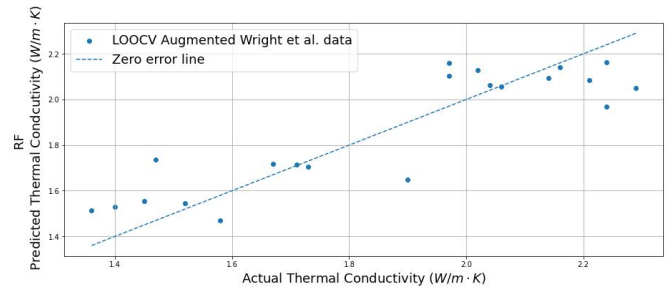
All three CGCNN models show significantly greater performance compared to the RF and GP models, with the MAE



**Figure 6:** Predicted vs actual thermal conductivity plot from leave-one-out-cross-validation of GP model using compositional descriptors

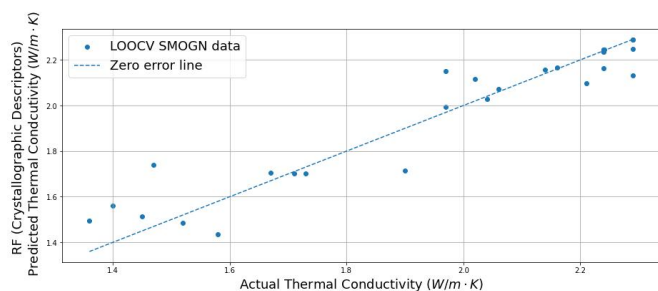


**Figure 7:** Predicted vs actual thermal conductivity plot from leave-one-out-cross-validation of RF model using compositional descriptors

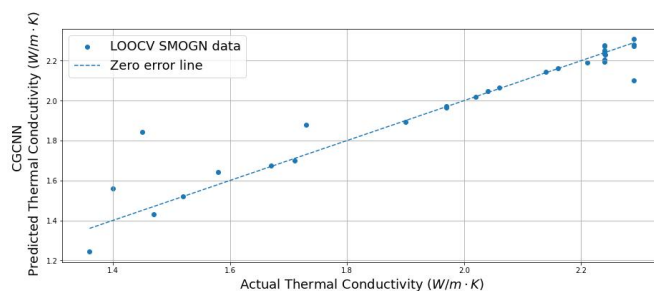


**Figure 8:** Predicted vs actual thermal conductivity plot from leave-one-out-cross-validation of RF model using crystallographic descriptors

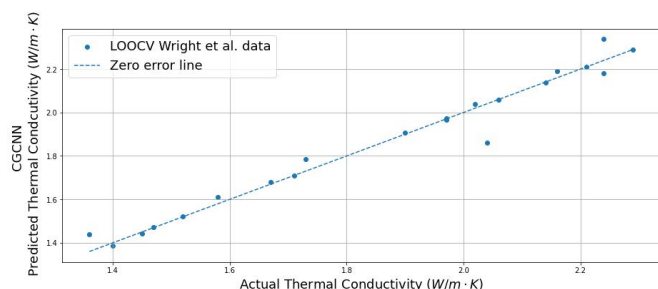
well below 0.1 and  $R^2$  exceeding 0.9. The augmented lattice-perturbation data offers a slight improvement in performance. The variation in prediction error for different lattice arrangements of the same composition is minimal, which suggests that the model can account for these slight variations in the lattice. The SMOBN data, when applied to the CGCNN algorithm, decreases the performance compared to model 7 (Figure 12), though it is comparable to model 5. The primary contributor to this decreased performance appears to be the composition  $(Sm_{1/4}Gd_{1/4}Eu_{1/4}Yb_{1/4})_2(Zr_{1/4}Hf_{1/4}Sn_{1/4}Ti_{1/4})_2O_7$ , which has an error of 27.1% compared to the average error of 2.8%. This data point is from the original data-set. It should be noted, that the compositions generated for the SMOBN data were not a perfect match to the crystallographic parameters, although



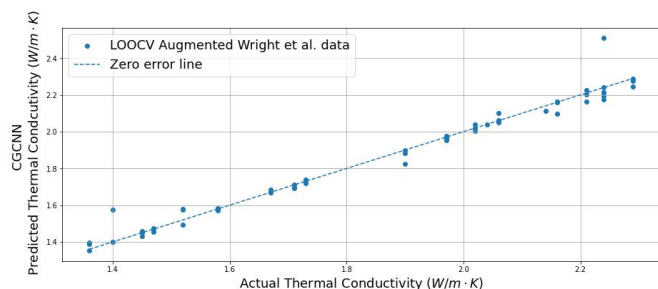
**Figure 9:** Predicted vs actual thermal conductivity plot from leave-one-out-cross-validation of RF model trained on SMOGN augmented data-set, using crystallographic descriptors



**Figure 12:** Predicted vs actual thermal conductivity plot from leave-one-out-cross-validation of CGCNN model trained on the SMOGN data-set



**Figure 10:** Predicted vs actual thermal conductivity plot from leave-one-out-cross-validation of CGCNN model



**Figure 11:** Predicted vs actual thermal conductivity plot from leave-one-out-cross-validation of CGCNN model trained on the augmented data-set

they were very close. This introduces an additional source of error in the training data. Furthermore, the TC of the generated data were roughly between 2.2-2.4 ( $W/mK$ ). For both the RF and CGCNN algorithms, the additional data appears to have improved the accuracy for this range of thermal conductivity. However, for the CGCNN algorithm, it has also significantly decreased performance in the lower range of thermal conductivity. Also, it should be noted that the TC of the generated data is not experimentally verified. Augmenting the data through lattice perturbation appears to be more effective, leading to a more consistent performance across the full range of TC. However, the latter method is still limited in that it is not able to provide data for sparse regions unlike SMOGN.

Table 6 shows the results of the Wilcoxon signed rank test, applied to compare the MAE difference between each model

trained on the original data-set. Values below 0.05 indicate a significant difference, while higher values indicate that the difference is of low significance. As stated before, the difference between the RF models using compositional and crystallographic descriptors is minimal. It is likely that a similar level of meaningful information can be gained from both sets of descriptors. The difference between CGCNN and the remaining models is, however, significant. The performance of the CGCNN models could be attributed to the fact that the graph-based representation of the data contains both compositional and crystallographic descriptors, together with the addition of spatial descriptors. Unlike the RF and GP models, the CGCNN models were also able to accurately predict the TC of the singular composition in the region of TC between 1.8 and 1.95  $W/mK$  ( $(Sm_{1/3}Gd_{1/3}Eu_{1/3})_2(Zr_{5/6}Hf_{5/6}Sn_{5/6}Ti_{3/4})_2O_7$ ). This holds promise in the ability of the CGCNN model to extrapolate for regions that are sparse in data.

## 4. Conclusions

This study compared three data representation methods for predicting the thermal conductivity (TC) of multi-component rare-earth pyrochlores; 1) chemical compositional descriptors, 2) simple crystallographic descriptors and 3) graph-based representations encoding the information from crystallographic information files (CIF). The first two were used to train Random Forest (RF) and Gaussian Process (GP) models, while the Crystal Graph Convolutional Neural Network (CGCNN) algorithm was applied to the graph-based representation. Leave-one-out cross-validation (LOOCV) showed CGCNN achieving the best performance with low MAE and high  $R^2$ , while GP had the worst performance with a negative  $R^2$ . The RF models trained on compositional and crystallographic descriptors performed similarly suggesting both methods contain comparable information.

To address potential over-fitting from the small dataset, two augmentation techniques were applied. The first, using the SMOGN algorithm, introduced random perturbations in the crystallographic descriptors and TC values. The RF model trained on method 2 showed an increase in performance values using the SMOGN generated data, but the CGCNN showed a reduction in performance.

The second augmentation method introduced small variations in atomic positions and site occupancies in the CIF files, without altering lattice parameters or the composition. GCNN models trained on this dataset showed slight performance improvements with minimal variation in TC predictions across copies. Both augmented and non-augmented CGCN models outperformed SMOGN augmented RF models, achieving  $R^2$  values above 0.97. Although this suggests possible over-fitting, both CGCNN models accurately predicted the TC of a composition from the sparsest region of the dataset with less than 1% error, indicating strong generalizability. In contrast RF and SMOGN-augmented RF models had errors of 10% and 16%, respectively.

Overall, graph-based representations offer greater predictive accuracy and generalisability than conventional descriptor-based approaches. Additionally, lattice perturbation shows promise as an augmentation strategy to enhance model performance in graph-based learning frameworks. It should be noted that training CGCNN on SMOGN data, required matching the new descriptors with estimated compositions. Furthermore, the TC values generated by SMOGN are not experimentally verified. Furthermore, the lattice parameters used to generate the CIF were calculated using Equation 1, which does introduce a degree of error.

## CRedit authorship contribution statement

**Amiya Chowdhury:** Conceptualisation, Methodology, Software, Formal Analysis, Data curation, Investigation, Writing - Original Draft Preparation. **Acacio Rincón Romero:** Conceptualisation, Methodology, Data curation, Supervision, Writing - Reviewing and Editing. **Eduardo Aguilar-Bejarano:** Methodology, Software. **Halar Memon:** Writing - Reviewing and Editing. **Grazziela Figueredo:** Conceptualisation, Methodology, Software, Supervision, Writing - Reviewing and Editing. **Tanvir Hussain:** Conceptualisation, Fund acquisition, Supervision, Writing - Reviewing and Editing.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) (grant number EP/V010093/1).

## Data availability

The data and code used for this study are available on a public repository on Github ([link here](#)).

## Bibliography

- , . Inorganic crystal structure database (icsd). <https://www.psds.ac.uk/icsd>. Accessed: 2024-04-22.
- Aidhy, D.S., 2024. Chemical randomness, lattice distortion and the wide distributions in the atomic level properties in high entropy alloys. *Computational Materials Science* 237, 112912. doi:10.1016/j.commatsci.2024.112912.
- Bakan, E., Vaßen, R., 2017. Ceramic top coats of plasma-sprayed thermal barrier coatings: Materials, processes, and properties. *Journal of Thermal Spray Technology* 26, 992–1010. doi:10.1007/s11666-017-0597-7.
- Cao, X., Vassen, R., Stoeber, D., 2004. Ceramic materials for thermal barrier coatings. *Journal of the European Ceramic Society* 24, 1–10. doi:10.1016/s0955-2219(03)00129-8.
- Carracedo-Reboredo, P., Liñares-Blanco, J., Rodríguez-Fernández, N., Cedrón, F., Novoa, F.J., Carballal, A., Maojo, V., Pazos, A., Fernandez-Lozano, C., 2021. A review on machine learning approaches and trends in drug discovery. *Computational and Structural Biotechnology Journal* 19, 4538–4558. doi:10.1016/j.csbj.2021.08.011.
- Centrich, X.T., Shehab, E., Sydor, P., Mackley, T., John, P., Harrison, A., 2014. An aerospace requirements setting model to improve system design. *Procedia CIRP* 22, 287–292. doi:10.1016/j.procir.2014.07.127.
- Choi, J.M., Lee, K., Kim, S., Moon, M., Jeong, W., Han, S., 2022. Accelerated computation of lattice thermal conductivity using neural network interatomic potentials. *Computational Materials Science* 211, 111472. doi:10.1016/j.commatsci.2022.111472.
- Darolia, R., 2013. Thermal barrier coatings technology: Critical review, progress update, remaining challenges and prospects. *International Materials Reviews* 58, 315–348. doi:10.1179/1743280413y.0000000019.
- Dippo, O.F., Vecchio, K.S., 2021. A universal configurational entropy metric for high-entropy materials. *Scripta Materialia* 201, 113974. doi:10.1016/j.scriptamat.2021.113974.
- Fergus, J.W., 2014. Zirconia and pyrochlore oxides for thermal barrier coatings in gas turbine engines. *Metallurgical and Materials Transactions E* 1, 118–131. doi:10.1007/s40553-014-0012-y.
- Fuentes, A.F., Montemayor, S.M., Maczka, M., Lang, M., Ewing, R.C., Amador, U., 2018. A critical review of existing criteria for the prediction of pyrochlore formation and stability. *Inorganic Chemistry* 57, 12093–12105. doi:10.1021/acs.inorgchem.8b01665.
- Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. *Journal of machine learning research* 3, 1157–1182.
- Jain, A., Ong, S.P., Hautier, G., Chen, W., Richards, W.D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G., et al., 2013. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Materials* 1. doi:10.1063/1.4812323.
- Jha, D., Ward, L., Paul, A., Liao, W.K., Choudhary, A., Wolverton, C., Agrawal, A., 2018. ElemNet: Deep learning the chemistry of materials from only elemental composition. *Sci. Rep.* 8, 17593.
- Jiang, D., Wu, Z., Hsieh, C.Y., Chen, G., Liao, B., Wang, Z., Shen, C., Cao, D., Wu, J., Hou, T., 2021. Could graph neural networks learn better molecular representation for drug discovery? a comparison study of descriptor-based and graph-based models. *Journal of Cheminformatics* 13. doi:10.1186/s13321-020-00479-8.
- Kohavi, R., John, G.H., 1997. Wrappers for feature subset selection. *Artif. Intell.* 97, 273–324.
- Kunz, N., 2020. SMOGN: Synthetic minority over-sampling technique for regression with gaussian noise. URL: <https://pypi.org/project/smogn/>.
- Leitner, J., Voňka, P., Sedmidubský, D., Svoboda, P., 2010. Application of neumann-kopp rule for the estimation of heat capacity of mixed oxides. *Thermochimica Acta* 497, 7–13. doi:10.1016/j.tca.2009.08.002.
- Liu, B., Wang, J., Zhou, Y., Liao, T., Li, F., 2007. Theoretical elastic stiffness, structure stability and thermal conductivity of  $la_2zr_2o_7$  pyrochlore. *Acta Materialia* 55, 2949–2957. doi:10.1016/j.actamat.2006.12.035.
- Liu, D., Shi, B., Geng, L., Wang, Y., Xu, B., Chen, Y., 2022. High-entropy rare-earth zirconate ceramics with low thermal conductivity for advanced thermal-barrier coatings. *Journal of Advanced Ceramics* 11, 961–973. doi:10.1007/s40145-022-0589-z.
- Liu, S., Zhao, L., Jiang, K., 2014. Thermophysical properties of lanthanum-gadolinium zirconate and gadolinia-stabilized zirconia nanocomposite



- from in situ reaction. *Advanced Engineering Materials* 17, 319–323. doi:10.1002/adem.201400164.
- Liu, Y., Zhao, T., Ju, W., Shi, S., 2017. Materials discovery and design using machine learning. *Journal of Materiomics* 3, 159–177. doi:10.1016/j.jmat.2017.08.002.
- Luo, Y., Li, M., Yuan, H., Liu, H., Fang, Y., 2023. Predicting lattice thermal conductivity via machine learning: A mini review. *npj Computational Materials* 9. doi:10.1038/s41524-023-00964-2.
- Mehboob, G., Liu, M.J., Xu, T., Hussain, S., Mehboob, G., Tahir, A., 2020. A review on failure mechanism of thermal barrier coatings and strategies to extend their lifetime. *Ceramics International* 46, 8497–8521. doi:10.1016/j.ceramint.2019.12.200.
- Miyazaki, H., Tamura, T., Mikami, M., Watanabe, K., Ide, N., Ozkendir, O.M., Nishino, Y., 2021. Machine learning based prediction of lattice thermal conductivity for half-Heusler compounds using atomic information. *Scientific Reports* 11. doi:10.1038/s41598-021-92030-4.
- Mouritz, A.P., 2012. 1.3.5 Superalloys. *American Institute of Aeronautics and Astronautics*.
- Mouta, R., Silva, R.X., Paschoal, C.W., 2013. Tolerance factor for pyrochlores and related structures. *Acta Crystallographica Section B Structural Science Crystal Engineering and Materials* 69, 439–445. doi:10.1107/s2052519213020514.
- Pakseresht, A., Sharifianjazi, F., Esmailkhanian, A., Bazli, L., Reisi Nafchi, M., Bazli, M., Kirubakaran, K., 2022. Failure mechanisms and structure tailoring of YSZ and new candidates for thermal barrier coatings: A systematic review. *Mater. Des.* 222, 111044.
- Pala, N., Abbas, A.N., Rockstuhl, C., Menzel, C., Mühlig, S., Lederer, F., Brown, J.J., Bright, V.M., Curley, S., Koh, Y.K., et al., 2012. Thermal conductivity and phonon transport. *Encyclopedia of Nanotechnology*, 2704–2711. doi:10.1007/978-90-481-9751-4\_277.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12, 2825–2830.
- Project, T.M., 2020. Materials data on eu2zr2o7 by materials project doi:10.17188/1284245.
- Ren, X., Wan, C., Zhao, M., Yang, J., Pan, W., 2015. Mechanical and thermal properties of fine-grained quasi-eutectoid  $(\text{La}_{(1-x)}\text{Yb}_x)_2\text{Zr}_2\text{O}_7$  ceramics. *Journal of the European Ceramic Society* 35, 3145–3154. doi:10.1016/j.jeurceramsoc.2015.04.024.
- Seko, A., Togo, A., Tanaka, I., 2018. Descriptors for machine learning of materials data. *Nanoinformatics*, 3–23. doi:10.1007/978-981-10-7617-6\_1.
- Stecura, S., 1978. Effects of compositional changes on the performance of a thermal barrier coating system. Technical Report NASA-TM-78976. NASA Lewis Research Center Cleveland, OH, United States.
- Suresh, G., Seenivasan, G., Krishnaiah, M., Srirama Murthi, P., 1997. Investigation of the thermal conductivity of selected compounds of gadolinium and lanthanum. *Journal of Nuclear Materials* 249, 259–261. doi:10.1016/s0022-3115(97)00235-3.
- Tsai, M.H., Yeh, J.W., 2014. High-entropy alloys: A critical review. *Materials Research Letters* 2, 107–123. doi:10.1080/21663831.2014.912690.
- Vassen, R., Cao, X., Tietz, F., Basu, D., Stöver, D., 2004. Zirconates as new materials for thermal barrier coatings. *Journal of the American Ceramic Society* 83, 2023–2028. doi:10.1111/j.1151-2916.2000.tb01506.x.
- Vaßen, R., Bakan, E., Mack, D.E., Guillon, O., 2022. A perspective on thermally sprayed thermal barrier coatings: Current status and trends. *Journal of Thermal Spray Technology* 31, 685–698. doi:10.1007/s11666-022-01330-2.
- Vaßen, R., Jarligo, M.O., Steinke, T., Mack, D.E., Stöver, D., 2010. Overview on advanced thermal barrier coatings. *Surface and Coatings Technology* 205, 938–942. doi:10.1016/j.surfcoat.2010.08.151.
- Wan, C.L., Pan, W., Xu, Q., Qin, Y.X., Wang, J.D., Qu, Z.X., Fang, M.H., 2006. Effect of point defects on the thermal transport properties of  $(\text{La}_x\text{Gd}_{(1-x)})_2\text{Zr}_2\text{O}_7$ : Experiment and theoretical model. *Physical Review B* 74. doi:10.1103/physrevb.74.144109.
- Ward, L., Agrawal, A., Choudhary, A., Wolverton, C., 2016. A general-purpose machine learning framework for predicting properties of inorganic materials. *Npj Comput. Mater.* 2.
- Wright, A.J., Wang, Q., Ko, S.T., Chung, K.M., Chen, R., Luo, J., 2019. Size disorder as a descriptor for predicting reduced thermal conductivity in medium- and high-entropy pyrochlore oxides. *Scripta Materialia*.
- Wright, A.J., Wang, Q., Ko, S.T., Chung, K.M., Chen, R., Luo, J., 2020. Size disorder as a descriptor for predicting reduced thermal conductivity in medium- and high-entropy pyrochlore oxides. *Scripta Materialia* 181, 76–81. doi:10.1016/j.scriptamat.2020.02.011.
- Wu, J., Wei, X., Padture, N.P., Klemens, P.G., Gell, M., García, E., Miranzo, P., Osendi, M.I., 2004. Low-thermal-conductivity rare-earth zirconates for potential thermal-barrier-coating applications. *Journal of the American Ceramic Society* 85, 3031–3035. doi:10.1111/j.1151-2916.2002.tb00574.x.
- Xie, T., Grossman, J.C., 2018. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* 120, 145301. URL: <https://link.aps.org/doi/10.1103/PhysRevLett.120.145301>, doi:10.1103/PhysRevLett.120.145301.
- Yang, L., Zhu, C., Sheng, Y., Nian, H., Li, Q., Song, P., Lu, W., Yang, J., Liu, B., 2018. Investigation of mechanical and thermal properties of rare earth pyrochlore oxides by first-principles calculations. *Journal of the American Ceramic Society* doi:10.1111/jace.16073.
- Zhang, Z., Chen, L., Zhong, F., Wang, D., Jiang, J., Zhang, S., Jiang, H., Zheng, M., Li, X., 2022. Graph neural network approaches for drug-target interactions. *Current Opinion in Structural Biology* 73, 102327. doi:10.1016/j.sbi.2021.102327.
- Zhu, J., Meng, X., Zhang, P., Li, Z., Xu, J., Reece, M.J., Gao, F., 2021a. Dual-phase rare-earth-zirconate high-entropy ceramics with glass-like thermal conductivity. *Journal of the European Ceramic Society* 41, 2861–2869. doi:10.1016/j.jeurceramsoc.2020.11.047.
- Zhu, T., He, R., Gong, S., Xie, T., Gorai, P., Nielsch, K., Grossman, J.C., 2021b. Charting lattice thermal conductivity for inorganic crystals and discovering rare earth chalcogenides for thermoelectrics. *Energy & Environmental Science* 14, 3559–3566. doi:10.1039/d1ee00442e.