Data-Centric Engineering (2020), 1: e8
doi:10.1017/dce.2020.7

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

Machine learning approaches to identify and design low thermal conductivity oxides for thermoelectric applications

Abhishek Tewari[1], Siddharth Dixit[2], Niteesh Sahni[2] and Stéphane P.A. Bordas[3,4]

[1]Department of Metallurgical and Materials Engineering, Indian Institute of Technology Roorkee, Hardiwar, India
[2]Department of Mathematics, Shiv Nadar University, Gautam Buddha Nagar, India
[3]Department of Engineering, Institute of Computational Engineering, University of Luxembourg, Esch-sur-Alzette, Luxembourg
[4]Institute of Mechanics and Advanced Materials, School of Engineering, Cardiff University, Cardiff, United Kingdom
*Corresponding author. E-mail: abhishek@mt.iitr.ac.in

(Received 23 April 2020; revised 11 July 2020; accepted 15 July 2020)

Keywords: Machine learning; oxides; rapid materials discovery; thermoelectric; thermal conductivity

Abstract

The search space for new thermoelectric oxides has been limited to the alloys of a few known systems, such as ZnO, $SrTiO_3$, and $CaMnO_3$. Notwithstanding the high power factor, their high thermal conductivity is a roadblock in achieving higher efficiency. In this paper, we apply machine learning (ML) models for discovering novel transition metal oxides with low lattice thermal conductivity (k_L). A two-step process is proposed to address the problem of small datasets frequently encountered in material informatics. First, a gradient-boosted tree classifier is learnt to categorize unknown compounds into three categories of k_L: low, medium, and high. In the second step, we fit regression models on the targeted class (i.e., low k_L) to estimate k_L with an $R^2 > 0.9$. Gradient boosted tree model was also used to identify key material properties influencing classification of k_L, namely lattice energy per atom, atom density, band gap, mass density, and ratio of oxygen by transition metal atoms. Only fundamental materials properties describing the crystal symmetry, compound chemistry, and interatomic bonding were used in the classification process, which can be readily used in the initial phases of materials design. The proposed two-step process addresses the problem of small datasets and improves the predictive accuracy. The ML approach adopted in the present work is generic in nature and can be combined with high-throughput computing for the rapid discovery of new materials for specific applications.

Impact Statement

Discovery of new materials is a complex and challenging task. Sequential nature of experimental route of investigating new materials makes it tedious and resource expensive. Application of data centric methods have shown a lot of promise in the recent past in the rapid discovery of new materials. Machine learning (ML) algorithms do not only predict the properties of interest, but also provide insight into the complex correlations between properties of materials. But the availability of large materials database is a challenge, which are usually required for these methods to attain high levels of predictive accuracy. In this work, a two-step ML process has been proposed to overcome the aforementioned challenge. The proposed method has been demonstrated using a dataset of transition metal oxides to predict their lattice thermal conductivity. Low thermal conductivity transition metal oxides are specially attractive for high temperature thermoelectric application because they exhibit excellent high temperature stability and have tunable electrical properties. The proposed

This article has earned badges for transparent research practices: Open Data and Open Materials. For details see the Data Availability Statement.

© The Author(s), 2020. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution, and reproduction in any medium, provided the original work is properly cited.

https://doi.org/10.1017/dce.2020.7 Published online by Cambridge University Press

method was able to provide most influencing fundamental materials properties, which can be readily used as design parameters in the early stages of materials selection. The method can be combined with high throughput computations to discover novel materials for specific applications.

1. Introduction

Thermoelectric effect has the potential to recover waste heat, which amounts to 50% of the global energy usage. Currently, the usage of technology is limited to niche applications (e.g., spacecraft) due to the low efficiency, rare elements, and toxic constituents (e.g., Bi, Te etc.) of thermoelectric (TE) materials. Oxide ceramics (e.g., ZnO, $CaMnO_3$, and $SrTiO_3$) have attracted a lot of attention for high-temperature applications, such as waste heat recovery in industrial power plants, automobiles, and so forth, due to their excellent high temperature stability, environment friendly constituents and cheaper mass production methods (Ovik et al., 2016; He et al., 2011; Koumoto et al., 2006). High electrical conductivity and Seebeck coefficient can be achieved in suitable oxide TE materials by transient element doping and band gap engineering (Ohta et al., 2008; Fergus, 2012). However, their inherent high thermal conductivity results into low thermoelectric efficiency ($ZT < 0.5$), which is defined as $ZT = \alpha^2 \sigma T/k$, where α, σ, k, and T are Seebeck coefficient, electrical conductivity, thermal conductivity, and temperature, respectively. For any practical applications, TE materials with $ZT > 1$ are required to make TE energy a commercially viable alternative for waste heat recovery.

On the experimental side, study of oxide TE materials has mostly focused on developing phonon glass–electron crystal structures (He et al., 2011), which allows decoupling of the electron and phonon transport properties. Enhancing the hierarchical scattering of phonons through nanostructuring mechanisms have been major focus in developing thermoelectric oxides. Incorporation of sintering additives has been commonly used strategy to introduce hierarchical phonon scattering (Wang et al., 2010; Buscaglia et al., 2014; Lan et al., 2012). Jood et al. (2011) reported a reduction upto 2 W/mK in the thermal conductivity of the Al-doped nanostructured ZnO. Azough et al. (2019) reported core-shell type of nanostructure formation within the grains in B doped $SrTiO_3$ ceramics leading to a low k_L value of 2.75 W/mK. Microstructural anisotropy introduced through Al-induced variations in oxygen stoichiometry can also enhance phonon scattering in preferential directions (Abutaha et al., 2013; Han et al., 2014). Hybrid superlattice type of structures have also been experimented for enhanced scattering of phonons at the interfaces in ZnO (Giri et al., 2016) and $SrTiO_3$ (Abutaha et al., 2015). Alvarez-Ruiz et al. (2018) reported unit cell twinning in Ga-doped ZnO, which start acting as phonon scattering centers. Introduction of structural defects using controlled synthesis methods can also help reduce k_L. Magnéli phases of TiO_2 have intrinsic, layered nanostructures defined by crystallographic shear planes, which act as scattering centers (Kieslich et al., 2016). Takemoto et al. (2014) reported the formation of a dense structure of three-dimensional (3D) stacking faults along the basal and pyramidal planes lowering k_L values upto 1.7 Wm/K in ZnO codoped with In and Ga. Zihua et al. (2018) introduced another level of nanostructuring by incorporating organic nanoparticles in the Co-doped ZnO.

In addition to experimental research, there are continuing efforts to bridge the gaps in our understanding of phonon scattering mechanisms using multiscale simulations to bring in the next generation of advances. Wu et al. (2016) applied first principle lattice dynamics to understand heat conduction mechanism in pure w-ZnO. Lower thermal conductivity of ZnO was attributed to smaller phonon group velocities, larger three-phonon scattering phase space, and larger anharmonicity in ZnO. It was also shown that ZnO possesses anisotropic thermal conductivity along the [1000] and [0001] directions, which has also been observed experimentally (Liang and Wang, 2020). Duda et al. (2012) conducted nonequilibrium molecular dynamics simulations to understand the effect of ordering of solid solutions. The results showed that ordering of solid solutions leads to change in the dominant scattering mechanism from impurity scattering to Umklapp three-phonon scattering. Wu et al. (2019) calculated the thermal conductivity of Silicon rich oxide layers inserted ZnO superlattice using the reverse nonequilibrium molecular dynamics method. Reduction in k_L was attributed to the phonon scattering at the ZnO/Si interface as well as the grain boundaries. Wang et al. (2017) studied the thermal conductivity of 2D ZnO monolayer and its

https://doi.org/10.1017/dce.2020.7 Published online by Cambridge University Press

anomalous temperature dependence using first principle density functional theory (DFT) simulations. Abnormally, slower fall in k_L with increasing T was found due to the significant contribution of optical phonon modes in overall thermal transport. Zhang and Koumoto (2013) showed that the thermal conductivity of $SrTiO_3$ superlattice decreases with decreasing grain size due to enhanced interface scattering.

High dimensionality of design space of thermoelectric materials makes the optimization of design parameters a nontrivial task. It is evident from the literature analysis that the class of materials explored for TE applications has been rather limited so far and our understanding of electronic and phonon transport of crystalline alloys is fairly limited (Minnich et al., 2009). On the other hand, rapid developments in the field of materials informatics has helped researchers explore new class of promising materials and establish correlations between design parameters and the thermoelectric properties (Wang et al., 2019). Wang et al. (2011) used high-throughput ab-initio calculations combined with regression analysis to show a positive correlation between power factor and the band gap and the charge carrier effective mass. Materials with large number of atoms per unit cell tend to have high power factor. Gaultois et al. (2013) conducted a data centric review of TE research literature creating a database of over 18,000 data points from over 100 publications. They used elaborate visualization techniques to extract the information of materials with promising thermoelectric properties along with their nature resource availability. They also designed a web-based recommendation engine based on random forest algorithm, which takes Seebeck coefficient, electrical conductivity, thermal conductivity, and band gap to evaluate the TE potential of a material (Oliynyk et al., 2016). High-throughput materials modeling combined with machine learning (ML) methods showed that large lattice parameter, band gap, and effective mass of holes are the key properties for high TE efficiency of nanograined half-heusler compounds (Carrete et al., 2014). Novel semiconductors with ultralow k_L values were proposed for further experimental studies (Carrete et al., 2014). McKinney et al. (2017) conducted high-throughput computational search for low Lorentz number materials for TE application. In addition to confirming existing TE materials, several new classes of materials were found, such as Zintl compounds and n-type ternary diamond-like semiconductors. Iwasaki et al. (2019) used supervised ML models to establish the key physical parameters controlling spin driven thermoelectric effect and proposed a novel material showing promising results. Oliynyk et al. (2016) found that electron count of B and difference in the atomic sizes of A and B are the most influential parameters in AB_2C type of compounds using random forest ML algorithm. Hou et al. (2019) used ML-based methods to optimize the Al/Si ratio in off-stochiometric $Al_{23.5+x}Fe_{36.5}Si_{40-x}$ compounds for achieving highest power factor. Miller et al. (2017) used high-throughput computations to screen 735 oxide materials for their thermoelectric properties and identified SnO as a potential n-type TE material. Measurements showed an extremely low k of 0.75 W/mK at moderate temperatures and ZT values of 0.22 in synthesized samples.

Application of ML algorithms on small datasets frequently encountered in materials science has been a key issue in materials informatics. Zhang and Ling (2018) proposed to include a crude estimate of the target property using low fidelity models as a way to improve the accuracy of ML models applied on small datasets. They achieved a high accuracy in predicting k_L by including empirical slack model values of k_L as a descriptor in the ML model. Singh and coworkers (Juneja et al., 2019; Juneja et al., 2020a; Juneja et al., 2020b) combined ML with high-throughput computing to build regression models for predicting the k_L of inorganic compounds. They also used maximum phonon frequency and integrated Gruneisen parameter as descriptors to build ML models for predicting k_L. Both the ML models to predict the k_L used complex derived properties as descriptors in their ML models, which restricts their utility in the initial phases of material selection and design. ML models based on characteristic materials properties are required to be used effectively in the discovery of new materials and reduce the time of design cycle.

In this work, we have applied a two-step ML-based process to first classify low k_L transition metal oxides and then predict their k_L values using regression methods. The proposed two-step process has been showed to be able to accurately predict the k_L values using a small dataset of transition metal oxides comprising 315 compounds. In this process, we were also able to define key fundamental material properties, which can be used for the screening of low k_L compounds in the initial stages of material design. The ML process has been described in detail in Section 2 and results are discussed in Section 3.

https://doi.org/10.1017/dce.2020.7 Published online by Cambridge University Press

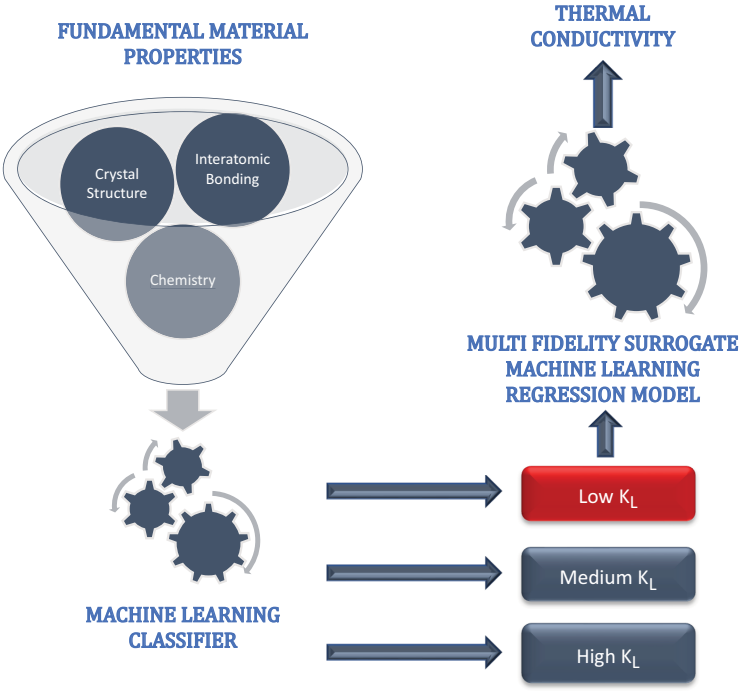

Figure 1. *Two-step machine learning process, where the first step filters low kL compounds using only fundamental material properties, such as details about crystal structure, interatomic bonding, and compound chemistry. In the second step, a multifidelity machine learning surrogate regression model is built to predict numerical kL values.*

2. Computational Methods

In this paper, the statistical methods employed are propelled mainly by the data gathered from Automatic-FLOW (AFLOW) for Materials Discovery database. The details of the database can be found elsewhere (Curtarolo et al., 2012). AFLOW uses the Gibbs implementation of quasiharmonic Debye–Gruneisen model to calculate the lattice thermal conductivity (k_L) of the compounds (Toher et al., 2017). First principle-based DFT calculations are performed for calculating the acoustic Debye temperature and Slater-Gamma method is used to calculate the Gruneisen parameter, which are then used in the calculation of k_L. Oxides and oxide alloys of transition metals, that is elements of groups 3–11 and periods 4–6 were considered in the present study as they have shown promise for high temperature thermoelectric applications (Ovik et al., 2016; Yin et al., 2017). The compounds considered in the present study had k_L ranging from 0.017 to 59.63 W/mK. The aim of the current study was to identify the most influencing fundamental properties affecting k_L of these compounds as well as build ML-based multifidelity surrogate models to predict the k_L of the transition metal oxides.

In order to do so, we implemented a two-step process (Figure 1): classification and regression. In the first step, we built a ML classifier to screen out unknown compounds having low k_L (<5). The same classifier model was also used to shortlist the most influencing fundamental material properties. The second step was to build a regression-based predictive model, which can be used to determine the numerical values of k_L of a compound. In the following, we describe both of these steps in detail.

2.1. Classification

The successful application of ML approaches on the modeling of material properties requires the selection of an appropriate set of modeling variables, namely the descriptors for the property of interest. In general, the descriptors are expected to be capable of both sufficiently distinguishing each of the modeled compounds/materials and determining the targeted property.

https://doi.org/10.1017/dce.2020.7 Published online by Cambridge University Press

In order to select the most influential features, we used following fundamental materials properties to build a classification model: mass density (ρ_m), ratio of oxygen to transition metal atom (O/M ratio), Bravias lattice type, atom density (ρ_n), electronic energy band gap (E_g), lattice energy per atom (e_L), point group order (O), and c/a ratio. Above mentioned parameters describe the crystal structure, compound chemistry, and interatomic bonding of an alloy. The idea was to use only fundamental crystal and materials properties so that they can be used as design parameters in the early stages of selection and shortlisting. Other ML models reported in the literature use complex derived properties, such as Gruneisen parameter, maximum phonon frequency, empirically estimated thermal conductivity (Juneja et al. (2019); Wang et al., 2019), which makes the utility of ML models rather limited. Classification model was built to segregate compounds into three categories of k_L viz. low ($k_L < 5$ W/mK), medium ($5 < k_L < 10$ W/mK), and high ($k_L > 10$ W/mK). In this way, the resulting model could potentially capture the underlying physical mechanisms after training, and thus offer reliable predictions for the chemistries beyond the training set. Around 30 different ML and deep learning models were built using the Caret library in R to solve this ternary classification task.

Best performance was achieved with Gradient Boosting Trees (details about the model training procedure are mentioned in Supplementary Appendix: Figure S5) in which the loss function to be optimized is in terms of trees grown on subsets of the predictor space. The algorithm XGBoost (Chen and Guestrin, 2016) achieves this task in a computationally efficient manner. This method has the potential to overfit data to any extent to give low prediction error rates. Xgboost algorithm has been widely used by data scientists in diverse problems involving classification and regression. Boosting is an ensemble technique where new tree-based models are added to correct the errors made by existing tree models. Tree-models are added sequentially until no further improvements can be made. Gradient boosting is an approach where new models are created that predict the residuals or errors of prior models and then added together to make the final prediction.

2.2. Regression

In this step, we used regression ML models to predict the absolute value of k_L of a compound. Since our dataset contained about 315 observations and 11 features (including the target variable), the dataset is relatively small. This may lead to higher variance in the least square estimates (Zhang and Ling, 2018). Regularization based regression techniques such as Lasso, Kernel-Ridge, Elastic net, and their modifications help us solve this problem by reducing the variance while managing negligible increase in bias. These models have been used extensively in the past by the computational materials community (Zhang and Ling, 2018; Hu et al., 2020) to build regression models with small sized datasets.

The Caret library in R (Kuhn et al., 2008) and the AutoML (H2O.ai, 2017) library from H_2O package were used to automate the process of building regression models, thus automating the end-to-end process of applying ML to real-world problems. AutoML tends to automate the maximum number of steps in an ML pipeline with a minimum amount of human effort without compromising the model's performance.

Due to the ease provided by these libraries, we were able to test the performance of a large number of predictive ML and deep learning models through cross validation on our dataset of thermoelectric oxides. Since we are interested in low k_L oxide alloys, we decided to build the regression model using the already classified data of only low k_L alloys, which contained 131 data points. The first regression model was built by considering the same descriptors, which were used to build the classification models in the first step. The best performing model (Random Forest) gave an R^2 value of 0.70. To improve the predictive accuracy of our model, we included two additional descriptors: Gruneisen parameter and Debye temperature, which have also been used earlier to build ML models for k_L prediction (Juneja et al., 2019). Best performance was achieved with cubist regression (details about the model training procedure are mentioned in Supplementary Appendix: Figure S6) which largely follows the model tree approach proposed by Quinlan et al. (1992). The basic idea behind the model tree approach is to use linear models instead of mere average of responses in the terminal leaves. This makes the method fit better than a Random Forest model in case the true responses are too large or too small. Even the splitting criteria

https://doi.org/10.1017/dce.2020.7 Published online by Cambridge University Press

chosen is differently as the expected reduction in the error of the node. Further, model trees deal with the problem of overfitting by incorporating a smoothing strategy devoled by Hastie and Pregibon (1990). The smoothing process adopted by Cubist is however more complex in comparison to model trees. For precise mathematical details the reader is refered to Kuhn and Johnson (2013).

3. Results and Discussions

A dataset of 315 compounds was obtained from AFLOW materials database. For the classification model, data for compounds was labeled based on the values of k_L as "low" (k_L <5 W/mK), "medium" ($0 < k_L < 10$ W/mK), and "high" ($k_L > 10$ W/mK). Equal instances of all the three classes were used to train the classifier to avoid class imbalance. The performance of top seven algorithms after training, testing, and hyperparameter tuning using 10-fold cross-validation is plotted in Figure 2. Using 10-fold cross validation while training the models allows us to divide the data into 10 parts out of which 9 are iteratively used for training and the last one for testing. Out of 30 different ML and deep learning models, both XGBoost and Random Forest models offer superior accuracy when compared to other counterparts such as Naïve Bayes, support vector machines (SVM), k Nearest Neighbors (kNN), linear discriminant analysis (LDA), and deep learning based classifiers. Higher value of cohens kappa coefficient also verifies their superiority. XGBoost was chosen for further analysis on the basis of higher mean class probability when compared to Random Forest. We notice that boosting (combining many weak learners to

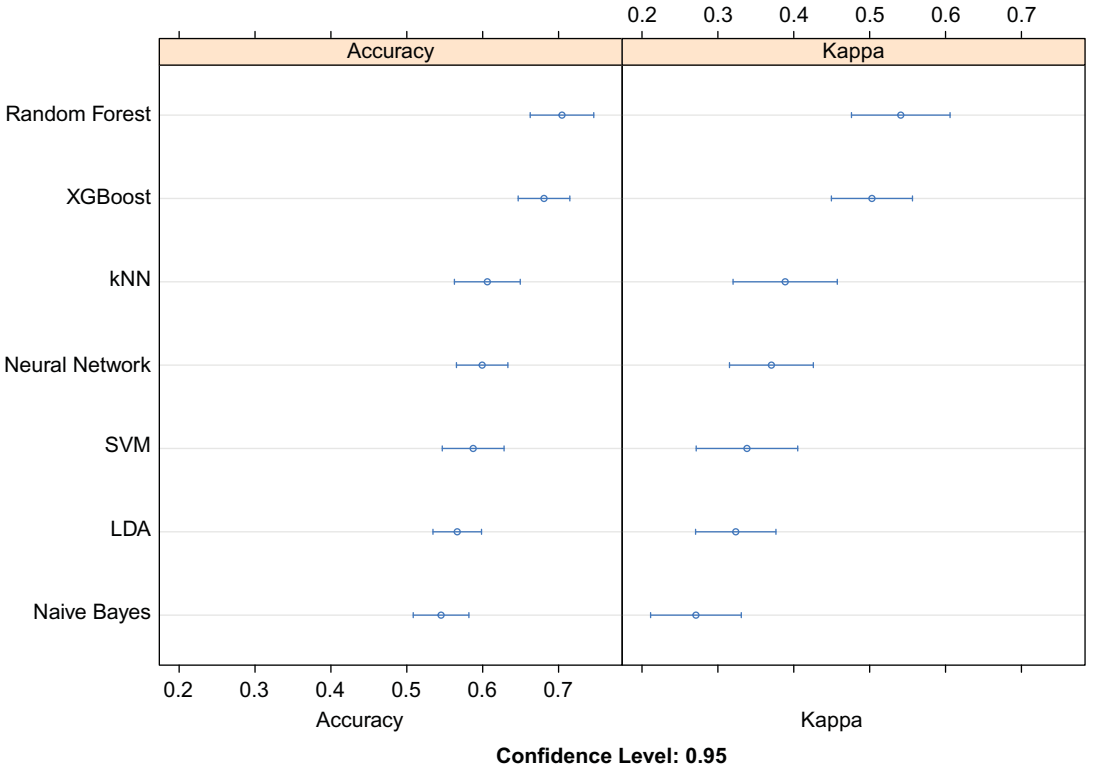

Figure 2. *Relative comparison of accuracy obtained using machine learning and deep learning classifiers. Here XGBoost and Random Forest surpass deep neural networks and other machine learning approaches to obtain the best classification accuracy. Cohen's kappa coefficient is also used to evaluate the different classification models amongst themselves. Abbreviations: kNN, k Nearest Neighbors; SVM, support vector machine with rbf kernel; LDA, linear discriminant analysis.*

https://doi.org/10.1017/dce.2020.7 Published online by Cambridge University Press

form a strong learner), plays a significant role in the success of tree-based classifiers for the limited data regime in materials sciences.

The XGBoost classifier correctly identified unknown compounds having Low k_L (sensitivity) with an accuracy of 81% and correctly identified unknown compounds not having Low k_L (specificity) with an accuracy of 82%. The balanced accuracy achieved while detecting compounds having high k_L was 84%. A dip in accuracy was observed while distinguishing low from medium k_L compounds and medium from high k_L compounds as the accuracy in such cases was 70%. The overall accuracy obtained while categorizing new compounds into the correct category was 72.13%. Table 1 represents the confusion matrix for model predictions on the validation set.

In order to assess the pros and cons of different classifiers, we went even further to calculate the probability of the class predicted by various methods for the given test observations. The average value of these probabilities are shown in Table 2. The class probabilities represent the confidence with which an unknown compound gets assigned to a particular class; therefore, higher values make the predictions trustworthy. Once again, XGBoost performs the best in this regard as well. We also computed the multi class area under the receiver operating characteristics (AUROC) values for the various methods according to the definition laid down by Hand and Till (2001). It tells how much model is capable of distinguishing between classes. Higher the AUROC, better the model is at distinguishing between classes. These numbers are summarized in Table 2. XGBoost clearly outscores the other methods in this regard.

The benefit of using XGBoost as the classifier is that after the boosted trees are constructed, it is relatively straightforward to retrieve importance scores for each attribute. Generally, importance provides a score that indicates how useful or valuable each feature was in the construction of the boosted decision trees within the model. The more an attribute is used to make key decisions with decision trees, the higher its relative importance. This importance is calculated explicitly for each attribute in the dataset, allowing attributes to be ranked and compared to each other.

Table 1. Confusion Matrix for predictions made on the validation set by the XGBoost classifier.

Precision	Low	Medium	High
Low	11	2	3
Medium	0	22	6
High	3	3	11

Table 2. A detailed comparison of different classifiers and their relative performance.

Model	Mean class probability	Sensitivity (low kL)	Specificity (low kL)	AUROC
XGBoost	0.85	0.81	0.82	0.96
Random Forests	0.78	0.84	0.87	0.98
Naive Bayes	0.73	0.63	0.85	0.84
kNN	0.67	0.85	0.73	0.83
Deep neural nets	0.63	0.59	0.74	0.74
SVM rbf kernel	0.54	0.78	0.82	0.84
LDA	0.52	0.63	0.70	0.75

Here mean class probability represents the average confidence with which the classifier assigns a particular compound to the predicted class. The corresponding sensitivities and specificities have also been mentioned.Abbreviation: AUROC, area under the receiver operating characteristics; kNN, k Nearest Neighbors; LDA, linear discriminant analysis; SVM, support vector machine.

https://doi.org/10.1017/dce.2020.7 Published online by Cambridge University Press

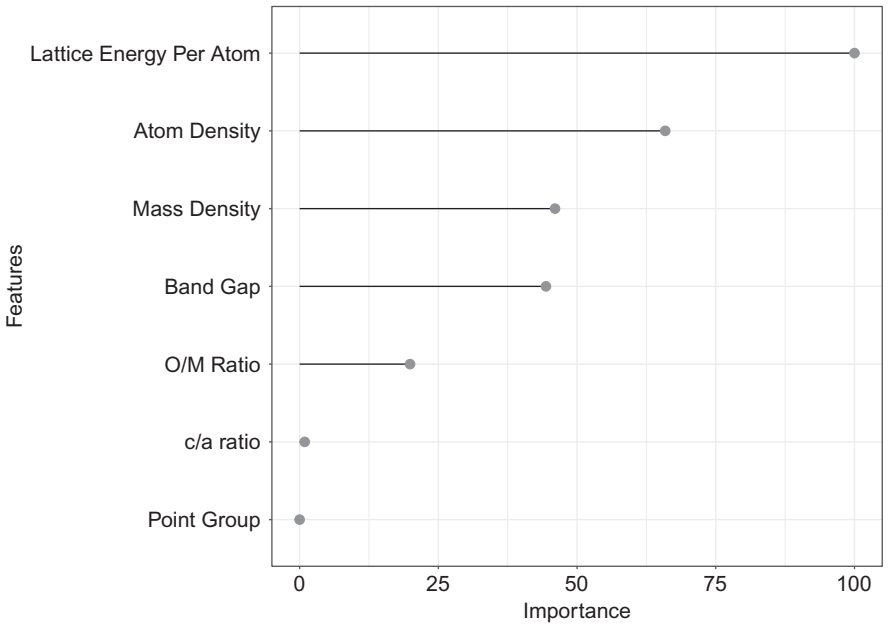

Figure 3. *Feature importance plot generated by the XGBoost Classifier. The relative importance of descriptors is calculated by how useful it was while making key decisions with Decision Trees.*

Figure 3 shows the relative importance of features output by the XGBoost classifier. Lattice energy per atom, atom density, electronic energy band gap, mass density, and O/M ratio were the key descriptors identified as important by our classifier. The idea behind performing the classification step was to shortlist the promising candidates based on their crystal structure, compound chemistry and interatomic bonding in the initial process of material design without having to calculate any complex derived properties. Some of the descriptors identified by the classification algorithm are intuitive and known from the physics of heat conduction. For example, Lattice energy per atom is a measure of interatomic bonding in a material. The lower the e_L, the higher the strength of interatomic bonding and higher the k_L. Similarly, classification model also predicts the importance of atom density and mass density, which have been reported earlier as influencing parameters in lattice thermal conductivity (Juneja et al., 2019). Electronic energy band gap and O/M ratio are two parameters, which are not directly linked to the physics of thermal conductivity, but are recognized as important classifying parameters in our study. The role of these parameters needs to be investigated further using atomistic modeling methods.

The second part of our formulated process, involved constructing regression-based predictive models. First, the regression models were built on the entire dataset, which was used for the classification. Random Forest models showed the highest accuracy with an cv – $R^2 = 0.44$ and cv – MAE = 3.2 on this dataset. The same model was then applied on already classified dataset of "low" k_L values with the cv – $R^2 = 0.70$ and cv – MAE = 0.71. The performance of the random forest model on two different datasets is plotted in Figure 4 a,b. An improvement in the accuracy might be attributed to the narrower spread of the classified data. Classification helps reduce the variance in the dataset according to the range of the material property of interest, which is used in the regression step. It helps in achieving greater predictive accuracy even with small dataset in the regression step. To further improve the predictive accuracy, Debye temperature and Gruneisen parameter were added as descriptors in the list of descriptors used in the classification step. Best performance was achieved by cubist model giving the cv – $R^2 = 0.96$ and a cv – MAE = 0.19. Gaussian process regression with polynomial kernel and Kernel ridge regression also performed well giving an accuracy of cv – $R^2 = 0.95$, cv – MAE = 0.26 and cv – $R^2 = 0.94$, cv – MAE = 0.23, respectively. Details about the performance of other models is given in Table 3.

https://doi.org/10.1017/dce.2020.7 Published online by Cambridge University Press

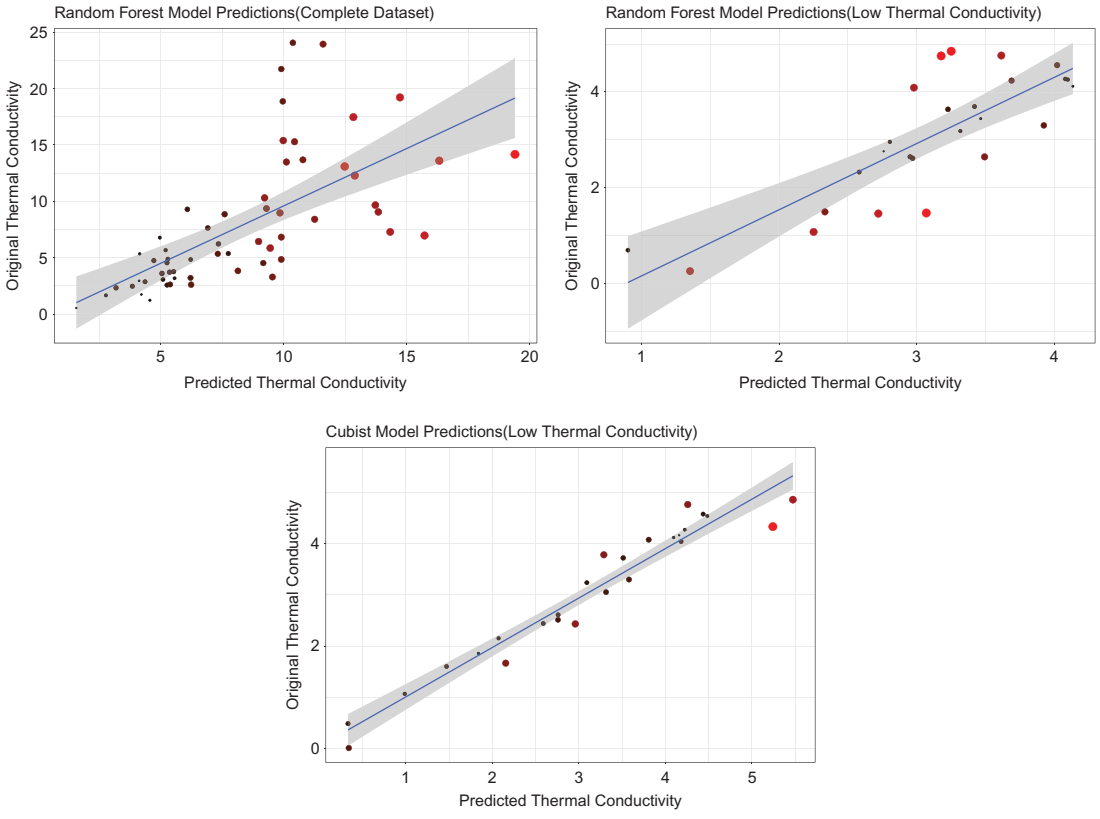

Figure 4. *Predictions of the regression models where lighter shades of red and bigger point sizes represent higher residuals, (a) random forest fitted on entire dataset, (b) random forest fitted on dataset of low kL compounds, and (c) Cubist model fitted on low kL compounds including Debye temperature and Gruneisen parameter.*

Table 3. Represents the 10-fold cross validation results obtained from the regression model including Gruneisen parameter and Debye temperature.

Model	cv-RMSE	cv-R2	cv-MAE
Cubist	0.27	0.96	0.19
GPR poly kernel	0.34	0.95	0.26
kNN	0.67	0.77	0.59
GBM	0.44	0.91	0.36
XGBoost	0.72	0.72	0.56
Random Forests	0.48	0.87	0.38
Kernel ridge	0.31	0.94	0.23
Deep neural nets	2.44	N.A.	2.14

Cubist model achieves the best predictive power. The bad performance of Neural Networks is justified by the lack of training data.Abbreviations: kNN, k Nearest Neighbors; RMSE, root mean square error; GBM, Gradient Boosted Machines; GPR, Gaussian Process Regression; MAE, mean absolute error.

https://doi.org/10.1017/dce.2020.7 Published online by Cambridge University Press

It is to be noted that the classifier can be used independently of the regression step in cases when the values of Gruneisen parameter and Debye temperature are not available for the compound. This is useful when we have only limited information about the compound, that is only values for the fundamental material properties mentioned earlier. Another important point is that the feature importance identified by the classifier is only applicable when the classifier is used independently of the regression step. We claim this because the regression step uses a different model, which means that the same features might not turn out to be important for the regression step. Therefore, the regression step uses all descriptors used in the classification step in addition to using Gruneisen parameter and Debye temperature as descriptors.

Further, it was also observed that ML approaches in general work better than deep learning when applied to datasets related to material science. This can be attributed to the fact that neural networks need high amounts of data to approximate the underlying function (Hornik et al., 1989) representing the k_L which is rarely available in cases of material science problems. Therefore, when a deep neural network with multiple permutations of the hidden layer (neurons) is built, it fails to converge to the optimal underlying function giving a large MAE of 2.14.

4. Conclusions

In this paper, a ML-based two-step process of discovering novel materials have been proposed. In the first step, classification is performed on the entire dataset to categorise the data, which is followed by fitting regression models to predict the numerical value of the property of interest. The proposed two-step process addresses the problem of small datasets in materials informatics by reducing the variance of the dataset using classification models according to the range of property of interest, which helps in achieving greater predictive accuracy in the regression step. The approach was applied on a dataset of transition metal oxides to classify and predict the k_L values of low k_L transition metal oxides. A high predictive accuracy of 95% was achieved using multiple ML-based regression algorithms, such as cubist model, kernel ridge and gaussian process. It was also shown that ML-based approach worked better in comparison to deep learning methods for problems involving small datasets. In addition, gradient boosted tree algorithm was able to identify key material properties namely: Lattice energy per atom, atom density, electronic energy band gap, mass density, and ratio of oxygen by transition metal atoms. Since the key descriptors can be derived from fundamental crystal structure, compound chemistry, and interatomic bonding, they can be easily utilized for the classification of compound in the early stages of materials selection, without needing to calculate computationally expensive derived complex properties. The two-step process proposed in the current work addresses a critical challenge in the materials informatics, which is the smaller sizes of datasets. The approach can be combined with high-throughput computing to discover novel materials for specific applications at lower computational cost. The work will be carried forward in that direction by demonstrating the proposed methodology on transition metal oxides to discover novel low k_L oxides.

Funding Statement. This research was supported by grants from the Science and Engineering Research Board, India through project number SRG/2019/000644. In addition, S.P.A. Bordas received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 811099 TWINNING Project DRIVEN for the University of Luxembourg: \url{https://2020driven.uni.lu/}

Competing Interests. The authors declare no competing interests exist.

Authorship Contributions. Conceptualization, A.T.; Methodology, A.T., S.D., N.S., and S.B.; Data curation, S.D; Data visualisation, S.D.; Writing-original draft: A.T. and S.D. All authors approved the final submitted draft.

Data Availability Statement. Replication data and code can be found on the github repository for this project: \url{https://github.com/Sid-darthvader/Machine-Learning-for-Thermoelectrics-Discovery}.

Ethical Standards. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

Supplementary Materials. To view supplementary material for this article, please visit http://dx.doi.org/10.1017/dce.2020.7.

https://doi.org/10.1017/dce.2020.7 Published online by Cambridge University Press

References

Abutaha AI, **Kumar SS**, **Li K**, **Dehkordi AM**, **Tritt TM and Alshareef HN** (2015) Enhanced thermoelectric figure-of-merit in thermally robust, nanostructured superlattices based on srtio3. *Chemistry of Materials 27*(6), 2165–2171.

Abutaha AI, **Sarath Kumar S and Alshareef HN** (2013) Crystal orientation dependent thermoelectric properties of highly oriented aluminum-doped zinc oxide thin films. *Applied Physics Letters 102*(5), 053507.

Alvarez-Ruiz DT, **Azough F**, **Hernandez-Maldonado D**, **Kepaptsoglou DM**, **Ramasse QM**, **Day SJ**, **Svec P**, **Svec Sr P and Freer R** (2018) Utilising unit-cell twinning operators to reduce lattice thermal conductivity in modular structures: Structure and thermoelectric properties of ga2o3 (zno) 9. *Journal of Alloys and Compounds 762*, 892–900.

Azough F, **Gholinia A**, **Alvarez-Ruiz DT**, **Duran E**, **Kepaptsoglou DM**, **Eggeman AS**, **Ramasse QM and Freer R** (2019) Self-nanostructuring in srtio3: A novel strategy for enhancement of thermoelectric response in oxides. *ACS Applied Materials & Interfaces 11*(36), 32833–32843.

Buscaglia MT, **Maglia F**, **Anselmi-Tamburini U**, **Marré D**, **Pallecchi I**, **Ianculescu A**, **Canu G**, **Viviani M**, **Fabrizio M and Buscaglia V** (2014) Effect of nanostructure on the thermal conductivity of la-doped srtio3 ceramics. *Journal of the European Ceramic Society 34*(2), 307–316.

Carrete J, **Li W**, **Mingo N**, **Wang S and Curtarolo S** (2014) Finding unprecedentedly low-thermal-conductivity half-heusler semiconductors via high-throughput materials modeling. *Physical Review X 4*(1), 011019.

Chen T and Guestrin C (2016) Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794.

Curtarolo S, **Setyawan W**, **Hart GL**, **Jahnatek M**, **Chepulskii RV**, **Taylor RH**, **Wang S**, **Xue J**, **Yang K**, **Levy O**, **Mehl MJ**, **Stokes HT**, **Demchenko DO and Morgan D** (2012) Aflow: An automatic framework for high-throughput materials discovery. *Computational Materials Science 58*, 218–226.

Duda JC, **English TS**, **Jordan DA**, **Norris PM and Soffa WA** (2012) Controlling thermal conductivity of alloys via atomic ordering. *Journal of Heat Transfer 134*(1), 014501-014501-4.

Fergus JW (2012) Oxide materials for high temperature thermoelectric energy conversion. *Journal of the European Ceramic Society 32*(3), 525–540.

Gaultois MW, **Sparks TD**, **Borg CK**, **Seshadri R**, **Bonificio WD and Clarke DR** (2013) Data-driven review of thermoelectric materials: Performance and resource considerations. *Chemistry of Materials 25*(15), 2911–2920.

Giri A, **Niemelä J-P**, **Tynell T**, **Gaskins JT**, **Donovan BF**, **Karppinen M and Hopkins PE** (2016) Heat-transport mechanisms in molecular building blocks of inorganic/organic hybrid superlattices. *Physical Review B 93*(11), 115310.

H2O.ai (2017) *H2O AutoML*. H2O version 3.30.0.1.

Han L, **Van Nong N**, **Zhang W**, **Holgate T**, **Tashiro K**, **Ohtaki M**, **Pryds N and Linderoth S** (2014) Effects of morphology on the thermoelectric properties of al-doped zno. *RSC Advances 4*(24), 12353–12361.

Hand DJ and Till RJ (2001) A simple generalisation of the area under the roc curve for multiple class classification problems. *Machine learning 45*(2), 171–186.

Hastie T and Pregibon D (1990) *Shrinking Trees*. AT & T Bell Laboratories. Stanford-Web.

He J, **Liu Y and Funahashi R** (2011) Oxide thermoelectrics: The challenges, progress, and outlook. *Journal of Materials Research 26*(15), 1762–1772.

Hornik K, **Stinchcombe M and White H** (1989) Multilayer feedforward networks are universal approximators. *Neural Networks 2*(5), 359–366.

Hou Z, **Takagiwa Y**, **Shinohara Y**, **Xu Y and Tsuda K** (2019) Machine-learning-assisted development and theoretical consideration for the al2fe3si3 thermoelectric material. *ACS Applied Materials & Interfaces 11*(12), 11545–11554.

Hu Y-J, **Zhao G**, **Zhang M**, **Bin B**, **Del Rose T**, **Zhao Q**, **Zu Q**, **Chen Y**, **Sun X**, **de Jong M and Qi L** (2020) Predicting densities and elastic moduli of sio 2-based glasses by machine learning. *NPJ Computational Materials 6*(1), 1–13.

Iwasaki Y, **Takeuchi I**, **Stanev V**, **Kusne AG**, **Ishida M**, **Kirihara A**, **Ihara K**, **Sawada R**, **Terashima K**, **Someya H**, **Uchida K**, **Saitoh E and Yorozu S** (2019) Machine-learning guided discovery of a new thermoelectric material. *Scientific Reports 9*(1), 1–7.

Jood P, **Mehta RJ**, **Zhang Y**, **Peleckis G**, **Wang X**, **Siegel RW**, **Borca-Tasciuc T**, **Dou SX and Ramanath G** (2011) Al-doped zinc oxide nanocomposites with enhanced thermoelectric properties. *Nano Letters 11*(10), 4337–4342.

Juneja R and Singh AK (2020a) Unraveling the role of bonding chemistry in connecting electronic and thermal transport by machine learning. *Journal of Materials Chemistry A 8*(17), 8716–8721.

Juneja R and Singh AK (2020b) Guided patchwork kriging to develop highly transferable thermal conductivity prediction models *Journal of Physics: Materials 3*(2), 024006.

Juneja R, **Yumnam G**, **Satsangi S and Singh AK** (2019) Coupling the high-throughput property map to machine learning for predicting lattice thermal conductivity. *Chemistry of Materials 31*(14), 5145–5151.

Kieslich G, **Cerretti G**, **Veremchuk I**, **Hermann RP**, **Panthöfer M**, **Grin J and Tremel W** (2016) A chemists view: Metal oxides with adaptive structures for thermoelectric applications. *Physica Status Solidi (A) 213*(3), 808–823.

Koumoto K, **Terasaki I and Funahashi R** (2006) Complex oxide materials for potential thermoelectric applications *MRS Bulletin 31*(3), 206–210.

Kuhn M and Johnson K (2013) *Applied Predictive Modeling*, Vol. *26*., New York, Springer.

Kuhn M (2008) Building predictive models in r using the caret package. *Journal of Statistical Software 28*(5), 1–26.

https://doi.org/10.1017/dce.2020.7 Published online by Cambridge University Press

Lan J, **Lin Y-H**, **Liu Y**, **Xu S and Nan C-W** (2012) High thermoelectric performance of nanostructured in 2 o 3-based ceramics. *Journal of the American Ceramic Society*, *95*(8):2465–2469.

Liang X and Wang C (2020) Electron and phonon transport anisotropy of zno at and above room temperature. *Applied Physics Letters 116*(4), 043903.

McKinney RW, **Gorai P**, **Stevanović V and Toberer ES** (2017) Search for new thermoelectric materials with low lorenz number. *Journal of Materials Chemistry A 5*(33), 17302–17311.

Miller SA, **Gorai P**, **Aydemir U**, **Mason TO**, **Stevanović V**, **Toberer ES and Snyder GJ** (2017) Sno as a potential oxide thermoelectric candidate. *Journal of Materials Chemistry C 5*(34), 8854–8861.

Minnich A, **Dresselhaus MS**, **Ren Z and Chen G** (2009) Bulk nanostructured thermoelectric materials: current research and future prospects. *Energy & Environmental Science*, *2*(5), 466–479.

Ohta H, **Sugiura K and Koumoto K** (2008) Recent progress in oxide thermoelectric materials: p-type ca3co4o9 and n-type srtio3. *Inorganic Chemistry 47*(19), 8429–8436.

Oliynyk AO, **Antono E**, **Sparks TD**, **Ghadbeigi L**, **Gaultois MW**, **Meredig B and Mar A** (2016) High-throughput machine-learning-driven synthesis of full-heusler compounds. *Chemistry of Materials 28*(20), 7324–7331.

Ovik R, **Long B**, **Barma M**, **Riaz M**, **Sabri M**, **Said S and Saidur R** (2016) A review on nanostructures of high-temperature thermoelectric materials for waste heat recovery. *Renewable and Sustainable Energy Reviews 64*, 635–659.

Quinlan JR (1992) Learning with continuous classes. In *5th Australian Joint Conference on Artificial Intelligence*, Vol. *92*, Singapore, World Scientific, pp. 343–348.

Takemoto H, **Fugane K**, **Yan P**, **Drennan J**, **Saito M**, **Mori T and Yamamura H** (2014) Reduction of thermal conductivity in dually doped zno by design of three-dimensional stacking faults. *RSC Advances 4*(6), 2661–2672.

Toher C, **Oses C**, **Plata JJ**, **Hicks D**, **Rose F**, **Levy O**, **de Jong M**, **Asta M**, **Fornari M**, **Nardelli MB and Curtarolo S** (2017) Combining the aflow gibbs and elastic libraries to efficiently and robustly screen thermomechanical properties of solids. *Physical Review Materials 1*(1), 015401.

Wang H, **Qin G**, **Li G**, **Wang Q and Hu M** (2017) Low thermal conductivity of monolayer zno and its anomalous temperature dependence. *Physical Chemistry Chemical Physics 19*(20), 12882–12889.

Wang N, **He H**, **Ba Y**, **Wan C and Koumoto K** (2010) Thermoelectric properties of nb-doped srtio3 ceramics enhanced by potassium titanate nanowires addition. *Journal of the Ceramic Society of Japan 118*(1383), 1098–1101.

Wang S, **Wang Z**, **Setyawan W**, **Mingo N and Curtarolo S** (2011) Assessing the thermoelectric properties of sintered compounds via high-throughput ab-initio calculations. *Physical Review X 1*(2), 021012.

Wang T, **Zhang C**, **Snoussi H and Zhang G** (2019) Machine learning approaches for thermoelectric materials research. *Advanced Functional Materials 30*(5), 1906041-1906041-14.

Wu H-T, **Su Y-C**, **Pao C-W and Shih C-F** (2019) Zno/silicon-rich oxide superlattices with high thermoelectric figure of merit: A comprehensive study by experiment and molecular dynamic simulation. *ACS Applied Materials & Interfaces 11*(14), 13507–13513.

Wu X, **Lee J**, **Varshney V**, **Wohlwend JL**, **Roy AK and Luo T** (2016) Thermal conductivity of wurtzite zinc-oxide from first-principles lattice dynamics—A comparative study with gallium nitride. *Scientific Reports 6*, 22504.

Yin Y, **Tudu B and Tiwari A** (2017) Recent advances in oxide thermoelectric materials and modules. *Vacuum 146*, 356–374.

Zhang R-Z and Koumoto K (2013) Grain-size-dependent thermoelectric properties of srtio 3 3d superlattice ceramics. *Journal of Electronic Materials 42*(7), 1568–1572.

Zhang Y and Ling C (2018) A strategy to apply machine learning to small datasets in materials science. *NPJ Computational Materials 4*(1), 1–8.

Zihua W, **Huaqing X**, **Yuanyuan W**, **Jiaojiao X and Jianhui M** (2018) Thermoelectric properties of co-doped zno by incorporating organic nanoparticles. *Rare Metal Materials and Engineering 47*(2), 452–456.

Cite this article: Tewari, A. Dixit, S. Sahni, N. and Bordas, S. P. A. 2020. Machine learning approaches to identify and design low thermal conductivity oxides for thermoelectric applications. *Data-Centric Engineering*, 1: e8. doi:https://doi.org/10.1017/dce.2020.7

https://doi.org/10.1017/dce.2020.7 Published online by Cambridge University Press