Machine Learning and Data Mining for Semiconductor Materials

MSE 395 Final Report

Group 26b.
Ruoyao Zhang, Yifan Yao, Matthew Truskowski, Jue Zhang, Jaerin Kim
May 9, 2017

Executive Summary

Overview

This experiment was designed to use machine learning and data mining tools to predict dielectric constants. These predictions can then be correlated to materials with known semiconductor behavior in order to find new semiconductor materials. Basic material properties were data mined from Materials Project and Mendeleev Database and separately computed higher order material parameters such as the Coulomb matrix were employed to describe the mined materials as well. An extensive literature review was employed to teach the machine learning model of the relationship between these material parameters and the dielectric constant.

Methodology

The first step of the process was to collect enough data points on historical experiments determining the dielectric constant for various materials. At least 50-100 materials is necessary for a machine learning model to run at all, and many more are required for it to accurately learn a complex relationship such as that between dielectric constant and the physical/atomic properties of a molecular compound. After that data was collected, Materials Project was queried to form descriptors for the materials included in that training set, and the machine learning model was trained. Finally, the trained model was given descriptors for every material on the Materials Project database and allowed to make a prediction. These predictions were organized according to the Frohlich model and can now and in the future be confirmed with density functional theory (DFT).

Findings and Implications

The machine learning model predicted that the materials with the best semiconductor properties including reduced electron-phonon coupling strength according to the Frohlich model (α~0.02-4) were mainly selenides and hydrides. Selenides are historically an important semiconductor, so the model is not entirely inaccurate. However, when broken down analytically in comparison to experimental values, there is still an error as high as 105% for the electron-phonon coupling strength calculations, leaving much room for improvement if more data points are used in the future to train the machine learning model.

1. Introduction

1.1 Objective

There is a constant search for new semiconductor materials with more efficient charge generation capabilities. Material properties such as the dielectric constant can be highly correlated with better semiconductor performance. Some materials likely have desirable semiconductor properties but have not yet been discovered or utilized. However, it is impossible to synthesize and test all potential materials. Such a hands-on approach can be time and resource intensive. Machine learning is a computational tool that can be used to analyze known experimental data for certain materials to train a relationship model. The machine learning model can then predict the target property (dielectric constant) for other candidate materials.

1.2 Mission statement

This project is focusing on gathering reference data for high and low frequency dielectric constants of different materials from the literature and other theory calculations, creating a well-fitted machine learning model based on the gathered data, and predicting the dielectric constants for many unknown materials. Based of the scope of this project, there are also several difficulties present so that we have to deal with. First, the gathered data for different materials are coming from different references, which means they have different experimental conditions such as temperature, pressure and frequency, so we need to set a rule to what sources should be used and how to manage those inconsistent data properly. Secondly, there are many models in data mining and machine learning fields, like the linear fitting, random forest, neural network and so on. We therefore need to test all the possible models for the data frame and find the one with the smallest error on the testing set. Thirdly, after the model has been trained, it needs to be evaluated by comparing with literature values and the DFT calculation so that the model can be used with confidence if the candidate materials perform well during the evaluation process. Last but not least, possible sources of error have to be analyzed so that improvements on such training models will be proposed to increase the accuracy of prediction.

2. Literature review

2.1 Electron-phonon coupling (ECP) and Frohlich model

The electron-phonon interaction is, besides the Coulomb interaction, one of the fundamental interactions of quasiparticles in solids. It plays an important role for a variety of physical phenomena. In particular in metals, low-energy electronic excitations are strongly modified by the coupling to lattice vibrations, which influences, e.g., their transport and thermodynamic properties. Electron-phonon coupling (EPC) also provides in a fundamental way an attractive electron-electron interaction, which is always present and, in many metals, is the origin of the electron pairing underlying the macroscopic quantum phenomenon of superconductivity. (Pavarini et al, 7)

When studying the interactions between electrons and atoms in a solid material, the concept of polaron as a quasiparticle is widely used in condensed matter physics. The concept is introduced to describe an electron moving in a dielectric crystal where the atoms move from their equilibrium positions to effectively screen the charge of an electron, known as a phonon cloud ("Polaron"). Frohlich proposed a model Hamiltonian for the "large" polaron through which its dynamics is treated quantum mechanically ,"Frohlich Hamiltonian", thus closely related to EPC in the material. The polarization, carried by the longitudinal optical (LO) phonons, is represented by a set of quantum oscillators with frequency ω_{LO}, the long-wavelength LO-phonon frequency, and the interaction between the charge and the polarization field is linear in the field: (Devreese)

Equ.1

where r is the position coordinate operator of the electron with band mass mb, p is its canonically conjugate momentum operator; $a_k^\dagger$ and a_k are the creation and annihilation operators for longitudinal optical phonons of wave vector k and energy $\hbar\omega_{LO}$. The V_k are Fourier components of the electron-phonon interaction:

Equ.2

The strength of the electron–phonon interaction is expressed by a dimensionless coupling constant α, which is defined as:

Equ.3

In this definition, ε_∞ and ε_0 are, respectively, the high-frequency and the low-frequency (static) dielectric constant of the polar crystal. We will refer α as the coupling strength or coupling coefficient. It is an essential parameter that can characterize the electron-phonon interaction using materials properties that can be experimentally obtained or calculated. In this project, our group focuses on the high- and low-frequency dielectric constants (ε_∞ and ε_0) while the group 26B focuses on optical phonon frequency (ω_{LO}).

2.2 Dielectric constant

The dielectric constant, ε, can be described as the ratio of the electric permeability material to the permittivity of free space (i.e., vacuum). More specifically, the dielectric constant describes the amount of charge needed to generate one unit of electric flux in a particular medium. Simply put, the dielectric constant measures the ability to store electrical energy under a certain E-field. This occurs through the aligning of the dipole moments. Inside a material, the molecules that compose our dielectric have dipole moments that are randomly oriented. In an electric field, the dipole moments will orient themselves according to our electric field and how much work the electric field must do to orient the dipole moments is proportional to the dielectric constant. So materials with high dielectric constants would be store higher electrical energy and thus have a higher capacitance.

The dielectric constant can be separated into two types of dielectric values, low frequency and high frequency dielectric constant values. Dielectric constant values are frequency dependent due to the fact that all the electric dipoles created by the E-field have various oscillation frequencies. The purpose for the high frequency and low frequency dielectric constant values was so that we could calculate the α (electron-phonon coupling strength constant) using equation 3, which uses low frequency and high frequency dielectric constant values.

The dielectric constant is an important parameter in electronic materials and semiconductors. For example, materials with a high dielectric constant value can be used to reduce the size of a capacitor whilst maintaining the same capacitance. In semiconductors, dielectric materials can be used to insulate transistor from each other. Experimentally calculating dielectric constant values requires time, capital, and perhaps synthesis of the material itself. So being able to identify high dielectric constant values using machine learning to circumvent the obstacles mentioned previously was a source of motivation for us.

2.3 Machine learning

There are many methods (models) to train a machine learning process. Commonly used models include linear regression and its variation (Lasso & Ridge), neural network, and random forest regressor, all of which we explored in the project. The Scikit Learn machine learning module for Python was chosen to train a model and use it to predict dielectric constants. The model generally takes two inputs to train: a set of known target values (dielectric constants from our literature search) and a complementary set of measured descriptor properties for each target value (material properties queried from Materials Project). In order to define these descriptor properties for a large dataset, it is necessary to write code that automates the process of pulling each property from the database for each material.

2.4 Data mining

The purpose of mining for data on dielectric constant values was so that we could be able to determine dielectric constant values for materials using machine learning. This is important, as mentioned before, because of the time-saving and money-saving incentives that come with calculating the dielectric constant values using a model. But before we can use a model, we have to first be able to have data to be able to test and modify the model. Data on the values of the dielectric constant (high frequency and low frequency) was collected from reputable sources such as Springer Materials, Research Gate, and other published literature experiments.

3. Design generation:

3.1 Data collection

When discussing the collection of dielectric constant values, several parameters that relate to the dielectric constant were also taken into account so that we could get the most optimal model to calculate for the α value. The frequency at which the experiment took place in, the temperature at which it took place in was all taken into account. To further supplement the dielectric constant values, for each material, we also linked the Materials ID from the Materials Project database, which included values of the band gap, lattice parameters, and structure, which were all used so that we could optimize our model.

One thing to take into account was that dielectric constant values have different values depending on the direction at which the dielectric constant was measured. This is because of the difference in the structure of the material in different direction which can affect the dielectric constant. Taking into account all the different dielectric constants at different directions for our model would have protracted the progress for our project. Thus, we averaged the dielectric constant values at different directions and used the averaged value as our dielectric constant (for high and low frequency values). If we ran the model, taking into account the different dielectric constant values at their respective direction axis, we might have been able to reduce our error of final results.

Another thing to note was that it was important to keep track of whether the dielectric constant value was high or low frequency and whether it was the actual dielectric constant value for the material. In order to do this, the dielectric constant values for materials were cross-referenced with dielectric constant values from different literature texts for both high and low frequency values. Approximately 200 data points of high and low frequency dielectric constant values were used for the machine learning model.

3.2 Python packages

Many Python packages are required in order to successful execute the machine learning and data mining process. In addition to basic computational packages such as matplotlib and numpy, the pandas library for .csv file handling and the mendeleev package for elemental data were used crucially in the code for descriptor generation. Materials Project also has a Python package for reading data from its website through a generated API key. Finally the Scikit Learn package allows simple and easy loading and execution of many different machine learning algorithms.

3.3 Descriptors

The key role of the descriptors is to find the actuating mechanisms of a certain property or function and describing it in terms of a set of physically meaningful parameters. In this project the descriptors are atomic and material properties such as bandgap, lattice parameter, density, etc., which are capable of accurate prediction of properties for virtually any stoichiometric inorganic crystalline material. (Isayev et al, 15279) The generalization of the discrete data set {P_i, d_i} to a continuous function P(d) has been traditionally achieved in terms of physical models, or mathematical fits.

$$P(d) = dc$$ Equ.4

where c is the Ω-dimensional (Ω=1, 2, ...) vector of coefficients. The coefficient c is determined by minimizing the loss function

$$\| P - Dc \|_2^2$$ Equ.5

where D is a matrix with each of the N rows being the descriptor d_i for each training data point. The sum of square of the differences between target value P_i and the estimated value f(d_i) can be expressed as

$$S = \sum_{i=1}^{n} (P_i - f(d_i))^2 = (P_i - d_i c)^2$$ Equ.6

This idea is the cornerstone of this project and the later machine learning algorithm heavily relies on the construction of the descriptors of materials for both training dataset and prediction target. The descriptors were generated by querying the Materials Project database and Mendeleev python package, and the training set high and low frequency dielectric constants were compiled from Springer Materials and other published literature experiments.

3.4 Coulomb matrix

If the descriptors mentioned above are the unprocessed properties that are directly related to interested materials, we can consider the coulomb matrix as a descriptor that contains the coulomb interactions with each compound. A three-dimensional molecular structure is converted to a numerical matrix using atomic coordinates R_i and nuclear charges Z_i. The matrix contains one row per atom, is symmetric, and requires no explicit bond information. The specific entry of the coulomb matrix is defined as (Hansen, 3404)

Equ.7

To characterize the constructed coulomb matrix, the eigenspectrum of the matrix is obtained and sorted as shown in the figure below. We can incorporate the sorted eigenspectrum along with the coulomb matrix to make the set of descriptors for interested materials.

Figure 1: Three different permutationally invariant representations of a molecule derived from its Coulomb matrix C: (a) eigenspectrum of the Coulomb matrix, (b) sorted Coulomb matrix, (c) set of randomly sorted Coulomb matrices.

3.5 Machine learning models

The purpose of this project is to predict unknown properties based on other known properties. However, it is currently unknown how the dielectric constant can be physically related to its atomic or chemical properties such as those in descriptors and its structure information such as the coulomb matrix. For the time frame of this project, three learning methods including linear regression and its variants, artificial neural network, and random forest regressor were explored, in aim to model the relationship between the dielectric constant and physical properties included in the descriptors at the high frequency mode and static mode. A schematics of the three methods are shown below.

3.5.1 Linear Regression

The linear regression model is approach to model the linear relationship between a scalar output and one or more independent variables. The equation generally used in this method follows:

$$Y = X\beta + \varepsilon \qquad \text{Equ.8}$$

in which y is the transpose of the scalar output such as y1,y2,y3,....ym, whereas x is a row of independent variables x1, x2,x3,...., xn. β ,therefore, is the transformative matrix with n columns and m rows. ε is the linear shift of the value between the output and input. Such a method utilizes the 'least sum of square of error' approach, which means it will find the optimal β and ε between independent variables and output with the smallest sum of square of error between the predicted data and real data, represented in the following graph. For the simplicity of this model, such a model is often used in a broad range of applications such as economics (Ehrenberg) but can be fundamentally limited to its inability to model relationships other than linearity. The limitations of the linear regression also include its sensitivity to the outliners. To improve on the accuracy, the model will inevitably have to be introduced the penalty function such as Lasso (L2 norm penalty) or Ridge (L1 norm penalty) so that the with each new input, β will be less likely to be affected. But a tradeoff is that a huge number of training set is required while the deficiency of the data can lead to the failure of convergence of such model.

Figure 2: Representation of a linear regression model

3.5.2 Artificial Neural Network

The Artificial neural network is a model that is structurally analogous to a neural network in an animal brain, represented in the figure below. In supervised learning, a column of data input are taken in as the first layer of neurons. After the first layer, each neuron represents a real value, containing the weighted sum of functions from the value the previous layer. Each arrow connecting two neurons is typically nonlinear function such as:

Equ.9

in which a_j^{i+1} is the jth node of the i+1 layer, a_k^i is the kth node of the i layer, b_j^{i+1} is the bias of the jth node in the i+1 layer, w_k^i is the kth node in the ith layer. Supposedly any non-linear, one-to-one, continuous function can be the transfer function f, but the one commonly used is given below:

Equ.10

Some user-imposed constraints are the number of layers and the number of nodes in each layer. Arguably, such a method can replicate a very complicated system such as the brain itself with the help of multiple hidden layers, each with a large set of nodes. ("Artificial Neural Network"). The weight and bias of each node will initialized randomly and optimized iteratively through the training process. However, it will expose risks to overfitting if with too many nodes.

Figure 3: Representation of a nerual network regressor

3.5.3 Random Forest

The random forest method is a tree-decision based approach but is applied to regression instead of classification. A representation of a tree decision is shown in the following diagram. X is as before the input set of independent parameters. Each red dot is a decision, based on which data are splitted. The decision process is optimized along the way learning proceeds. Splitted data are then processed into the next layer along the black arrow in the representation. If, by transferring, the data reach the bottom layer or belong to the target value set, it will become a leaf node Y and will no longer split. The whole learning process will continue as long as there are still data that have not reached the leaf node. The learning, however by itself, only includes the optimization of the splitting rule by minimization the error. Scikit-learn random forest algorithm utilizes the minimization of mean squared error. Some user-imposed constraints include the maximum number of nodes involved or the maximum number of layers to prevent the overcomplexity of model by overfitting the training set.

Figure 4: Representation of a random forest regressor

3.6 Density functional theory

Density functional theory is a theory for the ground state energy of a system as a function of the electron density instead of the wave function (Kohn, A1133). The theory firstly uses the pseudopotential of the atom by removing the tightly bound core electrons because those inner electrons do not involve in bonding. The orbitals are represented with a basis of plane waves or Gaussians. After constructing the original orbital, the algorithm iterates charge density and minimize the energy to self-consistency. As a trial run, the group choose silicon to verify the functionality of using DFT to calculate the dielectric constant. The convergence of k-point mesh and cut-off energy needs to be achieved to minimize the running time without the compromise of accuracy. As shown in the figures below, the minimized system energy converges with the increasing number of k-point mesh and cut-off energy. Essentially, we want to optimize the band structure and the number of plane waves in the basis set.

Figure 5: The plot of k-point convergence for silicon

Figure 6: The plot of cutoff energy convergence for silicon

Using the converged k-point mesh of (5 5 5 0 0 0) and E_{cutwfc} of 30 eV, we ran DFT calculation with PW package and obtained the high-frequency dielectric constant of silicon as 11.77, which is fairly close to the experimental value 11.97 at 300 K from literature. Therefore, we have exemplified the feasibility to run DFT to get theoretical values, which enables us to evaluate the performance of machine learning prediction.

3.7 Cross validation

Cross validation is evaluation method that used to evaluate the behavior of certain model, which is better than the residuals. And the reason why it behaves better than the residual evaluation is the residual evaluation does not give a prediction of how well the learner can do when it makes predictions for data that has not been seen while the cross validation can do that.

The basic idea of cross validation is that when the model is trying to train the data set, part of the data has been removed. And after the learning process is done, the removed data can be used to test the performance of the model. In other word, the data set is separated into two sets, called the training set and testing set. The advantage of this method is that it is usually preferable to the residual method and takes no longer to compute. However, this output may have large variance since the evaluation method would depend heavily on the data points that choose to be in the training set and the points in the test set. Therefore, the evaluation dependent significantly on the way the data is separated.

There is another way that can be used to improve the basic cross validation method, which is called K-fold cross validation. The data set is divided into K subsets, and all of the K subsets have been used as the test set in K times. And in each time, the other K-1 subsets are formed together to be the training set. The advantage of this method is that it dependent less on the division method and all the data contribute to the evaluation process. And the shortcoming about this method is the training algorithm has to be rerun for K times, which is kind of time consuming.

3.8 Design constraints

There are several constraints for this project, like the data collection and model convergence. First, all the literature data of the dielectric constant obtained from the resource for different compounds are under different situations, like the frequency, the pressure and the temperature. Second, there are so many machine learning models in the database, like the linear regression, random forest and so on. In addition, the dielectric constant has a complex relationship with other parameters like the band gap, number of valence electrons and other conditions. Choosing a simple model to represent this relationship is not accurate enough with so many varying dependencies. Moreover, the amount of data points obtained from the literature for the high and low frequency dielectric constant was only about 200, which is far too little to accurately predict the dielectric constant for other 60,000 other compounds. In an ideal machine learning experiment there would be the time and tools to collect thousands of data points for the model training set alone.

4. Results

4.1 Descriptor selection

A reliable prediction based on machine learning models is usually dependent on the abundance of training sets and the completeness yet independence of the descriptors. If the descriptors categories are linearly dependent, the overall effectiveness of the model will become futile, yet too many descriptors can cause the overfitting of the model. Therefore, it is necessary to do a pre-selection of descriptors. Descriptors can be selected by evaluating the strength of the monotonic relationship with the output. A good way to assess the correlation between the two is the rank-order difference of the output and input. Two examples of formation energy versus dielectric constant at high frequencies and bulk density versus dielectric constant at high frequencies in the cubic-structured materials are shown below.

Figure 7: Rank-order difference comparison between the dielectric constant at high frequencies for the cubic structured materials and formation energy (left) and bulk density (right). Left figure shows a low correlation while right shows a relatively high correlation.

To compare the correlation across the descriptors, it is more clear to have a numerical value to represent the strength of the correlation, expressed as the individual rank-order Spearman correlation coefficient (Spearman, 72):

Equ.11

in which d_i is the difference between the rank of the ith output and the rank of the ith descriptor s and n is the overall number of output. Values of Spearman Coefficient range from -1 to +1. A positive Spearman Coefficient means the two positively correlated statistically and vice versa. A greater absolute value of the coefficient corresponds to a stronger correlation: it is normally accepted that below 0.25, the two sets of values are not significantly correlated; between 0.25 and 0.5, the two sets of values are moderately correlated; a value above 0.5 shows two sets are significantly correlated. The Spearman Coefficient of various elemental, compound and lattice properties is shown in the following diagram with the dashed lines representing the level of significance of correlation. As predicted, the Spearman Coefficient of the formation energy is only -0.21 while for bulk density, it goes up to 0.68. It is also worth noting that such a list is not comprehensive since the spearman coefficient of some elements of descriptors such as the entire coulomb matrix and space groups cannot be calculated and for each category of descriptors related to the elemental properties such as the atomic mass, all of the average mass, reduced mass, standard deviation of masses, and the difference between the maximum and minimum values are taken into consideration when constructing the descriptors whereas only the average value is shown. Features with strong correlation include but not limit to bulk density, averaged atomic mass, the largest values of the eigenspectrum of the coulomb matrix, the averaged lattice parameter, the bandgap of the compound whereas the formation energy and the averaged total number of valence electrons have very weak correlation with the dielectric constant.

Figure 8: Spearman coefficient of elemental, compound, and lattice properties with the infinity dielectric constant, the dashed line representing the level of significance.

4.2 Machine learning model selection

As stated in the design generation section, we have explored and used three models including the linear regression, the artificial neural network, the random forest regressor to test the effectiveness of the machine learning methods on predicting the dielectric constants at the high frequency and the static dielectric constant. For the present study, since the number of valence electrons are given by each orbital s, p, d, and f, it is hard to compare light elements with d and f values being 0, with heavy elements with filled s and p inner orbitals. Therefore, to improve the correlation of the descriptors, only the formation energy is excluded from descriptors. For each learning model, we have implemented the 80--10--10 (fitting--cross-validation--testing) method. Below shows the result of (A) the linear regression model with Ridge penalty function when alpha = 0.3 applied to the dielectric constant at the high frequency of over 120 materials.

Figure 9: Literature values versus predicted values (training & testing) using (a) linear regression model with Ridge penalty function when alpha = 0.3. The dotted line represents a perfectly reliable model wherein predicted values exactly match real-life measurements.

From the above figure, it is apparent the linear regression model is more or less suitable for the training set with numerical values close to each other, shown by the clustering of the red dots of the training set from 0 to 30. However, it fails to predict the output when the literature value is over 60. To reduce its sensitivity to outliers, we have used the Ridge penalty function. The function will compensate for the deviation coming from the noise by a value called ɑ that will keep the sum of the size of the coefficient minimized. As shown in the literature, the sum of the size of the linear coefficients represents the complexity of the model. (Jain) If the size of ɑ is set, then the max complexity is set as well to eliminate the overfitting of the model on the outliers. For the current training set, we have partitioned ɑ from 0.001 to 1000 and found that 0.3 tends to have the best performance, as shown in the above figure.

Below shows the result of (A) the neural network model when the preset maximum number of layers is limited to 200 and alpha equals 0.001 (default in Scikit Learn), applied to the dielectric constant at the high frequency of over 120 materials. From the figure, it is also shown the model fails to capture the relationship even at the low dielectric constants range (smaller than 20), the predicted values deviates at a large magnitude even when the literature value centers around 5, which is clearly caused by the initial penalty function being two weak yet a penalty value smaller than 0.001 will inevitably render the model unable to converge by such a small amount of data compared to the size of neurons. A future work that can be improved is to find a much larger set of training set so that even without penalty (alpha=0), the model can still converge and trained as a suitable model.

Figure 10: Literature values versus predicted values (training & testing) using (b) neural network model when the preset maximum number of layers equal to 200, alpha = 0.001. The dotted line represents a perfectly reliable model wherein predicted values exactly match real-life measurements.

The last model to explore is the random forest regressor. Below shows the result of (C) the random forest regressor with the node size = 15, applied to the dielectric constant at the high frequency of over 120 materials.

Figure 11: Literature values versus predicted values (training & testing) using (c) random forest regression for the 1st iteration. The dotted line represents a perfectly reliable model wherein predicted values exactly match real-life measurements.

From the above figure, it is shown that the model can well predict the value of dielectric constants after training, for the training set. Such a close match between the predicted value and real value for the training set can extend to values much higher than neural network and linear regression model do. Even at the maximum of value of 113, the predicted value is much closer than the real value compared to the other two models. The node size was optimized by partitioning from 10 to 20 in which 15 generates the smallest error between the predicted values and real values of the training set. However, since the first node is initialized randomly at each time, it is inevitably the final value can be different even with the same training set. To stabilize the error and have an accurate estimation of how the model performs, we have increased the number of iteration to reduce such uncertainty. As shown in the following figure, the average error of training and testing set random forest model will first fluctuate between 0.56 and 0.50 for the first few times but will stabilize to 0.526 after 10 to 20 iterations.

Figure 12: The average error of the training set with training set versus number of iteration for the random forest regressor with node set to 15 for the static dielectric constants. The dashed line is a spined curve to fit the trend.

However, the real performance of the modeling can not be estimated on the training value, but on the set of data that is never touched by the algorithm called the testing data. The average error on the testing data versus predicted values using the three modeling methods are summarized in the following table. It is shown that the random forest is the most reliable for predicting the dielectric constant of materials with the error of dielectric constant at high frequencies being 78.0% and the error of static dielectric constant being 68.0% yet the median is around 44.5%, which means the average error is driven up by some outliers.

Table 1: Average Error of three learning models on the testing data of dielectric constants at high frequency and static dielectric constant

Learning Model	Linear Regression	Neural Network	Random Forest
Error of dielectric constant at high frequencies (%)	133.5	80.4	78.0
Error of static dielectric constant (%)	102.7	82.1	68.0

4.3 Material selection for reduced electron-phonon coupling effect

Within the Frohlich model, the strength of the electron-phonon coupling is characterized by the dimensionless coupling coefficient α as mentioned in the literature review. The predicted value for phonon frequency ω_{LO} is obtained from Group 26A while ε_{∞} and ε_0 are predicted by our group using random forest algorithm as described above. After assembly of the predicted ω_{LO}, ε_{∞} and ε_0 for each material, we obtained a distinct coupling coefficient α for each of the materials from the Materials Project database. The results are summarized in the histogram below.

Figure 13: Coupling strength of 39190 types of materials from the Material Project Database using the random forest regressor.

While the negative coupling strength coefficients from the prediction results cannot be physically interpreted, it is noteworthy that the center part of the distribution ranges from 0 to 4 and the typical values for semiconducting materials range from 0.02~4. Candidate materials generally include Se-based and H-based compound materials. Selenides are historically an important semiconductor, so this lends confidence to the basic predictive capabilities of our ML model (Yang, 125017). Some of the predicted types of compounds are also observed in the literature (Drozdov, 73) (Huang,311) for superconductor materials.

4.4 Error analysis and future work

By the propagation of error, the final of average error of the coupling strength can be as high as 105.4% given the accuracy of the optical phonon frequency ω_{LO} 45.8%. The lack of accuracy from our results for the prediction of electron phonon coupling strengths and the dielectric constants implies that there exists huge potential for us to make improvement on the current method. As mentioned in the data constraints, the collected dataset limits the accuracy due to varying experimental conditions from literature ranges from 1950s to recent years. For the experimental measurement of dielectric constants, the temperature and electric field frequency are mostly inconsistent, resulting in fluctuating values even for the same material. Consequently, the machine learning model yields high percent error based on these unreliable training data. Moving forward, we can improve the accuracy and consistency of the training data by replacing the ones with high percent error with corresponding

DFT-calculated values. Using this method, we can improve the accuracy of the training data without running DFT for all the materials, thus maintaining the cost.

Due to time constraint of the project, we have yet to verify the calculated electron phonon coupling strength using the assembled DFT results for phonon frequency, high- and low-frequency dielectric constants. To elaborate, the group should pick the materials with lowest predicted coupling strength α and run DFT to calculate theoretical values of ω_{LO}, ε_∞ and ε_0 for the material. Then, we can use Frohlich model to calculate the theoretical α for the target material. The final comparison between the theoretical value and the machine learning prediction can help us evaluate the effectiveness and performance of our overall model. However, the methodology our group has adopted exemplify the possibility of using statistical correlation to predict physical correlation in materials properties.

5. Conclusions

Training sets of approximately 200 literature values each were collected for high and low frequency dielectric constants along with a similarly sized set of optical phonon frequencies collected by Group 26a. The random forest regressor model was chosen as the most accurate for our descriptor/target relationship and dataset size. Uniting the ML predicted optical phonon frequency and dielectric constants with α returned several thousand materials in the semiconductor range, with selenides and hydrides being a chief appearance in low α predictions. Future work for this project would involve confirming these materials with DFT and reevaluating the size of the training data set as well as the sources of that data. Retraining the model with a much larger data set would almost certainly return more accurate predictions for the dielectric constant.

6. Reference

Pavarini, E; Koch, E; Scalettar, R; Martin,R The Physics of Correlated Insulator, Metals, and Superconductors Modeling and Simulation, pp.7, Forschungszentrum Julich, 2017, ISBN 978-3-95806-224-5

“Polaron” https://en.wikipedia.org/wiki/Polaron. Accessed: May 09, 2018

Devreese, J. “Frohlich Polarons. Lecture course including detailed theoretical derivations”. In: arXiv preprint arXiv:1611.06122 (2016)

Isayev, O; Oses, C; Toher, C; Gossett, E; Curtarolo,S; Tropsha, A “Universal Fragment Descriptors for Predicting Properties of Inorganic Crystals.” Nature Communications, vol. 8, pp.15679, DOI: 10.1038/ncomms1567

Hansen, K “Assessment and Validation of Machine Learning Methods for Predicting Molecular Atomization Energies”, J. Chem. Theory Comput., 2013, 9 (8), pp 3404–3419

Ehrenberg; S. Modern Labor Economics (10th international ed.). London: Addison-Wesley. (2008) ISBN: 9780321538963.

“Artificial Neural Network” Introduction to Dynamic Neural Networks – MATLAB & Simulink. www.mathworks.com. Retrieved 2017-06-15.

Kohn, W and Sham, L.J. “Self Consistent Equations Including Exchange and Correlation Effects”, Phys. Rev. 140, A1133 (1965).

Spearman, C. The Proof and Measurement of Association between Two Things. Am. J. Psychol. 15, pp. 72–101 (1904)

Jain, A, A Complete Tutorial on Ridge and Lasso Regression in Python, https://www.analyticsvidhya.com/blog/2016/01/complete-tutorial-ridge-lasso-regression-python/, Accessed May9, 2018

Yang, F; Xiong, S; Xia, Z; Liu, F; Han, C and Zhang, D, “Two-step synthesis of silver selenide semiconductor with a linear magnetoresistance effect” Semicond. Sci. Technol. 27, pp125017 (2012)

Drozdov, A; Eremets, M; Troyan, I; Ksenofontov, V; Shylin, S, “Conventional superconductivity at 203 kelvin at high pressures in the sulfur hydride system” Nature vol. 525, p. 73–76 (2015)

Huang, D and Hoffman, J.E. “Monolayer FeSe on SrTiO3.” Annual Review of Condensed Matter Physics 8 (1) (2017) pp.311–336. doi:10.1146/annurev-conmatphys031016-025242.

7. Appendix

7.1 Experimental Procedure

Codes that have been used in this project are stored in this online repository.
https://github.com/zhangruoyao68/MSE395

7.2 Hazard Analysis

Since this is a purely computational project, no lab experiments are conducted. There is no chemical hazard involved in this project.

7.2 Origin and Final Timeline

Figure 14: Original Project Timeline

Figure 15: Final Project Timeline