
Identifying Descriptors for the In-silico, High-throughput Discovery of the Thermal Insulators for Thermoelectric Applications

Master thesis by Bo Zhao

Date of submission: September 2, 2022

1. Review: Prof. Dr. Hongbin Zhang
2. Review: Prof. Dr. Matthias Scheffler
Darmstadt



TECHNISCHE
UNIVERSITÄT
DARMSTADT

Department of Materials and
Geosciences

Institute of Materials
Science

Abstract

In this thesis, we perform high-throughput screening calculations for the thermal conductivity (κ) at 300 K of 152 materials using first-principles methods. Due to the various approximations involved in these calculations, especially the low-order treatment of anharmonicity, a first set of 49 calculations is used to validate the computational approach. In general, this shows that the calculations agree relative well with experiments, but also that stronger anharmonic effects indeed lead to increased prediction errors for thermal insulators. In a second step, 103 additional materials are investigated, 83 of which are found to be potential thermal insulators with $\kappa < 10\text{W/mK}$. Since the degree of anharmonicity affects the accuracy of our computational predictions, we further investigate to which extent anharmonicity can be quantitatively measured. To do this, we compare between three metrics for anharmonicity (γ , σ^A and σ_{os}^A) and discuss their correlation with thermal conductivity. While σ^A and σ_{os}^A show a relatively promising correlation with κ , the Grüneisen parameters γ do not. To refine this finding and to find even better descriptors for the thermal conductivity, we eventually use a machine-learning based symbolic regression approach, i.e., the SISO (sure-independence screening and sparsifying operator) method. By applying it to our data, we can show that the models identified by the SISO approach show promising predictive accuracy and thus have the potential to facilitate the discovery of thermal insulators in future.

Contents

1. Introduction	5
1.1. Application Background	5
1.2. High-throughput Screening and Machine Learning	6
1.3. An Overview of Our Work	6
2. Theory	8
2.1. Lattice Dynamics from First Principles	8
2.1.1. Electronic structure	8
2.1.2. Density Functional Theory	12
2.1.3. Solid State Physics	17
2.2. Statistical Theories	36
2.2.1. LASSO (Least Absolute Shrinkage and Selection Operator) and Compressed Sensing	36
2.2.2. Symbolic Regression	37
2.2.3. SISSO (sure independence screening and sparsifying operators) . .	38
2.2.4. Cross Validations of Models	39
3. Results and Discussion	42
3.1. Numerical Settings	42
3.2. Preliminary Convergence Tests	42
3.3. Phono3py Results of Thermal Conductivity	47
3.4. Anharmonicity Measure	55
3.5. SISSO Regression for Thermal Conductivity	58
3.5.1. Regression including γ	61
3.5.2. Regression including σ^A , σ_{os}^A and γ	67
4. Conclusions	72
Appendices	78

A. Results of High-Throughput Screening	78
B. Cross Validation Results of SISSO	85

1. Introduction

1.1. Application Background

There are a number of applications in industry science and in our everyday life where thermal insulation plays an important role. Thermoelectrics is a particularly important application as it is not only a possible route towards saving energy, but also can replace standard chlorofluorocarbon refrigerants[7]. The efficiency of a thermoelectric material is described by the figure of merit

$$ZT = \frac{S^2 \sigma T}{\kappa}, \quad (1.1)$$

where S is the Seebeck coefficient, T is the temperature, σ is the electrical conductivity and κ is thermal conductivity. Eq. 1.1 suggests that efficient thermoelectrics should have as low κ as possible to ensure heat to be converted to electricity other than being transported.

Various attempts have been made to reduce the thermal conductivity. Intrinsically, one can start from choosing the right chemical composition. As a rule of thumbs, heavy elements, soft bonds, complex unit cell, etc. usually lead to low κ . Alternatively, thermal conductivity can be reduced extrinsically, e.g., by structure disorder and defects. For example, alloying and doping induced point defects have been used successfully in many thermoelectric materials system, such as $\text{PbTe}_x\text{S}_{1-x}$, SiGe and $\text{Cu}_2\text{Se}_{1-x}\text{I}_x$ [26]. Generally, such approaches are more successful if the materials has low κ already. It is thus important to identify more insulating materials and to identify reliable design rules that go beyond the aforementioned rules of thumb.

1.2. High-throughput Screening and Machine Learning

Despite the strong demand motivated by the vast applications, relatively few materials have been studied so far, since the accurate calculation of thermal conductivity, even for simple bulk materials, is still computationally challenging due to need of treating anharmonic effects. In this work, we use a low-order treatment of anharmonicity to screen for potential thermal insulators among a relatively large database of 152 materials. The aim of this is to reveal general trends for κ within different materials classes. To be able to investigate so many materials, a tradeoff between accuracy and the computational speed has to be taken into account. This tradeoff inaccuracy needs to be validated and checked against experimental results.

Once sufficient amount of thermal conductivity data is available, machine learning can be used for identifying descriptors formed by a set of parameters capturing the underlying mechanism of materials property. By this means, even more rapid estimations for κ can be obtained in a rapid fashion.

1.3. An Overview of Our Work

In this thesis, a systematic investigation will be carried out over five classes of materials, i.e., rock salt, zinc blende, fluorite, half-Heusler and chalcopyrite, for the understanding of thermal conductivity (κ) and for seeking promising thermal insulators ($\kappa < 10\text{W/m/K}$).

The thesis work contains the following four parts: A summary of the underlying theories, a description of the screening protocol and how we calculated the thermal conductivity of all materials, a comparison between different measures of the anharmonicity of a material, and finally the generation of machine learning models for identifying κ . In our second part, a high-throughput screening of thermal conductivity at 300 K over 152 materials is performed. Thermal conductivity is calculated from the phonon-phonon interaction using single mode relaxation time (SMRT) method. 49 of these materials are used to validate the approach by comparing with experimental results.

In the third part, a comparative study of three possible anharmonic metric, σ^A , σ_{os}^A and the Grüneisen parameter (γ) is performed, where σ^A is calculated from *ab initio* molecular dynamics (aiMD) and σ_{os}^A is calculated using a configuration generated via the

harmonic approximation. γ is obtained from the third order force constants same as the high-throughput screening for κ .

In the last part, a machine-learning technique, the sure-independence screening and sparsifying operator (SISSO) approach, is introduced to run regressions for the complex description of thermal conductivity using σ^A , σ_{os}^A , γ , Θ_D , etc. In this machine learning analysis, the 49 experimental values of thermal conductivity will be used as the property data set, while the training data set includes 16 physical quantities. These are used to form the feature space for the symbolic regression, before solving the fitting problem by compressed sensing. We run such regression for two purposes. One is to obtain better models of κ that are computationally cheaper than first principles calculations. Second, machine learning can help reveal what factors contribute most to thermal conductivity of materials.

2. Theory

2.1. Lattice Dynamics from First Principles

2.1.1. Electronic structure

In classical mechanics, motion of equation is described by Newton's second law. For solids, however, the interactions between any two particles (can be nucleus or electron) should not be neglected. The solution to such many-body problem is usually written as the form of the wave function and to be solved iteratively. Plane wave is a particular form efficient for dealing with periodic system with periodic potential.

Bloch Theorem and Periodic Potential

Plane wave can be viewed as the Fourier transform from real lattice space to reciprocal lattice space. The reciprocal lattice is known as the set of all wave vectors \mathbf{K} that yields plane waves with the periodicity of a given Bravais lattice vector \mathbf{R} . They are connected by the relation

$$e^{i\mathbf{K}\cdot\mathbf{R}} = 1 \quad (2.1)$$

According to the Bloch theorem, the eigenstates of the Hamiltonian takes the form

$$\psi_{n\mathbf{k}}(\mathbf{r}) = \exp(i\mathbf{k}\cdot\mathbf{r})u_{n\mathbf{k}}(\mathbf{r}), \quad (2.2)$$

where ψ is the eigenstate of the Hamiltonian, n is the band index, \mathbf{r} is lattice vector and \mathbf{k} is a wave vector. The cell periodic term $u_{\mathbf{k}}(\mathbf{r})$ can be expanded in a plane wave basis set

whose wave vectors are the reciprocal lattice vectors of the crystal,

$$u_{n\mathbf{k}}(\mathbf{r}) = \sum_{\mathbf{K}} c_{i\mathbf{K}} \exp(n\mathbf{K} \cdot \mathbf{r}) \quad (2.3)$$

where \mathbf{K} is the reciprocal lattice vector, and $c_{i\mathbf{K}}$ are expansion coefficients of $u_{n\mathbf{k}}$, while $u_{n\mathbf{k}}$ itself is periodic with respect to Bravais lattice vectors \mathbf{R} , $u_{n\mathbf{k}}(\mathbf{r} + \mathbf{R}) = u_{n\mathbf{k}}(\mathbf{r})$. With this relation we see that Bloch's theorem

$$\psi(\mathbf{r} + \mathbf{R}) = e^{i\mathbf{k} \cdot \mathbf{R}} \psi(\mathbf{r}) \quad (2.4)$$

holds.

For each Bravais lattice vector \mathbf{R} , we can define an translation operator \hat{T} used to shift the the wave function by \mathbf{R} such that $\hat{T}\psi(\mathbf{r}) = \psi(\mathbf{r} + \mathbf{R})$. If we apply \hat{T} to $H\psi$ and keep in mind that the Hamiltonian is periodic, we get

$$\hat{T}H\psi = H(\mathbf{r} + \mathbf{R})\psi(\mathbf{r} + \mathbf{R}) = H(\mathbf{r})\psi(\mathbf{r} + \mathbf{R}) = H\hat{T}\psi(\mathbf{r}) \quad (2.5)$$

Given that ψ is an arbitrary function, we find by comparing the first and the last terms of Eq. 2.5 that the translation operator \hat{T} commutes with the Hamiltonian. Therefore, Bloch wave functions $\psi_{n\mathbf{k}}$, which are eigenvectors of \hat{T} , are simultaneously the complete set of eigenvectors for Hamiltonian \hat{H} . This helps reduce the searching for eigenvectors of Hamiltonian from the entire Hilbert space down to N-dimensional eigenspaces, where N is the number of particles in the first Brillouin zone. The solutions to the wavefunctions for allowed \mathbf{k} in this way are superpositions of plane waves of wave vector \mathbf{k} and of wave vectors differing from \mathbf{k} by a reciprocal lattice vector.

Since for a given \mathbf{k} there are multiple solutions to the Schrödinger equation, subscript n in Eq. 2.2 and 2.3 is used to distinguish wave functions labeled with same \mathbf{k} . In the limit of large crystal, for each n the energy level $\varepsilon_n(\mathbf{k})$ is a continuous function that forms an energy band in the Brillouin zone.

Brillouin Zone Sampling

To evaluate some quantities one often has to get the weighted integral of the density. For instance, the total energy requires the integration over all the electron energy levels at all \mathbf{k} vectors. In this case, the density of states (DOS) given by integration of \mathbf{k} over the

Brillouin zone for all energy bands n

$$g(\varepsilon) = \sum_i \int \frac{d\mathbf{k}}{4\pi^3} \delta(\varepsilon - \varepsilon_n(\mathbf{k})) \quad (2.6)$$

can be helpful.

In practice, this integral is evaluated as a sum

$$g(\varepsilon) = \sum_{\mathbf{k}} \omega_{\mathbf{k}} g_{\mathbf{k}}(\varepsilon) \quad (2.7)$$

where $\omega_{\mathbf{k}}$ are integration weights that need to be carefully chosen. Using the mean-value theorem of integral calculus, one can always find a single point within an interval such that

$$g(\varepsilon) = \sum_{\mathbf{k}} \omega_{\mathbf{k}} g_{\mathbf{k}}(\varepsilon) = V \times g_{\mathbf{k}}(\varepsilon) \quad (2.8)$$

where V is the volume of that interval, usually a polyhedron in 3D. For sc, bcc and fcc lattices, these points are known. For more complex lattices there are methods to sample the \mathbf{k} -mesh. The Monkhorst-Pack mesh is one of those approaches to generate a uniform \mathbf{k} -mesh. It tells that the commensurate \mathbf{k} -points are given by

$$\vec{k}_{lmn} = \sum_i^3 \frac{2n_i - N_i - 1}{2N_i} \vec{G}_i \quad (2.9)$$

where N_i is the number of \mathbf{k} -points in each direction as specified by the \mathbf{k} -grid parameter in the calculations, $n_i = 1, \dots, N_i$, and \vec{G}_i is the reciprocal lattice vectors. \mathbf{k} -grid parameter has to be checked by doing convergence tests. Usually we want to evaluate the energy difference between two structures, or to calculate similar structures in different supercells. In both cases, \mathbf{k} -points density should stay the same, so that the not-fully-converged error can cancel out.

Born-Oppenheimer Approximation (BOA)

The Born-Oppenheimer (BO) approximation is based on the assumption that (1) nuclei are much heavier and that electrons hence move much faster, so that they always respond instantaneously to changes in the atomic positions. As a consequence, the many-particle Schrödinger equation can be separated into nuclear and electronic equations.

In more detail, the Hamiltonian for all particles (including electrons and nuclei) reads

$$\begin{aligned}
H = & T_{\text{el}} + T_{\text{nuc}} + U_{\text{el-el}} + U_{\text{nuc-el}} + U_{\text{nuc-nuc}} \\
& - \sum_{i=1}^{N^{\text{el}}} \frac{\hbar^2 \nabla_i^2}{2m} - \sum_{n=1}^N \frac{\hbar^2 \nabla_n^2}{2M_n} + \frac{1}{4\pi\epsilon_0} \frac{1}{2} \sum_{i,j=1; i \neq j}^{N^{\text{el}}} \frac{e^2}{|\mathbf{r}_i - \mathbf{r}_j|} \\
& - \frac{1}{4\pi\epsilon_0} \sum_{n=1}^N \sum_{i=1}^{N^{\text{el}}} \frac{Z_n e}{|\mathbf{r}_i - \mathbf{R}_n|} + \frac{1}{4\pi\epsilon_0} \frac{1}{2} \sum_{n,n'=1; n \neq n'}^N \frac{Z_n Z_{n'}}{|\mathbf{R}_n - \mathbf{R}_{n'}|}
\end{aligned} \tag{2.10}$$

where ∇_i^2 and ∇_n^2 are the Laplacians for electrons and nuclei. The first two terms of the equation are kinetic energies for nuclei and electrons, the last three terms describe Coulomb interactions between electrons, nuclei, and electron and nuclei.

The full, time-dependent Schrödinger equation subject to this Hamiltonian is given by

$$H\chi(\mathbf{x}, \mathbf{r}, t) = i\hbar \frac{\partial \chi}{\partial t} \tag{2.11}$$

where $\chi(\mathbf{x}, \mathbf{r}, t)$ is the full wave function that can be written in the separable form

$$\chi(\mathbf{x}, \mathbf{r}, t) = \psi^{\text{el}}(\mathbf{x}, \mathbf{r}) \psi^{\text{nuc}}(\mathbf{r}, t) \tag{2.12}$$

Due to BO approximation the electron wave functions $\psi^{\text{el}}(\mathbf{x}, \mathbf{r})$ do not depend explicitly on time. Instead they are functions of the nuclear coordinates \mathbf{r} . Substituting Eq. 2.12 into Eq. 2.11 leads to

$$\begin{aligned}
& - \frac{\hbar^2}{2} \left[\sum_{i=1}^{N^{\text{el}}} \frac{\psi^{\text{nuc}} \nabla_i^2 \psi^{\text{nuc}}}{m} + \sum_{n=1}^N \frac{\psi^{\text{nuc}} \nabla_n^2 \psi^{\text{el}}}{M_n} + \sum_n \frac{2 \nabla_n \psi^{\text{el}} \nabla_n \psi^{\text{nuc}}}{M_n} + \sum_{n=1}^N \frac{\psi^{\text{el}} \nabla_n^2 \psi^{\text{nuc}}}{M_n} \right] \\
& + (U^{ee} + U^{eZ} + U^{ZZ}) \psi^{\text{el}} \psi^{\text{nuc}} = i\hbar \psi^{\text{el}} \frac{\partial \psi^{\text{nuc}}}{\partial t}.
\end{aligned} \tag{2.13}$$

Here, the three potential terms in Eq. 2.11 are replaced by U^{ee} , U^{eZ} and U^{ZZ} for short. According to BOA, electrons are always staying at the equilibrium states and the third term in the squared bracket in Eq. 2.13 vanishes. The second term is much smaller compared to the first term for the mass of nuclei is much greater than that of electron under the BO approximation. This term is also negligible.

The time-independent electron part can be separated out from Eq. 2.13 as

$$-\frac{\hbar^2}{2} \sum_{i=1}^{N_{el}} \frac{\nabla_i^2 \psi^{el}}{m} + (U^{ee} + U^{ez})\psi^{el} = \varepsilon \psi^{el} \quad (2.14)$$

where $\varepsilon = \varepsilon_0(\mathbf{r})$ is the ground state energy of the electron. Inserting the right hand side into Eq. 2.13, we obtain the time-dependent nuclear Schrödinger equation, in which the electron energy enters as part of the partial field

$$-\frac{\hbar^2}{2} \sum_{n=1}^N \frac{\nabla_i^2 \psi^{nuc}}{M_n} + (U^{ZZ} + \varepsilon_0)\psi^{nuc} = i\hbar \frac{\partial \psi^{nuc}}{\partial t} \quad (2.15)$$

In other words, the potential $V(\mathbf{r}) = U^{ZZ} + \varepsilon_0$ ¹ depends only on the position of nuclei, which will be varied to find the ground state of the whole system.

2.1.2. Density Functional Theory

In density functional theory, the total energy of a system of many interacting particles can be expressed as a functional of the ground state density $n_0(\mathbf{r})$, and that the density $n(\mathbf{r})$ is obtained from an auxiliary function ψ_k , normally called Kohn-Sham wave function or orbital. They are related by

$$n(\mathbf{r}) = \sum_{k=1}^N |\psi_k(\mathbf{r})|^2 \quad (2.16)$$

The above statement is proven in the first and second Hohenberg-Kohn theorem. The first states that for any system of interacting electrons in an external potential, the potential is determined uniquely by the ground-state density. [9]. The second theorem states that the ground state energy is a functional of the density and assumes a minimum at the ground state.

Proof of theorem I. Assume there exist two different potentials $V_1(r) \neq V_2(r)$ that yield the same ground-state density. They lead to different Hamiltonians $\hat{H}_1 \neq \hat{H}_2$, and to different

¹We treat nuclei as classical particles and denote the potential of them as $V(\mathbf{r})$.

ground-state wave functions $\Psi_1 \neq \Psi_2$. Based on the variation principle, it follows that

$$\begin{aligned} E_1 &= \langle \Psi_1 | \hat{H}_1 | \Psi_1 \rangle < \langle \Psi_2 | \hat{H}_1 | \Psi_2 \rangle \\ E_2 &= \langle \Psi_2 | \hat{H}_2 | \Psi_2 \rangle < \langle \Psi_1 | \hat{H}_2 | \Psi_1 \rangle \end{aligned} \quad (2.17)$$

Then we have the following relations

$$\begin{aligned} E_1 &< \langle \Psi_2 | \hat{H}_2 - V_2 + V_1 | \Psi_2 \rangle = E_2 + \int d^3r (V_1(r) - V_2(r)) n_2(r) \\ E_2 &< \langle \Psi_1 | \hat{H}_1 - V_1 + V_2 | \Psi_1 \rangle = E_1 + \int d^3r (V_2(r) - V_1(r)) n_1(r) \end{aligned} \quad (2.18)$$

Summing over these two equations, we get the contradiction:

$$E_1 + E_2 < E_1 + E_2 \quad (2.19)$$

Proof of theorem II. For a given external potential $V_{ext}(\mathbf{r})$, we first define the associated total energy to be

$$\begin{aligned} E_{HK}[n] &= T[n] + E_{int}[n] + E_{ext}[n] \\ &= T[n] + E_{int}[n] + \int n(\mathbf{r}) V_{ext}(\mathbf{r}) d\mathbf{r} \end{aligned} \quad (2.20)$$

Assume n_0 is the ground state density with the corresponding Hamiltonian H_0 generated from the true external potential V_{ext} and hence its ground state wave function Ψ_0 . Let n_1 be some other density, corresponding to another wave function Ψ_1 . This density can also be as the wave function for H_0 , leading to the inequality given by

$$E_0 = E_{HK}[n_0] = \langle \Psi_0 | H_0 | \Psi_0 \rangle < \langle \Psi_1 | H_0 | \Psi_1 \rangle = E_{HK}[n_1] \quad (2.21)$$

In density functional theory, the Schrödinger equation with an effective-independent Hamiltonian for single-spin orbital is given by

$$\left[-\frac{1}{2} \nabla^2 - \sum_n \frac{Z_n}{|\mathbf{r} - \mathbf{R}_n|} + \int d^3r' n(\mathbf{r}') \frac{1}{|\mathbf{r} - \mathbf{r}'|} + V_{xc}[\mathbf{n}](\mathbf{r}) \right] \psi_k(\mathbf{r}) = \varepsilon_k \psi_k(\mathbf{r}) \quad (2.22)$$

where the first three terms are kinetic energy, the electrostatic interactions between electrons and nuclei and the electrostatic energy of the electron in the field generated by

the total electron density $n(\mathbf{r})$. The fourth term contains the many-body effects, forming together the so-called exchange-correlation potential.

The energy-functional for a many-electron system with electronic interactions included takes the form

$$E[n] = T[n] + \int d^3r n(\mathbf{r}) V_{ext}(\mathbf{r}) + \frac{1}{2} \int d^3r \int d^3r' n(\mathbf{r}') \frac{1}{|\mathbf{r} - \mathbf{r}'|} n(\mathbf{r}) + E_{xc}[n] \quad (2.23)$$

Exchange and Correlation Functional

The accuracy of the calculated total energy largely depends on the approximation of exchange and correction energy E_{xc} . The term $E_{xc}[n]$ Based on the assumption that the exchange and correlation is the functional of electron density, it is useful to express $E_{xc}[n]$ in the form

$$E_{xc}[n] = \int d\mathbf{r} n(\mathbf{r}) \epsilon_{xc}(n(\mathbf{r})) \quad (2.24)$$

where $\epsilon_{xc}(n(\mathbf{r}), n(\mathbf{r}))$ is the energy per electron at the position \mathbf{r} depends on the density in some neighborhood of \mathbf{r} . E_{xc} can be written in terms of density like potential term due to the exchange-correlation hole averaged over the interaction from fully independent to fully correlated.

The *Local Density Approximation* assumes that the the exchange and the correlation energies are the same as that in a homogeneous electron gas and that the exchange-correlation energy can thus simply be calculated as the integral of not necessarily homogeneous density,

$$\begin{aligned} E_{xc}^{\text{LDA}}[n(\mathbf{r})] &= E_x^{\text{LDA}}[n(\mathbf{r})] + E_c^{\text{LDA}}[n(\mathbf{r})] \\ &= \int d^3r \epsilon_x^{\text{LDA}}(n(\mathbf{r})) \cdot n(\mathbf{r}) + \int d^3r \epsilon_c^{\text{LDA}}(n(\mathbf{r})) \cdot n(\mathbf{r}) \end{aligned} \quad (2.25)$$

where the exchange energy per electron $\epsilon_x^{\text{LDA}}(n(\mathbf{r}))$ is

$$\epsilon_x^{\text{LDA}}(n(\mathbf{r})) = -\frac{3}{4} \left(\frac{3}{\pi} \right)^{1/3} n^{1/3}(\mathbf{r}) \quad (2.26)$$

and the correlation term $\epsilon_c^{\text{LDA}}(n(\mathbf{r}))$ is parameterized as

$$\epsilon_c^{\text{LDA}}(n(\mathbf{r})) = \frac{0.1423}{1 + 1.0529(3/4\pi n)^{1/6} + 0.3334(3/4\pi n)^{1/3}} \quad (2.27)$$

based on Quantum Monte Carlo methods[18].

Despite its success of homogeneous electron gas, LDA has the following shortcomings: (1) It is problematic for localized electrons, (2) it overestimates cohesive energies and underestimates the corresponding bond lengths and (3) erroneous self-interaction included [14].

The *Generalized Gradient Approximation (GGA)* is a more accurate approach in which not just the density, but also the derivatives of density are taken into account. GGA combines the idea of Taylor expansion with a generalized gradient given by

$$s = \frac{|\nabla n(\mathbf{r})|}{2k_F n(\mathbf{r})} \quad (2.28)$$

with $k_F = (3\pi^2 n)^{1/3}$ representing the magnitude of the local Fermi wave vector.

The GGA energy has the form as

$$E_{xc}^{\text{GGA}} = \int d^3r n(\mathbf{r}) \epsilon_{xc} F_{xc}[s] \quad (2.29)$$

where ϵ_{xc} is the exchange correlation energy per particle of the homogeneous gas as shown in Eq. 2.25 and F_{xc} is the enhancement factor. There is only one type of LDA, but there are several kinds of GGA due to different parameterizations, among which some are empirical, some semi-empirical, and others *ab initio*. Widely used GGA functionals are the Perdew and Yang (PW91) [4], the Becke (B88 [2], B97 [3]), Perdew, Burke, Ernzerhof (PBE) [19], just to name a few. One of the major drawback of both LDA and GGA is that the exchange correlation does not cancel the self-interaction present in the Hartree energy.

The enhancement factor $F_{xc}[s]$ for exchange term under GGA is given by

$$F_x = 1 + \mu s^2 + \dots \quad (2.30)$$

where the first term comes from L(S)DA, and the coefficient μ in the second term specifies

the s^2 contribution based on GGA. The GGA correlation energy takes the form

$$E_c[n] = \int d^3r n(\mathbf{r}) \{ \epsilon_c^{unif}(n(\mathbf{r})) + \beta t^2(\mathbf{r}) + \dots \} \quad (2.31)$$

where $\epsilon_c^{unif}(n(\mathbf{r}))$ is the correlation energy per particle of the uniform gas, β is the coefficient and t is the reduced density gradient for the correlation.

The calculations done in Chapter 3 used PBEsol as xc functional. PBEsol [6] stands for PBE functional revised for solids that restores the density-expansion gradient for exchange in solids, while the original PBE was designed in molecules. It is able to give more accurate results in describing the equilibrium properties of densely packed solids and their surfaces.

For PBEsol, $\mu = 2\mu_{GE}$ with $\mu_{GE} = 10/81 \approx 0.1235$ and $\beta = \beta_{GE} = 0.0667$, where μ_{GE} and β_{GE} are coefficients for gradient expansion. However, it does not well agree with densely packed solids. In PBEsol, $\beta = 0.046$ and $\mu = \mu_{GE}$ are chosen, making it exact for solids under intense compression and can give significantly better equilibrium lattice constants and surface energies. However, PBEsol is not expected to give good atomization energies, for which PBE is superior.

2.1.3. Solid State Physics

Harmonic Approximation of Lattice Dynamics

In real materials, atoms are not static, but oscillate around the equilibrium sites. If the assumption that for solid materials ions vibrate around their equilibrium positions, i.e., the Bravais lattice sites, with deviation much smaller than the interionic spacing holds, potential energy can be approximated by a Taylor expansion in terms of ion displacements $u(R)$ up to second order:

$$U(R) = \frac{N}{2} \left(\sum \Phi(R) + \sum u(R) \nabla \Phi(R) + \frac{1}{2} \sum u^2(R) \nabla^2 \Phi(R) + O(u^3(R)) \right) \quad (2.32)$$

$$\approx U^{eq} + \frac{1}{2} \sum_{\alpha, \beta}^N \sum_{\mu, \nu=x,y,z} u_{\alpha, \mu}(R) \Phi_{\mu, \nu}^{\alpha, \beta}(R - R') u_{\beta, \nu}(R') \quad (2.33)$$

The first order term in Eq. 2.32 vanishes due to the fact that the forces are zero when atoms are at the equilibrium positions, thus only the equilibrium and harmonic terms are left as shown in Eq. 2.33. Here, the equation is written in matrix form which is more convenient for higher dimensional cases. $\Phi_{\mu, \nu}^{\alpha, \beta}(R - R')$ is the second derivative of the potential with respect to the positions denoted by the subscripts $\{\mu, \nu\}$ and is given by

$$\Phi_{\mu, \nu}^{\alpha, \beta}(R - R') = \frac{\partial^2 \Phi^{\alpha, \beta}}{\partial u_{\alpha, \mu}(R) \partial u_{\beta, \nu}(R')} \quad (2.34)$$

where $\Phi^{\alpha, \beta}$ is the potential between atom α and β . The equilibrium potentials are nothing but the minimum of the potential energy surface (PES), i.e., the ground state energy obtained from *ab initio* calculations.

Using Eq. 2.33, one obtain the atomic equation of motion that reads

$$M_{\alpha} \ddot{u}_{\alpha, \mu}(t) = - \sum_{\beta, \nu} \Phi_{\mu, \nu}^{\alpha, \beta} u_{\beta, \nu} \quad (2.35)$$

where $u_{\alpha, \mu}^{\ddot{}}(t)$ is the second order time derivative of displacement for atom α . An ansatz for such a differential equation is an exponential function:

$$u_{\alpha, \mu} = u_{mq}(\alpha, \mu) e^{i\vec{q}\vec{R}_{\mu}} e^{-i\omega t} \quad (2.36)$$

Here, ω is the frequency of the normal modes of vibration. $u_{mq}(\alpha, \mu)$ is the polarization vector along μ direction. This form of displacement function is analogous to Bloch waves in describing electron wave function. Born-von Karman periodic condition is used to define the \vec{q} vector in Eq. 2.36, which requires that: $u_{\alpha, \mu}(R + N_i a_i) = u_{\alpha, \mu}(R)$, where N_i is any integer along i direction and a_i is the respective primitive lattice vector. Applying this to Eq. 2.35, one found that the \vec{q} take the form:

$$\vec{q} = \frac{n_1}{N_1} \mathbf{b}_1 + \frac{n_2}{N_2} \mathbf{b}_2 + \frac{n_3}{N_3} \mathbf{b}_3 \quad (2.37)$$

where \mathbf{b}_i are the reciprocal lattice vectors that satisfy $\mathbf{b}_i \mathbf{a}_i = 2\pi$.

Substituting Eq. 2.36 into the equation of motion (Eq. 2.35) and by cancels the time-dependent term $e^{-i\omega t}$ so that we end up with

$$M_\alpha \omega_{mq}^2 u_{mq}(\alpha\mu) e^{i\vec{q}\vec{R}_\kappa} = \sum_{\kappa'\beta\nu} \Phi_{\alpha\mu, \beta\nu}^{\kappa\kappa'} u_{mq}(\beta\nu) e^{i\vec{q}\vec{R}_{\kappa'}} \quad (2.38)$$

Here $\Phi_{\alpha\mu, \beta\nu}^{\kappa\kappa'}$ is the force constants of force acting on atom α with direction μ , induced by the displacement of atom β along direction ν . Multiplying with $e^{-i\vec{q}\vec{R}_\kappa}$ on both sides, gives

$$M_\alpha \omega_{mq}^2 u_{mq}(\alpha\mu) = \sum_{\beta\nu} \left\{ \sum_{\kappa'} \Phi_{\alpha\mu, \beta\nu}^{\kappa\kappa'} e^{i\vec{q}(\vec{R}_{\kappa'} - \vec{R}_\kappa)} \right\} u_{mq}(\beta\nu) \quad (2.39)$$

Eq. 2.39 is not a generalized form of eigenvalue equation due to the existence of mass matrix and can be solved by redefining the ansatz as

$$u_{\alpha, \mu} = \frac{1}{\sqrt{M_\alpha}} c_{mq}(\alpha, \mu) e^{i\vec{q}\vec{R}_\mu} e^{-i\omega t} \quad (2.40)$$

Now the equations of motion take the form:

$$\omega_{mq}^2 c_{mq}(\alpha\mu) = \sum_{\beta\nu} \left\{ \sum_{\kappa'} \frac{1}{\sqrt{M_\alpha M_\beta}} \Phi_{\alpha\mu, \beta\nu}^{\kappa\kappa'} e^{i\vec{q}(\vec{R}_{\kappa'} - \vec{R}_\kappa)} \right\} c_{mq}(\beta\nu) \quad (2.41)$$

The curly bracket on the right hand side converts the equation from \vec{R} dependent to \vec{q} dependent and can be replaced by a new variable called *dynamical matrix*

$$D_{\alpha\mu, \beta\nu}^{\kappa\kappa'} = \sum_{\kappa'} \frac{1}{\sqrt{M_\alpha M_\beta}} \Phi_{\alpha\mu, \beta\nu}^{\kappa\kappa'} e^{i\vec{q}(\vec{R}_{\kappa'} - \vec{R}_\kappa)} \quad (2.42)$$

Eq. 2.41 can now be rewritten as its most general form:

$$\omega_{mq}^2 c_{mq}(\alpha\mu) = \sum_{\beta,\nu} D_{\alpha\mu,\beta\nu}^{\kappa\kappa'} c_{mq}(\beta\nu) \quad (2.43)$$

By diagonalizing Eq. 2.43 we find out the eigenvalues for each \mathbf{q} , which are the solutions to ω .

Given the phonon dispersion $\omega(\mathbf{q})$, we can compute the phonon density of states $n_{vib}(\omega)$, the partition function $Z_{vib}(T)$ and the related quantities such as the internal energy $U_{vib}(T)$, the Helmholtz free energy $F_{vib}(T)$ and vibrational entropy $S_{vib}(T)$.

$$n_{vib}(\omega) = \int \frac{d\mathbf{q}}{(2\pi)^3} \delta(\omega - \omega(\mathbf{q})) \quad (2.44)$$

$$Z = \prod_{\mathbf{q}\nu} \frac{e^{-\hbar\omega_\nu(\mathbf{q})/2k_B T}}{1 - e^{-\hbar\omega_\nu(\mathbf{q})/k_B T}} \quad (2.45)$$

$$U_{vib} = -\frac{\partial}{\partial\beta} \ln Z_{vib} = \sum_{\mathbf{q}\nu} \left(\frac{1}{2} + \frac{1}{e^{\hbar\omega_\nu} - 1} \right) \quad (2.46)$$

$$F_{vib} = -k_B T \ln Z_{vib} = \frac{1}{2} \sum_{\mathbf{q}\nu} \hbar\omega(\mathbf{q}\nu) + k_B T \sum_{\mathbf{q}\nu} \ln(1 - e^{-\hbar\omega(\mathbf{q}\nu)/k_B T}) \quad (2.47)$$

$$S_{vib} = -\frac{\partial F_{vib}}{\partial T} \quad (2.48)$$

$$C_V = \frac{1}{k_B T^2} \frac{\partial^2 \ln Z}{\partial\beta^2} = \frac{1}{k_B T^2} (\langle U_{vib}^2 \rangle - \langle U_{vib} \rangle^2) \quad (2.49)$$

For quantum-mechanical harmonic oscillator, energies for each level are given by

$$E_n = \frac{1}{2} \hbar\omega + n\hbar\omega \quad (2.50)$$

where n is the excitation number of the oscillator. The density distribution of any ordinary classical gas P_n is given in equilibrium at temperature T by Maxwell-Boltzmann distribution such that $P_n \propto e^{-E_n/k_B T}$, where E_n is the energy at state n .

Average energy of an oscillator at temperature T is given by²

$$\begin{aligned}\varepsilon(\omega, T) &= \sum_{n=0}^{\infty} E_n P_n = (1 - e^{-\hbar\omega/k_B T}) \hbar\omega \sum_{n=0}^{\infty} (n + \frac{1}{2}) (e^{-\hbar\omega/k_B T})^n \\ &= \frac{1}{2} \hbar\omega + \hbar\omega \frac{1}{\exp(\hbar\omega/k_B T) - 1} = \frac{1}{2} \hbar\omega + \langle n \rangle \hbar\omega\end{aligned}\quad (2.51)$$

Next, two models will be discussed for estimating the phonon distribution to the heat capacity of a solid.

The *Einstein model* assumes that all atoms are in a similar potential and thus all harmonic oscillators have the same resonance frequency ω_E . It is often more convenient to examine the temperature derivative of energy, i.e., heat capacity, and not the absolute energy itself since it varies too slightly compared to the equilibrium value. The Einstein model of heat capacity under constant volume is given by

$$\begin{aligned}C_V &= \left(\frac{\partial U}{\partial T}\right)_V = 3N \frac{\partial}{\partial T} \langle \varepsilon(\omega, T) \rangle \\ &= 3N \frac{\partial}{\partial T} \sum_{n=0}^{\infty} P_n \varepsilon_n(\omega_E) \\ &= 3N \frac{\partial}{\partial T} \left(\frac{1}{2} \hbar\omega_E + \frac{\hbar\omega_E}{e^{\hbar\omega_E/k_B T} - 1} \right) \\ &= 3N k_B \left(\frac{\hbar\omega_E}{k_B T} \right)^2 \frac{e^{\hbar\omega_E/k_B T}}{(e^{\hbar\omega_E/k_B T} - 1)^2}\end{aligned}\quad (2.52)$$

If we define Einstein temperature $\Theta_E = \frac{\hbar\omega_E}{k_B}$, the heat capacity can be written in terms of Θ_E as

$$C_V = 3N k_B \left(\frac{\Theta_E}{T} \right)^2 \frac{e^{\Theta_E/T}}{(e^{\Theta_E/T} - 1)^2}\quad (2.53)$$

At large T, $C_V \approx 3N k_B$. While for low temperature, $C_V \propto \left(\frac{\Theta_E}{T}\right)^2 e^{-\Theta_E/T}$. Einstein model succeeds in describing heat capacity in intermediate and high temperature range. However, due to the inaccurate assumption that all oscillators share the same frequency, Einstein model fails to predict heat capacity under low temperature.

Debye model takes into account the variation of frequency over different lattice sites, or

²Here $P_n = \frac{e^{-n\hbar\omega/k_B T}}{1 - e^{-\hbar\omega/k_B T}}$ follows Bose-Einstein distribution.

different k-values. Debye assumes that ω and k follow a linear dispersion relation,

$$\omega = v_s k \quad (2.54)$$

where v_s is the sound velocity. Obviously, sound velocity differs from the longitudinal and the two transverse modes in diatomic case but it is not considered in Debye model. Similar to the way of how total number of free electrons is counted in Eq. ??, the number of vibration modes (or phonons) is given by

$$N = \frac{k^3}{6\pi^2} V \quad (2.55)$$

Then we get the density of states

$$g(k) = \frac{V k^2}{2\pi^2} \quad (2.56)$$

We can convert the density of states per polarization from the conservation law $g(\omega)d\omega \equiv g(k)dk$ and arrive at

$$g(\omega) = g(k) \frac{dk}{d\omega} = \frac{V \omega^2}{2\pi^2} \frac{1}{v_s^3} \quad (2.57)$$

The number of states is given by integrating the density of states up to some cutoff frequency ω_D

$$N = \int_0^{\omega_D} d\omega g(\omega) = \int_0^{\omega_D} d\omega \frac{V \omega^2}{2\pi^2} \frac{1}{v_s^3} = \frac{V \omega_D^3}{6\pi^2} \frac{1}{v_s^3} \quad (2.58)$$

where we can derive the cutoff frequency also called Debye frequency $\omega_D = 6\pi^2 v_s^3 \frac{N}{V}$ and Debye temperature $\Theta_D = \frac{\hbar \omega_D}{k_B} = \frac{\hbar v_s}{k_B} (6\pi^2 \frac{N}{V})^{1/3}$.

The internal energy is obtained by integrating over density of states times statistically weighted energy per state up to Debye frequency,

$$\begin{aligned} U &= \int_0^{\omega_D} d\omega g(\omega) \langle \varepsilon(\omega) \rangle \hbar \omega \\ &= \int_0^{\omega_D} d\omega \left(\frac{V \omega^2}{2\pi^2 v_s^3} \right) \left(\frac{\hbar \omega}{e^{\hbar \omega} - 1} \right) \\ &= 9N k_B T \left(\frac{T}{\Theta_D} \right)^3 \int_0^{x_D} dx \frac{x^3}{e^x - 1} \end{aligned} \quad (2.59)$$

where $x = \hbar \omega / k_B T$ and $x_D = \Theta_D / T$. The heat capacity is then given by time derivative

of internal energy

$$\begin{aligned}
C_V &= \left(\frac{\partial U}{\partial T}\right)_V = 3 \frac{V \hbar}{2\pi^2 v_s^3} \int_0^{\omega_D} d\omega \frac{\partial}{\partial T} \frac{\omega^3}{e^{\hbar\omega/k_B T} - 1} \\
&= \frac{3V \hbar^2}{2\pi^2 v_s^3 k_B T^2} \int_0^{\omega_D} d\omega \frac{\omega^4 \hbar\omega/k_B T}{(e^{\hbar\omega/k_B T})^2} \\
&= \frac{3V}{2\pi^2 v_s^3 k_B T^2} \int_0^{\omega_D} d\omega \frac{\omega^4 e^{\hbar\omega/k_B T}}{(e^{\hbar\omega/k_B T})^2} \\
&= 9Nk_B \left(\frac{T}{\Theta_D}\right)^3 \int_0^{x_D} dx \frac{x^4 e^x}{(e^x - 1)^2}
\end{aligned} \tag{2.60}$$

For large T,

$$C_V \approx 9Nk_B \left(\frac{T}{\Theta_D}\right)^3 \int_0^{x_D} dx x^2 = 3Nk_B \tag{2.61}$$

For T \rightarrow 0,

$$\begin{aligned}
U &\approx \frac{3\pi^4 Nk_B T^4}{5\Theta_D^3} \\
C_V &\approx \frac{12\pi^4 Nk_B}{5} \left(\frac{T}{\Theta_D}\right)^3 \propto T^3
\end{aligned} \tag{2.62}$$

Methods for Phonon Calculation

There are essentially two methods for calculating harmonic force constants within DFT, density functional perturbation theory (DFPT) and finite difference approaches.

Within DFT, the force induced by displacing an atom by a displacement, $u_{\alpha\mu}$, can be calculated using the Hellman-Feynman theorem

$$\frac{\partial E}{\partial u_{\alpha\mu}} = \int d^3r n(\mathbf{r}) \frac{\partial V_{ext}(\mathbf{r})}{\partial u_{\alpha\mu}} \tag{2.63}$$

For force constants, we need the second derivative, that is,

$$\frac{\partial^2 E}{\partial u_{\alpha\mu} \partial u_{\beta\nu}} = \int d^3r n(\mathbf{r}) \frac{\partial^2 V_{ext}(\mathbf{r})}{\partial u_{\alpha\mu} \partial u_{\beta\nu}} + \int d^3r \frac{\partial n(\mathbf{r})}{\partial u_{\beta\nu}} \frac{\partial V_{ext}(\mathbf{r})}{\partial u_{\alpha\mu}} \tag{2.64}$$

The derivative of the density in Eq. 2.64 can be obtained from perturbation theory, since according to Eq. 2.16, the calculation of perturbed density is equivalent to that of perturbed wave function. The regular way to get first-order responses is to calculate the expectation value of the perturbed operator in terms of unperturbed states, e.g.,

$$\Delta\psi_i^{(1)} = \sum_{j \neq i} \frac{\langle \psi_j^{(0)} | \Delta H | \psi_j^{(0)} \rangle}{\varepsilon_i^{(0)} - \varepsilon_j^{(0)}} \psi_j \quad (2.65)$$

where the superscript (0) denotes non-perturbed state and (1) denotes the first order perturbed state. Perturbed wave function can be obtained in a self-consistent way similar to solving the Kohn-Sham equation. From these, Eq. 2.64 then yields the force constants.

In the finite difference (also known as "frozen phonon") approach, the force constants needed to set up the dynamical matrix is accomplished by displacing one atom after another

$$\frac{\partial^2 E}{\partial u_{\alpha\mu} \partial u_{\beta\nu}} \approx \frac{F_{\alpha\mu}}{u_{\beta\nu}}, \quad (2.66)$$

with $\mu \rightarrow 0$.

The finite difference method involves calculations of supercells to get accurate force constants, so to sample phonon with $\mathbf{q} \neq 0$. To construct the supercell, we first define a $3 \times 3 \times 3$ supercell matrix \mathbf{P} that relates supercell vector ($\mathbf{a}_s \ \mathbf{b}_s \ \mathbf{c}_s$) and primitive cell matrix ($\mathbf{a}_p \ \mathbf{b}_p \ \mathbf{c}_p$) by

$$(\mathbf{a}_s \ \mathbf{b}_s \ \mathbf{c}_s) = (\mathbf{a}_p \ \mathbf{b}_p \ \mathbf{c}_p) \mathbf{P} \quad (2.67)$$

The total number of atom in the supercell is calculated by

$$N_s = N_p \cdot |\det(\mathbf{P})| \quad (2.68)$$

where $\det(\mathbf{P})$ is the determinant of \mathbf{P} . As one period of wave has to be bound by any of the lattice point in a unit cell, the number of commensurate q-points that are accurately calculated in phonon dispersion is equivalent to $|\det(\mathbf{P})|$.

These two methods lead to comparable accuracies. DFPT does not require the construction of supercells, but is typically only implemented in semilocal DFT calculations, but not for hybrid functional, nor beyond DFT methods. Conversely, the finite difference approach is compatible with all *ab initio* methods that allow a force calculation, but requires calculations in large supercells.

The Quasiharmonic Approximation

In experiments, what we are concerned about is usually not the Helmholtz energy and the related thermal quantities given above, but the Gibbs free energy since we are not controlling volume and temperature, but pressure and temperature. The Gibbs free energy is obtained via Legendre transform from

$$G(T, p) = \min_V [F(T, V) + pV] \quad (2.69)$$

where it means to find the minimum of the function of V in the bracket. One way to calculate $G(T, p)$ is the quasiharmonic approximation (QHA), where $G(T, p)$ is obtained by repeatedly computing $F(T, V)$ at various volumes. This is equivalent to redefining the phonon to be volume dependent so that the "extension" of the Helmholtz free energy given in Eq. 2.47 can be written as

$$F^{QHA}(T, V) = \frac{1}{2} \sum_{\mathbf{q}\nu} \hbar\omega(\mathbf{q}\nu V) + k_B T \sum_{\mathbf{q}\nu} \ln(1 - e^{-\hbar\omega(\mathbf{q}\nu V)/k_B T}) \quad (2.70)$$

Based on the two equations of states

$$P(V, T) = - \left(\frac{\partial F}{\partial V} \right)_T \quad (2.71)$$

$$V(P, T) = \left(\frac{\partial G}{\partial P} \right)_T \quad (2.72)$$

where F and G are Helmholtz and Gibbs free energy, we obtain the volume thermal expansion coefficient and the isothermal bulk modulus that are given by

$$\alpha = \frac{1}{V} \left(\frac{\partial V}{\partial T} \right)_P = \frac{1}{V} \left(\frac{\partial^2 G}{\partial T \partial P} \right)_{PT} \quad (2.73)$$

$$B_T = -V \left(\frac{\partial P}{\partial V} \right)_T = V \left(\frac{\partial^2 F}{\partial V^2} \right)_T \quad (2.74)$$

It can be seen that these two quantities are related by

$$\alpha B_T = - \left(\frac{\partial^2 F}{\partial V \partial T} \right)_{TV} \quad (2.75)$$

If we expand a lattice to V_1 against its equilibrium volume V_0 , Eq. 2.74 yields

$$\frac{V_1^{QHA}(T) - V_0}{V_0} = -\frac{1}{B_0} \left(\frac{\partial F}{\partial V} \right)_T \quad (2.76)$$

Now we introduce mode Grüneisen parameter defined as

$$\gamma_{q\nu} \equiv - \left(\frac{V}{\omega_{q\nu}} \frac{d\omega_{q\nu}}{dV} \right)_{V_0} \quad (2.77)$$

Combing Eq. 2.70, Eq. 2.76 and Eq. 2.83, we arrive at

$$\frac{V_1^{QHA}(T) - V_0}{V_0} = \frac{1}{B_0 V_0} \sum_{q\nu} \hbar \omega_{q\nu} \gamma_{q\nu} \left(n_{q\nu} + \frac{1}{2} \right) \quad (2.78)$$

Taking this equation back to the relation (Eq. 2.75) given above and replacing the temperature derivative with the heat capacity C_v , we obtain the Grüneisen equation of states

$$\alpha = \frac{\gamma C_v}{3B} \quad (2.79)$$

Eq. 2.79 implies that the thermal expansion shows similar trends as the heat capacity such that $\alpha \sim T^3$ when $T \rightarrow 0$ and that it converges to $k_B \gamma / B_0 V_0$ (γ is the overall Grüneisen parameter) when $T > \Theta_D$.

Perturbation Theory in Lattice Dynamics

In the harmonic approximation, we determine the stationary eigenstates of the harmonic Hamiltonian. It is however not sufficient in describing transport properties. One typical example is the thermal expansion which the harmonic approximation would never capture since it is assumed that the equilibrium size of the crystal is temperature independent. Also the phonon mean free path is infinite in the harmonic approximation (since no phonon interactions are included), and so is the thermal conductivity.

One approach to account for anharmonicity at least approximately is to obtain the third or higher order terms via Taylor expansion of the potential. The anharmonic term then

takes the form

$$U^{anh} = \sum_{n=3}^{\infty} \frac{1}{n!} \sum_{\mathbf{R}_1 \dots \mathbf{R}_n} \frac{\partial^n U}{\partial u_{\mu 1}(\mathbf{R}_1) \dots \partial u_{\mu n}(\mathbf{R}_n)} u_{\mu 1}(\mathbf{R}_1) \dots u_{\mu n}(\mathbf{R}_n) \quad (2.80)$$

In the following, only cubic anharmonicity is taken into account, though sometimes it is necessary to go further to higher orders (see [1], p. 489). Here, we first discuss the volume dependence. In the finite difference method, one now has to do three phonon calculations, one at equilibrium volume V_0 and the other two with slightly larger and smaller volume than V_0 . The volume derivative of dynamical matrix in Eq. 2.82 is computed as

$$\frac{\partial D}{\partial V} = \frac{D(V + \Delta V) - D(V - \Delta V)}{2(\Delta V)} \quad (2.81)$$

As defined in Eq. 2.83, the mode Grüneisen parameter can now be related to the dynamical matrix by

$$\gamma_{\mathbf{q}\nu} = -\frac{V}{2[\omega^2(\mathbf{q}\nu)]} \langle \varepsilon(\mathbf{q}\nu) | \frac{\partial D(\mathbf{q})}{\partial V} | \varepsilon(\mathbf{q}\nu) \rangle \quad (2.82)$$

where $\varepsilon(\mathbf{q}\nu)$ is the eigenvalue got from diagonalization of the dynamical matrix.

The overall Grüneisen parameter is calculated by summing over mode Grüneisen parameter weighted by its contribution to the specific heat

Grüneisen parameter defined in Eq.

$$\gamma = \frac{\sum_{\mathbf{q}\nu} \gamma_{\mathbf{q}\nu} C_{V\nu}(\mathbf{q})}{\sum_{\mathbf{q}\nu} C_{V\nu}(\mathbf{q})} \quad (2.83)$$

Alternatively, one can calculate Grüneisen parameter from third-order force constants. The volume derivative term $\frac{\partial D(\mathbf{q})}{\partial V}$ is related to the third order force constants by (subscripts are the same as used in Section 2.1.3)

$$\begin{aligned} \delta D_{\alpha\beta}(\kappa\kappa'; \mathbf{q}) &= \frac{1}{\sqrt{M_\kappa M_{\kappa'}}} \sum_{l'} \delta \Phi_{\mu\nu}(l\kappa; l'\kappa') \exp[i\mathbf{q} \cdot (r(l') - r(l))] \\ \delta \Phi_{\mu\nu}(l\kappa; l'\kappa') &= \sum_{l'', \kappa'', \lambda} \Phi_{\mu\nu\lambda}(l\kappa; l'\kappa'; l''\kappa'') r_\lambda(l''\kappa'') \end{aligned} \quad (2.84)$$

where $\delta D_{\alpha\beta}$ is the change of dynamical matrix due to the change of volume.

Lattice Thermal Conductivity

Transport phenomena are all typically characterized by the quantity being transported and the transport speed. Furthermore, the transport of (quasi-)particles is typically accompanied by collisions. For instance, electrons are often described in the Drude model, which assumes that the velocity between two collisions remains unchanged. The time interval, or relaxation time, of charge transport between two consecutive collisions is simply given by

$$\tau = \frac{m^* \sigma}{n e^2} \quad (2.85)$$

where m^* is the electronic mass, σ is the conductivity, n the charge density and e the charge. For thermal transport, these become heat and the velocity of the particles transporting the heat. Consider a simple heat transport model along the x -axis as shown in Fig. 2.1 with heat source and sink at both ends. Heat that arrives at T_2 experienced its last collision at either T_1 or T_3 . The heat balance at the mid position x_{T_2} with temperature T_2 is given by the difference between heat flow into and out of the domain.

$$j = \frac{1}{2} n v [\varepsilon(T[x_{T_2} - v\tau]) - \varepsilon(T[x_{T_2} + v\tau])] \quad (2.86)$$

When the mean free path $l = v\tau$ is very small, the energy density functions in Eq. 2.86 can be linearly expanded as

$$\begin{aligned} j &= \frac{1}{2} n v \varepsilon \left(T \left[x_{T_2} - v\tau \left(-\frac{dT}{dx} \right) \right] \right) - \frac{1}{2} n v \varepsilon \left(T \left[x_{T_2} + v\tau \left(-\frac{dT}{dx} \right) \right] \right) \\ &= \frac{1}{2} n v \left[\varepsilon(T(x_{T_2})) - v\tau \left(-\frac{dT}{dx} \right) \left(\frac{d\varepsilon}{dT} \right) \right] - \frac{1}{2} n v \left[\varepsilon(T(x_{T_2})) + v\tau \left(-\frac{dT}{dx} \right) \left(\frac{d\varepsilon}{dT} \right) \right] \\ &= n v^2 \tau \frac{d\varepsilon}{dT} \left(-\frac{dT}{dx} \right) \end{aligned} \quad (2.87)$$

For the three-dimensional case the velocity in Eq. 2.87 is replaced by the component along one coordinate, which is $\langle v_x^2 \rangle = \langle v_y^2 \rangle = \langle v_z^2 \rangle$. Note that here the mean square velocity is approximated as temperature-independent. With the energy derivative being replaced by heat capacity as well, the heat flux is then written as

$$\mathbf{j} = \frac{1}{3} v^2 \tau C_V (-\nabla T) \quad (2.88)$$

Comparing this results with Fourier's law, the thermal conductivity per phonon mode κ_λ is given by

$$\kappa_\lambda = \frac{1}{3} C_{V\lambda} v^2 \tau_\lambda = \frac{1}{3} C_{V\lambda} v l_\lambda \quad (2.89)$$

The relaxation time (or phonon lifetime) in the above equation remains to be the most tricky part of solving for the lattice thermal conductivity.

The phonon lifetime τ (or phonon scattering rate τ^{-1}) related to the phonon-phonon scattering has been approximated in different models.

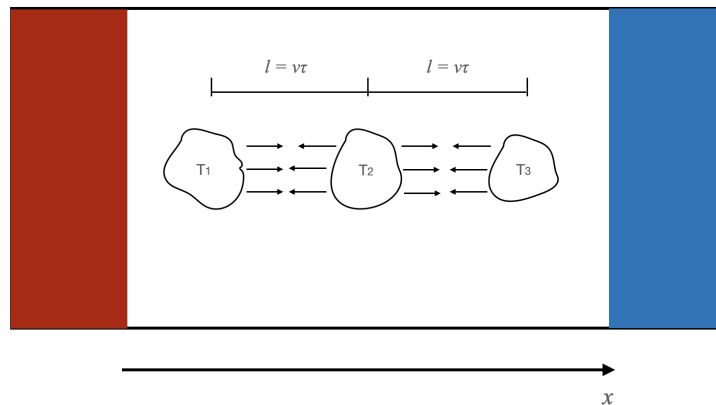


Figure 2.1.: Schematic diagram of heat transport

1. Slack Model and Debye-Callaway Model

The lattice thermal conductivity is quantitatively described by the early work of Debye and Peierls. It tells that (1) in the low temperature region thermal conductivity increases with heat capacity and the latter is T^3 dependent, (2) when temperature gets higher ($\sim 0.1\Theta_D$), the Umklapp processes start to dominant and the thermal conductivity starts to drop and (3) at the peak of the thermal conductivity, its value is sensitive to the crystal impurities.

Following these findings, the *Slack model* suggests that the lattice thermal conductivity can be approximated by

$$\kappa_L = \frac{k_B}{2\pi^2 v} \left(\frac{k_B T}{\hbar} \right)^3 \int_0^{\Theta_D/T} \frac{x^4 e^x}{\tau_c^{-1} (e^x - 1)^2} dx \quad (2.90)$$

where $x = \hbar\omega/k_B T$, Θ_D is the Debye temperature, v is the velocity of sound, and τ is the total phonon scattering time.

Similar to the Slack model, the *Debye-Callaway model* also only considers the contribution of acoustic modes to the thermal conductivity, which is sufficient for very simple materials. For the phonon scattering process, however, both normal and Umklapp processes will affect the heat transport. Umklapp process gives rise to the thermal resistance due to the reversed phonon flux, while the normal mode may redistribute momentum and energy among the phonons. Thus the total the scattering rate of phonon is expressed by $\tau_C^{-1} = \tau_N^{-1} + \tau_U^{-1}$. The partial thermal conductivities κ_i ($i = \text{LA, TA or TA' modes}$) are given by [25]

$$\kappa_i = \frac{1}{3} C_i T^3 \int_0^{\Theta_i/T} \frac{\tau_c^i(x) x^4 e^x}{(e^x - 1)^2} dx + \frac{\left[\int_0^{\Theta_i/T} \frac{\tau_c^i(x) x^4 e^x}{\tau_N^i (e^x - 1)^2} dx \right]^2}{\int_0^{\Theta_i/T} \frac{\tau_c^i(x) x^4 e^x}{\tau_N^i \tau_U^i (e^x - 1)^2} dx} \quad (2.91)$$

where Θ_i is the longitudinal (transverse) Debye temperature given by $\Theta = \hbar\omega_D/k_B$, $x = \hbar\omega/k_B T$ and C_i depends on the acoustic group velocity such that $C_i = k_B^4/2\pi^2 \hbar^3 v_i$. We notice that the Debye-Callaway model is a function of the Debye temperature (Θ), phonon velocity (v) and the phonon lifetime (τ) for both normal and Umklapp processes. Below are the fitting models for the scattering rate of different modes given in [15]. For Umklapp process, these are

$$[\tau_U^L(x)]^{-1} = B_U^L \left(\frac{k_B}{\hbar} \right)^2 x^2 T^3 e^{-\Theta_L/3T} \quad (2.92)$$

and

$$[\tau_U^T(x)]^{-1} = B_U^T \left(\frac{k_B}{\hbar} \right)^2 x^2 T^3 e^{-\Theta_T/3T}, \quad (2.93)$$

with

$$B_U^i \approx \frac{\hbar \gamma_i^2}{M v_i^2 \Theta_i} \quad (2.94)$$

For normal process, these are

$$[\tau_N^L(x)]^{-1} = B_N^L \left(\frac{k_B}{\hbar} \right)^2 x^2 T^5 \quad (2.95)$$

with

$$B_N^L \approx \frac{k_B^3 \gamma_L^2 V}{M \hbar^2 v_L^5} \quad (2.96)$$

and

$$[\tau_N^L(x)]^{-1} = B_N^L \left(\frac{k_B}{\hbar} \right)^2 x T^5 \quad (2.97)$$

with

$$B_N^T \approx \frac{k_B^4 \gamma_L^2 V}{M \hbar^3 v_T^5} \quad (2.98)$$

The three involved dispersion related parameters Debye temperature (Θ), phonon velocity (v) and Grüneisen parameter can be obtained either from experiments or *ab initio* calculations. There are also phonon-boundary scattering and phonon-isotope needed to be included but they are simply not shown here.

2. Three-Phonon-Interaction

In our study, the single-mode relaxation time (SMRT) method including the three-phonon interaction is used to calculate the thermal conductivity [22], where the phonon life time is obtained from the anharmonic term of the potential. Compared to the Slack model and the Debye-Callaway model, the phonon-phonon scattering is now treated as perturbation and the actual third-order anharmonicity is captured.

To do this, the Hamiltonian is expanded up to third order. Combining Eqs. 2.33 and 2.80, we get

$$\begin{aligned} U &= U^{eq} + U^{harm} + U^{anharm} \\ &= U^{eq} + \frac{1}{2} \sum_{\alpha, \beta}^N \sum_{\mu, \nu=x, y, z} u_{\alpha, \mu}(\mathbf{R}) \Phi_{\mu, \nu}^{\alpha, \beta}(\mathbf{R} - \mathbf{R}') u_{\beta, \nu}(\mathbf{R}') \\ &\quad + \frac{1}{6} \sum_{\alpha, \beta, \gamma}^N \sum_{\mu, \nu, \xi=x, y, z} \Phi_{\mu, \nu, \xi}^{\alpha, \beta, \gamma}(\mathbf{R}, \mathbf{R}', \mathbf{R}'') u_{\alpha, \mu}(\mathbf{R}) u_{\beta, \nu}(\mathbf{R}'') u_{\gamma, \xi}(\mathbf{R}''') \\ &= U^{eq} + \sum_{\mathbf{q}\nu} \left(\frac{1}{2} + n_{\mathbf{q}\nu} \right) \hbar \omega_{\nu}(\mathbf{q}) + \sum_{\lambda \lambda' \lambda''} \Phi_{\lambda \lambda' \lambda''} (\hat{a}_{\lambda} + \hat{a}_{-\lambda}^{\dagger}) (\hat{a}_{\lambda'} + \hat{a}_{-\lambda'}^{\dagger}) (\hat{a}_{\lambda''} + \hat{a}_{-\lambda''}^{\dagger}) \end{aligned} \quad (2.99)$$

The last equation in Eq. 2.99 is the quantum mechanical expression derived from the

second equation, where λ denotes the phonon mode (\mathbf{q}, ν) and $\Phi_{\lambda\lambda'\lambda''}$ represents the interaction between three phonons. \hat{a} and \hat{a}^\dagger are the creation and annihilation operators of the harmonic Hamiltonian respectively. $\Phi_{\lambda\lambda'\lambda''}$ is derived from the third order force constants and explicitly given by

$$\begin{aligned} \Phi_{\lambda\lambda'\lambda''} = & \frac{1}{\sqrt{N}} \frac{1}{6} \sum_{\kappa\kappa'\kappa''} \sum_{\alpha\beta\gamma} W_\alpha(\kappa, \lambda) W_\beta(\kappa', \lambda') W_\gamma(\kappa'', \lambda'') \sqrt{\frac{\hbar}{2m_\kappa\omega_\lambda}} \sqrt{\frac{\hbar}{2m_{\kappa'}\omega_{\lambda'}}} \sqrt{\frac{\hbar}{2m_{\kappa''}\omega_{\lambda''}}} \\ & \times \sum_{l'l''} \Phi_{\alpha\beta\gamma}(0\kappa, l'\kappa', l''\kappa'') e^{i\mathbf{q}'[l'\kappa' - \mathbf{0}\kappa]} e^{i\mathbf{q}''[l''\kappa'' - \mathbf{0}\kappa]} e^{i(\mathbf{q} + \mathbf{q}' + \mathbf{q}'')\mathbf{r}(0\kappa')} \Delta(\mathbf{q} + \mathbf{q}' + \mathbf{q}'') \end{aligned} \quad (2.100)$$

Here $\Delta(\mathbf{q} + \mathbf{q}' + \mathbf{q}'') = 1$ when $\mathbf{q} + \mathbf{q}' + \mathbf{q}''$ equals to reciprocal lattice vector and is zero otherwise. $W(\kappa, \lambda)$ are the polarization vectors corresponding to the eigenvectors of the Dynamical matrix. $\Phi_{\lambda\lambda'\lambda''}$ can then be used to calculate the imaginary part of the self-energy, which is given by

$$\begin{aligned} \Gamma_\lambda(\omega) = & \frac{18\pi}{\hbar^2} |\Phi_{-\lambda\lambda'\lambda''}|^2 \{ (n_{\lambda'} + n_{\lambda''} + 1) \delta(\omega - \omega_{\lambda'} - \omega_{\lambda''}) \\ & + (n_{\lambda'} - n_{\lambda''}) [\delta(\omega + \omega_{\lambda'} - \omega_{\lambda''}) - \delta(\omega - \omega_{\lambda'} + \omega_{\lambda''})] \} \end{aligned}$$

where n_λ is the phonon occupation number at the equilibrium. Phonon lifetime is related to the imaginary part of the self-energy via

$$\tau_\lambda = \frac{1}{2\Gamma_\lambda(\omega_\lambda)}. \quad (2.101)$$

Eq. 2.101 can be inserted into Eq 2.89 for calculating the mode-specific thermal conductivity

$$\kappa_{ph}^{\mu\mu}(\omega) = \frac{1}{\Omega N_q} \sum_\lambda C_\lambda v_\lambda^\mu v_\lambda^\mu \tau_\lambda \delta(\omega - \omega_\lambda) \quad (2.102)$$

Integrating over all mode frequencies gives the bulk thermal conductivity.

3. Green-Kubo Method

Using Green-Kubo method one is able to calculate thermal the conductivity in a non-perturbative manner via *ab initio* molecular dynamics. The Kubo relations [12] tell us that transport coefficients can be obtained by integrating the correlation function of the associated flux over time.

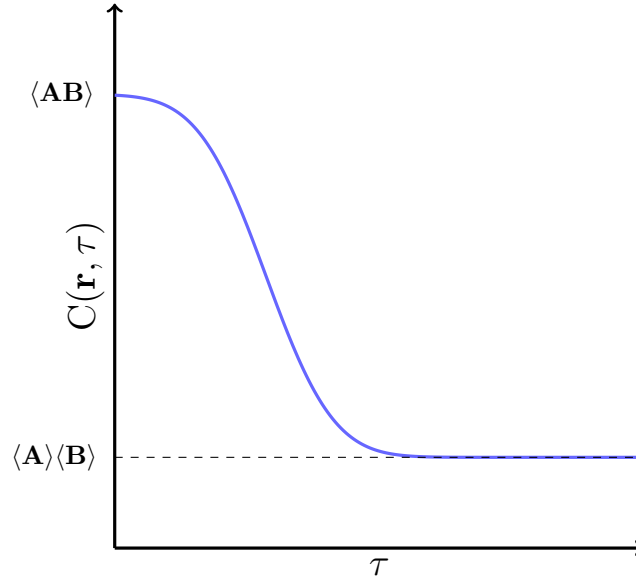


Figure 2.2.: Characteristic features of an auto-correlation function

Figure 2.2 shows the decay of an auto-correlation function $C(\mathbf{r}, \tau)$ over time, where A and B are two time-dependent signals. When the two signals are same quantity, the auto-correlation function will decay simply to zero. The thermal conductivity κ is related to the heat flux auto-correlation by [5]

$$\kappa_{\alpha} = \frac{1}{k_B T^2 V} \int_0^{\infty} \langle J_{\alpha}(t) J_{\alpha}(0) \rangle dt \quad (2.103)$$

where $\langle J_{\alpha}(t) J_{\alpha}(0) \rangle$ is the ensemble averaged heat flux auto-correlation function. Considering the continuity law

$$\frac{1}{V} \frac{\partial E(t)}{\partial t} + \nabla \mathbf{j} = 0 \quad (2.104)$$

and integrating both sides over position \mathbf{r} , we get the expression of the heat flux $\mathbf{J}(t)$ that is

$$\mathbf{J}(t) = \frac{1}{V} \frac{d}{dt} \sum_I \mathbf{R}_I E_I = \frac{1}{V} \sum_I \dot{\mathbf{R}}_I E_I + \frac{1}{V} \sum_I \mathbf{R}_I \dot{E}_I \quad (2.105)$$

where \mathbf{R}_I and E_I are the position and the energy for atom I. The last two terms in Eq. 2.105 imply that the heat flux can be decomposed into a convective term describing the diffusion of atoms which is negligible for heat transported in typical solids, and the conduction term describing the energy transfer between neighboring atoms. The conduction term can be rewritten in the following way.

$$\begin{aligned} \frac{1}{V} \sum_I \mathbf{R}_I \dot{E}_I &= \frac{1}{V} \sum_{I \neq J} (\mathbf{F}_{IJ} \cdot \mathbf{v}_I) \mathbf{R}_{IJ} \\ &= \frac{1}{V} \sum_I \sigma_I \mathbf{v}_I \end{aligned} \quad (2.106)$$

From Eq. 2.106 we see that the heat flux now depends on the stress tensor and the velocity of each atom, while no longer on the energy density directly. The solution to it is exact provided that the stress term can be formulated in a unique and well-defined way, which has been implemented in the all-electron code FHI-aims [11]. Then one is able to calculate the heat flux $\mathbf{J}(t)$ from the position and the velocity of each *ab initio* MD trajectory. Thermal conductivity κ is further obtained by the integral of the heat flux auto-correlation.

Measure for Anharmonicity

When comparing between MD-computed thermal conductivity and the perturbative ones, it is important to note that discrepancies arise for strongly anharmonic materials. It is thus important to characterize materials by anharmonicity.

Recently, Knoop *et al.* [10] developed a quantitative measure of the *degree of anharmonicity* of materials by computing the root mean square error (RMSE) of the difference between anharmonic and harmonic forces divided by the standard deviation of force distribution:

$$\sigma^A(T) \equiv \frac{\sigma [F^A]_T}{\sigma [F]_T} = \sqrt{\frac{\sum_{I,\alpha} \langle (F_{I,\alpha}^A)^2 \rangle_T}{\sum_{I,\alpha} \langle (F_{I,\alpha})^2 \rangle_T}} \quad (2.107)$$

where both $\sigma [F]_T$ and $\sigma [F^A]_T$ can be obtained from the time average, which are

$$\sigma [F]_T = \sqrt{\frac{1}{3N_I} \frac{1}{N_t} \sum_{t=1}^{N_t} F_{I,\alpha}^2(t)} \quad (2.108)$$

$$\sigma [F^A]_T = \sqrt{\frac{1}{3N_I} \frac{1}{N_t} \sum_{t=1}^{N_t} (F_{I,\alpha} - F_{I,\alpha}^{(2)})^2(t)} \quad (2.109)$$

The degree of anharmonicity can also be roughly predicted in a *one shot* way. This is done by calculating the atomic configurations from the distribution function predicted in the harmonic approximation instead of time averaging the MD results. For a given normal-mode coordinate $q_s = A_s(T) \cos(\omega_s t + \psi_s)$, the average kinetic energy in the thermal equilibrium is given by

$$\begin{aligned} \left\langle \frac{1}{2} \dot{q}_s \right\rangle &= \left\langle \frac{1}{2} \omega_s^2 A_s^2 \sin^2(\omega_s t + \psi_s) \right\rangle \\ &= \frac{1}{4} \omega_s^2 \langle A_s^2 \rangle = \frac{1}{2} k_B T \end{aligned} \quad (2.110)$$

The expectation value of the amplitude is $\langle A_s \rangle = \sqrt{2k_B T} / \omega_s$. If this holds for all s , then the energy of each mode is exactly $k_B T$ according to the equipartition theorem. Displacements can therefore be calculated from a random distribution given by

$$\Delta \mathbf{R}_I^\alpha = \frac{1}{\sqrt{M_I}} \sum_s \zeta_s \langle A_s \rangle \mathbf{e}_{sI}^\alpha \quad (2.111)$$

where \mathbf{e}_{sI}^α is the harmonic eigenvectors, and ζ_s is a normally distributed random number. If a material is fully harmonic, Eq. 2.111 would represent the true thermodynamic ensemble average. In this case, both σ^A and σ_{os}^A will be zero.

In calculating one-shot forces we simply set ζ_s to be ± 1 other than from a random distribution to approximate the harmonic turning points. Plugging this into Eq. 2.107 we get the one-shot metric σ_{os}^A . Note that for strongly anharmonic materials the approximated σ_{os}^A is not accurate and one still has to use σ^A from aiMD.

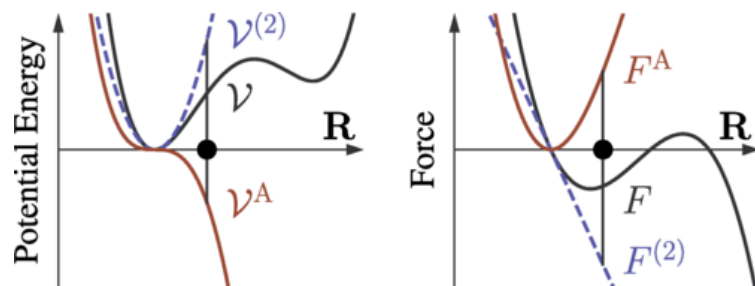


Figure 2.3.: These two plots show that for given potential energy and force, the anharmonic contribution of them is nothing but the difference between the total value and its harmonic approximation (figure from [10])

2.2. Statistical Theories

Once we are done with some calculations, we would like to gain some insights of structure-property relations by mapping the output results to the input data. Ideally, this should allow for skipping the costly computation step to determine the property and to only rely on simple, easy-to-compute properties. However, this mapping is usually not linear or explicit. Machine learning is a useful tool to find out the complex descriptor for us. Here in this section I will briefly introduce the approach, i.e, sure independence screening and sparsifying operator (SISSO), that is used to do the regression for the set thermal conductivity dataset obtained from *ab initio* calculations.

2.2.1. LASSO (Least Absolute Shrinkage and Selection Operator) and Compressed Sensing

Let's assume that there are some set of input data points $\{\mathbf{d}_{11}, \dots, \mathbf{d}_{1N}\}, \dots, \{\mathbf{d}_{N1}, \dots, \mathbf{d}_{MN}\}$ and a set of property data points $\{\mathbf{P}_1, \dots, \mathbf{P}_N\}$, \mathbf{d}_{MN} and $\mathbf{P}_N \in \mathbb{R}$. Our goal is to find a set of constants \mathbf{c} that can be the solution to the equation $\mathbf{P} = \mathbf{D} \cdot \mathbf{c}$, where \mathbf{P} is the property vector and \mathbf{D} is the input (or descriptor) matrix. *Least squares* is the most simple approach to solve this by minimizing the square of errors between the two sides, i.e.,

$$\arg \min_{\mathbf{c} \in \mathbb{R}^M} \|\mathbf{P} - \mathbf{D}\mathbf{c}\|_2^2 \quad (2.112)$$

where $\|\cdot\|_2^2$ denotes l_2 the square errors. The explicit solution to Eq. 2.112 is given by $\mathbf{c} = (\mathbf{D}^T \mathbf{D})^{-1} \mathbf{D}^T \mathbf{P}$. A problem of this approach is that for under-determined linear system, it would cause over-fitting if two entries of \mathbf{D} , inversely correlated to each other, are distributed with very large weights. Also, such an approach will usually lead to high-dimensional models with $c_i \neq 0 \forall i$.

l_2 Regularization

A better solution can be achieved by introducing an additional norm term (or penalty term) to Eq. 2.112. The most commonly used l_2 regularization is given by

$$\arg \min_{\mathbf{c} \in \mathbb{R}^M} (\|\mathbf{P} - \mathbf{D}\mathbf{c}\|_2^2 + \lambda \|\mathbf{c}\|_2^2) \quad (2.113)$$

where the coefficient λ controls the strength of the penalty and usually tuned by cross validation. However, the l_2 norm still tends to make use of all features, thus it is unable to give a sparsified solution that we want.

l_0 and l_1 Regularization

The sparsest solution can be obtained by replacing the second term of Eq. 2.113 with $\lambda\|\mathbf{c}\|_0$, which is nothing but the number of non-zero entries in \mathbf{D} . Solving such a problem is an NP-hard problem that scales non-polynomial with problem size. To accelerate the procedure, the l_1 norm regularization can simplify the regression by making it convex again, while still providing a sparse solution. Similar to the above, l_1 regularization is defined as

$$\arg \min_{\mathbf{c} \in \mathbb{R}^M} (\|\mathbf{P} - \mathbf{D}\mathbf{c}\|_2^2 + \lambda\|\mathbf{c}\|_1) \quad (2.114)$$

where $\|\mathbf{c}\|_1 = \sum_k |\mathbf{c}_k|$. This method is also known as LASSO (Least Absolute Shrinkage and Selection Operator). There exists a smallest $\lambda_0 > 0$, such that all entries of the solution \mathbf{c} are zero. With decreasing λ , more and more components become non-zero.

$$N \geq C\Omega \ln(M) \quad (2.115)$$

Eq. 2.115 quantitatively shows that there exists a constant C such that whenever this relation holds, a stable and robust recovery of \mathbf{c} from \mathbf{D} and \mathbf{P} is possible [8]. Compressed sensing provides a way to justify the construction of matrix \mathbf{D} and then the low-dimensional descriptor model is to be found via LASSO.

2.2.2. Symbolic Regression

Linear models of features are often not enough to describe most of the physical properties of materials. That's why non-linear models, e.g., kernel ridge regression, are often used for these applications. In our case, we use another approach based on symbolic regression to introduce non-linearity. It aims at generating sufficiently large feature space upon a small number of training data and mapping the regression output analytically to the input data. Features will be generated in a "building blocks" manner and based on this they will be further grouped into different rungs, i.e., Φ_1, Φ_2, Φ_3 etc., with each rung building up

out of features ("blocks") of the previous rung:

$$\Phi_n \equiv \bigcup_{i=1}^n \hat{\mathbf{H}}[\phi_1, \phi_2], \forall \phi_1, \phi_2 \in \Phi_{i-1} \quad (2.116)$$

In practice, features up to a few rung are sufficient and the maximum rung is one of the hyperparameters of the regression needed to be optimized. These feature spaces will be then mapped into dimension spaces by scoring the Pearson correlation with property vector \mathbf{P} . Optimal descriptors among them will be determined by the ℓ_0 method. Details of these two steps will be introduced in the following section.

2.2.3. SISSO (sure independence screening and sparsifying operators)

The goal of SISSO is to find the best low-dimensional linear models of a property given a set of non-linear, analytical expressions stored in Φ . The first step of this method is to use sure-independence screening (SIS) to screen all features in Φ to find the n_{sis} features that are most correlated to \mathbf{P} , using the Pearson correlation

$$\rho_{\mathbf{d}_i \mathbf{P}} = \sum_n \frac{[(\mathbf{d}_{ni} - \bar{\mathbf{d}}_i)(\mathbf{p}_n - \bar{\mathbf{p}}_n)]}{\sigma_{\mathbf{d}_{ni}} \sigma_{\mathbf{p}_n}} \quad (2.117)$$

where \mathbf{d}_{ni} , $\bar{\mathbf{d}}_i$ and $\sigma_{\mathbf{d}_{ni}}$ are the value of the n^{th} feature \mathbf{d}_i , mean value of \mathbf{d}_i and its standard deviation. \mathbf{p}_n , $\bar{\mathbf{p}}_n$ and $\sigma_{\mathbf{p}_n}$ are the respective values for \mathbf{p}_n . Once done these n_{sis} are stored inside of a subspace \mathbf{S}_{1D} and the best n_{res} models are trivially found using ℓ_0 -regularized least-squares regression. We then calculate the residuals Δ_{1D} ,

$$\Delta_{1D}^r \equiv \mathbf{P} - \mathbf{d}_{1D} \mathbf{c}_{1D}, \quad (2.118)$$

where r is the index of the found model, for each of the found models and then use those to perform the SIS selection for \mathbf{S}_{2D} . In this step we will be using multiple residuals, so the projection score has now changed to be the maximum correlation between the various residuals

$$\rho_{\mathbf{d}_i \mathbf{P}} = \max \left(\rho_{\mathbf{d}_i \mathbf{P}}^1, \dots, \rho_{\mathbf{d}_i \mathbf{P}}^{n_{res}} \right), \quad (2.119)$$

where $\rho_{\mathbf{d}_i \mathbf{P}}^r$ is the projection score associated with the residual of the r^{th} model. From here, we find the best two-dimensional model using ℓ_0 -regularized least-squares regression on the union of \mathbf{S}_{1D} , and \mathbf{S}_{2D} . This cycle is then repeated until some user-defined maximum

dimension is reached. In this algorithm the maximum dimension, n_{sis} , and n_{res} are all hyper-parameters that are optimized during cross validation.

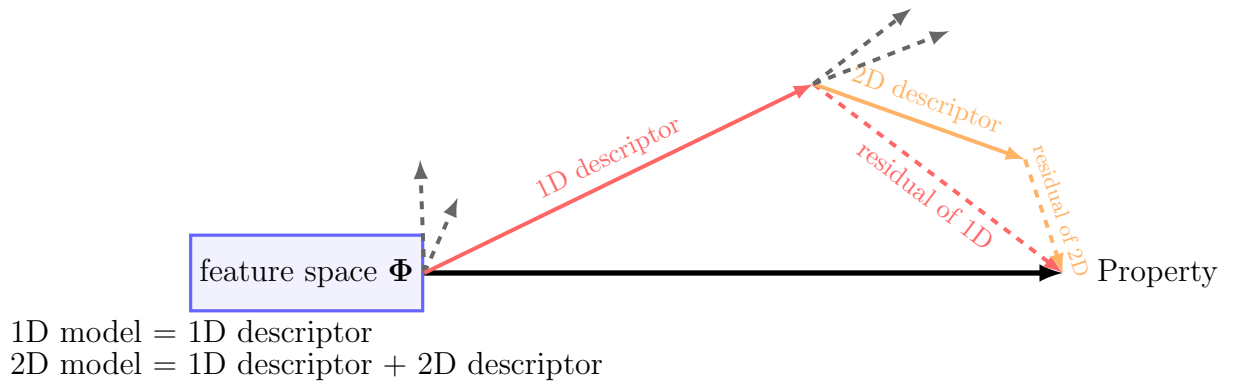


Figure 2.4.: Descriptors up to 2D are used to reduce the residual

2.2.4. Cross Validations of Models

Cross validations of model of the current dimension before moving on to the next dimension to ensure that the loss function - root mean square errors (RMSE) - is small enough so that the model is generalizable to be used in predicting property of a new material. Cross validation is performed by randomly leaving out a certain number of data from training data set. These left-out data form another test set to check the variance between real and fitting values of samples in the test set caused by over-fitting. Fig. 2.5 below shows the tradeoff between bias and variance over different complexities of model. It shows that the error is not always decreasing with increasing model complexity but that there exists a minimum.

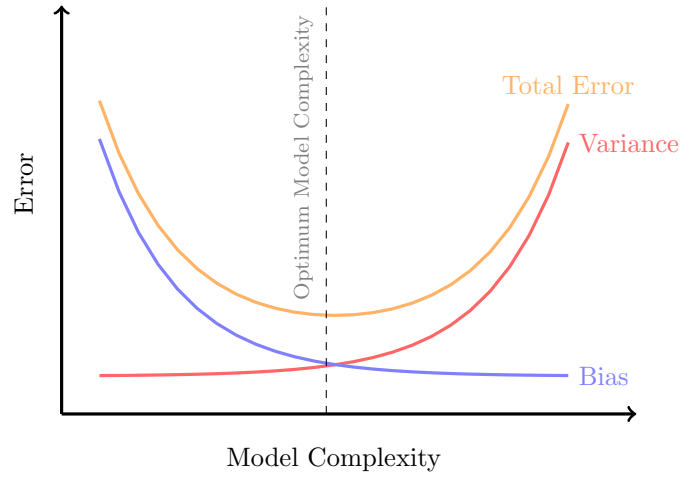


Figure 2.5.: Total Error = Bias + Variance

In 1D, the RMSE is defined by

$$RMSE_{1D} = \sqrt{\frac{1}{N} \sum_i^N (\mathbf{P}_i - \sum_j^M \mathbf{d}_{ij}^{1D} \mathbf{c}_j^{1D})^2} \quad (2.120)$$

where N and M are sizes of the property set and the feature set.

$$RMSE_{nD} = \sqrt{\frac{1}{N} \sum_i^N (\mathbf{P}_i - \sum_n^{n_{dim}} \sum_j^M \mathbf{d}_{ij}^{nD} \mathbf{c}_j^{nD})^2} \quad (2.121)$$

After running a number of cross validations for each set of “hyperparameters”³, a jackknife resampling will be done to check if the result of cross validations is converged. For jackknife resampling we simply take the average over all but the i-th data points which is called the jackknife replicate:

$$\bar{x}_{(i)} = \frac{1}{n-1} \sum_{j \in [n], j \neq i} x_j, \quad i = 1, \dots, n \quad (2.122)$$

³Hyperparameters will be discussed in detail in Sec. 3.5.

Obviously we will get n jackknife replicates and they can be used to calculate the jackknife estimate of the average:

$$\bar{x}_{jack} = \frac{1}{n} \sum_{i=1}^n \bar{x}_{(i)} \quad (2.123)$$

The jackknife estimate of variance of \bar{x} can be calculated from the variance of the jackknife replicates $\bar{x}_{(i)}$:

$$\hat{var}(\bar{x})_{jack} = \frac{n-1}{n} \sum_{i=1}^n (\bar{x}_{(i)} - \bar{x}_{jack})^2 = \frac{1}{n(n-1)} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (2.124)$$

We shall see that the larger size of data sample will give better estimated mean value with smaller variance. By performing jackknife resampling we are able to know whether the variance of RMSE is within the tolerance and that the estimate mean value can be used as the RMSE of the model, or we need more cross validations to give reliable estimate of RMSE.

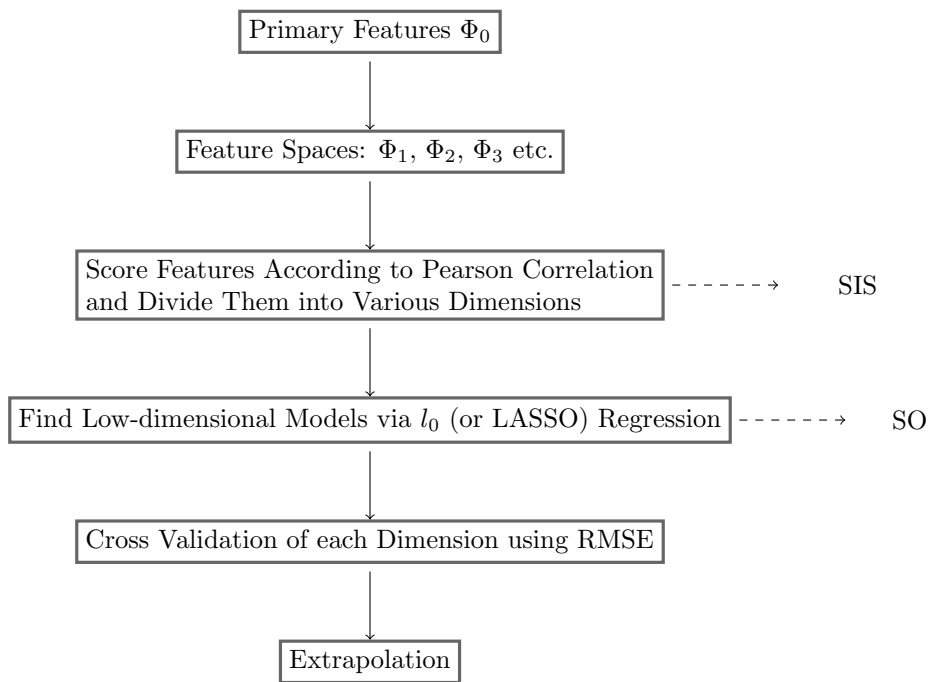


Figure 2.6.: Workflow of SISSO Regression

3. Results and Discussion

3.1. Numerical Settings

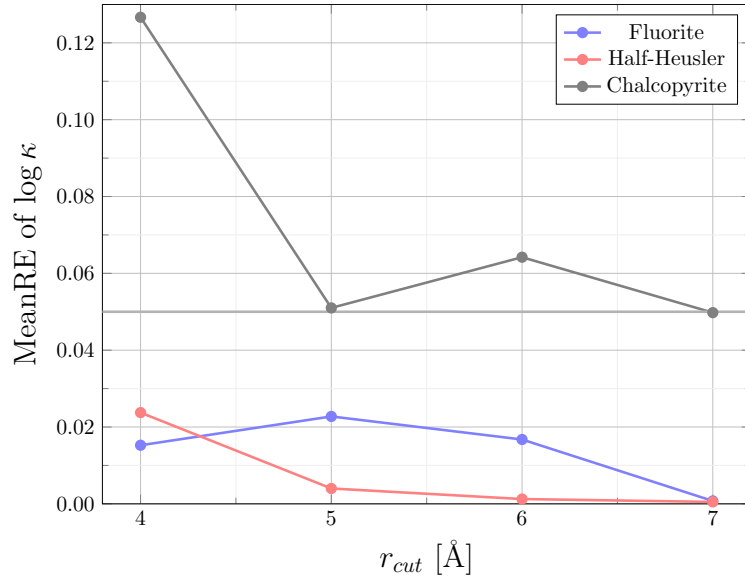
Ab initio forces used in the following sections are computed using the all-electron code FHI-aims with the *numerical atom-centered orbitals (NAOs)* method. PBEsol is used as the exchange-correlation functional and the SCF convergence criterion of density is set to be $10^{-7} \text{eV}/\text{\AA}^3$, k-point density is set as 2\AA^{-3} and *light* settings are used. Other numerical settings are set as the default in the code. All structures are fully relaxed using symmetry-preserving constraints[13], before force calculations are performed.

3.2. Preliminary Convergence Tests

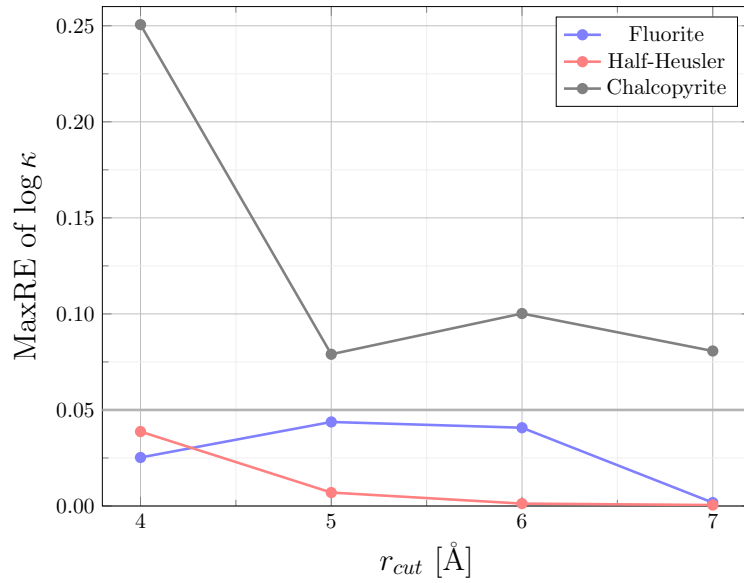
In this work, third order force constants are needed for the calculations of anharmonic quantities using the methods discussed in Sec. 2.1.3. To obtain third order force constants (fc3) by finite-differences, one has to compute the forces acting on each atom caused by displacing a pair of atoms from their equilibrium positions as shown in Eq. 2.80. The number of force pairs needed grows rapidly with the increasing supercell size and complexity of the structure. For instance, for Si supercell with 8 atoms, the displacement pairs needed are 16. However, this number goes up to 222 for the 64-atom Si supercell, and to 434 for the 216 one. Table 3.1 shows the materials to be calculated in our following high-throughput screening, along with the supercell sizes to be used and their corresponding number of atomic pair displacements needed for fc3 calculations.

Materials	Rocksalt	Zinblende	Fluorite	Half-Heusler	Chaclopyrite
Atom per supercell	64	64	96	96	144
Number of pairs	146	222	258	497	5192

Table 3.1.: Number of displaced configurations for five different structures

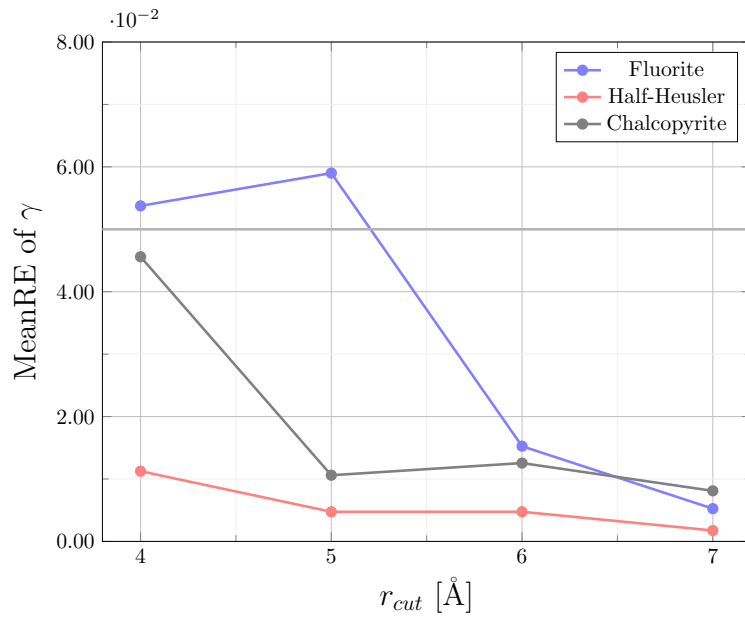


(a)

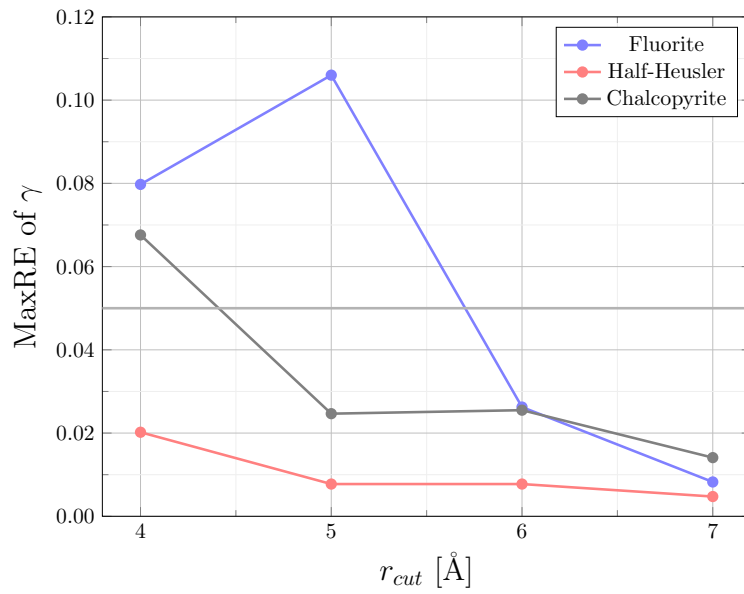


(b)

Figure 3.1.: Mean and Maximum Relative Errors of Thermal Conductivity at 300 K with Cutoff Radius from 4 to 7 Å. Mean relative error is defined as $\text{MeanRE} = \frac{\sum_i^N |x_i - x_{acc}| / x_{acc}}{N}$, where x_i are $\log \kappa$ and x_{acc} are the respective accurate values. MaxRE is the maximum among the set of relative errors.



(a)



(b)

Figure 3.2.: Mean and Maximum Relative Errors of the Grüneisen Parameter with Cutoff Radius from 4 to 7 Å.

To obtain third order force constants, two atoms are displaced simultaneously (\mathbf{r}_1 and \mathbf{r}_2) and such displaced configuration induces forces on all the other atoms given by

$$\mathbf{F}_3 = -\frac{\partial V}{\partial \mathbf{r}_3} \quad (3.1)$$

However, forces caused by the interaction of an atom pair may be shorter-ranged than the chosen supercell size which is determined by the harmonic approximation. It is possible to choose a cutoff radius d_{cut} (with an upper bound equal to distance between the two displaced atoms), such that only atomic pairs with $|\mathbf{r}_1 - \mathbf{r}_2| < d_{cut}$ are considered in the force calculations. The setting of cutoff radius helps reduce the number of pairs computed, but should not affect the accuracy of the third order force constants. Due to the large number of displacement configurations needed for the last three classes of materials in Table 3.1, a cutoff radius is particularly helpful in these calculations.

To check the accuracy, we perform a convergence test of Phono3py calculation for a total of fifteen materials, containing four fluorites (Li_2O , Mg_2Ge , Mg_2Si and Mg_2Sn), four half-heuslers (CoTiSb , FeVSb , NiSnTi and NiSnZr) and nine chalcopyrites (AgGaS_2 , CdAs_2Ge , CdGeP_2 , CuGaS_2 , CuGaSe_2 , CuGaTe_2 , ZnAs_2Ge , ZnAs_2Si and ZnGeP_2). The accuracy for each cutoff radius is evaluated by calculating the relative errors of the respective Grüneisen parameter and thermal conductivity calculated without cutoff radius. For chalcopyrite, we use results with $r_{cut} = 9 \text{ \AA}$ as the accurate values. This r_{cut} ensures a convergence of relative errors within 0.01. Results of the mean relative errors (MeanRE) and maximum relative errors (MaxRE) using different r_{cut} are shown in Fig. 3.1 and 3.2. We note that for thermal conductivity, both fluorite and half-heusler materials show very good convergence with MaxRe and MeanRe smaller than 0.05. Chalcopyrites, however, have larger calculation errors, where the MaxRE is unable to reach a value smaller than 0.05 at 7 \AA cutoff radius. For the results of Grüneisen parameter, the accuracy of the calculations of all the three classes of materials is greatly enhanced from cutoff 5 \AA to 6 \AA , and thus 6 \AA is a good choice. Given that for chalcopyrite materials 5192 displacements have to be computed to get the third order force constants without using a cutoff radius, it is helpful to use a cutoff radius of 6 \AA in the following high-throughput screening, since it halves the amount of displacements.

3.3. Phono3py Results of Thermal Conductivity

Here we perform a high-throughput screening over 152 materials from those five structures discussed in Section 3.2. Parameters for the calculations of *ab initio* forces are set in Sec. 3.1. The Phono3py code is used to calculate third order force constants. Cutoff pair distances of 6 Å are set for fluorite, half-heusler and chalcopyrite, as has been tested to give converged results in the previous section. Thermal conductivity (300 K), Grüneisen parameter and other thermodynamic properties can then be extracted. Other computational details are given in Sec. 3.1.

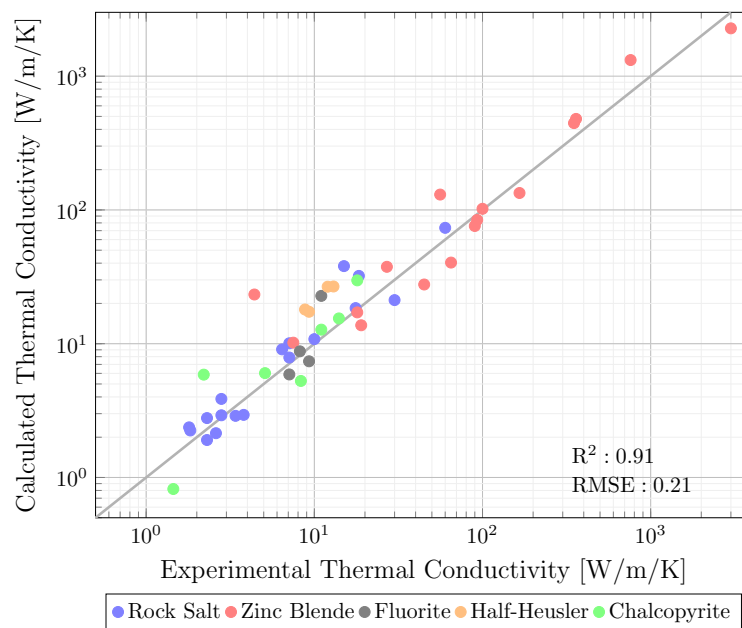


Figure 3.3.: Comparing Thermal Conductivities (300 K) calculated by Phono3py with the experimental values.

Structure	Rock Salt	Zinc Blende	Fluorite	Half-Heusler	Chalcopyrite
Number of materials	18	16	4	4	7
RMSE	0.15	0.24	0.17	0.31	0.22

Table 3.2.: Results of the number of materials for each structures and the RMSEs showing the relationship between experimental and calculated thermal conductivity.

Results for the thermal conductivity are shown in the parity plot (Fig 3.3) between computation and experiment results. Only 49 materials with measured values available are included in this plot. The overall R^2 value of 0.91 indicates relatively good agreement with the experiment results.

We note that part of the rock salt materials show a remarkably small thermal conductivity, which contradicts the common expectation that simple materials have high thermal conductivity [21]. Among them the one with lowest κ_{exp} is NaI ($\kappa_{exp} = 1.80\text{W/mK}$). Other halides are also found to have low thermal conductivity, e.g., RbI ($\kappa_{exp} = 2.30\text{W/mK}$) and KI ($\kappa_{exp} = 2.60\text{W/mK}$). This can be explained by their strong anharmonicity. For instance, their Grüneisen parameters are: 1.565 (NaI), 2.30 (RbI) and 1.587 (KI). Their mode Grüneisen parameters are shown in the Appendix (Fig. A.4).

Table 3.2 lists the number of the materials and the RMSE for each class. Among these five classes, only rock salt, zinc blende and chalcopyrite may have sufficient number of data to give a rough idea of the accuracy of the Phono3py calculations for each class, as measured by the RMSE. Conversely, calculations for fluorite and half-heusler materials are too few to give useful information regarding the accuracy of the calculation for these two individual classes. Nonetheless, we remark that **all** the half-heusler materials are severely overestimated in thermal conductivity, unlike other classes for which materials lie above and below the parity line.

calculated result of κ_{AgGaS_2} does not agree very well with its experimental value, as it shows a larger distance away from the parity line than the other blue points in Fig. 3.4. However, its significantly low thermal conductivity is of more interest to us. In the next part of this section we will investigate more about this material, using Si as a comparison for a good thermal conductor.

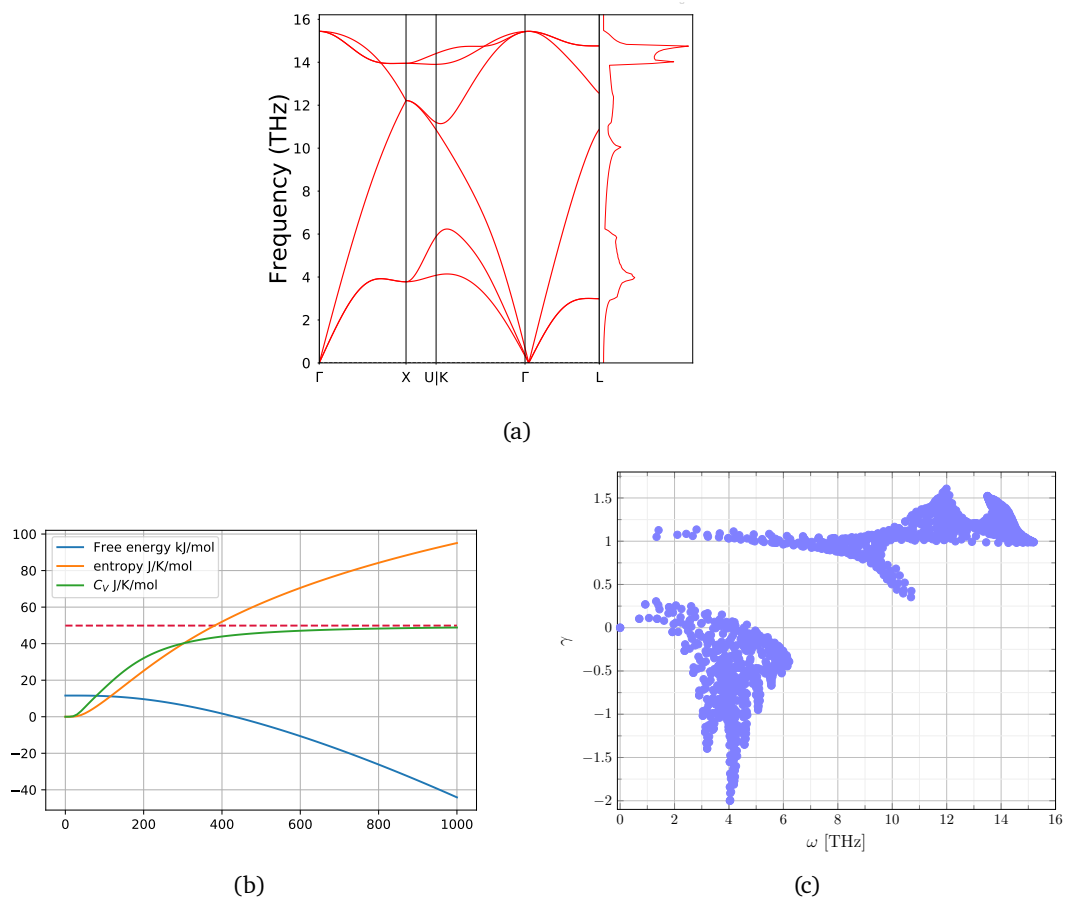


Figure 3.5.: (a) Phonon band structure and density of states, (b) thermal properties (Dashed horizontal line indicates the Dulong-Petit estimation of high temperature heat capacity) and (c) mode Grüneisen parameters of Si

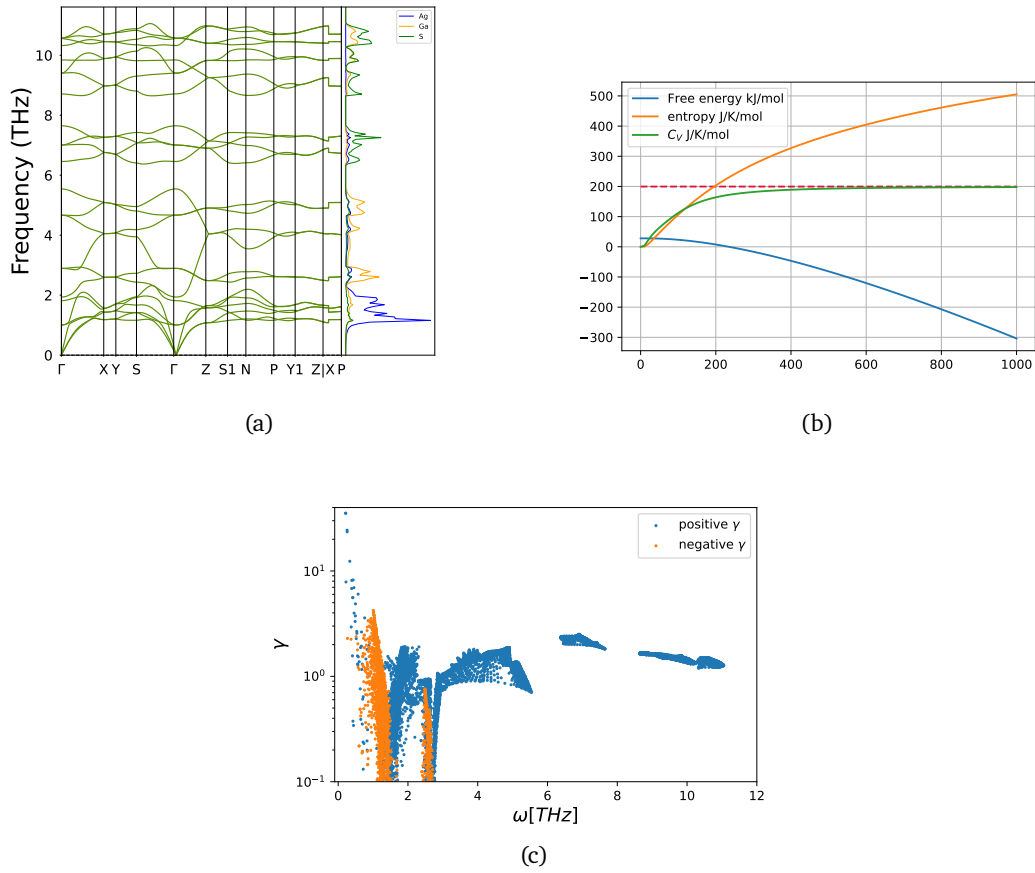


Figure 3.6.: (a) Phonon band structure and density of states, (b) thermal properties (Dashed horizontal line indicates the Dulong-Petit estimation of high temperature heat capacity) and (c) mode Grüneisen parameters of AgGaS_2

Figs. 3.5 show the phonon band structure and density of states, temperature dependent thermal properties and mode Grüneisen parameters of Si. They are important quantities for the evaluation of the dynamical stability of a material. There are two Si atoms in a primitive cell, resulting in six phonon dispersion bands in Fig. 3.5(a). Density of states (DOS) gives an overview of the phonon density distribution across some frequency range.

Fig. 3.5(b) shows the changes of the heat capacity, entropy and free energy under different temperatures. They are extracted from *ab initio* forces as derived in Eq. 2.44. Alternatively,

one can get these quantities from ensemble average by performing aiMD simulations. For free energy, another approach would be to do thermodynamic integration, that is, to integrate along the internal energy "path" between the reference state and the unknown state.

Fig. 3.5(c) shows the Grüneisen parameter of Si per mode. They are calculated from third order force constants via Phono3py. We can see clearly that those Grüneisen parameter points are divided into two clusters, one covers values from 0.25 to -2.0, while the other stays around 1.0. Grüneisen parameters in the first cluster all belong to the acoustic bands, which are the causes of Si's negative thermal expansion at low temperature.

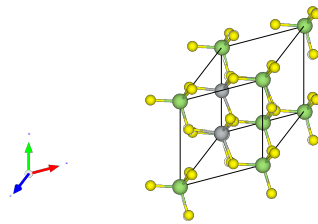
AgGaS₂ is one of the chalcopyrite semiconductors that are promising thermoelectric materials that can be used as both n- and p-type[20]. For chalcopyrite, value of thermal conductivity differs greatly between specific materials. As we found in our calculation, ZnGeO₂, LiBO₂, LiN₂P and SnZnP₂ have thermal conductivity over 20 W/mK, while all Ag-based chalcopyrites show low κ values around 2 W/mK.

Fig 3.6(a) shows the phonon band structure and the partial DOS of the chalcopyrite type (*I42d*) AgGaS₂. There are 8 atoms in the unit cell of AgGaS₂, resulting in 24 phonon branches (3 acoustic branches and 21 optic branches). The flat band shape leads to very small phonon sound velocity v_s of 1397 m/s¹ (for another chalcopyrite ZnGeP₂ with $\kappa = 29.374$ W/mK, $v_s = 5056$ m/s), in particular along X-P direction where all modes are double-degenerated[24]. Following the relaxation time approximation for lattice thermal conductivities, i.e., $\kappa_L = \frac{1}{3}C_v v_s^2 \tau$, the small sound velocity explains the low thermal conductivity. It is also noted that at low frequency region, the optical modes and the acoustic modes overlap, giving rise to a high density. The partial DOS plot on the right shows that density at low frequency is dominated by Ag atoms, while Ga atoms contribute in the intermediate to high frequency region and S atoms in the high frequency region. The partial DOS can then be mapped to other frequency dependent properties (e.g. mode Grüneisen parameter shown in Fig. 3.6(c)) to give an idea of which atom affects other properties.

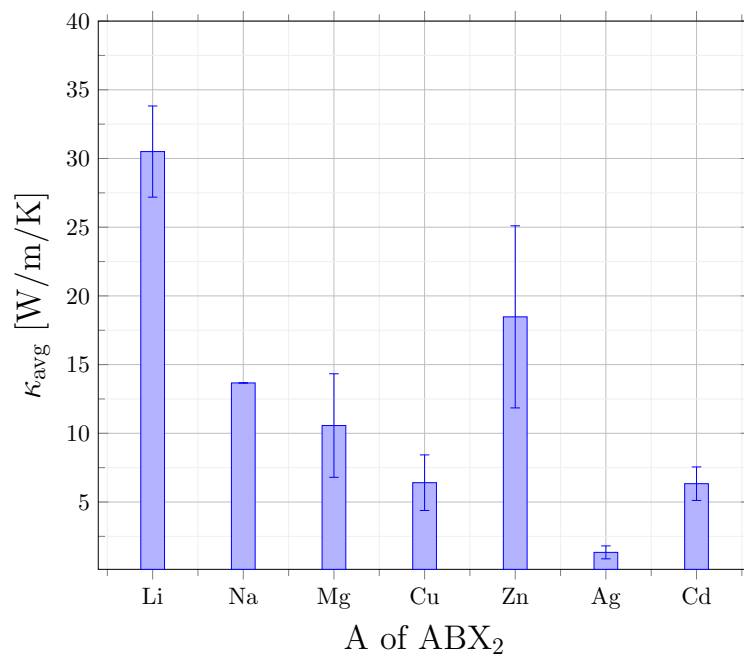
Three thermodynamic properties, i.e., the vibrational free energy E_F , entropy S and heat capacity C_V are plotted in Fig 3.6 (b) as functions of temperature. The absolute values of these three properties are all much greater than that of Si mainly due to the larger number of atoms in the unit cell. The heat capacity reaches a constant at roughly 400 K and follows Dulong-Petit law at higher temperature.

¹Phonon sound velocity is calculated via the derivative of frequency with respect to wave vector \mathbf{q} using finite difference method.

The mode Grüneisen parameter for AgGaS₂ is shown in Fig. 3.6(c). The overall Grüneisen parameter of AgGaS₂ in our calculation is 1.05 (for the relation between overall and mode Grüneisen parameter see Eq. 2.83), which is in good agreement with the reported γ of 1.02[23]. Two different colors represent the positive and negative parts of the mode Grüneisen parameters. The negative γ all come from low frequency branches (mainly acoustic branches), implying that these branches produce negative thermal expansion. What is different is that close to zero frequency, Grüneisen parameters are very large with a maximum of 35.31, from which it drops linearly till 1 THz. Mapping this back to the partial DOS of AgGaS₂ we find that most of these modes belong to Ag atoms which have the highest density at low frequencies.



(a)



(b)

Figure 3.7.: (a) Structure of chalcopyrite compound, (b) Average thermal conductivity (calculated) for different A-based chalcopyrites (A = Li, Na, Mg, Cu, Zn, Ag and Cd)

As observed in the case of AgGaS_2 , the low frequency Ag atoms with large Grüneisen parameters might explain the low thermal conductivity of this compound. Now we would like to see if this assumption can be generalized to the whole chalcopyrite space.

Fig. 3.7(a) shows the structure of the tetragonal unit cell of ternary chalcopyrite with

ABX₂ composition (I – III – VI₂ or II – IV – V₂). The yellow points stand for atom X (S, Se, Te, etc.), while the gray atoms (A = Cu, Ag, etc.) and the green atoms (B = Al, Ga, In, etc.) equally occupy the tetrahedron holes.

To reveal the A-atom dependent trend for the thermal conductivity in ABX₂ type chalcopyrite, we divide the 38 chalcopyrite materials into 7 groups based on the element on their A site. Here, GaLiTe₂ and CdSnP₂ are excluded since they do not follow the general trends. Of all the remaining 36 chalcopyrite materials, Li, Na, Cu and Ag based compounds correspond to I – III – VI₂ composition, while Mg, Zn and Cd correspond to II – IV – V₂ composition. The ionization number differs in these two cases (for I – III – VI₂ A is a +1 cation and for II – IV – V₂ A is a +2 cation). Then we calculate the averaged thermal conductivity for each group with results shown in Fig. 3.7(b). Generally speaking, thermal conductivity decreases with increasing atomic number/mass as shown from left to right in the bar chart. Ag-based chalcopyrite shows the lowest average thermal conductivity of 1.19 W/mK. Exceptions are found in Zn and Cd, where the average thermal conductivity of Zn-based compound is 18.5 W/mK and for Cd it is 6.3 W/mK. The descending trend of I – III – VI₂ compounds is in general consistent with the experimentally fitted model for thermal conductivity given in Ref. [23], which states that the relation $\kappa_L \propto \bar{M}\delta\Theta_D^3$ (\bar{M} is the average atomic mass, δ is the average volume occupied by one atom of the crystal and Θ_D is the Debye temperature) holds for each individual compositions, i.e., I – III – VI₂ and II – IV – V₂. However, the unexpected high average thermal conductivity of Zn-based compounds breaks this trend for II – IV – V₂ compounds.

3.4. Anharmonicity Measure

Can we obtain or at least have an initial guess of thermal conductivity without performing any costly Phono3py or MD calculations?

Of course one can do such predictions using machine learning if enough reliable data is available. But before that, we would like to first look for descriptors in physics domain. It is known that the thermal conductivity is finite due to anharmonicity. Here, we will present three anharmonicity metrics, i.e., σ^A , σ_{os}^A and Grüneisen parameter γ , and compare their performance as an anharmonicity measure by investigating their correlations with the thermal conductivity.

The idea behind σ^A and σ_{os}^A is based on Ref. [10]. They are extracted from the RMSE of normalized forces obtained via aiMD and harmonic sampling respectively (see the

derivation in Sec. 3). Both R^2 and RMSE are used to measure how well they can predict the thermal conductivity linearly. From Fig. 3.8(a) and 3.8(b) one can easily note that at the region near 10W/mK four materials are far off the regression line. They are the half-heusler materials (CoTiSb, FeVSb, NiSnTi and NiSnZr), corresponding to the four orange points well below the equality line in Fig. 3.3.

The greater R^2 and smaller RMSE of σ^A than that of σ_{os}^A indicate that σ^A is a better descriptor for thermal conductivity and anharmonicity. This can be explained from Fig. 3.8(a) where the data points of strongly anharmonic materials with larger σ^A are closer fitted to the regression line. As discussed in Sec. 3, σ_{os}^A is an approximation to σ^A based on the ensemble average of perfectly harmonic system. Errors thus must become larger for more anharmonic materials.

We further plot the correlation between κ_{exp} and γ as shown in Fig. 3.8(c). γ is correlated to the thermal conductivity via the Slack model and the Debye-Callaway model discussed in Sec. 3.3. They both tell that the phonon scattering rate $\tau^{-1} \propto \gamma^2$ (where the relaxation time $\tau \propto \gamma^{-2}$), and τ is proportional to the lattice thermal conductivity. The R^2 and RMSE show, however, that its correlation with the thermal conductivity is not particularly pronounced. This suggests that in both the Slack and the Debye-Callaway model, κ is dominated by Debye temperature Θ_D and other properties (atomic mass, volume, etc.) and that the Grüneisen parameter, however, plays only a minor role.

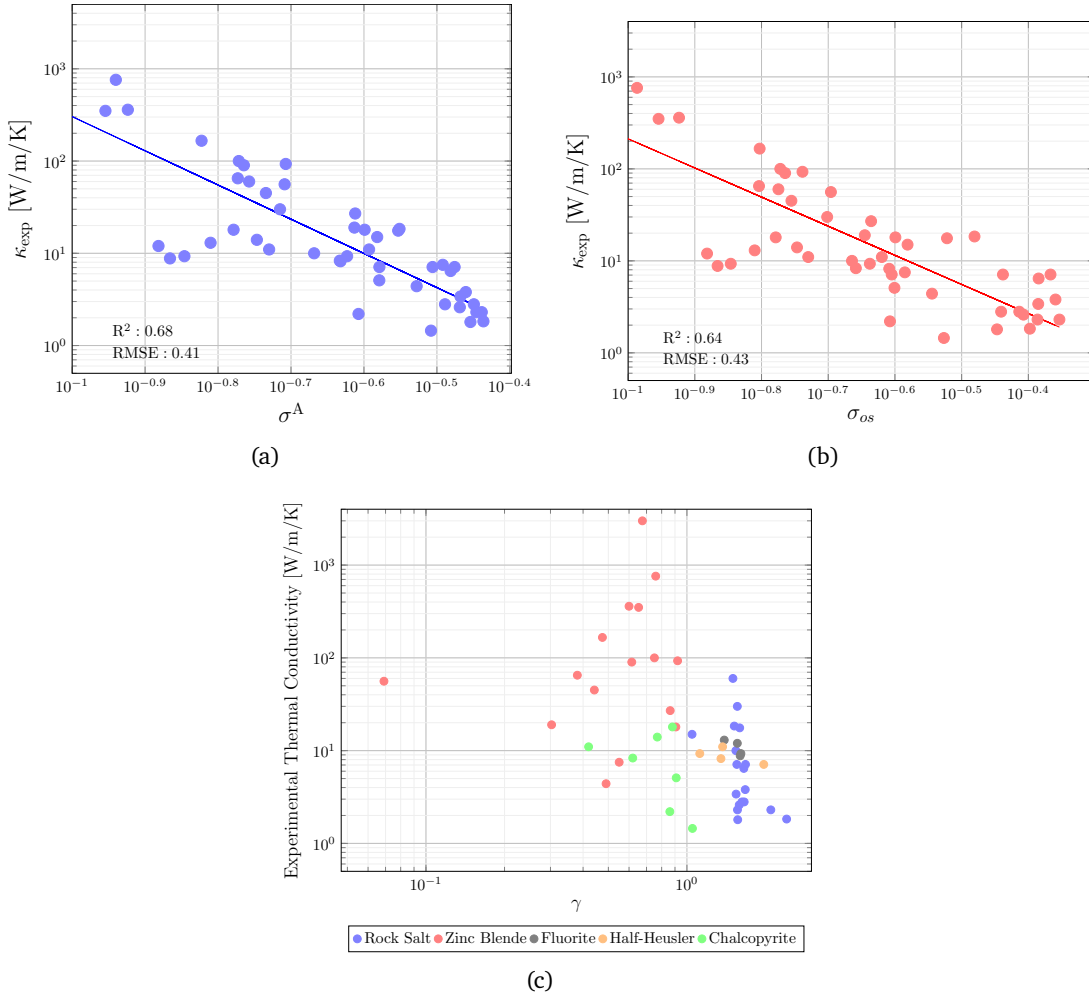


Figure 3.8.: Correlation of the three anharmonic metrics (at 300K) (a) σ^A , (b) σ_{os}^A and (c) γ^{-2} with the thermal conductivity, straight lines show the regression results (Due to bad correlation fitting line is removed from (c)).

3.5. SISSO Regression for Thermal Conductivity

As discussed in Sec. 3, both σ_{os}^A and γ don't encompass all aspects of thermal conductivity. For instance, the Slack model [21] for thermal conductivity reads

$$\kappa_L = A \frac{\bar{M} \Theta_D^3 \delta}{\gamma^2 T} \quad (3.2)$$

where \bar{M} is the average atomic mass, Θ_D is the Debye temperature, δ average volume per atom in the unit cell, γ is the Grüneisen parameter, T is temperature and A is a constant. This model has been proved to describe successfully in high- κ materials, e.g., Si, MgO, BP, etc. However, it fails in the case of strongly anharmonic materials. To check the accuracy of the Slack model, we plot the relation between κ_{exp} and the coefficient $M_a \Theta_a^3 \delta / \gamma^2$ using the κ data we calculated, as shown in Fig. 3.9. It shows that a roughly linear correlation can be reached for materials with κ_{exp} greater than 50 W/mK, whereas no linear trend is found for lower κ_{exp} .

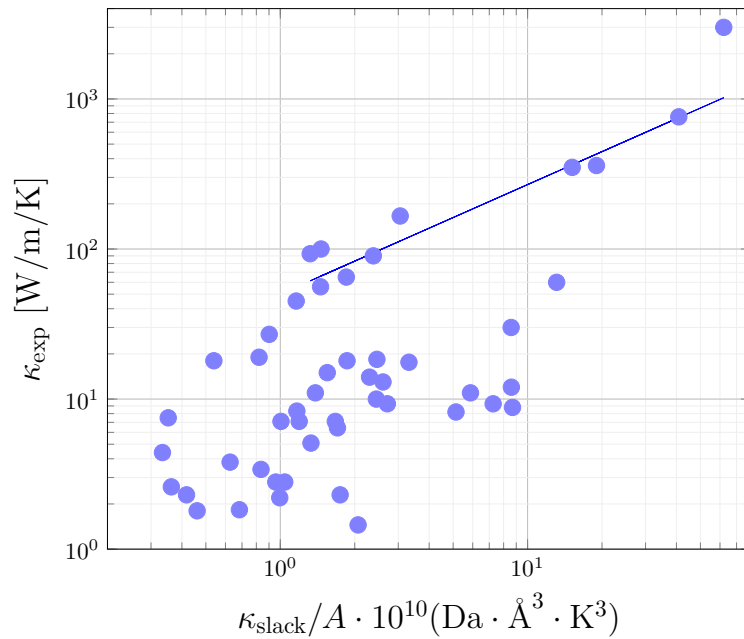


Figure 3.9.: Experimental values of thermal conductivity as a function of the coefficient $M_a \Theta_a^3 \delta / \gamma^2$ in the Slack model

Machine learning provides a way to include more features relevant to the other aspects of the thermal conductivity. As introduced in Sec. 2.2, SISSO (sure independence screening and sparsifying operator) is able to tackle immense and correlated feature spaces, and converges to the optimal solution from a combination of relevant to the materials' property [16].

In this work, the set of 49 experimental data shown in Sec. 3.3 will be used as the target property. As preprocessing step κ_{exp} will first be converted to $\log(\kappa_{exp})$ to reduce the range of possible values of κ . Table 3.3 shows the primary features (or training data) used to construct the feature spaces for the symbolic regression as discussed in Sec. 2.2.2, where some stem from *ab initio* calculations, some are structure-specific parameters. They have been selected as primary features since they represent relevant processes for the thermal conductivity. It is thus reasonable to assume that a combination of these physical features in the higher rungs of feature space can give a better description of thermal conductivity.

There are four parameters forming a set of hyperparameters that have to be adjusted and optimized during the training, which are

- (1) **n_sis_select**: number of features selected for each dimension;
- (2) **n_residual**: number of residuals of loss function taken to the calculation of RMSE for each dimension;
- (3) **max_rung**: feature space are built up to $\Phi_{\text{max_rung}}$ and
- (4) **desc_dim**: number of dimensions considered.

All hyperparameters in the following regressions are optimized using 50 or 100 iterations of leave-10% out cross validation (CV), i.e., 10% of the data in training set are used as test data in the CV.

Feature	Definition
γ	Grüneisen parameter
σ^A	Measure of degree of anharmonicity from MD [10]
σ_{os}^A	One shot measure of degree of anharmonicity [10]
$L_{avg_{prim}}$ (Å)	Average lattice constants of primitive cell
Θ_p (K)	Average phonon temperature [17]
$\Theta_{D\infty}$ (K)	Estimated high-temperature limits of Debye temperature [17]
Θ_a (K)	Debye temperature of the acoustic phonon branches
C_v (J/K)	Heat capacity
$\omega_{\Gamma_{max}}$ (THz)	Maximum frequency at Γ point
m_{min} (Da)	Minimum atomic mass
m_{max} (Da)	Maximum atomic mass
m_{avg} (Da)	Average atomic mass
μ (Da)	Reduced mass ¹
V_m (Å ³)	Molar volume
V_a (Å ³)	Volume per atom
ρ (Da/Å ³)	Density
n_atoms_prim (n_atom)	Number of atoms in the primitive cell
v_s (m/s)	Phonon sound velocity

¹ $\mu = \frac{1}{\sum_i \frac{1}{m_i}}$, where m_i are the masses of atoms in the primitive cell.

Table 3.3.: The primary features used in the following regressions. Definitions are given in the right column.

3.5.1. Regression including γ

In this section, SISSO regression is performed with γ included and σ^A/σ_{os}^A excluded. A total of 49 materials are taken to the training.

Cross Validations (CVs) of Models

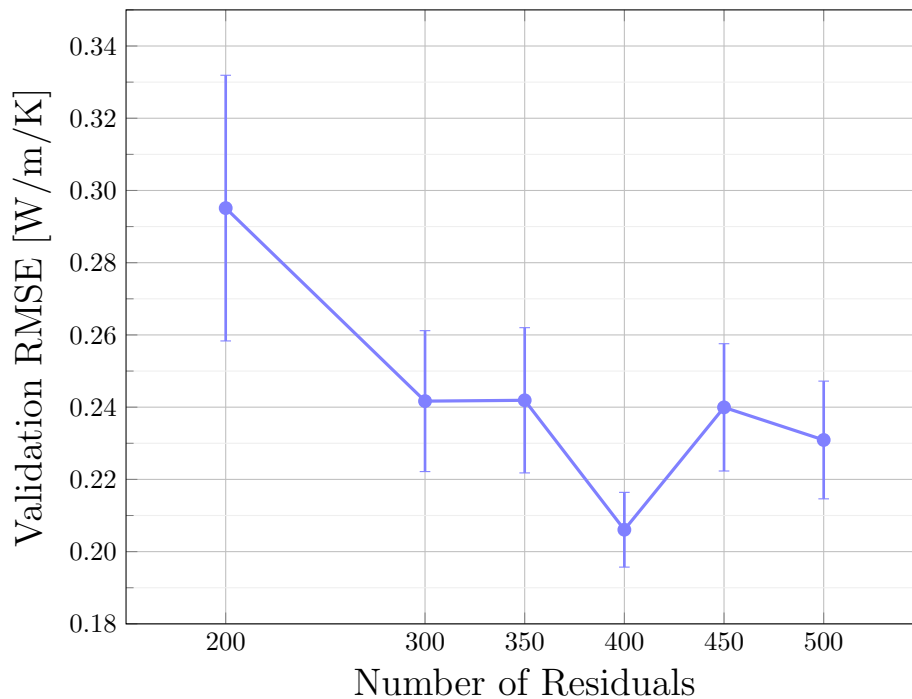


Figure 3.10.: Validation RMSE of 100 CVs. Other parameters are: `max_rung = 2`, `desc_dim = 3` and `n_sis_select = 500`.

Fig. 3.10 shows the RMSE obtained from 100 cross validations as the final step of the cross validations, with `n_residual` ranging from 200 to 500. A set of 50 CVs is performed first using the same hyperparameter settings, RMSE given in Fig. B.2, where the other parameters are already optimized through the initial cross validations with results shown in Fig. B.1 and B.3. In Fig. B.2 it is unclear whether 400 is the best parameter and then we increase the number of cross validations to 100 (Fig. 3.10). Based on this results we are

sure that n_{residual} should be 400, with the validation RMSE being 0.209. In Table B.1 we provide a list of the best model with smallest RMSE across all CVs for each dimension (hyperparameters are the same as the final training), which may give an idea of how models vary with dimension and cross validation.

Final Training

To finally determine a descriptor for this regression, we run another training with all materials included in the training set. The parity plot and the fitting errors are shown in Fig. 3.11 and Table 3.4². The maximum absolute error corresponds to LiH, whose thermal conductivity is underestimated by nearly 50%. Nevertheless, the overall regression performs well with $R^2 = 0.97$. Table 3.4 also lists the errors for each individual materials class, from which we get a rough idea of how well the model describes one certain class. Obviously, the consistently small errors confirms the accuracy of this model across the whole dataset.

The obtained equation for the descriptor is given by

$$\kappa_{est} = c_0 + a_0 \frac{\rho}{\sqrt{v_s}} + a_1 \frac{\sqrt[3]{m_{min}}}{L_{avg_prim}^2} + a_2 \Theta_p^2 V_a m_{avg} \quad (3.3)$$

where $c_0 = -1.743 \cdot 10^{-2}$, $a_0 = -17.97$, $a_1 = 5.515$ and $a_2 = -1.743e-08$.

From a physics point of view, the primary features entering the equation should be individually related to the thermal conductivity as: $\rho(-)$, $v_s(+)$, $m_{min}(-)$, $L_{avg_prim}(+)$, $\Theta_p(+)$, $V_a(+)$ and $m_{avg}(-)$ ³. Throughout the equation only the density ρ , the sound velocity v_s and the m_{avg} agree with this physical intuition. However, this ignores the correlations between the primary features, which may lead to different results. For example, V_a is actually inversely correlated to thermal conductivity because V_a and Θ_p are also inversely correlated to each other. The latter has strong positive impact on the thermal conductivity. The large differences between the coefficients might play a role, but this is in fact also related to the different order of magnitude of the primary features.

²Here both RMSE and R^2 are used in the error assessment. RMSE tells us the typical distance between the predicted value and the actual value, while R^2 tells us how well the predictor variables can explain the variation in the response variable.

³(+) stands for positive correlation, while (-) for negative correlation.

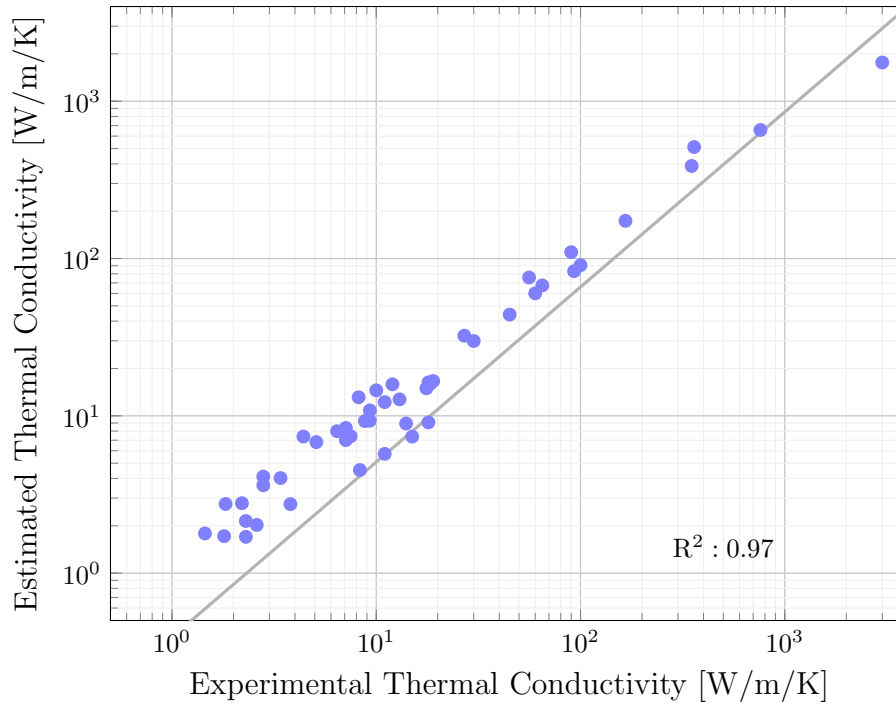


Figure 3.11.: Parity plot of the final training using the optimized hyperparameters with all materials included as the training data.

	MAE	MaxAE	RMSE
Overall	0.10	0.31	0.13
Rock salt	0.09	0.31	0.12
Zinc blende	0.09	0.30	0.13
Fluorite	0.08	0.20	0.11
Half-heusler	0.05	0.12	0.07
Chalcopyrite	0.16	0.28	0.18

Table 3.4.: Errors of the final training are presented in four ways, i.e, mean absolute error (MAE), maximum absolute error (MaxAE), root mean square error (RMSE) and R²

To make it easier to evaluate the contribution of each feature directly from the model equation, we rescale the primary features and the property before running the regression. Accordingly, all the data will be scaled to the range [0, 1]. This transformation is given by

$$x_{i_scaled} = \frac{x_i - x_{min}}{x_{max} - x_{min}} \quad (3.4)$$

where x_{min} and x_{max} are the minimum and the maximum number of each primary feature/property array.

The Model for the scaled training is given by

$$\kappa_{est} = c_0 + a_0(\omega_{\Gamma_{max}}^3) \cdot (\rho \cdot V_m) + a_1(\omega_{\Gamma_{max}}^3) \cdot (\rho \cdot V_a) + a_2(v_s^6) \cdot (\mu + m_{avg}) \quad (3.5)$$

where $c_0 = 1.57e-03$, $a_0 = -9.35$, $a_1 = 22.26$ and $a_2 = 4.45$. Now the coefficients are more comparable and we can state that the model depends more on the 2D descriptor containing $\omega_{\Gamma_{max}}$, ρ and V_a than the other two. All the three parameters are physically making sense, since a larger $\omega_{\Gamma_{max}}^3$ gives rise to higher Θ_D and thus higher κ , and ρ and V_a are inversely and positively correlated to κ respectively as discussed above. Fitting errors are listed in Table 3.5.

MAE	MaxAE	RMSE	R ²
0.002	0.01	0.003	0.99

Table 3.5.: Errors of the scaled training. Due to the small range of the data, regression errors look much better than in Table 3.4 for the original data.

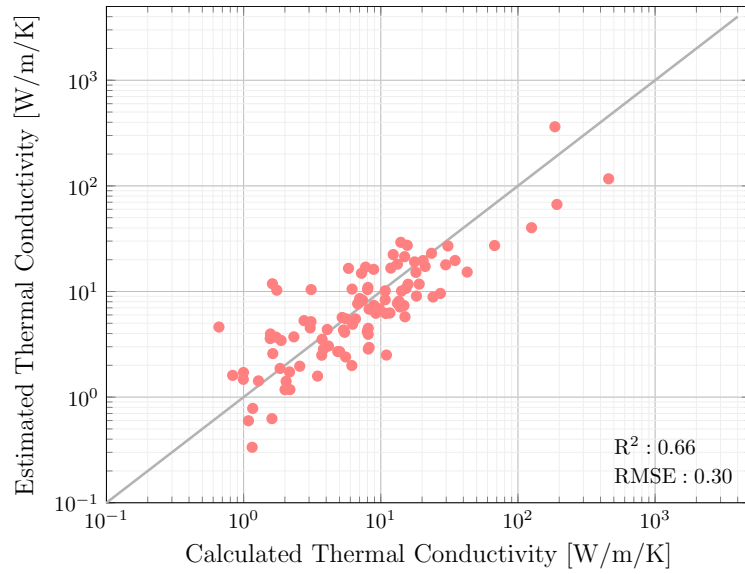
Predicting Property of New Materials

In the next step we will use these models to predict the thermal conductivity of the other 101 materials without measured data, but for which Phono3py data were produced. Two predictions using models obtained from direct and scaled regressions are performed. For the scaled one, primary features in the dataset are also scaled to [0, 1] based on Eq. 3.4 in advance. The estimated results are then be compared to their Phono3py results. The resulting parity plots for the direct and scaled regressions are shown in Fig. 3.12(a) and 3.12(b), respectively. R² and RMSE are both errors for the log values not the direct ones.

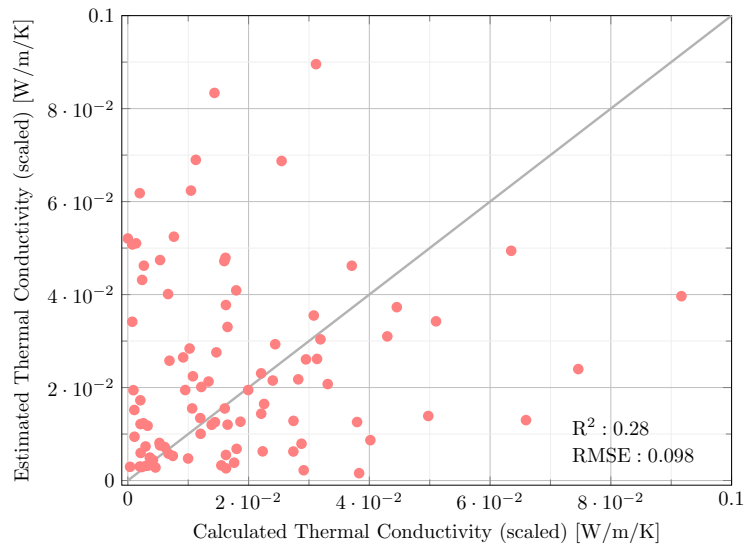
We can clearly see from these two figures that the descriptor obtained from the direct regression performs better in predicting thermal conductivity of new materials as most of the data points stay close to the parity line and the R^2 is much larger than the scaled one⁴. Note that in the case when data are of different scales, the comparison of RMSE is no longer making sense. For the direct regression, the mean absolute error (MAE) and maximum absolute error are 0.25 and 0.86. 70 of 101 materials have absolute error below RMSE (0.30), while 56 below MAE (0.25). The errors are in general normally distributed.

Based on the above results and discussion, we will continue using direct training data and $\log \kappa_{exp}$ as target property in the following regressions.

⁴The bad prediction of the scaled regression is also attributed to the lack of CVs which is the limitation of this study.



(a)



(b)

Figure 3.12.: Predictions of new set of materials using (a) direct regression and (b) scaled regression (for scaled regression, only points lying within $[0, 0.1]$ are shown on the plot)

3.5.2. Regression including σ^A , σ_{os}^A and γ

In this section we would like to check if the inclusion of σ^A and σ_{os}^A can give a better interpretation of the thermal conductivity, since in Sec. 3 they have been found as better descriptors of anharmonicity over γ .

Cross Validations of Models

Similar to the regression with γ , hyperparameters have to be fixed via cross validations. Previous regressions suggests that `max_rung` = 2 is already sufficient and will be adapted here. 50 CVs with leave-10% out fraction are performed for each setting. Cross validation results for `n_residual` and `desc_dim` are shown in Fig. 3.13. Best models with smallest RMSEs for each `n_residual` and `desc_dim` are listed in Table B.2. From the cross validations we have found the optimized hyperparameter set, which is

<code>max_rung</code>	<code>desc_dim</code>	<code>n_sis_select</code>	<code>n_residual</code>
2	2	500	500

Table 3.6.: Hyperparameters for the Final Training

Parameters given in Table 3.6 will be taken to the final training where all data are included in the training set.

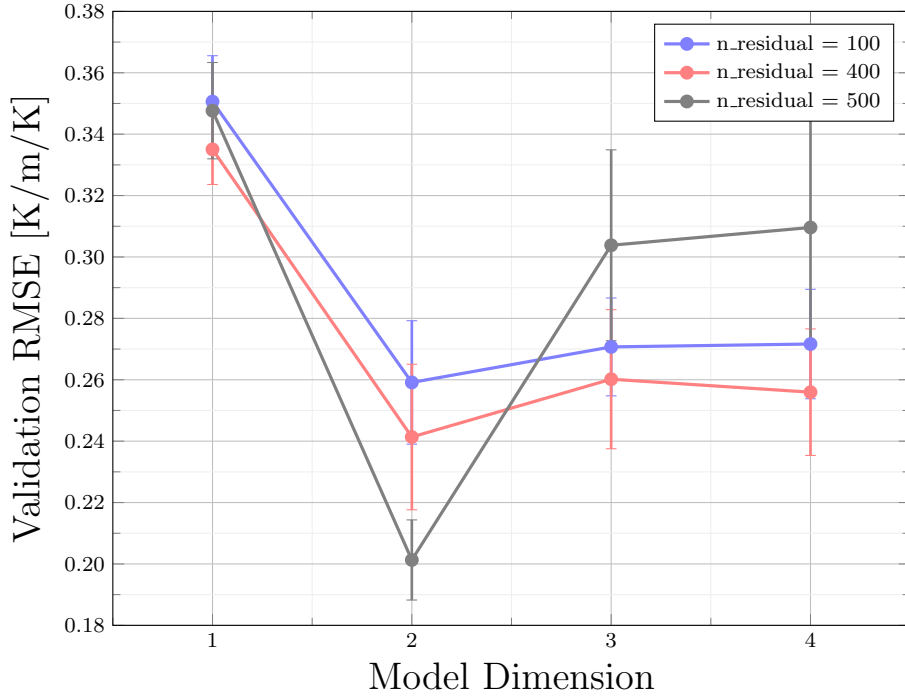


Figure 3.13.: Validation RMSEs of models containing 500 features per dimension. Other hyperparameters are `max_rung = 2` and `n_sis_select = 500`. The smallest RMSE (0.201) is found at dimension 2 with `n_residual = 500`.

Final Training

The resulting model is

$$\kappa_{est} = c_0 + a_0 \cdot ((\rho \cdot \Theta_a) \cdot (V_m \cdot \gamma)) + a_1 \cdot \frac{\ln(\sigma^A)}{\sqrt[3]{L_{avg_prim}}} \quad (3.6)$$

where $c_0 = -0.8043$, $a_0 = -8.478e-06$ and $a_1 = -2.4778$. In this model, both γ and σ^A enter the equation and are negatively correlated with κ_{est} . We also find that all the 2D models in Table B.2 (0.10 left out) and this one take the same form with same primary features included, though the coefficients are different.

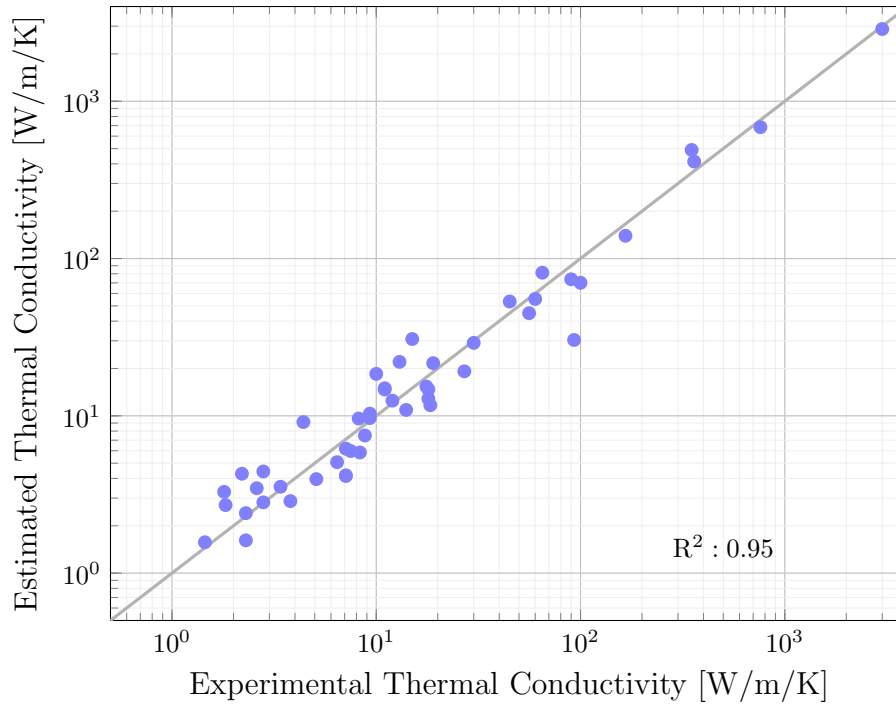


Figure 3.14.: Parity plot between experimental and estimated thermal conductivities. (n_{sis}=500, n_{res}=500)

	MAE	MaxAE	RMSE
Overall	0.13	0.49	0.16
Rock salt	0.13	0.31	0.16
Zinc blende	0.13	0.49	0.17
Fluorite	0.11	0.23	0.14
Half-heusler	0.09	0.23	0.12
Chalcopyrite	0.13	0.29	0.15

Table 3.7.: Errors of the final training, i.e, mean absolute error (MAE), maximum absolute error (MaxAE), root mean square error (RMSE) and R^2

Comparing Table 3.7 and Table 3.4, we find that the two models (the direction regression

in Sec. 3.5.1 and this one) give close regression results, though errors are slightly smaller in the first model. However, the second model including σ^A and σ_{os}^A has the advantage in describing thermal conductivity of chalcopyrite materials, where both the MAE (0.13) and the RMSE (0.15) are lower than the first one. Furthermore, the second model looks simpler with only two dimensions involved, and hence fewer primary features are required.

Predicting Property of New Materials

As a further validation of the model, a total of 90 new materials⁵ are taken to the prediction for their thermal conductivity. The dataset is composed of 20 rock salt, 6 zinc blende, 19 fluorite, 13 half-heusler and 32 chalcopyrite.

The prediction results are shown in Fig. 3.15. It is not surprising that this model is able to predict as well as the previous model (with only γ) does due to the close training errors of them. The slightly better R^2 may be explained by the smaller RMSE in cross validation step, where the overfitting problem is better removed. For the prediction, Phono3py results of κ are used as property dataset instead of the experiment data that includes the full anharmonicity (so does σ^A). The lack of full anharmonicity in the Phono3py results may be one of the causes for the slightly larger RMSE of this model than the previous one. Also, this model is able to give prediction with greater R^2 using a smaller dataset than the first model.

To summarize, both the two models we obtain through SISSO regression have the ability to accurately describe the thermal conductivity, while the model including σ^A as primary feature outperforms the one without σ^A for the reason that (1) the equation of the model takes a simpler form with feature space up to only 2D, (2) the better CV result means less prediction error, reflected by the larger R^2 value in predicting new materials and (3) the inclusion of σ^A makes it better in describing strongly anharmonic materials, as shown in the case of chalcopyrite.

⁵11 among all the 101 calculated materials do not have σ^A data, thus only 90 are used, which makes the following discussion not fully convincing. This is also a limitation of this work.

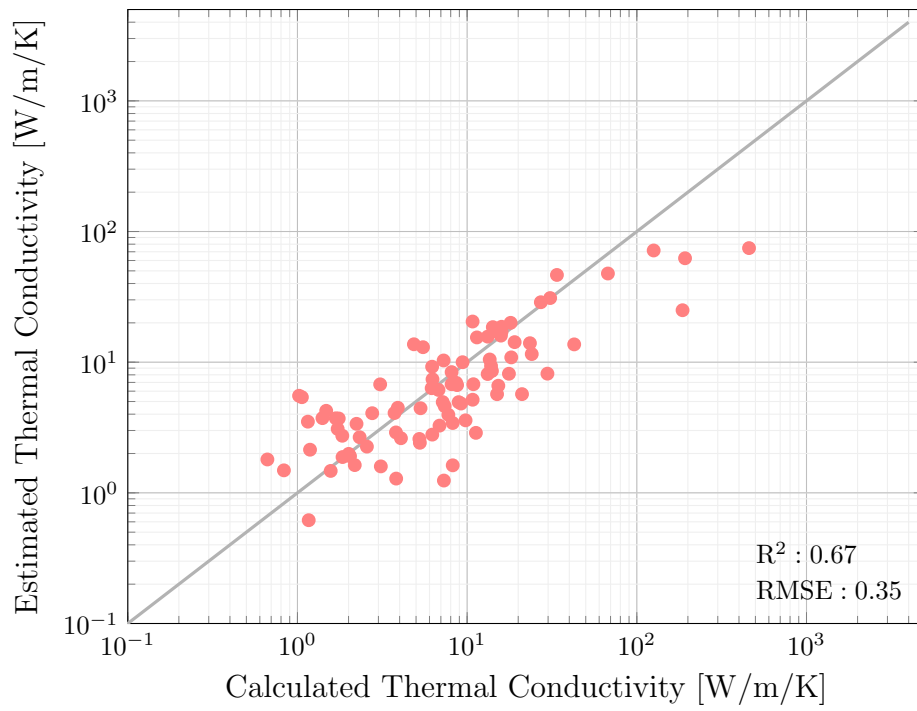


Figure 3.15.: Predictions of new set of materials using model shown in Eq. 3.6

4. Conclusions

In this thesis, we have focused on the screening for low- κ materials with potential for applications such as thermoelectrics. To this end, a systematic investigation is performed, starting from a high-throughput computation for the thermal conductivity including five classes of materials (rock salt, zinc blende, fluorite, half-heusler and chalcopyrite), followed by a SISSO regression for discovering descriptors to be used in the prediction of κ for materials without experimental values. The key findings are summarized as follows.

1. $r_{cut} = 6.0 \text{ \AA}$ is able to yield accurate third order force constants (fc3) for fluorite, half-heusler and chalcopyrite (convergence results see Sec. 3.2). This r_{cut} is set in the following high-throughput computation for the above materials. r_{cut} are not set in other two classes, rock salt and zinc blende, due to the low symmetry.
2. The high-throughput results are validated by comparing with experimental data. Good agreement has been reached in general, with $R^2 = 0.91$ and $RMSE = 0.21$ for log-scaled values, though all half-heusler materials are overestimated (Fig. 3.3). Based on the calculation, 83 out of 152 investigated materials are thermal insulators with $\kappa_{cal} < 10 \text{ W/mK}$ (see Table. A.1 and A.2).
3. A detailed study of AgGaS_2 as a low- κ , strong anharmonic material gives some insights into the origins of its low thermal conductivity as well as the strong anharmonicity. Based on the analysis of the band structure, DOS and mode Grüneisen parameter, the low thermal conductivity may have something to do with the Ag atom, and all other Ag-based chalcopyrites also show low thermal conductivity across the chalcopyrite space. This can give us some hints into what is going on overall. However, More intensive and systematic investigations are needed for a better understanding of these effects.
4. A comparative study of three possible anharmonic metrics (σ^A , σ_{os}^A and γ) reveals that there is a linear correlation of σ^A and σ_{os}^A (for definitions see Sec. 3) with thermal conductivity, where σ^A shows larger R^2 and smaller RMSE. The Grüneisen

parameter, γ , however, is not particularly suited to describe thermal conductivity (Fig. 3.8) at least on its own.

5. Eventually, two models are obtained via SISSO regressions and cross validations, one including only γ for the description of anharmonicity, and the other including both γ and σ^A . The second model performs better as the descriptor of thermal conductivity for the following reasons: (1) the equation (Eq. 3.6) of the model takes a simpler form with feature space up to only 2D and thus fewer primary features are required, (2) the better CV result leads to smaller prediction error, reflected by the larger R^2 value in predicting new materials and (3) the inclusion of σ^A makes it better in describing strongly anharmonic materials, as shown in the case of chalcopyrites.

Due to time and computational limits, only a relative small sets of materials covering only five materials was explored. Similarly, only a low order description of anharmonicity was used for calculating κ . This approximation is especially problematic for thermal insulators. In future, it would thus be interesting to study if the promising findings in this thesis hold up when a larger material space is explored and more accurate predictions for κ are used.



Acknowledgements


Thanks for the chance to do master thesis in NOMAD Laboratory of Fritz Haber Institute, Berlin. Thanks to Chris and Tom. The thesis would not be possible without their help.

感谢远在国内的家人。

Bibliography

- [1] Neil W Ashcroft, N David Mermin, et al. *Solid state physics*. 1976.
- [2] Axel D Becke. “Density-functional exchange-energy approximation with correct asymptotic behavior”. In: *Physical review A* 38.6 (1988), p. 3098.
- [3] Axel D Becke. “Density-functional thermochemistry. V. Systematic optimization of exchange-correlation functionals”. In: *The Journal of chemical physics* 107.20 (1997), pp. 8554–8560.
- [4] Kieron Burke, John P Perdew, and Yue Wang. “Derivation of a generalized gradient approximation: The PW91 density functional”. In: *Electronic density functional theory*. Springer, 1998, pp. 81–111.
- [5] Christian Carbogno, Rampi Ramprasad, and Matthias Scheffler. “Ab initio Green-Kubo approach for the thermal conductivity of solids”. In: *Physical review letters* 118.17 (2017), p. 175901.
- [6] Lucian A Constantin, John P Perdew, and José Maria Pitarke. “Exchange-correlation hole of a generalized gradient approximation for solids and surfaces”. In: *Physical Review B* 79.7 (2009), p. 075126.
- [7] Hossein Asnaashari Eivari et al. “Low thermal conductivity: fundamentals and theoretical aspects in thermoelectric applications”. In: *Materials Today Energy* (2021), p. 100744.
- [8] Luca M Ghiringhelli et al. “Learning physical descriptors for materials science by compressed sensing”. In: *New Journal of Physics* 19.2 (2017), p. 023017.
- [9] P Hohenberg and WJPR Kohn. “Density functional theory (DFT)”. In: *Phys. Rev* 136 (1964), B864.
- [10] Florian Knoop et al. “Anharmonicity measure for materials”. In: *Physical Review Materials* 4.8 (2020), p. 083809.

-
-
- [11] Franz Knuth et al. “All-electron formalism for total energy strain derivatives and stress tensor components for numeric atom-centered orbitals”. In: *Computer Physics Communications* 190 (2015), pp. 33–50.
- [12] Ryogo Kubo, Mario Yokota, and Sadao Nakajima. “Statistical-mechanical theory of irreversible processes. II. Response to thermal disturbance”. In: *Journal of the Physical Society of Japan* 12.11 (1957), pp. 1203–1211.
- [13] Maja-Olivia Lenz et al. “Parametrically constrained geometry relaxations for high-throughput materials science”. In: *npj Computational Materials* 5.1 (2019), pp. 1–10.
- [14] Richard M Martin. *Electronic structure: basic theory and practical methods*. Cambridge university press, 2020.
- [15] DT Morelli, JP Heremans, and GA Slack. “Estimation of the isotope effect on the lattice thermal conductivity of group IV and group III-V semiconductors”. In: *Physical Review B* 66.19 (2002), p. 195304.
- [16] Runhai Ouyang et al. “SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates”. In: *Physical Review Materials* 2.8 (2018), p. 083802.
- [17] Roland Pässler. “Basic moments of phonon density of states spectra and characteristic phonon temperatures of group IV, III–V, and II–VI materials”. In: *Journal of Applied Physics* 101.9 (2007), p. 093513.
- [18] John P Perdew and Alex Zunger. “Self-interaction correction to density-functional approximations for many-electron systems”. In: *Physical Review B* 23.10 (1981), p. 5048.
- [19] John P Perdew et al. “Relevance of the slowly varying electron gas to atoms, molecules, and solids”. In: *Physical review letters* 97.22 (2006), p. 223002.
- [20] Jose J Plata et al. “Charting the Lattice Thermal Conductivities of I–III–VI₂ Chalcopyrite Semiconductors”. In: *Chemistry of Materials* (2022).
- [21] Subhash L Shindé and Jitendra Goela. *High thermal conductivity materials*. Vol. 91. Springer, 2006.
- [22] Atsushi Togo, Laurent Chaput, and Isao Tanaka. “Distributions of phonon lifetimes in Brillouin zones”. In: *Physical Review B* 91.9 (2015), p. 094306.
- [23] ML Valeri-Gil and C Rincón. “Thermal conductivity of ternary chalcopyrite compounds”. In: *Materials Letters* 17.1-2 (1993), pp. 59–62.

-
- 
-
- [24] Jianhui Yang et al. “Pressure effect of the vibrational and thermodynamic properties of chalcopyrite-type compound AgGaS₂: A first-principles investigation”. In: *Materials* 11.12 (2018), p. 2370.
- [25] Yongsheng Zhang. “First-principles Debye–Callaway approach to lattice thermal conductivity”. In: *Journal of Materiomics* 2.3 (2016), pp. 237–247.
- [26] Qiye Zheng et al. “Advances in thermal conductivity for energy applications: a review”. In: *Progress in Energy* 3.1 (2021), p. 012002.



A. Results of High-Throughput Screening

Material	Type	κ_{cal} (W/mK)	Material	Type	κ_{cal} (W/mK)
AgAlS ₂	chalcopyrite	0.661	K ₂ O	fluorite	2.325
AgGaS ₂	chalcopyrite	0.794	NaI	rock salt	2.370
Rb ₂ Te	fluorite	0.832	CsCl	rock salt	2.566
AgAlSe ₂	chalcopyrite	0.995	K ₂ S	fluorite	2.758
AgInS ₂	chalcopyrite	0.998	BaO	rock salt	2.780
AgGaSe ₂	chalcopyrite	1.085	KBr	rock salt	2.889
AgInSe ₂	chalcopyrite	1.154	RbCl	rock salt	2.917
Rb ₂ O	fluorite	1.165	RbBr	rock salt	2.938
AgAlTe ₂	chalcopyrite	1.283	BaCl ₂	fluorite	3.060
InLiTe ₂	chalcopyrite	1.563	SrCl ₂	fluorite	3.100
K ₂ Te	fluorite	1.57	AlCuS ₂	fluorite	3.107
AgGaTe ₂	chalcopyrite	1.609	AgI	zinc blende	3.462
SnS	zinc blende	1.622	AlCuSe ₂	chalcopyrite	3.711
Rb ₂ S	fluorite	1.632	Na ₂ Te	fluorite	3.735
K ₂ Se	fluorite	1.726	CdF ₂	fluorite	3.818
AlLiTe ₂	chalcopyrite	1.746	NaBr	rock salt	3.866
LiI	rock salt	1.844	Na ₂ Se	fluorite	4.073
GaLiTe ₂	chalcopyrite	1.877	InLiSe ₂	chalcopyrite	4.154
RbI	rock salt	1.906	LiGeIn	half-heusler	4.856
CsF	rock salt	2.005	CdAs ₂ Ge	chalcopyrite	4.951
CsBr	rock salt	2.041	CdSnAs ₂	chalcopyrite	5.015
KI	rock salt	2.143	LiCl	rock salt	5.231
AgInTe ₂	chalcopyrite	2.160	CuInS ₂	chalcopyrite	5.345
CsI	rock salt	2.176	CdAs ₂ Si	chalcopyrite	3.449
LiBr	rock salt	2.246	CuInSe ₂	chalcopyrite	5.540
CsH	rock salt	2.298	MgAs ₂ Ge	chalcopyrite	5.597

Table A.1.: List 1 of thermal insulators ($\kappa_{cal} < 10\text{W/mK}$)

Material	Type	κ_{cal} (W/mK)	Material	Type	κ_{cal} (W/mK)
CdP ₂ Si	chalcopyrite	5.817	CuGaSe ₂	chalcopyrite	5.848
CuInTe ₂	chalcopyrite	6.152	BaS	rock salt	6.179
BaF ₂	fluorite	6.233	CuBSe ₂	chalcopyrite	6.535
LiAsMg	half-heusler	6.781	MgSe	rock salt	7.017
CuBS ₂	chalcopyrite	7.231	CaTe	rock salt	7.268
Li ₂ Te	fluorite	7.375	Mg ₂ Ge	fluorite	7.388
KH	rock salt	7.751	KCl	rock salt	7.883
AlCuTe ₂	chalcopyrite	7.975	CuN ₂ P	chalcopyrite	8.011
CdGeP ₂	chalcopyrite	8.065	LiAsZn	half-heusler	8.097
BaTe	rock salt	8.100	SnSe	rock salt	8.107
ZnPLi	half-heusler	8.219	CuI	zinc blende	8.224
Na ₂ S	fluorite	8.725	Mg ₂ Si	fluorite	8.771
MgAs ₂ Si	chalcopyrite	8.880	Na ₂ O	fluorite	8.911
KF	rock salt	9.069	ZnO	rock salt	9.202
SrF ₂	fluorite	9.789			

Table A.2.: List 2 of thermal insulators ($\kappa_{cal} < 10\text{W/mK}$)

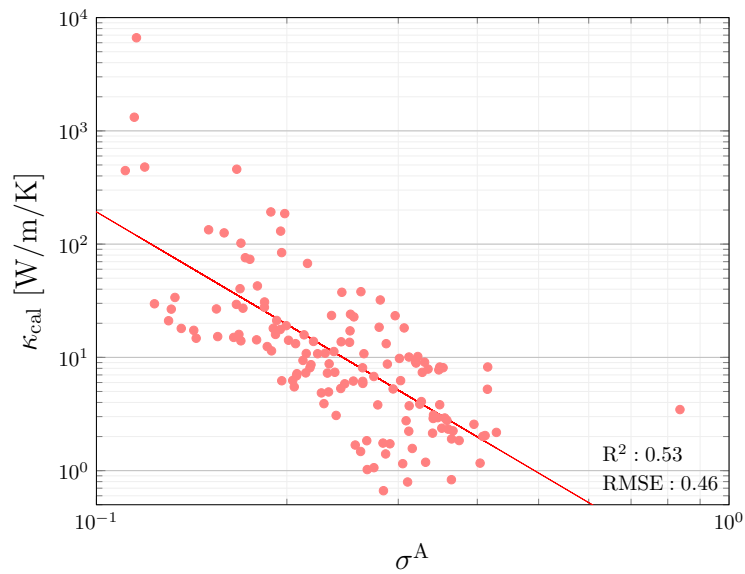


Figure A.1.: Calculated thermal conductivity (141 materials) vs. σ^A of them. Compared to Fig. 3.8(a) ($R^2 = 0.68$ and $\text{RMSE} = 0.41$) the correlation slightly gets worse. σ_{os} is calculated in the same way that describes the error between true and harmonic forces, which is what we define as the degree of anharmonicity, the only difference is that now we are using κ_{cal} calculated from third order force constants.

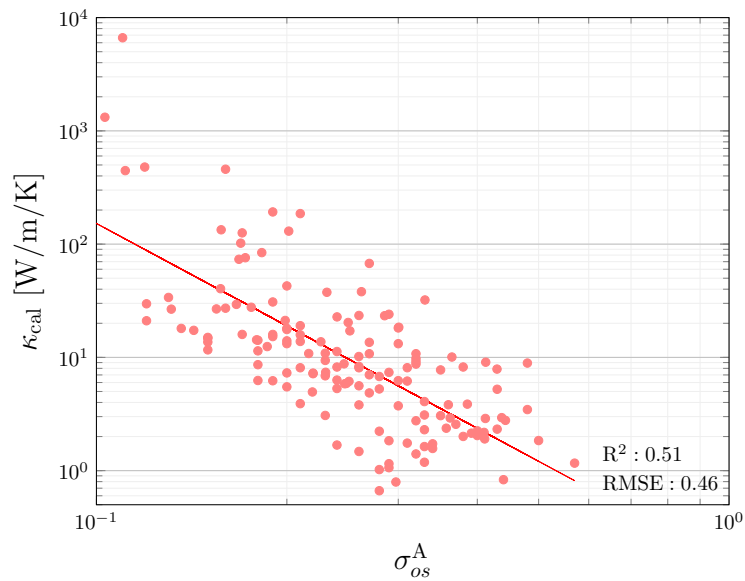


Figure A.2.: Calculated thermal conductivity of all the 152 materials vs. σ_{os}^A of them. Compared to Fig. 3.8(b) ($R^2 = 0.64$ and $RMSE = 0.43$) the correlation slightly gets worse. σ_{os} is calculated in the same way that describes the error between true and harmonic forces, which is what we define as the degree of anharmonicity, the only difference is that now we are using κ_{cal} calculated from third order force constants.

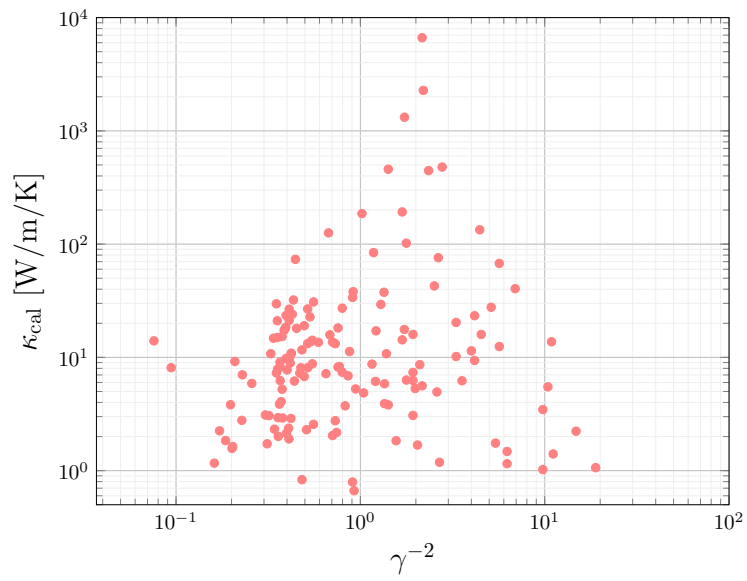
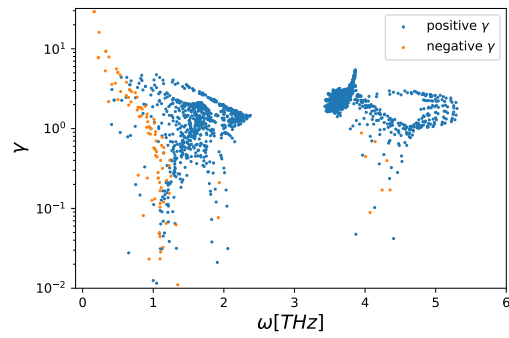
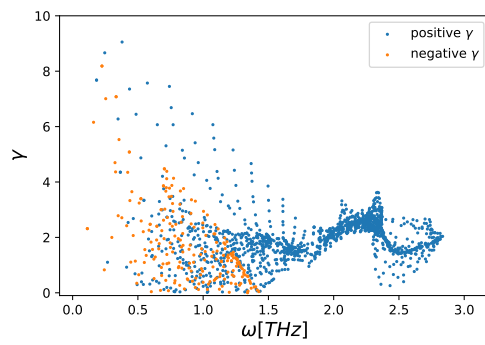


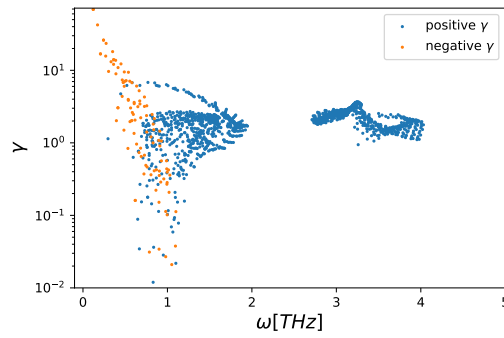
Figure A.3.: Calculated thermal conductivity of all the 152 materials vs. γ^{-2} of them



(a) NaI



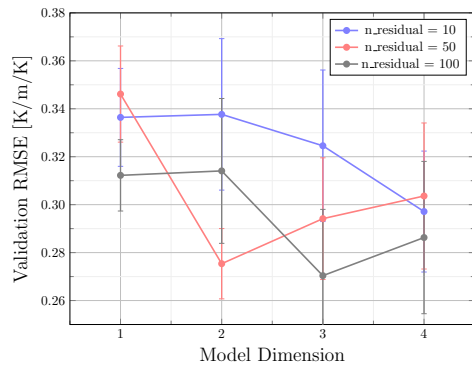
(b) RbI



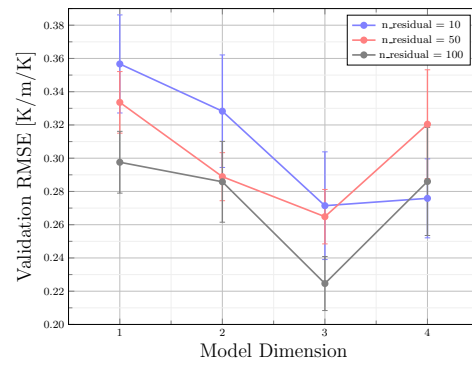
(c) KI

Figure A.4.: Mode Grüneisen parameters of NaI, RbI and KI

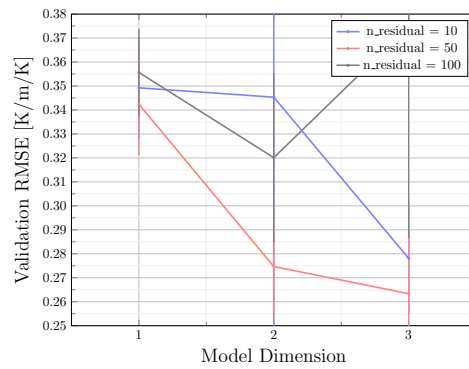
B. Cross Validation Results of SISSO



(a) $n_sis_select = 100$



(b) $n_sis_select = 500$



(c) $n_sis_select = 1000$

Figure B.1.: Three figures correspond to CV results of different n_sis_select as shown in the subcaptions. For each n_sis_select , three $n_residual$ are used, i.e., 10, 50, 100.

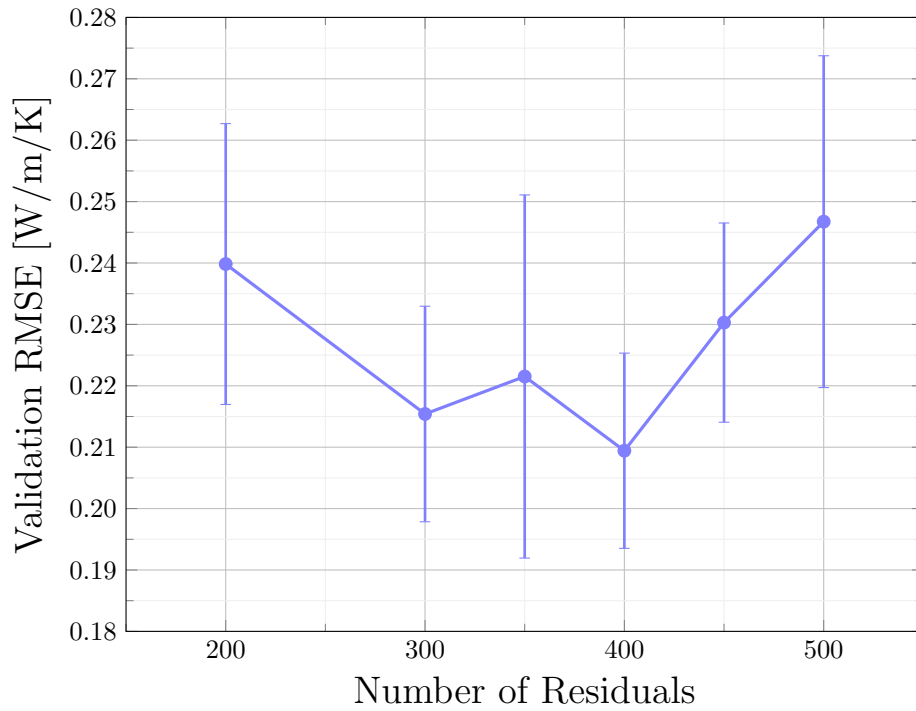


Figure B.2.: Validation RMSE of 50 CVs for the regression in Sec. 3.5.1. Other parameters are: `max_rung = 2`, `desc_dim = 3` and `n_sis_select = 500`. For 50 cross validation, regression using 400 residuals yields the smallest RMSE, but the error bar of it goes to somewhere larger than the average RMSE of 300 and 350 `n_residual`

Dimension	Models	Test RMSE
1	$0.3023 + 7.9631 \cdot 10^{-2} \cdot \left(\frac{\omega_{\Gamma_{max}}}{n_{atoms_prim}} \cdot \sqrt[3]{m_{min}} \right)$	0.0821
2	$-1.8640 - 7.4638 \cdot \frac{V_m/\mu}{\sqrt{v_s}} + 7.5694 \cdot 10^{-2} \cdot (\sqrt{\Theta_p} \cdot \sqrt[3]{V_a})$	0.0877
3	$0.2965 - 1.0857 \cdot 10^2 \cdot \frac{\rho \cdot m_{avg}}{v_s \cdot \mu}$ $+ 3.5644e-09 \cdot ((m_{avg} \cdot \Theta_{D\infty}) \cdot (V_a \cdot \Theta_{D\infty}))$ $+ 1.4598e-03 \cdot \left(\sqrt[3]{m_{min}} \cdot \frac{\Theta_p}{n_{atoms_prim}} \right)$	0.0513

Table B.1.: Models with smallest test RMSE of each dimension (n_sis = 500, n_residual = 400)

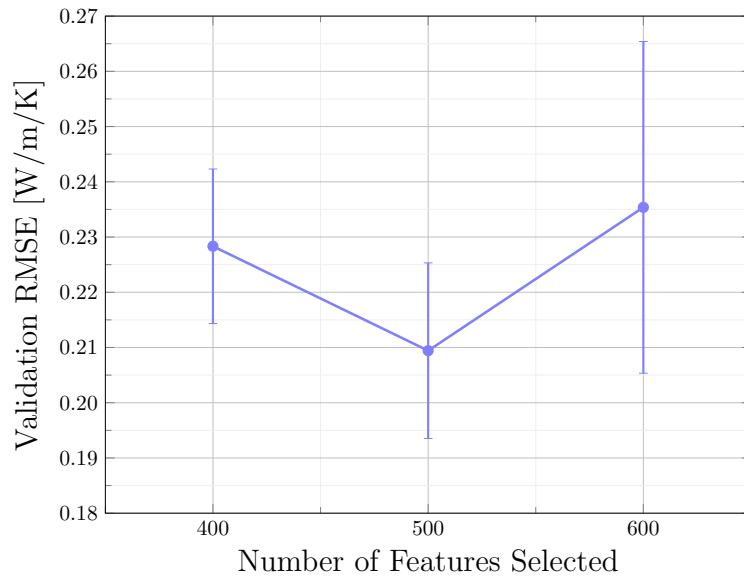


Figure B.3.: Cross validations for `n_sis_select` = 400, 500, 600, with `max_rung` = 2, `n_residual` = 400 and `desc_dim` = 3 as optimized earlier. Results show that the one with 500 features selected gives the smallest RMSE.

Dimension	Models	Test RMSE
1	$0.2956 + 0.0807 \cdot \sqrt[3]{m_{min}} \cdot \frac{\omega_{\Gamma max}}{n_{atoms_prim}}$	0.1251
2	$-0.7913 - 8.5600e-6 \cdot ((\rho \cdot \gamma) \cdot (V_m \cdot \Theta_a)) - 2.4722 \cdot \frac{\ln(\sigma)}{\sqrt[3]{L_{avg_prim}}}$	0.0674
3	$-0.2686 + 3.7887e-05 \cdot \frac{V_a}{\sigma_{os}^2}$ $+ 3.7887e-5 \cdot (\frac{m_{avg}}{\Theta_a} - 2.1196 \frac{m_{max}}{\Theta_p}) + 1.9414e-03 \cdot \sqrt[3]{m_{min}} \cdot \frac{\Theta_p}{n_{atoms_prim}}$	0.1000
4	$1.3778 - 7.8360e-06 \cdot ((\Theta_a - \Theta_{D\infty}) \cdot (m_{max} \cdot L_{avg_prim}))$ $- 5.1820e+02 \cdot \frac{m_{max}/V_a}{V_a \cdot \Theta_a}$ $- 0.9481 \cdot (\sqrt[3]{C_v} \cdot \sigma) + 1.5690e-04 \cdot (\sqrt[3]{m_{min}} \cdot \frac{v_s}{n_{atoms_prim}})$	0.0979
1	$0.2979 + 7.9679e-02 \cdot (\sqrt[3]{m_{min}} * \frac{\omega_{\Gamma max}}{n_{atoms_prim}})$	0.1508
2	$-0.7863 - 2.4636 \cdot ((\rho \cdot \gamma)(V_m \cdot \Theta_a)) - 8.4830e-06 \cdot (\frac{\ln(\sigma)}{\sqrt[3]{L_{avg_prim}}})$	0.0871
3	$-0.3054 + 14.4497 \cdot \frac{n_{atoms_prim} \cdot \rho}{m_{min} \cdot \Theta_{D\infty}}$ $+ 1.3506e-02 \cdot \frac{\Theta_a/n_{atoms_prim}}{\sqrt{\omega_{\Gamma max}}} - 3.3912e+02 \cdot \frac{\omega_{\Gamma max}/\Theta_a}{\sigma_{os} + \sigma}$	0.0764
4	$-0.2919 + 6.3967e-06 \cdot \frac{V_a \cdot \Theta_{D\infty}}{\Theta_a \cdot \sigma_{os}}$ $+ 6.3107 \cdot (\frac{m_{avg}}{\Theta_a} - \frac{m_{max}}{\Theta_p}) - 1.8714 \cdot \frac{\Theta_a/\Theta_{D\infty}}{n_{atoms_prim} \cdot L_{avg_prim}}$ $+ 4.1979e-03 \cdot \frac{v_s \cdot V_a}{L_{avg_prim} \cdot \sigma_{os}}$	0.0667
1	$0.2996 + 8.0321e-02 \cdot \sqrt[3]{m_{min}} \cdot \frac{\omega_{\Gamma max}}{n_{atoms_prim}}$	0.1839
2	$-0.8213 - 2.4931 \cdot ((\rho \cdot \gamma) \cdot (V_m \cdot \Theta_a)) - 8.4887e-06 \cdot \frac{\ln(\sigma)}{\sqrt[3]{L_{avg_prim}}}$	0.0660
3	$0.9199 + 3.4929e-02 \cdot (\frac{m_{avg}}{\Theta_a} - \frac{m_{max}}{\Theta_p})$ $- 1.6305 \cdot \frac{\omega_{\Gamma max} \cdot L_{avg_prim}}{\sqrt{n_{atoms_prim}}}$	0.0769
4	$0.8159 + 5.1109e-05 \cdot \frac{C_v \cdot \sigma_{os}}{\sqrt{\mu}}$ $+ 50.9457 \cdot (\frac{m_{avg}}{\Theta_a} - \frac{m_{max}}{\Theta_p}) - 1.7517 * \frac{m_{min}/\rho}{L_{avg_prim}^6}$ $- 5.4608e-02 \cdot ((\Theta_{D\infty} - \Theta_p) \cdot \frac{V_a}{\sigma_{os}})$	0.0732

Table B.2.: Models with smallest test RMSE corresponding to the regression in Sec. 3.5.2. From top to bottom are $n_{residual} = 100, 400$ and 500 .