

**DISCOVERY OF THERMOELECTRIC INORGANIC HALIDE
PEROVSKITES MATERIALS USING MACHINE LEARNING**

BY

HASSAN ALI ALSALMAN

A Thesis Presented to the
DEANSHIP OF GRADUATE STUDIES

KING FAHD UNIVERSITY OF PETROLEUM & MINERALS

DHAHRAN, SAUDI ARABIA

In Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

In

ELECTRICAL ENGINEERING

April 2021

**DISCOVERY OF THERMOELECTRIC INORGANIC HALIDE
PEROVSKITES MATERIALS USING MACHINE LEARNING**

Hassan Alsalman

ELECTRICAL ENGINEERING

April 2021

KING FAHD UNIVERSITY OF PETROLEUM & MINERALS

DHAHRAN- 31261, SAUDI ARABIA

DEANSHIP OF GRADUATE STUDIES

This thesis, written by Hassan Als Salman under the direction of his thesis advisor and approved by his thesis committee, has been presented and accepted by the Dean of Graduate Studies, in partial fulfillment of the requirements for the degree of **MASTER OF SCIENCE IN ELECTRICAL ENGINEERING**.



Prof. Fahhad Alharbi
(Advisor)



Dr. Abdallah Al-Ahmari
Department Chairman



Prof. Ali Muqaibel
(Member)



Dr. Suliman Saleh Al-Homidan
Dean of Graduate Studies



Dr. Abduljabar Alsayoud
(Member)

4/7/2021

Date



© Hassan Alsalman

2021

To those who live in my heart |

ACKNOWLEDGMENTS

I thank Allah for the grace he bestowed upon me, for the privilege of being able to continue my education and to pursue a life dream of mine.

I would like to express my gratitude to my advisor Prof. Fahhad Alharbi for his continuous guidance, his teachings, patient, time, and kindness. My gratitude to the committee members Prof. Ali Muqaibel and Dr. Abduljabar Alsayoud for their insightful comments, which helped me creating a better version of my work. I also like to thank the help offered by the Computational Group members, Dr.Saad Alqahtani and Ibrahim Alsayoud for their insightful contributions in my work. To Dr.Saad Al-Ahamdi , Dr. Samir Al-Ghadban, Dr. Azzeddine Zerguine, Dr. Sharif Sheikh and Dr. Maohamemd Sharwi for their encouragement to pursue this degree. To my friends Eng.Adulellah Alhanabi. Eng. Ali Alkhalifa and Abdulmonem Almohaini, thanks for the most joyous of times and the support in my academic years.

Finally, most sincerities of gratitude for my grandparents, parents, and my brothers (Hashem, Qasim and Husien). Their support is indescribable, time and time again they motivate me to do my best. |

|

TABLE OF CONTENTS

ACKNOWLEDGMENTS	V
TABLE OF CONTENTS.....	VI
LIST OF TABLES.....	IX
LIST OF FIGURES.....	X
LIST OF ABBREVIATIONS.....	XI
ABSTRACT	XII
ملخص الرسالة	XIV
CHAPTER 1 INTRODUCTION.....	1
1.1 Thermoelectric Effect	1
1.2 Motivation	2
1.3 Objectives	3
1.4 Thesis Structure	4
CHAPTER 2 OVERVIEW OF THERMOELECTRICITY AND ML APPROACHES.....	6
2.1 Overview of Thermoelectrical Materials	6
2.2 Data-Driven Modelling.....	9
2.2.1 Linear Regression	10
2.2.2 Regularization	12
2.3 Machine Learning for Thermoelectricity	13
2.3.1 TE Effect Features Selection	13
2.3.2 TE material discovery using ML Classification.....	17
2.3.3 Predicting the $\mathbb{Z}T$ using ML	18

CHAPTER 3 PREDICTING THERMAL CONDUCTIVITY	20
3.1 Introduction	21
3.2 Data	23
3.2.1 Data Filtering	24
3.2.2 Feature Analysis.....	25
3.3 Models for Lattice Thermal Conductivity	28
3.4 Results and Discussion	29
3.5 Conclusion	31
 CHAPTER 4 ML MODELS FOR ELECTRICAL TRANSPORTS	 33
4.1 Introduction	34
4.2 Data	35
4.3 Electrical Conductivity.....	38
4.3.1 Feature Analysis.....	38
4.3.2 Modeling the Electrical Conductivity.....	39
4.3.3 Results and Discussion	40
4.4 Seebeck Coefficient.....	41
4.4.1 Feature Analysis.....	41
4.4.2 Modeling the Seebeck Coefficient.....	44
4.4.3 Results and Discussion	46
4.5 Conclusion	47
 CHAPTER 5 $\mathbb{Z}T$ PREDICTION.....	 48
5.1 Gathering Data.....	48
5.2 Results	49

CHAPTER 6 CONCLUSION AND FUTURE WORK	53
6.1 Conclusion	53
6.2 Suggestions for Future Work.....	54
REFERENCES.....	56
VITAE	62

LIST OF TABLES

Table 3-1 Considered space groups	25
Table 3-2 Results summary of training and testing for the generalized models.....	29
Table 3-3 Results summary of training and testing for all models.	30
Table 3-4 Considered space groups	36
Table 5-1 The prediction of electrical and thermal transports at $n=10^{20}/\text{cm}^3$	50
Table 5-2 The prediction of electrical and thermal transports at $n=10^{18}/\text{cm}^3$	51

LIST OF FIGURES

Fig 2-1 Electrical and thermal transport schematic for carrier concentration 10^{18} - 10^{22} cm ⁻³	7
Fig 2-2 ML workflow	9
Fig 2-3 The considered features and descriptors used to develop a machine learning model(s) to estimate ZT	15
Fig 3-1 Data filtering process	24
Fig 3-2 Thermal conductivity vs. a) the bulk modulus, b) the shear modulus, c) Young's modulus, d) the speed of sound, e) the density, f) Poisson's ratio, g) the unit cell volume per atom, and h) the reciprocal of log of density multiplied by unit cell volume.....	27
Fig 3-3 Correlation coefficient matrix	27
Fig 4-1 Data filtering process	36
Fig 4-2 Electrical conductivity vs. a) the reciprocal of effective mass, b) the Bandgap, c) volume ,and d) mass density	38
Fig 4-3 Correlation matrix for σ	39
Fig 4-4 The predicted VS Actual values of the Electrical conductivity	40
Fig 4-5 Seebeck coefficient vs. a) the effective mass, b) the bandgap, c) volume/(number of sites/5), d) mass density, and e) the Pauling electronegativity of the B atom.	42
Fig 4-6 The correlation matrix of the $ S $	42
Fig 4-7 Schematic of Band Filling.....	43
Fig 4-8 N_V estimation process.....	45
Fig 4-9 The predicted Vs Actual values of the Seebeck coefficient.....	46

LIST OF ABBREVIATIONS

TE	:	Thermoelectric
DFT	:	Density functional theory
ML	:	Machine learning
DT	:	Decision tree
RF	:	Random forest
IHP	:	Inorganic halide perovskites
DOS	:	Density of state

ABSTRACT

Full Name : Hassan Ali Alsalman
Thesis Title : Discovery of Thermoelectric Inorganic Halide Perovskites Materials Using Machine Learning
Major Field : Electrical Engineering
Date of Degree : [April 2021]

In the path of finding efficient renewable energy sources, thermoelectricity is considered among the viable options with high potential. To realize it, there is a need to “find” new efficient thermoelectric materials. Currently, the most commonly used set of these materials has low efficiency. Such efficiency is assessed through a figure of merit ZT . The standard ways of estimating this figure are either through experiments or expensive atomic scale calculations, which makes the discovery of new materials extremely challenging. However, with the vast existing accessible data, the problem can be approached alternatively by machine learning. For a thermoelectric semiconductor, varying the temperature or carrier concentration introduces complex ZT behavior, besides the fact that many physical mechanisms will be involved at microscale level. By decomposing the ZT to its constituents and studying their behaviors at specific temperature and carrier concentration, a better understanding for these complex trends is acquired. By doing so, the required variables to describe the electrical and thermal transports of a thermoelectric are reduced. These considerations eventually make machine learning modeling a viable approach for this problem. In this work, we aim to develop highly accurate predictive models for a specific family of materials (inorganic halide perovskites). These materials exhibit an interesting combination of ultra-low thermal conductivity and high Seebeck

coefficient, which can lead to high ZT values. For these materials, the ZT constituents: electrical, thermal conductivities and Seebeck coefficient are predicted using elastic properties and some electronic transport features. From these parameters the ZT is predicted, which shall allow a quick screening mechanism to discover novel thermoelectric materials.

ملخص الرسالة

الاسم الكامل: حسن علي السلطان

عنوان الرسالة: اكتشاف مواد بيروفسكايت هالية غير عضوية و كهروحرارية باستخدام التعلم الآلي

التخصص: الهندسة الكهربائية

تاريخ الدرجة العلمية: رمضان 1442هـ

خلال السنوات الأخيرة، كثرت البحوث عن مصادر فعالة للطاقة متجددة. و تعتبر مصادر الطاقة الكهروحرارية أحد الخيارات الواعد لما تقدمه من الإمكانيات العالية والتطبيقات المتعددة. مصادر الطاقة الكهروحرارية تعتمد على المواد الكهروحرارية المستخدمة في عملية توليد الطاقة، مما خلق الحاجة إلى إيجاد مواد كهرو حرارية جديدة ذات كفاءة عالية. بينت البحوث الحالية أن أغلب المواد الكهرو حرارية تعتبر موادا منخفضة الكفاءة. و يتم تقييم هذه الكفاءة من خلال معيار تم استخدامه من العلماء و يسمى (ZT) figure of merit. الطرق القياسية لتقدير هذا الرقم تكون إما من خلال التجارب العملية أو عن طريق الحسابات الذرية المكلفة. مما يجعل اكتشاف مواد كهروحرارية جديدة أمراً صعباً للغاية. و لكن مع وجود البيانات الهائلة المتاحة عن خصائص المواد الغير العضوية، يمكن التعامل معها عن طريق تحليل البيانات واستخدام التعلم الآلي كحل بديل.

في هذا العمل، نهدف إلى تطوير معادلات للتنبؤ بمعيار الكفاءة ZT بدقة عالية لعائلة معينة من المواد الكهروحرارية (بيروفسكايت هالية غير عضوية). لتطوير هذه المعادلات، سوف نتطرق إلى تحليل معيار الكفاءة و علاقته بخصائص هذه المواد الكهربائية و الحرارية و خصائص المرونة باستخدام التعلم الآلي. لأن هذه المواد تنتمي لأشباه الموصلات، سوف يتم دراسة هذه المواد تحت مستوى موحد لكثافة الناقلات الكهربائية و عند درجة حرارة الغرفة. استخدام هذه المعادلات سيسمح بتطوير آلية مسح سريعة لاكتشاف مواد كهروحرارية جديدة.

CHAPTER 1

INTRODUCTION

1.1 Thermoelectric Effect

In general, thermal energy is a by-product of any physical operating system, often thought of as waste. One way to enhance overall system efficiency is to utilize this inevitably existing energy. This led to the idea of thermoelectric (TE) generators. Basically, when two different connected conductors are subjugated to a temperature gradient, current will flow, and this phenomenon is known as the Seebeck effect [1]. This effect is quantified by the Seebeck coefficient (S), which is:

$$S = -\frac{\Delta V_{Th}}{\Delta T} \quad (1.1)$$

where T is the temperature and V_{Th} is the induced thermoelectric voltage. Conceptually, this coefficient is phenomenological and does not describe or quantify a material TE performance. Thus, it is not a proper measure of potentiality and there is a need for another “measure” that better describes the applicability and efficiency of a TE material [2-4]. The standardly used “measure” is the TE figure of merit and it is

$$\mathbb{Z}T = \frac{S^2 \sigma T}{k} \quad (1.2)$$

where k is thermal conductivity and σ is electrical conductivity. From Eq. 2, it is clear that to enhance $\mathbb{Z}T$, σ must be large and/or k must be small; physically, this means maintaining the temperature gradient as long as possible. However, tuning these parameters independently is extremely difficult because they are highly correlated [5].

In order to search for efficient TE materials, we need to focus on the three main parameters: namely, the electrical, Seebeck coefficient and thermal conductivities. These parameters can be measured experimentally for known material or estimated computationally for novel materials. However, both of these approaches are very expensive and computationally demanding. Alternatively, we can seek a surrogate approach based on machine learning and data mining. For that, σ , S & k should be adequately predicted using simple data-driven models.

1.2 Motivation

As aforementioned, thermoelectricity is considered among the viable options for renewable energy with high potential. For that, it is important to identify abundant and efficient TE materials. The search is done either by experiments, computations, or ML applications. Since the first two approaches require heavy resources and generally long times, one can use the latter. Most of the research of TE materials using ML is general for semiconductors as a whole or specific for a particular family of materials [2,4]. When it comes to TE in the considered family of inorganic halide perovskites (IHP), there is a short of focus and less efforts [13].

Incontrovertibly, there is no general-purpose ML predictive models for all families of materials due to infinite materials configuration space. Initially, the materials are

categorized in families and then each family will have its own ML models. To the best of our knowledge, there is no ML TE models for the important family of inorganic halide perovskites. This group of materials demonstrate experimentally a great potential in solar cells [7-9]. To explore it further, a $\mathbb{ZT}$ predictor is developed specifically for this group of materials. Taking this approach will hopefully reduce anomalies in the prediction. Furthermore, this approach is expected to yield higher accuracy compared to a non-categorized prediction. Another benefit from developing a family specific predictive model is the focused feature analysis, which facilitates the study of materials behavior and exploring the physic behind it. The effectiveness of such prediction can be enhanced further by considering a specific operating condition, such as room temperature and specific doping levels.

1.3 Objectives

To use ML approaches in the discovery of novel TE materials, one needs to understand $\mathbb{ZT}$ and its constituent behavior with respect to the variation of other parameters. Thus, creating predictive models for the thermal and electrical transports of IHP will require the consideration and collection of possible descriptors. Having this prospective, the objectives of this work are listed as follow:

1. To study the thermal transports, electrical transports and $\mathbb{ZT}$ for IHP at room temperature and carrier concentration of $10^{18}/\text{cm}^3$.
2. To gather sufficient data for the feature analysis and modeling processes.
3. Creating a reduced feature space using high level material descriptors.

4. To develop predictive models for the electrical conductivity, thermal conductivity and the Seebeck coefficient.
5. To evaluate the model's performance and limitations on different data sets and higher materials carrier concentration.
6. Conduct a data mining to screen potential IHP TE materials.

The ML work is conducted using MATLAB, Alteryx and Google Colab. Data gathering and parsing is conducted using Alteryx software and python codes to access different databases.

1.4 Thesis Structure

This thesis consists of six chapters. Chapter 1 is a general concise introduction, describing the thermoelectric effect, the motivation of this work, the objectives, and the main objectives. Chapter 2 presents an overview of thermoelectricity and ML applications in the field. The chapter also discusses recent advancement in TE modeling and previous contribution to the field. Then, Chapter 3 discusses the developed lattice thermal conductivity models and related results. Modeling electrical transports of TE materials is then discussed in Chapter 4. Chapter 5 presents the results of data mining and the prediction of the ZT . Finally, Chapter 6 is a conclusion chapter that summarizes the work findings and discusses future lookouts.

I would like to raise to the reader attention that Chapter 3 and Chapter 4 are considered for future publications separately and inclusively. Therefore, they are written in manuscript

format. Consequently, there could be some unavoidable repetitions and abbreviation redefinitions.

CHAPTER 2

OVERVIEW OF THERMOELECTRICITY AND ML

APPROACHES

2.1 Overview of Thermoelectrical Materials

As aforementioned, TE materials exhibit the Seebeck effect, which can utilize a temperature gradient to generate electric potentials [5]. This property has promising applications in utilizing wasted energy, such as TE harvesting devices [5] and Thermophotovoltaic generation [7,8]. For these applications to be practical, an efficient TE material is needed. A lot of work was done to understand the effect of the TE efficiency and its relationship with the thermal and electrical transports [2,5,13]. For a TE material to be efficient, it must possess high electrical transports and low thermal transports. However, these parameters are coupled. For instance, increasing electrical conductivity reduces the Seebeck coefficient and increases thermal conductivity [5,13].

The problem further escalates when considering different doping levels. The most promising TE materials are doped at a carrier concentration in the range of 10^{19} - 10^{21} cm⁻³ [5,13]. This is illustrated in Fig. 2-1 (after Haque et. al. [13] where the shown peak position of ZT is just for illustration and not general). Clearly, the electrical and the thermal transports exhibit nonlinear relations with the doping level. Heavily doped materials show

even more complex behavior [15,26], demonstrating the difficulty of tuning these parameters and the navigating optimum doping level.

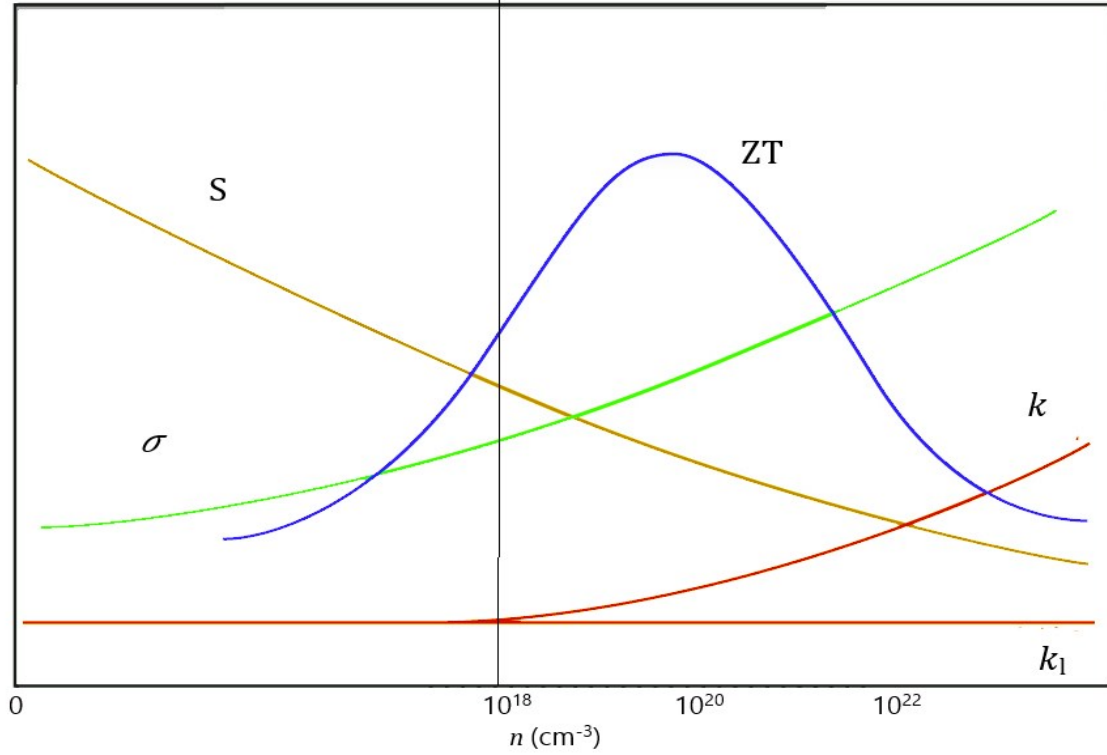


Fig 2-1 Electrical and thermal transport schematic for carrier concentration 10^{18} - 10^{22} cm^{-3}

An interesting sub-group of TE materials is skutterudites. These materials exhibit what is known as Phonon Glass Electron Crystal (PGEC) behavior [5] resulting in decoupling electrical and thermal transports. This phenomenon is related to the complexity of the compound and its heterogeneity at atomic level [5]. As a result, skutterudites can have high ZT up to 1.7.

Another interesting group of the TE materials are the half-Heusler compounds [2]. This group of materials show thermal conductivity less than 5 (W/m K) at room temperature [2]. Half-Heusler compounds attract attention due to the abundance of high ZT materials

in this group. In a seminal work by Yang and coworkers [5], they showed “computationally” that 11,858 out of 79,057 half-Heusler compounds have $ZT > 2$ at high temperature, which is the main reason for considering half-Heusler for high temperature TE applications.

In the past decade, halide perovskites have received momentous interest as a promising group for photovoltaics and probably for TE materials [7-9,13]. Halide perovskites are crystals with a chemical composition ABX_3 , where A is a +1 cation, B is a +2 metal cation, and X is a halide anion [13]. Similar to skutterudites compounds [5], halide perovskites exhibit a PGEC behavior [13]. Many of this group materials exhibits an interesting combination of ultra-low thermal conductivity and high Seebeck coefficient [13]. However, the obtained ZT values of IHP are still considered low. This is – mistakenly – attributed to “their” low electrical conductivity [13,48-49]. However, the considered IHPs are just a small subset of the whole family. Also, many IHPs are known for their high mobilities and conductivities [49].

This issue is illustrated in two interesting materials, $CsSnI_3$ and $CsPbI_3$ having a lattice conductivity 0.54 (W/m K) and 0.25 (W/m K) at room temperature, respectively [13]. For these two materials at $T=1000$ K and at a doping level of 10^{19} cm^{-3} , ZT of the range of 0.4-0.65 were obtained [13]. However, for Pb-based halide perovskite, there are concerns of using it, as Pb is toxic. Thus, the immediate alternative consideration is Sn-based halide perovskites, which exhibit relatively high electrical conductivity. However, Sn-based perovskite suffers from chemical instability issue [13]. Although these studies and many others illustrates the potential of IHP as promising TE materials, the covered IHP subspace

is still small and further investigations to explore their TE potential and to consider possible enhancement are required [13].

2.2 Data-Driven Modelling

Arguably, data driven modeling is one of the most convenient tools to analyze and correlate different aspects of complex systems captured in datasets. This approach takes a general process, which can be illustrated as follows:

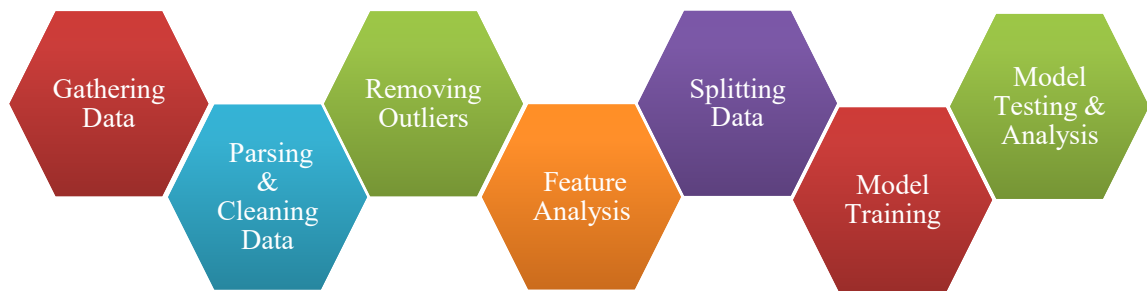


Fig 2-2 ML workflow

The process starts with gathering the required datasets. Usually for materials database, the data is stored in formats that are efficient or suitable for visualization while supporting proper querying process. However, these formats do not support data processing. Thus, there is a need for a parsing mechanism. The parsed data may require some cleaning and proper formatting. Once the data is cleaned, outliers can be detected and removed. This is one of the most critical steps in ML, outliers can influence and mislead the results at the model training stage. After preparing the dataset, a feature analysis is conducted. features

are inspected using visualization methods and/or correlation matrix to select possible features to model. The data then is split and trained using the training subset to optimize its performance. After obtaining a proper ML model, it is tested on a separate testing data and then evaluated for consistency and possible enhancements.

The general workflow of the used ML process is similar for most ML problems. However, there are several statistical learning approaches to consider for any problem. Choosing one (or more) depends on the nature of the data, the complexity of the relation between the variables within dataset, computational resources, and targeted accuracy. Machine learning methods can be categorized into supervised, unsupervised, and reinforcement learnings. Of the most famous ML methods are generalized linear regression, neural networks (NN), decision trees, random forest, boosted models, and many others [2,4]. Each can be deployed as an automated predictor, resulting in a time efficient process providing that the prediction accuracy is good. It is often the case that there are different possible methods that can be used for a given problem. Thus, it is a common practice to develop different ML models then comparing their performances [4]. In doing so, one must consider the accuracy, complexity, number of features, and susceptibility for overfitting, which influences the credibility of ML models as the dataset expands. In the following subsections, some of the utilized ML methods in this work are described generally and briefly.

2.2.1 Linear Regression

Regression is the process of statistically constructing a relation between set of inputs and a response (or more). As a matter of fact, the regression task is one of the earliest and most

fundamentals ML approaches [2]. Overtime, many regression methods at different levels were developed. The peculiar application depends on the nature of the studied problem and data spaces (features and size). Thus, it is obvious that different problems need different regressions [2]. The concept of regression can be extended to classification as in the case of logistic regression. In this work, we utilize linear regression with pre-designed features (descriptors). In general, linear regressions are not demanding in term of computational resources. They are easier to develop and to interpret compared to more advanced models such as the ones in kernel-based and deep learning methods [2,3].

In linear regression, the response $y \in \mathbb{R}$ is related to the inputs vector $\mathbf{x} \in \mathbb{R}^{p+1}$ where p features are used, and the extra term is for the intercept. The relation takes the following form:

$$y = \mathbf{x} \cdot \mathbf{w} + \epsilon \quad (2.1)$$

where $\mathbf{w}$ is the vector determining the model and it is the one to be found statistically and ϵ is an assumed Gaussian error with zero mean. By having N training data points where Eq. 2.1 is applied, the resulted equations system can be combined in

$$\mathbf{y} = \mathbf{X} \mathbf{w} + \boldsymbol{\epsilon} \quad (2.2)$$

where $\mathbf{y} \in \mathbb{R}^N$ is the responses vector, $\mathbf{X} \in \mathbb{R}^{N \times p+1}$ is the features matrix, and $\boldsymbol{\epsilon}$ is error vector. In full form, Eq. 2.2 becomes

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} 1 & X_1 & \dots & X_{1p} \\ 1 & X_2 & \dots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{N1} & \dots & X_{Np} \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_N \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{bmatrix} \quad (2.3)$$

The fitter seeks to obtain the best linear relation by maximizing the reward indicated by the cost function. One of the most commonly used cost function is the mean square error (MSE) J , which is:

$$J = \|\mathbf{X} \mathbf{w} - \mathbf{y}\|_2^2 \quad (2.4)$$

The reward here is optimized by minimizing the error J by varying $\mathbf{w}$. Linear models are often used in explanatory studies [2], as they are easy to interpret and their low computational cost. For these advantages, linearization is pursued for response that needs explanation. Linearization can be realized by transforming the input features, enabling the usage of linear model.

2.2.2 Regularization

To avoid overfitting, it is customary to seek regularized models. These models use adjusted cost functions to prevent (or at least to reduce) overfitting. Among the commonly used regularizations are ridge, lasso, and elastic net methods. Ridge regression accounts to the difference of feature correlations to a targeted variable. This is done by introducing the L2 regularization in the cost function [2] such that

$$J = \|\mathbf{X} \mathbf{w} - \mathbf{y}\|_2^2 + \alpha \|\mathbf{w}\|^2 \quad (2.5)$$

Regularization term imposes a lower weight for irrelevant features, reducing their contribution – and hence the model complexity – to the targeted variable.

Similarly, the lasso regression introduces instead a L1 regularization term to the cost function [2]

$$J = \|\mathbf{X} \mathbf{w} - \mathbf{y}\|_2^2 + \alpha \|\mathbf{w}\| \quad (2.6)$$

Lasso regression reduces the contribution of unimportant features by setting their weights to zero. Due to this property, lasso regression can be used in features selection stage. Such that only relevant features will remain in the model and the rest are discarded [2].

The third considered regularization is elastic net regression. It utilizes a weighted combination (through α) of both the ridge and lasso regressions [3]. Thus, resulting in a more comprehensive cost function

$$J = \|\mathbf{X} \mathbf{w} - \mathbf{y}\|_2^2 + \alpha \|\mathbf{w}\|^2 + (1 - \alpha) \|\mathbf{w}\| \quad (2.7)$$

As it accounts for both the second and the first norms, in many cases it offers a balanced model.

2.3 Machine Learning for Thermoelectricity

2.3.1 TE Effect Features Selection

In materials informatics, a descriptor is a quantity (or an extremely simple model) connecting microscopic properties to continuum macroscopic behavior. The possible analytical methods to develop and study descriptors vary depending on the field of the problem. Practically, these descriptors are used directly or indirectly as “features” in ML implementation. For thermoelectricity and depending on the analyses’ rationales, the implementation method of features and descriptors can be classified in three categories [5]. One interesting category emphasizes feature reduction, while another one allows having

all the obtainable basis microscopic features. The third category seeks compact derived features or descriptors.

There are tons of features and descriptors; however, most of them are basic, and only few of them are relevant to the problem at hand; namely, density, cell's size descriptors, cell's volume, atomic mass, spin momentum, and angular momentum [3,9]. It is observed that most of the descriptors are chemical ones, and they show high correlation with TE figure of merit. However, they do not solely dominate the TE behavior of three-dimensional (3D) materials. Thus, there is a need to expand the lookout and to incorporate additional features (or descriptors) [2,3].

In this work, it is sought to regroup the incorporated features in accordance with the $\mathbb{Z}T$ figure of merit as depicted in Fig. 2-2. Accordingly, the model of the $\mathbb{Z}T$ figure of merit can be represented as:

$$\mathbb{Z}T = \frac{S^2 \sigma T}{k_l + k_e} \quad (2.8)$$

where the thermal conductivity is split into two components. The first one k_e relates to the electronic contribution and the other is the lattice thermal conductivity denoted by k_l .

Furthermore, the reduced list of features can be categorized further into three parts. The first part contains electronic transport properties, including the carrier mobility, effective mass, and Pauling electronegativity [6,11]. This part is related to the electrical conductivity and the Seebeck coefficient [2]. In a similar way, another set of features are grouped in a part related to lattice thermal conductivity [2,11-12], which is the dominant part of the total thermal conductivity of semiconductors at low doping level [13]. Finally,

to comprehensively describe the $\mathbb{Z}T$ values, one must consider additional general and interactional features such as elastic properties and electron–phonon interactions [2]. Similar classification is used in materials discovery initiative for Heusler compounds [2,10] where an accuracy of 94% using hybrid ML-density functional theory (DFT) process was achieved. However, this reported accuracy measure does not consider false negatives and hence it could be misleading [14]. That issue is addressed as one of the problems with the classification of TE materials [2,4,10].

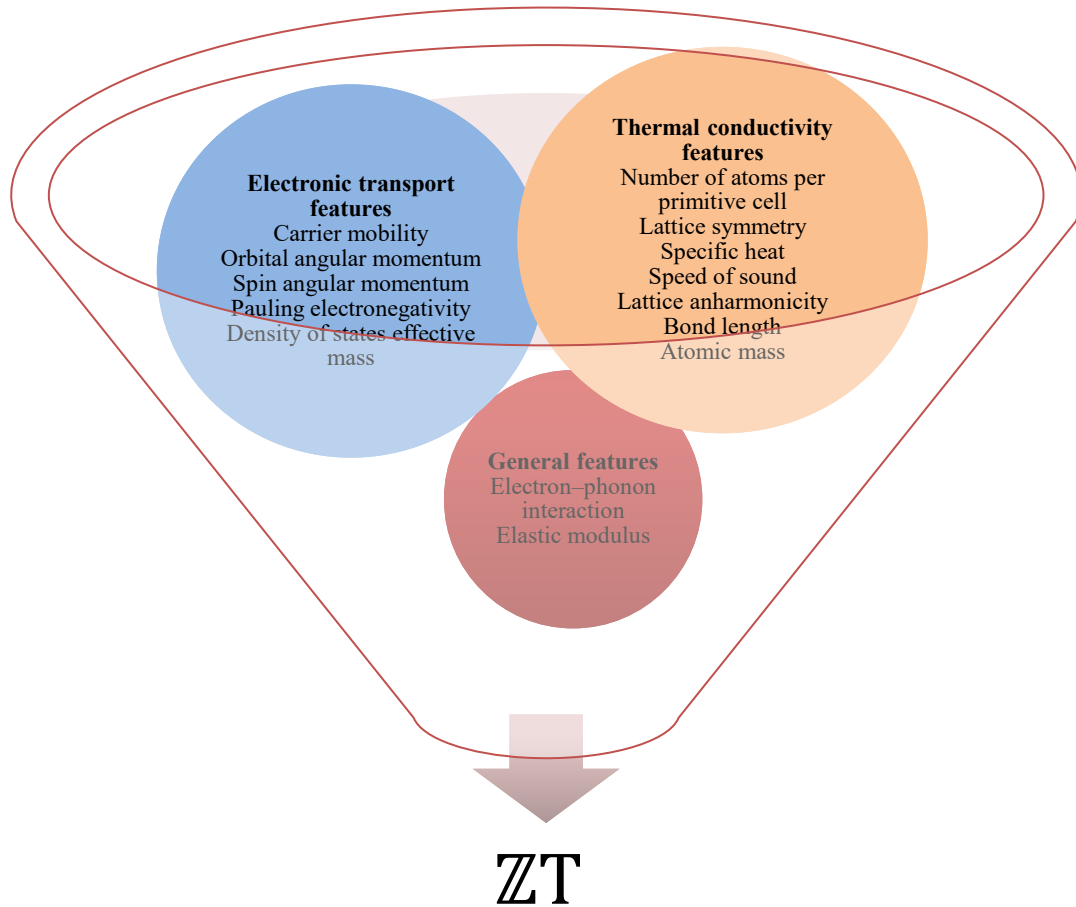


Fig 2-3 The considered features and descriptors used to develop a machine learning model(s) to estimate $\mathbb{Z}T$

Lately and by adopting physics-guided techniques, it was proposed that using highly compact and high-level descriptors [6,12,16] can enhance predictability and robustness in the predicting $\mathbb{Z}T$. Currently, three of these descriptors are identified. The first descriptor is a model for carrier charge mobility [6,11]:

$$\mu_0 = A_0 B^s m_b^{*-t} \quad (2.8)$$

where A_0, s and t are fitted parameters under the constant relaxation time approximation. B is the bulk modulus and m_b^* is the effective mass. The second descriptor is the lattice thermal conductivity, which is modeled under the consideration of two dominant scattering mechanisms: namely acoustical and optical [6,11,17]. The resulted model is

$$k_l = A_1 \frac{M v_s^3}{V_n^3 n^{1/3}} + A_2 \frac{(1-n^{-1/3})v_s}{V_n^3} \quad (2.9)$$

where A_1 and A_2 are fitted parameters, V_n is average volume per atom, M is the average atomic mass, n is the number of atoms in the primitive cell and v_s is the speed of sound [6,11]. The above features are used partially to develop the 3rd descriptor which is an alternative semiempirical figure of merit for thermoelectricity (β_{se}) that is derived from the $\mathbb{Z}T$. β_{se} is

$$\beta_{se} \propto \frac{\mu_0}{k_l} \left(\frac{m_{DOS}^*}{m_e} \right)^{3/2} \quad (2.10)$$

where μ_0 is the intrinsic charge carrier mobility, m_{DOS}^* is the effective mass of density of state (DOS) and m_e is electron mass. This method is applicable for both n -type and p -type materials [6,11]. The descriptor β_{se} accounts for the Seebeck coefficient, using the ratio of

the DOS and the electron masses [15]. The method that uses the β_{se} is reported as a high throughput screening method and computationally inexpensive one [6]. However, this approach was applied to only 412 materials not including inorganic halide perovskites. Also, the lattice conductivity and the mobility predictors were only trained for 45 materials [6].

2.3.2 TE material discovery using ML Classification

To determine the applicability of a material for thermoelectric generation, general purpose ML is not possible, and hence the materials should be classified in families first. Thus, ML classification is sometimes used. There are few approaches commonly used such as random forest and decision tree [10]. However, most often in literature, the work is carried out using the decision tree [4,10]. The classifiers consider features like bandgap, electrical conductivity, Seebeck coefficient and thermal conductivity in the classification [4]. One can consider different descriptors, depending on the structure of the material, dimensionality, and doping type. Consequently, a different set of ranges for these features are determined by the decision tree [2,4,10].

An interesting case is observed when dealing with half Heusler compounds. The decision tree drastically changes when comparing the materials at $T=300$ K with the ones at $T=1000$ K [2]. This is mainly due to structural phase transitions. For the first case, only one feature is needed that is the Pettifor's chemical scale, whereas it is not enough for the latter case and additional features are needed [2]. Thus, a specific problem setup is needed to ensure the success of the ML classification model.

Furthermore, it is known that classifications are associated with the quality of the data and the considered family of materials. For example, for Heusler compounds, an accuracy of 94% was reported using the hybrid ML-DFT method. In contrast, a range of accuracy between 84%-96% obtained for the case of the general 3D bulk TE screening [10]. The limitation of using a general model is due to large variation in behavior across different material spaces. The possible material space is extremely vast. Thus, it can be concluded that splitting materials into specific families at specified operation conditions and training them individually proved to be useful in such materials related ML work.

2.3.3 Predicting the $\mathbb{Z}T$ using ML

There are two commonly used ML approaches to predict the $\mathbb{Z}T$ figure of merit. The first one is based on ANN [2] while the second is linear regression approaches with engineered features [2,6,18]. In an attempt to predict the $\mathbb{Z}T$ figure of merit of $\text{Cu}_x\text{Bi}_2\text{Te}_{2.85+y}\text{Se}_{0.15}$ materials, ANNs was used with 12 features [2,19]. The achieved prediction accuracy was 98% [2], for the different permutations of a specific material. Another work was conducted by classifying compounds into three material-based families, before applying ML regression separately [18]. The mean absolute error of the applied approach was 0.0734 [18], which shows that even without using the complicated and uninterpretable ANN, proper pre-classification of materials and conditioning can be critical.

In an aspect related to this thesis, it is noted that the previous predictions were successful even at a generalized temperature, which cannot be applied to some materials such as perovskites [20]. The main reason behind this limitation is the fact that inorganic

perovskites go through successive structural phase transitions. Different phases require different considerations, due to their nonlinear TE behavior [13]

CHAPTER 3

PREDICTING THERMAL CONDUCTIVITY

Manuscript Title: Lattice Thermal Conductivity Prediction of Inorganic Halide Perovskites

In the past few decades, thermoelectric (TE) materials received considerable attention. Thermoelectric technologies offer a solution for the seekers of sustainable energy sources. However, the applicability of such technologies depends on having efficient thermoelectric material. Generally, such materials have low efficiency. Thus, a better understanding of the governing aspects of its performance is needed. This can be assessed through the figure of merit ZT , which illustrates the necessity of having low thermal conductivity. Such conductivity is commonly evaluated using experimental means or quantum mechanical computational approach. However, calculating such conductivity is difficult due to the huge number of involved physical mechanisms. Considering the expenses of these methods and by having many accessible databases, alternative approaches based on machine learning are considered. The aim of this work is to develop an accurate predictor for the lattice thermal conductivity using structural and elastic properties. The focus of this work lies on a specific family of materials; namely, inorganic halide perovskites. This reemerging group of thermoelectric materials demonstrates interesting thermal transports values. The initial results of using a predictive model demonstrate a promising potential in the discovery of novel thermoelectric materials with low thermal conductivity.

3.1 Introduction

In the path of finding sustainable energy sources, thermoelectric technologies offer a viable option. Often thought of as a waste, thermal energy is the residual of any physical operating system. Thermoelectric (TE) materials can convert this wasted energy into electrical energy source using the Seebeck effect [2-4]. For such application to be practical, this process must exhibit a reasonable efficiency. Thus, an efficient TE material is the key to ensure an efficient TE power conversion. The efficiency of TE materials is influenced by its electrical and thermal transport behavior [2-5,13]. Generally, these transports are coupled [5]. Yet, to simplify estimating the TE effect, one may attempt to separate these parameters. This work explains the conditions in which these parameters are separated. The focus of this work is the thermal transports, mainly the lattice thermal conductivity.

Thermal conductivity is a key descriptor of a material for many applications. Low thermal conductivity is sought for different applications such as TE generators [16]. Its effect on efficiency of TE generator is characterized by the $\mathbb{Z}T$ figure of merit [2-4].

$$\mathbb{Z}T = \frac{S^2 \sigma T}{k} \quad (3.1)$$

where k is thermal conductivity, S is the Seebeck coefficient, σ is electrical conductivity, and T is the temperature. The thermal conductivity can be decomposed into two parts: lattice thermal conductivity and electronic thermal conductivity [25-26]. Evaluating these parameters – especially the lattice thermal conductivity – imposes an obstacle. For experimental approaches, such evaluation requires heavy resources and time [4] while for computational based approaches, transport calculations are demanding in both computational resources and time [4]. This issue is further complicated with increased

dimensionality of materials. Therefore, a lot of work was done in 2-dimensional (2D) materials [5,10]. Since this work consider 3D materials, an alternative approach is required. It is proposed to use Machine Learning (ML) data-driven approach with density functional theory (DFT) computed structural features, which are relatively less computationally demanding and vast computational databases are accessible.

For low level of doping concentrations, a major contribution to the thermal conductivity is dominated by the lattice [13,25]. Under such condition, one can use the following approximation:

$$k \approx k_l \quad (3.2)$$

where k_l is the lattice thermal conductivity. This parameter is influenced by the acoustic phonons and the optical phonon scattering, contributing to the total lattice vibration of the material. These effects are described in the Debye-Callaway Model for thermal conductivity [6,11-12].

$$k_l = A_1 \frac{M v_s^3}{TV^{\frac{2}{3}} n^{\frac{1}{3}}} + A_2 \frac{(1-n^{\frac{2}{3}})v_s}{V^{\frac{2}{3}}} \quad (3.3)$$

where A_1 and A_2 are fitted parameters, V is average volume per atom, M is the average atomic mass, T is the temperature, n is the number of atoms in the primitive cell, and v_s is the speed of sound. The model shows an association between the elastic parameters and the lattice thermal conductivity. The Debye-Callaway Model suggests an interesting correlation between elastic parameters and lattice thermal conductivity. However, this model is sometime not consistent with experimental results [12]. Thus, there is a need to investigate and develop better lattice thermal conductivity models.

As standardly done in materials informatics [24], there are very few general-purpose ML predictive models and alternatively related ML models are developed for families of materials. Chosen for their promising performance in photovoltaics [7-8,13], the considered materials family is inorganic halide perovskites (IHP). This group of materials demonstrate ultra-low lattice thermal conductivity values [13] and – to the best of our knowledge – there is no ML predictor made for IHP. This work investigates different elastic parameters and different linear ML models in order to create an efficient predictive model for lattice thermal conductivity.

From the feature analysis conducted in this work, there are strong correlations between the elastic parameters and the lattice thermal conductivity. The highest correlation is observed in the Young modulus. By using only four features; Young modulus, Poisson's ratio, density, and unit cell volume, an ML predictive model was created. The accuracy of the model shows promising results of applying a data-driven method in the prediction of the lattice thermal conductivity. The model is compared with some existing models and demonstrates a higher accuracy.

3.2 Data

IHP materials have the composition of ABX_3 , where A is an alkali metal, B is a metal, and X is a halide. The needed data was extracted using RESTful application programming interface (API) from AFLOWlib [21]. The query of the database was conducted by limiting the species to those which follow the composition of IHP and selecting the target parameters. Lastly using a python code, these data are converted to a data frame and saved as an excel file. The data contains the density, volume, bulk modulus, shear modulus,

Poisson's ratio, and the thermal conductivity of 48 IHP compounds at 300 K computed using DFT post calculations. Both the Young's modulus and the speed of sound are then calculated using

$$E = 2 G(1 + \nu) \quad (3.4)$$

$$v_s = \sqrt{\frac{E}{\rho}} \quad (3.5)$$

where E is Young's modulus, G is the shear modulus, ν is Poisson's ratio, v_s is the speed of sound, and ρ is the density.

3.2.1 Data Filtering

After acquiring the data, it undergoes a series of filtration process to prepare the data.

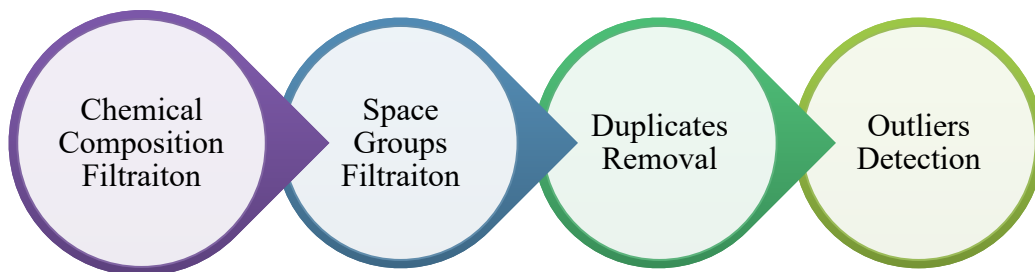


Fig 3-1 Data filtering process

The first applied filter is based on the chemical composition, only materials with the chemical formula ABX_3 are selected. Since the chemical compositions are not stored in simplest form or the proper order in the dataset, the compositions are parsed, arranged,

and stored in the simplest form. The materials are filtered further based on specific space groups, which are in the scope of Glazer's classification of tilted octahedra [24,27]. These space groups are summarized in the below table:

Table 3-1 Considered space groups [24,27].

	space-group	space-group #		space-group	space-group #
1	$\text{Pm}\bar{3}\text{m}$	221	9	C2/c	15
2	I4/mcm	140	10	$\text{P}\bar{1}$	2
3	P4/mbm	127	11	Pnma	62
4	Imma	74	12	$\text{P2}_1/\text{m}$	11
5	C2/m	12	13	$\text{P4}_2/\text{nmc}$	137
6	Cmcm	63	14	$\text{Im}\bar{3}$	204
7	I4/mmm	139	15	Immm	71
8	$\text{R}\bar{3}\text{c}$	167			

The dataset is filtered for duplicates, where repeated materials with the same chemical composition and space group are removed. The data then is inspected for outliers using Inter Quartile Ratio (IQR). Data cleansing is also performed based on the IQR limits. Finally, the data is split randomly into a training set and testing set with a ratio of 80% to 20%.

3.2.2 Feature Analysis

The Debye-Callaway Model is an insightful model for the lattice thermal conductivity, which inspired other works to investigate its constituent using ML approaches [12]. The

result of that work suggests considering the density, volume, bulk modulus, shear modulus, Poisson's ratio, hardness, number of atoms per unit cell, and Young's modulus as possible features for the lattice thermal conductivity prediction [12]. The reported results showed an R^2 accuracy of 0.946 using only three features [12]; namely total number of atoms in unit cell, volume per atom, and Young's modulus.

This section investigates the correlations between some of the suggested features and lattice thermal conductivity. To visualize and capture the trends between the targeted variable and the features, the following scattering plot in Fig. 3-1 is used. It is clear that thermal conductivity is proportional (not necessarily linearly) to bulk modulus, shear modulus, Young's modulus, and speed of sound. On the contrary, it is inversely proportional to Poisson's ratio, density, and volume. These facts will be utilized in the suggested model forms in the next section. These behaviors are expected as reported by many [2,12,16]. Also, the correlations between the considered features are investigated to avoid redundancies and to reduce the number of needed features. These are shown in the correlation matrix (capturing interconnectedness) in Fig. 3-2. It is clear – and not surprisingly – that the bulk modulus, shear modulus, and Young's modulus are highly correlated with their correlation coefficients exceeds 95%. Thus, it will not be useful to use multiple of them in a single model.

Furthermore, it is clear that the speed of sound is a critical feature for thermal conductivity. The resulted correlation factor is high (92% as shown in Fig. 3-3). The of Poisson's ratio, the density, and volume correlations scores with the conductivity are lower compared to elastic parameters, but they are still significant and hence considered for modeling in the next subsection.

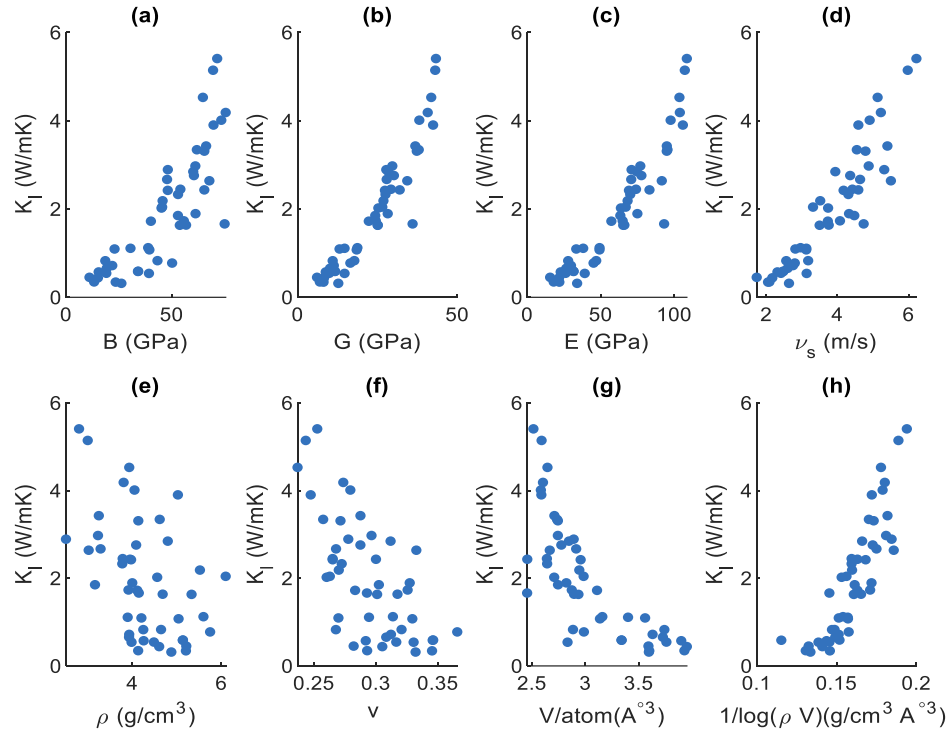


Fig 3-2 Thermal conductivity vs. a) the bulk modulus, b) the shear modulus, c) Young's modulus, d) the speed of sound, e) the density, f) Poisson's ratio, g) the unit cell volume per atom, and h) the reciprocal of log of density multiplied by unit cell volume.

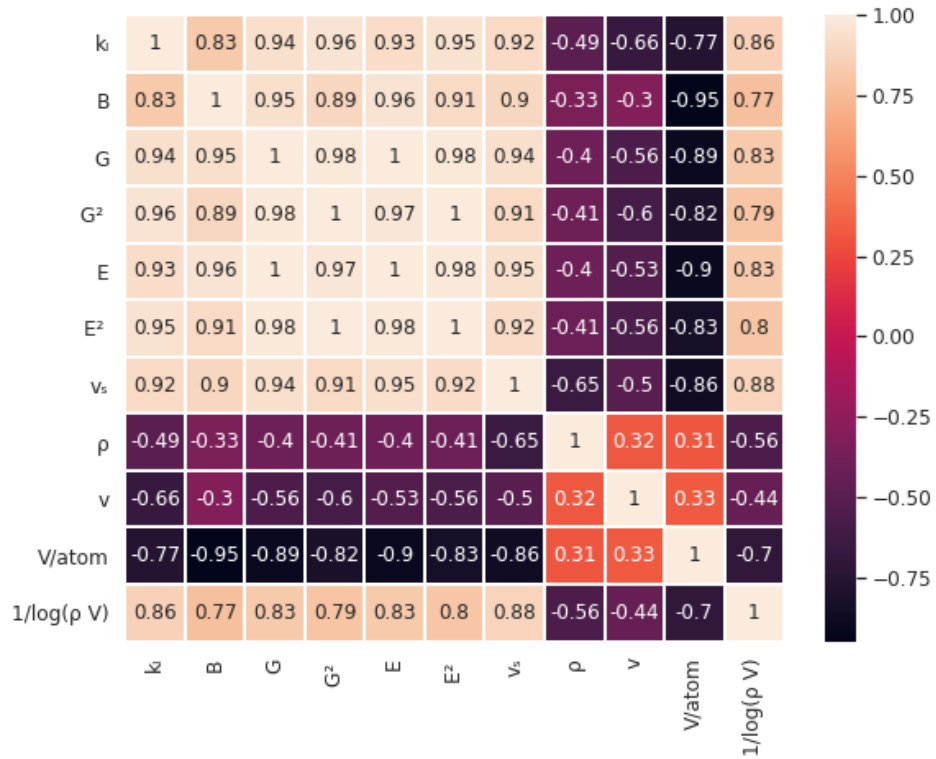


Fig 3-3 Correlation coefficient matrix of the considered features and thermal conductivity.

3.3 Models for Lattice Thermal Conductivity

The filtered-out dataset is small and hence prone to overfitting. Thus, elastic net linear generalized regression is used with feature transformation to create possible ML models. This approach is meant to restrict the model's freedom of fitting. As discussed in the previous subsection, looking at the relationship between the targeted variable and the inputs in Fig. 3-2, one can observe almost linear relations between most features and the response. As for nonlinear relations, they are captured by using some features transforming. Thus, the of consideration of G^2 , E^2 and $\log(\rho V)$. This data preprocessing enhances the performance of simple predictor like linear regression.

In summary, the feature analysis highlighted eight considerable features: B , G , G^2 , E , E^2 , v_s , v , and $\log(\rho V)$, each with a correlation factor over 80%. These features are used to create different possible models. However, from Fig. 3-3, there is a significant correlation between B , G , E and v_s . Thus, prior to train the model, the feature space will contain only one of these parameters in addition to v , and $\log(\rho V)$. Hence, limiting the number of dependent features in a model and possible overfitting.

Since G and E have the highest correlation factors and they are intercorrelated, one must be chosen. To do so, general models are generated in the forms of:

$$k_l = 0.0004 E^2 - 0.0186 E - 8.806 v + \frac{24.945}{\log(\rho V)} \quad (3.6)$$

$$k_l = 0.002 G^2 - 0.042 G - 7.932 v + \frac{22.774}{\log(\rho V)} \quad (3.7)$$

Table 3-2 Results summary of training and testing for the generalized models

	Training		Testing	
Model	RMSE	R^2	RMSE	R^2
Eq(3.6)	0.273	0.961	0.326	0.933
Eq(3.7)	0.251	0.967	0.358	0.916

Since the model that utilizes E has higher average accuracy, E is considered instead of G . Using E, E^2, v , and $\log(\rho V)$ four models are created using different combinations of these parameters. The four proposed models are listed in Eqs. 3.8-3.11.

$$k_l = 0.0002E^2 - 9.153 v + \frac{22.340}{\log(\rho V)} \quad (3.8)$$

$$k_l = 0.0003E^2 - 0.0103 E - 0.0453v \quad (3.9)$$

$$k_l = 0.0274 E - 11.27v + \frac{22.473}{\log(\rho V)} \quad (3.10)$$

$$k_l = 0.0005E^2 - 0.0264E + \frac{7.1525}{\log(\rho V)} \quad (3.11)$$

3.4 Results and Discussion

The four models are trained on the same dataset after data shuffling. Overall, one can observe that the effectiveness of using elastic parameters to predict the lattice thermal conductivity. However, among the four models, the third model shown in Eq. 3.10 is the most accurate one. This can be related to the high correlation between the Young's modulus and the lattice thermal conductivity and lower variance in the interrelation (Fig.3-2c). Young's modulus can be considered as a descriptor containing the shear and the bulk

moduli. Thus, it is a more comprehensive descriptor of both the speed of sound and the lattice vibrations. After testing all the predictors, a summary of the training and testing accuracies are generated and listed in Table 3-3 below beside some of the best reported values in the literature.

Table 3-3 Results summary of training and testing for all models.

	Number of features	Training		Testing	
Model		RMSE	R^2	RMSE	R^2
Eq(3.8)	4	0.260	0.965	0.398	0.890
Eq(3.9)	2	0.363	0.937	0.393	0.928
Eq(3.10)	4	0.334	0.942	0.422	0.924
Eq(3.11)	3	0.407	0.914	0.419	0.876
Loftis et. Al. [12]	3				0.946
Tewari et. Al.[28]	11				0.95

It can be seen that even with few features, highly accurate models both in R^2 and the RMSE measures can be developed. The model in Eq. 3.10 shows the best performance with an accuracy higher than (or comparable to) some of the best previously proposed models [11-12,28], which confirms the effectiveness of considering only a specific family of materials. It is worth noting that only 4 features are used while larger number of features are used in the cited models utilizing up to 6 features [11,12] or in some work 11 & 12 [28]. The corresponding results of these two works are shown in the last two rows in Table

3-3. It also worth noting that the proposed model has the simplest form and was created using a simple ML method in comparison to the models proposed on some works [12,28].

Furthermore, the volume per number of atom and density showed a consistent significance in most of the models. This can be related to the effect of the structural properties of a material and the lattice thermal conductivity [5,11-12]. It is noted that some materials in the data set showed an ultra-low thermal conductivity $k_l < 0.4$ W/m K. This highlights the importance of investigating IHP as a promising family of TE materials. Further assessment for the models developed in this work is needed. However, there is a shortage of data to train the models and to compare our results to. This is related to the lack of focus on IHP compounds for TE [13].

3.5 Conclusion

In this work, the thermal conductivity of IHP is investigated and estimated using ML approaches. These materials demonstrate interesting thermal transport values and hence high potential for TE applications. IHP material finds its way in different applications, mainly in photovoltaics. The developed models aim to predict the lattice thermal conductivity with low electronic contribution. This is achieved by studying materials at lower doping level at which the electronic contribution to the total thermal conductivity is negligible. The work exploited the usage of data driven approaches to avoid the demanding computational or experimental means. By investigating the relationship between structural properties and the elastic parameters of materials, an accurate ML model to estimate lattice thermal conductivity for IHP was developed. These new models were developed for this family mainly, showing promising results and accuracy with a small number of features,

which could be crucial for TE material discovery. Further data is needed to expand the developed ML models. However, after investigating different databases there is a shortage of TE related data for IHP and data limitation for certain temperatures. Thus, more work is needed to further explore the potential of IHP in TE applications.

CHAPTER 4

ML MODELS FOR ELECTRICAL TRANSPORTS

Prediction of Inorganic Halide Perovskites Electrical Transports Using ML Approaches

In the recent years, significant effort has been dedicated to understanding the performance of thermoelectric (TE) materials. Thermoelectricity offers a promising application as a renewable energy source. However, the efficiency of such source relies on using an efficient thermoelectric material. So far, the achieved energy conversion efficiencies are low. To overcome this issue, more insights on TE materials performance are needed. This can be quantified by using the figure of merit ZT , which illustrates the necessity of having excellent electronic transports. Having full account to estimate electronic transports is difficult due to the large number of involved mechanisms. Furthermore, another obstacle facing the study of TE material performance is the extremely complex evaluation of such transports. Experimental assessments and quantum mechanical based computational approaches are commonly used to measure and evaluate the ZT figure of merit. Due to the considerable expenses of these methods, one may consider an alternative approach. With the recent developments in data analytics, machine learning modeling has become a viable approach, due to the accessible materials databases. This work aims to develop accurate predictors for the electrical conductivity and the Seebeck coefficient of inorganic halide perovskites using the relation between structural properties and parameters obtained from

the electronic band structure. The prediction can be used to search for novel thermoelectric materials.

4.1 Introduction

In the constant search for viable renewable energy sources, thermoelectricity has become a promising source in recent years [2]. For efficient thermoelectric (TE) generators, it is crucial to utilize efficient TE materials. Therefore, the search for such materials became a key in evaluating the viability of this energy source [5]. For TE generator, both electrical and thermal transports are used to evaluate its efficiency. The relation between these parameters and efficiency is quantified using the ZT figure of merit [2-4] as show in Eq 4.1.

$$ZT = \frac{S^2 \sigma T}{k} \quad (4.1)$$

where S is the Seebeck coefficient, σ is electrical conductivity, T is the temperature, and k is thermal conductivity. The estimations of electrical and thermal transports are commonly done by experiments, computations, or empirical models [5]. Recently, machine learning (ML) modeling has provided an alternative venue due to the remarkable developments in materials informatics and databases.

In this work, we developed an efficient and extremely compact model to estimate the electronic transport of inorganic halide perovskite (IHP) materials tailored for TE applications. Due to their low thermal conductivity and high Seebeck coefficient [13], IHP attracted the attention of the thermoelectricity (TE) community recently [7-9]. Besides their success in solar cells [13], this family of materials shows a good potential in TE

applications [13]. As commonly done in TE materials ML modeling, only a specific group of materials is considered for modeling [7,13]. To the best of our knowledge, there is no machine learning TE models for this important materials family of IHP. In this work, the electrical conductivity is modeled in term of charges carrier mobility. The used dataset is based on density functional theory (DFT) calculations for stable IHP materials. Many features are used in the analysis such as, the effective mass, band gap, unit cell volume per atom, and the density [2].

Unsurprisingly, it turned out that only the effective mass is needed to capture extremely well electronic conductivity [28-31]. Furthermore, the Seebeck coefficient model showed a promising accuracy, using only the approximated density of state effective mass and the band gap. The density of state effective mass was approximated using linear regression for materials with similar band degeneracy factor. This factor can be related to unit cell volume per number of sites and the band gap. Thus, justifying the usage of the linear model and its high accuracy.

4.2 Data

The data contains the density, volume, number of sites, lattice parameters, band gap, effective mass, Seebeck coefficient, electrical conductivity of *n*-type materials of 32 compounds. All the electrical transport data are evaluated at 300 K and carrier concentration of $10^{18}/\text{cm}^3$ using the DFT computed band structure. The electrical transports data were obtained from Materials Project Contributions (MPContribs) [32]. As for the structural data and space group information, they were obtained directly from the Material project [23] database itself. This process is carried through the usage of the pymatgen

module and MPRester querying [23] function offered by Material project. The data is then stored as a data frame and analyzed.

After acquiring the data, it passes through several stages to prepare the data for processing. The first stage is selecting materials with the chemical composition ABX_3 . The materials are then filtered based on specific space groups, which are based on Glazer's classifications of perovskite tilting systems [24,27]. These space groups are summarized in Table 4.1.



Fig 4-1 Data filtering process

Table 4-1 Considered space groups [24,27].

	space-group	space-group #		space-group	space-group #
1	$Pm\bar{3}m$	221	9	$C2/c$	15
2	$I4/mcm$	140	10	$P\bar{1}$	2
3	$P4/mbm$	127	11	$Pnma$	62
4	$Imma$	74	12	$P2_1/m$	11
5	$C2/m$	12	13	$P4_2/nmc$	137
6	$Cmcm$	63	14	$Im\bar{3}$	204
7	$I4/mmm$	139	15	$Immm$	71
8	$R\bar{3}c$	167			

In this work, only stable perovskites are considered [33-37]. The tolerance factor (TF) is used to describe the phase stability of the perovskite structure. This factor presents an essential feature for the fabrication of TE materials and performance consistency [37]. This issue is commonly highlighted in the application of IHP materials in solar cells [33, 38], where some materials show promising performance. However, due to these materials' instability, their performance is limited [38]. This work accounts for this issue in the process of discovering promising TE materials with excellent electrical transports. Thus, only materials with $0.7 < TF \leq 1$ are considered [38]. To calculate the tolerance factor for different orientations of supercells with different number of atomic sites, TF equation can be expanded as:

$$TF = \begin{cases} \frac{a}{b}, & n = 5 \\ \frac{a+b}{2c}, & n = 10 \\ \frac{a+b}{\sqrt{2}c}, & n = 20 \end{cases} \quad (4.2)$$

where n is the number of atomic sites, TF is the tolerance factor, a , b , and c are the lattice parameters. For structures where $b > c$, the two parameters are swapped and evaluated according to the Eq. 4.2. However, the lattice parameters data requires parsing and proper storing. The dataset is filtered for duplicates, where repeated materials with the same chemical composition and space group are removed. The data then is inspected for outliers using Inter Quartile Ratio (IQR). Data cleansing is also performed based on the IQR limits. Finally, the data is split into a training set and testing set with a ratio of 4:1.

4.3 Electrical Conductivity

4.3.1 Feature Analysis

To capture the trends between the targeted variable and the features, the scattered plot (Fig. 4-2) and correlation matrix (Fig. 4-3) are created. Clearly, the highest correlation between the electrical conductivity is with the effective mass. This follows the connection between the intrinsic mobility and the electrical conductivity through the effective mass [6,9]. This linear correlation was also observed in some materials at certain carrier concentration in previous work [32]. Furthermore, it is clear that there is a linear fit between the electrical conductivity and the effective mass.

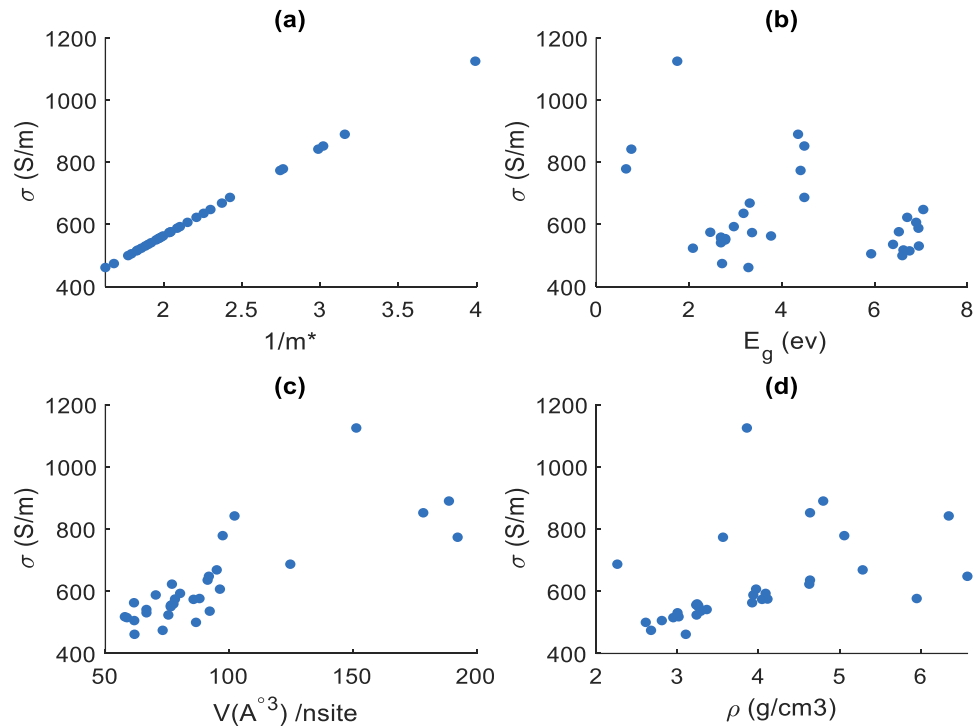


Fig 4-2 Electrical conductivity vs. a) the reciprocal of effective mass, b) the Bandgap, c) volume ,and d) mass density



Fig 4-3 Correlation matrix for electrical conductivity and the considered physical features.

An interesting correlation between the structural features and the electrical conductivity are shown in Fig. 4-2 (c) and Fig. 4-2 (d). This may suggest the study of the elongation and the angle variance within the octahedron structure [36]. However as observed with different databases, the atom positions were stored for specific structural visualization software and parsers. Thus, parsing these data and extracting the elongation and the angle variance becomes highly difficult and inaccurate and it will be approached in future work.

4.3.2 Modeling the Electrical Conductivity

Since small data sets are prone to overfitting, elastic net linear regression is used with feature transformation to create possible ML models. This approach is meant to restrict the model's freedom of fitting. The model is constructed using only the effective mass, due to its dominant correlation (almost 1 as shown in Fig. 4-3). The resulted model is

$$\sigma = \frac{282}{m^*} \quad (4.5)$$

where σ is the electrical conductivity and m^* is the effective mass. The model shows confirmation with the intrinsic mobility under constant relaxation time [6,29].

$$\mu_o = \frac{e\tau}{m^*} \quad (4.6)$$

where μ_o is the intrinsic mobility and m^* is the effective mass, τ is the relaxation time.

Using the effective mass as a predictor is very common as can be found in different works [6,11], which align with the work proposed here.

4.3.3 Results and Discussion

The training result of the electrical conductivity model shows a steady and highly accurate performance across the data set as shown in Fig. 4-4.

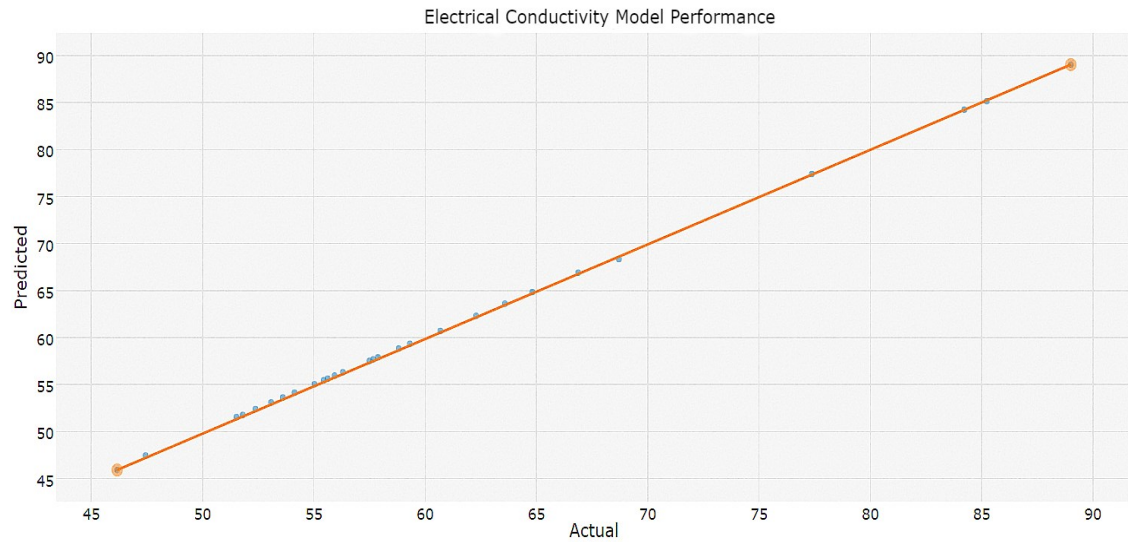


Fig 4-4 The predicted vs. actual values of the electrical conductivity

The predictor shows a training accuracy of 0.975 R^2 with 2.311 absolute RMSE. The corresponding testing accuracy is 0.99 R^2 with 0.3 RMSE. Although the model is linear and requires only one feature, the predictor results are excellent. This can be attributed to different possible explanations. One explanation is the linear relationship observed between the reciprocal of effective mass and electrical conductivity. This is confirmed in the DFT calculations, as some materials following certain carrier concentration doping show this linear trend [42]. Another reason is the quality of the calculated mobility under a constant relaxation time [43] and hence limiting the effect of different scattering mechanisms [43]. Another indirect reason is the fact that we only considered the contribution of stable IHP at room temperature in this work. So, this step resulted in removing many outliers beforehand.

4.4 Seebeck Coefficient

4.4.1 Feature Analysis

To conduct the visual feature analysis of the Seebeck coefficient, we use the following scatter plot (Fig. 4-5) and the correlation matrix (Fig. 4-6) considering the same features used earlier beside Pauling electronegativity of the B site. This follows from some previously suggested works [2]. Clearly, the effective mass and energy gap have the highest correlation with S . Hence, they are used to develop the model. Another interesting aspect is the effect of states filling on S .

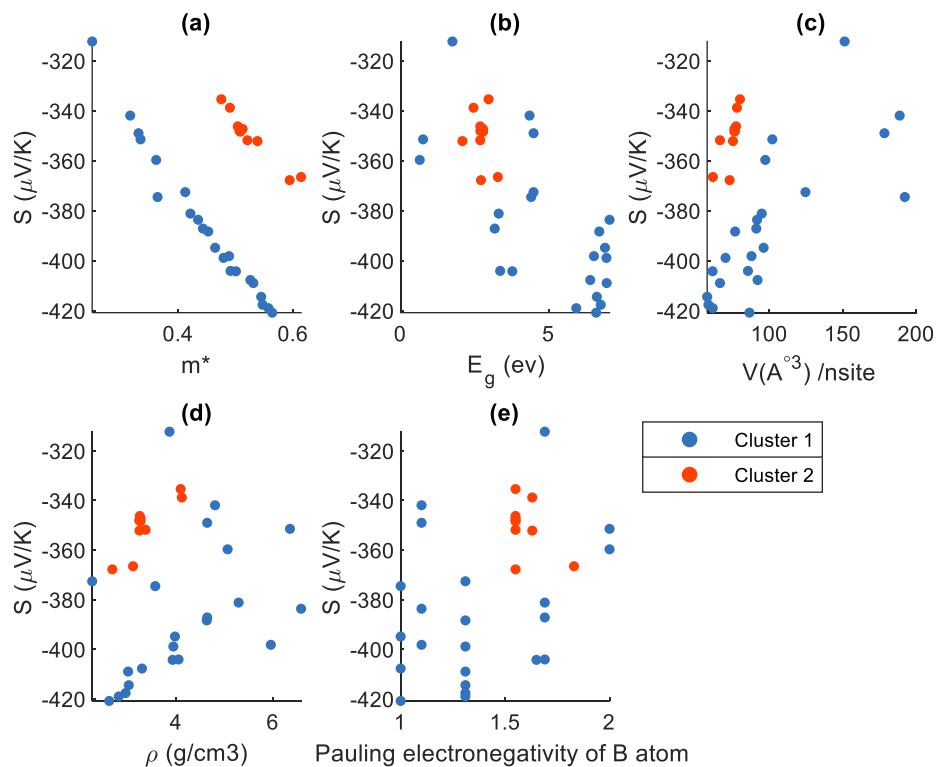


Fig 4-5 Seebeck coefficient vs. a) the effective mass, b) the bandgap, c) volume/(number of sites/5), d) mass density, and e) the Pauling electronegativity of the B atom.

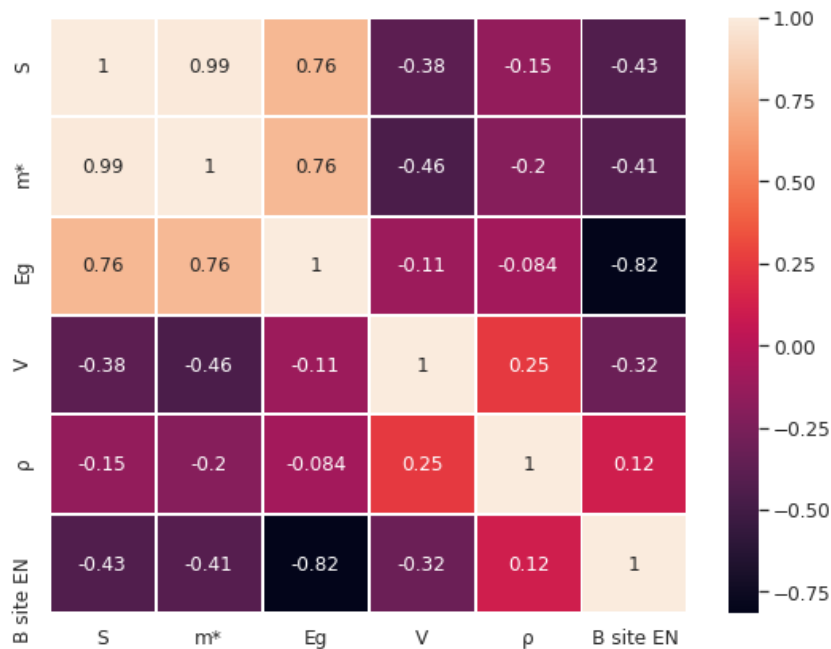


Fig 4-6 The correlation matrix of the $|S|$

As can be seen in Fig. 4-5(a), there are two distinct linear behaviors. As known, this is related to the resulted valley degeneracy factor and the band complicity as shown in Eq. 4.7 [26,40-46]. The effective mass introduced earlier is also referred to as the conduction band effective mass, which can be related to the DOS effective mass through the below expression:

$$N_V K = \left(\frac{m_b^*}{m^*} \right)^{\frac{3}{2}} \quad (4.7)$$

where N_V is the valley degeneracy factor, K is the anisotropy factor and m_b^* is DOS effective mass. The term $N_V K$ relates to the band utilization, which can be related to the band filling as shown below:

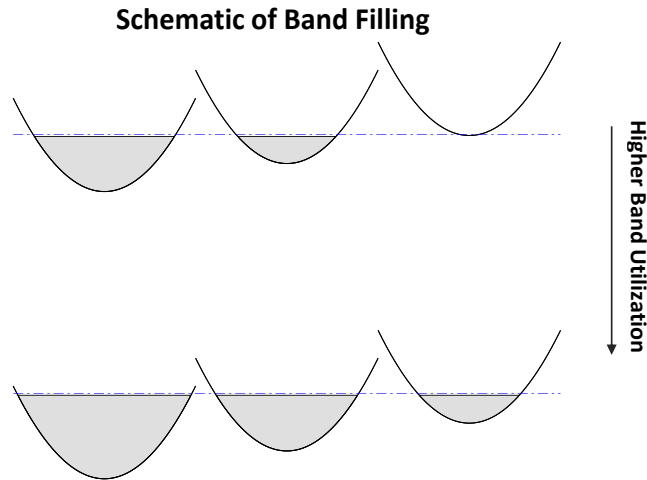


Fig 4-7 Schematic of Band Filling

Band filling is a result of increasing the carrier concentration. However, two materials at the same doping level can have different $N_V K$ [26], this is due to the different curvature of

the band. Steeper bands fill faster, while flatter bands fill at a lower rate [26]. This explains the behavior in Fig. 4-5(a), where clearly there are two groups of materials with different valley degeneracy factors. The two groups can be classified based on bandgap and unit cell volume suggesting that materials in cluster 2 are of a higher valley degeneracy factor compared to the ones at cluster 1. Furthermore, Fig. 4-5 (b) shows a resemblance of a linear correlation between the bandgap and the Seebeck coefficient. As for the density, volume per number of sites, and the Pauling electronegativity of the B atom, there is no clear association between them, and the targeted variable as reflected in the small correlation coefficients as depicted in Fig.4-6. It is to be noted, that some feature may not convey a strong relationship with the targeted variable. However, one must account to the limited variability of the features due to the small dataset.

4.4.2 Modeling the Seebeck Coefficient

Elastic net linear regression is used to predict the Seebeck coefficient, to maintain the performance of the predictor for all the electrical transports. Since the data obtained have a carrier concentration less than $10^{19}/\text{cm}^3$, the materials can be assumed to be degenerate [13]. Thus, limiting the possible band degeneracy and the number of bands contributing to the DOS effective mass [26,42]. As for the anisotropy factor K , it is reduced to 1 for all the materials here. This is the result of performing the effective mass calculations at constant relaxation time [39]. To evaluate the valley degeneracy factor N_V for the different clusters, data points were classified as shown in the illustration of Fig. 4-8.

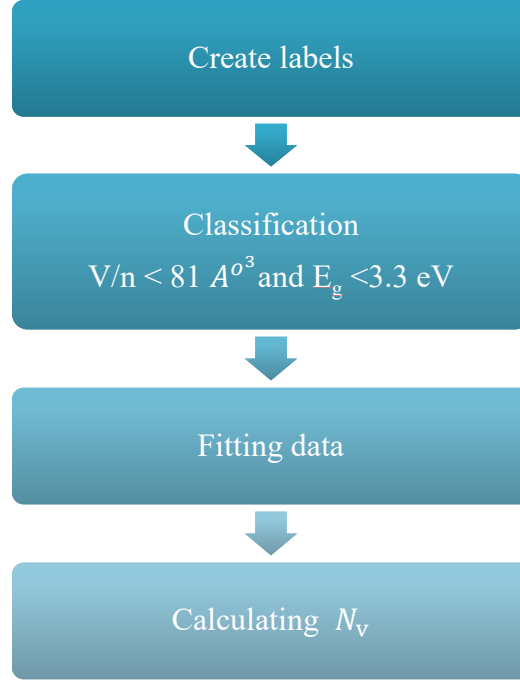


Fig 4-8 N_v estimation process

The classification is conducted based on bandgap and volume per number of sites. For the different groups, a linear fitter is used to predict the slope of the effective mass and the Seebeck coefficient and to create labels. Finally, the ratio between the two slopes was taken to approximate the valley degeneracy factor and the DOS effective mass. Using the DOS effective mass, a linear model is created to predict the Seebeck coefficient.

$$S = 344.774 m_b^* + 0.039 E_g + 231.641 \quad (4.8)$$

where S is the Seebeck coefficient and m_b^* is the DOS effective mass, E_g is the bandgap.

4.4.3 Results and Discussion

The training performance diagnosis shown in Fig 4-9 presents promising results. The model training accuracy is 96.6% with 4.729 RMSE and the testing result is 98.8% with 4.430 RMSE.

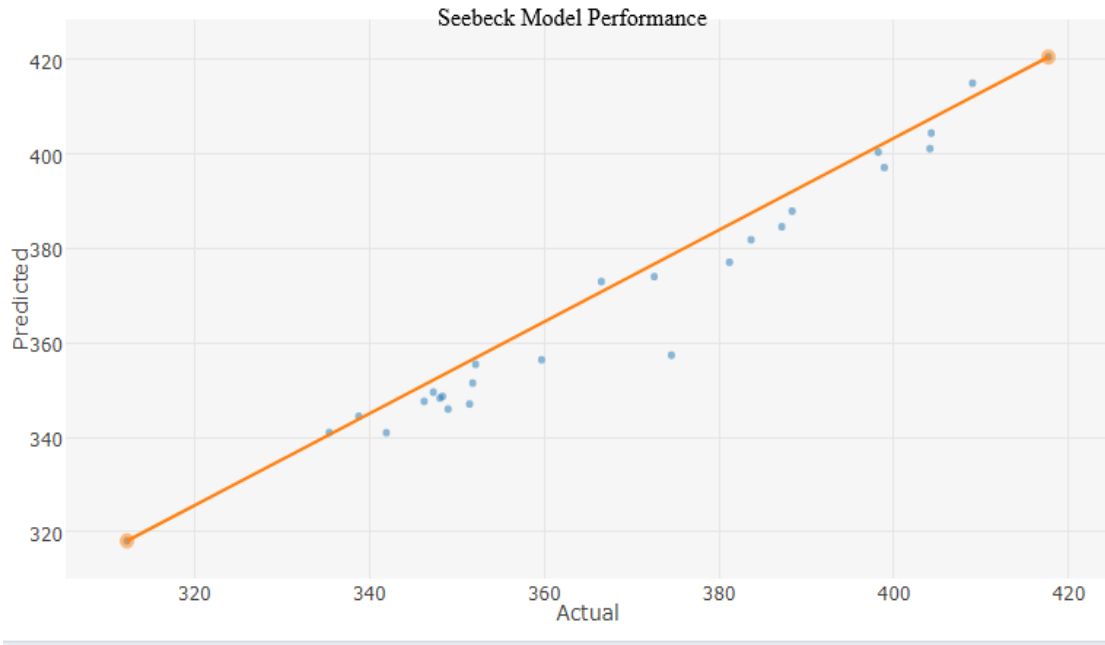


Fig 4-9 The predicted Vs Actual values of the Seebeck coefficient

The effectiveness of a linear model can be related to the range of effective masses and the small data set. Another reason is the outliers filtering for stable inorganic halides. The band anisotropy and the band degeneracy are critical aspects for high ZT TE materials [41]. The usage of constant relaxation time limits the variance contributions of these parameters. However, one can observe the trend between the band degeneracy factor and the magnitude of the Seebeck coefficient as inversely proportional. This conforms the general trend presented in the literature [5,11,26]. Considering IHP at carrier concentration $10^{18}/\text{cm}^3$

simplified the analysis of the Seebeck coefficient, which can be compared to the studies in previous works at higher carrier concentrations [13,26,42].

4.5 Conclusion

In this work, an accurate ML model is developed to predict electrical transports for IHP, which exhibits interesting electrical transport values and hence considered for TE applications. The developed models aim to predict the electrical conductivity and the Seebeck coefficient with low carrier concentration at room temperature and constant relaxation time. This is achieved by studying materials at doping level of $10^{18}/\text{cm}^3$ at which the band degeneracy factor is reduced. By considering a constant relaxation time, a significant simplification is introduced from the reduced band anisotropy. Thus, reducing the complexity between two key features; the DOS effective mass and the conduction band effective mass. This work utilizes data-driven approaches to avoid the demanding computational or experimental means. By investigating the relationship between structural properties and parameters obtained from the electronic band structure of the materials, ML predictors for the electrical conductivity and the Seebeck coefficient were developed. These new models were developed for IHP mainly, showing promising results and accuracy with few features, which could be essential for TE material screening and discovery. To further test the electrical transport predictors, larger datasets are needed. However, the shortage of data and the quality of the computed features create a critical obstacle. Another obstacle is distinguishing between anomalies behavior and calculation error of the DFT computed dataset. Thus, more work is needed to further explore the potential of IHP in TE applications.

CHAPTER 5

ZT PREDICTION

In the previous chapters, the ZT model was studied and analyzed with respect to electrical and thermal transports. In Chapter 3, a model for the lattice thermal conductivity was created. In Chapter 4, the electrical and the Seebeck coefficient were developed. By combining these models, a ZT predictor is created. In this chapter this model is used to predict the ZT of two data sets for possible discovery of TE materials.

5.1 Gathering Data

To search for new IHP materials for TE applications, two different data sets were obtained. The first one is of IHP with a high carrier concentration ($10^{20}/\text{cm}^3$) obtained from Jarvis DFT [47]. The data was pulled from the Jarvis data base for 3D materials through an API access. The features obtained from this data set are DFT calculated. Upon further inspection, some features have abnormal values (mobilities in 10^6 order), which cause an error in the predicted ZT value. These are considered anomalies and removed. The second data set is from Material project, which are actually used for Jarvis data. The data set contains structural, electrical, and elastic properties of IHP at a doping level of $10^{18}/\text{cm}^3$ at 300 K. The predictors developed in Eq. 3.9, Eq. 4.5 and Eq. 4.8 are used to evaluate the Seebeck coefficient, the electrical conductivity, and the thermal conductivity. Finally, from these parameters, the ZT is calculated.

For TE materials at doping level higher than $10^{18}/\text{cm}^3$, some adjustments are applied to the models. The electrical conductivity is scaled by the ratio of the doping level divided by $10^{18}/\text{cm}^3$. This is related to the scaling of mobility due to the change of carrier concentration [17] as shown in Eq. 5.1.

$$\mu = ne\mu_o \quad (5.1)$$

where n here is the carrier concentration level and μ_o is the intrinsic mobility. For doping levels higher than $10^{18}/\text{cm}^3$, the contribution of electronic thermal conductivity become more significant than lattice thermal conductivity. This can be calculated by using electrical conductivity through the Wiedemann-Franz law in [39]

$$k_e = \sigma LT \quad (5.2)$$

k_e is the electronic thermal conductivity, $L = 2.4 \cdot 10^{-8} \text{ W}\Omega\text{K}^{-2}$ is Lorenz number. Finally, by using Eq. 3.2 the total thermal conductivity is calculated.

5.2 Results

The results of evaluating the electrical and thermal transports at a doping level of $10^{20}/\text{cm}^3$ are shown in Table 5-1. For very small effective mass, there is a factor of 10 in the difference between some of the predicted σ and the obtained from DFT calculation. This could be due to the evaluation of the electrical conductivity at relaxation time lower than 10^{-14} s. The overall results show that the electrical conductivity model could be extrapolated to higher doping levels.

The main error in evaluating the $\mathbb{Z}T$ for this dataset is due to the error in predicting the Seebeck coefficient. This is expected since the relation between the Seebeck coefficient and the carrier concentration is not linear [26]. Band filling caused by higher carrier concentrations will increase the band utilization [26], which will increase the valley degeneracy factor. The Seebeck coefficient predictor is designed for IHP material with carrier concentrations of $10^{18}/\text{cm}^3$ and limited valley degeneracy factors, which explains the deviation shown in Table 5-1.

Table 5-1The prediction of electrical and thermal transports at $n=10^{20}/\text{cm}^3$

Material	m^*	σ (S/m)	Predicted σ (S/m)	S (μ V/K)	Predicted S (μ V/K)	k (W/m K)
KCdF₃	0.561903	50,443	50,176	-134	-423	2.571267
KMgF₃	0.629636	45,237	44,778	-146	-447	5.432402
KZnF₃	0.496899	53,460	56,739	-125	-402	3.688521
RbCaF₃	0.511916	50,729	55,075	-127	-408	3.66654
RbCdF₃	0.519840	53,826	54,236	-128	-409	2.650499
RbHgF₃	0.343885	72,124	81,986	-96	-350	1.430299
RbMgF₃	0.602691	49,315	46,780	-138	-438	4.976816
RbPbF₃	0.199682	57,586	141,193	-240	-303	2.74659

As for the second data sets, the predicted results in Table 5-2 show a promising consistency. The $\mathbb{Z}T$ values generally are not optimum at this carrier concentration [5,13,26] and temperature [13]. However, some of the results of the predicted lattice thermal conductivity are highly accurate. For example, CsSnI₃ thermal conductivity is reported to be in the range 0.34-0.42 W/m K obtained experimentally [48-49] compared to 0.31 W/m K as obtained by the developed model. Also, of the highest $\mathbb{Z}T$ are Ge- based halides (CsGeBr₃), which agrees with the experimental studies reported in the literature [50].

Table 5-2 The prediction of electrical and thermal transports at $n=10^{18}/\text{cm}^3$

Material	σ (S/m)	S (μ V/K)	k_l (W/m K)	$\mathbb{ZT}$
NaFeF₃	540	-409	1.57	0.017
SbOF₃	300	-543	0.74	0.036
KMnF₃	550	-406	1.76	0.015
KMnF₃	550	-406	1.80	0.015
ZnAgF₃	680	-327	0.99	0.022
NaVF₃	490	-364	2.73	0.007
CsPbCl₃	490	-427	0.24	0.110
CsSnI₃	520	-414	0.31	0.086
CsGeBr₃	3780	-261	0.10	0.773
RbGeI₃	4630	-257	0.32	0.284
CsSnI₃	520	-414	0.31	0.087
CsSnI₃	520	-414	0.31	0.086
LiCuF₃	5380	-253	2.30	0.045
LiVF₃	380	-479	2.99	0.009
LiMnF₃	460	-440	2.73	0.010
LiFeF₃	470	-435	2.28	0.012
LiFeF₃	470	-435	3.22	0.008
LiFeF₃	470	-435	2.00	0.013
RbCaF₃	536.3	-412	3.060	0.009
KHgF₃	779.2	-356	0.779	0.038
CsYbF₃	648.4	-382	3.102	0.009

Further investigation is required beyond this work. However, there is a lack of experimental and computed data for inorganic halide perovskites limiting the ability to compare the models to other references. It is to be noted that the repeated materials in Table 5-1 and Table 5-2 are of different space groups.

CHAPTER 6

CONCLUSION AND FUTURE WORK

In the past years, TE materials received a considerable attention due to their TE effects. Thermal energy is a by-product of any physical operating system, often thought of as a waste. Thermoelectric technologies can turn this waste to a renewable energy source. For this process to be practical it must be efficient. Thus, the need for efficient TE material with high ZT values. Generally, to evaluate ZT for novel materials, the electrical conductivity, Seebeck coefficient and the thermal conductivity must be evaluated first. These parameters can be evaluated through experimental approaches or using computational estimation. Since both approaches are very expensive and computationally demanding, an alternative approach is sought. ML methods are used in the discovery of TE materials.

6.1 Conclusion

This work seeks to bridge the gap between experimental and data driven approaches. ML provides a feasible venue to discover novel promising TE materials. The key for proper ML modeling is analyzing the thermoelectric effect with respect to its electrical and thermal transports. This is very difficult, due to the complex behavior caused by the changes in temperature or carrier concentration besides the fact that many physical mechanisms are involved at the microscale. Thus, using specific material conditions and focusing on a single group of materials facilitate the understanding of these complex trends

and hence affect preparing the problem for a proper ML modeling and feature analysis. This process presents three models, each one produces promising results and requires few features as inputs. These models are created to predict the electrical conductivity, Seebeck coefficient and the thermal conductivity for a IHP semiconductor of a doping level of $10^{18}/\text{cm}^3$ at room temperature. Using these models, the $\mathbb{ZT}$ is predicted for these materials. The $\mathbb{ZT}$ values for IHP at this doping level and temperature are generally low and there are higher achievable $\mathbb{ZT}$ values under higher temperature and doping level. However, these models can be used in an initial screening process.

6.2 Suggestions for Future Work

In this work, high-level descriptors were mainly used. However, there are other simple descriptors worth investigation. In previous works, bond length showed an interesting effect on thermal conductivity. Bond length can be used to calculate the elongation term, which is considered as a measure of the octahedral deformation for perovskites. With that term considered, one must consider the octahedron angle variance. Changing these parameters will alter $[\text{BX}_6]^{1-}$ octahedra, which can affect the electrical transports. Thus, studying these parameters and the relation between them and $\mathbb{ZT}$ will further enrich our understanding of these materials.

Throughout this work, only IHP at $T=300\text{K}$ and $n=10^{20}/\text{cm}^3$ are considered to facilitate the prediction of the $\mathbb{ZT}$ and the screening process. However, higher $\mathbb{ZT}$ values were reported at higher temperature and doping level. Some of the predictive models show a possible extrapolation for IHP at higher doping level, while other don't. To further expand these models to predict a higher range of $\mathbb{ZT}$ values, more data is required to continue this

progress. Obtaining more data will help assess the models and observe their limitations. Additionally, further studies are required to understand the electrical transports and thermal transports behavior of IHP at high ZT values.

REFERENCES

- [1] Apertet, Y., Ouerdane, H., Goupil, C. et al, A note on the electrochemical nature of the thermoelectric power, (2016).
- [2] Wang, T., Zhang, C., Snoussi, H., Zhang, G., Machine Learning Approaches for thermoelectric Materials Research. *Adv. Funct. Mater.*, (2020).
- [3] Iwasaki, Y., Takeuchi, I., Stanev, V. et al. Machine-learning guided discovery of a new thermoelectric material, (2019).
- [4] Gaultois, M., Oliynyk, A. O., Mar, A., Sparks, T. D., Mulholland, G. J., & Meredig, B, Perspective: Web-based machine learning models for real-time screening of thermoelectric materials properties, (2016).
- [5] Yang, J., Xi, L., Qiu, W. et al. On the tuning of electrical and thermal transport in thermoelectric: an integrated theory–experiment perspective, (2016).
- [6] Yan, Jun, Gorai, Prashun, Ortiz, Brenden, Miller, Sam, Barnett, Scott, Mason, Thomas, Stevanovic, Vladan , Toberer, Eric, Material descriptors for predicting TE, (2014).
- [7] Xie, Hongyao, Hao, Shiqiang, Bao, Jin-Ke, Slade, Tyler, Snyder, G. , Wolverton, Christopher, Kanatzidis, Mercouri. All-Inorganic Halide Perovskites as Potential TEMaterials: Dynamic Cation Off-Centering Induces Ultralow Thermal Conductivity, (2020).
- [8] Li, Yuheng, Yang, Kesong. High-throughput computational design of organic–inorganic hybrid halide semiconductors beyond perovskites for optoelectronics. *Energy & Environmental Science*, (2019).
- [9] Hongyao Xie, Shiqiang Hao, Jinke Bao, Tyler J. Slade, G. Jeffrey Snyder, Christopher Wolverton, and Mercouri G. Kanatzidis, All-Inorganic Halide Perovskites as Potential Thermoelectric Materials: Dynamic Cation off-Centering Induces Ultralow Thermal Conductivity, (2020).

- [10] Choudhary, K., Garrity, K.F., Sharma, V. et al. High-throughput density functional perturbation theory and machine learning predictions of infrared, piezoelectric, and dielectric responses. *npj Comput Mater* 6, 64, (2020).
- [11] Gorai, P., Stevanović, V., Toberer, E. Computationally guided discovery of thermoelectric materials. (2017).
- [12] Loftis, Christian, Yuan, Kunpeng, Zhao, Yong, Hu, Ming , Hu, Jianjun. (Lattice Thermal Conductivity Prediction using Symbolic Regression and Machine Learning, (2020).
- [13] Haque, M. A., Kee, S., Villalva, D. R., Ong, W.-L., Baran, D., Halide Perovskites: Thermal Transport and Prospects for Thermoelectricity. *Adv. Sci*, (2020).
- [14] Chicco, Davide, Jurman, Giuseppe. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, (2020).
- [15] Norouzzadeh, P., Vashaee, D. Classification of Valleytronics in Thermoelectricity, (2016).
- [16] Chen, Lihua, H. Tran, Rohit Batra, C. Kim and R. Ramprasad “Machine Learning Models for the Lattice Thermal Conductivity Prediction of Inorganic Materials, (2019).
- [17] Kumar, Anup , Muralidharan, Bhaskaran , Choi, Jung-Hae , Lee, Seung-Cheol , Bhattacharjee, Satadeep. AMMCR: Ab-initio model for mobility and conductivity calculation by using Rode Algorithm, (2019).
- [18] Parv Katyal, Madhav Rathi, Piyush Mehra and Amrish K. Panwar. “Evaluation of Figure of Merit of thermoelectric Materials Using Machine Learning”, (2020).
- [19] Wang, Z.-L., Adachi, Y. and Chen, Z.-C, Processing Optimization and Property Predictions of Hot-Extruded Bi–Te–Se Thermoelectric Materials via Machine Learning, (2020)

- [20] Li, Z., Xu, Q., Sun, Q., Hou, Z., Yin, W.-J. Thermodynamic Stability Landscape of Halide Double Perovskites via High-Throughput Computing and Machine Learning, (2019)
- [21] H. Taylor, Frisco Rose, Cormac Toher, Ohad Levy, Kesong Yang, Marco Buongiorno Nardelli, Stefano Curtarolo. A RESTful API for exchanging materials data in the AFLOWLIB.org consortium. Computational Materials Science, Volume 93, Pages 178-192. Richard, October (2014).
- [22] Verma, A. , Kumar, A. Bulk modulus of cubic perovskites, (2012).
- [23] S. P. Ong, S. Cholia, A. Jain, M. Brafman, D. Gunter, G. Ceder, and K. A. Persson. The Materials Application Programming Interface (API): A simple, flexible and efficient API for materials data based on REpresentational State Transfer (REST)principles.Computational Materials Science, (2015).
- [24] Glazer, A. M.. The classification of tilted octahedra in perovskites, (1972).
- [25] Abdel-Motaleb, I. and Qadri, S.M. Thermoelectric Devices: Principles and Future Trends, (2017).
- [26] Suwardi, Ady, Daniil Bash, H. K. Ng, J. R. Gómez, D. Repaka, P. Kumar and K. Hippalgaonkar. “Inertial effective mass as an effective descriptor for thermoelectric via data-driven evaluation.” Journal of Materials Chemistry 7, (2019).
- [27] Stability Trend of Tilted Perovskites, Fazel Shojaei and Wan-Jian Yin, The Journal of Physical Chemistry C, (2018).
- [28] Tewari, A., Dixit, S., Sahni, N., , Bordas, S. Machine learning approaches to identify and design low thermal conductivity oxides for thermoelectric applications. Data-Centric Engineering, (2020).
- [29] Niedermeier, Christian A. et al. “Electron effective mass and mobility limits in degenerate perovskite stannate BaSnO_3 .” Physical Review B 95, (2017).

- [30] S. R. Bhandari, D. K. Yadav, B. P. Belbase, M. Zeeshan, B. Sadhukhan, D. P. Rai, R. K. Thapa, G. C. Kaphle and M. P. Ghimire, Electronic, magnetic, optical and thermoelectric properties of $\text{Ca}_2\text{Cr}_{1-x}\text{Ni}_x\text{OsO}_6$ double perovskites, RSC Adv, (2020).
- [31] Faghaninia, Alireza et al. “Ab initio electronic transport model with explicit solution to the linearized Boltzmann transport equation.” Physical Review B 91, (2015).
- [32] Ricci, F. Dryad digital repository <https://doi.org/10.5061/dryad.gn001> (2017).
- [33] Bartel, Christopher, Sutton, Christopher, Goldsmith, Bryan, Ouyang, Runhai, Musgrave, Charles, Ghiringhelli, Luca, Scheffler, Matthias. New Tolerance Factor to Predict the Stability of Perovskite Oxides and Halides, (2019).
- [34] Kieslich, Gregor, Sun, Shijing, Cheetham, Tony. An Extended Tolerance Factor Approach for Organic-Inorganic Perovskites, (2015).
- [35] Travis, Will, Glover, Emily, Bronstein, Hugo, Scanlon, David, Palgrave, Robert. On the Application of the Tolerance Factor to Inorganic and Hybrid Halide Perovskites: A Revised System, (2016).
- [36] Jong, Un-Gi, Yu, Chol-Jun, Kye, Yun-Hyok, Jumabekov, Askhat, Utegulov, Zhandos, Islam, M.. Computational Prediction of Structural, Electronic, Optical Properties and Phase Stability of Double Perovskites K_2SnX_6 ($\text{X} = \text{I}, \text{Br}, \text{Cl}$), (2019).
- [37] Tao, Q., Xu, P., Li, M. et al. Machine learning for perovskite materials design and discovery. npj Comput Mater 7, 23, (2021).
- [38] Heesoo Park, Raghvendra Mall, Fahhad H. Alharbi, Stefano Sanvito, Nouar Tabet, Halima Bensmail, and Fedwa El-Mellouhi, The Journal of Physical Chemistry A, (2019).
- [39] Ricci, F., Chen, W., Aydemir, U. et al. An ab initio electronic transport database for inorganic materials. Sci Data 4, 170085, (2017)
- [40] Lili Xi, Jiong Yang, Lihua Wu, Jihui Yang, Wenqing Zhang, Band engineering and rational design of high-performance thermoelectric materials by first-principles, Journal of Materiomics, Volume 2, Issue 2, (2016).

- [41] Jeong, Changwook , Kim, Raseong , Luisier, Mathieu , Datta, Supriyo , Lundstrom, M.s. On Landauer versus Boltzmann and full band versus effective mass evaluation of thermoelectric transport coefficients. *Journal of Applied Physics*, (2010).
- [42] Yang, Dongwen , Lv, Jian , Zhao, Xingang , Qiaoling, Xu , Yuhao, Fu , Zhan, Yiqiang , Zunger, Alex , Zhang, Lijun. Functionality-Directed Screening of Pb-Free Hybrid Organic-Inorganic Perovskites with Desired Intrinsic Photovoltaic Functionalities. *Chemistry of Materials*, (2017).
- [43] Gibbs, Zachary , Ricci, Francesco , Li, Guodong , Zhu, Hong , Persson, Kristin , Ceder, Gerbrand , Hautier, Geoffroy , Jain, Anubhav , Snyder, G. Effective mass and Fermi surface complexity factor from ab initio band structure calculations. *npj Computational Materials*, (2017).
- [44] Witkoske, Evan , Wang, Xufeng , Askarpour, Vahid , Maassen, Jesse , Lundstrom, M.s. Thermoelectric Band Engineering: The Role of Carrier Scattering. *Journal of Applied Physics*. 122, (2017).
- [45] Alam, Hilaal. A review on the enhancement of figure of merit from bulk to nano-thermoelectric materials, (2013).
- [46] Gibbs, Zachary , Kim, Hyun-Sik , Wang, Heng , Snyder, G. Band gap estimation from temperature dependent Seebeck measurement—Deviations from the $2e|S|_{\max}T_{\max}$ relation. *Applied Physics Letters*, (2015)
- [47] Choudhary, Kamal , Garrity, Kevin , Reid, Andrew , DeCost, Brian , Biacchi, Adam , Hight Walker, Angela , Trautt, Zachary , Hattrick-Simpers, Jason , Kusne, A. , Centrone, Andrea , Davydov, Albert , Jiang, Jie , Pachter, Ruth , Cheon, Gwoon , Reed, Evan , Agrawal, Ankit , Qian, Xiaofeng , Sharma, Vinit , Zhuang, Houlong , Tavazza, Francesca. JARVIS: An Integrated Infrastructure for Data-driven Materials Design. (2020).

- [48] Lee, Woochul et al. "Ultralow thermal conductivity in all-inorganic halide perovskites." *Proceedings of the National Academy of Sciences of the United States of America* vol. 114,33, (2017).
- [49] Haeger, Tobias , Heiderhoff, Ralf , Riedl, Thomas. Thermal properties of metal-halide perovskites. *Journal of Materials Chemistry*, (2020).
- [50] Jong, Un-Gi , Yu, Chol-Jun , Kye, Yun-Hyok , Hong, Song-Nam , Kim, Hyon-Gyong. Manifestation of the thermoelectric properties in Ge-based halide perovskites, (2020). |

Vitae

Name	:Hassan Ali Alsalman
Nationality	:Saudi
Date of Birth	:2/5/1996
Email	:alsalman.a.hasan@gmail.com
Address	:36343- AlAhsaa, -Eastern province- KSA
Academic Background	:Hassan A. Alsalman received his B.Sc. degree in Electrical Engineering from King Fahd University of Petroleum and Minerals Dhahran, Saudi Arabia, in 2018. He is currently perusing M.Sc. degree in Electrical Engineering at King Fahd University of Petroleum and Minerals. His interests include: Machine learning applications in material discovery and seismology, Fiber sensors and application in seismic reflection survey.