

Faculty of Material Science and Ceramics
DEPARTMENT OF INORGANIC CHEMISTRY



MASTER OF SCIENCE THESIS

KACPER PELC

**SEARCHING FOR LOW THERMAL CONDUCTIVITY
MATERIALS BY MACHINE LEARNING ALGORITHMS**

SUPERVISORS:

Prof. Dr. Krzysztof Wojciechowski

AGH University of Science and Technology

Krakow 2023

Contents

1. Abstract.....	4
2. Introduction	5
3. Thermal Conductivity.....	6
3.1. Thermal conductivity metals.....	8
3.2. Thermal conductivity of other crystal structures	9
3.3. Density functional theory (DFT).....	12
3.4. Molecular Dynamic (MD).....	13
3.5. Green–Kubo method	14
3.6. Slack model Automatic Gibbs Library (AGL).....	15
3.7. Applications of materials with low thermal conductivity	16
3.7.1. Thermoelectric materials	16
3.7.2. Thermal coating barriers.....	17
4. Machine Learning.....	18
4.1. Data Validation	18
4.2. Data Preprocessing.....	19
4.3. Dimensionality reduction	19
4.4. Unsupervised Correlation Method	20
4.5. Scaling.....	21
4.6. Metrics for Machine Learning	22
4.6.1. Mean Squared Error	22
4.6.2. Mean Absolute Error	22
4.6.3. Coefficient of determination (R^2).....	23
5. Algorithms.....	25
5.1. Least Absolute Shrinkage and Selection Operator LASSO.....	25
5.2. Support Vector Regression (SVR).....	26

5.3.	Kernel Ridge Regression (KRR).....	27
5.4.	eXtreme Gradient Boosting (XGBoost, XGB)	28
6.	Review of literature about Predicting lattice thermal conductivity via machine learning	31
6.1.	Datasets	31
6.2.	Descriptors	32
6.3.	Algorithms.....	35
6.4.	Conclusion.....	36
7.	Learning Procedure	37
7.1.	Acquisition of data and preprocessing	37
7.2.	Descriptors	37
7.3.	Multifactor dimensionality reduction.....	40
7.4.	Discussion of results	52
7.5.	Laboratory data	57
8.	Conclusions	Error! Bookmark not defined.
9.	Supplementary Files	65
10.	Bibliography	66

1. Abstract

This thesis develops descriptors and a machine learning model to predict lattice thermal conductivity (κ_L). The dataset comes from the AFLOW repository and contains 5664 samples. Following the calculations, the conclusion was reached that the best algorithm from KRR, LASSO, SVR and XGBoost is the last one. For the test set, the correlation coefficient R^2 was 0.87 and the mean absolute error (MAE) was 2.50 W/(m·K) and 73% of samples were within the bounds of this error. In order to find out which features were the most important to the model, a multidimensionality reduction was performed by creating a correlation matrix and discarding the features that were strongly correlated with the other features. For the best model, the biggest collapse in the quality of results occurred when the last five features began to be removed. After analysis, these features were found to be the most important: thermal conductivities of the elements, cell volume, group, and space group of crystal structure. An attempt was then made to predict the thermal conductivity of the materials tested in the laboratory, but the results turned out to be quite different from the actual measurements. The flowchart of the study can be found in Figure 1.

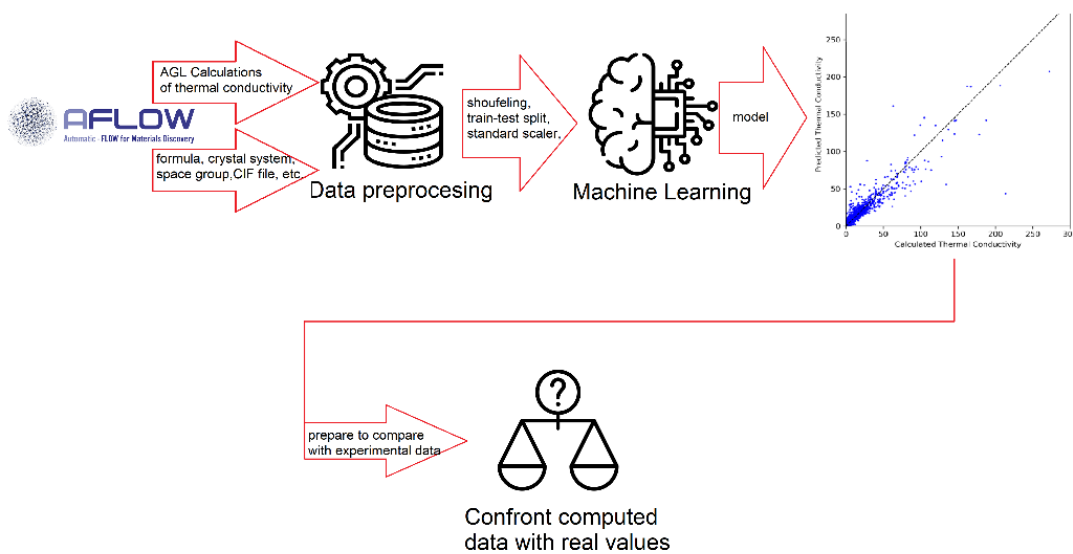


Figure 1 Flowchart of the machine learning process [Own source, using icons from flaticon.com]

2. Introduction

Thermal conductivity is one of the crucial features of materials in several technological applications. For example, a low lattice thermal conductivity (LTC) κ_L value translates to a high-performance storage heat converter based on thermoelectric materials or thermal barrier coatings. On the other hand, materials with high thermal conductivity could be used in heat sinks or optoelectronic devices where heat dissipation is one of the crucial parameters. Screening crystalline materials for those applications with classical methods of predicting thermal conductivity like the first-principles calculations, performed using Density Functional Theory (DFT), conductivity could be computationally expensive. In recent years machine learning (ML) techniques have gained significant relevance in the field of materials discovery and design. These techniques have enabled the avoidance of the complex and expensive traditional loops used for creating new materials for various applications. This is largely due to the advancements in ML algorithms and the availability of powerful computing facilities.

3. Thermal Conductivity

Thermal conductivity is the ability of a material to conduct heat. It is the measure of how quickly heat would flow through a substance when a temperature difference is applied across it. The higher the material's thermal conductivity, the easier it can transfer heat.

The basic law of heat conduction is Fourier's law, which states that the density of heat flux Q is proportional to the temperature gradient T in an isotropic body $Q = -\lambda \text{grad}T$ [1]. The proportional constant λ (Eq. 1) is the thermal conductivity. The minus sign indicates that the temperature drops in the direction of the heat transport; therefore, the temperature gradient is a negative quantity. The Figure 2 shows a simplified diagram of heat conduction. Thermal conductivity unit is $W/(m \cdot K)$.

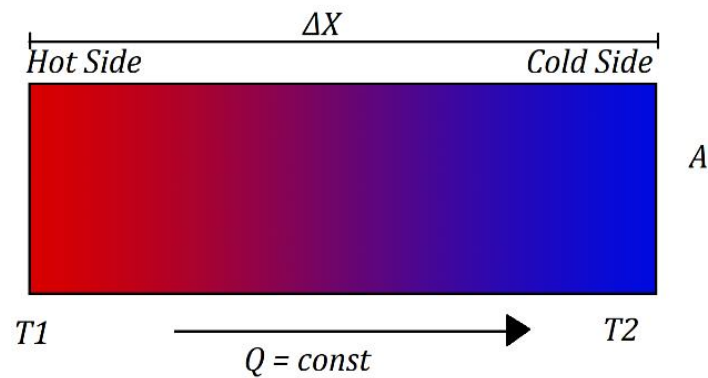


Figure 2 Schematic view of heat current, through a sample that has a temperature difference between its two ends [the hot (red) region with $T1$ and the cold (blue) region with $T2$]. The linear gradient of colour from red to blue indicates the temperature gradient.

$$\lambda = -\frac{Qx}{tA\Delta T} \quad (\text{Eq. 1})$$

Where:

λ – thermal conductivity coefficient

Q – heat flow

t – flow time

A – area through which the heat flows

x – element length

λT – temperature difference

Heat energy can be transmitted through solids via various carriers such as electrons, lattice waves, electromagnetic waves, spin waves, and other excitation sources (Eq. 2) [2].

$$\kappa = \sum_{\alpha} \kappa_{\alpha} \quad (\text{Eq. 2})$$

Where:

α – excitation

κ – thermal conductivity

Thermal conductivity of solids varies significantly from one material to another due to a variety of factors, such as sample sizes for single crystals or grain sizes for polycrystalline samples, lattice defects or imperfections, dislocations, anharmonicity of the lattice forces, carrier concentrations, interactions between carriers and lattice waves, interactions between magnetic ions and lattice waves, and more. The diverse range of processes involved in thermal conductivity makes it an interesting area of study both experimentally and theoretically.

Most materials are very nearly homogeneous; therefore, we can usually write $\kappa = \kappa(T)$. Similar definitions are associated with thermal conductivities in the y- and z-directions (κ_y, κ_z), but for an isotropic material the thermal conductivity is independent of the direction of transfer, $\kappa_x = \kappa_y = \kappa_z = \kappa$.

From the foregoing equation it follows that the conduction heat flux increases with increasing thermal conductivity and increases with increasing temperature difference. In general, the thermal conductivity of a solid is larger than that of a liquid, which is larger than that of a gas.

This trend is largely due to differences in intermolecular spacing characteristic of the two states of matter.

3.1. Thermal conductivity metals

Metals are solids and have a crystalline structure where the ions occupy equivalent positions in the crystal lattice. They generally have high electrical conductivity and high thermal conductivity. As a result, thermal energy can be transported mainly through two effects [2]:

- free electrons migration
- lattice vibrational waves (phonons).

Metals have a unique feature in their structure that is the presence of charge carriers, specifically electrons. The electrical and thermal conductivities of metals originate from the fact that their outer electrons are delocalized. Their contribution to thermal conductivity is referred to as electronic thermal conductivity (κ_e). In pure metals such as gold, silver, copper, or aluminium, the heat current associated with the flow of electrons by far exceeds the contribution of the flow of phonons. In contrast, in alloys, the contribution of phonons to κ is no longer negligible.

In metals, heat conductivity is primarily attributed to free electron transfer. The correlation between electrical conductivity and thermal conductivity is proportional, but raising the temperature increases the thermal conductivity while decreasing the electrical conductivity. This behaviour is quantified in the Wiedemann–Franz law [3].

$$\frac{\kappa}{\sigma} = LT \quad (\text{Eq. 3})$$

Where:

σ – electrical conductivity [S/m]

κ – thermal conductivity [W/(m·K)]

L – Lorentz number [$2.44 \times 10^{-8} \text{V}^2/\text{K}^2$]

T – Temperature [K]

This relation is based on the fact that heat and electrical transport both involve free electrons in the metal. The electrical conductivity decreases as the particle velocity increases, because collisions divert electrons from the forward transport of charge. However, the thermal conductivity increases with the average particle velocity since that increases the forward transport of energy. The Wiedemann-Franz law is generally well-obeyed at high temperatures. However, in the low and intermediate temperature regions, the law fails due to the inelastic scattering of the charge carriers.

It should be noted that this general correlation between electrical and thermal conductance does not hold for other materials due to the increased importance of phonon carriers on heat transport in non-metals.

3.2. Thermal conductivity of other crystal structures

In non-magnetic insulating ceramic materials heat conduction is typically dominated by the propagation of phonons. However, the fundamental understanding of phonon heat conduction in solid-state materials has not advanced significantly since the 1960s, when the Debye model for gas heat transfer was adapted for this purpose. There has been little motivation to update this model, except for minor mathematical symbols and notation changes. Thus, the thermal conductivity of pure crystals is often expressed using this model [2, p. 115].

$$\kappa = \frac{1}{3} v l C_v \quad (\text{Eq. 4})$$

Where:

κ – thermal conductivity

v – the average sound velocity

Cv – specific heat

l – phonon mean free path

Thermal conductivity is largely determined by the phonon mean free path, which is a crucial factor. The integration involved in determining the mean free path is performed across all relevant phonon frequencies for a given material.

The equation explains that in perfect materials with a flawless lattice structure and harmonic vibration the thermal conductivity is infinite, as phonon propagation is unrestricted, and phonon mean free path is unlimited. However, real materials with a lattice structure which is not ideal have limitations in phonon mean free path and thermal conductivity due to the scattering effects of phonons caused by defects, boundaries, and other factors.

The expression for thermal conductivity, according to Callaway (Eq 5-10), is based on the assumption that all phonon scattering processes can be described using relaxation times[4], [5].

$$\kappa = \kappa_L + 2\kappa_T \quad (\text{Eq. 5})$$

$$\kappa_L = \kappa_{L1} + \kappa_{L2} \quad (\text{Eq. 6})$$

$$\kappa_{L1} = \frac{1}{3} C_L T^3 \int_0^{\Theta_{L/T}} \frac{\tau_C^L(x) x^4 e^x}{(e^x - 1)^2} dx \quad (\text{Eq. 7})$$

$$\kappa_{L2} = \frac{1}{3} C_L T^3 \frac{\left[\int_0^{\Theta_{L/T}} \frac{\tau_C^L(x) x^4 e^x}{\tau_N(x) (e^x - 1)^2} dx \right]^2}{\int_0^{\Theta_{L/T}} \frac{\tau_C^L(x) x^4 e^x}{\tau_N^L(x) \tau_R^L(x) (e^x - 1)^2} dx} \quad (\text{Eq. 8})$$

$$\kappa_{T1} = \frac{1}{3} C_L T^3 \int_0^{\Theta_{L/T}} \frac{\tau_C^L(x) x^4 e^x}{(e^x - 1)^2} dx \quad (\text{Eq. 9})$$

$$\kappa_{T2} = \frac{1}{3} C_L T^3 \frac{\left[\int_0^{\Theta_{T/T}} \frac{\tau_C^L(x) x^4 e^x}{\tau_N(x) (e^x - 1)^2} dx \right]^2}{\int_0^{\Theta_{T/T}} \frac{\tau_C^T(x) x^4 e^x}{\tau_N^T(x) \tau_R^T(x) (e^x - 1)^2} dx} \quad (\text{Eq. 10})$$

Where:

L longitudinal phonons

T transverse phonons

$$C_{L/T} = \frac{k_B^4}{2\pi^2 \hbar^3 v_{L/T}}$$

$$x = \frac{\hbar \omega}{k_B T}$$

T – Temperature

\hbar – Planck constant

k_B – Boltzmann constant

ω – phonon frequency

$(\tau_N)^{-1}$ is the scattering rate for normal phonon processes, $(\tau_R)^{-1}$ is the sum of all resistive scattering processes, and $(\tau_C)^{-1} = (\tau_N)^{-1} + (\tau_R)^{-1}$

k_B – Boltzmann constant

$\Theta_{L/T}$ – Debye temperature for longitudinal and transverse

$v_{L/T}$ – sound velocity

In general, phonon-boundary scattering can be disregarded, as the grain size is typically much larger than the phonon mean free path, which is limited by phonon-phonon and phonon-defect scatterings in bulk materials over the entire temperature range.

3.3. Density Functional Theory (DFT)

Density Functional Theory (DFT) is a theory of electronic structure based on the electron density distribution instead of the many-electron wave function $\Psi(r_1, r_2, r_3, \dots)$ [6].

DFPT (Density Functional Perturbation Theory) is a computational method that utilizes linear response theory to analyse the effects of small perturbations on the solutions of the Kohn-Sham equations, which are used to determine the electronic charge density. This approach allows for the investigation of various phenomena, such as phonons, electric field response, and phonon-electron coupling.

The calculations are performed in reciprocal space within the confines of the unit cell, enabling the determination of dynamical matrices across the entire phonon Brillouin zone without requiring supercells. To obtain the force constants in real space, an inverse Fourier transform is applied. However, DFPT calculations exhibit a computational scaling that grows with the fourth power of the number of atoms in the unit cell, making them more suitable for small unit cell crystals. Many Density Functional Theory packages provide the capability to calculate harmonic force constants using DFT (for example Vienna Ab Initio Simulation Package VASP [7], or Quantum Espresso [8]). The theoretical framework and implementation of DFPT for cubic force constants are currently being developed and integrated into DFT packages. It is possible to solve Boltzmann Transport Equation (BTE) (Eq. 11) [9] Via DFT phonon calculations.

$$\kappa_{\alpha} = \sum_{q,v} c \left(\begin{matrix} q \\ v \end{matrix} \right) v_{g,\alpha}^2 \left(\begin{matrix} q \\ v \end{matrix} \right) \tau_{\alpha} \left(\begin{matrix} q \\ v \end{matrix} \right) \quad (\text{Eq. 11})$$

Where:

κ – thermal conductivity

q – a phonon mode with wave vector

v – polarization

c – the volumetric specific heat, which is a function of temperature and frequency

τ_α – the lifetime (which is also called the scattering time or the transport lifetime) when the heat flux is applied in the α

α – direction.

3.4. Molecular Dynamic (MD)

The molecular dynamic (MD) method is a numerical simulation technique that can be used to study the behavior of atoms and molecules in materials at finite temperatures. It is a powerful tool for simulating material properties such as transport properties. The atomic positions and momenta of a sample can be evolved by numerically integrating the Newtonian equations of motion [10]. Heat flow J (Eq. 12) was developed by Ikeshoji & Hafskjold [11].

$$J = \frac{\Delta E_k}{\Delta t} \quad (\text{Eq. 12})$$

Where:

ΔE_k – kinetic energy

Δt – time step

If J is known, it is possible to compute the thermal conductivity (k) using different form of Fourier law (Eq 13) :

$$J_\alpha = \sum_\beta \kappa_{\alpha\beta} (\nabla T)_\beta \quad (\text{Eq. 13})$$

Where:

∇T – Thermal gradient

J_α – the α -component of heat current

$\kappa_{\alpha\beta}$ – the second-order tensor of thermal conductivity

The nonequilibrium, steady-state MD approach is required to simulate ∇T using the direct method. However, this approach is difficult to implement in the first-principles codes [12].

3.5.Green–Kubo method

The Green-Kubo formalism allows scientists to calculate many dynamic properties such as viscosity, thermal conductivity, and electrical conductivity from equilibrium simulations of atomic systems, as opposed to non-equilibrium molecular dynamics (NEMD). Either way, calculations of κ generally require large cells and long-time scales to converge the statistical sampling. As a consequence, the thermal conductivity of real materials has been calculated so far only for a restricted set of substances and at best with a semi-empirical description of the interatomic interactions [13].

In the Green-Kubo method, the lattice thermal conductivity κ_α along a particular direction α is given by (Eq. 14) [10]:

$$\kappa_\alpha = \frac{1}{k_b T^2 V} \int_0^\infty \langle J_\alpha(t) J_\alpha(0) \rangle dt \quad (\text{Eq. 14})$$

Where:

t – the time,

V – the system volume

T – temperature,

J_α – the α component of the lattice heat current vector J

$\langle J_\alpha(t) J_\alpha(0) \rangle$ – the ensemble averaged heat current auto-correlation function

3.6. Slack model Automatic Gibbs Library (AGL)

AGL (Automatic Gibbs Library) is a software that implements the “GIBBS” [14] method in AFLOW [15] and Materials Project [16] platform using C++ and Python framework. The library includes automatic error handling and correction to facilitate high-throughput computation of the material’s thermal properties [17]. In the AGL, the thermal conductivity is calculated by the method proposed by Slack using the Debye temperature and the Grüneisen parameter (Eq. 15).

$$\kappa_l(\theta_a) = \frac{0.849 \times 3 \sqrt[3]{4}}{20\pi^3(1 - 0.514\gamma^{-1} + 0.228\gamma^{-2})} \times \left(\frac{k_B \theta_a}{h}\right)^2 \frac{k_B m V^{1/3}}{h\gamma^2} \quad (\text{Eq. 15})$$

Where:

V – the volume of the unit cell

m – the average atomic mass

θ_a – Debye temperature (but this parameter is slightly different from the traditional Debye temperature - it is obtained by only considering the acoustic modes, based on the assumption that the optical phonon modes in crystals do not contribute to heat transport)

γ – Grüneisen parameter

Grüneisen parameter describes the effect that changing the volume of a crystal lattice has on its vibrational properties, and, as a consequence, the effect that changing temperature has on the size or dynamics of the lattice. The Grüneisen parameter is one of the most valuable quantities in thermodynamics, which links together the material properties of bulk modulus, heat capacity at constant volume, thermal expansion coefficient, and volume. [18]

The Debye temperature is a characteristic temperature scale for the lattice vibrations in a solid. It represents the temperature at which the solid has vibrational modes with the same frequency as those of a three-dimensional rigid body [19].

Authors claim that the correlation with experimental results is ~0.9

3.7.Applications of materials with low thermal conductivity

Materials with low thermal conductivity are desirable for many technological applications such as thermal management of electronic and photonic devices, heat exchangers, energy converters and thermal isolations.

3.7.1. Thermoelectric materials

Thermoelectric (TE) are the materials that can generate electricity from the application of a temperature gradient, or vice versa, through the thermoelectric effect. By exploiting this coupling between thermal and electrical properties, the thermoelectric devices can be made that carry heat from the cold to the hot side (refrigeration) or that generate electricity from heat flows.

Thermoelectric generators (TEGs) are solid-state devices that convert heat flux (temperature differences) directly into electrical energy through a phenomenon called the Seebeck effect. The dimensionless figure of merit (zT) is used to quantify TE performance and is related to the conversion efficiency (η). This energy conversion process is also highly reversible; thermoelectric materials can be used for power generation as well as solid-state refrigeration.

Thermal conductivity is one of the key features of the zT (Eq. 16) parameter. It is defined as [20].

$$zT = \frac{S^2 \sigma}{\kappa} T \quad (\text{Eq. 16})$$

Where:

S – the Seebeck coefficient,

σ – the electrical conductivity,

κ – the thermal conductivity

T – the absolute temperature

A higher zT value indicates a better thermoelectric material.

3.7.2. Thermal coating barriers

Thermal barrier coating (TBC) is a type of coating applied to metallic surfaces that operate at high temperatures, such as gas turbine engine parts, to protect them from heat damage and extend their life. TBCs are produced in several ways, including electron beam physical vapour deposition (EBPVD), air plasma spray (APS), high-velocity oxygen fuel (HVOF), electrostatic spray-assisted vapour deposition (ESAVD), and direct vapour deposition [21].

4. Machine Learning

Machine learning is a field of computer science and artificial intelligence. It is focused on solving problems by self-learning algorithms. In contrast to the classical approach, where a human indicates the way to solve a problem, in machine learning, the algorithm chooses the way to solve the problem, based on trial and error. The field of machine learning has gained a lot of popularity in recent years due to its wide range of applications. As a result, the leaders of the IT market began to invest heavily in this sector [22].

Machine learning is a powerful tool in materials research. It has been used to predict the corrosion behavior as well as the tensile and compressive strengths of the fiber/matrix interfaces in ceramic-matrix composites; and make faster and accurate predictions (using past historical data) of phase diagrams, crystal structures, and materials properties [23]. It is mostly concerned with supervised learning. Machine learning constructs models for specific material properties and quickly achieves the prediction of material properties [24].

Emerging materials informatics tools also offer tremendous potential and new avenues of mining for structure-property-processing linkages from aggregated and curated materials data sets [25].

4.1.Data Validation

Data validation is an important step in machine learning. It is the process of ensuring that the data used to train a machine learning model is accurate and reliable. This is important because if the data is inaccurate or unreliable, the model will not be able to make accurate predictions. Data validity significantly affects the quality of the generated model.

Firstly, a very common problem is insufficient data. Unfortunately, to solve even the simplest problems, one needs hundreds or even thousands of samples representing the discussed issue.

Secondly, a large amount of data of poor quality may pose another obstacle. The problem here may be an incorrect description of the data, outliers, differences in units, scales of the given data, errors in the training data, or incomplete data. In this case, it is necessary to clean up the data by spending a considerable amount of time detecting inconsistencies and cleaning the data from undesirable fragments that prevent or hinder training.

Another problem may be the non-representativeness of the training data. This means that the data set may not include examples that could significantly affect the result of learning, or there is a disproportionately large amount of one type of data as compared to the others, which will make the algorithm lean towards more representation of data.

4.2.Data Preprocessing

Data preprocessing is the first step in machine learning. It refers to the technique of preparing (cleaning and organizing) the raw data to make it suitable for building and training machine learning models. In other words, data preprocessing is a data mining technique that transforms raw data into an understandable and readable format [26].

4.3.Dimensionality reduction

Dimensionality reduction is the process of reducing the number of features in a dataset while retaining as much information as possible. It is used to overcome the 'curse of multi-dimensionality', which refers to the fact that high-dimensional data is often sparse and difficult to work with. The curse of dimensionality is a common problem in machine learning where the performance of the model deteriorates as the number of features increases. This is because the complexity of the model increases with the number of features, and it becomes more difficult to find a good solution. In addition, high-dimensional data can also lead to overfitting, where the model fits the training data too closely and does not generalize well to new data. There are several techniques for dimensionality reduction including principal component analysis (PCA), singular value decomposition (SVD), and *t*-distributed stochastic neighbor embedding (t-SNE) [27].

Advantages and Disadvantages of Dimensionality Reduction include [27]:

- It helps in data compression, and hence reduces storage space.
- It reduces computation time.
- It also helps remove redundant features, if there are any.
- Improved Visualization: High dimensional data is difficult to visualize, and dimensionality reduction techniques can help in visualizing the data in 2D or 3D, which can help in better understanding and analysis.

- **Overfitting Prevention:** High dimensional data may lead to overfitting in machine learning models, which can lead to poor generalization performance. Dimensionality reduction can help in reducing the complexity of the data, and hence prevent overfitting.
- **Feature Extraction:** Dimensionality reduction can help in extracting important features from high dimensional data, which can be useful in feature selection for machine learning models.
- **Improved Performance:** Dimensionality reduction can help in improving the performance of machine learning models by reducing the complexity of the data, and hence reducing the noise and irrelevant information in the data.
- It may lead to some amount of data loss.
- **Interpretability:** The reduced dimensions may not be easily interpretable, and it may be difficult to understand the relationship between the original features and the reduced dimensions.
- **Overfitting:** In some cases, dimensionality reduction may lead to overfitting, especially when the number of components is chosen based on the training data.
- **Sensitivity to outliers:** Some dimensionality reduction techniques are sensitive to outliers, which can result in a biased representation of the data.

The simplest way to do the reduction is feature selection which is a technique for selecting a subset of the original features that are most relevant to the problem at hand. The goal is to reduce the dimensionality of the dataset while retaining the most important features. There are several methods for feature selection including filter methods, wrapper methods, and embedded methods.

4.4.Unsupervised Correlation Method

Unsupervised Correlation Method is based on creating a matrix of correlation. Afterward the selection of the threshold value all data with correlation higher than the threshold (except the diagonal) are dropped. Examples in Fig 3.

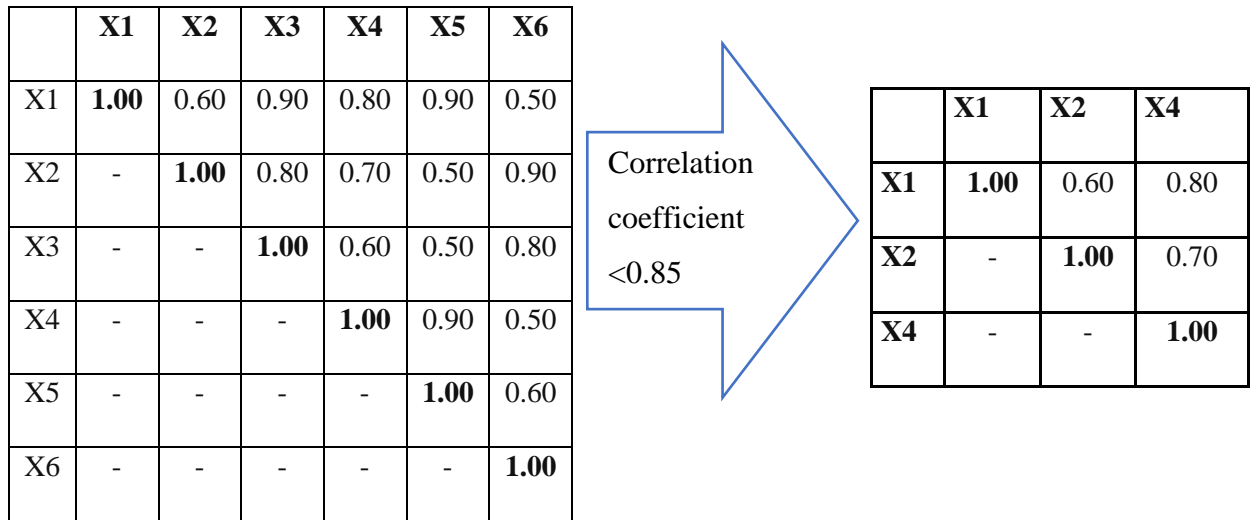


Figure 3. Visualisation of Dimensionality Reduction by dropping features with correlation higher than threshold value

4.5. Scaling

Another aspect may concern the different scales of individual data. When there are significant differences in the values of individual data, the algorithm can pay more attention to the data with higher values, ignoring insignificant changes in the data, e.g., orders of magnitude lower, even if they have a significant impact on the result. The purpose of the data scaling is to avoid issues with the features having different units or ranges, which artificially adds more weight to certain features in the model. It also helps improve the performance of predictive modeling algorithms [28]. It can be dealt with this in two ways:

- **Min-max scaling** is a simple way to set values between 0 and 1. This is done by subtracting the minimum value from the given value and then dividing it by the difference between the minimum and maximum values.
- **Standardization** – involves subtracting the mean value from each sample and then dividing it by the standard deviation. As a result, the dataset has unit variance, and the mean is always zero. The advantage of this method is high insensitivity to outliers.

4.6. Metrics for Machine Learning

Regression metrics are used to evaluate the performance of regression models. The most popular metric for regression algorithms is Mean Squared Error (MSE). Other regression metrics include Mean Absolute Error (MAE) and Coefficient of determination R^2 (R-squared).

4.6.1. Mean Squared Error

The MSE (Eq. 17) is calculated as the mean or average of the squared differences between predicted and expected target values in a dataset [29].

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (\text{Eq. 17})$$

Where:

MSE – the mean squared error.

n – the number of observations.

y_i – the observed value of the dependent variable for observation i .

\hat{y}_i – the predicted value of the dependent variable for observation i

4.6.2. Mean Absolute Error

In Regression Metrics, Mean Absolute Error (MAE) (Eq. 18) is the average of the absolute differences between the actual value and the model's predicted value [29]. It is a common metric used to evaluate regression models. A lower MAE indicates that the model is better at predicting the values of the dependent variable.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (\text{Eq. 18})$$

Where:

MAE – the mean absolute error.

n – the number of observations.

y_i – the observed value of the dependent variable for observation *i*.

y_i^{hat} – the predicted value of the dependent variable for observation *i*

4.6.3. Coefficient of determination (R²)

R² (Eq. 19) or Coefficient of Determination is a prevalent metric that uses two mean squared error calculations. While the former is the mean square of each real value versus the average of observations, the latter is the mean squared error of the actual value versus the predicted one [29].

Best possible score is 1.0 and it can be negative (because the model can be arbitrarily worse). In the general case when the true value is non-constant, a constant model that always predicts the average value disregarding the input features would get a score of 0.0 [30].

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (\text{Eq. 19})$$

Where:

R² – the coefficient of determination.

n – the number of observations.

y_i – the observed value of the dependent variable for observation *i*.

y_i^{hat} – the predicted value of the dependent variable for observation *i*.

y^{bar} – the mean value of the dependent variable.

The R² equation measures how much of the variance in the dependent variable can be explained by the independent variable. A high R² value indicates that there is a strong correlation between

the independent variable and the dependent variable. However, a high R^2 value does not necessarily mean that there is a causal relationship between the variables.

5. Algorithms

Machine learning algorithms are the programs that can learn the hidden patterns from the data, predict the output, and improve the performance from experiences on their own. In simple terms, a machine learning algorithm is like a recipe that allows computers to learn and make predictions from data. Instead of explicitly telling the computer what to do, we provide it with a large amount of data and let it discover patterns, relationships, and insights on its own [31].

Regression analysis is a fundamental concept in the field of machine learning. It falls under supervised learning wherein the algorithm is trained with both input features and output labels. It helps in establishing a relationship among the variables by estimating how one variable affects the other. Regression in machine learning consists of mathematical methods that allow data scientists to predict a continuous outcome (y) based on the value of one or more predictor variables (x) [32], [33].

One of the key issues is the choice of algorithm. Some of these may focus on different aspects of the problem.

5.1. Least Absolute Shrinkage and Selection Operator (LASSO)

LASSO (Least Absolute Shrinkage and Selection Operator) is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the resulting statistical model. It is an adaptation of the popular and widely used linear regression algorithm. It enhances regular linear regression by slightly changing its cost function, which results in less overfit models [34].

LASSO is used when there are many predictors in a model and some of them are irrelevant or redundant. It shrinks the coefficients of the irrelevant predictors to zero, effectively removing them from the model. This helps to reduce overfitting and improve the generalization performance of the model.

The target can be a 2-dimensional array, resulting in the optimization of the following objective (Eq. 20) [35]:

$$MIN MSE(y, y_{pred}) + \alpha \sum_{i=1}^n |\theta_i| \quad (\text{Eq. 20})$$

Where:

y_i – is the target,

θ_i – is the coefficient

y_{pred} – is the predicted feature

n – number of model parameters

θ_i – parameter vector

5.2.Support Vector Regression (SVR)

Support Vector Regression (SVR) is a regression algorithm that uses Support Vector Machines (SVMs) to perform regression. SVMs are supervised learning models that analyze data for classification and regression analysis.

In most linear regression models, the objective is to minimize the sum of squared errors. Take Ordinary Least Squares (OLS) for example. The objective function for OLS with one predictor (feature) is as follows (Eq. 21) [36]:

$$MIN \sum_{i=1}^n (y_i - w_i x_i)^2 \quad (\text{Eq. 21})$$

Where:

y_i – is the target,

w_i – is the coefficient

x_i – is the predictor (feature).

Lasso, Ridge, and SVR are extensions of a simple equation used in linear regression. They add an additional penalty parameter that aims to minimize complexity and/or reduce the number of features used in the final model. Regardless, the aim — as with many models — is to reduce the error of the test set.

SVR is implemented using libsvm [37] and uses a parameter to control the number of support vectors.

5.3. Kernel Ridge Regression (KRR)

Kernel ridge regression combines ridge regression (linear least squares with l2-norm regularization) with the kernel trick. It thus learns a linear function in the space induced by the respective kernel and the data. For non-linear kernels, this corresponds to a non-linear function in the original space.

The form of the model learned by KRR is identical to support vector regression (SVR). However, different loss functions are used: KRR uses squared error loss while support vector regression uses epsilon-insensitive loss, both combined with l2 regularization. In contrast to SVR, fitting a KRR model can be done in closed form and is typically faster for medium-sized datasets. On the other hand, the learned model is non-sparse and thus slower than SVR, which learns a sparse model for $\epsilon > 0$, at prediction time [38].

Main function is minimizing target shown as (Eq. 22):

$$\text{MIN } MSE(y, y_{pred}) + \alpha \sum_{i=1}^n \theta_i^2 \quad (\text{Eq. 22})$$

Where:

y_i – is the target,

θ_i – is the coefficient

y_{pred} – is the predicted feature

n – number of model parameters

θ_i – parameter vector

5.4.eXtreme Gradient Boosting (XGBoost, XGB)

Extreme Gradient Boosting is a tree-based algorithm, which sits under the supervised branch of Machine Learning. Boosting is an ensemble technique that combines predictions from multiple models into one. It accomplishes this by sequentially modelling each predictor based on the errors of its predecessor (assigning greater weight to predictors with better performance).

Basics of this algorithm:

- **Tree-based algorithms** – XGBoost use decision trees as base estimators.
- **Prediction target** – the trees are built using residuals, not the actual class labels. Hence, despite us focusing on classification problems, the base estimators in this algorithm is

regression trees and not classification trees. This is because residuals are continuous and not discrete.

- **Tree depth** – algorithm allows to control the maximum size of the trees to minimize the risk of overfitting the data.
- **Ensemble methods** – like Random Forest or AdaBoost, this algorithm builds many trees in the process. In the end, the final prediction is based on all of the trees.
- **Learning rate** – the value of each tree is scaled by the learning rate. This enables the algorithm to have a more gradual and steady improvement at each step.

Process map — finally, here is a simple illustration of the process used by XGBoost Fig 4.

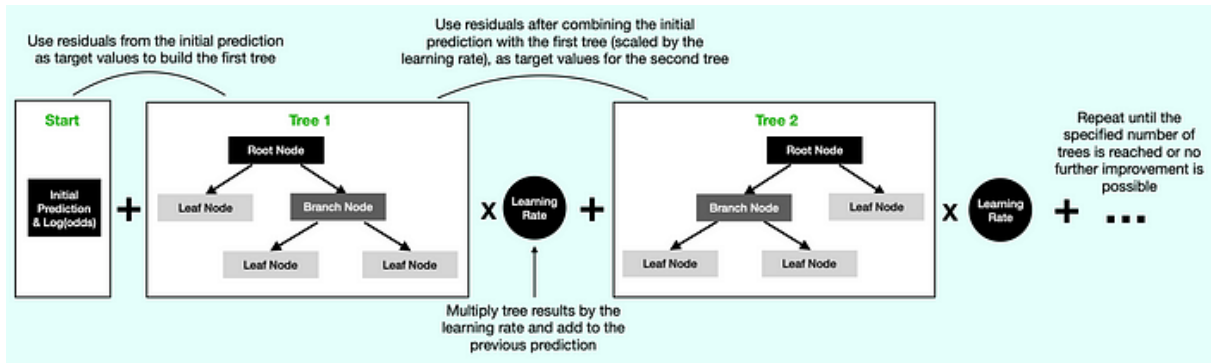


Figure 4 The process map for Gradient Boosting and XGBoost algorithms. Image by [38]

Gradient boosting is a type of boosting that minimizes the loss function using a gradient descent algorithm [39].

XGBoost uses its own method of building trees where the Similarity Score (Eq. 23) and Gain determine the best node splits.

$$\text{Similarity Score} = \frac{(\sum_{i=1}^n \text{Residual}_i)^2}{\sum_{i=1}^n [\text{Previous probability}_i (1 - \text{Previous probability}_i)] + \lambda} \quad (\text{Eq. 23})$$

Residual is actual (observed) value - predicted value.

Previous probability is the probability of an event calculated at a previous step. The initial probability is assumed to be 0.5 for every observation, which is used to build the first tree. For any subsequent trees, the previous probability is recalculated based on initial prediction and predictions from all prior trees, as shown in the process map.

Lambda is a regularization parameter. Increasing lambda disproportionately reduces the influence of small leaves (the ones with few observations) while having only a minor impact on larger leaves (the ones with many observations).

6. Review of literature about Predicting lattice thermal conductivity via machine learning

In recent decades, two main approaches to predicting lattice thermal conductivity (κ_L) have been molecular dynamics simulations and first-principles calculations. However, these methods are limited by either low accuracy or high computational cost. To address these issues, machine learning (ML) has been successfully employed to accurately predict κ_L in a high-throughput manner. Y. Luo et al. [40] create a mini review on this topic.

6.1. Datasets

The κ_L (lattice thermal conductivity) can be collected from first-principles calculations, molecular dynamics simulations, experimental measurements, or materials databases. The dataset is usually divided into two subsets, including the training and testing sets, and principal component analysis (PCA) can be used to identify distinct features in the dataset. The input features for ML algorithms usually contain information about structural properties, elemental properties, and phonon properties. To solve the problem of representing systems of different sizes using feature vectors of fixed length, statistical values of elemental properties can be adopted. It is important to screen out features closely related to the target property and to check feature relevancy and redundancy before any training process is to begin. Some approaches, such as the active sample selection based on PCA and recursive feature elimination (RFE), can improve the model performance and reduce the dimensionality of the feature vector.

There are a few ways to collect that type of data. For example, using databases such as AFLOW [15] or Materials Project [16]. However, in most papers, the size of the data sets is relatively small by machine learning standards. Only in 3 out of 11 articles, are there more than 1000 chemical compounds in the data set. Moreover, some of those are focused only on one type of structure, for example Half-Heusler (HH) compounds. Almost all of them relate to κ_L the temperature of 300K.

Values of κ_L could be calculated via Automatic Gibbs Library (AGL) method [41], Slack model, DFT calculations, MD, or another first-principle method. Those values should be compared with experimental data to check if it is close to real thermal conductivity.

6.2.Descriptors

Descriptors are designed as the input data for machine learning to bridge the gap between the atomic structure and particular properties of a system. The vector that represents descriptors must have the same length for each sample.

Input features could be collected in clusters representing e.g.

- **Compound** – properties of the atoms that make up the compound e.g. density, boiling point, atomic mass, atomic radius, electronegativity, etc.
- **Structure of compound** – properties related to crystal structure e.g., volume of the cell, number of atoms in one cell, space group, lattice type, number of species, etc.
- **Thermal properties of the compound** – e.g., Debye temperature, specific heat capacity, Gruneisen parameter, zero-point energy.

A comparison of descriptors (features) from five papers can be found in the supplementary file.

Descriptors based on the **weighted average** of individual properties of atoms in a chemical compound:

- **Atomic number** – the number of a chemical element is the number of protons in the nucleus of an atom of the element. It determines the identity of an element and many of its chemical properties.
- **Mendeleev number** – in 1984 Pettifor [42] suggested that a well-structured chemical space can be derived by changing the sequence of the elements in the Periodic. He proposed a chemical scale that determines the “distance” between the elements on a one-dimensional axis and a Mendeleev number (MN) — an integer showing the position of an element in the sequence. Pettifor claimed that binary compounds with the same structure type occupy the same region in a two-dimensional map plotted using the MNs (the Pettifor map) [42].
- **Period and Group** – The periodic table is arranged in rows and columns. The rows are called periods and the columns are called groups. Elements in the same group have similar chemical properties because they have the same number of valence electrons. Elements in the same period have the same number of electron shells.

- **Atomic mass** – The atomic mass of an element is the mass of an atom of that element, measured in atomic mass units (AMU). It is expressed as a multiple of one-twelfth the mass of the carbon-12 atom, which is assigned an atomic mass of 12 units. Some scientists have postulated a strong relationship between the atomic mass and the conductivity of the element.
- **Atomic density** – (N ; atoms/cm³) is the number of atoms of a given type per unit volume (V ; cm³) of the material.
- **Valence electrons** – Valence electrons are the electrons in the outermost shell of an atom that participate in chemical reactions. They are the electrons that are involved in forming bonds with other atoms.
- **Absolute radius of the atom** – it is difficult to measure because the electron cloud surrounding the nucleus does not have a clearly defined boundary. However, the atomic radius is defined as half the distance between the nuclei of two identical atoms which are bonded together.
- **Covalent radius** – it is defined as half of the distance between the nuclei of two identical atoms that are bonded together by a single covalent bond. Covalent radii are usually smaller than the atomic radii because the atoms tend to form bonds with other atoms to achieve a more stable electron configuration.
- **Van der Waals radii** – The Van der Waals radius is the distance between the nuclei of two atoms when they are joined by a van der Waals bond. It is the volume “occupied” by an individual atom (or molecule). The van der Waals volume may be calculated if the van der Waals radii (and, for molecules, the inter-atomic distances and angles) are known.
- **Electron affinity** – the energy change that occurs when an electron is added to a neutral atom to form a negative ion. It is a measure of how much an atom “wants” to gain an electron. The greater the electron affinity, the more likely an atom is to gain an electron.
- **Electronegativity** – it is a measure of the tendency of an atom to attract a shared pair of electrons towards itself when it is bonded to another atom. It is a relative measure and is based on the electronegativity scale developed by Linus Pauling. The higher the electronegativity value of an atom, the more strongly it attracts electrons towards itself.

- **First ionization energy** – is required to remove the outermost electron from a neutral atom in the gas phase. It is a measure of how tightly an electron is held by an atom. The higher the first ionization energy of an atom, the more difficult it is to remove an electron from that atom.
- **Polarizability** – the ability of an atom or molecule to be distorted by an external electric field. It is a measure of how easily the electron cloud of an atom or molecule can be distorted. The greater the polarizability of an atom or molecule, the more easily it can be distorted by an external electric field. boiling point.
- **Melting point of an element** – the temperature at which it changes from a solid to a liquid state. It is a measure of the strength of the forces holding the atoms or molecules together in the solid state. The higher the melting point of an element, the stronger the forces holding its atoms or molecules together in the solid state.
- **Molar volume** – is the volume occupied by one mole of a substance at a given temperature and pressure. It is equal to the molar mass divided by the mass density.
- **Thermal conductivity of an element** – the ability of a material to conduct heat. The thermal conductivity of a compound depends on its chemical composition and crystal structure. Some factors that affect the thermal conductivity of a compound include: the number of atoms in the molecule, the strength of the bonds between the atoms, and the arrangement of the atoms in the crystal lattice.
- **The orbital exponent of Slater-type orbitals** – it is a parameter used in quantum chemistry to describe the shape of atomic orbitals. Slater-type orbitals are a type of basis function used in quantum chemical calculations to describe the electronic structure of molecules. The orbital exponent determines the orbital's size and shape and affects the accuracy of the calculation.
- **Global hardness** – it is a measure of the resistance of a molecule against deformation. It is a measure of the energy required to remove an electron from the highest occupied molecular orbital (HOMO) and to add an electron to the lowest unoccupied molecular orbital (LUMO). The global hardness of a molecule is related to its reactivity and stability. The higher the global hardness of a molecule, the less reactive it is and the more stable it is.

- **Electrophilicity indices** – are used to measure the ability of a molecule to accept electrons. It is a measure of the electrophilic character of a molecule.
- **Atomization enthalpy** – the amount of energy required to separate all the atoms in one mole of a compound into individual atoms in the gas phase. It is also known as the heat of atomization or enthalpy of atomization.
- **Fusion enthalpy** – the amount of energy required to melt a given amount of a substance at its melting point.
- **Vaporization enthalpy** – the amount of energy required to vaporize a given amount of a substance at its boiling point.
- **Binding energy** – the smallest amount of energy required to remove a particle from a system of particles or to disassemble a system of particles into the individual parts. Binding energy is especially applicable to subatomic particles in atomic nuclei, to electrons bound to nuclei in atoms, and to atoms and ions bound together in crystals

Features based on the structure of the material under study:

- **Space group** – In crystallography, a space group is the symmetry group of an object in space, usually in three dimensions. The elements of a space group (its symmetry operations) are the rigid transformations of an object that leave it unchanged. A descriptor has to represent one of the 230 space groups [43].
- **Cell volume** – In crystallography, cell volume refers to the volume of the unit cell, which is the smallest repeating unit of a crystal lattice.
- **Atom volume** – it is the volume of a unit cell divided by the number of atoms that are in it.

6.3. Algorithms

Due to the fast growth of artificial intelligence, numerous machine-learning algorithms have been suggested, including:

- Bayesian Optimization (BO)[44]
- eXtreme Gradient Boosting (XGBoost) [45]
- Neural Network (NN) [45]
- Kernel Ridge Regression (KRR) [45]

- Least Absolute Shrinkage and Selection Operator (LASSO)
- Sure Independence Screening and Sparsifying Operator (SISSO) [46], [47]
- Gaussian Process Regression (GPR) [47]

6.4. Conclusion

With the rapid advancement of artificial intelligence, various machine learning (ML) algorithms have been proposed for the direct prediction of thermal conductivity (κ_L) in high-throughput models. These models typically utilize input features that encompass fundamental physical properties related to the systems under investigation, such as lattice constants, phonon frequencies, and atomic masses. Such data-driven models have shown promising results in rapid screening and inverse design of materials with desired κ_L , even beyond the training sets. Developing efficient ML models often requires calculating energies, forces, and stresses of a large number of atomic configurations using density functional theory (DFT), which can be time-consuming for systems with large unit cells and complex compositions. Additionally, the employed features usually capture only the atomic environment within a certain cutoff radius, neglecting long-range interactions that may be crucial for thermal transport properties in some cases. It is expected that the efficiency and accuracy of ML can be further enhanced by careful selection and optimization of input features and learning schemes.

ML models could be also improved by increasing information in one sample, by for example creating new descriptors, which represent crystal structure or previewing unverified features such as bonding inhomogeneity.

The hypothesis after analyzing that review is to find which descriptors are the most important for predicting thermal conductivity, to analyze possible new descriptors for describing materials and test their effectiveness in predicting κ_L for materials analyzed in the AGH laboratory.

7. Learning Procedure

7.1. Acquisition of data and preprocessing

The data used in this work was downloaded from the AFLOW repository [48]. The set consisted of 5664 samples with the following properties:

- chemical formula
- auid (aflow ID)
- aurl (aflow internet address)
- Space group
- AGL thermal conductivity
- Volume of atom in cell
- Volume of cell

The Grüneisen parameter as well as Debye temperature in the AFLOW database was calculated via the AGL method as well as the lattice thermal conductivity. Because of that, those parameters should not be applied in the machine learning method. It could decrease the range of materials which could be processed via the ML model, and it increases the risk of overfitting [49].

7.2. Descriptors

Some of the descriptors were created on the basis of their chemical composition. It was done in part by using both the code and the property database from an article by X. Wang et al [45].

In the end, there were 28 features left for each material (Tab. 1).

Table 1 List of used features.

Num.	Feature Name	Num.	Feature Name
1	Atomic number	15	Boiling point
2	Space group	16	Fusion enthalpy
3	Cell volume	17	Covalent radii
4	Mendeleev number	18	Atomization enthalpy
5	Thermal conductivity of element	19	Global hardness
6	Valence electrons	20	Atomic density
7	Electron affinity	21	Melting point
8	Molar volume	22	Van der Waals radii
9	Orbital exponent of Slater-type orbitals	23	First ionization energy
10	Polarizability	24	Electrophilicity indices
11	Electronegativity	25	Binding energy
12	Atom volume	26	Atomic mass
13	Group	27	Vaporization enthalpy
14	Absolute radii	28	Period

The distribution of the data is shown on the histogram (see Supplementary Files).

After downloading the data and creating descriptors based on the chemical formula, the dataset was pre-processed. Values have been normalized, scaled, shuffled, and split to train and test sets (90% train, 10% test). The ML models were built by training the dataset containing the target property and the pre-processed features. During the training process, a 10-fold cross-validation process was repeated 10 times with shuffled datasets to assess the model's viability and predictive capability. The R^2 , Mean Absolute Error score and Mean Square Error score for

each fold were chosen as the metrics for assessing the models in this work and were then averaged alongside each repetition. The processed data was introduced into the algorithms. After calculations fitted models were applied for experimental data.

Table 2. shows a statistical summary of the results of the tested algorithms. From the data presented below it can be concluded that Extreme Gradient Boosting is the best-tested algorithm for test dataset with an R^2 score of 0.89 and an MAE 2.50 W/(m·K) and 73% of samples were within the bounds of this error. It constitutes a significant improvement compared to SVR and LASSO models, and on this problem, it performs slightly better than the KRR model. However, R^2 equal to 1 may indicate an overfitting of the algorithm. The best result achieved over the course of this work is slightly worse than the one presented in [45], where on similar datasets they achieve $R^2 = 0.90$ and $MAE = 2.13$

Table 2 Statistical summary of the prediction performance on the training/testing data set for the ML models generated from all used descriptors.

Algorithm	KPI	Train	Test
KRR	R^2	0.85	0.75
	MSE	34.7	78.84
	MAE	1.77	2.82
SVR	R^2	0.76	0.69
	MSE	56.49	95.37
	MAE	1.26	2.86
XGB	R^2	1	0.89
	MSE	6.22	35.22
	MAE	1.38	2.50
LASSO	R^2	0.34	0.34
	MSE	154.82	205.24
	MAE	5.25	5.65

The Figure 5 plots the results of the predicted thermal conductivity for test set versus the AGL-calculated values using the XGBoost model, which has the best performance among the four thermal conductivity regression models as identified from Table 2.

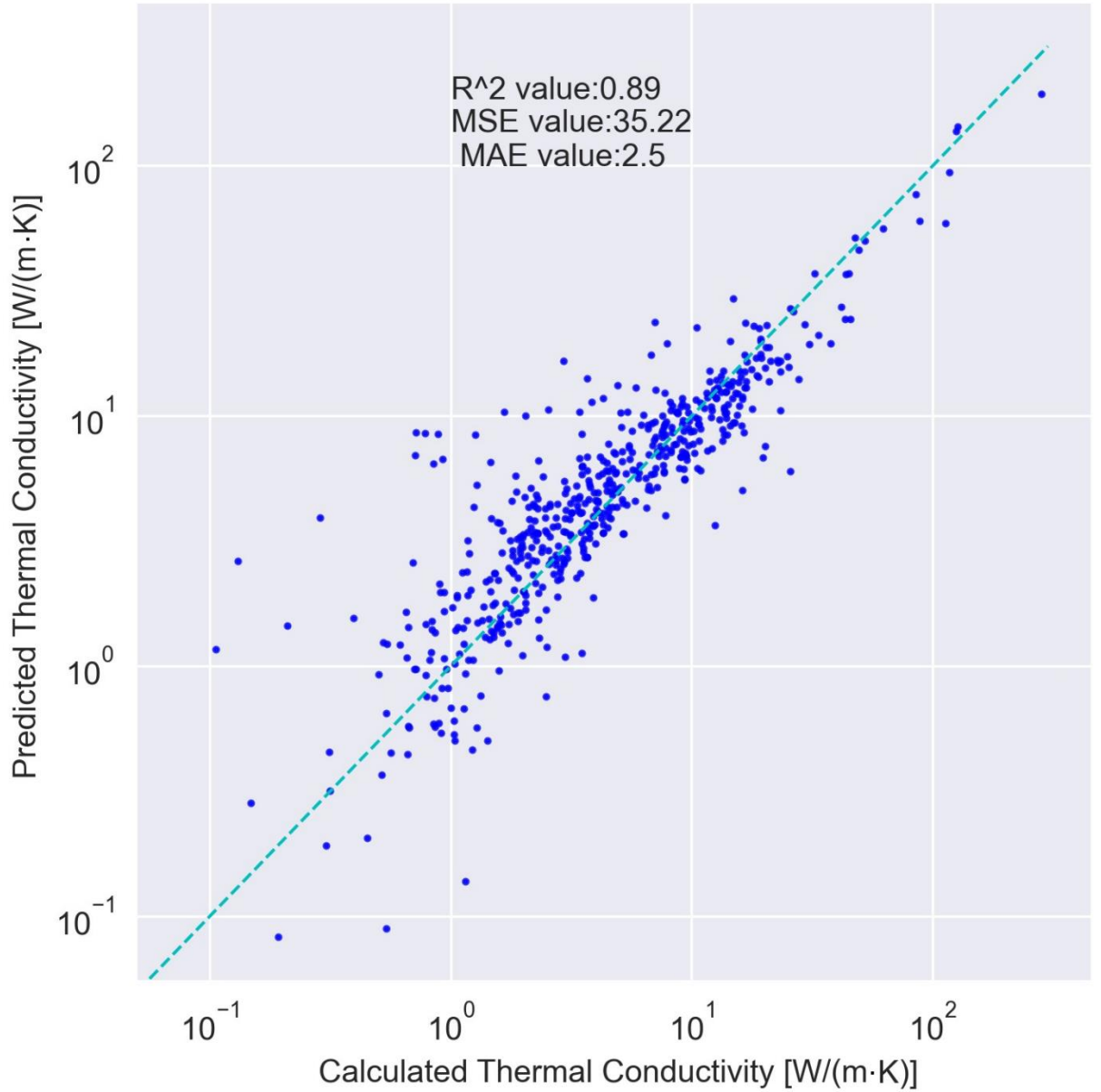


Figure 5 The XGBoost model predicted κ_l results versus calculated values on the test set containing 567 materials (log scale).

7.3. Multifactor dimensionality reduction

Figure 6 presents a heatmap of all descriptors used. From this matrix, it is possible to find a correlation between specific features. A pair plot, which is more accurate because it represents a dependency graph between each feature, is attached in the Supplementary file.

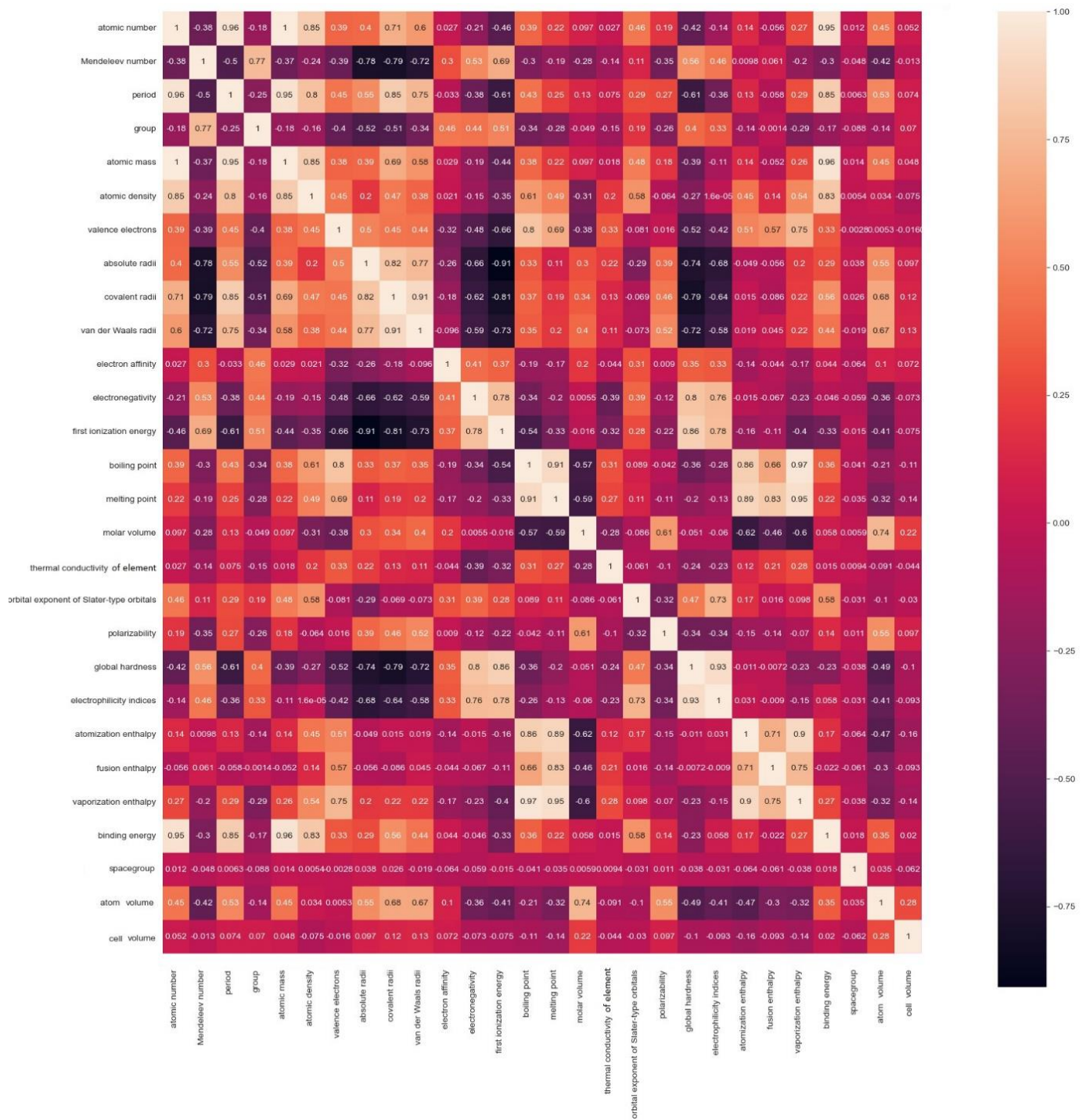


Figure 6 Heatmap of correlations between features

After analysis, a conclusion was drawn that quite a few features are significantly correlated with each other. It was decided on the basis of the correlation matrix to reject values that have surpassed the threshold values. From 100% correlation to 30% in 5% increments. The results

are shown in Figures 7-10. Calculations were performed for the two best performing algorithms XGB and KRR. MAE and R^2 are presented as key values.

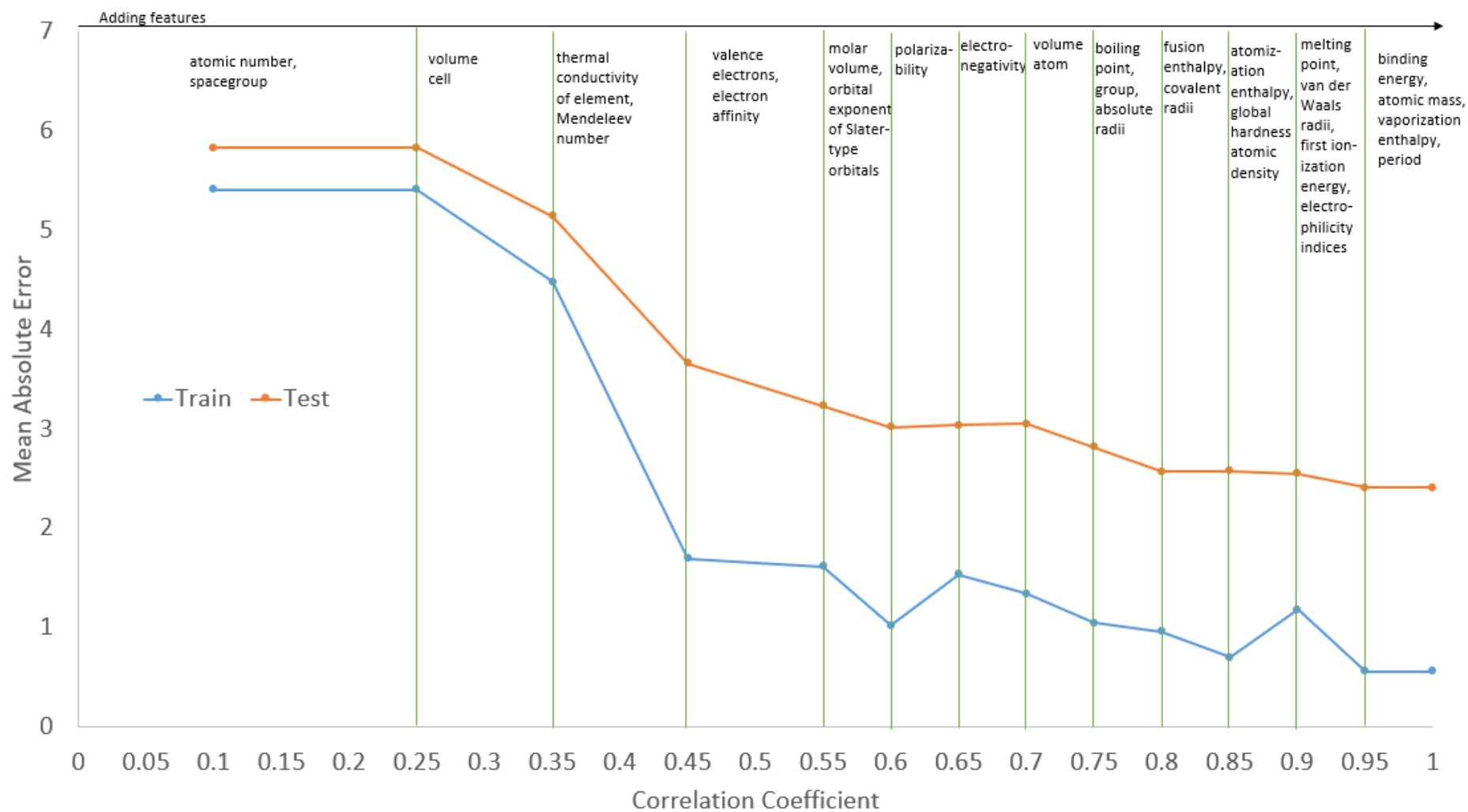


Figure 7 MAE function from a correlation of dropped values for XGB algorithm. (Less is better)

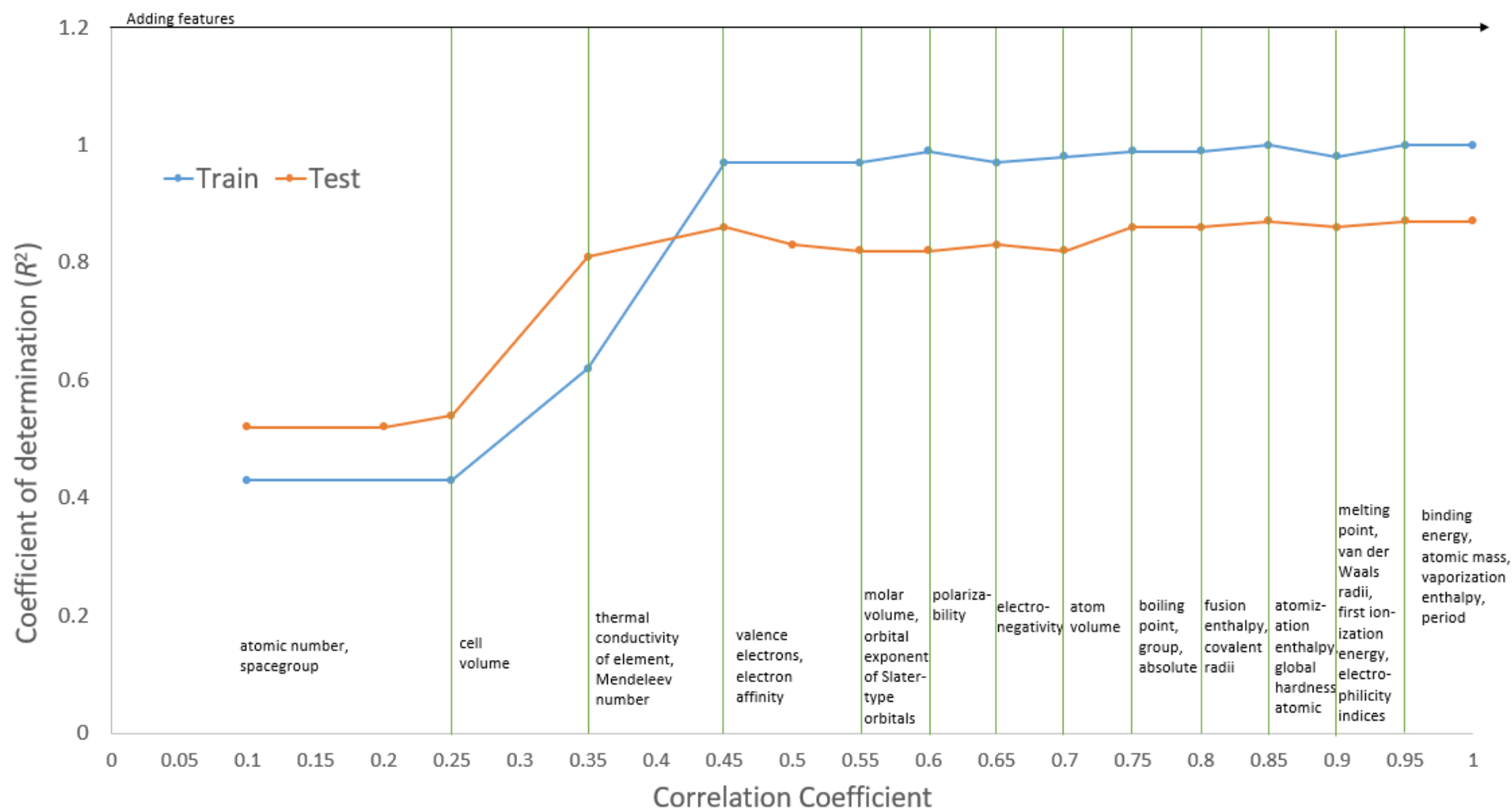


Figure 8 R^2 function from correlation of dropped values for XGB algorithm. (Higher is better)

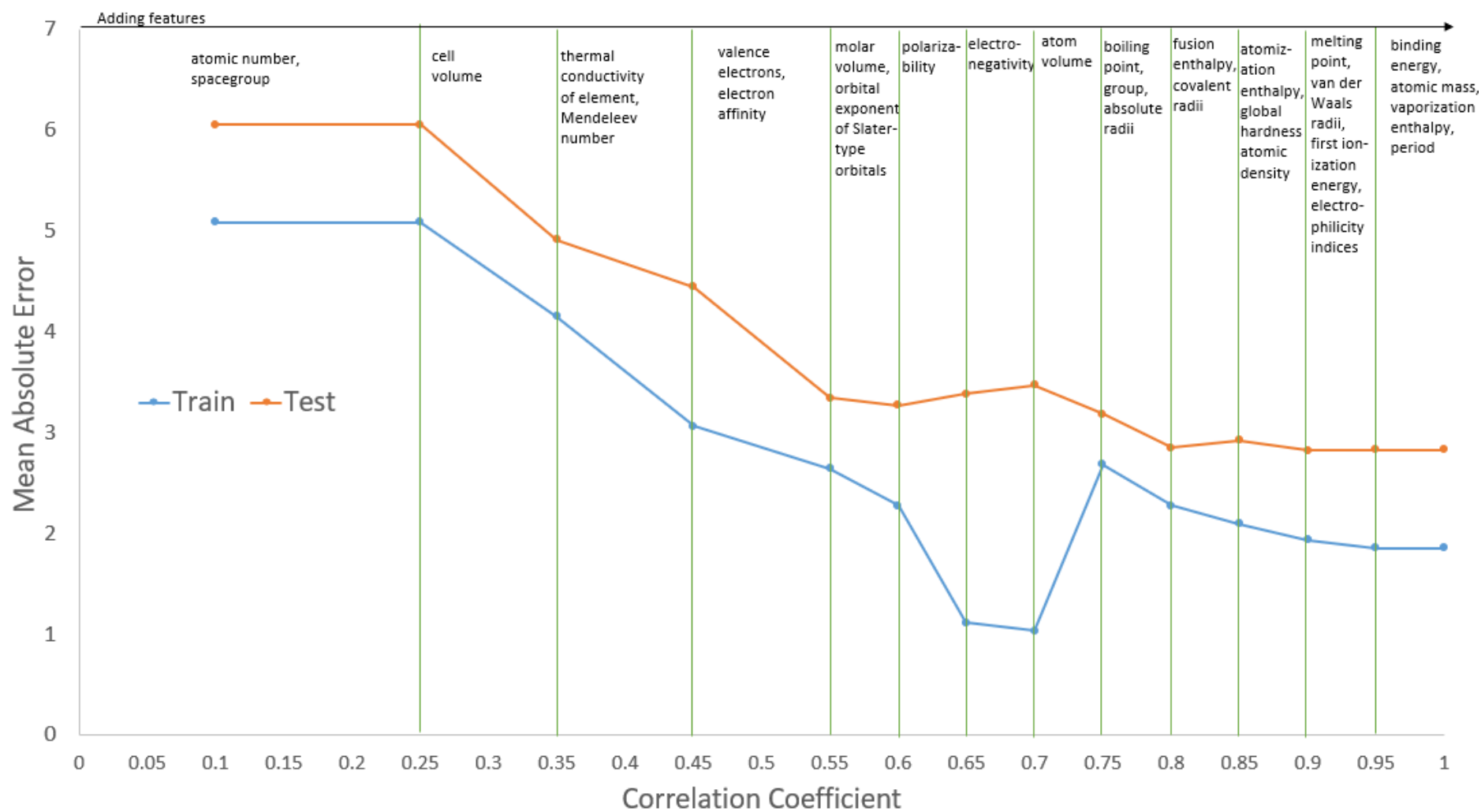


Figure 9 MAE function from correlation of dropped values for KRR algorithm. (Less is better)

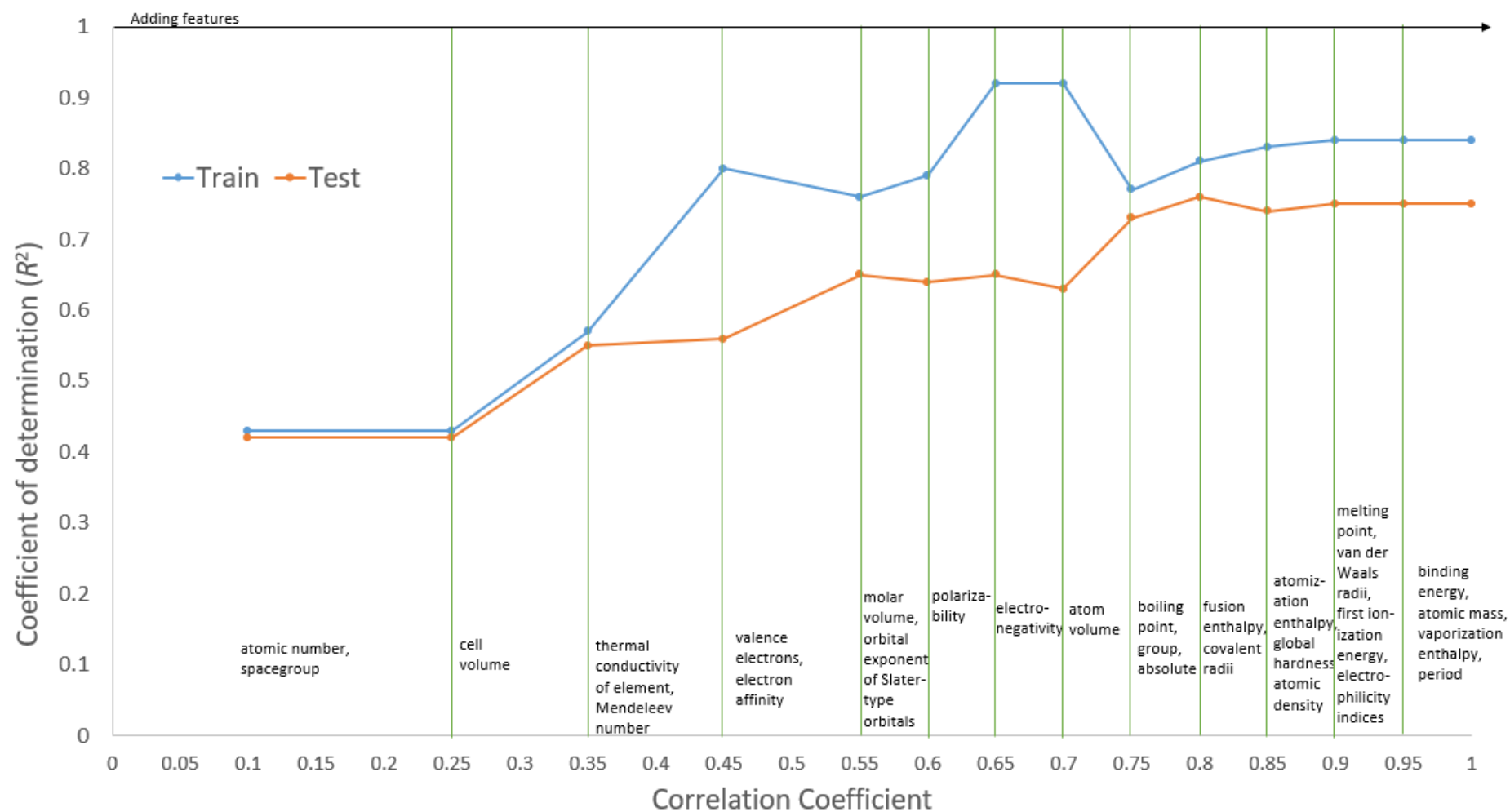


Figure 10 R^2 function from correlation of dropped values for KRR algorithm. (Higher is better)

After these calculations, it was decided to remove the Mendeleev number feature because it is a non-empirical feature and has other features implemented in it. The results for both algorithms are shown in Figures 11 -14

From these graph's results, despite the removal of 18 of the 28 features (60% correlation), there was no significant decrease in MAE Growth and R^2 decline for the XGB algorithm. The removal of the thermal conductivity of element resulted in a significant worsening of the KPI; the removal of the volume cell feature had the biggest impact on the decrease of the results.

The KRR algorithm does not provide such obvious results (Fig 13 and 14). First, removing features above a 70% correlation resulted in a decrease in KPI for the test set and an increase for the training set. It is not known what this anomaly can be attributed to. After removing more features, the R^2 and MSE values returned to the trend from before the removal of that feature in 70% correlation for the training set. However, for this algorithm too, the removal of Thermal conductivity and Mendeleev number resulted in a deterioration of the results which worsened after discarding the volume cell feature.

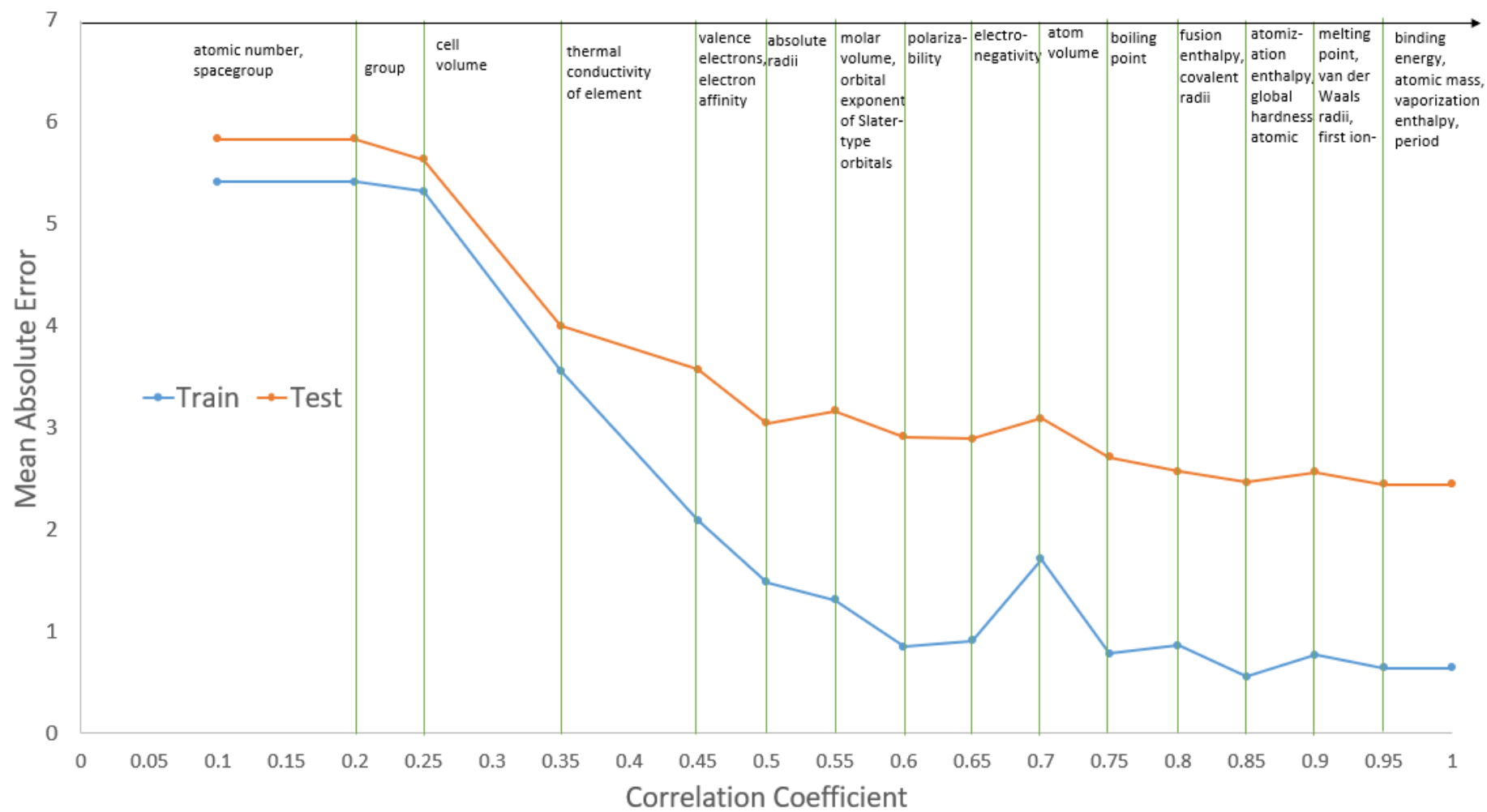


Figure 11 MAE function from correlation of dropped values for XGB algorithm. (Less is better) Without Mendelev Number.

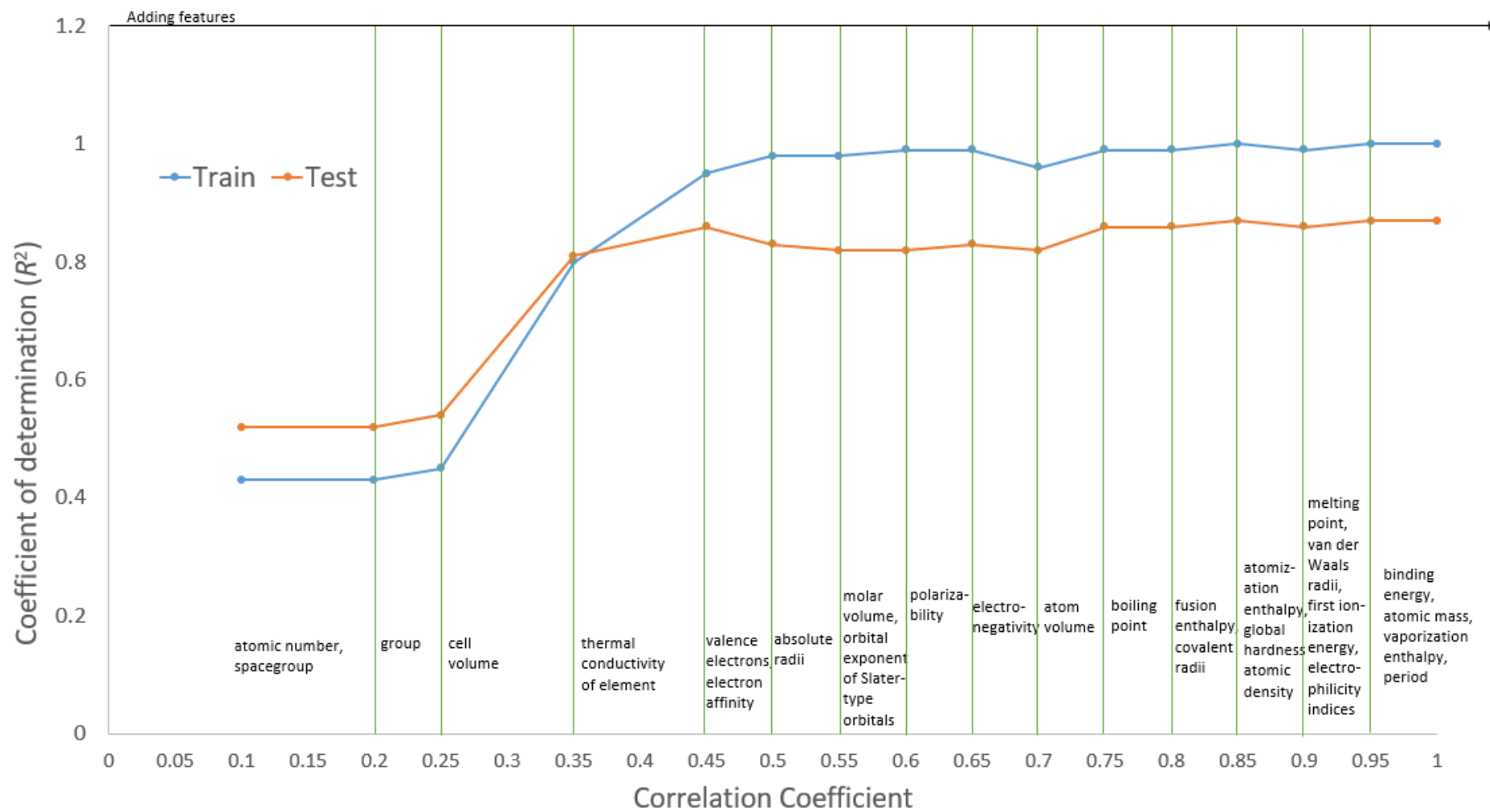


Figure 12 R^2 function from correlation of dropped values for XGB algorithm. (Higher is better) Without Mendeleev Number.

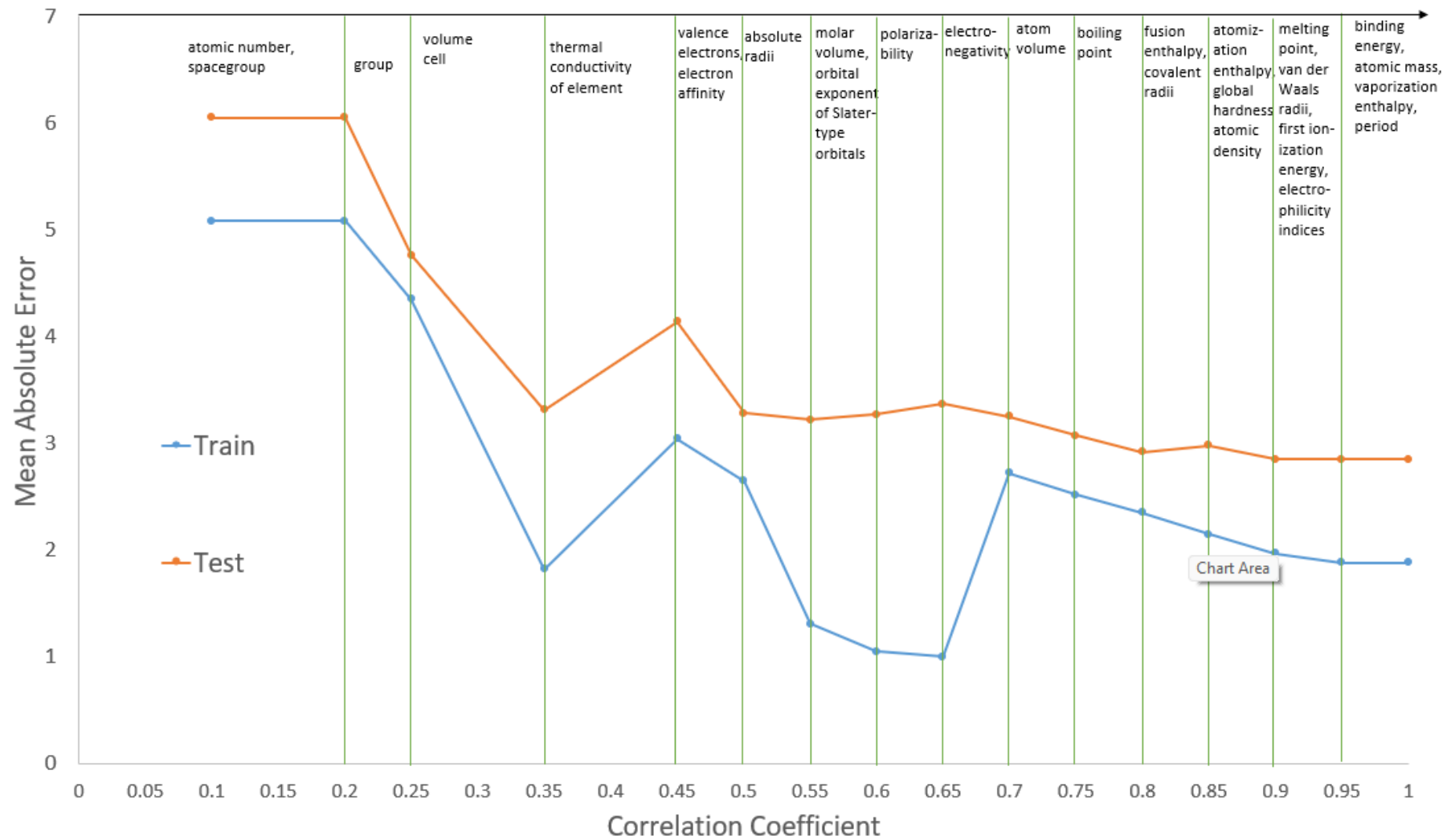


Figure 13 MAE function from correlation of dropped values for KRR algorithm. (Less is better) Without Mendeleev Number.

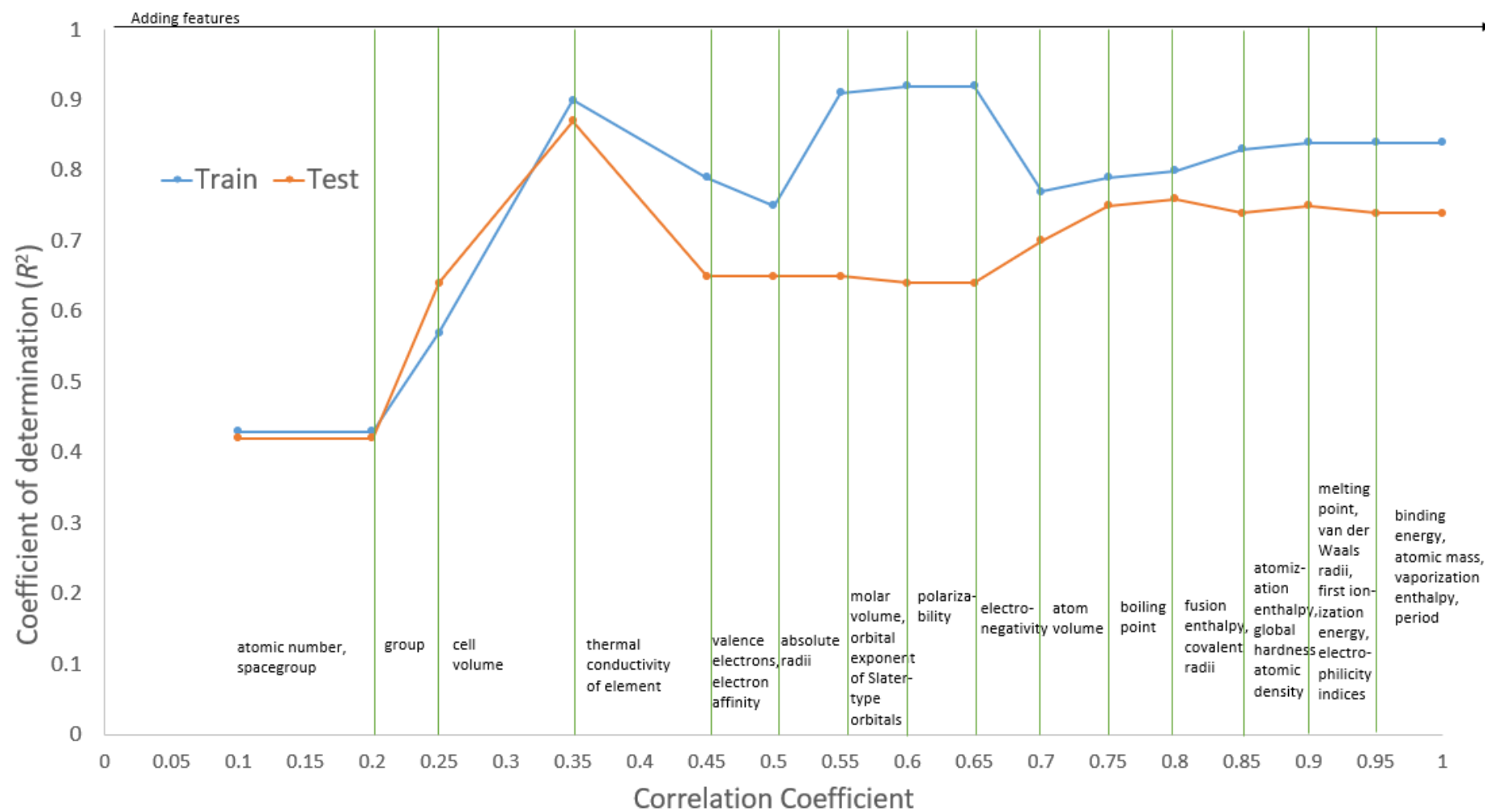


Figure 14 R^2 function from correlation of dropped values for KRR algorithm. (Higher is better) Without Mendeleev Number

After removing the Mendeleev's Number, the results did not turn out to be significantly worse. However, the trait "group" gained heavily (it moved toward low correlation values), and KPIs fell after its removal. It should also be noted that for KRR there again appeared some anomalies in the form of improving the results following the removal of a large number of features. The flow of the graphs for XGB is much closer to the version containing the Mendeleev Number.

After the calculations, it can be concluded that some features can be dropped without negatively affecting the accuracy of the calculations. I suggest that 14 features with a correlation of more than 75% with the rest of the data can be removed if the XGB algorithm is used. It is:

- boiling point,
- group,
- absolute radii,
- fusion enthalpy,
- covalent radii,
- atomization enthalpy,
- global hardness,
- atomic density,
- melting point,
- Van der Waals radii,
- first ionization energy,
- electrophilicity indices,
- binding energy,
- atomic mass,
- vaporization enthalpy,
- period

The charts with higher resolutions and containing mean absolute error can be found in Supplementary files.

7.4. Discussion of results

This article [45] uses similar algorithms and a sub-similar dataset, but both differ when it comes to feature validity inferences. They used two methods: the first involves quantifying the dataset

by changing one descriptor and keeping others invariant to observe the corresponding MAEs. The definition of it is as follows (Eq. 24):

$$p(x_k) = \frac{1}{l-1} \sum_{i=1}^{l-1} \frac{|\delta_{k,i+1} - \delta_{k,i}|}{\delta_{k,i}} \quad (\text{Eq. 24})$$

Where:

$p(x_k)$ - importance of x_k

x_k - corresponding MAEs.

And the results of this method are presented on Fig 15.

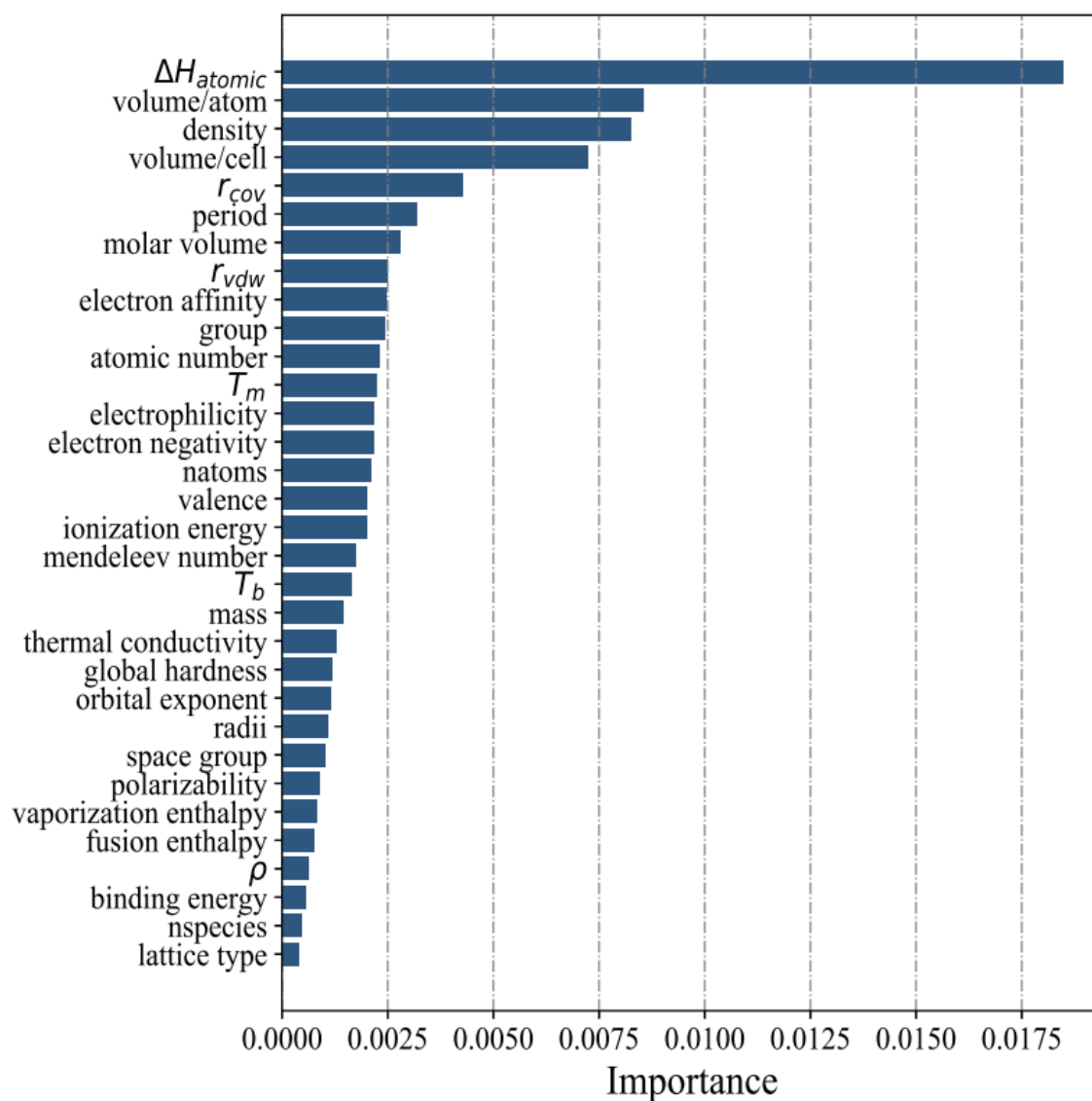


Figure 15 Feature importance for the 32 features of the XGB model [Supporting Information for [45]] (Higher is more important)

The second is based on the average gain across all splits using the given features. And the results are presented on Fig 16.

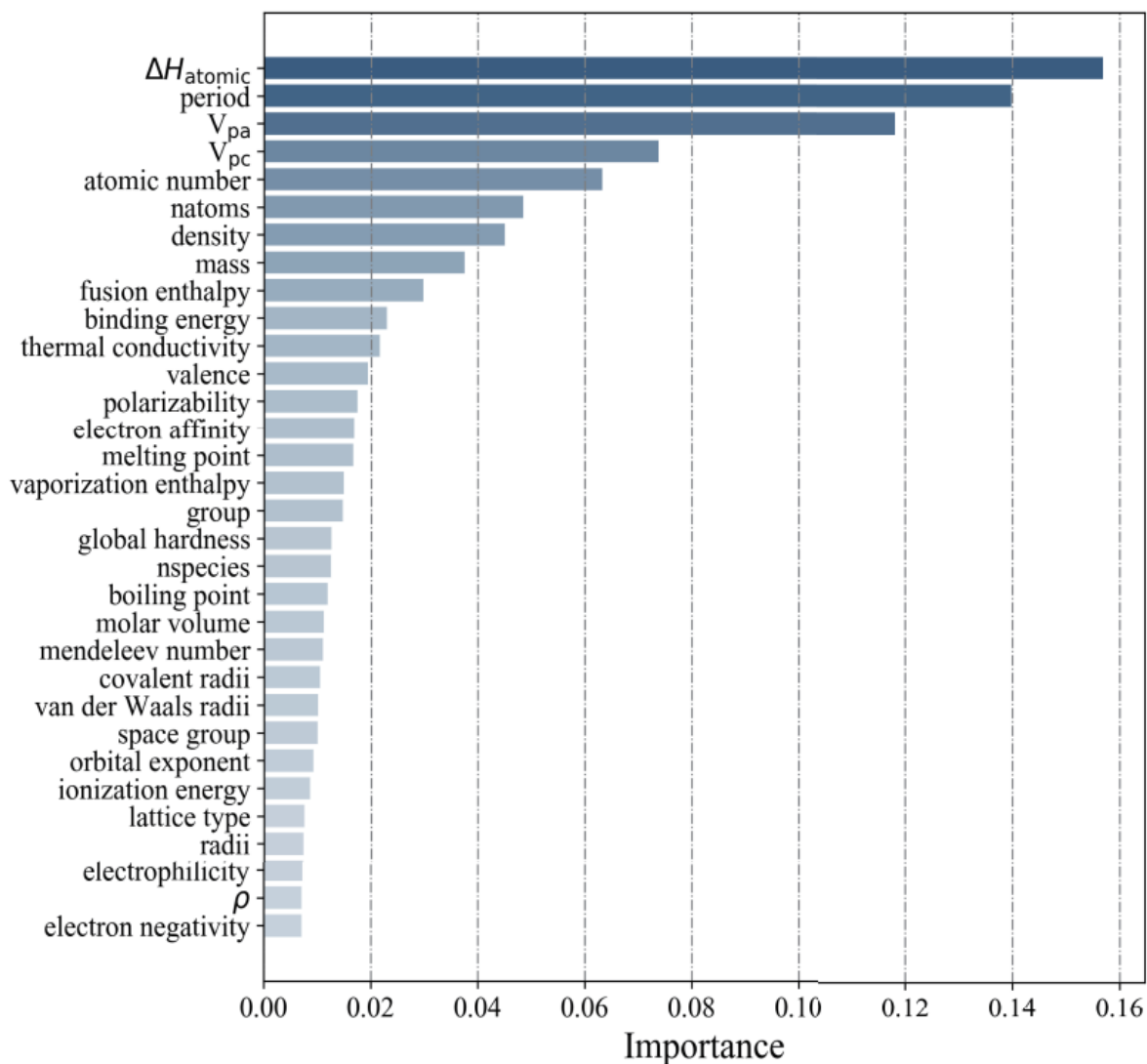


Figure 16 Feature importance for the 32 features of the XGBoost model. The definition of it is the average gain across all splits using the features. [Supporting Information for [45]] (Higher is more important)

After analysing these charts, one can conclude that the most important features according to that article are as follows:

- ΔH_{atomic}
- Cell Volume
- Atom Volume
- Period
- Atomic number

Contrary to the results of that paper this work determines the importance of the features after applying the method based on the elimination of highly correlated features (Fig 17).

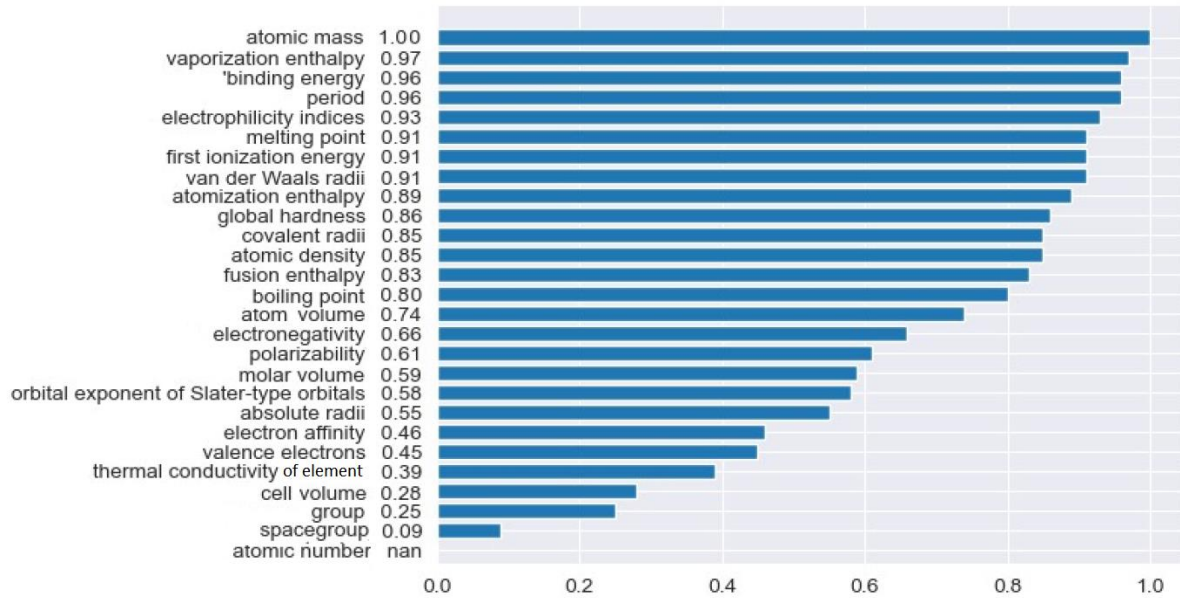


Figure 17 Feature correlation chart (A smaller value means a higher value for the algorithm)

According to the analysis based on the correlation of characteristics, one can fairly conclude that the most important are:

- Atomic number
- Space group
- Cell Volume
- Group
- Thermal conductivity of element

The most surprising difference is that Wang. et Al. asserts that the average enthalpy of the atoms in a compound constitutes its most important feature. That article noted that the characteristic of the average thermal conductivity of atoms in an elementary cell and space group is quite important for the thermal conductivity.

Some of the characteristics listed above have already been identified by other researchers as affecting thermal conductivity, e.g. by Rong Sun and Mary Anne White [2, p. 239].

The effect of elemental cell volume on the thermal conductivity of a material is not surprising. In the Green-Kubo method (Eq. 14), it is one of the components of the formula for thermal conductivity, according to which the larger the volume, the lower the conductivity is. One can try to explain this by the fact that in larger cells there is a scattering of phonons which are carriers of thermal energy and thus the thermal conductivity decreases.

The space group also seems to be quite important for thermal conductivity. It may turn out that a particular arrangement of atoms in space may be related to an effect on the propagation of phonons, which may be related to an effect on thermal conductivity.

High atomic mass has also been pointed out by Rong Sun and Mary Anne White [2, p. 239] as influencing low thermal conductivity. This may be due to the greater energy it takes to excite vibrations in a heavy atom.

The group of elements affects the volume of the atom which can affect the length of the bonds. Longer bonds can adversely affect the spread of heat carriers. On top of that, a larger atomic volume can affect a larger cell volume as described above.

It seems puzzling that thermal conductivity is also affected by the thermal conductivity of the elements in the structure. Mostly, the mechanism of conductivity in materials with metallic structures is quite different from that in the compounds studied. However, in the KRR algorithm, after removing Mendeleev number, one can see a deterioration of KPI after adding the thermal conductivity of the elements when with XGB one can see a slight increase. This may suggest that, however, this feature is not that important and is simply low correlated with the other descriptors, but does not affect the low conductivity of the compound.

7.5. Laboratory data

In Thermoelectric Research Laboratory at AGH University of Science and Technology, advanced research on thermoelectric modules was conducted. The conductive properties of materials were measured. The result for the experimental data for the best algorithm (XGB) is shown in Fig 18. R^2 score is 0.12 and an MAE is 5.72 W/(m·K) and 65% of samples were within the bounds of this error.

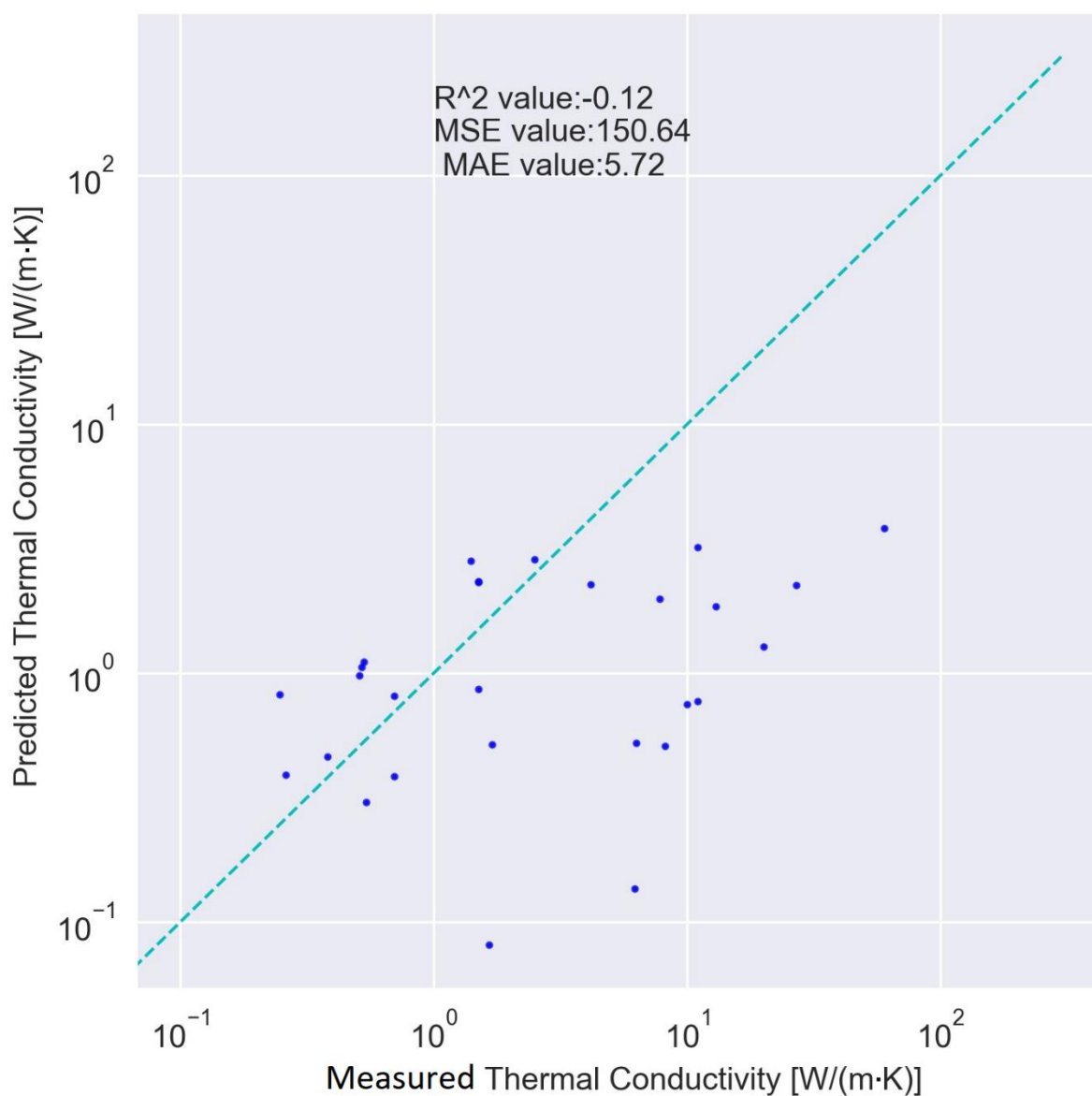


Figure 18 Predicted kl vs Measured Thermal Conductivity for experimental data, XGB algorithm.

Selected compounds with the values both measured and predicted from the best algorithm are presented in Table 3. The values in green have an error smaller than the MAE and the values in orange have a larger one.

Table 3 Low thermal conductivity compounds measured in TRL compared with predicted by ML

Name	Thermal Conductivity measured κ_{exp} [W/(m·K)]	Thermal Conductivity predicted κ_{pred} [W/(m·K)]	Absolute error Δ [W/(m·K)]
Ag ₉ AlSe ₆	0.29	-1.57	1.86
CaGa ₆ Te ₁₀	0.53	-0.54	1.07
Cu ₂ CoTi ₃ S ₈	1.40	1.13	0.27
Cu ₂ Se	0.54	0.63	0.09
Cu ₂ Te	0.70	0.08	0.62
Cu ₇ PS ₆	0.24	0.38	0.14
Cu ₈ GeSe ₆	0.24	-1.22	1.46
Cu ₈ SiSe ₆	0.26	0.19	0.07
CuCo ₂ S ₄	1.50	0.81	0.69
Ge	60.00	0.44	59.56
PbGa ₆ Te ₁₀	0.51	0.64	0.13
NaAgGa ₆ Te ₁₀	0.25	0.07	0.18
PbIn ₆ Te ₁₀	0.38	-1.42	1.80
SnGa ₆ Te ₁₀	0.52	-0.42	0.93
ZnTe	11.00	-0.38	11.38
ZnSe	13.00	0.51	12.49
ZnS	27.00	1.19	25.81
Cu ₂ CoSnSe ₄	1.50	1.95	0.45

Cu ₂ CoSnS ₄	4.18	1.55	2.63
SnSe	1.50	2.09	0.59
Mg ₂ Si	11.00	1.18	9.82
CuTi ₂ S ₄	2.50	1.52	0.98
CuInTe ₂	6.20	1.26	4.94
CuGaTe ₂	8.20	-0.53	8.73
CuFeS ₂	7.80	1.78	6.02
Cu ₂ HgGeTe ₄	1.70	-0.26	1.96
CdTe	6.30	-0.64	6.94
CdSe	10.00	-0.33	10.33
CdS	20.00	0.53	19.47
AgSbTe ₂	0.70	0.46	0.24
AgGaTe ₂	1.66	-0.80	2.46

It can be noted that some of the values predicted by the algorithm are negative which is impossible to achieve under laboratory conditions, but often the error is not very large. In addition, in more than half of the cases, the error is small and does not exceed the MAE. The biggest difference can be noticed for the Germanium structure. To understand why the algorithm resulted in such large errors for some of the compounds, histograms (Fig. 19-20) were created for the 5 most important features that emerged through the multivariate reduction from the previous chapter.

The underestimation of germanium's relatively high thermal conductivity may be due to its semi-metallic properties and increased electron conductivity, which the algorithm may not take into account.

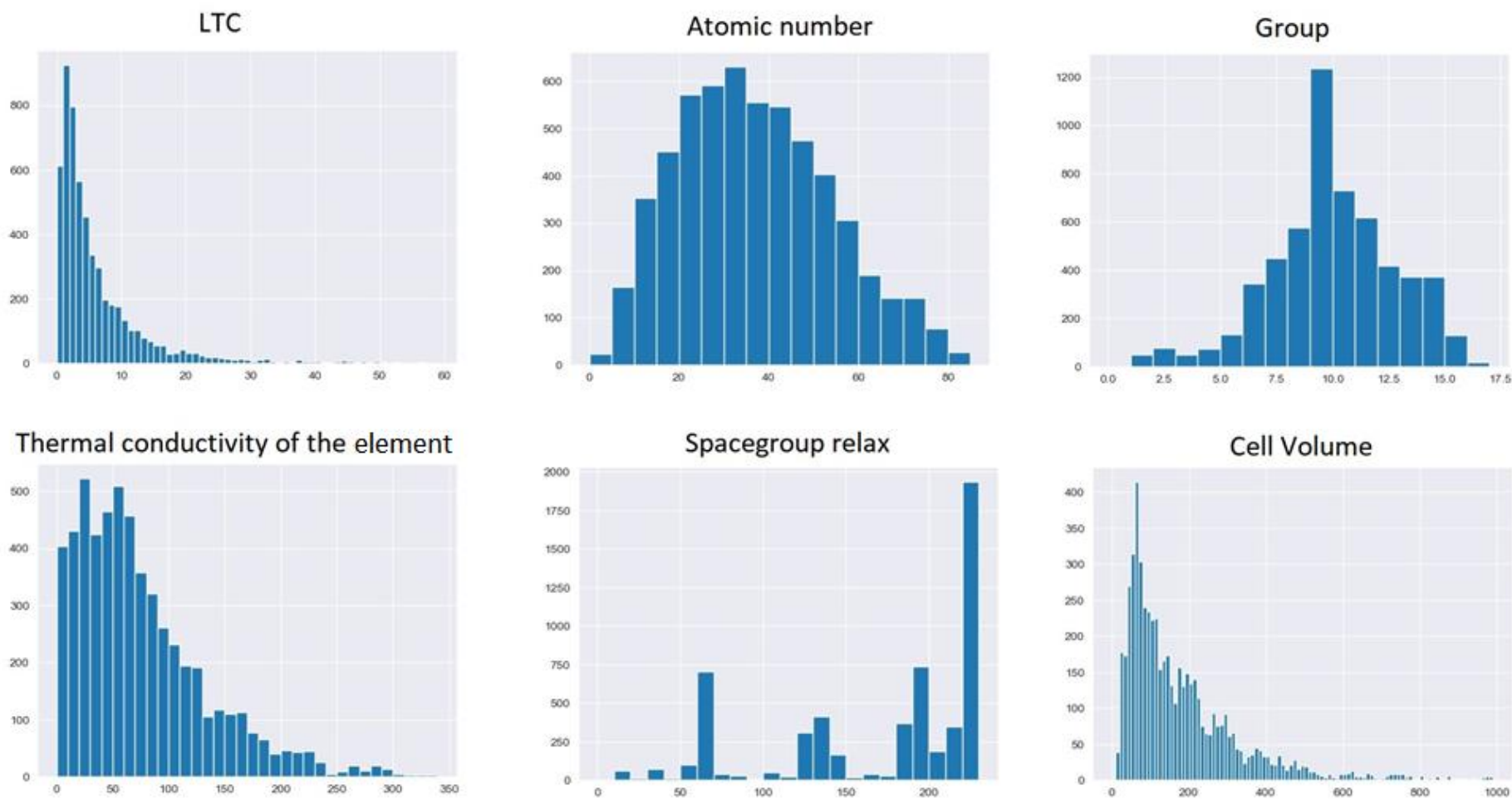
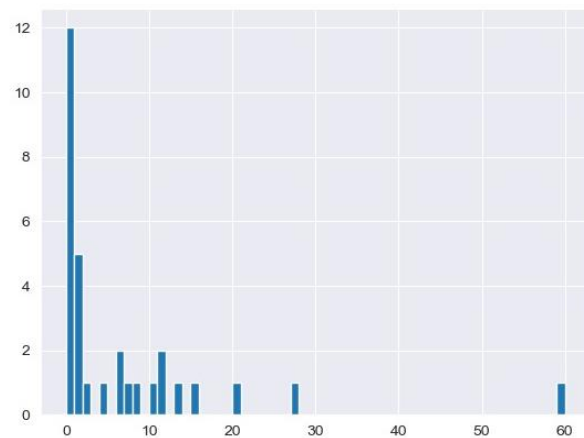
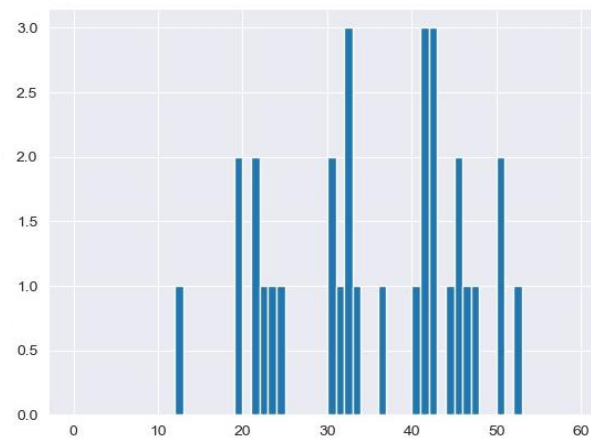


Figure 19 Histogram showing data from AFLOW

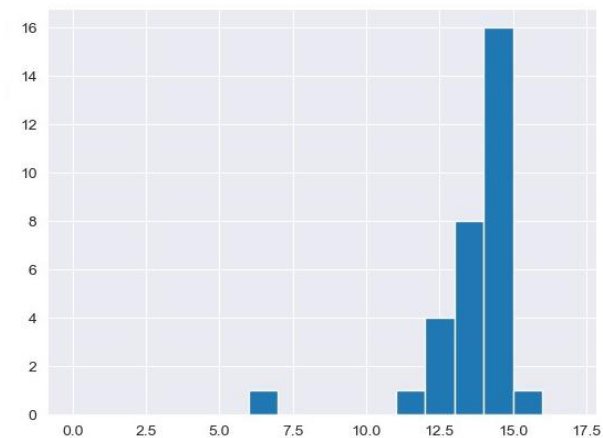
Mesured thermal conductivity



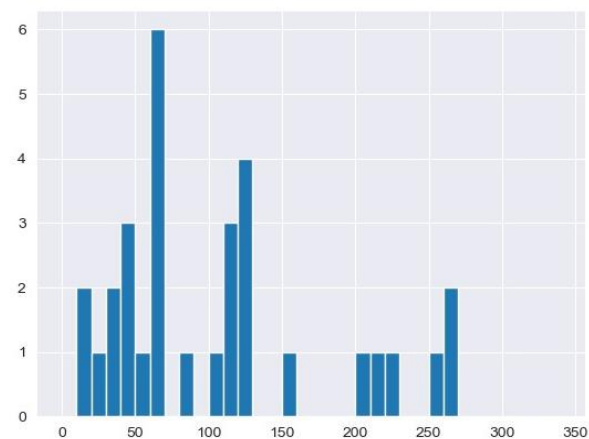
Atomic number



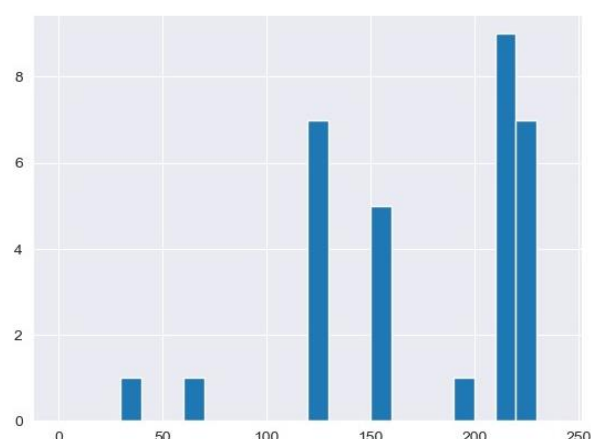
Group



Thermal conductivity of the compound



Spacegroup relax



Cell Volume

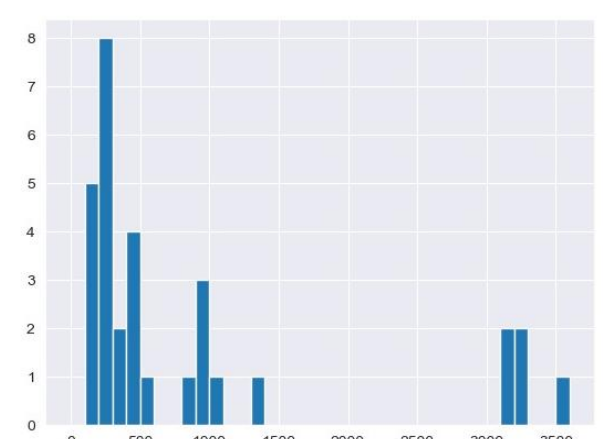


Figure 20 Histogram showing experimental data

In the experimental data set, there are elementary cells more than almost 3 times larger than the largest in the training set. In addition, there are relatively more samples from the group 15 in the experimental set than in the training set. To see if the trained algorithm does accurately predict the actual thermal conductivity of the materials, a larger experimental data set would need to be created. Additionally, one might consider not including the structures consisting of only one type of atoms.

8. Summary

In summary, the machine learning models such as SVR, KRR, XGBoost, and LASSO were used to predict the thermal conductivity of crystal materials based on a set of universal descriptors of materials. The performance of different machine learning models based on different types of descriptors using a subset of the complete set of 28 descriptors was tested. The correlation coefficient analysis shown for XGB proved to be the most important of the available algorithms.

For selected machine learning algorithms, the following features have proven to be the most important using multifactor dimensionality reduction:

- **Atomic number**
- **Group**
- **Space group**
- **Thermal conductivity of compound**
- **Cell Volume**

Attempts to calculate thermal conductivity for compounds tested in the laboratory have been unsuccessful for about 1/3 of the dataset. Inefficiencies may be due to the small data set and differences between the test and experimental sets. Further research is required.

In summary, the results indicate that the machine learning method could be useful for guiding the search for potential thermoelectric materials. However, it requires some refinement. Perhaps the use of other hyperparameters, neural networks, or descriptors that describe the structure more accurately than the space group itself, could improve the accuracy of the predictions.

Once refined, one can try searching huge databases of data like Materials Project or AFLOW to find the materials with low thermal conductivity that may have other applications e.g., in the search for new thermoelectric materials.

9. Supplementary Files


Files containing high-resolution graphs, data for calculations, and source code can be found at

<https://github.com/HaltRiv/>

10. Bibliography

- [1] ‘What is Fourier’s Law of Thermal Conduction - Definition’, *Thermal Engineering*, May 22, 2019. <https://www.thermal-engineering.org/what-is-fouriers-law-of-thermal-conduction-definition/> (accessed May 31, 2023).
- [2] T. M. Tritt, Ed., *Thermal conductivity: theory, properties, and applications*. in Physics of solids and liquids. New York: Kluwer Academic/Plenum Publishers, 2004.
- [3] W. Jones, *Theoretical solid state physics*. New York: Dover Publications, 1985. Accessed: May 31, 2023. [Online]. Available: <http://archive.org/details/theoreticalsolid0001jone>
- [4] J. Callaway, ‘Model for Lattice Thermal Conductivity at Low Temperatures’, *Phys. Rev.*, vol. 113, no. 4, pp. 1046–1051, Feb. 1959, doi: 10.1103/PhysRev.113.1046.
- [5] D. T. Morelli, J. P. Heremans, and G. A. Slack, ‘Estimation of the isotope effect on the lattice thermal conductivity of group IV and group III-V semiconductors’, *Phys. Rev. B*, vol. 66, no. 19, p. 195304, Nov. 2002, doi: 10.1103/PhysRevB.66.195304.
- [6] W. Kohn, A. D. Becke, and R. G. Parr, ‘Density Functional Theory of Electronic Structure’, *J. Phys. Chem.*, vol. 100, no. 31, pp. 12974–12980, Jan. 1996, doi: 10.1021/jp960669l.
- [7] ‘VASP - Vienna Ab initio Simulation Package’. <https://www.vasp.at/> (accessed Jun. 01, 2023).
- [8] ‘Home Page’, *Quantum Espresso*. <https://www.quantum-espresso.org/> (accessed Jun. 01, 2023).
- [9] A. J. H. McGaughey, A. Jain, H.-Y. Kim, and B. (傅博) Fu, ‘Phonon properties and thermal conductivity from first principles, lattice dynamics, and the Boltzmann transport equation’, *J. Appl. Phys.*, vol. 125, no. 1, p. 011101, Jan. 2019, doi: 10.1063/1.5064602.
- [10] J. Kang and L.-W. Wang, ‘A First-Principles Green-Kubo Method for Thermal Conductivity Calculation’.
- [11] T. Ikeshoji and B. Hafskjold, ‘Non-equilibrium molecular dynamics calculation of heat conduction in liquid and through liquid-gas interface’, *Mol. Phys.*, vol. 81, no. 2, pp. 251–261, Feb. 1994, doi: 10.1080/00268979400100171.
- [12] ‘First-principles Debye–Callaway approach to lattice thermal conductivity’, *J. Materiomics*, vol. 2, no. 3, pp. 237–247, Sep. 2016, doi: 10.1016/j.jmat.2016.06.004.
- [13] ‘Thermal Conductivity of Argon from the Green-Kubo Method’. <https://courses.physics.illinois.edu/phys466/sp2013/projects/2004/Team1/index.html> (accessed May 29, 2023).
- [14] M. A. Blanco, E. Francisco, and V. Luaña, ‘GIBBS: isothermal-isobaric thermodynamics of solids from energy curves using a quasi-harmonic Debye model’, *Comput. Phys. Commun.*, vol. 158, no. 1, pp. 57–72, Mar. 2004, doi: 10.1016/j.comphy.2003.12.001.
- [15] ‘Aflow - Automatic FLOW for Materials Discovery’. <https://www.aflow.org/> (accessed Mar. 12, 2023).
- [16] ‘Materials Project - Home’, *Materials Project*. <https://materialsproject.org/> (accessed Apr. 18, 2023).
- [17] C. Toher *et al.*, ‘High-throughput computational screening of thermal conductivity, Debye temperature, and Grüneisen parameter using a quasiharmonic Debye model’, *Phys. Rev. B*, vol. 90, no. 17, p. 174107, Nov. 2014, doi: 10.1103/PhysRevB.90.174107.

- [18] X. Zhang, S. Sun, T. Xu, and T. Zhang, ‘Temperature dependent Grüneisen parameter’, *Sci. China Technol. Sci.*, vol. 62, no. 9, pp. 1565–1576, Sep. 2019, doi: 10.1007/s11431-019-9526-3.
- [19] ‘Debye Model For Specific Heat’, *Engineering LibreTexts*, Jul. 28, 2016. [https://eng.libretexts.org/Bookshelves/Materials_Science/Supplemental_Modules_\(Materials_Science\)/Electronic_Properties/Debye_Model_For_Specific_Heat](https://eng.libretexts.org/Bookshelves/Materials_Science/Supplemental_Modules_(Materials_Science)/Electronic_Properties/Debye_Model_For_Specific_Heat) (accessed Apr. 18, 2023).
- [20] T. Parashchuk, Z. Dashevsky, and K. Wojciechowski, ‘Feasibility of a high stable PbTe:In semiconductor for thermoelectric energy applications’, *J. Appl. Phys.*, vol. 125, no. 24, p. 245103, Jun. 2019, doi: 10.1063/1.5106422.
- [21] K. Mondal, L. I. Nuñez, C. M. Downey, and I. J. van Rooyen, ‘Thermal Barrier Coatings Overview: Design, Manufacturing, and Applications in High-Temperature Industries’, *Ind. Eng. Chem. Res.*, vol. 60, no. 17, pp. 6061–6077, May 2021, doi: 10.1021/acs.iecr.1c00788.
- [22] ‘The state of AI in 2022—and a half decade in review | McKinsey’. <https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2022-and-a-half-decade-in-review> (accessed Jun. 02, 2023).
- [23] Y. Liu, T. Zhao, W. Ju, and S. Shi, ‘Materials discovery and design using machine learning’, *J. Materiomics*, vol. 3, no. 3, pp. 159–177, Sep. 2017, doi: 10.1016/j.jmat.2017.08.002.
- [24] J. Fang *et al.*, ‘Machine learning accelerates the materials discovery’, *Mater. Today Commun.*, vol. 33, p. 104900, Dec. 2022, doi: 10.1016/j.mtcomm.2022.104900.
- [25] R. Ramprasad, R. Batra, G. Piliand, A. Mannodi-Kanakkithodi, and C. Kim, ‘Machine learning in materials informatics: recent applications and prospects’, *Npj Comput. Mater.*, vol. 3, no. 1, Art. no. 1, Dec. 2017, doi: 10.1038/s41524-017-0056-5.
- [26] ‘Data Preprocessing in Machine Learning: 7 Easy Steps To Follow | upGrad blog’. <https://www.upgrad.com/blog/data-preprocessing-in-machine-learning/> (accessed Jun. 02, 2023).
- [27] J. Brownlee, ‘Introduction to Dimensionality Reduction for Machine Learning’, *MachineLearningMastery.com*, May 05, 2020. <https://machinelearningmastery.com/dimensionality-reduction-for-machine-learning/> (accessed Jun. 02, 2023).
- [28] ‘How to scale data for Machine Learning: standardize features - Marcos del Cueto’, Oct. 22, 2021. <https://www.mdeldcueto.com/blog/scale-data-for-machine-learning-standardize-features/> (accessed Jun. 02, 2023).
- [29] ‘Regression Algorithms: which Machine Learning Metrics?’, *MyDataModels*, Oct. 21, 2020. <https://www.mydatamodels.com/blog/regression-metrics/> (accessed May 27, 2023).
- [30] ‘sklearn.metrics.r2_score — scikit-learn 1.2.2 documentation’. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html (accessed May 27, 2023).
- [31] ‘10 Machine Learning Algorithms to Know in 2023’, *Coursera*, May 17, 2023. <https://www.coursera.org/articles/machine-learning-algorithms> (accessed Jun. 02, 2023).
- [32] A. Anwar, ‘A Beginner’s Guide to Regression Analysis in Machine Learning’, *Medium*, Jun. 07, 2021. <https://towardsdatascience.com/a-beginners-guide-to-regression-analysis-in-machine-learning-8a828b491bbf> (accessed Jun. 02, 2023).
- [33] ‘Machine Learning Algorithms - Javatpoint’, *www.javatpoint.com*. <https://www.javatpoint.com/machine-learning-algorithms> (accessed Jun. 02, 2023).

- [34] ‘Lasso Regression Explained, Step by Step’. https://machinelearningcompass.com/machine_learning_models/lasso_regression/ (accessed Jun. 05, 2023).
- [35] ‘sklearn.linear_model.Lasso’, *scikit-learn*. https://scikit-learn/stable/modules/generated/sklearn.linear_model.Lasso.html (accessed Jun. 05, 2023).
- [36] T. Sharp , ‘An Introduction to Support Vector Regression (SVR)’, *Medium*, Apr. 03, 2023. <https://towardsdatascience.com/an-introduction-to-support-vector-regression-svr-a3ebc1672c2> (accessed Jun. 02, 2023).
- [37] ‘LIBSVM -- A Library for Support Vector Machines’. <https://www.csie.ntu.edu.tw/~cjlin/libsvm/> (accessed Jun. 23, 2023).
- [38] ‘sklearn.kernel_ridge.KernelRidge’, *scikit-learn*. https://scikit-learn/stable/modules/generated/sklearn.kernel_ridge.KernelRidge.html (accessed Apr. 19, 2023).
- [39] S. Dobilas, ‘XGBoost: Extreme Gradient Boosting — How to Improve on Regular Gradient Boosting?’, *Medium*, Feb. 05, 2022. <https://towardsdatascience.com/xgboost-extreme-gradient-boosting-how-to-improve-on-regular-gradient-boosting-5c6acf66c70a> (accessed May 27, 2023).
- [40] Y. Luo, M. Li, H. Yuan, H. Liu, and Y. Fang, ‘Predicting lattice thermal conductivity via machine learning: a mini review’, *Npj Comput. Mater.*, vol. 9, no. 1, Art. no. 1, Jan. 2023, doi: 10.1038/s41524-023-00964-2.
- [41] C. Toher *et al.*, ‘Combining the AFLOW GIBBS and elastic libraries to efficiently and robustly screen thermomechanical properties of solids’, *Phys. Rev. Mater.*, vol. 1, no. 1, p. 015401, Jun. 2017, doi: 10.1103/PhysRevMaterials.1.015401.
- [42] Z. Allahyari and A. R. Oganov, ‘Nonempirical Definition of the Mendeleev Numbers: Organizing the Chemical Space’, *J. Phys. Chem. C*, vol. 124, no. 43, pp. 23867–23878, Oct. 2020, doi: 10.1021/acs.jpcc.0c07857.
- [43] S. Curtarolo *et al.*, ‘AFLOWLIB.ORG: A distributed materials properties repository from high-throughput ab initio calculations’, *Comput. Mater. Sci.*, vol. 58, pp. 227–235, Jun. 2012, doi: 10.1016/j.commatsci.2012.02.002.
- [44] W. Wang, ‘Bayesian Optimization Concept Explained in Layman Terms’, *Medium*, Mar. 22, 2022. <https://towardsdatascience.com/bayesian-optimization-concept-explained-in-layman-terms-1d2bcdeaf12f> (accessed Jul. 03, 2023).
- [45] X. Wang, S. Zeng, Z. Wang, and J. Ni, ‘Identification of Crystalline Materials with Ultra-Low Thermal Conductivity Based on Machine Learning Study’, *J. Phys. Chem. C*, vol. 124, no. 16, pp. 8488–8495, Apr. 2020, doi: 10.1021/acs.jpcc.9b11610.
- [46] T. Zhu *et al.*, ‘Charting lattice thermal conductivity for inorganic crystals and discovering rare earth chalcogenides for thermoelectrics’, *Energy Environ. Sci.*, vol. 14, no. 6, pp. 3559–3566, Jun. 2021, doi: 10.1039/D1EE00442E.
- [47] R. Juneja, G. Yumnam, S. Satsangi, and A. Singh, ‘Coupling High-throughput Property Map to Machine Learning for Predicting Lattice Thermal Conductivity’, *Chem. Mater.*, vol. 31, Jun. 2019, doi: 10.1021/acs.chemmater.9b01046.
- [48] D. Hicks *et al.*, ‘AFLOW-SYM: Platform for the complete, automatic and self-consistent symmetry analysis of crystals’. arXiv, Feb. 22, 2018. doi: 10.48550/arXiv.1802.07977.
- [49] J. Brownlee, ‘Overfitting and Underfitting With Machine Learning Algorithms’, *MachineLearningMastery.com*, Mar. 20, 2016. <https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/> (accessed Apr. 18, 2023).