COMPUTATIONAL EXPLORATION OF RE-SI-O CHEMICAL SPACE USING DENSITY FUNCTIONAL THEORY AND TRUSTWORTHY MACHINE LEARNING

Mukil Venthan Ayyasamy

A Dissertation submitted to the Graduate Faculty
of the University of Virginia in Candidacy for the Degree of
Doctor of Philosophy

Department of Materials Science and Engineering

University of Virginia
December 2023

Dr. Prasanna V. Balachandran, Advisor
Dr. Bi-Cheng Zhou, Chair
Dr. Elizabeth Opila
Dr. Stephen Baek
Dr. Tianxi Li

Computational Exploration of RE-Si-O Chemical Space using Density Functional Theory and Trustworthy Machine Learning

Mukil Venthan Ayyasamy

(ABSTRACT)

The coefficient of thermal expansion (CTE) is a critical material property that quantifies the degree to which a material expands or contracts upon heating. Despite progress in theory, computations and empirical model development, existing knowledge has limitations in predicting the CTE of complex materials such as the compounds that form in the rare earth-silicon-oxygen (RE-Si-O) ternary space, which are candidate materials for environmental barrier coatings (EBCs). This thesis aims to bridge this gap by leveraging computational methods based on density functional theory (DFT) calculations and machine learning (ML) techniques. While well-trained ML models are good at generating predictions, they often serve as black boxes that limit their interpretability. This limitation is particularly problematic in materials science, where the available data sets are often limited and sparse.

To address these challenges, this dissertation targets three key goals: (1) Accelerating the design of novel compounds in the complex RE-Si-O chemical space with targeted CTE values by establishing previously unknown quantitative structure-property relationships using DFT and ML; (2) Building trust in ML models (in a narrow sense) by accurately gauging when and where they can succeed or fail, thereby facilitating well-informed design choices; and (3) Uncovering the insights behind ML model

predictions to better comprehend the science underlying the quantified structure-property relationships. This integrated approach was applied to three major material classes in the RE-Si-O chemical space: RE disilicates, RE monosilicates, and RE silicate apatites. Finally, a holistic CTE model was developed that quantitatively captures the relationship between structure and volumetric CTE across these diverse material classes, including some of the high entropy variants.

Throughout the thesis, I focused on two critical attributes of RE-Si-O systems that are essential for the development of EBCs: the DFT total energies and the polyhedral description of the crystal structures. Key contributions are summarized below:

- In *RE disilicates*, I used DFT calculations to generate the total energy difference (ΔE) data that offered key insights into the energetics favoring polymorph formation. The calculations also provide optimized crystal structures from which one can generate two types of descriptors: (1) Unit cell parameters (more accessible to the experimental community) and (2) Polyhedral descriptors (may carry mechanistic insights that can be correlated with CTE). I trained an ensemble of ML models to rapidly predict the volumetric CTE, along with the associated uncertainties. Experiments from our collaborators validated the CTE predictions for the $Sm_2Si_2O_7$ compound in $P4_1$ space group.

- In *RE monosilicates*, I focused on CTE anisotropy. Using DFT and density of states calculations, I uncovered a previously unidentified trend that correlated the *d*-orbital bandwidth and RE-O effective coordination numbers in iso-electronic Sc_2SiO_5, Y_2SiO_5, and La_2SiO_5 compounds with the measured CTE anisotropy data (taken from the literature). The crystal structures were constrained to the $C2/c$ space group in these calculations.

- In *RE silicate apatites*, I calculated ΔE from DFT to reveal the energetics trend across the different RE silicate apatites: non-stoichiometric ($RE_{9.33}Si_6O_{26}$) and RE silicate apatite bearing alkali metals ($RE_9A_1Si_6O_{26}$, where A=Li, Na, K, Rb, and Cs monovalent cations) and alkaline earth metals ($RE_8AE_2Si_6O_{26}$, where AE=Be, Mg, Ca, Sr, and Ba divalent cations). Unlike the disilicates and monosilicates, the literature data on CTE for this materials class is sparse. Therefore, I did not investigate the CTE property for this materials class. Nonetheless, my calculations led to the development of polyhedral descriptors which I hypothesize as a more meaningful representation of the crystal chemistry when compared to the traditional ionic radii description.

- I built a holistic model that has the capability to predict the volumetric CTE of RE disilicate, RE monosilicate and RE silicate apatites as a function of chemical composition and crystal structure. Intriguingly, the model that was trained only using single-component compounds was also able to predict the CTE trend for two four-component systems: β-$C2/m$ $(Y, Yb, Lu, La)_2Si_2O_7$ and β-$C2/m$ $(Y, Yb, Er, Dy)_2Si_2O_7$ that were not part of the training data. Post hoc model interpretation of the trained model revealed the critical role of RE-O bond length, SiO_4 polyhedral volume and bond angle variance within SiO_4 polyhedral units.

This thesis lays the foundation for rational design of volumetric CTE in RE-Si-O compounds.

Acknowledgments

The journey towards the completion of this PhD thesis has been a long and often challenging one. It is a journey that I did not undertake alone, and there are many who have helped me along the way to whom I owe my deepest gratitude.

Firstly, I would like to express my heartfelt appreciation to my advisor, Dr. Prasanna Balachandran. Your guidance has been nothing short of transformational. You not only taught me the intricacies of research but also instilled in me the ethos of scientific inquiry. I have learned the importance of collaboration and the significance of detail-oriented research through your mentorship. Your patience in addressing my mistakes during my time as a graduate student has played a crucial role in my academic development. This encouraged me to seek your guidance repeatedly, particularly when tackling unfamiliar concepts, which significantly contributed to my growth as a researcher. During my personal struggles, you became an unwavering pillar of support, demonstrating not only your exceptional guidance in academic matters but also your remarkable patience and understanding as I navigated through my health issues. Your belief in my abilities, even during my most challenging moments, has left an permanent impact on my journey, and for that, I am deeply grateful.

Next, I wish to express my sincere gratitude to my committee members: Dr. Elizabeth Opila, Dr. Bi-Cheng Zhou, Dr. Stephen Baek, and Dr. Tianxi Li. Your insightful feedback and meticulous reviews have profoundly enriched my work. I want to specifically highlight the immense benefit I've received from Dr. Elizabeth Opila and her research group. Their experimental work has often served as a guiding light and a validation reference for my research. My collaboration with Dr. Elizabeth Opila and

Dr. Bi-Cheng Zhou has been particularly valuable, teaching me the true essence of collaboration in research. I must also acknowledge the invaluable contributions of Dr. Tianxi Li and Dr. Stephen Baek. Their expertise in data science and statistics provided essential guidance, particularly in the realm of machine learning, a pivotal aspect of my thesis. Their insights shaped the direction of this study and enabled me to navigate complex analytical tasks.

I've had the privilege of collaborating with a talented group of research colleagues, an experience I hold dear. Regrettably, I can't list everyone, but your recognition is owed. The problems we tackled together, the brainstorming sessions we engaged in, and the shared celebrations of small victories are all cherished memories. Our collaborative projects not only marked milestones in my academic path but also etched fond recollections that will accompany me forward.

On a personal note, I owe a tremendous debt of gratitude to my friends, Pramod and Siddarth, and my family. Throughout this challenging journey, you've served as my emotional anchors. From late-night pep talks to essential breaks, you've provided the stability that kept me grounded. I must particularly acknowledge my parents, whose unwavering love and encouragement have been my constant driving force, even during my toughest moments.

To each of you, I extend my heartfelt thanks for being integral to this journey. This thesis encapsulates not only my own efforts but stands as a testament to the immeasurable support you've provided.

Contents

Symbols and Acronyms xi

List of Figures xv

List of Tables xxx

1 Introduction 1

1.1 Background and Motivation 1

1.2 Research Topics Discussed 4

1.3 Specific Aims 7

1.4 Thesis Organization 9

2 Background 12

2.1 CTE fundamentals 12

2.1.1 Microscopic Thermal Expansion Fundamentals 12

2.1.2 Macroscopic Thermal Expansion Fundamentals for Pure Materials 15

2.1.3 Thermal Expansion Mismatch 17

2.2 Prospective materials for EBC applications 19

2.3 RE-Si-O Crystal Chemistry and Descriptors 21

3 Methods 30

3.1 An Overview of Density Functional Theory 30

3.1.1 Structure Optimization 33

3.1.2 Density of States Calculation 36

3.1.3 Formation Energy Calculation 37

3.1.4 Special Quasirandom Structures (SQS) 38

3.2 Machine Learning 40

3.2.1 The Experience 41

3.2.2 The Task 41

3.2.3 The Performance Measure 48

3.2.4 Uncertainty Quantification 49

3.2.5 Explainable artificial Intelligence 61

4 Crystal Chemistry of $\mathbf{RE_2Si_2O_7}$ 70

4.1 Chapter Organization 70

4.2 Introduction 71

4.3 DFT calculations of the $RE_2Si_2O_7$ crystal chemistry 74

4.4 Structure-CTE relationship of $RE_2Si_2O_7$ using Machine Learning . . 77

4.4.1 Dataset and Descriptor Generation 77

4.4.2 Machine Learning and Bootstrap Resampling 82

4.5 Experimental validation of $Sm_2Si_2O_7$ 85

4.6 Summary 86

5 Crystal Chemistry of RE_2SiO_5 91

5.1 Chapter Organization 92

5.2 Introduction 92

5.3 DFT study of the RE_2SiO_5 crystal chemistry 95

5.4 Exploring Structure-CTE relationship of RE_2SiO_5 using Machine Learning 96

5.4.1 Dataset and Descriptor Generation for Training Machine Learning Models 96

5.4.2 Machine Learning and Bootstrap Resampling 99

5.5 Anisotropy trends in RE_2SiO_5 101

5.6 Summary 105

6 Holistic CTE model for the RE-Si-O crystal chemistry 111

6.1 Chapter Organization 112

6.2 Introduction 112

6.2.1 DFT calculations of the Rare-Earth Silicate Apatites 114

6.3 Structure-CTE relationship of RE-Si-O using a holistic ML model . . 119

6.3.1 Dataset and Descriptor Generation 119

6.3.2 Machine Learning and Bootstrap Resampling 124

6.4 Summary 134

7 Conclusion 140

8 Future Work 145

8.1 Equation-Informed Bayesian Models 145

8.2 Calibration for prediction interval algorithm 145

8.3 Influence of defects on electronic structure trends 146

Appendices 148

A Trustworthy ML 148

A.1 Trustworthy ML algorithms 148

B Crystal Chemistry of $RE_2Si_2O_7$ 152

B.1 Total energy difference of $RE_2Si_2O_7$ from DFT 152

C RE Silicate Apatites 153

C.1 Correlation analysis of polyhedral descriptor set 153

C.2 DFT total energy data of RE silicate apatites 154

D Publications 165

Bibliography 167

Symbols and Acronyms

E_H Hartree energy. 30

E_{tot} Total electronic energy of a system. 30

E_{xc} Exchange-correlation energy. 30, 31

R^2 Coefficient of determination. 125

T_s Kohn-Sham kinetic energy term. 30

ΔE DFT total energy difference with respect to the ground state structure. iv, v, xx, xxi, xxiii, xxvi–xxxi, 35, 36, 74–77, 84, 86, 90, 95, 96, 102, 105, 108, 116–118, 142, 155–164

ϕ Kohn-Sham orbitals. 30

ρ Electron density. 30, 31

σ Standard error from bootstrap. xxxi, 128

v_{ext} External potential energy. 30

A Alkali Element. 117

ABCTE Apparent Bulk Coefficient of Thermal Expansion. xix, xxi, xxv, xxx, xxxi, 78, 79, 83, 84, 89, 97, 100, 107, 120, 124, 125, 127, 128, 133, 134, 137, 143

AE Alkaline Earth Element. xxiii, 115–117, 119

CMAS Calcium-Magnesium-Alumino-Silicate. 20, 22

CTE Coefficient of Thermal Expansion. iv, v, xxiv, xxv, xxxi, 3–5, 7–11, 111–114, 119, 121, 123, 128–134, 139–144

DFT Density Functional Theory. xxiii, xxv, xxxi, 4, 5, 9–11, 24, 28, 30, 31, 116–118, 125–128, 130, 139, 141, 153

DOS Total density of states. 36, 37

EBC Environmental Barrier Coating. 4, 5, 11, 111, 128, 141

eSVR Ensemble Support Vector Regression. xxi, xxv, xxxi, 99, 107, 124, 125, 128, 129, 132, 134, 137, 138, 143

GGA Generalized gradient approximation. 31

ICE Individual conditional expectation. 67, 68

IFC Interatomic Force Constant. 130

LDA Local density approximation. 31

ML Machine Learning. xxv, xxxi, 38, 111, 112, 114, 119, 124, 126–128, 134, 137, 140, 143, 153

OOB Out of bag. 54–57

PCC Pearson Correlation Coefficient. 153

PDOS Projected density of states. 37

PDP Partial Dependence Plot. 67, 129, 130, 132, 134, 143, 144

PI Prediction interval. xxx, xxxi, 51–53, 55, 59, 60, 128

PICP Prediction interval coverage probability. xxx, 59, 60

QHA Quasi-Harmonic Approximation. xxv, xxx, xxxi, 3, 73, 83, 84, 93, 125, 128, 139

RE Rare Earth. 141–143

RE-Si-O Rare Earth-Silicon-Oxygen. 4, 5, 7–11, 20, 23, 24, 28, 29, 112, 113, 119, 121, 128, 134, 135, 140–142, 144

RMSE root mean squared error. 49, 65, 83, 125

SD Standard Deviation. xvi, 25

SHAP Shapley additive explanations. 61, 66, 67

SQS Special Quasirandom Structures. 9, 38–40, 126, 127

SVM Support Vector Machine. 44, 46

SVR Support Vector Regression. 45, 124, 129

UQ Uncertainty Quantification. 50, 51

UVa University of Virginia. 59

UVa-MSE University of Virginia, Materials Science and Engineering. 126, 127

XAI Explainable Artificial Intelligence. 49, 61, 132, 140

List of Figures

1.1 A schematic illustration of a typical ML approach used in materials science. The approach generally involves collection of data and knowledge regarding the problem at hand to build a dataset, which then will typically be used to build a black-box model that outputs certain predictions. However, the human user (or the domain expert) will not know when and why to trust the prediction with respect to the input. 6

1.2 A schematic of the overarching goals for this thesis. (a) I gather data and incorporate domain knowledge, (b) Using the data and knowledge, I then build a trustworthy ML approach where the obbjective is to not only train a generalizable black-box model that mimics the nature but also address the question of when and why to trust the model using uncertainty quantification and (c) Model explanation, where I peek into the black-box and uncover the hidden patterns. This is crucial for developing an understanding of the underlying scientific phenomena. . 8

2.1 The asymmetric potential energy well for the separation of two atoms, which fundamentally describes the primary mechanism of thermal expansion [38]. 13

2.2 The ball and spring model of atoms and bonds, respectively, describing the nature of atoms in a lattice [39]. 14

2.3 Bravais lattice crystal systems and the corresponding coefficient of thermal expansion tensor. Taken from Ref. [5]. 17

2.4 The crystal structures of some of the observed rare-earth disilicates (including their seven polymorphs), monosilicates (two polymorphs) and silicate apatite (hexagonal crystal structure). 21

2.5 Schematic showing the logic for descriptor generation. The RE-O and Si-O polyhedral units are highlighted. The average and standard deviation (SD) of each polyhedral attribute extracted from VESTA for RE-O and Si-O polyhedra are extracted for each compound. Along with formation energy we have an initial set of 25 descriptors. Using correlation analysis, seven descriptors (Si_avg_bond_length, Si_poly_volume, RE_avg_bond_length, AK_avg_bond_length, RE_avg_bond_length_sd, Si_bondangle_var) and form_Energy are shortlisted. 25

3.1 Scheme of the self-consistent solution of the Kohn-Sham equations. [86, 87] 32

3.2 Support Vector Machines for Classification. Individual points are colored according to their classification. Dashed red and blue lines are the support vectors with the black line being the optimized decision boundary. 44

3.3 Support Vector Regression. Here there is no classification, blue lines represent the support vectors used to identify the black line or optimized regression line. 45

3.4 The graphic shows the different sources of uncertainty. (a) Depicts the uncertainty due to lack of data (called the model error). The model error reduces as we collect more data in the regions that lack them. (b) Depicts the uncertainty due to conflicting data for the same input space (called the sample noise). This is an irreducible error and can only be quantified but not reduced. 50

3.5 A schematic of our proposed bootstrap prediction interval algorithm. The algorithm is a modification of the prediction interval algorithm published by Srivatsava and Kumar [111]. Our modification lies in the approximation of observation noise distribution using the .632 measure [112] of bootstrap residuals instead of training residuals as proposed by Srivatsava and Kumar. This is similar to the recent work by Mougan and Nielsen. 54

3.6 A general overview and categorization of Model explanation. The schematic shows the broad classification of model explanations including glass box models and black box explanations or post-hoc model explanations. This thesis focusses only on the black box explainers which can further be classified into two types: 1) Variable attribution and 2) What-if plots. They both can be used to provide model explanations in different granularities (local and global). While local explanation involves a single observation, global explanation explains the model as a whole considering all the observations in question. . . 62

3.7 The accuracy-explainability tradeoff in AI. The accuracy vs explainability (notational) graph depicts the lack of explainability or transparency in accurate models and the lack of accuracy in explainable models. The goal of employing black-box explainers is to harness the advantages of both ends of the spectrum. 63

3.8 The flow chart for ML and model explainability approach for explaining model behaviour at local and intermediate levels. (Adapted from our recently published paper [129].) 68

4.1 The crystal structures of seven polymorphs in the $RE_2Si_2O_7$ family of compounds. The RE-site, SiO_4 tetrahedral unit, and the disilicate unit is highlighted. 72

4.2 Pair-wise statistical correlation analysis for the eight descriptors from DFT calculations for 16 $RE_2Si_2O_7$ compounds in seven polymorphs. Dark red and dark blue indicate strong positive and negative correlation, respectively. For example, a strong positive correlation (> 0.9) is found between the lattice constant c, the number of atoms in the unit cell (#atoms), and the volume. Since correlation indicates redundancy, we did not consider the lattice constant c and #atoms as input for building ML models. The remaining descriptors, a, b, α, β, γ, and volume were used as inputs for building ML models. 78

4.3 Decision trees for classification of polymorphs in RE disilicates based on GINI impurity using the training data shown in Table 6.1. At each node, a check occurs and if true proceeds to the left and vice-versa until it reaches the leaf node that displays the classification. Under each classification, the compounds belonging to the space group are listed. 79

4.4 Graphical representation of polyhedral metrics (Y-axes) in relation to rare-earth (RE) ionic radii in Angstroms (Å) (X-axes), focusing on a charge state of 3+ and a coordination number of 8. The plots are representative of all 112 RE disilicates. The secondary X-axis bears the annotations of RE cation in the compound. 88

4.5 Performance of the trained ensemble of 25 SVR models. The X- and Y-axes are the known and ML predicted ABCTE data, respectively. The error bar represent the standard deviation from the ensemble of ML models. The red dashed line represents the X=Y line and the data points falling on this line indicate perfect agreement between the ML models and the known data. The black dots are the experimental data points used to train the models. The blue diamond and green squares represent the test data that were not used to train the ML models. The true values for the blue diamond and green square data points are taken from the experimental measurements and published DFT-QHA calculations [77], respectively. 89

4.6 Results from (a) X-ray Diffraction (XRD), (b) SEM image, and (c) Dilatometry measurements for the synthesized $Sm_2Si_2O_7$ ceramic sample. The majority phase in XRD is the tetragonal $P4_1$ A-phase, which was predicted as metastable from DFT ΔE calculations. The SEM image shows the sample as predominantly $Sm_2Si_2O_7$ with a minor amount of partially reacted Sm_2SiO_5 and unreacted SiO_2 together with porosity. The CTE from dilatometry for the temperature range of 300°-975° C is measured as $11.55 \pm 0.18 \times 10^{-6}$ K^{-1}, which is in agreement with the ML prediction of $10.39 \pm 1.41 \times 10^{-6}$ K^{-1}. 90

5.1 The crystal structure of Y_2SiO_5 in (a) $C2/c$ (ground state) and (b) $P2_1/c$ (metastable) space groups. The two crystallographically unique Y-atoms (Y1 and Y2) and the Si atoms are labeled for clarify. 93

5.2 Decision trees for classification of polymorphs in RE monosilicates based on GINI impurity using the training data shown in table Table 5.2. At each node, a check occurs and if true proceeds to the left and vice-versa until it reaches the leaf node that displays the classification. Under each classification, the compounds belonging to the space group are listed. 98

5.3 Pearson correlation coefficient plot for the dataset comprising 32 RE_2SiO_5 compounds. Dark red indicate strong positive correlation. The plot shows that the variables “b” and “volume” are the only two variables that are not highly correlated and thus are not redundant and suitable inputs for the ML model. 99

5.4 Graphical representation of polyhedral metrics (Y-axes) in relation to Rare Earth (RE) Ionic Radii in Angstroms (Å) (X-axes), focusing on a charge state of 3+ and a coordination number of 8. The secondary x-axis bears the annotations of RE cation in the compound. The plots are representative of all 32 RE monosilicates. 100

5.5 Performance of the trained eSVR model. The X- and Y-axes are the known and ML predicted ABCTE data, respectively. The error bar represent the standard deviation from the ensemble of ML models. The red dashed line represents the X=Y line and the data points falling on this line indicate perfect agreement between the ML models and the known data. The black dots are the experimental data points used to train the models. The red diamonds represent the validation data that were not used to train the ML models. The validation data points include X2-$C2/c$ Er_2SiO_5 reported by Khan *et al.* [184] and Ridley *et al.* [24], X2-$C2/c$ Tb_2SiO_5 by Ridley *et al.* [24] and X1-$P2_1/c$ Sm_2SiO_5 reported by Ogura *et al.* [185]. 107

5.6 The ML predicted CTE with uncertainties from bootstrapping for all the RE_2SiO_5 compounds.The polymorphs predicted to be ground state (ΔE=0) by DFT calculations are marked by asterisks. The polymorphs marked in red are included in the training data and green marks validation set. The ones marked in black are unexplored compounds. . . 108

5.7 CTE Anisotropy trends in RE_2SiO_5 (RE=Sc,Y,Yb,Nd,Er,Dy) reported by Ridley et al [24]. 108

5.8 RE-O coordination environment in (a) Sc_2SiO_5 and (b) Y_2SiO_5 in their respective ground state $C2/c$ structures. In (c), we show the local La-O environment for the La_2SiO_5 compound in its metastable $C2/c$ structure. The longest bond lengths that define the *seventh* RE-O bond are highlighted in the Figure for the Y_2SiO_5 and La_2SiO_5 compounds. 109

5.9 Total and local DOS for (a) Sc_2SiO_5 and (b) Y_2SiO_5 in their respective ground state $C2/c$ structures. In (c), we show the total and local DOS for La_2SiO_5 in its metastable $C2/c$ structure. E_F is the Fermi energy. 109

5.10 RE-O coordination environment in (a) Y-substituted Sc_2SiO_5 in the hypothetical $C2/c$ structure with a smaller unit cell volume, (b) Y_2SiO_5 in its ground state $C2/c$ structure with equilibrium volume, and (c) Y-substituted La_2SiO_5 in its hypothetical $C2/c$ structure with a larger unit cell volume. The longest bond lengths that define the seventh RE-O bond are highlighted in the Figure. 110

5.11 Total and local DOS for (a) Y-substituted Sc_2SiO_5 in the hypothetical $C2/c$ structure with a smaller unit cell volume, (b) Y_2SiO_5 in its ground state $C2/c$ structure with equilibrium volume, and (c) Y-substituted La_2SiO_5 in the hypothetical $C2/c$ structure with a larger unit cell volume. E_F is the Fermi energy. 110

6.1 Total energy difference, in from DFT calculations with respect to the lowest energy structure in $RE_8AE_2Si_6O_{26}$ for AE=Ba,where space groups given in parentheses indicate the final converged structure when the tolerance is set at 0.0001 or lower in FINDSYM. ΔE=0 signifies the ground structure. When ΔE is within a range of approximately 5 meV/atom, it suggests a close energetic competition with the ground state structure. 116

6.2 Total energy difference, in from DFT calculations with respect to the lowest energy structure in $RE_8AE_2Si_6O_{26}$ for AE=Ca,where space groups given in parentheses indicate the final converged structure when the tolerance is set at 0.0001 or lower in FINDSYM. ΔE=0 signifies the ground structure. When ΔE is within a range of approximately 5 meV/atom, it suggests a close energetic competition with the ground state structure. 117

6.3 Total energy difference, in from DFT calculations with respect to the lowest energy structure in $RE_9A_1Si_6O_{26}$ for A=Na,where space groups given in parentheses indicate the final converged structure when the tolerance is set at 0.0001 or lower in FINDSYM. ΔE=0 signifies the ground structure. When ΔE is within a range of approximately 5 meV/atom, it suggests a close energetic competition with the ground state structure. 118

6.4 Decision trees for determining CTE based on GINI impurity on the training data shown in table Table 6.1. At each node, a check occurs and if true proceeds to the left and vice-versa until it reaches the leaf node that displays the classification. Under each classification, the compounds belonging to the corresponding material classes are listed. 121

6.5 Visualization showing two-dimensional projections of our high-dimensional dataset based on the t-distributed stochastic neighborhood embedding (t-SNE) method [203]. 123

6.6 Graphical representation of polyhedral metrics (Y-axes) in relation to Rare Earth (RE) Ionic Radii in Angstroms (Å) (X-axes), focusing on a charge state of 3+ and a coordination number of 8. The plots are representative of the training data including all three material classes in RE-Si-O space. The variable AK_avg_bond_length is ignored as we only have 2 alkali earth/alkaline earth metal bearing apatites in the dataset. The secondary x-axis bears the annotations of RE cation in the compound. 136

6.7 Performance of the eSVR model trained on the dataset including all three material classes in RE-Si-O space. The X- and Y-axes are the known and ML predicted ABCTE data, respectively. The error bar represent the standard error from the ensemble of eSVR model. The grey lines represent the prediction interval limits constructed using the doubt algorithm [113]. The red dashed line represents the X=Y line and the data points falling on this line indicate perfect agreement between the ML models and the known data. The black dots are the data points used to train the models. The red diamonds represent the test data that were not used to train the ML models. 137

6.8 Results of global feature importance analysis for the eSVR model trained on the dataset including all three material classes in RE-Si-O space. Blue bar shows average impact, across all the models in the ensemble, on the dropout loss. Orange error bar depicts the standard deviation of the dropout loss across all the models. 138

6.9 PDPs showing the relationship between the descriptors (x-axis) and CTE (y-axis). The shaded regions indicate the uncertainties based on standard deviation. The plots of representative of the training dataset including all three material classes in RE-Si-O space. 139

6.10 Plots showing inverse correlation between RE_avg_bond_length and interatomic force constants of RE-O polyhedron [77] obtained by DFT-QHA calculations for a select RE disilicate compounds. 139

C.1 Pair-wise statistical correlation analysis for the training data used for the holistic CTE model. Dark red and dark blue indicate strong positive and negative correlation, respectively. We remove redundancy by only considering variables whose PCC was less than 0.7 for building ML models. 154

C.2 Total energy difference, in from DFT calculations with respect to the lowest energy structure in $RE_8AE_2Si_6O_{26}$ for AE=Ba,where space groups given in parentheses indicate the final converged structure when the tolerance is set at 0.0001 or lower in FINDSYM. ΔE=0 signifies the ground structure. When ΔE is within a range of approximately 5 meV/atom, it suggests a close energetic competition with the ground state structure. 155

C.3 Total energy difference, in from DFT calculations with respect to the lowest energy structure in $RE_8AE_2Si_6O_{26}$ for AE=Be,where space groups given in parentheses indicate the final converged structure when the tolerance is set at 0.0001 or lower in FINDSYM. ΔE=0 signifies the ground structure. When ΔE is within a range of approximately 5 meV/atom, it suggests a close energetic competition with the ground state structure. 156

C.4 Total energy difference, in from DFT calculations with respect to the lowest energy structure in $RE_8AE_2Si_6O_{26}$ for AE=Sr,where space groups given in parentheses indicate the final converged structure when the tolerance is set at 0.0001 or lower in FINDSYM. ΔE=0 signifies the ground structure. When ΔE is within a range of approximately 5 meV/atom, it suggests a close energetic competition with the ground state structure. 157

C.5 Total energy difference, in from DFT calculations with respect to the lowest energy structure in $RE_8AE_2Si_6O_{26}$ for AE=Ca,where space groups given in parentheses indicate the final converged structure when the tolerance is set at 0.0001 or lower in FINDSYM. ΔE=0 signifies the ground structure. When ΔE is within a range of approximately 5 meV/atom, it suggests a close energetic competition with the ground state structure. 158

C.6 Total energy difference, in from DFT calculations with respect to the lowest energy structure in $RE_8AE_2Si_6O_{26}$ for AE=Mg,where space groups given in parentheses indicate the final converged structure when the tolerance is set at 0.0001 or lower in FINDSYM. ΔE=0 signifies the ground structure. When ΔE is within a range of approximately 5 meV/atom, it suggests a close energetic competition with the ground state structure. 159

C.7 Total energy difference, in from DFT calculations with respect to the lowest energy structure in $RE_9A_1Si_6O_{26}$ for A=Cs,where space groups given in parentheses indicate the final converged structure when the tolerance is set at 0.0001 or lower in FINDSYM. ΔE=0 signifies the ground structure. When ΔE is within a range of approximately 5 meV/atom, it suggests a close energetic competition with the ground state structure. 160

C.8 Total energy difference, in from DFT calculations with respect to the lowest energy structure in $RE_9A_1Si_6O_{26}$ for A=K,where space groups given in parentheses indicate the final converged structure when the tolerance is set at 0.0001 or lower in FINDSYM. ΔE=0 signifies the ground structure. When ΔE is within a range of approximately 5 meV/atom, it suggests a close energetic competition with the ground state structure.161

C.9 Total energy difference, in from DFT calculations with respect to the lowest energy structure in $RE_9A_1Si_6O_{26}$ for A=Li,where space groups given in parentheses indicate the final converged structure when the tolerance is set at 0.0001 or lower in FINDSYM. ΔE=0 signifies the ground structure. When ΔE is within a range of approximately 5 meV/atom, it suggests a close energetic competition with the ground state structure. 162

C.10 Total energy difference, in from DFT calculations with respect to the lowest energy structure in $RE_9A_1Si_6O_{26}$ for A=Na,where space groups given in parentheses indicate the final converged structure when the tolerance is set at 0.0001 or lower in FINDSYM. ΔE=0 signifies the ground structure. When ΔE is within a range of approximately 5 meV/atom, it suggests a close energetic competition with the ground state structure. 163

C.11 Total energy difference, in from DFT calculations with respect to the lowest energy structure in $RE_9A_1Si_6O_{26}$ for A=Rb,where space groups given in parentheses indicate the final converged structure when the tolerance is set at 0.0001 or lower in FINDSYM. ΔE=0 signifies the ground structure. When ΔE is within a range of approximately 5 meV/atom, it suggests a close energetic competition with the ground state structure. 164

List of Tables

2.1 List of polyhedral descriptors explored in this thesis along with their descriptions. 26

3.1 A comparison of our PI algorithm against the "doubt" method by Mougan and Nielsen [113], where the metric used is the prediction interval coverage probability (PICP) for various prescribed confidence levels. Parity of PICPs with the confidence levels is considered to be the best result. The PICPs show that "doubt" outperforms our approach. 59

4.1 Total energy difference (ΔE, meV/atom) from DFT calculations with respect to the lowest energy structure in $RE_2Si_2O_7$, where RE=La, Ce, Sm, Gd, Nd, Pr, Yb and Y. Space groups given in parentheses indicate the final converged structure when the lattice constant tolerance is set at 0.0001 decimal places or lower in `FINDSYM`[157]. 75

4.2 The compounds used for training and testing the ML models is given. The ABCTE predictions from ML, along with the uncertainty (σ), is also given. All experimental (training and test data) and DFT-QHA (test data) ABCTE data are taken from Fernández-Carrión *et al.* [140], Dolan *et al.* [142] and Luo *et al.* [77], respectively. The temperature ranges used for the CTE determination in both experiments and DFT-QHA calculations are also given. (Data published in Ref. Ayyasamy$_RE2Si2O7$) 84

5.1 Total energy difference (ΔE, meV/atom) from DFT calculations with respect to the lowest energy structure in RE_2SiO_5. Space groups were determined using the `FINDSYM` [157] web application. Structures with ΔE=0 (bold face font) represent the ground state structure for that compound. 96

5.2 Thermal Expansion Coefficients Data for RE monosilicate collected from Literature. 97

5.3 Total energy difference (ΔE, meV/atom) from DFT calculations with respect to the lowest energy structure in RE_2SiO_5, where RE=Sc, Y, and La. Space groups were determined using the `FINDSYM` web application. Structures with ΔE=0 (bold face font) represent the ground state structure for that compound. 102

6.1 Thermal Expansion Coefficients Data From Literature. 120

6.2 The compounds used for testing and experimentally validating the eSVR models are given. The ML prediction for ABCTE $\times 10^{-6}$ K^{-1}, along with the bootstrap standard error (σ) and the prediction interval (PI) from the doubt algorithm [113], are given. The temperature ranges used for the CTE determination in both experiments and DFT-QHA calculations are also given. 128

B.1 Total energy difference, in from DFT calculations with respect to the lowest energy structure in, where Space groups given in parentheses indicate the final converged structure when the tolerance is set at 0.0001 or lower in FINDSYM [157]. 152

Chapter 1

Introduction

1.1 Background and Motivation

The coefficient of thermal expansion (CTE) is an intrinsic material property that is indicative of the extent to which a material expands upon heating.[1] From a historical perspective, heat change dimensions has been around since the first century B.C.; however, more attention was focused on thermal expansion with the invention of a thermometer in the late 1500s to early 1600s [2]. It was not until the discovery of Invar, Einstein's harmonic oscillators, and Grüneisen's rule that thermal expansion was given serious scientific thought in the early 1900s [2, 3, 4]. Most of the current body of thermal expansion knowledge was generated between 1950 and 1980, and it has significantly slowed down afterwards [2, 3].

The fundamental origin of thermal expansion can be directly traced to the anharmonic nature of the interatomic forces in solids [5]. If the forces are purely harmonic, there would be no thermal expansion because atoms vibrate about their equilibrium positions symmetrically whatever be the amplitude. However, the thermal vibrations in any real solids are to some extent anharmonic, which leads to the phenomenon of thermal expansion. The linear thermal expansion coefficient (α_l) and volume expan-

[1] Some materials with *negative* thermal expansion contract upon heating [1]. However, such materials will not be the focus of this thesis.

sion coefficient (α_v) are two important parameters used to quantitatively characterize the extent of thermal expansion in any given material. The α_l or α_v is defined as per unit change in length or volume with per unit change in temperature, respectively, and is normally expressed as $\alpha_l = dl/(ldT)$ or $\alpha_v = dV/(VdT)$, where l is length, V is volume and T is temperature. If the thermal expansion is isotropic, then $\alpha_v = 3\alpha_l$. Usually, an average linear thermal expansion coefficient (α_l) is measured and given over a temperature range in the literature. Until now, numerous microscopic models have been undertaken for estimating the thermal expansion coefficient. A full-fledged and well known theory of thermal expansion was firstly proposed by Grüneisen [6]. It is proved that the thermal expansion coefficient of a solid has a direct relation with the heat capacity and inverse relation with the bulk modulus through the Grüneisen parameter. Most of the later developments in theories of thermal expansion are related to a more sophisticated definition of Grüneisen parameter through different lattice dynamical calculations [7, 8].

There has been a long-standing interest in establishing quantitative crystal structure–electronic structure–CTE relationships in solids [9, 10, 11, 12, 1, 13, 14]. Thermal expansion has strong correlation with chemical bonding. It is generally believed that materials with stronger bonds will have lower thermal expansion because of their rigid structures [11]. As a result, several empirical models have been proposed to estimate CTE as a function of chemical bonding parameters. Megaw *et al.* deduced an empirical equation to qualitatively relate linear CTE to Pauling valence of chemical bond for simple compounds in cubic crystal structures [15]. Cameron *et al.* have correlated CTE of metal-oxygen bonds with their stretching frequency and reduced mass [16]. Hazen *et al.* initially developed an empirical relationship between the CTE and Pauling valence in several metal-oxygen coordination polyhedra [17]. Later, the authors provided a general formula (given in Equation 1.1) between the CTE (α) and chemical bonding parameters that include an empirical ionicity factor (S), valence of cation (Z_c) and anion (Z_a), and coordination number of cation (CN), [18]

$$\alpha = 4 \times 10^{-6} \frac{CN}{S^2 Z_c Z_a} \tag{1.1}$$

This formula was then used to predict the CTE of several binary and ternary oxides. Roy *et al.* [11] argue that the α described by Equation 1.1 should be treated as a zeroth-order approximation because it fails to describe the CTE for complex structures with two or more cation-centered polyhedral units in non-trivial topology and irregular coordination (e.g. monosilicates, disilicates and silicate apatites explored in this thesis). Moreover, the model for α also lacks an uncertainty quantification capability.

The state-of-the-art computational efforts to realize a predictive understanding of CTE require performing first principles calculations within the quasi-harmonic approximation (QHA) to account for temperature-dependent properties like CTE [7, 8, 19, 20, 21]. However, first principles based lattice dynamical calculations have the limitation of increased computational cost for complex structures [22]. Moreover, when there are instabilities in the phonon dispersion curves (e.g. a phonon mode with imaginary frequency) then there is no consensus about how to treat such data within the QHA approximation. Phonon instabilities are common when a crystal structure is not dynamically stable at 0 K [23]. This limits the universal applicability of first principles calculations to predict CTE. There is a need for developing alternative approach to complement the first principles calculations and guide experimental efforts in accelerating new materials discovery for targeted applications.

1.2 Research Topics Discussed

This thesis demonstrates how a combination of first principles calculations and data-driven machine learning (ML) methods can be used to develop a quantitative and predictive understanding of the relationship between the crystal structures and physical properties in the RE-Si-O crystal chemistry space. More specifically, I focus on the volumetric CTE property because calculation of CTE from first principles is expensive and CTE is one of the critical properties of interest to the environmental barrier coatings (EBC) community [24]. In addition to volumetric CTE, I have also focused on developing an understanding of factors that govern polymorph stability in the rare-earth disilicate ($RE_2Si_2O_7$) and rare-earth monosilicate (RE_2SiO_5) materials. All polymorphs in the $RE_2Si_2O_7$ and RE_2SiO_5 crystal chemistry belong to non-cubic crystal structures. As a result, CTE is anisotropic in these materials. Developing a quantitative understanding of CTE anisotropy is an open question that is beyond the scope of this thesis. Nonetheless, I investigated a specific aspect of CTE anisotropy in this thesis that captures the role of electronic structure for a smaller subset of RE_2SiO_5 compounds [25]. To investigate these important properties, I synergistically integrated first principles calculations based on density functional theory (DFT) with ML.

This thesis addresses the following research objectives. Initially, the thesis delves into analyzing the DFT total energy differences for the expansive RE-Si-O domain, which incorporates monosilicates, disilicates, and apatites. These calculations not only indicate which polymorph of a compound is energetically favorable, but also gives us the optimized structures. This data paves the way for our subsequent objective: rapidly estimating the CTE through a ML model that takes the optimized structure as input. In this context, we can establish descriptors on two fronts: (1)

the unit cell parameters and (2) polyhedral descriptors. It is important to probe the relationships between structure and property at both these levels, investigating the similarities and contrasts they hold. Although computationally, either level could be our starting point, the unit cell parameters are more accessible to the experimental community. I demonstrated the predictive power of the unit cell parameters for the CTE within RE disilicates and monosilicates. Leveraging the experiments conducted by our colleagues, we validate our models and demonstrate the efficacy of our strategy for EBC applications. Although predicting the apparent bulk CTE (or volumetric CTE) with ML is beneficial, the design of EBCs cannot overlook the tensorial nature of CTE's anisotropy. While there are experimental findings that highlight unique anisotropy patterns in compounds [24], a deeper insight into these patterns would enhance property customization. As a result, I utilized DFT calculations to shed light on the connection between electronic structures and the anisotropy of CTE.

The second way to describe the relationship between structure and CTE would be via polyhedral descriptors. It has a distinct advantages compared to the unit cell parameters: (1) It may carry mechanistic information that can be correlated with phonon dispersion curves and phonon density of states and (2) It can serve as descriptors that are unique to different structures across the entire RE-Si-O space including disilicates, monosilicates and apatites. This will enable us to accomplish our ultimate task of building a holistic data-driven ML model that will establish a quantitative relationship between the polyhedral descriptors and the CTE for the entire RE-Si-O space. I developed a novel representation scheme based on polyhedral descriptors that uniquely fingerprints both single- and multi-component compounds.

A typical ML approach (Figure 1.1) involves using a database of known materials that is constructed by compiling available data and domain knowledge about some

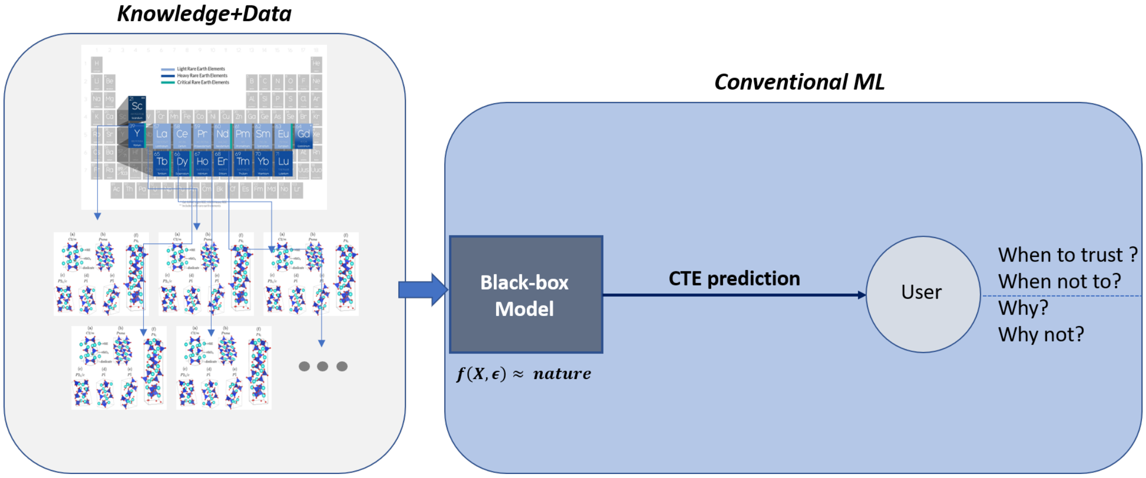


Figure 1.1: A schematic illustration of a typical ML approach used in materials science. The approach generally involves collection of data and knowledge regarding the problem at hand to build a dataset, which then will typically be used to build a black-box model that outputs certain predictions. However, the human user (or the domain expert) will not know when and why to trust the prediction with respect to the input.

phenomena of interest as input, fitting a linear or non-linear function that mimics the underlying nature of the problem, and generate predictions (with or without uncertainties) on new and previously unseen data. There is no component in this approach that was initially integrated in the formalism that will help the domain expert ascertain whether the predictions will adhere to the underlying science. More recently, the use of physics-informed ML or scientific ML addresses some of the concern [26, 27]. Also, the domain expert will not know when not to trust the model with respect to the input space.

This approach may be sufficient if accurate prediction is the only goal. However, it is imperative that we explain the black-box model, and the model confidence in order to understand and explore the underlying science. This is especially true in the case of small-sized datasets (less than a few hundred datapoints, which is typical in materials science). To address this problem, one of the promising research directions is to focus on developing a trustworthy ML formalism (Figure 1.2). Two key properties can be attributed to the trustworthy ML approach: (i) uncertainty quantification - to know when the model makes predictions that it is confident about (Figure 1.2b) , and (ii) model explanation - to know why (even approximately) if the prediction made by the model is backed by science (Figure 1.2c). I discuss a rigorous uncertainty estimation algorithm that quantifies the prediction intervals using bootstrap-based statistical methods. There is a rapidly growing interest in the application of ML-based approaches to accelerate new materials discovery [28, 29]. Many of these approaches rely on uncertainties from ML predictions to meaningfully explore the search space. There aren't many studies in the literature that have explored the promise of prediction intervals for materials exploration [30]. In this thesis, I demonstrate a novel prediction interval algorithm, study the efficacy of prediction intervals on benchmark materials science datasets and then apply it to the models used in this thesis. I also demonstrate novel post hoc model interpretation algorithms that can explain the predictions from the black-box models to provide an understanding of the hidden model behavior. The purpose is to address the following question: when and why can we trust on a ML model to make decisions, especially when trained on small and sparse datasets? Addressing these questions bridges the gap between black-box ML models and domain experts.

1.3 Specific Aims

This thesis addresses the following central questions:

1. Can we accelerate the design of novel compounds in the complex RE-Si-O chemical space with targeted CTE? This requires us to establish a hitherto

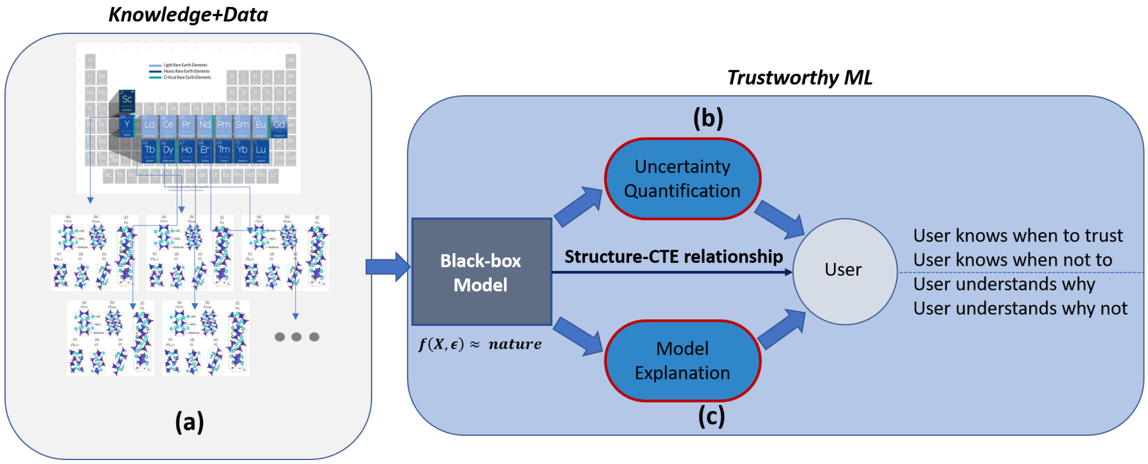


Figure 1.2: A schematic of the overarching goals for this thesis. (a) I gather data and incorporate domain knowledge, (b) Using the data and knowledge, I then build a trustworthy ML approach where the obbjective is to not only train a generalizable black-box model that mimics the nature but also address the question of when and why to trust the model using uncertainty quantification and (c) Model explanation, where I peek into the black-box and uncover the hidden patterns. This is crucial for developing an understanding of the underlying scientific phenomena.

unknown quantitative structure-property relationships in the RE-Si-O family of compounds using computational and ML approaches.

2. Can we ensure trust in the ML approach by capturing insights into when/where a ML model can succeed and/or fail (see Figure 1.2b). Associating the CTE prediction with an accurate measure of model confidence will help the community make informed design decisions.

3. Can we understand why a black-box ML model makes a certain prediction? Structure-property relationships in materials can be better understood if the predictive understanding is backed up by science hidden in the model (see Figure 1.2c).

1.4 Thesis Organization

In this thesis, I have compiled a total of eight chapters (including Introduction). The rest of the chapters are organized as follows:

- Chapter 2 reviews the foundational understanding of CTE from both microscopic and macroscopic viewpoints. Building on this groundwork, the chapter scrutinizes the ramifications of mismatches in thermal expansion on the mechanical integrity of systems where this is a concern. The narrative then pivots to well-researched use-case: EBCs, and explores potential materials in the RE-Si-O family that are suitable for this application. Emphasis is placed on the need for establishing a link between crystal structure and CTE in these silicate compounds. The chapter wraps up by motivating bond geometrical descriptors based on RE-O and Si-O polyhedral units as information carriers of CTE across different materials classes in the RE-Si-O chemical space.

- Chapter 3 is partitioned into two main sections. The first section centers around DFT and its applications. It covers topics such as the optimization of crystal structures, mapping out the total energy as a function of various polymorphs, and calculating density of states, which will later be important for exploring anisotropic trends in CTE in Chapter 5. The section also delves into the calculation of formation energy, one of the variables I used for understanding the relationship between structure and CTE, and finally describes the special quasirandom structures (SQS) method [31] needed for simulating multi-component solid solutions. The second section is devoted to ML methodologies. It opens with an overview of prevalent ML techniques, followed by a critique of commonly used performance and uncertainty metrics, highlighting their limitations in building user trust. The "black box" nature of certain ML algorithms is also discussed. The chapter further elaborates on uncertainty quantification and explainable artificial intelligence as two approaches that can enhance trust in ML models. The algorithms presented in this section are model- and data-agnostic, but their application to models in subsequent chapters will be determined based on the specific needs of each study. For the ML models focused on $RE_2Si_2O_7$ (Chapter 4) and RE_2SiO_5 (Chapter 5), I'll use standard error for uncertainty assessment. For the holistic ML model, a more stringent prediction interval method was used. Additionally, post hoc explanations will be applied to the holistic model to better understand the link between structure and CTE in RE-Si-O crystals.

- Chapter 4 discusses the crucial role of $RE_2Si_2O_7$ in EBCs and outline its design challenges. I used DFT calculations to assess the total energy of its polymorphs. Next, we present ML algorithms that utilize unit cell parameters obtained from DFT optimized crystal structures to establish a relationship between structure and CTE, enabling quick CTE predictions along with uncertainties across the entire $RE_2Si_2O_7$ chemical space. Additionally, I employed polyhedral descriptors to offer insights into structure-CTE relationships that simpler descriptors like RE ionic radii may overlook. The chapter concludes with experimental validation (conducted by our collaborators), thereby reinforcing the credibility of our combined DFT and ML approaches.

- Chapter 5 discusses the DFT and ML approach applied to RE monosilicates. Similar to Chapter 4, I calculated the total energy of RE_2SiO_5 compounds in the two polymorphs and utilized ML models to predict their CTE based on DFT optimized unit cell parameters. Recognizing the importance of CTE anisotropy in EBC applications, the later part of the chapter focuses on DFT calculations linking the *d*-orbital bandwidth and RE-O effective coordination numbers in isoelectronic Sc_2SiO_5, Y_2SiO_5, and La_2SiO_5compounds whose CTE anisotropy data was taken from the published literature [24].

- Chapter 6 aims to achieve the overarching goal of the thesis—a holistic ML model for describing CTE across the entire RE-Si-O chemical space. To accomplish this, the chapter first introduces data related to RE apatites, a class of compounds not previously discussed in the thesis. It then explores the potential energy landscape pertinent to these compounds. Transitioning to machine learning, the chapter delves into the construction of a holistic ML model for CTE. The model employs polyhedral descriptors as training data, providing insights that go beyond traditional descriptors like RE ionic radii. To instill trust in the model, I apply a prediction interval algorithm for uncertainty quantification and employ post hoc model interpretability methods. These methods help to reveal the functional relationship between polyhedral descriptors and CTE, offering new insights to the community. The consistency of the results with other physical parameters further validates the approach.

- In Chapter 7, a summary of the research findings and key results are provided to conclude the thesis.

- In Chapter 8, future research directions based on the thesis outcomes are outlined.

Chapter 2

Background

2.1 CTE fundamentals

This chapter delves into the core principles of thermal expansion, starting with an exploration of the microscopic and macroscopic perspectives on the CTE in crystalline materials. Following this foundation, the chapter examines the impact of thermal expansion mismatches on the mechanical stability of systems where such issues are prevalent. I then shift my focus to a commonly studied application, the EBC system, and the prospective materials within the rare earth-silicon-oxygen (RE-Si-O) spectrum for this application. The chapter emphasizes the importance of developing a quantitative relationship between structure and CTE for these silicate materials. I conclude by identifying the descriptors that could facilitate a mechanistic understanding of thermal expansion.

2.1.1 Microscopic Thermal Expansion Fundamentals

At the microscopic level, crystalline solids consist of atoms arranged in a periodically ordered lattice or repeating structure [32]. These atoms are not fixed but continually vibrate or oscillate around an average position [32, 33]. This system of atoms can be represented through a Boltzmann energy distribution, initially formulated around

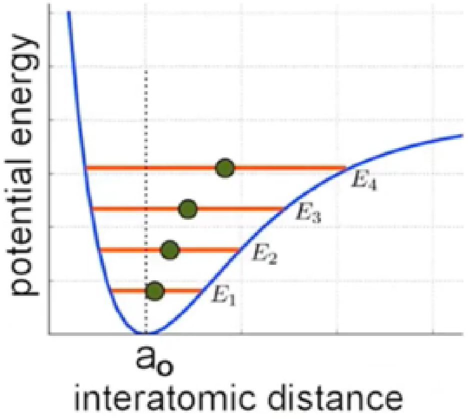


Figure 2.1: The asymmetric potential energy well for the separation of two atoms, which fundamentally describes the primary mechanism of thermal expansion [38].

Einstein's theory of vibratory harmonic oscillators, using Planck's energy quantization [32, 34, 35]. While the harmonic term of the energy function can estimate many crystalline properties, it does not predict thermal expansion [32]. If only harmonic oscillations were considered, the vibration amplitude would increase with temperature, but the atoms' average position would remain unchanged [36, 32]. In the most basic scenario, the anharmonicity of the energy well between two atoms offers a fundamental mechanism for describing thermal expansion, as depicted in Figure 2.1 [37, 32, 35, 38].

This model of an asymmetric potential energy well can be extended to a straightforward one-dimensional model for vibratory motion, using a ball and spring model to represent atoms in the lattice, as illustrated in Figure 2.2. However, atoms in the lattice move not only in alignment with the springs but also perpendicularly to them, creating a wave-like three-dimensional motion [37, 39]. These lattice vibrations' energy can be quantized as phonons, serving as one of the primary mechanisms for thermal transport [32, 35]. The characteristics of the bonds between atoms in the lattice influence atomic spacing and, consequently, thermal expansion behavior [32].

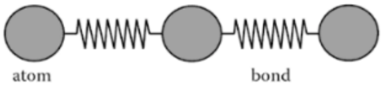


Figure 2.2: The ball and spring model of atoms and bonds, respectively, describing the nature of atoms in a lattice [39].

On a microscopic level, thermal expansion theories are often expressed using thermodynamic quantities [37, 36]. Mie [37] initially predicted that the ratio of the volumetric CTE to the isothermal compressibility would remain constant. Grüneisen observed that the ratio of the volumetric CTE over the heat capacity was consistent for many metals, leading to his rule [37, 36], expressed by Equation 2.1 [36]:

$$\frac{3\alpha}{\chi_T} = \frac{\gamma C_V}{V} \tag{2.1}$$

where α is the linear CTE, χ_T is the isothermal compressibility, C_V is the specific heat, V is the volume, and γ is the Grüneisen constant [36]. This parameter remains constant with temperature change across many crystalline solids, reflecting the anharmonicity of the potential energy well for a pair of atoms [37]. While Grüneisen's rule does not account for phase transitions or anisotropic materials, it formed the foundation for many subsequent theories [37]. More thermodynamic details and equations for microscopic thermal expansion can be found in references [40, 41], including the contributions to the CTE from electronic or magnetic effects. It is crucial to recognize that any alteration to the crystal lattice can lead to significant changes in the thermal expansion coefficient. For instance, a discontinuity in the CTE occurs as the temperature nears a structural phase transition, where the crystal lattice rearranges its structure, or at a ferromagnetic transition when the attraction or repulsion of atoms alters, thus modifying the atoms' equilibrium positions [37]. Estimates of the CTE can be derived from the microscopic understanding of thermal expansion. The melting point of a solid at the microscopic level happens when thermal vibrations are intense enough to break the lattice bonds [32]. The volumetric expansion of the crystal lattice from 0 K to the melting point for many metals is roughly 8% [32]. Therefore, as a general guideline, the higher the melting point, the less expansion occurs over each temperature segment, reducing the thermal expansion [32]. Stronger bonds shorten the bond length and deepen and symmetrize the potential energy well, lowering a material's CTE [32]. Similarly, an increase in bond valence leads to a decrease in the thermal expansion coefficient [32].

While these generalizations are applicable to pure metals, most engineered systems utilize structural materials that are alloys, introducing additional complexities. A prominent example is Invar, an alloy composed of 36 wt.% nickel and 64 wt.% iron [32]. Below the Curie point, Invar exhibits ferromagnetic properties, experiencing a volumetric magnetostriction that restricts thermal expansion due to the magnetic field. Additionally, at temperatures beneath the Debye temperature, not all vibrational phonon modes are active, leading to nonlinear expansion behavior in Invar at colder temperatures [38].

2.1.2 Macroscopic Thermal Expansion Fundamentals for Pure Materials

Thermal expansion is typically assessed and documented through the macroscopic coefficient of linear thermal expansion, α_{ij}, a second-rank symmetric tensor property [36, 32]. This coefficient connects the strain, or change in length per original length, to a temperature change, as expressed in Equation 2.2 [32]:

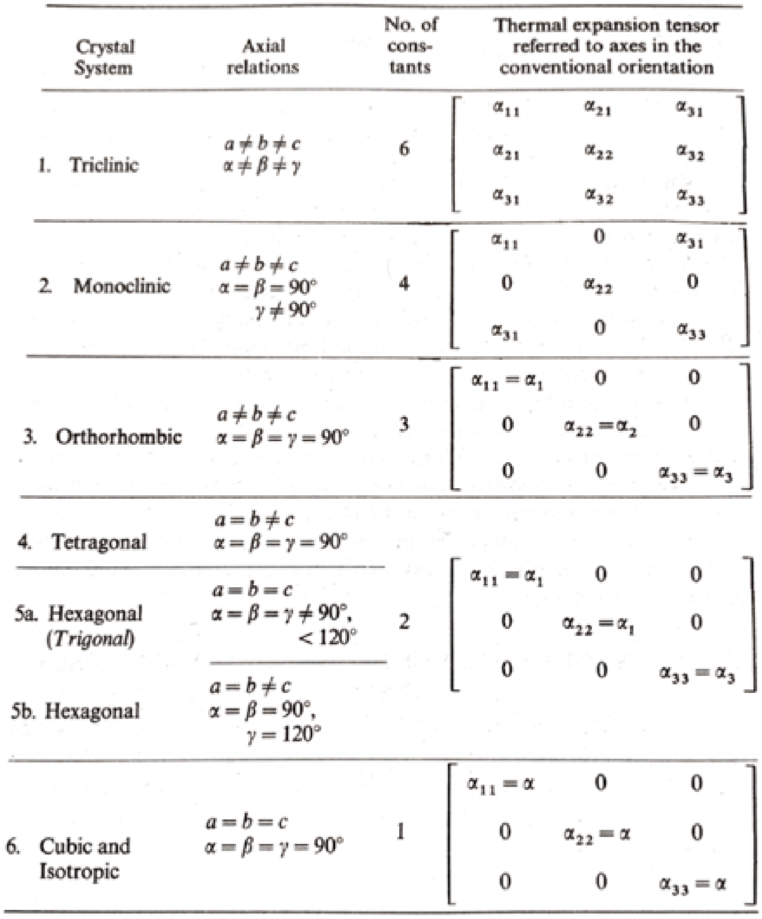
$$\epsilon_{ij} = \alpha_{ij}\Delta T \tag{2.2}$$

Here, the strain is denoted by ϵ_{ij}, resulting from the linear coefficient of thermal expansion, α_{ij}, due to a temperature change, ΔT. Since α_{ij} is a second-rank tensor property, it can exhibit anisotropic behavior, leading to varying expansions along different directions, with up to 6 independent directions based on the minimum symmetry [36, 32]. Each crystal system characterizes the inherent symmetry of atoms in the material, describing the spatial relationship of material properties. The number of independent thermal expansion constants required to define the CTE for each crystal system is summarized in Figure 2.3.

Another approach to describing a material's thermal expansion is through the volumetric thermal expansion coefficient, β, measured relative to a reference volume, as shown in Equation 2.3 [37, 32]:

$$\beta = \frac{1}{V}\left(\frac{\delta V}{\delta T}\right)_P \tag{2.3}$$

Here, V represents the volume, and T the temperature, taken at constant pressure, P. The volumetric thermal expansion for isotropic materials is simply three times the linear CTE and is the sum of the principal axes of the linear CTE for anisotropic materials [37, 32]. Although the CTE is inherently a tensor property, it is often reduced to a scalar quantity, disregarding the anisotropy, for most materials [32]. This simplification is evident in textbooks and handbooks that list nominal CTE values for various materials as scalars.



Crystal System	Axial relations	No. of constants	Thermal expansion tensor referred to axes in the conventional orientation
1. Triclinic	$a \neq b \neq c$ $\alpha \neq \beta \neq \gamma$	6	$\begin{bmatrix} \alpha_{11} & \alpha_{21} & \alpha_{31} \\ \alpha_{21} & \alpha_{22} & \alpha_{32} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} \end{bmatrix}$
2. Monoclinic	$a \neq b \neq c$ $\alpha = \beta = 90°$ $\gamma \neq 90°$	4	$\begin{bmatrix} \alpha_{11} & 0 & \alpha_{31} \\ 0 & \alpha_{22} & 0 \\ \alpha_{31} & 0 & \alpha_{33} \end{bmatrix}$
3. Orthorhombic	$a \neq b \neq c$ $\alpha = \beta = \gamma = 90°$	3	$\begin{bmatrix} \alpha_{11} = \alpha_1 & 0 & 0 \\ 0 & \alpha_{22} = \alpha_2 & 0 \\ 0 & 0 & \alpha_{33} = \alpha_3 \end{bmatrix}$
4. Tetragonal	$a = b \neq c$ $\alpha = \beta = \gamma = 90°$	2	$\begin{bmatrix} \alpha_{11} = \alpha_1 & 0 & 0 \\ 0 & \alpha_{22} = \alpha_1 & 0 \\ 0 & 0 & \alpha_{33} = \alpha_3 \end{bmatrix}$
5a. Hexagonal (*Trigonal*)	$a = b = c$ $\alpha = \beta = \gamma \neq 90°$, $< 120°$		
5b. Hexagonal	$a = b \neq c$ $\alpha = \beta = 90°$, $\gamma = 120°$		
6. Cubic and Isotropic	$a = b = c$ $\alpha = \beta = \gamma = 90°$	1	$\begin{bmatrix} \alpha_{11} = \alpha & 0 & 0 \\ 0 & \alpha_{22} = \alpha & 0 \\ 0 & 0 & \alpha_{33} = \alpha \end{bmatrix}$

Figure 2.3: Bravais lattice crystal systems and the corresponding coefficient of thermal expansion tensor. Taken from Ref. [5].

2.1.3 Thermal Expansion Mismatch

A critical design challenge in the context of CTE is the thermal expansion mismatch. This phenomenon is intricately connected to the properties and behavior of various materials, especially in applications requiring careful thermal management across heterogeneous interfaces. The mismatch of the CTE between two dissimilar materials is found in many systems/applications related to composites, brazing, functionally graded materials, coatings, etc. [42, 43, 44].

One of the applications of interest in this thesis is the thermal or environmental barrier coatings (T/EBCs) applied to the ceramic matrix composites (CMCs) in aero-engine environments. From a materials engineering perspective, the integration of EBCs and CMCs represents a transformative technology, poised to revolutionize the manufacturing of ceramic-based gas turbines. This innovation is driven by the ever-increasing need for more energy-efficient and environmentally sustainable propulsion and energy generation systems [45, 46]. These advancements are crucial in meeting the pressing societal demands of the 21st century.

The primary objective of EBCs is to shield EBC/CMC systems from various modes of degradation, thereby extending the lifespan of EBC/CMC systems. One of the important requirements for EBCs is having low residual stresses that necessitates a suitable thermal expansion match with CMC [47, 48]. In T/EBC systems that undergo a temperature change above 1000°C, a CTE difference of merely 1 ppm/ °C can induce internal stresses reaching hundreds of megapascals at elevated temperatures [49].

EBC Stresses: The stress exerted by EBC (denoted as σ_{EBC}) on CMC substrates can be broken down into three components (Equation 2.4):

$$\sigma_{EBC} = \sigma_t + \sigma_a + \sigma_g \tag{2.4}$$

where, σ_t is the thermal mismatch stress (given in Equation 2.5), σ_a is the aging stress, and σ_g is the growth stress (Ref [50]).

$$\sigma_t = \frac{\left(\alpha_{\mathrm{EBC}} - \alpha_{\mathrm{CMC}}\right) E_{\mathrm{EBC}} \Delta T}{\left(1 - \nu_{\mathrm{EBC}}\right)} \tag{2.5}$$

where α_{EBC} and α_{CMC} are the CTE for EBC and the CMC substrate, respectively,

E_{EBC} is the Young's modulus of EBC, and ν_c is the Poisson's ratio of EBC [50].

The aging stress is a stress due to the changes in physical, mechanical, and chemical properties of EBC that are induced by thermal exposures. Factors causing these changes include oxidation, chemical reactions, phase transformations, and sintering. The growth stress is a stress that develops during the EBC deposition. There are other stresses not addressed in Equation 2.4 such as the thermal shock stress induced by temperature gradient.

The challenges posed by thermal expansion mismatch are not limited to EBCs but extend to various other applications where precise control over thermal properties is essential. Understanding and addressing thermal expansion mismatch is vital for the design and application of materials in various high-temperature environments, emphasizing the need for innovative materials and methods to alleviate the CTE mismatch and enhance the performance and longevity of these critical systems. It is essential to recognize that the considerations around thermal expansion mismatch often involve not just the volumetric CTE but also the tensor property of CTE. This complexity adds another layer to the challenge, especially when non-cubic compounds are being considered for specific applications, underscoring the need for a comprehensive understanding of anisotropy in CTE.

2.2 Prospective materials for EBC applications

For industrial use, EBC materials must fulfill certain requirements. These include (a) phase stability at high temperatures to avoid coating breakdown from phase transformation; (b) CTEs that match with Si-based ceramics or CMCs to mitigate risks of delamination or cracking due to CTE disparities; (c) chemical compatibility with the base material to overcome harmful reactions; (d) high resistance to hot-corrosion, particularly from water vapor and calcium-magnesium-alumino silicate (CMAS) glass deposits; and (e) low thermal conductivity for better temperature management of the substrate [51].

EBC technology has evolved through about three generations. The first-gen mullite/YSZ coatings fell out of favor due to CTE mismatch between YSZ and mullite, which led to cracking and diffusion routes [52, 53, 54, 55]. Second-gen Si/mullite-BSAS top coats have a tendency to detach during long-term high-temperature operation, likely because of glass phase formation [56, 57, 58, 59].

In contrast, rare-earth silicates nearly fulfill all the requirements for EBCs, gaining significant interest for their CTEs compatibility with Si-based ceramics, high-temperature phase stability, chemical compatibility, low thermal conductivity (especially the high entropy variants), and excellent resistance to hot corrosion by water vapor and CMAS molten salts [60, 61, 62, 63]. Therefore, rare-earth silicates are considered the most promising materials for topcoat applications in an integrated EBC structure with a Si bond layer. Unlike the previous two generations, its failure mainly stems from cracks or fractures caused by stresses, resulting from volume strain due to phase transition during corrosion, significant differences between CTEs of rare-earth silicates and penetrated CMAS glass, and the growth of thermally grown oxide in the Si bond layer. This thesis primarily only explores topics associated with the crystal chemistry of RE-Si-O. In the next section, we investigate the crystal chemistry of RE-Si-O and how its structure can be distinctly identified to comprehend the connection between CTE and structure.

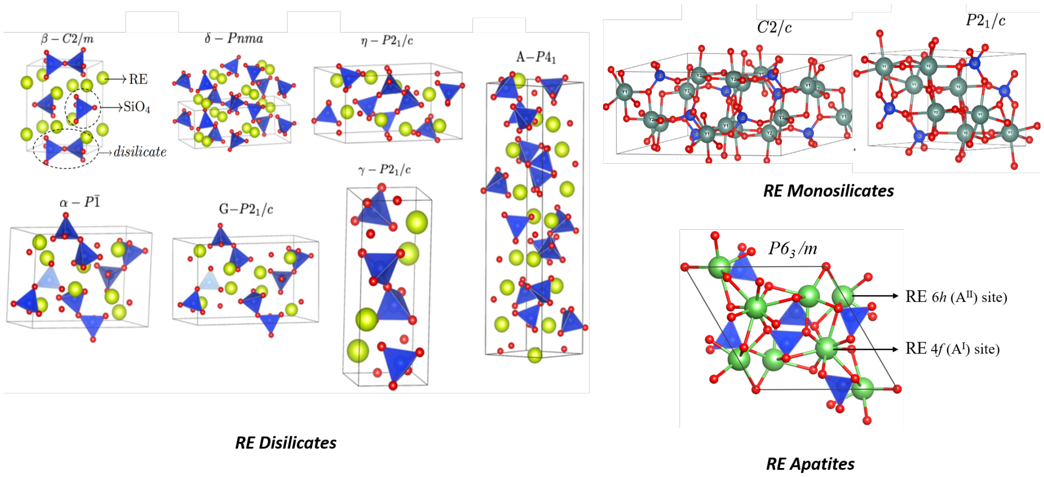


Figure 2.4: The crystal structures of some of the observed rare-earth disilicates (including their seven polymorphs), monosilicates (two polymorphs) and silicate apatite (hexagonal crystal structure).

2.3 RE-Si-O Crystal Chemistry and Descriptors

Some of the rare-earth silicates formed from RE_2O_3 and SiO_2, include RE_2SiO_5 monosilicates, $RE_2Si_2O_7$ disilicates, and $RE_{9.33}Si_6O_{26}$ silicate oxyapatites. The crystalline structure of these silicates varies based on the ratio of RE_2O_3 to SiO_2 as well as the size of the rare-earth cations. For a RE_2O_3:SiO_2 ratio of 1:1, two distinct crystal structures emerge, while seven polymorphs are observed for a ratio of 1:2. On the other hand, $RE_{9.33}Si_6O_{26}$ has only one crystal structure, irrespective of the type of rare-earth cation involved [64].

Some of the representative crystal structures are shown in Figure 2.4. The disilicates have seven known polymorphs: $C2/m$ (β), $Pnma$ (δ), $P2_1/c$ (η), $P\bar{1}$ (α), $P2_1/c$ (G), $P2_1/c$ (γ), and $P4_1$ (A). Wherein, CTEs of rare-earth disilicates with the same crystal structures are similar. The smallest and largest unit cells are made of 11 and 88 atoms, respectively. The smaller sized polymorphs β-$RE_2Si_2O_7$ and γ-$RE_2Si_2O_7$ have average linear CTEs close to that of the Si-based ceramics $(3.5 - 4.5 \times 10^{-6} K^{-1}$

in the temperature range of 300-1400K).

On the other hand, the monosilicates have been reported in one of the two low-symmetry monoclinic crystal structures in “X1” and “X2” [65, 66, 67, 68, 69].

The space group of X1-RE_2SiO_5 is $P2_1/c$, and it comprises rare-earth cations with the larger ionic radii found in La–Gd. Conversely, the space group of X2-RE_2SiO_5 is $C2/c$, which includes rare-earth cations with the smaller ionic radii of Tb–Lu. Notably, Tb is an exception, as it is known to exist in both $C2/c$ and $P2_1/c$. The coordination numbers of rare-earth cations are 7 and 9 in X1-RE_2SiO_5, as well as 6 and 7 in X2-RE_2SiO_5. The X2 and X1 unit cells have 64 and 32, respectively. The average linear CTE of rare-earth monosilicate is increasing when increasing temperature, and reach $6.94 - 8.84 \times 10^{-6} K^{-1}$ at 1473 K.

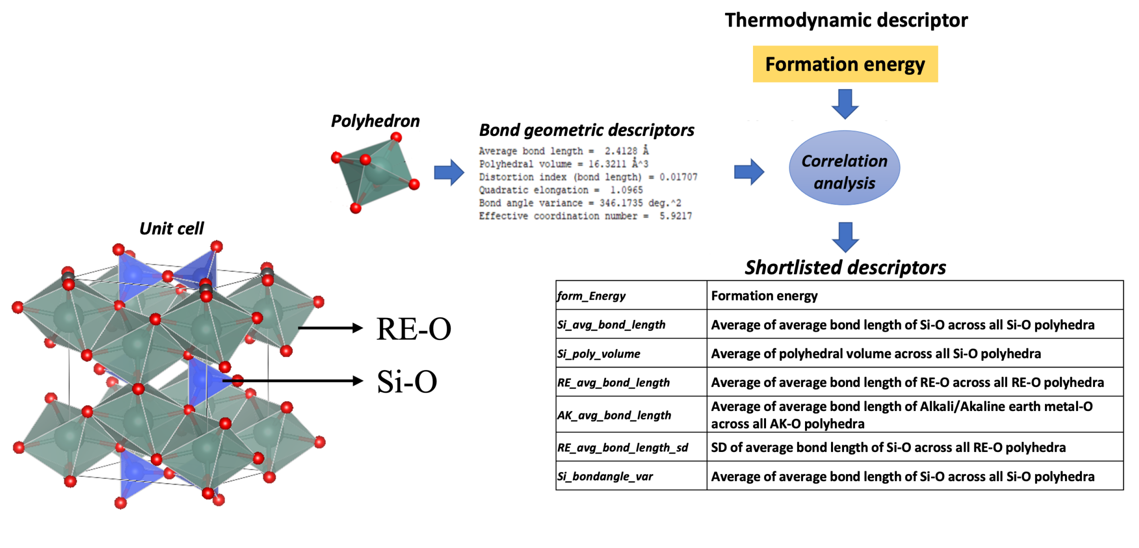
Both disilicates and monosilicates are nominally stoichiometric. The apatite silicate structure-type is non-stoichiometric ($RE_{9.33}Si_6O_{26}$) and have alkali metals ($RE_9A_1Si_6O_{26}$, where A=Li, Na, K, Rb, and Cs monovalent cations) and alkaline earth metals ($RE_8AE_2Si_6O_{26}$, where AE=Be, Mg, Ca, Sr, and Ba divalent cations) as cation substitutions in the lattice, which adds complexity. $RE_{9.33}Si_6O_{26}$) typically forms in the hexagonal $P6_3/m$ space group [70]. The unit cell contains 40-44 atoms, but supercells (120-126 atoms) are needed to accommodate the vacancy point defects in the RE-site. Usually, the CTEs of $RE_{9.33}Si_6O_{26}$ are much higher than those of Si-based ceramics, which may lead to its disuse as EBC materials [71, 72]. However, the outstanding resistance of $RE_{9.33}Si_6O_{26}$ to CMAS corrosion may attract attention on its potential of $RE_{9.33}Si_6O_{26}$ as a type of EBC material [73].

From the foregoing discussion, it is evident that the crystal structure plays a pivotal role in determining the CTE of rare-earth silicates. Even when different rare-earth cations are present, silicates with identical crystal structures tend to have comparable CTEs. The unifying structural feature across these silicates is the SiO_4 tetrahedral framework. These materials are made up of REO_x polyhedra and SiO_4 tetrahedra, with the coordination numbers for the rare-earth cations ranging between 6 and 9. Within these silicates, it is the thermal deformation resistance of the REO_x polyhedra that primarily dictates the CTEs [74, 75, 76].

In light of the key role that crystal structure and REO_x polyhedra play in determining the CTE of rare-earth silicates, polyhedral descriptors emerge as valuable tools for a more nuanced understanding [17]. Descriptors based on polyhedra can serve as mathematical or geometrical constructs that characterize the shape, size, and symmetry of a polyhedron encompassing a central atom or ion in a crystal or molecule. These descriptors quantitatively portray the local environment that comprises (REO_x) polyhedra and (SiO_4) tetrahedra, offering insights into how this environment could modulate the CTE of the compounds [77]. Their universal applicability ensures they can be leveraged across diverse materials, ranging from inorganic compounds to intricate molecular systems. Commonly used polyhedral descriptors include the average bond length, polyhedral volume, effective coordination number, distortion index, which gauges the deviation of a polyhedron from a regular shape, quadratic elongation, which measures the polyhedron's elongation or compression, and bond angle variance, which assesses the variability in bond angles within the polyhedron [78].

While RE ionic radii remain the most commonly used descriptor for understanding the CTE of materials in the RE-Si-O chemical space, it operates on a fairly straightforward principle. The strength of the RE metal-oxide bond diminishes as the ionic radius increases for a given charged ion and coordination number. This leads to an increase in the thermal expansion of disilicates as the radius of the RE cation increases. Although intuitive, this descriptor may lack the depth needed to understand the underlying mechanistic insight affecting CTE. Another alternative lies in descriptors based on unit cell parameters, such as volume and the lattice constants a, b, c along with the angles between them. These parameters are often highly correlated with ionic radii and offer the advantages of simplicity and accessibility. However, like ionic radii, these descriptors may also lack the capability to provide mechanistic insights. This brings us to the unique value of polyhedral descriptors. Unlike ionic radii and unit cell parameters, polyhedral descriptors can capture the intricate details of the local atomic environment, including the Si-O-Si bond angles and anharmonicity of lattice vibrations, both of which directly influence CTE. This provides a level of mechanistic understanding that is not easily achievable with simpler descriptors. Therefore, incorporating polyhedral descriptors into our analytical toolbox is worthwhile. By comparing these with simpler, more accessible descriptors like ionic radii, we can discern whether polyhedral descriptors offer additional, perhaps more nuanced, information that is not captured by ionic radii. This comparative approach is instrumental for deepening our understanding of the structure-CTE relationship in these complex materials. In this context, the upcoming discussion will focus on developing a set of polyhedral descriptors tailored for the RE-Si-O chemical space. The aim is to devise descriptors that offer not just mechanistic information correlating with CTE but also a level of universality. This universality is crucial for a comprehensive description of CTE across all three classes of materials in the RE-Si-O system, namely disilicates, monosilicates, and apatites.

Figure 2.5 shows a snapshot of a typical DFT optimized structure, where I have illustrated the descriptor generation logic. I divided the crystal structure into two subunits based on (1) RE-O polyhedra and (2) Si-O polyhedra. Using VESTA [79] (a ver-



```
Average bond length =  2.4128 Å
Polyhedral volume = 16.3211 Å^3
Distortion index (bond length) = 0.01707
Quadratic elongation =  1.0965
Bond angle variance = 346.1735 deg.^2
Effective coordination number =  5.9217
```

form_Energy	Formation energy
Si_avg_bond_length	Average of average bond length of Si-O across all Si-O polyhedra
Si_poly_volume	Average of polyhedral volume across all Si-O polyhedra
RE_avg_bond_length	Average of average bond length of RE-O across all RE-O polyhedra
AK_avg_bond_length	Average of average bond length of Alkali/Akaline earth metal-O across all AK-O polyhedra
RE_avg_bond_length_sd	SD of average bond length of Si-O across all RE-O polyhedra
Si_bondangle_var	Average of average bond length of Si-O across all Si-O polyhedra

Figure 2.5: Schematic showing the logic for descriptor generation. The RE-O and Si-O polyhedral units are highlighted. The average and standard deviation (SD) of each polyhedral attribute extracted from VESTA for RE-O and Si-O polyhedra are extracted for each compound. Along with formation energy we have an initial set of 25 descriptors. Using correlation analysis, seven descriptors (Si_avg_bond_length, Si_poly_volume, RE_avg_bond_length, AK_avg_bond_length, RE_avg_bond_length_sd, Si_bondangle_var) and form_Energy are shortlisted.

satile software tool designed for visualizing and analyzing crystallographic structure data), I generated the following six polyhedral descriptors for each unique polyhedron in the crystal structure: (1) average bond length, (2) polyhedral volume, (3) effective coordination number, (4) distortion index: $D = \frac{1}{n}\sum_{i=1}^{n}\frac{|l_i - l_{\text{av}}|}{l_{\text{av}}}$, where l_i is the distance from the central atom to the ith coordination atom and l_{av} is the average bond length, (5) bond angle variance (only for Si-O polyhedra): $\sigma^2 = \frac{1}{m-1}\sum_{i=1}^{m}(\phi_i - \phi_0)^2$, where m is (number of faces in the polyhedron) × $\frac{3}{2}$ (i.e., number of bond angles), ϕ_i is the i^{th} bond angle, and ϕ_0 is the ideal bond angle for a regular polyhedron (for example, 90° for an octahedron or $109°28'$ for a tetrahedron) and (6) quadratic elongation (only for Si-O polyhedra): $\langle\lambda\rangle = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{l_i}{l_0}\right)^2$, where l_0 is the center-to-vertex distance of a regular polyhedron of the same volume. $\langle\lambda\rangle$ is dimensionless, giving a quantitative measure of polyhedral distortion which is independent of the effective size of the polyhedron. I note that these polyhedral descriptors are default in VESTA and have been used before in the literature to study crystal chemistry trends [80, 81, 82, 83]. Given that there are multiple polyhedral units in each polymorph, I employed the mean and standard deviation of these descriptors across the different, distinct RE-O and Si-O polyhedral entities for descriptor development. This approach results in a standardized representation in which every polymorph, regardless of its complex RE-O or Si-O coordination configurations, can be articulated using a set of 24 descriptors. This set comprises 12 descriptors based on the Si polyhedron (average and standard deviation of six polyhedral descriptors), 8 related to the RE polyhedron (average and standard deviation of four polyhedral descriptors), four for the alkali/alkaline earth metal polyhedron (average of four polyhedral descriptors, as these do not possess more than one unique polyhedron). The complete array of these descriptors, along with their individual explanations, is presented in Table 2.1.

Table 2.1: List of polyhedral descriptors explored in this thesis along with their descriptions.

Variable	**Description**
Si_avg_bond_length	Average of average bond length of Si-O across all Si-O polyhedra
Si_poly_volume	Average of polyhedral volume across all Si-O polyhedra
Si_distortion	Average of polyhedral distortion across all Si-O polyhedra
Si_eff_coord_num	Average of effective coordination number across all Si-O polyhedra
Si_quad_elong	Average of quadratic elongation across all Si-O polyhedra

Si_bondangle_var	Average of variance in bond angles across all Si-O polyhedra
RE_avg_bond_length	Average of average bond length of RE-O across all RE-O polyhedra
RE_poly_volume	Average of polyhedral volume across all RE-O polyhedra
RE_distortion	Average of polyhedral distortion across all RE-O polyhedra
RE_eff_coord_num	Average of effective coordination number across all RE-O polyhedra
AK_avg_bond_length	Average bond length of Alkali/Alkaline Earth-O across all Alkali/Alkaline Earth-O polyhedra
AK_poly_volume	Average of polyhedral volume across all Alkali/Alkaline Earth-O polyhedra
AK_distortion	Average of polyhedral distortion across all Alkali/Alkaline Earth-O polyhedra
AK_eff_coord_num	Average of effective coordination number across all Alkali/Alkaline Earth-O polyhedra
Si_avg_bond_length_sd	SD of average bond length of Si-O across all Si-O polyhedra
Si_poly_volume_sd	SD of polyhedral volume across all Si-O polyhedra
Si_distortion_sd	SD of polyhedral distortion across all Si-O polyhedra
Si_eff_coord_num_sd	SD of effective coordination number across all Si-O polyhedra
Si_quad_elong_sd	SD of quadratic elongation across all Si-O polyhedra

Si_bondangle_var_sd	SD of variance in bond angles across all Si-O polyhedra
RE_avg_bond_length_sd	SD of average bond length of RE-O across all RE-O polyhedra
RE_poly_volume_sd	SD of polyhedral volume across all RE-O polyhedra
RE_distortion_sd	SD of polyhedral distortion across all RE-O polyhedra
RE_eff_coord_num_sd	SD of effective coordination number across all RE-O polyhedra

One other useful descriptor would be lattice energy because various research studies have explored its impact in predicting the CTE of materials [84, 85]. For instance, a study by Zhang *et al.* [84] employed lattice energy along with polyhedral descriptors to semi-empirically describe the CTE. Lattice energy measures the strength of the electrostatic bonds within a crystal, essentially serving as a yardstick for the thermodynamic stability. High lattice energy is often correlated with a resistance to structural changes, including thermal expansion. While lattice energy provides valuable insights, its calculation can be resource-intensive, particularly for complex materials. Formation energy, on the other hand, represents the total energy change that occurs when a compound is formed from its elemental constituents. This metric can be obtained from DFT calculations and offers a glimpse into the thermodynamic stability of a compound. Thus, I hypothesize that combining formation energy with local polyhedral descriptors (comprising 24 descriptors) offer a coherent framework for investigating the relationship between structure and CTE in crystals that form in the RE-Si-O phase space. This gives us an initial set of 25 descriptors. However, based on the correlation analysis detailed in the Appendix (section C.1), only 7 descriptors are required to convey the same information without redundancy. They include: Si_avg_bond_length, Si_poly_volume, RE_avg_bond_length, AK_avg_bond_length, RE_avg_bond_length_sd, Si_bondangle_var and form_Energy. Exploring structure-property relationships through the various descriptors mentioned in this chapter is essential for understanding both their commonalities and differences. Therefore, our forthcoming investigation into the crystal chemistry of RE-Si-O will be based on a diverse set of descriptors including ionic radii, formation energy, polyhedral features, and unit cell parameters. Before diving into that, the next chapter will provide an in-depth look at the methodologies that I have explored in this thesis.

Chapter 3

Methods

3.1 An Overview of Density Functional Theory

Density functional theory (DFT) offers a practical approach to solving the Schrödinger's equation, enabling the calculation of material properties from fundamental principles. Rooted in the theories of Hohenberg, Kohn, Sham, Thomas, and Fermi, DFT asserts that the total electronic energy of a system (E_{tot}) can be expressed as a functional of the system's electronic charge density (ρ). This density is computed as the sum of Kohn-Sham orbitals (ϕ), as described in Equation 3.1.

$$\rho(r) = \sum_{i}^{N} |\phi(r)|^2 \tag{3.1}$$

The Kohn-Sham version of DFT calculates the energy for a given ϕ using the eigenvalue equation in Equation 3.2, known as the Kohn-Sham equation. This equation consists of several components: T_s, the Kohn-Sham kinetic energy, given by Equation 3.3; E_H, the Hartree energy, as in Equation 3.4; v_{ext}, the external potential; and E_{xc}, the electron-electron exchange correlation term [86, 87].

$$\epsilon_i \phi_i(r) = [T_s + E_H + v_{ext} + E_{xc}]\phi(r) \tag{3.2}$$

$$T_s[\rho(r)] = \sum_{i=1}^{N} \int \rho^*(r)(-\frac{\hbar^2}{2m}\nabla^2)\rho(r) \tag{3.3}$$

$$E_H[\rho(r)] = \frac{e^2}{2} \int dr \int dr' \frac{\rho(r)\rho(r')}{|r-r'|} \tag{3.4}$$

The term E_{xc} is particularly challenging to solve directly within the Kohn-Sham framework and requires numerical approximations. Despite this complexity, many approximations yield results close to experimental values. However, E_{xc} can still be a significant source of error in DFT, and selecting the correct numerical approximation is crucial.

Approximations for E_{xc} are generally categorized into LDA and GGA. LDA, as expressed in Equation 3.5, assumes that E_{xc} depends solely on ρ at each spatial point. In contrast, GGA includes the first derivative of ρ, as shown in Equation 3.6. Further refinements, such as meta-GGA functionals and Hybrid functionals, can also be employed.

$$E_{XC}^{LDA} = \int \epsilon_{xc}(\rho(r))\rho(r)d^3r \tag{3.5}$$

$$E_{XC}^{GGA} = \int \epsilon_{xc}(\rho(r), \nabla\rho(r))\rho(r)d^3r \tag{3.6}$$

The choice of a basis set for the wave function is another essential aspect of DFT calculations. For solid materials, Bloch planewaves are commonly used due to their parameterizable nature and ability to represent the periodicity of solids. Other options, such as Slater type orbitals and Gaussian basis sets, are more suited for molecular calculations.

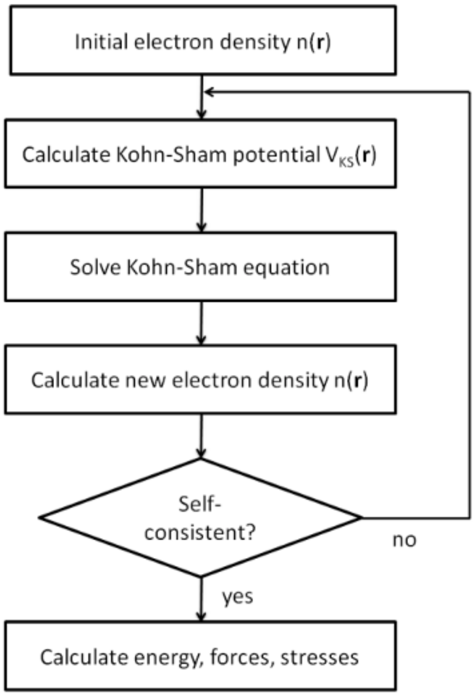

Initial electron density n(**r**)
Calculate Kohn-Sham potential V_{KS}(**r**)
Solve Kohn-Sham equation
Calculate new electron density n(**r**)
Self-consistent?
no
yes
Calculate energy, forces, stresses

Figure 3.1: Scheme of the self-consistent solution of the Kohn-Sham equations. [86, 87]

The Kohn-Sham (KS) equations are solved through a self-consistent iterative process, as depicted in Figure 3.1. The procedure begins with an initial or trial value for the electron density $n(r)$. With a given exchange-correlation (XC) functional, the KS potential $V_{KS}(r)$ is derived from this initial density. Solving the KS equations then produces the KS orbitals $\phi_i(r)$, leading to a new electron density $n(r)$ and a corresponding total energy. This newly computed density is then used as the input for the next iteration. The process continues, repeating these steps until self-consistency is achieved. In practical terms, self-consistency is determined by the convergence of the total energy. When consecutive total energy values differ by less than a chosen convergence criterion, the calculation is considered complete, and the final total energy, forces, stresses, and other properties can be reported.

3.1.1 Structure Optimization

Optimizing structures using DFT is a well-known approach in computational materials science. It involves finding the equilibrium geometry of a system that minimizes the total energy. The Hellmann-Feynman theorem plays a crucial role in this optimization, providing a theoretical foundation for calculating forces and guiding the system to its minimum energy configuration. The Hellmann-Feynman theorem states that the force on a nucleus is equal to the negative gradient of the total energy with respect to the nuclear coordinates. Mathematically, it can be expressed as [86, 87]:

$$F_i = -\frac{\partial E}{\partial R_i} \tag{3.7}$$

where F_i is the force on the i-th nucleus, E is the total energy, and R_i is the position of the i-th nucleus. This theorem simplifies the calculation of forces in a system by relating them directly to the derivative of the energy. It allows for efficient computation of forces without explicitly considering the wave function's response to nuclear motion. By iteratively calculating total energy and forces and updating the geometry, the system is guided to its equilibrium structure. The Hellmann-Feynman theorem's elegant relationship between forces and energy derivatives streamlines this process, enabling efficient and accurate optimization. The steps involved are:

Initial Structure and Parameters

The optimization begins with an initial guess for the structure, including atomic positions and lattice parameters. This initial structure is often derived from experimental data or previous calculations.

Total Energy Calculation

The total energy of the initial structure is calculated using DFT. This involves solving the Kohn-Sham equations self-consistently to obtain the electronic structure and total energy.

Force Calculation

Using the Hellmann-Feynman theorem, the forces on the nuclei are calculated from the derivative of the total energy with respect to the nuclear positions. These forces indicate how the atoms should move to reduce the total energy.

Geometry Update

The atomic positions are updated based on the calculated forces, typically using optimization algorithms like the steepest descent or conjugate gradient methods. The updated geometry is expected to be closer to the equilibrium structure.

Convergence Check

The optimization process is iteratively repeated, recalculating the total energy and forces and updating the geometry until convergence criteria are met. Common convergence criteria include a threshold for the maximum force on any atom and a tolerance for changes in total energy. Once convergence is achieved, the final optimized structure represents the equilibrium geometry of the system, corresponding to the minimum total energy.

DFT calculations not only provide the optimized structure of a material but also yield the total energy for that structure. This total energy is a critical quantity that can be used to explore the potential energy landscape and understand the energetic competition among various possible structures a compound can form. By calculating the energy difference ΔE (in meV/atom) with respect to the lowest energy structure, one can gain insights into the stability and likelihood of different structural configurations.

The ground state structure of a compound is the configuration with the lowest total energy. In terms of ΔE, the ground state structure has ΔE= 0 meV/atom, indicating that it is the most stable and energetically favorable configuration. This structure represents the natural state of the compound under given conditions and serves as a reference point for evaluating other possible structures.

Structures with a ΔE greater than 5 meV/atom are generally considered unlikely to form [86, 87]. The higher ΔE indicates that these configurations are energetically unfavorable compared to the ground state and metastable structures. While not entirely ruled out, these higher ΔE structures are less likely to be observed in practice, and their formation may require specific non-equilibrium conditions or external influences [88].

3.1.2 Density of States Calculation

The density of states (DOS) is a key descriptor of chemical bonding and contains information related to its electronic properties. The DOS (measure of how many states are available for occupation at each energy level) signifies the electronic structure by illustrating the likelihood of finding a state at a specific energy level, with each state having the capacity to be occupied by two electrons. The quantum mechanical nature of electrons means that there is a probability associated with an electron existing at a given energy, resulting in a spectral curve with intricate features. The DOS serves as a visual representation of the electronic structure, where any alteration in the DOS corresponds to a change in the available states for an electron. Consequently, this shift is indicative of variations in the electronic structure and, presumably, electronic properties. While the DOS curve encapsulates a statistical distribution of electronic states, deciphering the factors and their interplay that constitute this distribution is challenging. The complexity of these interactions leads to a DOS curve that is often difficult to interpret. Since the DOS symbolizes the electronic structure, scrutinizing the changes in DOS due to variations in crystal structure and chemistry contributes to the formulation of enhanced relationships between electronic structure, crystal structure, and chemistry.

The calculation begins by solving the Kohn-Sham equations to obtain the eigenvalues, which represent the energy levels of the system [86, 87]. The energy levels are then grouped into bins or intervals, and the number of states in each bin is counted. To obtain a smooth DOS curve, a smearing or broadening technique (e.g., Gaussian smearing) is often applied to the binned data. The DOS is typically normalized to ensure that the integral over all energies equals the total number of states. The partial density of states (PDOS) provides insight into the contribution of specific atoms or orbitals to the overall DOS. It is essential for understanding the local electronic structure. The PDOS is calculated by projecting the Kohn-Sham wave functions onto a set of localized atomic orbitals. This allows for the decomposition of the DOS into contributions from specific atoms or orbitals. By analyzing the PDOS, one can understand how different atomic species or orbital characters contribute to the electronic structure, bonding, and other properties.

3.1.3 Formation Energy Calculation

Formation energy is a critical parameter in understanding the stability and properties of materials, particularly in the context of defects, alloys, and novel compounds. In this section, I outline the methodology employed to calculate the formation energy using Density Functional Theory (DFT) calculations. The first step in calculating formation energy is to determine the total energy of the system. This involves solving the Kohn-Sham equations for the system of interest, considering both the electronic and ionic contributions. The reference state energy is calculated for each constituent element in its most stable form (e.g., bulk phase). This involves performing separate DFT calculations for each element in a standardized condition, such as zero pressure and temperature.

The formation energy (E_f) of a compound is calculated using the following formula:

$$E_\mathrm{f} = E_\mathrm{total} - \sum_i n_i E_i^\mathrm{ref}$$

where E_total is the total energy of the compound, n_i is the number of atoms of element i in the compound, and E_i^ref is the energy of element i in its reference state.

3.1.4 Special Quasirandom Structures (SQS)

Modeling disordered systems using conventional supercells can be computationally expensive and may not accurately capture the random nature of the atom distribution. The SQS method overcomes these challenges by generating a structure that statistically represents the randomness of the disordered system [89]. The SQS method allows for the simulation of random structures by creating a representative structure that mimics the statistical properties of a truly random system. The SQS method has been widely used in DFT calculations for various applications alloy modeling, defect studies and calculation of thermodynamic properties such as free energy by accurately representing configurational entropy. In this thesis, I use SQS for optimizing structures of multi-component rare-earth disilicate compounds which are used for validating the holistic ML model.

The SQS method is designed to emulate the correlation functions of a random alloy up to a specific range. This means that the SQS captures the essential statistical properties of the random system, such as pair correlations and higher-order correlations, within a defined range. This statistical representation ensures that the SQS behaves similarly to the random system it represents, providing a reliable model for computational studies. One of the key advantages of the SQS method is its ability to represent a random structure using a relatively small supercell. Unlike conventional methods that may require large supercells to capture randomness, the SQS method efficiently models the disordered system without the need for extensive computational resources. This size efficiency not only reduces computational cost but also makes the method more accessible for complex simulations. Moreover, an accurate representation of the statistical properties of a random system enables the calculation of configurational averages without the need for extensive sampling. By capturing the essential statistical behavior of the system, the SQS allows for meaningful averages to be computed, providing insights into the system's overall behavior. This ability to perform configurational averaging streamlines the study of disordered systems, making the SQS method a valuable tool in computational materials science.

The construction of an SQS begins with defining the target correlation functions for the random system. These functions are based on the desired composition and structural properties of the system and serve as the reference for constructing the SQS. By carefully defining these target functions, the SQS can be tailored to represent the specific random system under study. Once the target correlation functions are defined, various candidate structures are generated by randomly distributing atoms within a supercell. These candidate structures represent different possible realizations of the random system and serve as the starting point for finding the optimal SQS. The generation of diverse candidate structures ensures that the optimization process explores a wide range of possibilities. The candidate structures are then evaluated based on how closely their correlation functions match the target correlation functions. Optimization algorithms are employed to find the structure that best represents the random system. This optimization process involves iteratively adjusting the candidate structures and evaluating their correlation functions until the best match is found. The result is an SQS that accurately captures the statistical properties of the random system. The final step in constructing an SQS is validation. The selected SQS is rigorously validated to ensure that it accurately represents the statistical properties of the random system. This validation process may involve comparing the SQS's correlation functions with those of the target system and assessing how well the SQS reproduces the system's behavior. Successful validation confirms that the SQS is a reliable model for the random system, ready for use in subsequent computational studies.

3.2 Machine Learning

In this section, I initially explore some of the most commonly used ML algorithms and approaches in the realm of material science. I will be using a definition of ML that is both accessible and provided by Tim Mitchell in his book "Machine Learning" [90]. Mitchell states: "*A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E*" This definition lays the basis for our subsequent discussion on ML. Moreover, I will pinpoint areas of improvement specifically in the application of ML to material science and suggest strategies for enhancement.

3.2.1 The Experience

At the heart of this definition lies the concept of “experience”. In the context of ML, experience is a synthesis of knowledge and datasets. This knowledge is often patterns or representations the model derives from the data it is exposed to. This experiential learning can be broadly categorized into two primary ML paradigms: supervised and unsupervised learning based on the nature of the examples they receive during the training period. Unsupervised learning uses a dataset of examples with only features (quantifiable characteristics of the subject under scrutiny) denoted as x, and it learns some meaningful relationships among the examples, such as the probability distribution $p(x)$. Conversely, supervised learning employs a dataset of examples with both features x and corresponding labels y, which are the “appropriate” values associated to the features. In this scenario, the learning algorithm determines the probability distribution $p(y|x)$ or the expected outcome of a regression $E(y|x)$.

3.2.2 The Task

ML can be utilized for a variety of tasks. These tasks typically involve the manipulation of examples (or data points) provided to the algorithm. These examples consist of variables/features/descriptors (such as chemical structures, chemical compositions, reaction temperature, etc.) which are typically compiled into a vector x to illustrate a material’s physiochemical properties, structural attributes, composition properties, or synthesis process conditions. The subsequent sections will elucidate various algorithms, primarily falling under two broad categories: supervised and unsupervised ML. While there have been extensive efforts invested in the continuous search for a “better” model, there is no universally superior algorithm for all problems [91].

Therefore, the aim is not to find a universally superior algorithm but to identify an algorithm that is best suited for a specific problem. With this perspective, it becomes paramount to experiment with a variety of algorithms tailored to the unique facets of the dataset and problem in question. Through meticulous evaluation and testing, I identified specific algorithms that excelled with our datasets. The subsequent sections will focus on these algorithms. Armed with this foundational understanding of these algorithms, readers will be better positioned to appreciate the depth of our investigation in the subsequent chapters.

Classification

The purpose of an algorithm is to identify the category or class to which a specific observation belongs. This is achieved by learning a function $F : \mathbb{R}^n \longrightarrow \{1, \dots, m\}$ that maps the feature vector x to one specific class among m distinct classes. Instead of settling on a single class, the function can also provide a probability distribution across all classes, where each entry in the output vector $y = f(x)$ represents the likelihood that the example belongs to a particular class. ML algorithms have been effectively used to tackle materials classification tasks. For instance, given a set of synthesis conditions, a classification model can predict whether synthesized materials will form successfully or which sections of synthesized materials will exhibit defects.

Decision trees Decision trees (DTs) are models based on sequential logic. A thorough review of DTs is given by Murthy [92], and recent advancements with various algorithms are also available [93]. DTs offer a series of logical rules that examine an object's features. Within a DT, each node symbolizes the feature being classified, and the branches descending from the parent node signify the possible values that the parent node's feature can assume. The parent node is then recursively divided into child nodes, each testing their features, until a stopping condition is reached, such as reaching the tree's maximum depth. However, DTs are susceptible to overfitting. If a tree is deep and each child node contains only a small amount of data, the DT model's generalizability will be weak. Extensive research in DT has focused on finding efficient algorithms that also reduce the tree's size to avoid overfitting [93]. Another method to prevent overfitting is to employ ensemble techniques, such as Random Forest (RF) [94] and Gradient Boosting (GB) [95]. These methods make predictions by combining the outputs of separately trained DTs, differing only in how the collection of DTs is assembled. Random Forest trains each DT independently using a random subset of the original data, obtained through bagging, where random samples are drawn with replacement. To ease computational complexity, each DT can only select a random subset of the entire feature set. An RF model makes a prediction on an input vector by averaging individual models (regression) or choosing the class with the majority votes (classification). Conversely, Gradient Boosting constructs base-learners (individual models like DT or SVM that comprise the ensemble) sequentially [95]. At each step, the next model's parameter is selected to correlate most with the negative gradient of the loss function, enhancing the ensemble's performance with each additional base-learner. It has been demonstrated that Gradient Boosting will consistently outperform Random Forest if the parameters are meticulously adjusted [93]. However, GB is generally slower than RF and more susceptible to noisy data. One of the fundamental advantages of DT is its interpretability.

Regression

Another typical task is regression, where the algorithm strives to learn a function $F : \mathbb{R}^n \longrightarrow \mathbb{R}$ that determines a continuous value y or a group of continuous values

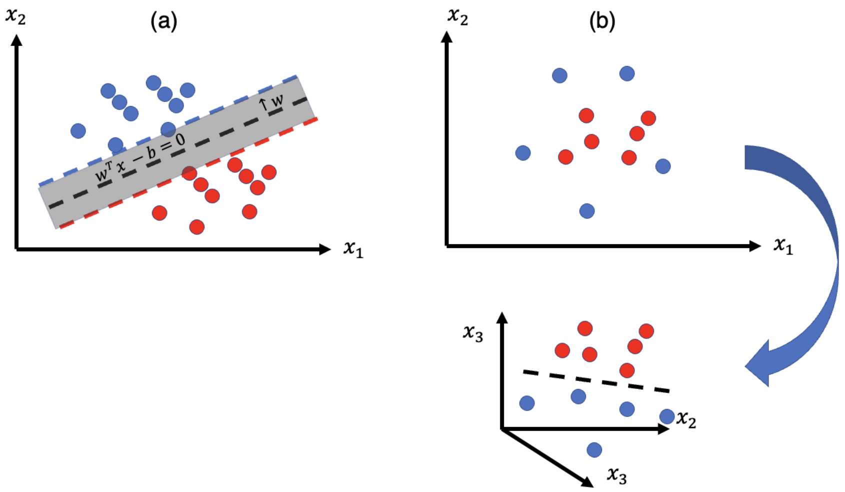


Figure 3.2: Support Vector Machines for Classification. Individual points are colored according to their classification. Dashed red and blue lines are the support vectors with the black line being the optimized decision boundary.

represented as a vector $\mathbf{y}$.

Support Vector Machines SVM are a popular method for nonlinear classification problems. They are perhaps easier to understand in the context of classification where hyperplanes function as decision boundaries making the problem and solution easier to visualize. For clarity I will start by introducing the classification methodology and then transfer that logic to the regression problem.

Given a data set of $(x_1, x_2, y)_1, ..., (x_1, x_2, y)_N$, where y is either -1 or 1 we would like to identify a decision boundary which is maximally distant from both $y = 1$ and $y = -1$. This decision boundary is referred to as a hyperplane and satisfies the relationship $w^T x - b = 0$, where w is normal to the hyperplane. We would like this hyperplane to maximize the distance from the nearest point $(x_1, x_2, y)_i$. If the points are linearly separable, as they are in Figure 3.2 (a), this can be accomplished by first selecting 2 initial hyperplanes, beyond which y takes on the value of either 1 or -1, which are maximally distant from each other. These hyperplanes are termed support

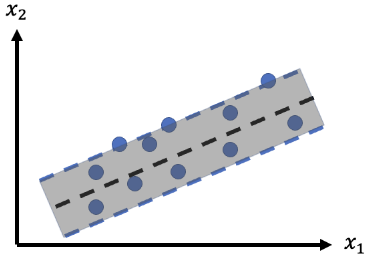


Figure 3.3: Support Vector Regression. Here there is no classification, blue lines represent the support vectors used to identify the black line or optimized regression line.

vectors. taking the parallel hyperplane that bisects the distance between these two support vectors gives us the desired decision boundary.

$$\langle x_i, x_j \rangle = exp(\frac{||x_i - x_j||^2}{2\sigma^2}) \tag{3.8}$$

To treat problems where the data is not clearly, similar to what is shown in Figure 3.2 (b), we can attempt to transform the data into higher dimensionality with the hope that it is separable there [96]. A visualization of this methodology is shown Figure 3.2 (b). It is computationally expensive to map individual observations into higher dimensional space, instead we can describe the relationships between the data though some nonlinear kernel (most commonly a radial basis function shown in Equation 3.8) and map this to a higher dimensional space. The decision boundary can then be identified based on the kernel description and mapped back to the original problem space.

In SVR we still rely on hyperplanes defined by mapping the problem to a higher dimension through the kernel trick. The difference is, instead of trying to maximize the difference between two support vectors, we are trying to maximize the number of points which fall inside support vectors. This is shown schematically in Figure 3.3.

The distance between the decision boundary and the support vectors (the other hyperplanes) is referred to as ϵ and is a hyperparameter of SVM. Only points which fall inside of this window will be considered during the regression, which has the effect of making SVM less susceptible to outliers relative to other nonlinear modeling techniques [97].

Clustering

This is a type of unsupervised learning task that is beneficial when dealing with a large volume of unlabeled data. The objective is to group data points into clusters where items within the same cluster are more "similar" to each other compared to those in another cluster. The definition of "similar" depends on the context and requirements. By grouping items into clusters, one can derive meaningful insights from the data even when the dataset is unlabeled. Various clustering methods exist, each designed for specific purposes. The two most prevalent forms of clustering are partitioning clustering and hierarchical clustering.

Partitioning clustering is focused on dividing the items into k clusters [37]. Crisp clustering ensures that each item is a member of only one class, with an output of 0,1, while fuzzy clustering permits items to belong to clusters to varying degrees, with values in the range [0,1]. Standard methods of partitioning clustering encompass k-means clustering and k-medians clustering.

Hierarchical clustering, on the other hand, is designed to illustrate how clusters are interconnected through tree-like structures [37]. The tree can be constructed downwards, breaking the top cluster into progressively smaller clusters (using divisive algorithms), or upwards by starting with numerous small clusters and merging them (using agglomerative algorithms). Unlike partitioning clustering, which provides one set of clusters for a single value of k, hierarchical clustering offers clusters for a range of values of k. This allows hierarchical clustering to reveal how smaller clusters relate to larger ones.

Dimension Reduction & Visualization

When developing a model, there may be instances where the number of features is overwhelming. In such cases, it's advantageous to decrease the dimension of the features by transforming them into a lower dimension, while retaining as much information as possible. This can enhance computational efficiency, potentially improve model performance, prevent overfitting, and aid in uncovering insights for specific tasks. Dimension reduction has been employed in material discovery to optimize prediction results, such as condensing extensive long-time dynamic information into lower dimension data for superior performance. Furthermore, by minimizing less relevant features, dimension reduction can help reveal the fundamental physics/chemistry of a material model. Additionally, the conversion of high-dimensional data into 2D or 3D plots for visualization is a crucial aspect of dimension reduction. This allows for valuable insights to be gleaned from an understandable plot and has been used in material discovery to visualize the high-dimensional material design space.

In essence, determining the task to be solved is the initial step in effectively applying ML in material discovery. However, the task categories discussed so far are not always standalone. For instance, efficient searching can be integrated with classification or regression to create a closed design loop for finding high-temperature ferroelectric perovskites, and dimension reduction can be utilized to build a superior feature set for training the models [98].

t-Distributed Stochastic Neighbor Embedding (t-SNE) This is a widely-used unsupervised learning technique for dimension reduction and data visualization, consisting of two stages [99]. First, in the original high-dimensional space, t-SNE constructs a probability distribution to determine how likely two pairs of points would be chosen. Points that are nearer to each other (for example, with a smaller Euclidean norm) are assigned higher probabilities, and vice versa. Second, in the lower-dimensional space, a probability distribution is defined over all points. The algorithm's goal is to minimize the Kullback-Leibler divergence, ensuring that the probability distributions in both the high and low-dimensional spaces are similar to each other. This process enables the mapping of points from a high-dimensional space to a lower one.

3.2.3 The Performance Measure

This is employed to evaluate an algorithm's effectiveness on a specific task. For classification, the algorithm's performance can be gauged based on accuracy (percentage of correct output), error rate (percentage of incorrect output), or a more intricate metric derived from the confusion matrix. For algorithms that produce a probability distribution, the log-probability can be computed. For regression, it's common to use mean squared errors (MSE), mean absolute errors (MSE) or some forms of norm errors.

The dataset for a specific task is usually divided into three parts: training dataset, validation dataset, and test dataset. The algorithm is trained on the training dataset and further optimized (by adjusting hyperparameters such as learning rate and structure of the algorithm model) based on the performance on the validation dataset. The performance of the optimized supervised learning algorithm is typically measured by its ability to perform on the test dataset, known as the test performance, which serves as an indicator of the model's generalization capability.

So far in our exploration, we have delved into the elements and parts of conventional ML which as per Mitchel's definition encapsulates algorithms designed to adapt and improve over time based on experience. However, looking ahead, my focus shifts towards transcending the limitations of conventional approaches. In the following two sections, I discuss the innovative tools and methodologies that enable us to go beyond the mere performance measure "P". By incorporating elements that facilitate a deeper understanding of where and why a model fails, I aim to construct more robust and insightful models. This includes the application of uncertainty quantification, which provides insights into the reliability of predictions, and the utilization of Explainable Artificial Intelligence (XAI) techniques, which shed light on the underlying reasons for a model's behavior. Together, these advanced approaches promise to enhance our ability to build, interpret, and trust ML models, taking us a step closer to fully realizing the potential of this dynamic field.

3.2.4 Uncertainty Quantification

A vast majority of the materials informatics literature only deals with the usage of the average performance metrics of trained ML models such as the mean average error (MAE), the mean square error (MSE), or the root-mean-square error (RMSE) as the uncertainty metric to assess the quality of the trained regression models [100, 30]. These average estimates are insufficient to determine uncertainties associated with individual instances. For example, when applying an ML model for predicting the property of an unexplored material, we want to know how certain the model is

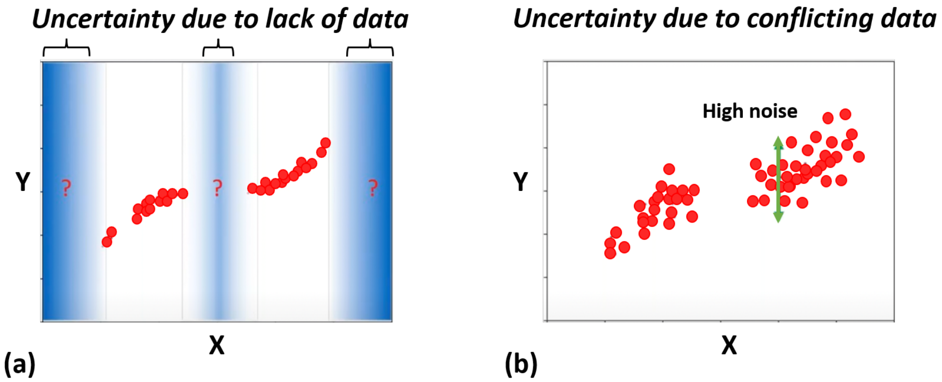


Figure 3.4: The graphic shows the different sources of uncertainty. (a) Depicts the uncertainty due to lack of data (called the model error). The model error reduces as we collect more data in the regions that lack them. (b) Depicts the uncertainty due to conflicting data for the same input space (called the sample noise). This is an irreducible error and can only be quantified but not reduced.

about the prediction for that particular material, and an evaluation of the average performance of the model over the entire dataset is not necessarily significant. Also, such average error metrics are not reflective of the different sources of the uncertainty: (1) Aleatoric (or data noise) uncertainty caused by the stochastic nature of data generation process due to the inherent variations (or randomness) in the data. The problem is referred to as the data consistency challenge in the field of materials science, and can be viewed as aleatoric (or irreducible) uncertainty. (2) Epistemic (or model error) uncertainty that stems from the sparsity in materials data due to the expensive nature of the experiments. As a result, the trained ML model is ignorant due to the lack of knowledge about the best model.

There are several other Uncertainty Quantification (UQ) methods available that model the epistemic uncertainty using confidence intervals [101, 102]. But evaluating the uncertainty for a new observation that the model has not seen using confidence intervals (i.e. ignoring aleatoric contribution) leaves us with a false sense of accuracy. The aleatoric contribution also has its impact on active learning [103] and anomaly detection problems [104]. The total uncertainty metric for a new observation, which we refer to as prediction intervals in this work, can be evaluated by accounting for the aleatoric sources along with the epistemic sources of data uncertainty. As a result, prediction intervals are wider than confidence intervals.

Despite the importance of the prediction intervals in ML based materials modeling and the availability of UQ methods to choose from, the literature on this area is sparse. The reason behind the sparsity in their implementation towards materials informatics problems can be due to additional computational effort beyond the training of the model [30]. This is an exception for Bayesian methods like Gaussian process (GP) since uncertainty estimation is innate to the method itself. GPs are Bayesian models in which a prior distribution is first specified and then updated given observations to yield a posterior distribution. The mean of this posterior distribution is used for regression, and the covariance matrix is used for UQ [105].

However, GP regression has the disadvantage of normality assumptions regrading the prior distribution and sample noise distribution which leads to loss of information such as skew of the error. This makes the uncertainty metric of GP (called the credible intervals) symmetric which is rarely the reality. Also, GP can not be used for high dimensional datasets efficiently due to the cubic time complexity involved with the inversion of the kernel matrix.

A few distribution-free PI algorithms have been introduced in the materials informatics literature in the recent years. One of the popular approaches is the probabilistic quantile regression method, where the mean squared error (MSE) loss function is modified to predict conditional quantiles rather than conditional means [30, 106]. As shown in the work of Choudhary *et al.* and Osman *et al.* [30, 106], one of the strengths of quantile regression is that one has to fit the model only once. However, one notable weakness of the quantile prediction intervals is that the model is quantifying its own uncertainty. This means that one is reliant on the model being able to correctly fit the data. So, in a case where the conditional means of the data follow a linear trend but the quantiles don't, one would then have to choose a non-linear model to get correct prediction intervals. Further, if we're overfitting the training data then the prediction intervals will also become overfitted [107]. Another approach to quantify PIs was introduced by Choudhary *et al.* [30], where the authors used two independent ML models to fit the data: one to predict the actual value of the property (the "base" model) and another to fit the residuals (the "error" model). While interesting, this PI approach models the total error term ($\sigma^2_{model} + \sigma^2_{noise}$) without distinguishing the relative contributions from σ^2_{model} and σ^2_{noise}. Furthermore, it appears that the authors have assumed that the total error term can be modeled as a function of the independent variable $\mathbf{x}$. This particular formalism has a connection with the Bayesian calibration of computer models introduced by Kennedy and O'Hagan [108], who defined the error model term as the discrepancy function. Unlike Kennedy and O'Hagan, in the approach developed by Choudhary *et al.*, the contribution of σ^2_{noise} is not clearly discerned. To train these models, the data must be split into three groups: one for fitting the base model, one for fitting the error model and also validating the base model, and the last for validating the error model. Although the method shows promising PI properties, like the quantile regression, the PI quality relies on the model fit. Moreover, the division of the data into three sets can be a significant drawback when the amount of data is limited which is typical in material science domain.

Ensemble-based ML algorithms offer yet another route to quantify PIs in y. Recent work by Lu *et al.* [109] and Roy *et al.* [110] are examples of random forest based

ML methods, where PIs are constructed based on quantiles method, leave-one-out method, shortest PI method and highest density region method. In another work, Sricharan & Ashok [111] introduced a novel non-parametric bootstrap-based approach for evaluating both the sources of uncertainties. We identify two key advantages with this approach: (1) The algorithm is free from any assumptions about the data or noise model, and (2) It is model agnostic. As this method is foundational to our approach in constructing prediction intervals, I will dive deeper into the algorithm.

In examining the underlying methodology of this prediction interval algorithm, the authors present the *0*th measured target, denoted by $y(x_0)$, as follows:

$$\begin{aligned} y_0 = y(x_0) &= \psi(x_0) + \epsilon(x_0) \qquad (3.9) \\ &= \hat{y}_r(x_0) + \psi(x_0) - \hat{y}_r(x_0) + \epsilon(x_0) \qquad (3.10) \\ &= \hat{y}_r(x_0) + \eta_r(x_0) + \epsilon(x_0) \qquad (3.11) \end{aligned}$$

where, $\hat{y}_r(x_0)$ is the model for a given input x_0, $\psi(x_0)$ is the true regression mean and $\eta_r(x_0)$, defined as $\psi(x_0) - \hat{y}_r(x_0)$, signifies the model's error. $\epsilon(x_0)$ is the error that shifts the target from its true regression mean, $\psi(x_0)$, to the measured value, $y(x_0)$. The prediction error centralized around $\hat{y}_r(x_0)$ is then categorized into two distinct, independent sources: (i) Model error, symbolized as $\eta_r(x)$ and (ii) Observation noise, represented as $\epsilon(x)$.

The authors employ bootstrapping techniques to deduce the realizations of the model error $\eta_r(x_0)$. Bootstrapping as seen in Figure 3.5 is a resampling technique used in statistics to estimate the distribution of a statistic by repeatedly sampling with re-

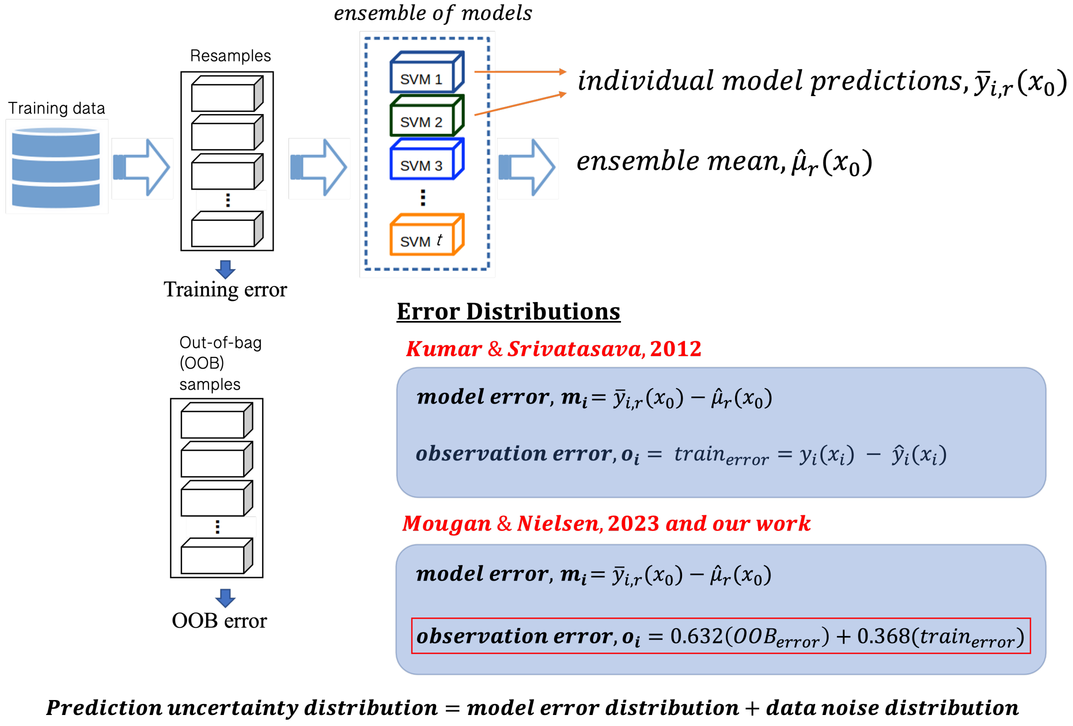

ensemble of models
Resamples
Training data
SVM 1
SVM 2
SVM 3
SVM t
individual model predictions, $\bar{y}_{i,r}(x_0)$
ensemble mean, $\hat{\mu}_r(x_0)$
Training error
Error Distributions
Kumar & Srivatasava, 2012
model error, $m_i = \bar{y}_{i,r}(x_0) - \hat{\mu}_r(x_0)$
observation error, $o_i = train_{error} = y_i(x_i) - \hat{y}_i(x_i)$
Out-of-bag (OOB) samples
OOB error
Mougan & Nielsen, 2023 and our work
model error, $m_i = \bar{y}_{i,r}(x_0) - \hat{\mu}_r(x_0)$
observation error, $o_i = 0.632(OOB_{error}) + 0.368(train_{error})$
Prediction uncertainty distribution = model error distribution + data noise distribution

Figure 3.5: A schematic of our proposed bootstrap prediction interval algorithm. The algorithm is a modification of the prediction interval algorithm published by Srivatsava and Kumar [111]. Our modification lies in the approximation of observation noise distribution using the .632 measure [112] of bootstrap residuals instead of training residuals as proposed by Srivatsava and Kumar. This is similar to the recent work by Mougan and Nielsen.

placement from the observed data. By doing so, it's possible to simulate the behavior of a statistic over many "hypothetical" samples drawn from the underlying population. During the bootstrapping process, not all observations are sampled in each resample due to the nature of random sampling with replacement. The data points that are not included in a particular bootstrap sample are termed as out-of-bag (OOB) samples. These OOB samples are unique to each bootstrap iteration and can be used as a validation set. On average, about 36.8% of the observations end up as OOB samples in any given bootstrap iteration. This proportion arises from the properties of random sampling with replacement, where any specific observation has a $(1 - \frac{1}{n})^n$

chance of being excluded, which approaches $\frac{1}{e}$ as n becomes large.

From the bootstrapping procedure, the authors construct the model error as: $m_i = \bar{y}_{i,r}(x_0) - \hat{\mu}_r(x_0)$, where, $\bar{y}_{i,r}(x_0)$ denotes the predictions from each individual boostrap resample and $\hat{\mu}_r(x_0)$ is the bootstrap mean for m boostrap resamples, defined as:

$$\hat{\mu}_r(x_0) = \frac{\sum_{i=1}^{m} \bar{y}_{i,r}(x_0)}{m} \tag{3.12}$$

The observation error, termed as o_i are then approximated as the training error/residuals,

$$o_i = y_i(x_i) - \hat{y}_i(x_i)$$

One of the disadvantages of this definition is that the training errors is known to be too low because of possible overfitting. This will cause the PI to be too narrow or optimistic. To overcome this issue, we use the 0.632 estimate (Friedman *et al.* [112]), which is a weighted average of the training error and the OOB error. Specifically, it is calculated as:

$$0.632 \times (\text{OOB error}) + 0.368 \times (\text{training error}) \tag{3.13}$$

The rationale behind this weighting is that the training estimate often exhibits an optimistic bias since the model is evaluated on the data it was trained on. On the other hand, the OOB error can sometimes be overly pessimistic. The 0.632 estimate uses the fact that only 36.8% of the observations end up as OOB samples and thus attempts to strike a balance between these two. A recent paper by Mougan and Neilsen [113] employing this modification shows significant improvement in the prediction interval quality.

The 0.632 estimate seeks to combine the optimism of the training error with the pessimism of the OOB error. However, in practice, the degree of overfitting can vary depending on the dataset and the model's complexity. A fixed weight of 0.632 might not always provide the most accurate estimate of the true error, especially when the degree of overfitting is either very low or very high. This is where the Freidman's [112] validation weight comes into play.

The validation weight is designed to dynamically adjust based on the observed overfitting. When there's no overfitting, the model's performance on both the training set and unseen data (OOB samples) should be very similar. In this situation, the validation weight reduces to the standard 0.632 estimate. This is because the training error provides a reasonably accurate depiction of the model's performance on new data. If a model severely overfits the training data, its performance on the training set would be much better than on unseen data. In such cases, the training error can be highly misleading. To counteract this, when overfitting is detected to be severe, the validation weight approaches 1. This effectively prioritizes the OOB error, which is a more trustworthy indicator of the model's performance on new data.

In essence, the validation weight acts as a corrective factor. By adjusting the emphasis between training and OOB errors based on the observed overfitting, it aims to provide a more realistic and reliable estimate of the model's true error.

Mougan and Nielsen use the No Information Error Rate which serves as a baseline error rate for calculating the relative overfitting rate and then the validation weight. In the context of regression, No Information Error Rate is equivalent to the error one would obtain if predictions were made without any information about the predictors. A common approach is to predict the mean of the target variable for all observations. The error of this naive model, which simply predicts the mean response regardless of the input, represents the No Information Error Rate. It essentially captures the inherent variability in the data.

The Relative Overfitting Rate compares the performance difference between the model on the bootstrap samples (training error) and on the out-of-bag samples (OOB error) to the difference between the naive model (No Information Error Rate) and the training error. It's calculated as:

$$\frac{\text{(OOB error) - (training error)}}{\text{(No Information Relative Error Rate) - (training error)}} \tag{3.14}$$

A relative overfitting rate close to 0 suggests minimal overfitting, while a rate close to 1 indicates substantial overfitting.

The validation weight can then be used to adjust the weight between the OOB error and the training error based on the observed degree of overfitting. It is defined as:

$$\text{validation weight} = \frac{0.632}{1-(1-0.632)\times \text{overfitting Rate}} \tag{3.15}$$

In scenarios with no overfitting, the validation weight equals 0.632, aligning with the standard 0.632 bootstrap estimate. This gives a balanced error estimate that's neither too optimistic (as with the training error) nor too pessimistic (as with the OOB error). In cases of severe overfitting, the validation weight approaches 1. This means the model's performance on the OOB samples (new, unseen data) is given full priority, providing a more realistic error estimate.

Now the observation error o_i can be written as:

$$\text{(validation weight)} \times \text{(OOB error)} + \text{(1-validation weight)} \times \text{(training error)} \quad (3.16)$$

The Prediction Error Distribution is the distribution of errors made by the predictive model, and it encompasses both the Model Error and Observation Error. For constructing the prediction intervals, Mougan and Nielsen [113] (and similarly Srivatsava and Kumar [111]) take the approach of combining the Model and Observation Error distributions by convolving the two distributions. In other words, they sum the two error distributions, which gives a new error distribution that represents the total error (or total uncertainty) in the predictions. From this combined distribution, prediction intervals are constructed by taking quantiles at the desired confidence level. This is a straightforward approach and is computationally less intensive.

In contrast to the convolution-based approach, I introduce an additional sampling loop to draw samples from the total error distribution and adjust them using Rademacher variable. Rademacher variables are random variables that take the values +1 or -1 with equal probability. By adjusting the sampled errors using the Rademacher variable, they effectively introduce a form of randomness that mimics the flipping or mirroring of errors. This adjustment can help in providing a more robust estimate of the prediction intervals, especially when the underlying error distribution has heavy tails or is not symmetric. This approach is inspired by the textbook on "Bootstrap methods and their application" by Davison and Hinkley [114], as well as by a paper authored by Lins *et al.* [115]. Our prediction interval algorithm can be seen in Appendix A.1 . Our modifications of the prediction interval algorithm by Srivatsava and Kumar [111] discussed in this section was developed in collaboration with Dr. Tianxi

Li from UVa-Department of Statistics, whose expertise in statistics was pivotal for understanding the relevant literature and algorithm development.

I evaluate the performance of our approach in three different materials science based regression benchmark data sets curated by Henderson *et al.* [116]. I chose datasets of different sizes that involves prediction of: (1) Thermal hysteresis (ΔT) of NiTiCuFePd alloys (22 observations),(2) Curie temperature (T_c) of perovskites (117 observations), (3) Effective thermal conductivity (K_{eff}) (720 observations) using structure and thermodynamics based descriptors. For further details, readers are directed to refer to Henderson *et al.* [116]

Dataset	**Method**	**PI for $\alpha =$**					
		0.95	**0.85**	**0.75**	**0.5**	**0.25**	**0.05**
Thermal hysteresis	Our Method	1.00	0.76	0.69	0.5	0.29	0.08
	Doubt	0.94	0.83	0.74	0.51	0.26	0.049
Curie temperature	Our Method	1.00	0.73	0.67	0.55	0.24	0.063
	Doubt	0.93	0.84	0.73	0.52	0.27	0.048
Effective thermal conductivity	Our Method	0.99	0.78	0.62	0.45	0.28	0.04
	Doubt	0.95	0.86	0.76	0.49	0.25	0.051

Table 3.1: A comparison of our PI algorithm against the "doubt" method by Mougan and Nielsen [113], where the metric used is the prediction interval coverage probability (PICP) for various prescribed confidence levels. Parity of PICPs with the confidence levels is considered to be the best result. The PICPs show that "doubt" outperforms our approach.

I choose the coverage probability (PICP) as our performance metric which is the most important characteristic of PIs, and it is determined by counting the number of target values that the constructed PIs cover.

$$\text{PICP} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} c_i \tag{3.17}$$

where,

$$c_i = \begin{cases} 1, & \text{if } t_i \in [L_i, U_i] \\ 0, & \text{if } t_i \notin [L_i, U_i] \end{cases} \tag{3.18}$$

In this equation, n_{test} represents the number of samples included in the test set, while L_i and U_i correspond to the lower and upper boundaries of the i^{th} PI. It is desirable for the coverage probability (PICP) to be equal to or greater than the nominal confidence level associated with the PIs.

Although our algorithm is model agnostic, in this work, I only demonstrate our algorithm using the ϵ-support vector regression with a non-linear Gaussian radial basis function kernel (SVR_{RBF}) [117]. During the training and comparison process, I split the data into 100 different 90%/10% train/test splits using different sampling seeds. Then the PICP was evaluated for different prescribed confidence levels 0.95,0.85,0.75,0.5,0.25,0.05 and the results are shown in Table 3.1. It is evident the the algorithm's PICP does not does not meet the prescribed confidence levels in all the datasets. The PICP is found to be suboptimal for all confidence intervals, except for the 0.05 level. Specifically, there are noticeable underestimations at the 0.85, 0.75, and 0.5 confidence levels, while the 0.95 level consistently shows overestimations. In contrast, the algorithm developed by Mougan and Nielsen [113], known as "doubt", performs well across the board for all confidence levels. As a result, I integrate the "doubt" algorithm into our holistic CTE model for RE-Si-O. The underlying reasons for the shortcomings of our original approach, as well as potential improvements, will be the focus of future research.

The next section will cover the explainable artificial intelligence methods used in this thesis that can help us ensure trustworthiness of the ML models.

3.2.5 Explainable artificial Intelligence

Explainable artificial intelligence (XAI) refers to the methods and techniques within the field of AI that make the logic, decisions, and actions of ML models understandable to human experts. The need for explainability stems from the increasing complexity of AI models, which can act as "black boxes", making it difficult to understand how decisions are made. XAI aims to bridge this gap, fostering trust and facilitating the integration of AI systems in critical areas like science, healthcare, and finance.

The taxonomy of XAI ((Figure 3.6)) can be broadly categorized into two main areas:
Explainable Models: These are models (also known as glass box models) designed to be inherently explainable. They include simpler models like linear regression, decision trees, and rule-based systems where the relationship between input and output is transparent.
Post Hoc Explainability (Black-box explainers): This involves applying techniques to explain the decisions of accurate black-box models after they have been made. Methods like breakDown [118], SHAP [119], and what-if plots [118] fall under this category.

The accuracy-explainability tradeoff (Figure 3.7) represents a fundamental dilemma in AI. On one hand, complex models such as deep neural networks and ensemble methods often deliver superior predictive accuracy. They can capture intricate patterns and nonlinear relationships in data that simpler models might miss. However, their complexity makes them opaque and difficult to understand, even for experts. On the

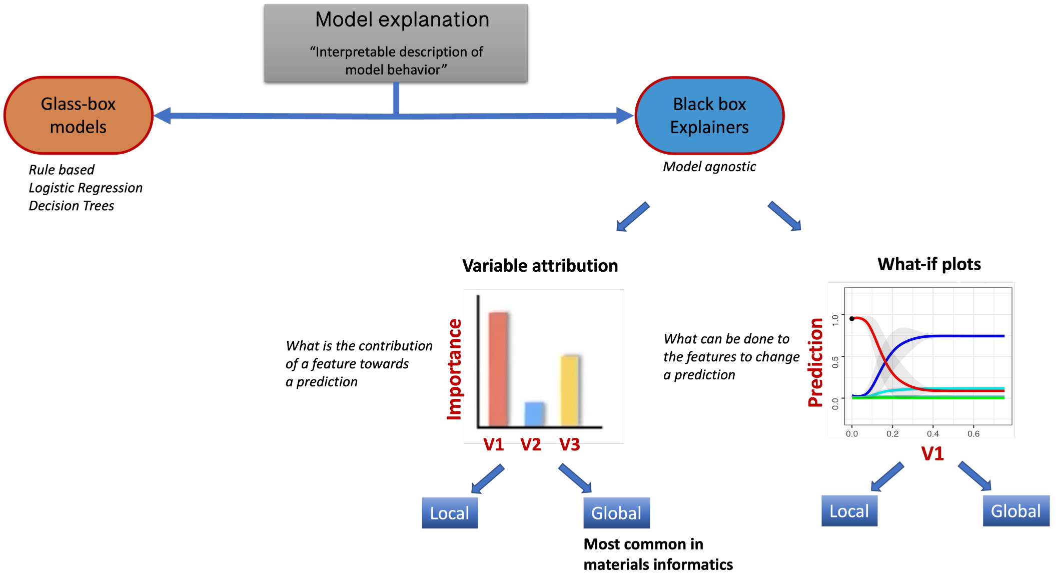


Figure 3.6: A general overview and categorization of Model explanation. The schematic shows the broad classification of model explanations including glass box models and black box explanations or post-hoc model explanations. This thesis focusses only on the black box explainers which can further be classified into two types: 1) Variable attribution and 2) What-if plots. They both can be used to provide model explanations in different granularities (local and global). While local explanation involves a single observation, global explanation explains the model as a whole considering all the observations in question.

other hand, simpler models like linear regression or decision trees are more transparent and interpretable. They allow users to see the exact relationships between input variables and the predicted output. Yet, this transparency often comes at the cost of predictive accuracy, especially when dealing with complex or high-dimensional data like those I deal with in this thesis. Therefore, I strike balance between transparency and accuracy by choosing the post-hoc model explanations for the non-linear black box models known for their accuracy.

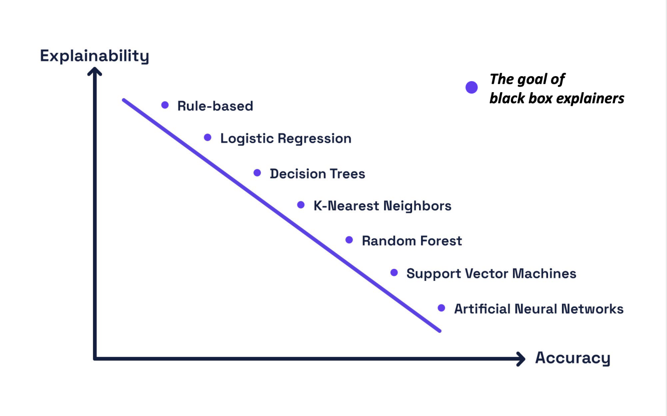


Figure 3.7: The accuracy-explainability tradeoff in AI. The accuracy vs explainability (notational) graph depicts the lack of explainability or transparency in accurate models and the lack of accuracy in explainable models. The goal of employing black-box explainers is to harness the advantages of both ends of the spectrum.

Post-hoc Explainability

Post hoc model explanation can be defined as the human interpretable descriptions of the relationship between the input and its corresponding prediction. For tabular data, human interpretable description is often variable attribution values that additively build towards the prediction. Most of the variable attribution methods work based on a common general idea. The intuition behind the idea is that if breaking the link between a variable (X_j) and the target (y), increases in prediction error, then the variable j is consider important for making a prediction. Thus, its is given a high variable attribution value. This approach aids in understanding which variables or features are most influential in a model's predictions, which is a cornerstone of explainability in ML.

Distinct from mere variable importance, which offers a broad overview of influential features, counterfactuals represent another type of model explanation. Often referred to as "what-if scenarios", they are plots that show the marginal effect of a single feature on the prediction of a previously fit model [120]. In addition to offering insights on variable importance, they also provide specific, actionable recommendations. Therefore, they typically resonate more with non-experts, presenting tangible scenarios that demystify model decisions in relatable terms.

Both variable attribution and counterfactuals can be further classified into: i) global explanation and ii) local explanation [121, 120]. The global explanation provides an interpretable description of the whole model behaviour given the entire input. It helps the user to understand the big picture of the underlying function. Whereas, the local explanation provides an interpretable description of the model behaviour specific to a target neighborhood (instance-wise). In a typical materials science dataset, this might often mean the description the input-output relationship for a specific material and/or a specific structure.

The literature on application of explainable ML methods to engineering problems is dense. However, the idea of incorporating explainable ML methods into the current ML-based materials design and discovery workflow is still in its infancy. The efforts based on tree-based ML models used for materials design often include global interpretation methods since its innate to decision tree method [122, 123, 124]. However, only a handful of materials science papers that uses local explanation methods are found in the literature [122, 125, 28, 29].

Along with the sparsity of local interpretability methods being used in materials informatics, I also identify the following knowledge gaps in the application of explainable ML in materials science:

i) What-if scenario: Although variable attribution methods give an idea of which variable is important towards a prediction, it does not help the domain expert visualize the functional relationship between the variables and the prediction. To do this, what-if plots like the partial dependence plot (global method) and individual conditional expectation plot (local method) can be used [121, 120]. They describe the change in the prediction as we change a variable with other variables held constant.

ii) Local, global and intermediate level explanation: While most of the efforts in explainable ML focus on either local or global explanation, a framework that can provide explanation at the intermediate level is crucial for making decisions [121, 120]. This is especially important in materials science as often there will be a class/group of instances that correspond to a material class or a specific structure. Local explanation can address one of the compounds in the material class. But its also crucial to uncover the underlying mechanisms of materials phenomena with respect to types of materials or structure.

iii) Uncertainty of explanations: As explained in the previous section, uncertainty is crucial to understand when we can trust an ML output. Arguably, this applies to model explanations too. However, there are no papers in the materials informatics literature that report model explanation with quantified uncertainties.

The global variable importance provides users with an overarching perspective on the influence of predictor variables on the model's predictions, encompassing the entire dataset. The algorithmic approach adopted by DALEX [126] to compute this global variable importance is rooted in permutation. For each predictor variable, the algorithm shuffles or permutes its values and subsequently gauges the impact of this permutation on the model's performance. Essentially, the deterioration in the model's performance, often measured using metrics like accuracy for classification tasks or root mean squared error (RMSE) for regression, signifies the importance of that particular variable. A substantial dip in performance upon permutation is indicative of the variable's pivotal role in the model's predictions. Conversely, a negligible change suggests that the variable might not be as influential. They can discern which features are paramount to the model and which might be superfluous. Furthermore, understanding these influential variables can enhance the model's interpretability, making it more palatable to stakeholders and facilitating more informed decision-making.

While global variable importance offers a broad understanding, it's equally crucial to delve into the intricacies of individual predictions. Local variable attribution methods address this need, providing insights into how each feature contributes to specific instance predictions. Among these local methods, SHapley Additive exPlanations (SHAP), developed by Lundberg and Lee [127], stands out as one of the most acclaimed techniques. SHAP's strength lies in its foundation on cooperative game theory, where SHAP values were initially conceptualized to determine individual players' contributions to a collective outcome [119]. In predictive models, the contribution of each variable can be calculated by averaging over every possible ordering of variables using SHAP, allowing to locally analyze the importance of each input feature for a given instance prediction.

In SHAP analysis, the variable importance measure of the j-th variable (or the first j variables) on an instance $\underline{x}_*$ is articulated as $\varphi(j, \underline{x}_*) = \frac{1}{p!} \sum_J \Delta^{[j|\pi(J,j)]}(\underline{x}_*)$, where the summation spans all $p!$ permutations of input variables. The term $\pi(J, j)$ designates the set of input variables that precede the j-th variable in set J. In essence, $\varphi(j, \underline{x}_*)$ encapsulates the average importance of the variable across all conceivable orderings, underscoring SHAP's thoroughness in model interpretation. In the ML context, SHAP offers a granular look into feature contributions by averaging over all potential feature orderings. This meticulous approach ensures a comprehensive understanding of a feature's influence, making SHAP particularly adept at pinpointing the drivers behind individual predictions.

While variable attributions provide insights about the more general questions such as whether a feature should be included in the model in the first place, or which features are most important to a model's prediction, what-if plots provide useful insights into the relationship between each feature and the predicted response for each instance or compositions in our dataset. The what-if plots for a particular observation (local) are generated by determining the marginal effect of a feature, $\hat{f}(x_S^{(i)}, x_C^{(i)})$, i.e., the change in model prediction as $x_S^{(i)}$, the value of a feature under consideration, increases or decreases [128]. In the function described, $x_C^{(i)}$ are actual values of other features from the dataset. This localized perspective is known as Individual Conditional Expectation (ICE).

Building on the ICE concept, the Partial Dependence Plot (PDP) can be understood as an average of ICE plots over all data points. Specifically, the PDP for a feature is calculated as $\text{PDP}(x_S) = \frac{1}{N}\sum_{i=1}^{N} \hat{f}(x_S^{(i)}, x_C^{(i)})$, where N represents the total number of instances in the dataset. By taking this average, PDP offers a holistic understanding of a feature's influence across the entire dataset, revealing general trends that may not be apparent when examining individual data points through ICE. Together, ICE and PDP provide a comprehensive lens, with ICE detailing instance-specific insights and PDP giving an aggregated overview, ensuring a rich and nuanced understanding of model behaviors.

Up to this point, I've looked into both local and global techniques for variable attribution and what-if analyses. However, in the realm of materials science, examining clusters or groups of data points can reveal hidden patterns and trends. To tap into these insights, I merge the perspectives of global and local explanations by employ-

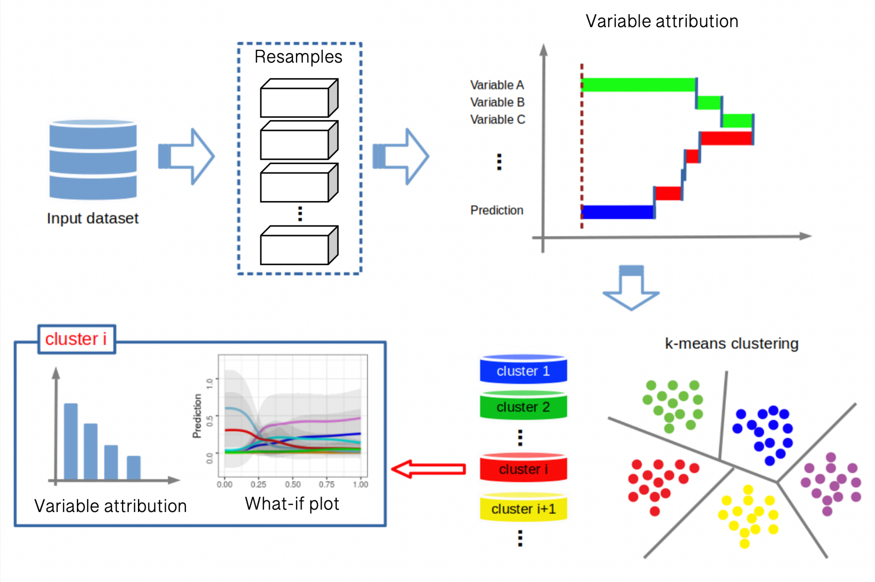


Figure 3.8: The flow chart for ML and model explainability approach for explaining model behaviour at local and intermediate levels. (Adapted from our recently published paper [129].)

ing unsupervised clustering. The foundational idea is to use variable attributions as embeddings, encapsulating the sensitivity profile of each data point. When I cluster based on these embeddings, I group data points that the model perceives similarly in terms of feature importance. Such clustering can unearth patterns that might be latent in the raw data, revealing cohorts of data points for which the model has consistent sensitivities. I perform this step by clustering (k-means method) the instances using treating the local variable importance values as embeddings (as shown in Figure 3.8). With these clusters in hand, applying ICE analyses becomes a powerful next step. By exploring counterfactual scenarios for entire clusters, we can discern collective patterns in how slight perturbations to inputs might affect model predictions for groups of data points. This aggregated view offers a balance between the granularity of individual data point explanations and the broader strokes of global model behavior. In essence, this combined approach harnesses the strengths of both variable attribution and counterfactual explanations. It proposes a method that operates at an intermediate level of granularity, offering insights that are both detailed and broadly applicable.

In the context of ML, elucidating model behavior is only half the challenge. Equally crucial is the understanding of the uncertainties associated with these interpretations. As we delve into variable attributions and counterfactual explanations, assessing the reliability and variance of the derived insights becomes paramount. Bootstrap standard errors emerge as a robust statistical tool to address this. As discussed in the previous section, bootstrapping involves repeatedly drawing samples (with replacement) from the original dataset and recalculating the desired metric for each sample. For our context, it means recalculating variable attributions and counterfactual scenarios over these resampled datasets. The mean and standard error observed across these recalculations provides the aggregate measure and measure of uncertainty respectively. All the methods and approaches I've discussed, encompassing the explainability techniques, our unique clustering strategy and uncertainty quantification, are concisely compiled and included in Appendix A.2.

Our algorithm capable of explaining the ML model on global, local and intermediate levels is shown to be useful for understanding a prediction. Two of my recent publications on the use of explainable ML methods on high entropy alloys show that the developed algorithms are portable between datasets [129, 130]. The promise of these methods are demonstrated on the basis of the post hoc model explanations to check if the model reflects the underlying science. This adds trust to our data-driven approach.

Chapter 4

Crystal Chemistry of $RE_2Si_2O_7$

Rare-earth disilicates ($RE_2Si_2O_7$) represent one of the three major material classes in the rare-earth, silicon, and oxygen (RE-Si-O) chemical space. They serve as an ideal subject for examining the coefficient of thermal expansion (CTE) due to their role in Environmental Barrier Coatings (EBCs). EBCs, designed to shield high-performance materials from harsh thermal and chemical conditions, make CTE an indispensable factor for ensuring long-lasting performance. Nonetheless, the engineering of disilicates is complex, largely due to their polymorphic nature. The existence of various energetically viable crystal forms complicates the design process. Consequently, two key attributes become essential for the application of disilicates in EBCs: (1) The DFT total energy difference is invaluable for assessing the relative stability of multiple polymorphs, and (2) The relationship between structure and CTE that can accelerate the design of these compounds.

4.1 Chapter Organization

The chapter begins by contextualizing the significance of $RE_2Si_2O_7$ in the EBC application. I discuss why this material is pivotal for high-performance applications and what challenges are currently faced in its implementation. I then dive into the DFT total energy calculations of $RE_2Si_2O_7$. This section provides a comprehensive survey on how energetically competitive various polymorphs are, based on their stability and formation energies. The DFT calculations form the cornerstone of our theoretical investigation. This section elaborates on the methodology used for these calculations and discusses the results, emphasizing the presence of several energetically competing crystal structures. This phenomenon is rationalized as one of the reasons for observing polymorphism in $RE_2Si_2O_7$. In the next section, we introduce the ML models employed to establish a relationship between the crystal structures and their CTE. Finally, I present experimental data to validate the computational findings. Collaborative experiments conducted by Dr. Deijkers and Dr. Wadley from UVa-MSE serve to corroborate the relationship between the crystal structure and CTE, thereby validating the ML predictions.

4.2 Introduction

Metal oxide and silicate coatings are commonly used to provide thermal and environmental protection for high temperature gas turbine engine components [65, 66, 67, 68]. In applications such as aviation jet engines, where temperatures up to 1500°C are common in the hottest sections, environmental barrier coatings (EBCs) protect the Si-based ceramic matrix composite (CMC) gas turbine components from reacting with oxygen and corrosive water vapor. Thus, there is interest in discovering novel EBC oxides and silicates for gas turbine applications that will enable higher temperature operation for the future jet propulsion technology [131]. $RE_2Si_2O_7$, where RE is a rare-earth element, are identified as candidate materials for EBC applications because of their thermal stability at high temperatures, high resistance to oxidation and good match in the coefficient of thermal expansion (CTE) with the Si-based

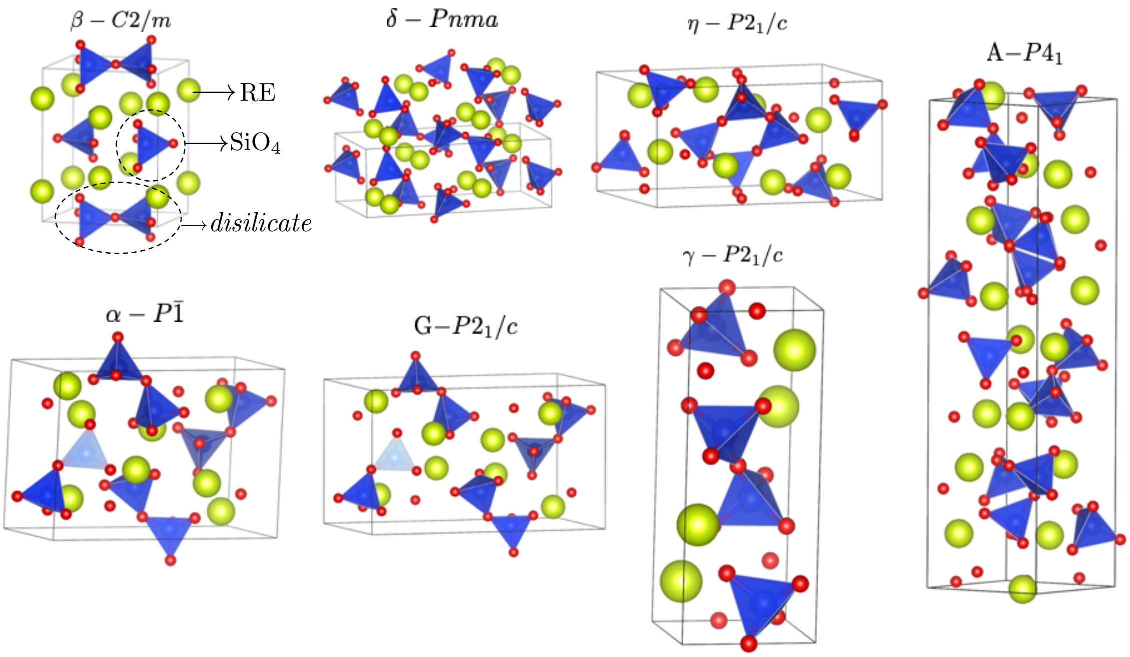


Figure 4.1: The crystal structures of seven polymorphs in the $RE_2Si_2O_7$ family of compounds. The RE-site, SiO_4 tetrahedral unit, and the disilicate unit is highlighted.

CMCs, such as SiC fiber reinforced SiC composites [132, 133]. But, the disilicates are relatively volatile in steam conditions at 1400–1500° C [134]. The steam volatility problem can be reduced through the use of rare-earth monosilicates (RE_2SiO_5) [134], however these materials have a higher CTE than the Si-based CMCs [135] leading to coating cracking and loss of protection [136]. Thus, there is no optimal EBC material available that satisfies all requirements for operation under extreme conditions and coating materials design remains an important challenge [131, 77, 137, 138].

The focus of this study is on the $RE_2Si_2O_7$ family of compounds because the chemical diversity and structural flexibility (via polymorphism) offer a fertile playground to tailor the thermophysical properties for a targeted application. The phase stability of $RE_2Si_2O_7$ family has also been shown to be a strong function of the RE^{3+} chemistry, temperature, and processing history [139, 140, 141]. Some of the experimentally observed polymorphs are shown in Figure 4.1. Intriguingly, not all polymorphs have a CTE in the desired range for EBC applications (3–5.5×10^{-6} K) [132]. Depending on the RE^{3+} cation and the associated $RE_2Si_2O_7$ crystal structure, their linear coefficient of thermal expansion (CTE) range from 3.9–14×10^{-6} K^{-1} [140, 142].

The objective of this work is two-fold. First, to explore the total energy trend of $RE_2Si_2O_7$ materials as the rare-earth is varied using DFT calculations, and to extract insights into the polymorphism exhibited by these compounds. A quantitative understanding of the energetics and its relationship with structural phase transformations is important for tailoring the properties of $RE_2Si_2O_7$-type compounds as EBC materials [132]. This work aims to provide thermodynamic insights using DFT calculations. Although theoretical work based on DFT calculations exist on $RE_2Si_2O_7$ compounds [77, 138, 143, 144, 145, 146, 147], most of the effort has focused on a subset of $RE_2Si_2O_7$ crystal structure space.

The vast majority of the DFT studies have focused on determining the CTE from phonon calculations using quasi-harmonic applications (QHA) [77, 138]. The CTE of a material can be defined as the fractional increase in length (linear dimension) per unit rise in temperature [148]. To date, experimental CTE values are known for only 18 out of 112 possible $RE_2Si_2O_7$ structures [77, 140, 142]. Performing high-throughput phonon calculations to calculate the CTE using QHA on the entire search space is computationally challenging because some of the polymorphs have as many as 88 atoms in the unit cell. An alternative approach is needed that will enable accelerated screening for quantities, such as CTE, that are difficult to calculate using first principles methods.

The second objective of this work is to explore the potential of machine learning (ML) methods to rapidly predict the CTE for unexplored $RE_2Si_2O_7$ polymorphs. Although a few ML-based attempts at CTE prediction have been explored in the literature [149, 150, 151], no ML work exists on the prediction of CTE for $RE_2Si_2O_7$ polymorphs.

4.3 DFT calculations of the $RE_2Si_2O_7$ crystal chemistry

I started this research by calculating the ΔE (in meV/atom) with respect to the lowest energy structure using DFT. The DFT calculations are performed using the planewave pseudopotential code `Quantum ESPRESSO` [152]. The PBEsol exchange-correlation functional [153] was used and the core and valence electrons were treated with ultrasoft pseudopotentials [154]. The Brillouin zone integration was performed using a Monkhorst-Pack [155] k-point mesh centered at Γ and 60 Ry plane-wave cutoff for wavefunctions (600 Ry kinetic energy cutoff for charge density and potential). The scalar relativistic pseudopotentials were taken from the PSLIBRARY [156]. The atomic positions and the cell volume were allowed to relax until an energy convergence threshold of 10^{-8} eV and Hellmann-Feynman forces less than 2 meV/Å, respectively, were achieved. The $4f$-states for the rare-earth elements are considered as core states in our calculations. The converged crystal structures were visualized in `VESTA` [79] and the space groups were determined using `FINDSYM` [157].

In Table 5.3, the total energy difference per atom (ΔE) with respect to the lowest energy structure is given for a selected set of eight $RE_2Si_2O_7$ compounds. An interesting feature about these compounds (except RE=Yb and Y, which will be discussed later) is that each compound has at least three or more crystal structures within a ΔE threshold of 5 meV/atom, indicating close energetic competition between the

Table 4.1: Total energy difference (ΔE, meV/atom) from DFT calculations with respect to the lowest energy structure in $RE_2Si_2O_7$, where RE=La, Ce, Sm, Gd, Nd, Pr, Yb and Y. Space groups given in parentheses indicate the final converged structure when the lattice constant tolerance is set at 0.0001 decimal places or lower in `FINDSYM`[157].

Space group	ΔE (meV/atom)							
	$La_2Si_2O_7$	$Ce_2Si_2O_7$	$Sm_2Si_2O_7$	$Gd_2Si_2O_7$	$Nd_2Si_2O_7$	$Pr_2Si_2O_7$	$Yb_2Si_2O_7$	$Y_2Si_2O_7$
$C2/m$ (β)	67.08	52.60	12.91	**0.00**	28.95	40.49	**0.00**	**0.00**
$Pnma$ (δ)	24.74	19.43	4.68	2.05	9.46	14.50	29.79	10.47
$P2_1/c$ (η)	0.63	0.24	6.27	14.01	0.60	0.22	71.59	33.05
$P\bar{1}$ (α)	**0.00**	2.75	8.83	1.40	**0.00**	1.75	69.05	32.88
$P2_1/c$ (G)	3.11 ($P\bar{1}$)	2.75	**0.00 ($P\bar{1}$)**	1.40 ($P\bar{1}$)	0.001 ($P\bar{1}$)	1.76	36.54	14.04
$P2_1/c$ (γ)	63.03	49.76	13.15	1.32	27.87	38.60	3.51	1.99
$P4_1$ (A)	0.34	**0.00**	4.72	11.20	0.09	**0.00**	64.90	27.02

polymorphs. For example, $La_2Si_2O_7$ has four unique structures that are within the 5 meV/atom threshold. The lowest energy structure is found to be the triclinic α-$P\bar{1}$ space group. The monoclinic (η-$P2_1/c$) and tetragonal (A-$P4_1$) structures are within 1 meV/atom above the α-$P\bar{1}$ structure (within the intrinsic errors of DFT total energy). The G-$P2_1/c$ monoclinic structure is 3.11 meV/atom above the ground state structure. The $La_2Si_2O_7$ compound has been synthesized in both G-$P2_1/c$ and A-$P4_1$ crystal structures and both of their experimental CTE values are also known [139, 140].

Similar to $La_2Si_2O_7$, the $Ce_2Si_2O_7$ compound has been synthesized in both A-$P4_1$ (at lower temperature) and G-$P2_1/c$ (at higher temperature) structures [158]. The DFT data given in Table 5.3 indicate that the lowest energy structure of $Ce_2Si_2O_7$ is A-$P4_1$ and the total energy of the monoclinic G-$P2_1/c$ structure is 2.75 meV/atom higher with respect to A-$P4_1$ structure. Although there is another structure (η-$P2_1/c$) that is only 0.24 meV/atom higher in energy above that of A-$P4_1$, it is not discussed here because this structure has not been reported in the experimental literature. Experimentally, $Ce_2Si_2O_7$ is known to undergo a "sluggish" phase transition from A-$P4_1$ to G-$P2_1/c$ at 1274° C [158, 159]. While these results indicate that the ΔE data from DFT can be used to find the thermodynamical contribution to the free energy, the rates of the phase changes upon cooling are not addressed. On the basis of ΔE data given in Table 5.3, the observed polymorphic behavior in $Pr_2Si_2O_7$, $Nd_2Si_2O_7$, $Sm_2Si_2O_7$ and $Gd_2Si_2O_7$ can be rationalized [131, 139, 141, 160].

The $Yb_2Si_2O_7$ paints a different picture. The lowest energy structure belongs to the monoclinic β-phase with a $C2/m$ space group. The next most favorable phase is another monoclinic structure (γ-$P2_1/c$), which is 3.51 meV/atom higher in energy relative to β-$C2/m$. The γ-$P2_1/c$ structure has not been reported in the experimental literature. The third lowest energy structure is the orthorhombic δ-$Pnma$ phase, which is 29.8 meV/atom higher in energy compared to the β-$C2/m$ phase. In the experimental literature, crystal structure and CTE data has only been found for the β-$C2/m$ phase.

There is one compound of $Y_2Si_2O_7$ that eluded the simple thermodynamic description based on the DFT ΔE data. The $Y_2Si_2O_7$ compound exists in multiple polymorphs (α-, β-, γ-, and δ-phases) [140], however the ΔE data fails to capture this behavior. The ΔE data predicts an energetic competition only between the β-$C2/m$ and γ-$P2_1/c$ phases. We conjecture that kinetics (e.g. activation barrier for the phase transformation) is relatively more important in the $Y_2Si_2O_7$ crystal chemistry compared to other compounds in the $RE_2Si_2O_7$ family. The Gibbs free energy of formation based on CALPHAD (CALculation of PHAse Diagram) has been shown to reproduce the phase diagram of $Y_2Si_2O_7$ with sufficient accuracy, which further confirms the important role of kinetics in describing the phase equilibria of the $Y_2Si_2O_7$ compound [161, 162, 163].

The ΔE data for the remaining eight $RE_2Si_2O_7$ (RE=Dy, Eu, Ho, Lu, Tm, Sc, Er and Tb) compounds are given in the Appendix section B.1. The calculated ΔE trend for these compounds are similar to those of $Yb_2Si_2O_7$ and $Y_2Si_2O_7$ indicating relatively less propensity for polymorph formation. It is noted that the $Sm_2Si_2O_7$ has been synthesized before in at least two polymorphs [139, 160]: α-$P\bar{1}$ and A-$P4_1$ phases. Table 5.3 shows that both polymorphs are within the ΔE=5 meV/atom energy threshold.

4.4 Structure-CTE relationship of $RE_2Si_2O_7$ using Machine Learning

In addition to crystal structure, CTE is also important to assess the viability of $RE_2Si_2O_7$ compounds for EBC applications. I built an ML model that can predict CTE using descriptors based on the structure and following the sections that discusses details of the dataset, descriptor generation, model architecture and results.

4.4.1 Dataset and Descriptor Generation

The input to ML models are the descriptors that represent each $RE_2Si_2O_7$ in their crystal structure. The DFT-optimized unit cell parameters (three axial distances and three axial angles), volume, and number of atoms in the unit cell are considered as inputs. The descriptor set, comprising these 8 variables for all 112 $RE_2Si_2O_7$ compounds, is subjected to pairwise statistical correlation analysis using pair-wise correlation analysis (Pearson correlation). The correlation plot Figure 4.2 shows that the number of atoms in the unit cell and lattice constant c are strongly correlated with the unit cell volume. Therefore, the number of atoms in the unit cell and lattice constant c are not considered as input for ML.

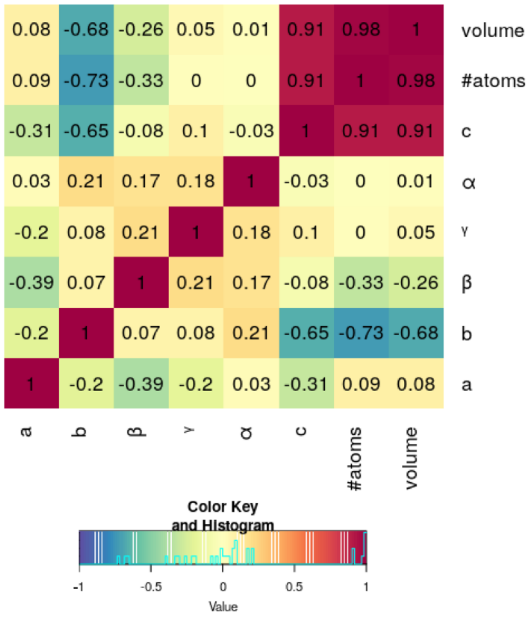


Figure 4.2: Pair-wise statistical correlation analysis for the eight descriptors from DFT calculations for 16 $RE_2Si_2O_7$ compounds in seven polymorphs. Dark red and dark blue indicate strong positive and negative correlation, respectively. For example, a strong positive correlation (> 0.9) is found between the lattice constant c, the number of atoms in the unit cell (#atoms), and the volume. Since correlation indicates redundancy, we did not consider the lattice constant c and #atoms as input for building ML models. The remaining descriptors, a, b, α, β, γ, and volume were used as inputs for building ML models.

The dependent variable in the dataset is the experimentally measured volumetric CTE, to be specific, ABCTE (apparent bulk coefficient of thermal expansion) data compiled from surveying the published literature [77, 140, 142]. Fernández-Carrión *et al.* documented at least three different methods for extracting the CTE data from the literature: (i) ABCTE from X-ray diffraction (XRD) data, (ii) coefficient of mean linear thermal expansion from XRD using matrix algebra analysis, and (iii) average linear CTE from dilatometric measurements. Therefore, some amount of variability is expected in the CTE data used for ML. However, we were unable to quantify the variation because the measurement uncertainties are not reported in the original pa-

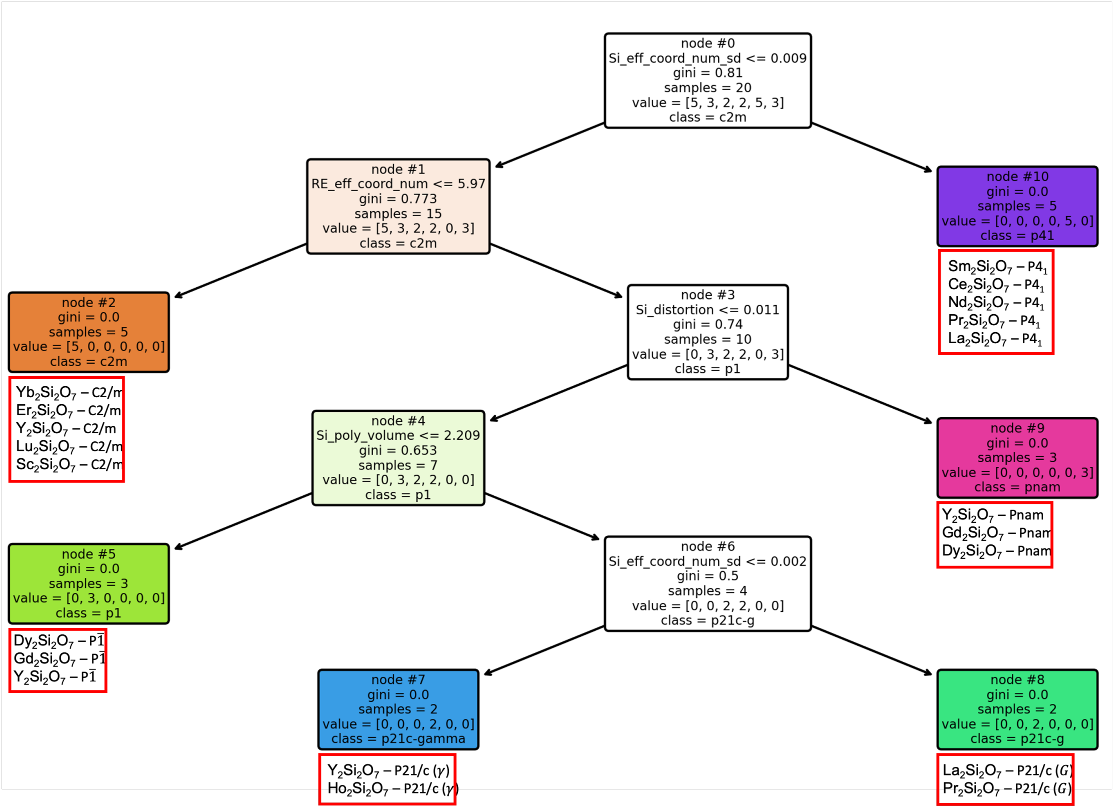


Figure 4.3: Decision trees for classification of polymorphs in RE disilicates based on GINI impurity using the training data shown in Table 6.1. At each node, a check occurs and if true proceeds to the left and vice-versa until it reaches the leaf node that displays the classification. Under each classification, the compounds belonging to the space group are listed.

pers. In addition, it is also not uncommon to have secondary phases and non-uniform stress distrbution in the samples that were used for CTE measurements, which introduces another source of measurement uncertainty that is difficult to quantify using traditional analysis. Therefore, due to the lack of data on measurement uncertainty, I considered the experimental ABCTE as a point estimate for training ML models.

Although unit cell parameters are the main descriptors used for ML model training, mainly because they are relatively easy to generate, leveraging polyhedral descriptors in data visualization provides additional insights. Specifically, it aids in understanding the relationship between these descriptors and the various polymorphs. The details on these descriptors is included in Table 2.1. The dataset can be visualized using decision trees built for classification of the different polymorphs in RE disilicates. The primary objective of this study is not to create a generalized predictive model, but rather to gain insights into the underlying data patterns and generate meaningful decision rules. Decision trees are renowned for their interpretability, making them an ideal choice for comprehending complex datasets. By deliberately overfitting the model, I aim to capture fine-grained patterns, which may reveal underlying relationships among the features and classes. The results of this analysis provide valuable domain knowledge that can be leveraged to make informed decisions and guide further data exploration.

The tree initiates its decision-making process at the root node by evaluating the feature Si_eff_coord_num_sd. It further unfolds through several decision nodes, generating unique decision paths that culminate in the final classification into one of six different polymorphs in the training data. If Si_eff_coord_num_sd > 0.009, the compounds are directly classified into $P4_1$. This path serves as a decisive factor for RE disilicate compounds, emphasizing the pivotal role of variance in Si effective coordination number in their classification. The pathway to $C2/m$ involves compounds where Si_eff_coord_num_sd ≤ 0.009 and RE_eff_coord_num ≤ 5.97. It can be noted that $P4_1$ and $C2/m$ are classified just using effective coordination number based descriptors. Compounds are classified into $Pnam$ if Si_eff_coord_num_sd ≤ 0.009, RE_eff_coord_num > 5.97 and Si_distortion > 0.011. This path is indicative of the significance of both Si distortion and coordination environment of the RE and Si polyhedra. If Si_eff_coord_num_sd ≤ 0.009, RE_eff_coord_num > 5.97, Si_distortion ≤ 0.011 and Si_poly_volume ≤ 2.209 the compounds are classified as $P\bar{1}$. The classification of $P2_1-\gamma$ and $P2_1-G$ involves the use of Si_eff_coord_num_sd twice. If Si_eff_coord_num_sd ≤ 0.009, RE_eff_coord_num > 5.97, Si_distortion ≤ 0.011, Si_poly_volume > 2.209 and Si_eff_coord_num_sd ≤ 0.002 the compounds are classified as $P2_1 - \gamma$ and $P2_1 - G$ follows a similar path Si_eff_coord_num_sd ≤ 0.009, RE_eff_coord_num > 5.97, Si_distortion ≤ 0.011, Si_poly_volume > 2.209 and Si_eff_coord_num_sd > 0.002.

The next visualization that I used to understand the dataset is the plot of our polyhedral descriptors against the RE ionic radii. The visualizations illuminate the specific trends exhibited by each structural type, as well as the unique insights provided by the polyhedral descriptors that are not captured solely by ionic radii, the conventional descriptor for CTE. From the plots Figure 4.4, it can be inferred that the variables correlate with ionic radii in different strengths. While RE_avg_bond_length is strongly correlated (positive) to ionic radii Si_avg_bond_length, Si_poly_volume has only a moderate positive correlation. RE_avg_bond_length_sd is the only variable that doesn't show any correlation. The form_Energy which has a strong positive correlation to ionic radii is the only variable with several outliers in every space group. There are also other variables with a few outliers RE_avg_bond_length and Si_bondangle_var. This indicates that the descriptors are likely dictated by a complex interplay of factors that ionic radii alone cannot capture. Si_bondangle_var often shows a strong negative correlation with ionic radii, especially in space groups like $Pnma$, $P4_1$ and $P\bar{1}$. Overall, although it can see that many of the variables have correlation with ionic radii, the presence of some uncorrelated variables and outliers suggest that the ionic radii alone is not sufficient for capturing the full scope of influences on the material's CTE.

4.4.2 Machine Learning and Bootstrap Resampling

An ensemble of SVR_{RBF}-based ML models is used to establish a relationship between the unit cell parameters ($\mathbf{X}$) from DFT and the continuous dependent variable CTE (Y). Given a sample of data ($\mathbf{X}, Y$), the regression problem can be formulated as follows, $Y = f(\mathbf{X}) + \eta$, where η is the random error term. Although there are many methods available for determining f from data, we chose the ϵ-support vector regression with a non-linear Gaussian radial basis function kernel (SVR_{RBF}) [117] because of its improved generalization ability [164, 165]. The ϵ-SVR contain "hyperparameters" such as the penalty parameter, the insensitive loss function parameter, and a coefficient of the kernel function that control the model complexity and help balance the bias-variance tradeoff. The SVR hyperparameters were adjusted to optimize the leave-one-out error. We utilized the ϵ-SVR method as implemented in the `e1071` package [166] within the `RSTUDIO` environment [167].

Since the training set (containing 15 observations) is only a small sample of the population, we lack complete information on the probability distribution of model parameters. This introduces uncertainty in quantifying the model output distribution. Therefore, estimating error bars for each prediction is as important as estimating the mean prediction itself. We employed the bootstrap resampling technique [168] for assessing uncertainty, a method particularly well-suited for scenarios with limited training data [169, 170, 171, 172, 173, 174]. This approach involves generating numerous ”pseudo training sets” by randomly selecting samples from the original training set, with replacement. Consequently, some data points may be duplicated, while others might be omitted. Using these pseudo sets, we build an ensemble of SVR_{RBF} models. The ensemble's mean and standard error serve as indicators of the expected CTE and its uncertainty, respectively.

There are two important tuning variables associated with the procedure: (i) The size of the data points to be resampled (n) and (ii) the number of resamples (B). Here, the n was fixed at 15 (number of unique samples in the training data), whereas 10 different values for B were explored (between 10–200). Every resampling created two kinds of samples: in-bag and out-of-bag. The data points in the in-bag samples were used for training the ML models. The out-of-bag samples were used to test the performance of the trained models. The optimal value for B was determined by calculating the out-of-bag root mean squared error (RMSE) associated with each B.

A total of 15 $RE_2Si_2O_7$ compounds were used to train the ML models and eight independent $RE_2Si_2O_7$ were used for testing. The ensemble SVR models with B=25 had the smallest out-of-bag RMSE. Thus, an ensemble of 25 SVR models was used to train the models and the trained models were subsequently used to predict the ABCTE for the remaining 97 $RE_2Si_2O_7$ compounds not considered in the training set.

The performance of the ensemble of 25 SVR models on the training and test data is shown in Figure 4.5. The training data is shown as black dots and test data is shown as blue diamond and green squares. The test data includes both experimentally measured and DFT-QHA computed ABCTE values. In Table 6.1, the compounds used for training and testing, along with the ML predictions, are also given. It is important to note that the test data was not used to train the ensemble of ML models. A vast majority of the data points, especially the test points, either lie close to or on the X=Y line that indicates good training performance.

The trained models were then used to predict the ABCTE for the remaining $RE_2Si_2O_7$ compounds not considered in the training set. We chose $Sm_2Si_2O_7$ for experimentally validating our predictions since there is no experimental or theoretical report of CTE

Table 4.2: The compounds used for training and testing the ML models is given. The ABCTE predictions from ML, along with the uncertainty (σ), is also given. All experimental (training and test data) and DFT-QHA (test data) ABCTE data are taken from Fernández-Carrión *et al.* [140], Dolan *et al.* [142] and Luo *et al.* [77], respectively. The temperature ranges used for the CTE determination in both experiments and DFT-QHA calculations are also given. (Data published in Ref. Ayyasamy$_{R}E2Si2O7$)

$RE_2Si_2O_7$	Experimental ABCTE ($\times 10^{-6}$ K^{-1})	Temperature range (K)	ML Predicted ABCTE $\pm$ σ ($\times 10^{-6}$ K^{-1})
Training data (Experimental from literature)			
A-$P4_1$ $La_2Si_2O_7$	14	303-1373	12.4 ± 1.17
A-$P4_1$ $Pr_2Si_2O_7$	11.8	303-1573	11.8 ± 0.86
A-$P4_1$ $Nd_2Si_2O_7$	10.5	303-1473	10.9 ± 1.08
δ-$Pnma$ $Gd_2Si_2O_7$	7.3	303-1873	7.6 ± 0.46
δ-$Pnma$ $Dy_2Si_2O_7$	7.7	303-1423	7.6 ± 0.44
β-$C2/m$ $Er_2Si_2O_7$	3.9	303-1873	4.2 ± 0.14
β-$C2/m$ $Yb_2Si_2O_7$	4	303-1873	4.3 ± 0.11
β-$C2/m$ $Lu_2Si_2O_7$	4.2	303-1823	4.3 ± 0.14
β-$C2/m$ $Sc_2Si_2O_7$	5.4	303-1873	4.6 ± 0.52
γ-$P2_1/c$ $Ho_2Si_2O_7$	4.2	303-1748	4.3 ± 0.78
γ-$P2_1/c$ $Y_2Si_2O_7$	3.9	293-1473	4.4 ± 0.77
G-$P2_1/c$ $La_2Si_2O_7$	6.4	303-1073	6.6 ± 0.40
G-$P2_1/c$ $Pr_2Si_2O_7$	6.8	303-1648	6.7 ± 0.41
α-$P\bar{1}$ $Gd_2Si_2O_7$	8.3	303-1573	8.2 ± 0.50
α-$P\bar{1}$ $Dy_2Si_2O_7$	8.5	303-1648	8.1 ± 0.45
Test data (Experimental from literature)			
β-$C2/m$ $Y_2Si_2O_7$	4.1	293-1673	4.2 ± 0.35
α-$P\bar{1}$ $Y_2Si_2O_7$	8	293-1473	8.1 ± 0.69
δ-$Pnma$ $Y_2Si_2O_7$	8.1	293-1673	7.6 ± 0.44
Test data (DFT-QHA from literature)			
β-$C2/m$ $Ho_2Si_2O_7$	4.09	300-1700	4.2 ± 0.19
β-$C2/m$ $Tm_2Si_2O_7$	3.92	300-1700	4.2 ± 0.11
γ-$P2_1/c$ $Er_2Si_2O_7$	4.03	300-1700	4.3 ± 0.78
δ-$Pnma$ $Tb_2Si_2O_7$	8.27	300-1700	7.6 ± 0.44
δ-$Pnma$ $Ho_2Si_2O_7$	8.57	300-1700	7.5 ± 0.44
New prediction and Experimental validation (This work)			
A-$P4_1$ $Sm_2Si_2O_7$	11.55 ± 18	573-1248	10.39 ± 1.41

data for $Sm_2Si_2O_7$ in any of the polymorphs. Also, $Sm_2Si_2O_7$ is one of the compounds with multiple competing polymorphs as per our ΔE data (Table 5.3) indicating that $Sm_2Si_2O_7$ could form in either G-$P2_1/c$, δ-$Pnma$ or A-$P4_1$ phase. Therefore, the experimental validation of our CTE and ΔE predictions for this compound will be valuable for the community. The experiments were performed by our collaborators Dr. Deijkers and Prof. Wadley from UVa-MSE department.

4.5 Experimental validation of $Sm_2Si_2O_7$

The ceramic $Sm_2Si_2O_7$ sample was synthesized using high-energy ball milling. Samples annealed at 950°C for 144 hours did not form any $Sm_2Si_2O_7$ phase. On the other hand, the sample annealed at 1400°C for 20-hours formed $Sm_2Si_2O_7$ phase, but there was also a significant fraction of hexagonal $Sm_{9.33}Si_6O_{26}$ apatite phase. Finally, the sample annealed at 1500°C for 20 hours formed a majority $Sm_2Si_2O_7$ A-phase in the $P4_1$ space group (see Figure 4.6a, where small amounts of α-SiO_2 cristobalite and $P2_1/c$-Sm_2SiO_5 phases are also found). The tetragonal A-$P4_1$ phase is not the lowest energy structure from DFT (Table 5.3). The $P4_1$ structure is 4.71 meV/atom higher in energy relative to the ground state structure. The scanning electron microscopy (SEM) image (shown in Figure 4.6b) also indicates the samples majority phase as $Sm_2Si_2O_7$ with partially reacted Sm_2SiO_5 and SiO_2 together with a small volume fraction of porosity.

The CTE value for the sample was measured using dilatometry in the temperature range 373-1273 K. The heating and cooling cycle of a test sequence is shown in Figure 4.6c. The CTE is measured as the $\frac{dL}{L_0} \cdot \frac{1}{dT}$, where L_0 is the original length. The dilatometry data also shows a small kink in the CTE curve in the 200°-300°C temperature range, where the α- to β-SiO_2 transition is known to occur [175]. The contribution from the α- to β-SiO_2 phase change towards CTE can be ignored by not considering any dilatometry data below 300°C. In the 300°-975°C temperature range, CTE can be determined from the heating and cooling cycle as 11.73 and 11.37 $\times 10^{-6}$ K^{-1}, respectively. The average CTE can then be calculated as $11.55 \pm 0.18 \times 10^{-6}$ K^{-1}. However, resolving the role of α-SiO_2 alone does not address the problem, because the sample still contains some volume fraction of the β-SiO_2 and Sm_2SiO_5 phase. Thus, even in the refined temperature range (300°-975°C), our measured CTE data corresponds to that of a composite sample. Disentangling the contribution from Sm_2SiO_5 was found to be non-trivial due to the large spread in the literature CTE data for Sm_2SiO_5. For example, He *et al.* [176] report a wide range of CTE values between 9 and 12 $\times$ 10^{-6} K^{-1} for the Sm_2SiO_5 compound. Tian *et al.* [177] report a value between 7 and 8 $\times$ 10^{-6} K^{-1} for Sm_2SiO_5 in the $C2/c$ structure. Thus, we conclude that the CTE value of $11.55 \pm 0.18 \times 10^{-6}$ K^{-1} can be considered as a lower bound limit for the $Sm_2Si_2O_7$ compound in the A-$P4_1$ structure. This result falls within the uncertainty of the ML prediction of $10.39 \pm 1.41 \times 10^{-6}$ K^{-1}, thus validating the model.

4.6 Summary

In conclusion, this chapter presents a detailed analysis of the trends between RE ionic radii and polyhedral descriptors for RE disilicate compounds. Our observations indicate that some of the polyhedral descriptors like RE_avg_bond_sd were not highly correlated with the ionic radii. This suggests that these descriptors carry information not captured by the ionic radii alone, broadening our perspective on the factors influencing the CTE. We also delve into the potential of unit cell parameters as another descriptor set for machine learning-based CTE prediction. I employed an integrated strategy that fuses DFT calculations with ML techniques to effectively identify and predict the CTE of $RE_2Si_2O_7$ materials. These materials are critically important for the engineering of EBCs. The DFT data provide valuable information about ΔE that favor the formation of specific polymorphs. Further, our machine learning models, incorporating unit cell parameters derived from DFT, successfully capture CTE trends across a range of $RE_2Si_2O_7$ compounds. The validity of these models is corroborated through comparisons with both existing literature and our own experimental results for $Sm_2Si_2O_7$. Overall, the chapter outlines an approach that spans computational modeling, ML, and experimental validation. This multi-pronged strategy aims to accelerate the design and discovery of $RE_2Si_2O_7$ compounds with optimal characteristics for EBC applications. I acknowledge that understanding the volumetric CTE is not entirely sufficient because CTE anisotropy plays an important role in these compounds [24]. Even so, the advancements highlighted in our discussion represent a pivotal movement in this area of research. It's noteworthy that comprehensive data on CTE anisotropy has only recently started appearing in the literature, making the development of a machine learning model to describe anisotropy a challenging endeavor. However, in the upcoming chapter, we delve into an interesting trend in CTE anisotropy as presented by Ridley *et al.* [24], exploring it through electronic structure calculations.

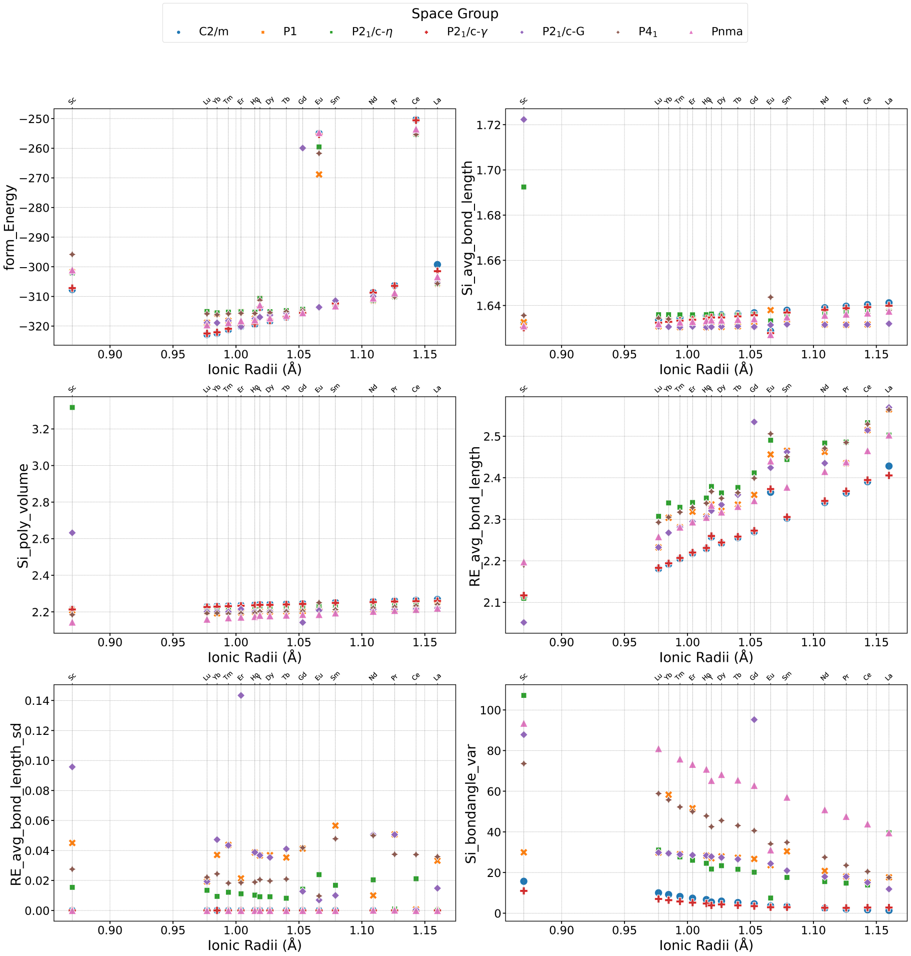


Figure 4.4: Graphical representation of polyhedral metrics (Y-axes) in relation to rare-earth (RE) ionic radii in Angstroms (Å) (X-axes), focusing on a charge state of 3+ and a coordination number of 8. The plots are representative of all 112 RE disilicates. The secondary X-axis bears the annotations of RE cation in the compound.

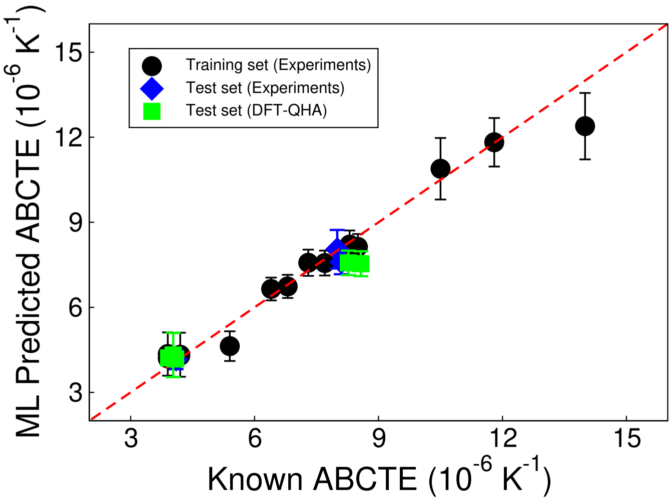


Figure 4.5: Performance of the trained ensemble of 25 SVR models. The X- and Y-axes are the known and ML predicted ABCTE data, respectively. The error bar represent the standard deviation from the ensemble of ML models. The red dashed line represents the X=Y line and the data points falling on this line indicate perfect agreement between the ML models and the known data. The black dots are the experimental data points used to train the models. The blue diamond and green squares represent the test data that were not used to train the ML models. The true values for the blue diamond and green square data points are taken from the experimental measurements and published DFT-QHA calculations [77], respectively.

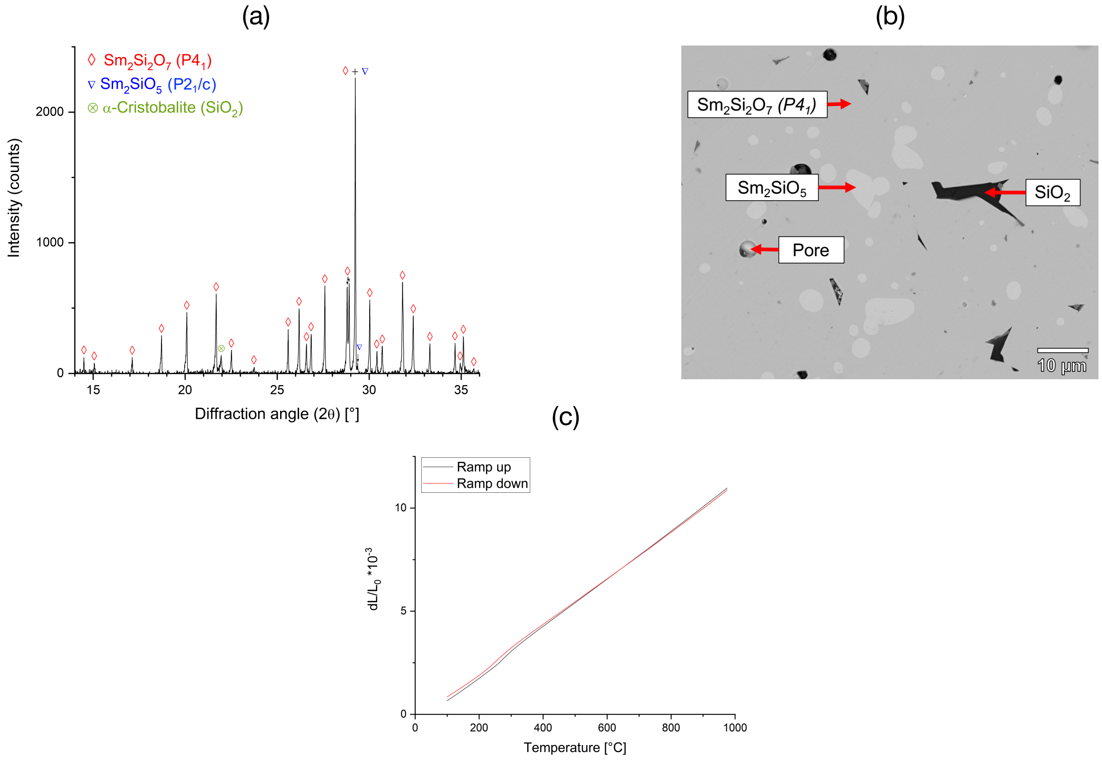


Figure 4.6: Results from (a) X-ray Diffraction (XRD), (b) SEM image, and (c) Dilatometry measurements for the synthesized $Sm_2Si_2O_7$ ceramic sample. The majority phase in XRD is the tetragonal $P4_1$ A-phase, which was predicted as metastable from DFT ΔE calculations. The SEM image shows the sample as predominantly $Sm_2Si_2O_7$ with a minor amount of partially reacted Sm_2SiO_5 and unreacted SiO_2 together with porosity. The CTE from dilatometry for the temperature range of 300°-975° C is measured as $11.55 \pm 0.18 \times 10^{-6}$ K^{-1}, which is in agreement with the ML prediction of $10.39 \pm 1.41 \times 10^{-6}$ K^{-1}.

Chapter 5

Crystal Chemistry of RE_2SiO_5

This chapter delves into the rare-earth monosilicates (RE_2SiO_5), which is yet another important materials family within the RE-Si-O crystal chemistry domain. Similar to disilicates, these compounds have emerged as promising candidates for Environmental Barrier Coatings (EBC) applications, owing to their exceptional thermal stability and resistance to chemical attack. The challenges encountered in the design of disilicates for EBC applications are equally applicable to monosilicates. Consequently, a comprehensive survey of the DFT total energy, coupled with the rapid prediction of the coefficient of thermal expansion (CTE) across the entire search space included in this chapter is of significant value to the EBC community. A recent experimental study conducted by Dr. Ridley and Dr. Opila from UVa-MSE has shed new light into the CTE values of select monosilicate compounds. The intriguing anisotropy trends reported in their work have inspired a more focused examination of anisotropy from the perspective of electronic structure within my research. Understanding the anisotropy of CTE is pivotal in formulating design philosophies for EBC materials, as it provides insights into the directional dependencies that may influence the material's behavior and performance in specific applications.

5.1 Chapter Organization

Similar to Chapter 4, I will initially survey the DFT total energy difference of the RE_2SiO_5 space to understand how energetically competitive the polymorphs are in this materials family. I then discuss DFT calculations and the results predicting the presence of several energetically competing crystal structures, which is rationalized as one of the reasons for observing polymorphism. Then I use ML methods to establish a relationship between the crystal structures of RE_2SiO_5 and their volumetric CTE. Finally, I show the DFT work used to unravel the correlation between electronic structure and anisotropy trends observed in key RE_2SiO_5 compounds.

5.2 Introduction

The RE_2SiO_5 compounds are promising candidates for EBC application at extreme temperature environments, such as the high temperature gas turbine engine components as they offer superior thermal stability and resistance to chemical attack compared to the other RE silicate material classes [65, 66, 67, 68, 69] These compounds are known to form in one of the two low-symmetry monoclinic crystal structures in space groups $C2/c$ (also referred to as "X2" in the literature) or $P2_1/c$ (X1). The unit cell and the local coordination environment surrounding the cations for the two structures are shown in Figure 5.1. The $C2/c$ and $P2_1/c$ structures are preferred for smaller (Tb–Lu) and larger RE^{3+} cations (La–Gd), respectively (except Tb which exists in both $C2/c$ and $P2_1/c$) [178, 24]. The conventional unit cells of $C2/c$ and $P2_1/c$ structures contain 64 and 32 atoms, respectively. In both the $C2/c$ and $P2_1/c$ structures, there are two unique crystallographic sites for the RE-element and one

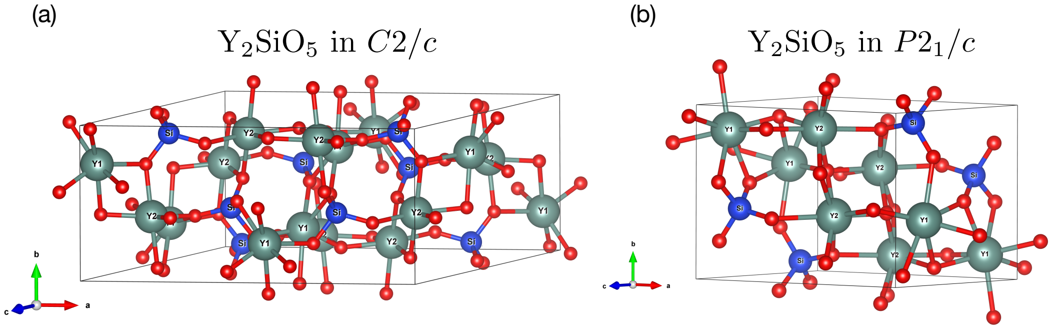


Figure 5.1: The crystal structure of Y_2SiO_5 in (a) $C2/c$ (ground state) and (b) $P2_1/c$ (metastable) space groups. The two crystallographically unique Y-atoms (Y1 and Y2) and the Si atoms are labeled for clarify.

unique crystallographic site for the Si-atom. In both structures, the Si-O atoms are coordinated by a Si-centered tetrahedron; whereas there are two unique, irregular RE-cation centered RE-O polyhedral units in the unit cell (with a local coordination environment containing RE-O_6, RE-O_7 or RE-O_8 oxygen neighbors depending on the RE-cation size).

Like $RE_2Si_2O_7$, the study of RE_2SiO_5 polymorphs and their relationship to CTE and potential energy landscape is equally vital, given their potential for tailoring thermophysical properties. Once again, the challenge lies in the limited availability of phase stability and CTE data, as experimental approaches are costly, and high-throughput phonon calculations using the Quasiharmonic Approximation (QHA) are computationally demanding, especially for polymorphs with large unit cells. An ML based approach is required to understand the linkage between structure and properties and thus expedite the screening of compounds. First, I aim to analyze the total energy trends across different rare-earth elements using DFT calculations, shedding light on the polymorphism within these compounds. A detailed understanding of the energy dynamics and its correlation with structural phase changes is vital for customizing RE_2SiO_5 as EBC materials. While some theoretical insights based on experiments are available, the focus has been limited to a narrow subset of RE_2SiO_5 compounds, mirroring the challenges faced with $RE_2Si_2O_7$. Secondly, I harness the capabilities of ML to swiftly forecast the CTE for unexplored RE_2SiO_5 polymorphs. Although ML has been employed for CTE prediction in other contexts, its application to RE_2SiO_5 is an unexplored area, paralleling the opportunities and challenges identified for $RE_2Si_2O_7$. This novel approach could pave the way for a more efficient exploration of these complex materials. As discussed in Chapter 2, in cubic systems, CTE is an isotropic quantity. However, the RE silicate materials that form in non-cubic space groups, the CTE can have different magnitudes (and signs) along each unique axis. A popular example in the RE_2SiO_5 materials family is the Y_2SiO_5. This compound is one of the promising candidates for EBC application due to its excellent high-temperature stability, low oxygen permeability, and low thermal conductivity. [179, 69] However, recent experimental measurements indicate a large CTE anisotropy in phase-pure Y_2SiO_5 bulk compound, which is not desired for EBC applications. [24] In contrast, the degree of CTE anisotropy in Sc_2SiO_5 was found to be notably low. Both Y_2SiO_5 and Sc_2SiO_5 form in the same $C2/c$ crystal structure. In addition, both Y and Sc have similar valence electron configuration of $nd^1(n+1)s^2$, where n is the principal quantum number (n=3 and 4 in Sc and Y, respectively). There is no data in the literature that provide insights into the plausible reasons behind the large difference in the CTE anisotropy between Y_2SiO_5 and Sc_2SiO_5. It is common in the literature to explore a solid solution approach (eg., mix Sc_2SiO_5 with Y_2SiO_5) to reduce the CTE anisotropy problem. However, Sc-containing compounds are more expensive compared to that of the Y-containing compounds. Therefore, their extensive use as an EBC material is uncertain. Developing a fundamental understanding is important to rationally explore strategies that will enable one to lower the degree of CTE anisotropy in the Y_2SiO_5 compound.

5.3 DFT study of the RE_2SiO_5 crystal chemistry

To describe the energetics of the RE_2SiO_5 crystal chemistry, I calculated the ΔE (in meV/atom) with respect to the lowest energy structure using DFT calculations. The DFT calculations were performed using the planewave pseudopotential code `Quantum ESPRESSO` [152]. The PBEsol exchange-correlation functional [153] was used and the core and valence electrons were treated with ultrasoft pseudopotentials [154]. The Brillouin zone integration was performed using a Monkhorst-Pack [155] k-point mesh centered at Γ and 60 Ry plane-wave cutoff for wavefunctions (600 Ry kinetic energy cutoff for charge density and potential). The scalar relativistic pseudopotentials were taken from the PSLibrary [156]. The atomic positions and the cell volume were allowed to relax until an energy convergence threshold of 10^{-8} eV and Hellmann-Feynman forces less than 2 meV/Å, respectively, were achieved. The $4f$-states for the rare-earth elements are considered as core states in our calculations. The converged crystal structures were visualized in `VESTA` [79] and the space groups were determined using `FINDSYM` [157].

Unlike disilicates, the interpretation of ΔE in monosilicates is relatively simple because it has been known to form in only two polymorphs X1 ($P2_1/c$) and X2 ($C2/c$). In Table 5.1, I show the ΔE values obtained from the DFT calculations for the entire monosilicate chemical space. The stable structure for larger rare-earth cations (La-Gd) is X1 ($P2_1/c$) and the stable structure for smaller cations (Tb-Lu) is X2 ($C2/c$) [24] with an exception of $Tb_2Si_2O_7$ which exists in both the structures. The DFT ΔE calculations capture this trend well except for $Gd_2Si_2O_7$ whose ground state is predicted to be X2 ($C2/c$). However, the X1 ($P2_1/c$) phase having a ΔE value 2.95 meV/atom shows a close energetic competition between the two $Gd_2Si_2O_7$ phases. Similarly, the close ΔE values for the two structures of $Tb_2Si_2O_7$ support the compound's existence in both configurations. Recently, the RE monosilicates (RE=Sc, Y, Dy, Er, and Yb) were synthesized in X2 ($C2/c$) structure by Ridley et al [24], confirming the DFT predictions.

Table 5.1: Total energy difference (ΔE, meV/atom) from DFT calculations with respect to the lowest energy structure in RE_2SiO_5. Space groups were determined using the `FINDSYM` [157] web application. Structures with ΔE=0 (bold face font) represent the ground state structure for that compound.

	ΔE (meV/atom)							
Space Group	La_2SiO_5	Ce_2SiO_5	Pr_2SiO_5	Nd_2SiO_5	Sm_2SiO_5	Eu_2SiO_5	Gd_2SiO_5	Tb_2SiO_5
X1-$P2_1/c$	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	2.95	6.39
X2-$C2/c$	24.93	22.03	18.64	14.67	6.03	38.65	**0.00**	**0.00**

	ΔE (meV/atom)							
Space Group	Dy_2SiO_5	Ho_2SiO_5	Er_2SiO_5	Tm_2SiO_5	Yb_2SiO_5	Sc_2SiO_5	Y_2SiO_5	Lu_2SiO_5
X1-$P2_1/c$	13.19	11.80	13.19	13.12	12.31	7.63	6.98	11.39
X2-$C2/c$	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**

5.4 Exploring Structure-CTE relationship of RE_2SiO_5 using Machine Learning

I follow the same approach as in the case of $RE_2Si_2O_7$ for building a model for predicting the CTE for all 32 possible RE_2SiO_5 compounds.

5.4.1 Dataset and Descriptor Generation for Training Machine Learning Models

The dataset for training the ML model to predict CTE for RE silicates is shown in Table 6.1. The data contains 11 RE monosilicates including three X1 structures and 8 X2 structures. The source of the CTE data is literature based on experimental characterization as shown in the Table 6.1.

Table 5.2: Thermal Expansion Coefficients Data for RE monosilicate collected from Literature.

Compound	Experimental ABCTE ($\times 10^{-6}$ K^{-1})	Temperature range (K)	Reference
X1-$P2_1/c$ Gd_2SiO_5	10.3	473–1623	Al Nasiri *et al.* [180]
X1-$P2_1/c$ Nd_2SiO_5	8.91	303-1373	Ridley *et al.* [181]
X1-$P2_1/c$ Y_2SiO_5	8.7	473–1623	Fukuda *et al.* [182]
X2-$C2/c$ Dy_2SiO_5	7.6	303-1373	Tian *et al.* [183]
X2-$C2/c$ Ho_2SiO_5	7.38	303-1373	Tian *et al.* [183]
X2-$C2/c$ Lu_2SiO_5	6.7	473–1623	Al Nasiri *et al.* [180]
X2-$C2/c$ Sc_2SiO_5	6.17	303-1373	Ridley *et al.* [24]
X2-$C2/c$ Tb_2SiO_5	8.9	303-1373	Tian *et al.* [183]
X2-$C2/c$ Tm_2SiO_5	7.64	303-1373	Tian *et al.* [183]
X2-$C2/c$ Y_2SiO_5	7.7	303-1373	Fukuda *et al.* [182]
X2-$C2/c$ Yb_2SiO_5	7.2	473–1623	Al Nasiri *et al.* [180]

In the preceding chapter on RE disilicates, I utilized decision trees for data exploration and visualization. The interpretability of decision trees makes them equally apt for examining the complexities within RE monosilicates. Just as before, My aim is not predictive accuracy but rather understanding the intricacies of the data. Unlike the RE disilicates, the decision tree for monosilicates is simple. The polymorphs are simply split by the polyhedral descriptor Si_eff_coord_num (the average effective coordination number of Si polyhedra in the structure). The decision rule for "X1"-$P2_1/c$ structure is Si_eff_coord_num ≤ 3.989 and it is Si_eff_coord_num > 3.989 for the structure "X2"-$C2/c$. The result of the Pearson correlation analysis performed on the dataset (11 training and the rest 21 observations not included in training) with 8 variables including the unit cell parameters a, b, c, α, β, γ, volume, number of atoms (natoms) shown in Figure 5.3 suggest that the training data should be reduced to 11 rows and 2 columns represented by just the "volume" and "b" axis cell parameter. It is important to note that variable "b" in turn is highly correlated to the polyhedral descriptor Si_eff_coord_num that splits this dataset.

Continuing along the lines of the investigative approach employed in the RE disilicates chapters, the next visualization we perform involves plotting polyhedral descriptors

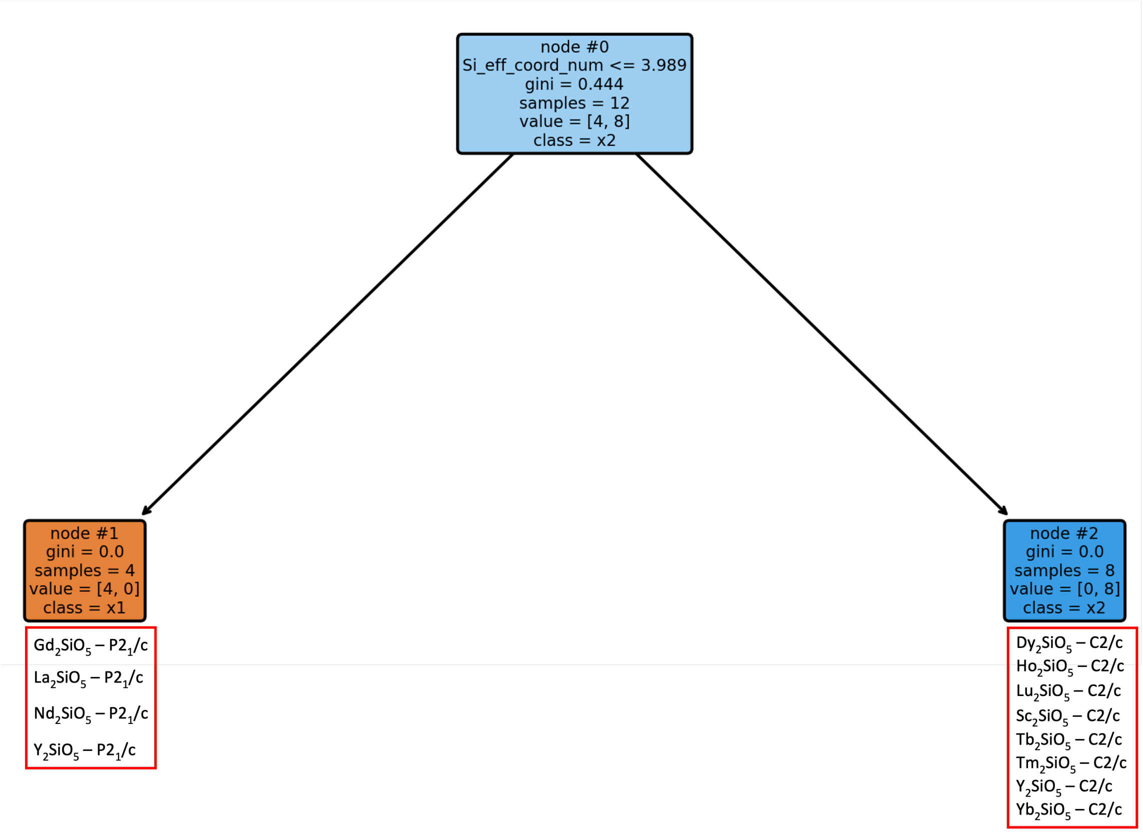


Figure 5.2: Decision trees for classification of polymorphs in RE monosilicates based on GINI impurity using the training data shown in table Table 5.2. At each node, a check occurs and if true proceeds to the left and vice-versa until it reaches the leaf node that displays the classification. Under each classification, the compounds belonging to the space group are listed.

against RE ionic radii. These visual representations, referenced in Figure Figure 5.4, expose the unique patterns associated with each space group. From these plots, it's evident that the P2$_1$/c and C2/c structural types both exhibit similar trends in variables such as form_Energy, RE_avg_bond_length, and Si_bondangle_var. Specifically, while form_Energy and RE_avg_bond_length display a strong positive relationship with ionic radii, Si_bondangle_var shows a negative one, with the trend being more pronounced for $P2_1/c$.

In contrast, when considering Si_avg_bond_length, Si_poly_volume, and RE_avg_bond_length_sd,

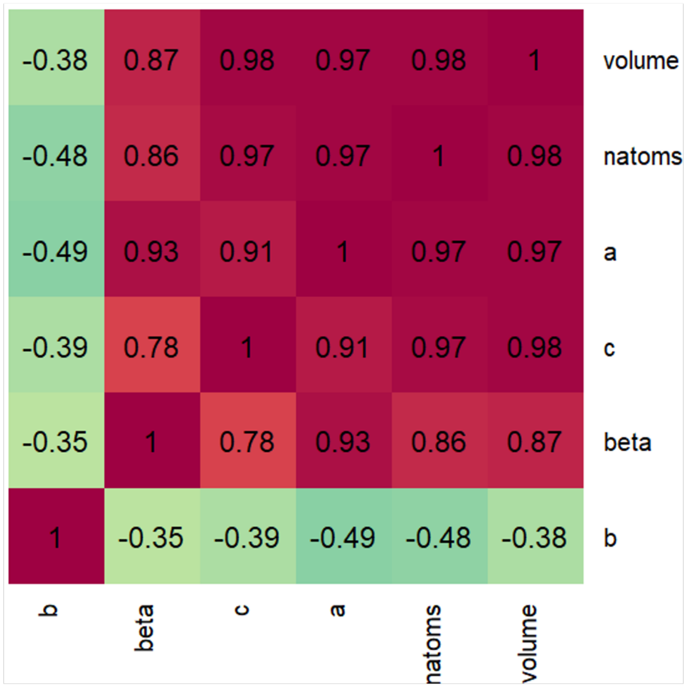


Figure 5.3: Pearson correlation coefficient plot for the dataset comprising 32 RE_2SiO_5 compounds. Dark red indicate strong positive correlation. The plot shows that the variables “b” and “volume” are the only two variables that are not highly correlated and thus are not redundant and suitable inputs for the ML model.

the trends between $P2_1/c$ and $C2/c$ diverge substantially. Particularly for $C2/c$, these variables demonstrate a high degree of non-linearity and lack of correlation. This stands in stark contrast to the disilicates, where space groups largely followed parallel trends. For the $C2/c$ space group, Si_avg_bond_length and Si_poly_volume are strongly positively correlated with ionic radii, while RE_avg_bond_length_sd shows a negative correlation.

5.4.2 Machine Learning and Bootstrap Resampling

I adopt the same eSVR based architecture as seen in Chapter 4 to establish a quantitative relationship between the unit cell parameters ($\mathbf{X}$) from DFT and CTE (Y)

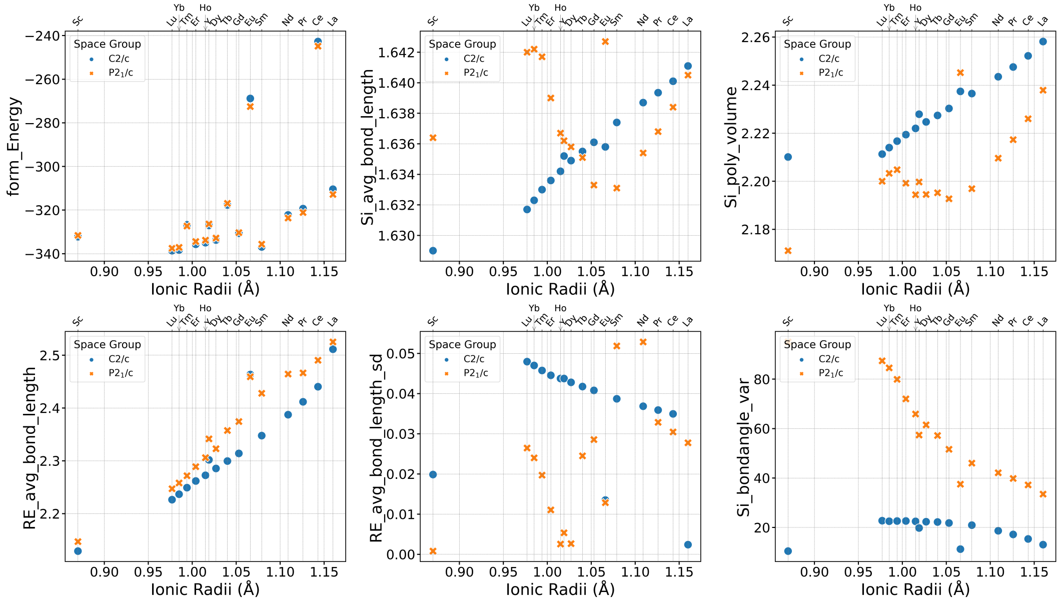


Figure 5.4: Graphical representation of polyhedral metrics (Y-axes) in relation to Rare Earth (*RE*) Ionic Radii in Angstroms (Å) (X-axes), focusing on a charge state of 3+ and a coordination number of 8. The secondary x-axis bears the annotations of RE cation in the compound. The plots are representative of all 32 RE monosilicates.

assembled from surveying the published experimental literature of known RE_2SiO_5 compounds. The only change is the number of bootstrap resamples (B) used in the ensemble which is optimised (based on smallest out-of-bag RMSE) to be 10 in the case of RE_2SiO_5 training dataset.

A total of 11 RE_2SiO_5 compounds were used to train the ML models and 3 independent RE_2SiO_5 (RE=Er,Tb,Sm) were used for validation from literature. The performance of the model is shown in Figure 5.5. A vast majority of the data points, especially the test points, either lie close to or on the X=Y line that indicates good training performance. The trained models were then used to predict the ABCTE for the remaining 21 RE_2SiO_5 compounds not considered in the training set. The predictions for all 32 RE_2SiO_5 compounds, including specific markings (asterisks) for the ground state polymorphs, are illustrated in Figure 5.6. These predictions serve as a valuable tool in making informed design decisions tailored to EBC applications.

However, it is imperative to recognize that the provided predictions pertains solely to volumetric CTE. Given that CTE is inherently a tensor property, the potential anisotropy in CTE must be taken into account. This consideration becomes particularly significant when contemplating the utilization of non-cubic compounds for EBC applications, as the directional dependence of thermal expansion may influence material's mechanical behavior and performance.

5.5 Anisotropy trends in RE_2SiO_5

Figure 5.7 shows one of the interesting results to note from the work by Ridley *et al.* is a large CTE anisotropy in phase-pure Y_2SiO_5 bulk compound, which is not desired for EBC applications [24]. In contrast, the degree of CTE anisotropy in Sc_2SiO_5 was found to be notably low as we can see in Figure 5.7. Both Y_2SiO_5 and Sc_2SiO_5 form in the same $C2/c$ crystal structure. In addition, both Y and Sc have similar valence electron configuration of $nd^1(n+1)s^2$, where n is the principal quantum number (n=3 and 4 in Sc and Y, respectively). There is no data in the literature that provide insights into the plausible reasons behind the large difference in the CTE anisotropy between Y_2SiO_5 and Sc_2SiO_5. An understanding of factors that govern the CTE anisotropy is crucial for tailoring the material for targeted applications. [186, 187, 24].

We hypothesise that the difference in the degree of CTE anisotropy between Y_2SiO_5 and Sc_2SiO_5 is due to the electronic structure difference originating from the $4d^15s^2$ and $3d^14s^2$ valence electronic configurations of Y and Sc, respectively. I tested the hypothesis by performing density functional theory (DFT) calculations to uncover

Table 5.3: Total energy difference (ΔE, meV/atom) from DFT calculations with respect to the lowest energy structure in RE_2SiO_5, where RE=Sc, Y, and La. Space groups were determined using the `FINDSYM` web application. Structures with ΔE=0 (bold face font) represent the ground state structure for that compound.

Space Group	Sc_2SiO_5	Y_2SiO_5	La_2SiO_5
$C2/c$	**0**	**0**	24.93
$P2_1/c$	7.63	6.98	**0**

hitherto unknown electronic structure trends with implications in the design of EBC materials for protecting high temperature gas turbine engine components. The key data of interest is the electronic structure of Sc_2SiO_5 and Y_2SiO_5 compounds. For completeness, we also consider the La_2SiO_5 compound in the analysis. The valence electron configuration of La can be written as $5d^16s^2$, which is also isoelectronic to Y and Sc. More specifically, we calculate the total and atom-projected (local) density of states (DOS) in the ground state and hypothetical structures. While the ground state structures will contain atoms in their equilibrium volume and fully relaxed positions, the hypothetical structures carry a unique meaning in this work. In the two hypothetical structures that we have considered, we fix the unit cell volume to be that of Sc_2SiO_5 (741.19 Å^3) and La_2SiO_5 (980.98 Å^3). In contrast, the unit cell volume for Y_2SiO_5 in its equilibrium structure is 844.59 Å^3. We then fully substitute the Sc- and La-atoms with that of Y-atoms and relax only their atomic positions until the total interatomic forces are negligibly small. These calculations reveal the electronic structure difference between Y-4d and Sc-3d orbitals in the $C2/c$ crystal structures, as well as the dependence of Y-4d orbital bandwidth on the unit cell volume.

In Table 5.3, we show the total energy difference (ΔE, in meV/atom) between the two monoclinic crystal structures for Sc_2SiO_5, Y_2SiO_5 and La_2SiO_5. In the case of Sc_2SiO_5 and Y_2SiO_5, the $C2/c$ space group is found to be the lowest energy structure, which is in agreement with the experimental data. [24] However in La_2SiO_5, the $P2/c$ space group is the lowest energy structure, which is also in agreement with the experimental data. The total energy difference data for the rest of the RE_2SiO_5 compounds are given in the Supplemental Information.

In Figure 5.8, we show the diverse RE-O coordination environment surrounding the RE1-O and RE2-O atoms in the $C2/c$ structure, where RE=Sc, Y, and La. In the case of Sc_2SiO_5 compound, we have two six-coordinated Sc1-O and Sc2-O polyhedra. In Y_2SiO_5, where the ionic radius of Y^{3+} is greater than that of Sc^{3+}, the Y1-O is a seven-coordinated polyhedron whereas the Y2-O remains six-coordinated (similar to the Sc2-O environment). Finally, in La_2SiO_5 (where the ionic radius of La^{3+} is largest of the three cations) both La1- and La2-sites are seven-coordinated to the neighboring O atoms.

In Figure 5.9, we show the total and local DOS for Sc_2SiO_5 and Y_2SiO_5 in their ground state $C2/c$ structures. We have also included the DOS for La_2SiO_5 in the higher energy $C2/c$ structure (see Table 5.3). All three compounds are predicted as wide band gap insulators. The DFT-GGA level band gaps for Sc_2SiO_5, Y_2SiO_5 and (hypothetical) La_2SiO_5 in the $C2/c$ structures are predicted as 4.6, 5.0, and 4.7 eV, respectively. We believe this is an underestimation of the true or experimentally measured band gap due to the well-documented limitations of the semi-local functionals to adequately describe the excited state property. [188, 189] Nonetheless, this discrepancy does not affect the key goals of this work. From Figure 5.9, we infer that the the top of the valence bands and the bottom of the conduction bands are dominated by the oxygen-2p and RE-d states, respectively. One of the insights gleaned from the local DOS is the width of the RE-d orbitals in the conduction bands. Our definition of the d-orbital bandwidth is given in Figure 5.9. As expected, the Sc-3d orbitals in the conduction bands have a narrow bandwidth (Figure 5.9a) compared to that of the Y-4*d* (Figure 5.9b) and La-5*d* orbitals (Figure 5.9c). Intriguingly, the Y-4*d* and La-5*d* orbitals in the conduction bands have similar bandwidths. We expected the Y-4*d* orbitals to have a relatively narrower bandwidth compared to that of the more spatially extended La-5*d* orbitals. We conjecture that the large Y-4*d* orbital bandwidth in the conduction bands is one of the key reasons behind the relatively large degree of CTE anisotropy experimentally observed in the Y_2SiO_5 compound compared to that of the Sc_2SiO_5 counterpart. [24]

I also performed two additional DFT calculations to test whether one can reduce the Y-4*d* bandwidth in the conduction bands. First, I fully substituted Y in the place of Sc in Sc_2SiO_5 but we constrained the unit cell volume to that of the Sc_2SiO_5 compound (mimicking hydrostatic pressure). I relaxed the internal coordinates until the interatomic forces were small. This constrained simulation cell was 73.3 meV/atom higher in energy compared to the Y_2SiO_5 compound in its equilibrium volume. In Figure 5.10a, I show the local Y-O coordination environment that defines the two crystallographically independent Y-sites for this hypothetical compound. In the ground state structure (shown in Figure 5.10b), only the Y1-O polyhedron was coordinated to seven O-atoms, whereas in this hypothetical compound both Y-O polyhedra are seven-coordinated. The increase in effective number of neighboring O-atoms is likely due to the smaller unit cell volume, which forces the atoms to get closer to one another. We then calculated the total and local DOS for the hypothetical Y_2SiO_5 compound in the reduced unit cell volume. The spectra is shown in Figure 5.11a. We find that the Y-4*d* bandwidth increased marginally relative to that of the Y_2SiO_5 in its ground state structure (see Figure 5.11b). These results indicate that the hydrostatic pressure is not a viable route to reduce the Y-4*d* bandwidth in the conduction bands.

Alternatively, I also fully substituted Y in the place of La in the metastable La_2SiO_5 $C2/c$ structure. In this case, I constrained the unit cell volume to that of the La_2SiO_5 compound. Similar to the previous simulation, we only relaxed the internal coordinates until the interatomic forces were small. This hypothetical structure was 84.1 meV/atom higher in energy than the equilibrium Y_2SiO_5 structure. In Figure 5.10c, we show the local Y-O coordination environment for this hypothetical compound. Both Y-O polyhedra reduced its effective number of neighboring O atoms to six in this structure. The larger unit cell volume did not require the atoms to get closer to one another. I then calculated the total and local DOS. The results are shown in Figure 5.11c. We find that the Y-4d bandwidth reduced when compared to that of the ground state Y_2SiO_5, which is encouraging. However, the reduction was not dramatic to match that of the Sc-3d bandwidth in the Sc_2SiO_5 compound.

5.6 Summary

In summary, the approach combining DFT and ML is demonstrated as a method to accelerate the discovery of EBC materials in the RE_2SiO_5 search space. The DFT calculations reveal that the ΔE data contain insights that are correlated with the energetics promoting polymorph formation. The ML work highlighted the potential of data-driven techniques in rapidly predicting CTE, a task that typically requires significant computational resources. Acknowledging the fact that anisotropy in CTE could be an additional cause for microcracking [186, 187], I used DFT to gain insights on the anisotropy patterns in RE_2SiO_5. The DFT calculations to reveal a previously unknown correlation between the d-orbital bandwidth and unit cell volume in the $C2/c$ structure for three RE_2SiO_5 compounds (Sc_2SiO_5, Y_2SiO_5 and La_2SiO_5). In order to reduce the Y-4*d* bandwidth, which I conjecture will also reduce the degree of CTE anisotropy, the Y_2SiO_5 compound should form in an open structure with a reduced Y-O effective coordination number in both polyhedral units. This will likely require materials synthesis and processing using non-equilibrium techniques. Future work can focus on the coupling between unit cell volume and point defects (eg., introduce Y- and O-vacancies), which can further affect the local DOS and modify the Y-4*d* bandwidth in the conduction bands.

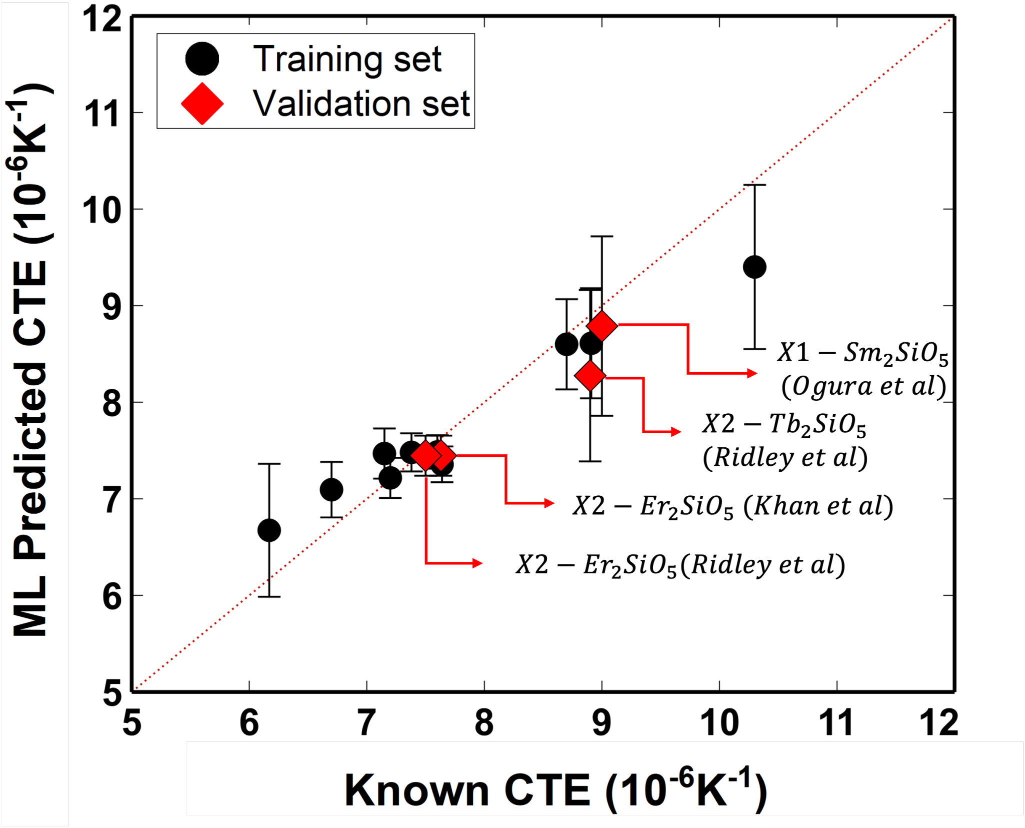


Figure 5.5: Performance of the trained eSVR model. The X- and Y-axes are the known and ML predicted ABCTE data, respectively. The error bar represent the standard deviation from the ensemble of ML models. The red dashed line represents the X=Y line and the data points falling on this line indicate perfect agreement between the ML models and the known data. The black dots are the experimental data points used to train the models. The red diamonds represent the validation data that were not used to train the ML models. The validation data points include X2-$C2/c$ Er_2SiO_5 reported by Khan *et al.* [184] and Ridley *et al.* [24], X2-$C2/c$ Tb_2SiO_5 by Ridley *et al.* [24] and X1-$P2_1/c$ Sm_2SiO_5 reported by Ogura *et al.* [185].

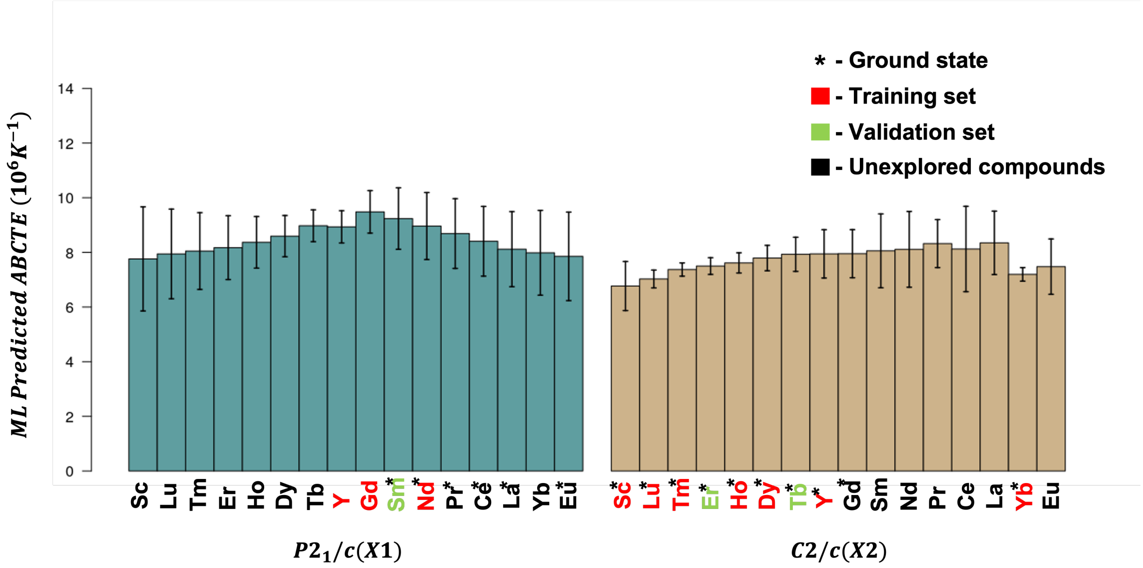


Figure 5.6: The ML predicted CTE with uncertainties from bootstrapping for all the RE_2SiO_5 compounds.The polymorphs predicted to be ground state (ΔE=0) by DFT calculations are marked by asterisks. The polymorphs marked in red are included in the training data and green marks validation set. The ones marked in black are unexplored compounds.

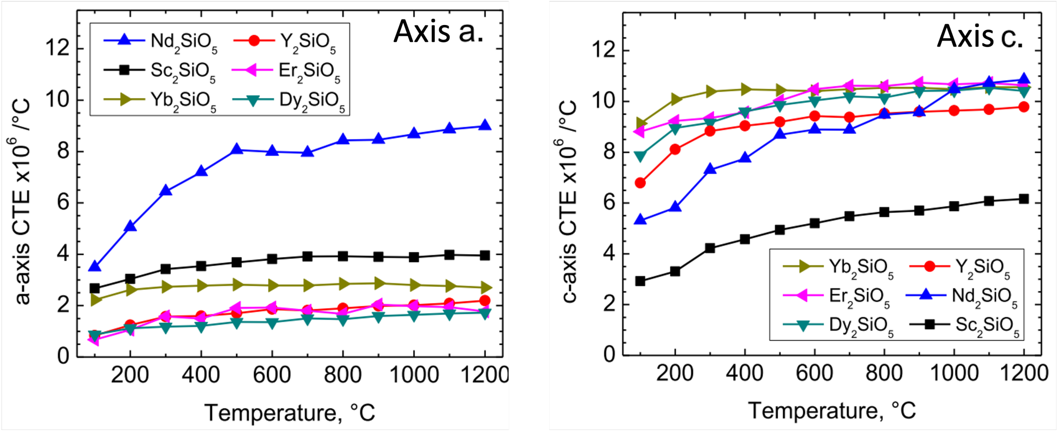


Figure 5.7: CTE Anisotropy trends in RE_2SiO_5 (RE=Sc,Y,Yb,Nd,Er,Dy) reported by Ridley et al [24].

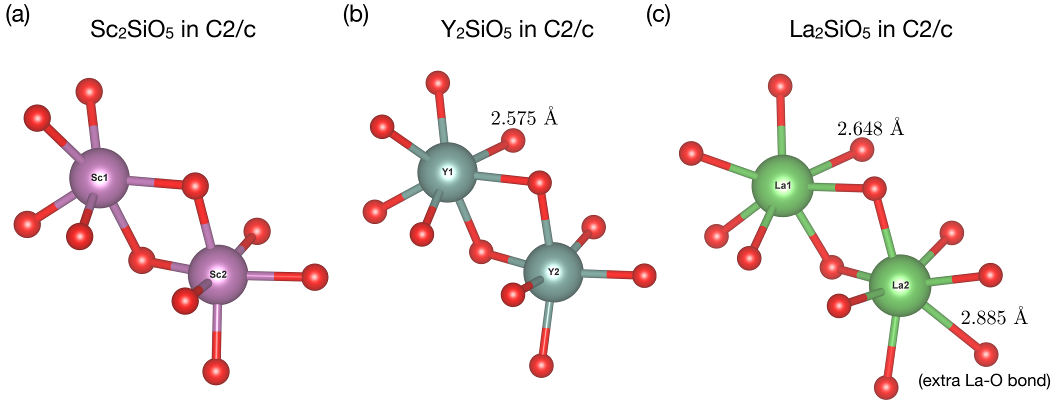


Figure 5.8: RE-O coordination environment in (a) Sc_2SiO_5 and (b) Y_2SiO_5 in their respective ground state $C2/c$ structures. In (c), we show the local La-O environment for the La_2SiO_5 compound in its metastable $C2/c$ structure. The longest bond lengths that define the *seventh* RE-O bond are highlighted in the Figure for the Y_2SiO_5 and La_2SiO_5 compounds.

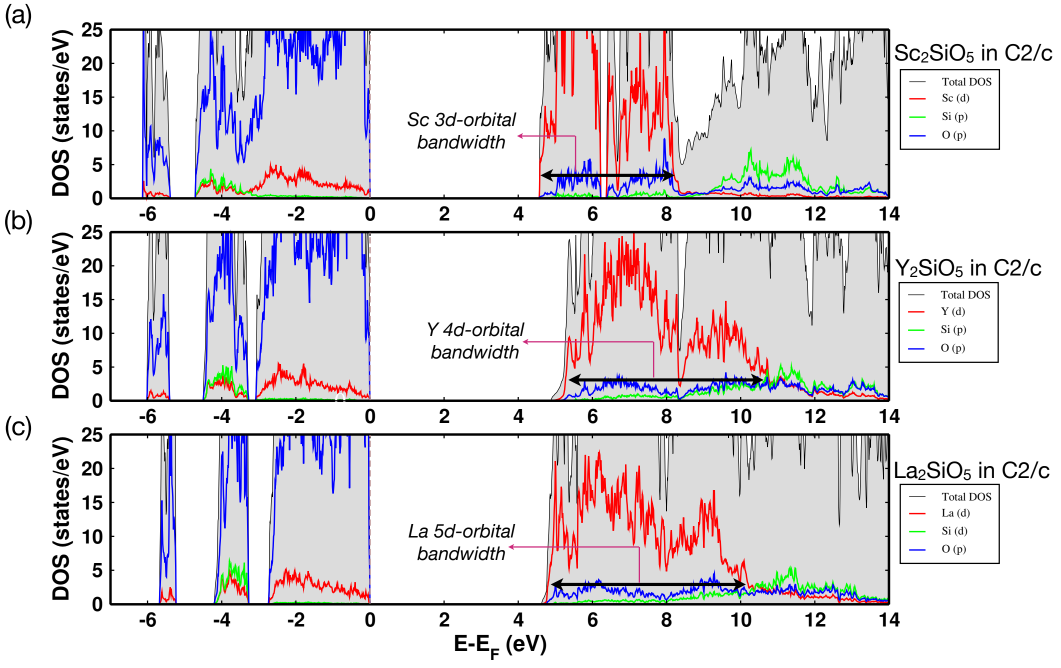


Figure 5.9: Total and local DOS for (a) Sc_2SiO_5 and (b) Y_2SiO_5 in their respective ground state $C2/c$ structures. In (c), we show the total and local DOS for La_2SiO_5 in its metastable $C2/c$ structure. E_F is the Fermi energy.

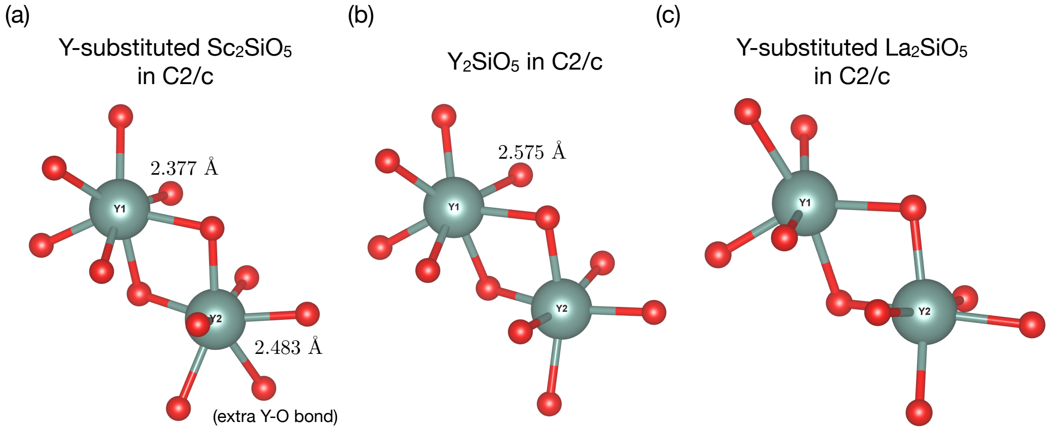


Figure 5.10: RE-O coordination environment in (a) Y-substituted Sc_2SiO_5 in the hypothetical $C2/c$ structure with a smaller unit cell volume, (b) Y_2SiO_5 in its ground state $C2/c$ structure with equilibrium volume, and (c) Y-substituted La_2SiO_5 in its hypothetical $C2/c$ structure with a larger unit cell volume. The longest bond lengths that define the seventh RE-O bond are highlighted in the Figure.

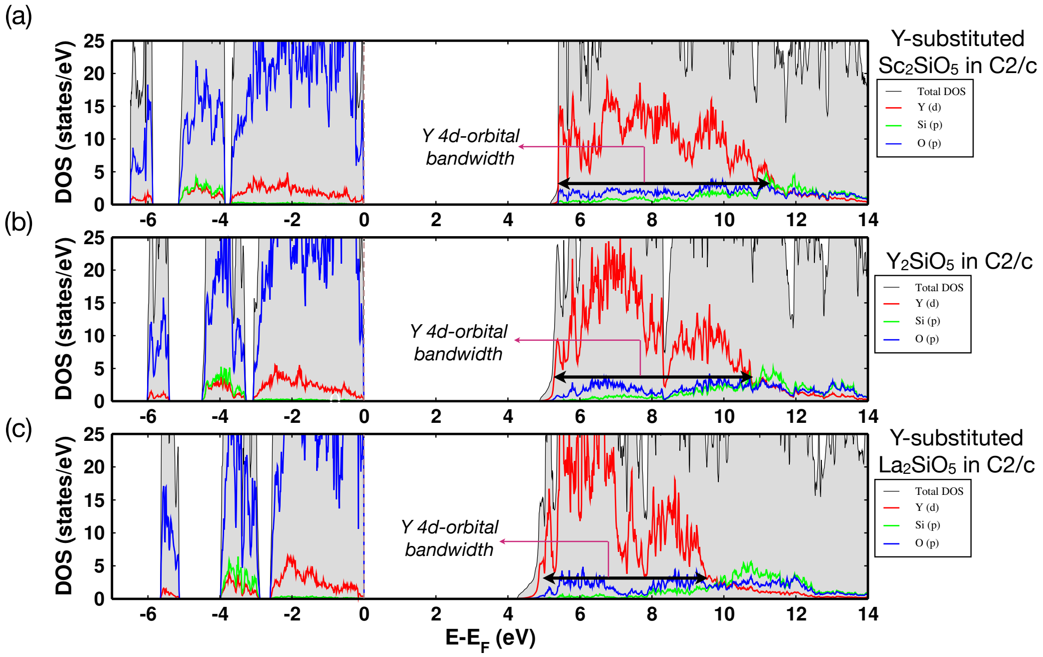


Figure 5.11: Total and local DOS for (a) Y-substituted Sc_2SiO_5 in the hypothetical $C2/c$ structure with a smaller unit cell volume, (b) Y_2SiO_5 in its ground state $C2/c$ structure with equilibrium volume, and (c) Y-substituted La_2SiO_5 in the hypothetical $C2/c$ structure with a larger unit cell volume. E_F is the Fermi energy.

Chapter 6

Holistic CTE model for the RE-Si-O crystal chemistry

This chapter focuses on establishing a holistic description of the structure-CTE relationship for the three major materials classes in the RE-Si-O crystal chemistry (that include disilicates, monosilicates, and silicate apatites). Such a description provides a framework for developing a fundamental understanding of key factors that impact the CTE. Deriving a unified theoretical equation/model that can comprehensively describe the relationship across various material classes is a non-trivial task. Currently, such a comprehensive model is absent from the literature. To this end, I will demonstrate a holistic ML model developed based on polyhedral features that can be used across different compounds (both single- and multi-component) belonging to different material classes. Through advanced interpretability methods, this model not only predicts but also elucidates the functional relationships governing the thermal expansion behavior in RE-Si-O crystal systems. The research discussed in this chapter is an important step towards design optimization of the RE-Si-O crystal chemistry for EBC applications.

6.1 Chapter Organization

First, I motivate the need for a holistic description of CTE. Then, I discuss the RE silicate apatites. I discuss the DFT total energy trends relevant to these compounds before transitioning to the task of constructing a holistic ML model for CTE. I then apply the prediction intervals method, as described in subsection 3.2.4 to quantify the uncertainty associated with the predictions. I then use post hoc model interpretability methods to reveal the learned relationship between the polyhedral descriptors and the CTE. Lastly, an interpretable mathematical equation for CTE as a function of polyhedral descriptors is formulated using grammatical evolution [190, 191, 192], guided by the insights from post hoc model interpretability.

6.2 Introduction

In preceding chapters, I discussed in some detail two distinct RE-Si-O material classes: $RE_2Si_2O_7$ (Chapter 4) and RE_2SiO_5 (Chapter 5). I demonstrated that one can establish structure-CTE relationship within each of those materials family based on unit cell parameters. The predictive capability of these models was also demonstrated. While predictive models are invaluable, they provide little insights into the understanding of CTE. Such an understanding would two purposes: (1) It would not only enhance the model's predictive accuracy but also provide design rules that can in essence, uncovering physical insight would enable researchers to dissect the complex interplay between crystal structure and CTE, thus paving the way for more targeted research and development. Building on this need for mechanistic understanding, the central challenge then hinges on finding a common representation for CTE that can effectively capture the interplay between structure and CTE. This representation should encapsulate mechanistic information related to phonon dispersion curves, which fundamentally dictates the CTE. However, the challenge is magnified by the rich structural diversity within the RE-Si-O crystal chemistry, known for its extensive polymorphism. Finding a descriptor set that is both mechanistically informative and universally applicable across different RE-Si-O material classes is a daunting but necessary task.

In pursuit of this objective, numerous empirical models in existing literature have made commendable progress. These models often capitalize on the fundamental relationship between thermal expansion and the strength of chemical bonds. Models by Megaw, Cameron, and Hazen have been particularly noteworthy in this context, incorporating key parameters like valence, stretching frequency of bonds, ionicity factor, and coordination number to predict CTE [15, 16, 17, 18]. Adding to this body of work, Zhang *et al.* recently introduced a semi-empirical approach grounded in principles of chemical bonding. Their model effectively predicts thermal expansion coefficients for both simple and complex crystals by considering lattice energy and bond geometric descriptors [84]. However, these existing approaches exhibit a notable limitation: they often rely on pairwise approximations involving binary compounds to make predictions for more complex, multi-component systems [84]. This approach, albeit simple and an excellent starting point, can become unreliable for complex materials, especially in the case of entropy stabilized compounds, when the multicomponent compound lacks features or characteristics that are present in the binary compounds. This limitation underscores the need for a descriptor set that is both informative at the mechanistic level and universally applicable across the diverse and complex landscape of RE-Si-O materials.

I hypothesize that the local polyhedral features can serve as the common descriptor that has a more meaningful representation of the crystal chemistry when compared to the traditional RE ionic radii description of CTE. This approach is akin to the semi-empirical models I mentioned earlier. I recognize that a lack of extensive data for less-studied material classes is a significant hurdle, a challenge that could be addressed by crafting a CTE model applicable across diverse classes. For example, there are only five known compounds in the RE silicate apatites class for which CTE data is available in the existing literature. Moreover, multi-component compounds in the $RE_2Si_2O_7$ and RE_2SiO_5 have recently garnered interest due to their potential for reduced thermal conductivity via the high entropy effect [193, 24, 194, 195]. Therefore, I consider generating the data for RE silicate apatites class as an important step towards our goals. Before diving deep into the intricacies of the holistic ML model, I will discuss the data I have gathered on RE silicate apatites and DFT total energy results.

6.2.1 DFT calculations of the Rare-Earth Silicate Apatites

The apatite silicate structure is inherently non-stoichiometric, with a general formula of $RE_{9.33}Si_6O_{26}$. This structure type offers a fertile playground for cation substitutions, allowing for both alkali metals (with formula $RE_9A_1Si_6O_{26}$, where A can be Li, Na, K, Rb, or Cs) and alkaline earth metals (with formula $RE_8AE_2Si_6O_{26}$, where AE is Be, Mg, Ca, Sr, or Ba) to substitute the RE-site in the lattice. These compounds typically crystallizes in the hexagonal $P6_3/m$ space group [70]. For simulating $RE_{9.33}Si_6O_{26}$, I constructed a 1×1×3 supercell containing 124 atoms in order to accommodate the non-stoichiometry. I used the unit cell containing 42 atoms for the $RE_9A_1Si_6O_{26}$ and $RE_8AE_2Si_6O_{26}$crystal chemistry. I calculated ΔE (in meV/atom) with respect to the lowest energy structure using DFT calculations. The DFT calculations were performed using the planewave pseudopotential code `Quantum ESPRESSO` [152]. The PBEsol exchange-correlation functional [153] was used and the core and valence electrons were treated with ultrasoft pseudopotentials [154]. The Brillouin zone integration was performed using a Monkhorst-Pack [155] k-point mesh centered at Γ and 60 Ry plane-wave cutoff for wavefunctions (600 Ry kinetic energy cutoff for charge density and potential). The scalar relativistic pseudopotentials were taken from the PSLibrary [156]. The atomic positions and the cell volume were allowed to relax until an energy convergence threshold of 10^{-8} eV and Hellmann-Feynman forces less than 2 meV/Å, respectively, were achieved. The $4f$-states for the rare-earth elements are considered as core states in our calculations. The converged crystal structures were visualized in `VESTA` [79] and the space groups were determined using `FINDSYM` [157].

The calculations show that all the $RE_{9.33}Si_6O_{26}$ compounds converged in the $P3$ space group. However, the $RE_8AE_2Si_6O_{26}$ converged in three subgroups of $P6_3/m$: $P\bar{6}$, $P\bar{3}$ and $P6_3$. Whereas the $RE_9A_1Si_6O_{26}$ converged in $P\bar{6}$ and Pm space groups. One of the reasons for the discrepancy could be attributed to the smaller size of the supercells and/or unit cells that are not sufficient to mimic the disorder. The plots describing the energetic competition for the all of the AE groups (AE=Ba, Be, Ca, Mg, Sr) and A groups (A=Li, Na, K, Rb, Cs) are included in the Appendix (section C.2). Here, I discuss only a few plots that are representative of the trends the compounds follow. In the compounds of $RE_8AE_2Si_6O_{26}$, the influence of AE ions on the ground state and metastable structures are significant. In Figure 6.1, the total energy difference per atom (ΔE) with respect to the lowest energy structure is shown for $RE_8AE_2Si_6O_{26}$ (AE=Ba). In the figure, ΔE=0 mev/atom is the lowest energy structure (ground

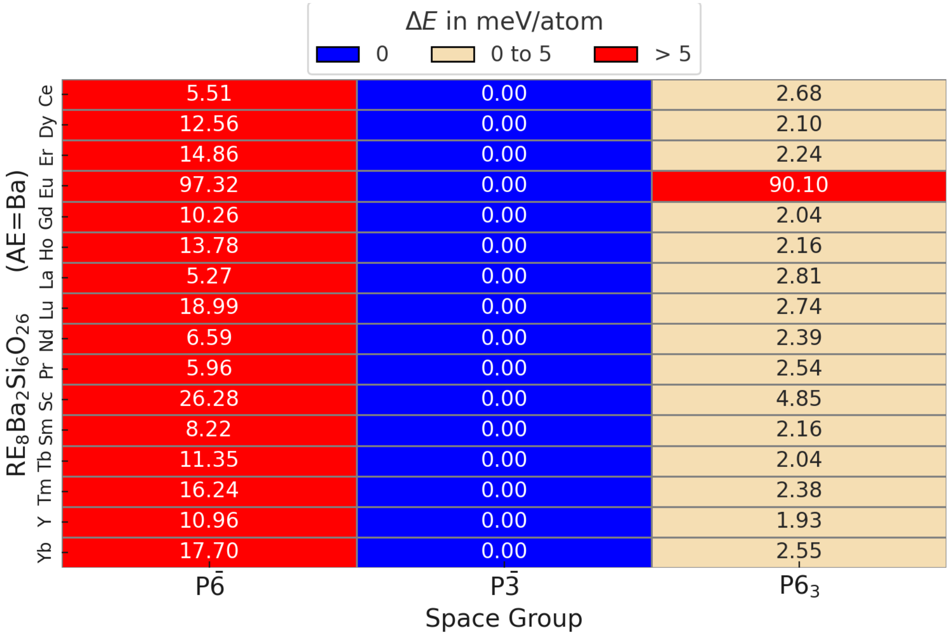


Figure 6.1: Total energy difference, in from DFT calculations with respect to the lowest energy structure in $RE_8AE_2Si_6O_{26}$ for AE=Ba,where space groups given in parentheses indicate the final converged structure when the tolerance is set at 0.0001 or lower in FINDSYM. ΔE=0 signifies the ground structure. When ΔE is within a range of approximately 5 meV/atom, it suggests a close energetic competition with the ground state structure.

structure).

For AE=Ba, the predominant ground state structure across different RE is $P\bar{3}$, with almost every compound having a closely competing metastable structure in $P6_3$. A similar trend can also be seen in the case of AE=Sr. On the other hand, when AE=Be and Mg, the predominant ground state structure shifts to $P6_3$ with $P\bar{3}$ being the close competitor. AE=Ca (Figure 6.2), is the only case where there is no clear dominant energetic preference. Here, out of the 16 compounds, 10 of them (RE=Nd, Dy, Er, Eu, Ho, Pr, Sm, Tb, Y, Gd) have $P\bar{3}$ structure as their lowest energy configuration. Meanwhile, the $P6_3$ structure is the ground state for the remaining 6 compounds (RE=La, Ce, Sc, Tm, Yb, Lu).

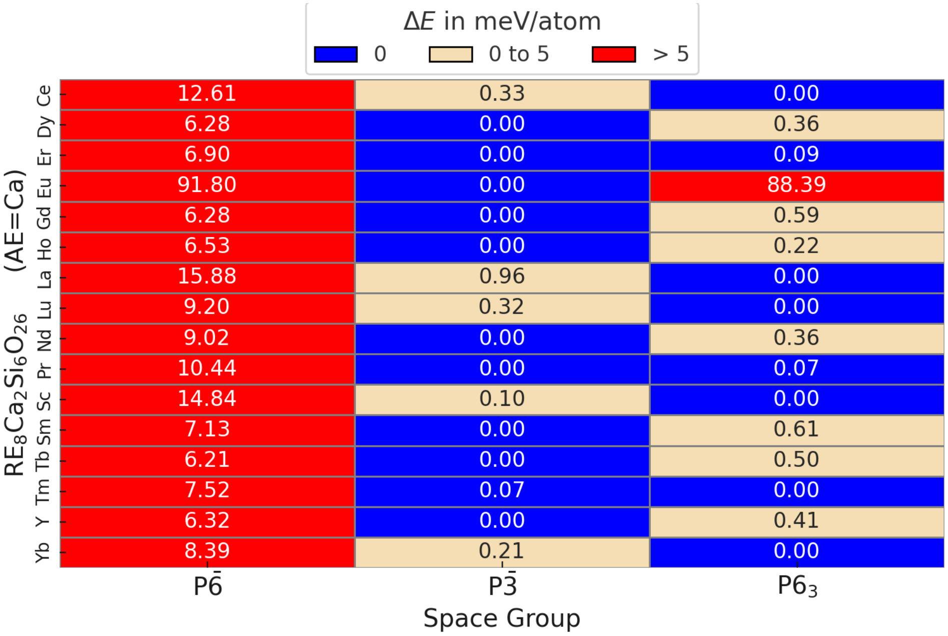


Figure 6.2: Total energy difference, in from DFT calculations with respect to the lowest energy structure in $RE_8AE_2Si_6O_{26}$ for AE=Ca,where space groups given in parentheses indicate the final converged structure when the tolerance is set at 0.0001 or lower in FINDSYM. ΔE=0 signifies the ground structure. When ΔE is within a range of approximately 5 meV/atom, it suggests a close energetic competition with the ground state structure.

In contrast to $RE_8AE_2Si_6O_{26}$, for the compounds of $RE_9A_1Si_6O_{26}$, where A includes alkali metals, the predominant ground state structure is Pm across all types of A ions, without any close competing metastable structures. Exceptions to this general trend include $RE_9Rb_1Si_6O_{26}$ for RE elements Gd, Sm, and Y, as well as $RE_9Na_1Si_6O_{26}$ for RE elements La and Y (shown in Figure 6.3). In these specific compounds, the ground state structure is $P\bar{6}$.

Overall, within the $RE_8AE_2Si_6O_{26}$ series, the AE ions play a significant role in determining the stable structure. For instance, the presence of AE ions such as Ba and Sr predominantly leads to the stabilization of the $P\bar{3}$ structure, with a close competitor in the $P6_3$ phase. A shift is observed for AE ions like Be and Mg, which predomi-

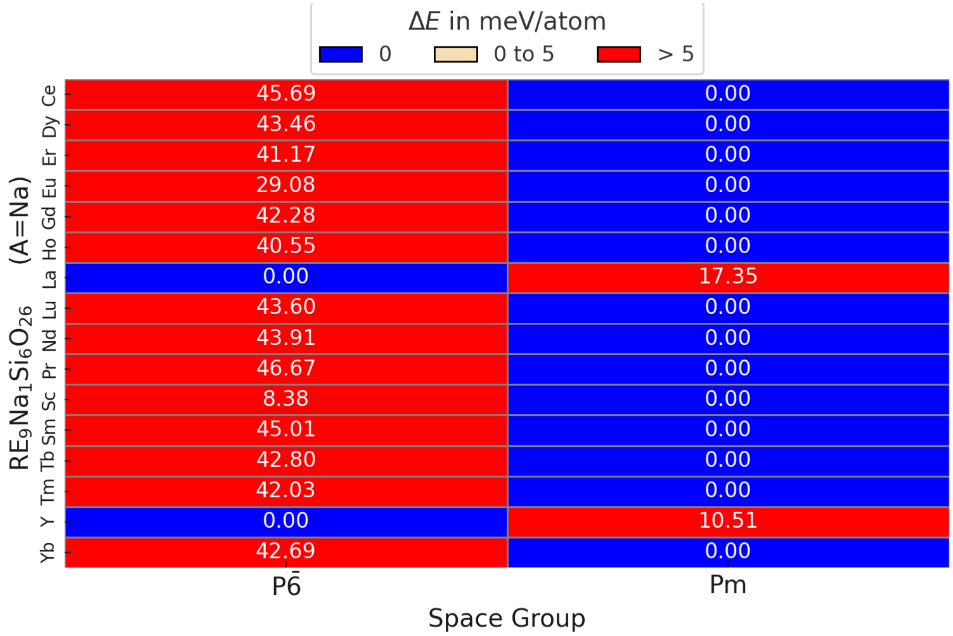


Figure 6.3: Total energy difference, in from DFT calculations with respect to the lowest energy structure in RE$_9$A$_1$Si$_6$O$_{26}$ for A=Na,where space groups given in parentheses indicate the final converged structure when the tolerance is set at 0.0001 or lower in FINDSYM. ΔE=0 signifies the ground structure. When ΔE is within a range of approximately 5 meV/atom, it suggests a close energetic competition with the ground state structure.

nantly stabilize in the $P6_3$ structure. However, for AE=Ca, there's a mix, with no clear dominant structure, showcasing both $P\bar{3}$ and $P6_3$ as ground states.

On the other hand, in the RE$_9$A$_1$Si$_6$O$_{26}$ series, where A represents alkali metals, the influence of the A ion on the structural preference is less pronounced. The series demonstrates a more consistent preference for the Pm structure across different alkali metals. Exceptions arise, notably with A=Rb and Na, where certain RE elements deviate and favor the $P\bar{6}$ structure. These observations underscore the more dominant influence of AE in determining structural preferences compared to A ions. While overarching patterns are evident, specific elemental combinations can introduce variations, underscoring the complex interplay of factors in these compounds.

6.3 Structure-CTE relationship of RE-Si-O using a holistic ML model

This section will discuss the details of the descriptor generation, ML model and the results that include the predictions for single and multi-component compounds in RE-Si-O, post-hoc model explanation and equation generation for volumetric CTE in RE-Si-O.

6.3.1 Dataset and Descriptor Generation

The dataset for training the holistic ML model to predict CTE for multiple classes of RE silicates is shown in Table 6.1. The data contain 21 RE disilicates, 11 RE monosilicates, 3 RE-Si apatites (RE= La, Dy, Nd) and 2 alkaline earth (AE) metal bearing RE-Si apatites (AE=Ca and RE=Y, Yb) for which experimental volumetric CTE values are available. I restricted the data collection to only consider the single component compounds because I did not perform DFT calculations to mimic pseudo-binary or pseudo-ternary solid solutions. The literature source of the CTE data is shown in the Table 6.1 for each compound.

Next, I visualize the dataset using decision trees. While the previous chapters focused on classification of polymorphs within individual classes of RE disilicates, RE monosilicates, and RE containing apatites, this chapter takes a more holistic approach. Here, the objective shifts to classifying the overarching material classes themselves. Despite the change in focus, the value of interpretability remains constant, allowing us to understand relationships not just within but also between these distinct material classes. In this chapter, I aim to draw broader conclusions that can inform decisions

Table 6.1: Thermal Expansion Coefficients Data From Literature.

Compound	Experimental ABCTE ($\times 10^{-6}$ K^{-1})	Temperature range (K)	Reference
A-$P4_1$ $La_2Si_2O_7$	14	303-1373	Fernández-Carrión et al. [196]
A-$P4_1$ $Pr_2Si_2O_7$	11.8	303-1573	Fernández-Carrión et al. [196]
A-$P4_1$ $Nd_2Si_2O_7$	10.5	303-1473	Fernández-Carrión et al. [196]
A-$P4_1$ $Sm_2Si_2O_7$	11.55	573-1248	Ayyasamy *et al.* [197]
A-$P4_1$ $Ce_2Si_2O_7$	12.4	573-1248	Strzelecki *et al.* [198]
δ-$Pnma$ $Gd_2Si_2O_7$	7.3	303-1873	Fernández-Carrión *et al.* [196]
δ-$Pnma$ $Dy_2Si_2O_7$	7.7	303-1423	Fernández-Carrión *et al.* [196]
δ-$Pnma$ $Y_2Si_2O_7$	8.1	293-1673	Fernández-Carrión *et al.* [196]
β-$C2/m$ $Er_2Si_2O_7$	3.9	303-1873	Fernández-Carrión *et al.* [196]
β-$C2/m$ $Yb_2Si_2O_7$	4	303-1873	Fernández-Carrión *et al.* [196]
β-$C2/m$ $Lu_2Si_2O_7$	4.2	303-1823	Fernández-Carrión *et al.* [196]
β-$C2/m$ $Sc_2Si_2O_7$	5.4	303-1873	Fernández-Carrión *et al.* [196]
β-$C2/m$ $Y_2Si_2O_7$	4.1	293-1673	Fernández-Carrión *et al.* [196]
γ-$P2_1/c$ $Ho_2Si_2O_7$	4.2	303-1748	Fernández-Carrión *et al.* [142]
γ-$P2_1/c$ $Y_2Si_2O_7$	3.9	293-1473	Fernández-Carrión *et al.* [142]
G-$P2_1/c$ $La_2Si_2O_7$	6.4	303-1073	Fernández-Carrión *et al.* [196]
G-$P2_1/c$ $Pr_2Si_2O_7$	6.8	303-1648	Fernández-Carrión *et al.* [196]
α-$P\bar{1}$ $Gd_2Si_2O_7$	8.3	303-1573	Fernández-Carrión *et al.* [196]
α-$P\bar{1}$ $Dy_2Si_2O_7$	8.5	303-1648	Fernández-Carrión *et al.* [196]
α-$P\bar{1}$ $Y_2Si_2O_7$	8	293-1473	Fernández-Carrión *et al.* [196]
X1-$P2_1/c$ Gd_2SiO_5	10.3	473–1623	Al Nasiri *et al.* [180]
X1-$P2_1/c$ Nd_2SiO_5	8.91	303-1373	Ridley *et al.* [181]
X1-$P2_1/c$ Y_2SiO_5	8.7	473–1623	Fukuda *et al.* [182]
X2-$C2/c$ Dy_2SiO_5	7.6	303-1373	Tian *et al.* [183]
X2-$C2/c$ Ho_2SiO_5	7.38	303-1373	Tian *et al.* [183]
X2-$C2/c$ Lu_2SiO_5	6.7	473–1623	Al Nasiri *et al.* [180]
X2-$C2/c$ Sc_2SiO_5	6.17	303-1373	Ridley *et al.* [24]
X2-$C2/c$ Tb_2SiO_5	8.9	303-1373	Tian *et al.* [183]
X2-$C2/c$ Tm_2SiO_5	7.64	303-1373	Tian *et al.* [183]
X2-$C2/c$ Y_2SiO_5	7.7	303-1373	Fukuda *et al.* [182]
X2-$C2/c$ Yb_2SiO_5	7.2	473–1623	Al Nasiri *et al.* [180]
$P3$ $Nd_{9.33}Si_6O_{26}$	9.4	303-1373	Okudera *et al.* [199]
$P3$ $Dy_{9.33}Si_6O_{26}$	9.64	303-1373	Misture *et al.* [200]
$P3$ $La_{9.33}Si_6O_{26}$	9.4	303-1373	Fukuda *et al.* [201]
$P6_3$ $Ca_2Y_8(SiO_4)_6O_2$	8.7	303-1373	Stokes *et al.* [202]
$P6_3$ $Ca_2Yb_8(SiO_4)_6O_2$	8.54	303-1373	Stokes *et al.* [202]

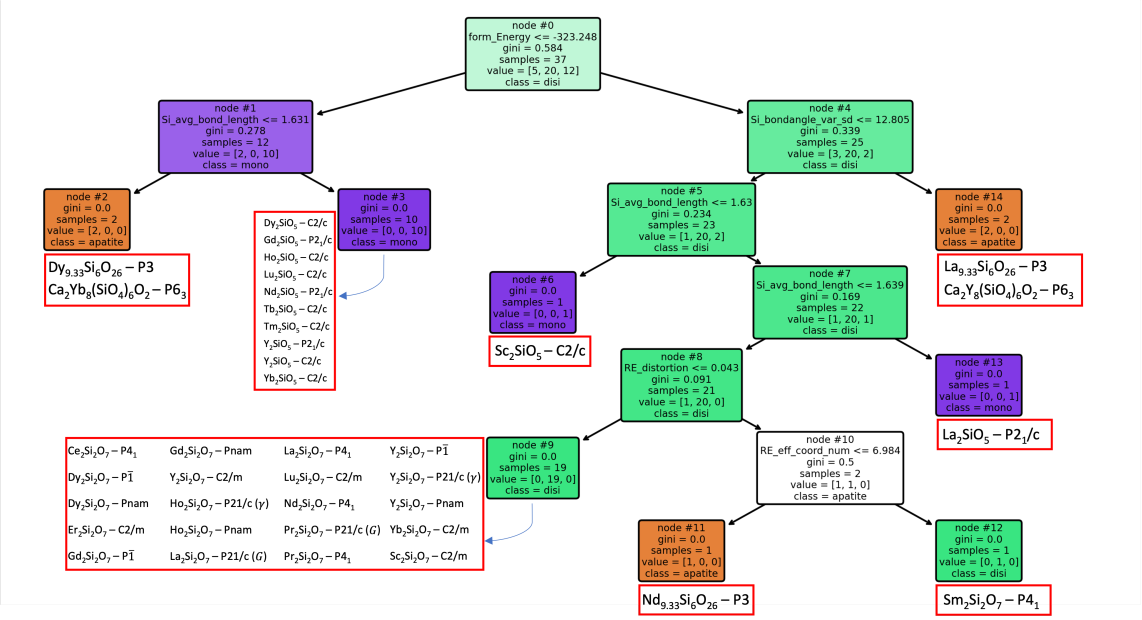


Figure 6.4: Decision trees for determining CTE based on GINI impurity on the training data shown in table Table 6.1. At each node, a check occurs and if true proceeds to the left and vice-versa until it reaches the leaf node that displays the classification. Under each classification, the compounds belonging to the corresponding material classes are listed.

and strategies in multiple domains of RE-Si-O crystal chemistry.

The decision tree built on this dataset classified the three classes using formation energy as the root node and 6 decision nodes generating a total of 8 unique decision paths. Classification of disilicates involves one path: [If form_Energy is > -323.25 AND Si_bondangle_var_sd ≤ 12.805 AND RE_eff_coord_num > 5.772 AND Si_avg_bond_length ≤ 1.639 AND RE_distortion $\leq 0.043 \rightarrow$ Disilicates]. $Sm_2Si_2O_7$ in $P4_1$ is the only compound that does not follow this path. RE_distortion ≤ 0.043 is the decision node that separates this compound from the rest of the disilicates. Similar to disilicates, the classification of monosilicates predominantly involves one path: [If form_Energy ≤ -323.24 AND Si_avg_bond_length $> 1.631 \rightarrow$ Monosilicates]. However, Sc_2SiO_5-X2 and La_2SiO_5-X1 does not follow this path as they are separated from the rest of the monosilicates by the decision node form_Energy $\leq$ -323.248.

Unlike the other two classes, the classification of apatites does not involve one single prominent decision path. In the training data, 40% of the apatites are classified using Si_avg_bondlength $>$ 1.631 condition; 40% are classified using Si_bondangle_var_sd $>$ 12.805 condition; the rest 20% ($Nd_{9.33}Si_6O_{26}$) are classified using a common decision path as $Sm_2Si_2O_7$ $P4_1$ with the only difference being Si_eff_coord_num_sd threshold. This may be due to the proximity of $Sm_2Si_2O_7$ $P4_1$ to the apatite compound $Nd_{9.33}Si_6O_{26}$ in the high-dimensional space, which is shown in the t-SNE plot [203] (Figure 6.5). The t-SNE plot also justifies the compounds Sc_2SiO_5-X2 and La_2SiO_5-X1 not following the path that the rest of the monosilicates follow by highlighting their proximity to the disilicate clusters. Interestingly, alkaline/alkali earth metal bearing RE-Si-apatites are not clustered together with common decision paths, which is contrary to the t-SNE result. Instead, $Dy_{9.33}Si_6O_{26}$ and $Ca_2Yb_8(SiO_4)_6O_2$ $P6_3$ follow the path: [If form_Energy ≤ -323.24 AND Si_avg_bond_length $\leq 1.631 \rightarrow$ Apatite], while $La_{9.33}Si_6O_{26}$ $P3$ and $Ca_2Y_8(SiO_4)_6O_2$ $P6_3$ follow the path: [If form_Energy > -323.25 AND Si_bondangle_var_sd $> 12.805 \rightarrow$ Apatite].

Next I analyze the trends between these descriptors and the RE ionic radii. The ionic radii reported are for the 3+ oxidation state and for a coordination number of 8. The trends shown in Figure 6.6 are representative of the training dataset built for the holistic ML model. I ignored the descriptor AK_avg_bond_length because there are only three apatite data points for which it is non-zero in the training set. From the plots, it can be seen that RE_avg_bond_length is the only variable that shows highly correlated increase with RE ionic radii. This is consistent with the fact that the RE metal-oxide bond strength decreases with increasing ionic radius, which could result in longer bond lengths. The plots of the Si_avg_bond_length, Si_poly_volume

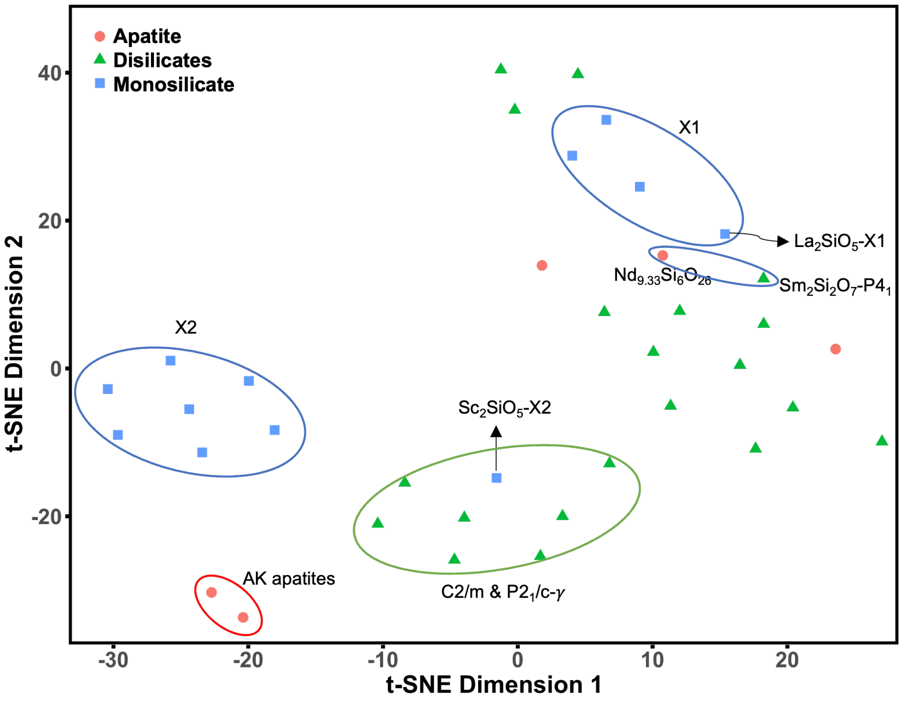


Figure 6.5: Visualization showing two-dimensional projections of our high-dimensional dataset based on the t-distributed stochastic neighborhood embedding (t-SNE) method [203].

and form_Energy (ignoring outliers) show a moderately correlated trends. Decrease in form_Energy with increasing ionic radius could be related to the fact that larger ions have a higher coordination number and can form more stable, lower-energy structures. Si_avg_bond_length increase with larger radii. We see a similar rise in Si_poly_volume as ionic radii increases. The plot of Si_bondangle_var and RE_avg_bond_length_sd is scattered with no clear trend, implying that the bond angle variability is not solely dependent on ionic sizes. Overall, while RE ionic radii provide a useful first-order approximation for predicting CTE, the polyhedral variables offer additional dimensions of information that can be crucial for a more comprehensive understanding of the material properties.

6.3.2 Machine Learning and Bootstrap Resampling

ML is used to establish a quantitative relationship between the down-selected 7 descriptors ($\mathbf{X}$) and ABCTE (Y) assembled from surveying the published experimental literature of known RE disilicate, monosilicate and apatite compounds [77, 140, 142]. Since ABCTE is a numerical quantity, I used regression-based ML methods in this problem. Given a sample of data $(\mathbf{X}, Y)$, the regression problem can be formulated as follows, $Y = f(\mathbf{X}) + \eta$, where η is the random error term. The regression learning was performed using the support vector regression (SVR) algorithm. We used the ϵ-support vector regression with a non-linear Gaussian radial basis function kernel [117] because of its improved generalization ability [164, 165]. SVR hyperparameters such as the penalty term and the insensitive loss function were adjusted to optimize the leave-one-out error. We used the ϵ -SVR method implemented in the `e1071` package [166] within the `RSTUDIO` environment [167].

I built an ensemble of SVR (eSVR) models using the bootstrap resampling method [168] to make predictions as well as evaluate the error bars based on standard error. And, in addition to the standard errors, I also constructed the prediction intervals using the "doubt" algorithm discussed in subsection 3.2.4 which is also based on bootstrapping. The mean and standard deviation of the predictions from the eSVR models are then used as an estimate of the ABCTE and its associated standard errors, respectively. The prediction intervals can be calculated using the quantile of the total error distribution which is essentially the sum of model (Equation 3.12) and observation error Equation 3.16 defined in subsection 3.2.4. Each individual SVR model in the ensemble goes through hyperparameter optimization using a grid-search method consisting of γ (distance penalty) = (0.001, 0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1), and cost = (0.001, 0.01, 0.1, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 100) with

10-fold cross validation provided by the e1071 package in R. [166]

I divided the dataset into training and testing subsets. This was achieved through the common practice of a random train/test split, which helps to prevent overfitting and assess model performance on unseen data. I ensured reproducibility by setting a random seed. I experimented with different random seeds (e.g., 42, 123, and 789) to assess the robustness of our model. The models trained on the different training data were used to predict the ABCTE along with the error bars for the different corresponding test data points that were not included in the training. The performance was characterized by calculating the R^2 value along with the RMSE between the predicted and true values in the independent test set.

Prediction of CTE for single component RE-Si-O compounds

The best performing training data chosen based on the performance on the test data includes a total of 33 data points including the data contains 18 RE disilicates, 11 RE monosilicates, 2 RE-Si apatites and 2 alkaline earth metal bearing RE-Si apatites. The performance of the eSVR model on the training and test data is shown in Figure 6.7. The training data is shown as black dots and test data is shown as blue diamond (experimental measurements) and red triangles (DFT-QHA literature data [77]). The PIs (gray lines) being wider than the standard error is indicative of the prediction interval algorithm accounting for observation error in addition to the model error. In Table 6.2, the compounds used for testing and validation, along with the ML predictions and the two uncertainty metrics (standard error and prediction intervals) are also given. A vast majority of the data points, especially the test points, either lie close to or on the X=Y line indicating good training performance.

Prediction of multi-component $RE_2Si_2O_7$ compounds

To demonstrate the capability of the trained model to make predictions for multi-component compounds in spite of being trained only on single component compounds, it was used to predict the ABCTE for multi-component $(La,Lu,Y,Yb)_2Si_2O_7$ - β-$C2/m$ and $(Dy,Er,Y,Yb)_2Si_2O_7$ - β-$C2/m$ structures obtained from the SQS-DFT calculations performed by Dr. Kyungtae Lee (former postdoc fellow from UVa-MSE materials informatics group). DFT calculations are performed using the planewave pseudopotential code `Quantum ESPRESSO` [152]. The PBEsol exchange-correlation functional [153] was used and the core and valence electrons were treated with ultrasoft pseudopotentials [154]. The Brillouin zone integration was performed using a Monkhorst-Pack [155] k-point mesh centered at Γ. We used the k-mesh size of 5×3×1, 5×3×1 and 3×2×2 for β-$C2/m$, δ-$Pnma$ and G-$P2_1/c$ structures, respectively. The planewave cutoffs were set at 60 Ry and 720 Ry for wavefunctions and kinetic energy, respectively. The scalar relativistic pseudopotentials were taken from the PSLibrary [156]. The atomic positions and the cell volume were allowed to relax until an energy convergence threshold of 10^{-8} eV and Hellmann-Feynman forces less than 3 meV/Å, respectively, were achieved. The $4f$-states for the rare-earth elements are considered as core states in our calculations. The converged crystal structures were visualized in `VESTA` [79] and the space groups were determined using `FINDSYM` [157]. We also use `VESTA` to generate the polyhedral descriptors as input for training ML models, which is common in the materials science literature[204, 205, 206].

To model disordered substitutional solid solution, the SQS structrues were constructed using the `ATAT` code [207, 208]. We considered a total of six elements (four RE, Si and O) to formulate each solid solution based on the Monte Carlo simulated annealing method [209]. The SQS structures were made with the clusters consisting of pair and triple points with various intersite distances ranging from 4 to 10 Å for every 1 Å increase. The optimal SQS structures were identified by finding the structures with a minimum value of objective function, which is a measure to evaluate similarity to the correlation functions of a disordered state. The geometry optimization of the resulting SQSs was conducted using the `Quantum ESPRESSO`. We constructed the SQS structures for β-$C2/m$ (110 atoms) and optimized them (*i.e.,* relax both internal coordinates and cell parameters) using DFT. These calculations were done by my colleague Dr. Kyungtae Lee (former postdoc fellow from UVa-MSE materials informatics group)

The ML ABCTE predictions with standard error for $(Y_{0.25}Yb_{0.25}Lu_{0.25}La_{0.25})_2Si_2O_7$ - β-$C2/m$ and $(Y_{0.25}Yb_{0.25}Er_{0.25}Dy_{0.25})_2Si_2O_7$ - β-$C2/m$ (in units of 10^{-6} K^{-1}) are 5.57±1.15 and 4.37± 0.58. When accounting for prediction intervals, the estimates are 5.57±2.19 and 4.37± 1.99. The ABCTE of $(Y_{0.25}Yb_{0.25}Lu_{0.25}La_{0.25})_2Si_2O_7$ in the two-phase structure (β-$C2/m$+G-$P2_1/c$) was experimentally measured to be 5.6 $\times10^{-6}$ K^{-1} using dilatometry in the temperature range of 300-950 °C by our colleagues Dr. Deijkers and Dr. Wadley from UVa-MSE. Also the ABCTE of β-$C2/m$ $(Y_{0.25}Yb_{0.25}Er_{0.25}Dy_{0.25})_2Si_2O_7$ was experimentally measured to be 3.28 by Salanova *et al.* [210] (UVa-MSE PhD student in Dr. Ihlefeld group). The experimental values which fall within the error bar of the ML prediction validate the model and demonstrate the capability of the model to extend the predictions to multi-component compounds. To comprehend the significance of this achievement, it is essential to consider the broader context. There are at least 15 common RE elements with a nominal 3+ charge state in the periodic table (Sc, Y and 13 lanthanides, excluding Pm) that can occupy the RE-site in $RE_2Si_2O_7$. If we consider forming four-, five-, and six-component $RE_2Si_2O_7$ solid solutions with equiatomic RE concentrations, then we

Table 6.2: The compounds used for testing and experimentally validating the eSVR models are given. The ML prediction for ABCTE $\times 10^{-6}$ K^{-1}, along with the bootstrap standard error (σ) and the prediction interval (PI) from the doubt algorithm [113], are given. The temperature ranges used for the CTE determination in both experiments and DFT-QHA calculations are also given.

Compound	Experimental	Temperature	ML Predicted
	ABCTE	range (K)	ABCTE ± [σ],[PI]
Test data			
γ-$P2_1/c$ $Ho_2Si_2O_7$	4.2	303-1748	4.2 ± 0.09,2.01
β-$C2/m$ $Sc_2Si_2O_7$	5.4	303-1873	5.4 ± 0.36,2.11
X2-$C2/c$ $Er_2Si_2O_7$	7.6	303-1373	7.43 ± 0.17,2.10
$P3$ $La_{9.33}Si_6O_{26}$	9.4	303-1373	11.78 ± 2.29,2.42
Test data (DFT-QHA from literature) [77]			
β-$C2/m$ $Ho_2Si_2O_7$	4.09	300-1700	4.09 ± 0.08,2.02
β-$C2/m$ $Tm_2Si_2O_7$	3.92	300-1700	4.2 ± 0.03,2.01
γ-$P2_1/c$ $Er_2Si_2O_7$	4.03	300-1700	4.13 ± 0.05,2.08
δ-$Pnma$ $Tb_2Si_2O_7$	8.27	300-1700	7.4 ± 0.13,2.10
δ-$Pnma$ $Ho_2Si_2O_7$	8.57	300-1700	7.5 ± 0.13,2.12
Literature validation for multi-component $RE_2Si_2O_7$			
β-$C2/m$ (Dy, Er, Y, Yb)$_2$Si$_2$O$_7$ [210]	3.28	573-1248	4.37 ± 1.27,2.56
New prediction and Experimental validation (This work)			
β-$C2/m$ (Y, Yb, Lu, La)$_2$Si$_2$O$_7$	5.6	573-1248	5.57 ± 1.15,2.32

have a total of 9,373 (1365 + 3003 + 5005) unique compositions. This number pertains solely to $RE_2Si_2O_7$ solid solutions and does not take into account other material classes. Navigating such an extensive chemical space using traditional trial-and-error methods is impractical. The ability of the holistic ML model to navigate through this vast search space is a critical outcome. This capability can efficiently guide further experimental exploration of new multi-component RE-Si-O solid solutions, thereby accelerating the design of EBC materials with targeted volumetric CTE.

Post hoc Model Interpretability

ML is a powerful tool for establishing quantitative relationships between composition and properties. However, it is often viewed as a "black box" method due to the difficulty in examining and explaining the model behavior. Without comprehending what a trained model has "learned", it becomes a challenge to determine if the model is accurately reflecting the system's physics. To gain a deeper understanding of the model's behavior, I employed a post hoc interpretability approach (elaborated in subsection 3.2.5) that leverages both variable attribution and what-if plots. While variable attribution offers insight into the model's behavior for each observation in the dataset by illustrating the contribution of each descriptor to the predicted value, what-if plots reveal the insights about the functional relationship between a descriptor and the response variable.

I used a well-known global what-if plot method referred to as partial dependence plot (PDPs) to reveal the pattern that the black-box eSVR models have learned [211]. PDPs capture insights about the relationship between a descriptor and the response by showing how the model prediction would be affected if we changed a value of one variable while keeping all other variables unchanged [112]. Since I trained an ensemble of B SVR models, I evaluated the partial dependence estimates for each of the trained SVR models in the ensemble. The mean and standard deviation of the estimates are considered as the partial dependence estimates of the ensemble model and the associated uncertainty, respectively. I employed the `pdp` package [211] as implemented in `R`-language for visualizing the final PDP plots.

While PDPs are useful for offering a broad qualitative perspective on how a feature impacts a model's outcomes, they shouldn't be leaned on for exact quantitative analysis. This is due to their inherent linearity assumption and the tendency to average out descriptor interactions. The PDPs (Figure 6.9) illustrating the functional relationships between the predicted CTE and the descriptor show two different trends. One in which the predicted CTE decreases with the increase in descriptor value. Si_poly_volume is the only descriptor following this inverse relationship. The rest of the descriptors follow the trend in which the predicted

CTE increases with the increase in the descriptor value with RE_avg_bond_length, RE_avg_bond_length_sd, Si_poly_volume, Si_bondangle_var being more sensitive than form_Energy and AK_avg_bond_length. The sensitivity of the PDP curves can be interpreted as the importance or the contribution of the descriptor towards the prediction. The global feature importance analysis also seems to indicate similar importance ranking for the descriptors. Figure 6.8 shows that form_Energy and AK_avg_bond_length are descriptors belonging to the least important descriptors while RE_avg_bond_length, RE_avg_bond_length_sd, Si_bondangle_var and Si_poly_volume are the top four important descriptors.

The PDP shows that the CTE increases with an increase in RE_avg_bond_length. I interpret it using the general idea that longer RE-O bonds are generally weaker than shorter ones. This is because the outer shell electrons, which are involved in bonding, are farther from the nucleus in longer bonds, leading to weaker electrostatic attraction between the RE ions and the surrounding oxygen atoms. Weaker bonds are easier to break and reform, which would make the material more susceptible to thermal expansion. This is also why we can say that the RE metal-oxide bond strength decreases with increasing ionic radius for a given charged ion and coordination number, leading to the thermal expansion of disilicates increasing with the RE cation radius. In simpler terms, longer RE-O bonds make it easier for the material to expand when heated, resulting in a higher CTE.

To corroborate this functional relationship, I plot RE_avg_bond_length against the RE-O interatomic force constants (IFC) as shown in Figure 6.10. The IFC is obtained from a recent study on $RE_2Si_2O_7$ using DFT-DFT calculations [77]. The plot shows that RE_avg_bond_length inversely correlates with the interatomic force constant of RE-O polyhedra serves as a strong confirmation of the trend that PDP shows. The interatomic force constant is a measure of bond strength or stiffness. A lower value for the interatomic force constant implies weaker or less stiff bonds. Since RE_avg_bond_length and the interatomic force constant are inversely related, this tells us that longer RE-O bonds are indeed weaker, corroborating the explanation for the first trend. This inverse correlation essentially reinforces the idea that longer RE_avg_bond_length contributes to a higher CTE. The weaker or less stiff RE-O bonds don't hold the lattice as tightly together, providing the material with greater flexibility to expand or contract, which manifests as a higher CTE. This shows that our model mimic one of the physical descriptors that dictate the CTE.

The increase in CTE with an increase in Si_bondangle_var (variance in the Si-O-Si bond angles) can imply that a greater variance in these bond angles leads to a more flexible molecular structure. This flexibility allows the crystal lattice to accommodate more thermal vibrations, leading to a higher CTE. While factors like bond angle variance, bond length and bond length variability contribute to greater thermal expansion, an increase in Si_poly_volume appears to have the opposite effect, making the structure more rigid and less prone to thermal expansion. A larger Si_poly_volume could mean that the polyhedra are filling up the available space within the crystal lattice more efficiently, leaving less "free volume" for the material to expand into when heated. This effective space-filling would make the structure more rigid, reducing its ability to accommodate thermal vibrations and thereby leading to a lower CTE. This highlights the complex interplay of factors that determine a material's thermal properties and shows that the relationship between microscopic structural parameters and macroscopic properties like CTE can be quite nuanced.

Analytical Equation for CTE

While post-hoc model explanations of ML models are insightful, generating an explicit equation goes a step further: it crystallizes these insights into a form that is both interpretable and generalizable. An equation reveals the mathematical relationships between variables, allowing for more direct scientific interpretation and offering a blueprint for the targeted experimental design. This melding of predictive accuracy with interpretability creates a powerful tool for material science, enabling us to more efficiently explore and understand complex systems.

To this end, I use a new method that is built upon the foundations of the XAI algorithms discussed in this thesis. The algorithm uses grammatical evolution [191]. It begins with a population of randomly generated mathematical expressions and iteratively refines them through processes analogous to natural selection, crossover, and mutation in biological systems. The predefined set of grammar rules ensures that the resulting equations are syntactically correct and logically coherent. Traditional grammatical evolution often produces equations that are mathematically correct but difficult to interpret or lacking physical meaning. To address this, we added a layer of constraint based on PDP and the functional relationships they reveal between polyhedral descriptors and CTE. This novel algorithm was developed by my colleague Mr. Shunshun Liu.

I used the same training data used for the eSVR model and trained the grammatical evolution model using the ***gramEvol*** [212]. Additionally, we informed the grammar using our PDP functional relationships of all 7 variables form_Energy, Si_poly_volume, Si_bondangle_var, RE_avg_bond_length, RE_avg_bond_length_sd, Si_avg_bond_length_sd, and AK_avg_bond_length. Out of these variables, the model omitted form_Energy and

AK_avg_bond_length, which is consistent with their global feature importance values (see Figure 6.8). The equation evolved by the model using the remaining 5 variables is:

$$
\begin{aligned}
\text{CTE} &= 8.905 \times \text{RE\_avg\_bond\_length} - 7.214 \times \text{Term1} - 12.644 \qquad (6.1)\\
\text{Term1} &= \frac{\text{Si\_poly\_volume}}{\text{Si\_bondangle\_var} \times \text{Term2}} \qquad (6.2)\\
\text{Term2} &= 1 - \text{Si\_poly\_volume} \times \text{Si\_bondangle\_var}\\
&\quad \times (\text{RE\_avg\_bond\_length\_sd} + \text{AK\_avg\_bond\_length}) \qquad (6.3)
\end{aligned}
$$

In the evolved equation, RE_avg_bond_length has a positive correlation with CTE, indicating that an increase in this parameter generally results in a higher CTE. Term1, which combines Si_poly_volume and Si_bondangle_var, RE_avg_bond_length_sd and AK_avg_bond_length also modulate CTE, but their effects are more complex and are embedded within Term1.

The ABCTE predictions using the evolved equation for $(Y_{0.25}Yb_{0.25}Lu_{0.25}La_{0.25})_2Si_2O_7$ - β-$C2/m$ and $(Y_{0.25}Yb_{0.25}Er_{0.25}Dy_{0.25})_2Si_2O_7$ - β-$C2/m$ (in units of 10^{-6} K^{-1}) are 5.67 and 4.8 respectively and the experimental ABCTE of the two compounds are 5.6 and 3.28 [210] (in units of 10^{-6} K^{-1}) respectively. The close alignment between our model's predictions and experimental data serves as a strong validation of the evolved equation, marking a key milestone of this thesis. The unique strength of this approach is its generalizability and accuracy. Unlike traditional models that use semi-empirical equations based on local bond geometric features, our algorithm can be applied to a broader array of materials, including complex hign entropy compounds. This opens up new avenues for accelerating the design of a diverse range of compounds in the

RE-Si-O crystal chemistry.

6.4 Summary

In this chapter, I addressed a gap in the current literature by developing a holistic ML model that describes the structure-CTE relationship across a range of material classes within the RE-Si-O crystal chemistry. This holistic ML model offers a unified framework that is applicable to both single and multi-component compounds. This adaptability is achieved through an innovative approach to descriptor generation, which leverages polyhedral features. An eSVR model using these descriptors was built to rapidly predict the ABCTE of all the compounds in RE-Si-O crystal chemistry. One of the key results is the ability of the model trained solely on single-component RE silicates to predict the ABCTE of 2 different four-component β-$C2/m$ (Y, Yb, Lu, La)$_2$Si$_2$O$_7$ and β-$C2/m$ (Y, Yb, Er, Dy)$_2$Si$_2$O$_7$ with prediction uncertainties quantified. The agreement between ML prediction and experimental validation adds confidence to the approach. Post hoc model interpretation of the validated eSVR model revealed the functional relationship between the polyhedral descriptors and the ABCTE. The PDP trends show that variables like Si_bondangle_var, RE_avg_bond_length_sd, and Si_avg_bond_length positively affect CTE, while Si_poly_volume negatively impacts it. Among these, RE_avg_bond_length stands out for its validated significance through its inverse correlation with the interatomic force constant, emphasizing its crucial role in determining thermal expansion behavior. While post hoc model interpretation offer valuable insights, generating an explicit equation like the one evolved through grammatical evolution provides an even deeper understanding. This equation not only encapsulates complex relationships between variables but also offers a highly reliable and generalizable framework for understanding and predicting CTE in RE-Si-O crystal chemistry. Importantly, the close match between the model's outputs and experimental results serve as a key validation of the developed ideas.

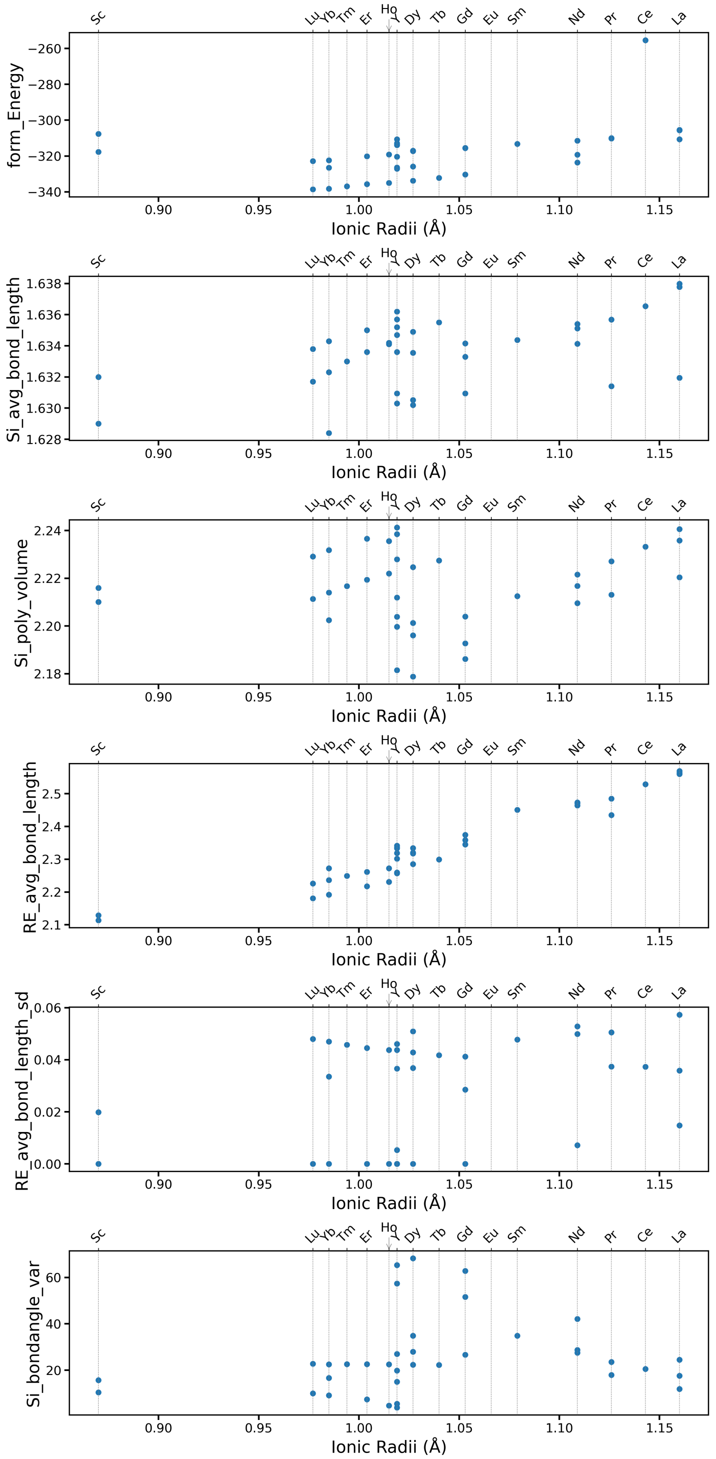


Figure 6.6: Graphical representation of polyhedral metrics (Y-axes) in relation to Rare Earth (*RE*) Ionic Radii in Angstroms (Å) (X-axes), focusing on a charge state of 3+ and a coordination number of 8. The plots are representative of the training data including all three material classes in RE-Si-O space. The variable AK_avg_bond_length is ignored as we only have 2 alkali earth/alkaline earth metal bearing apatites in the dataset. The secondary x-axis bears the annotations of RE cation in the compound.

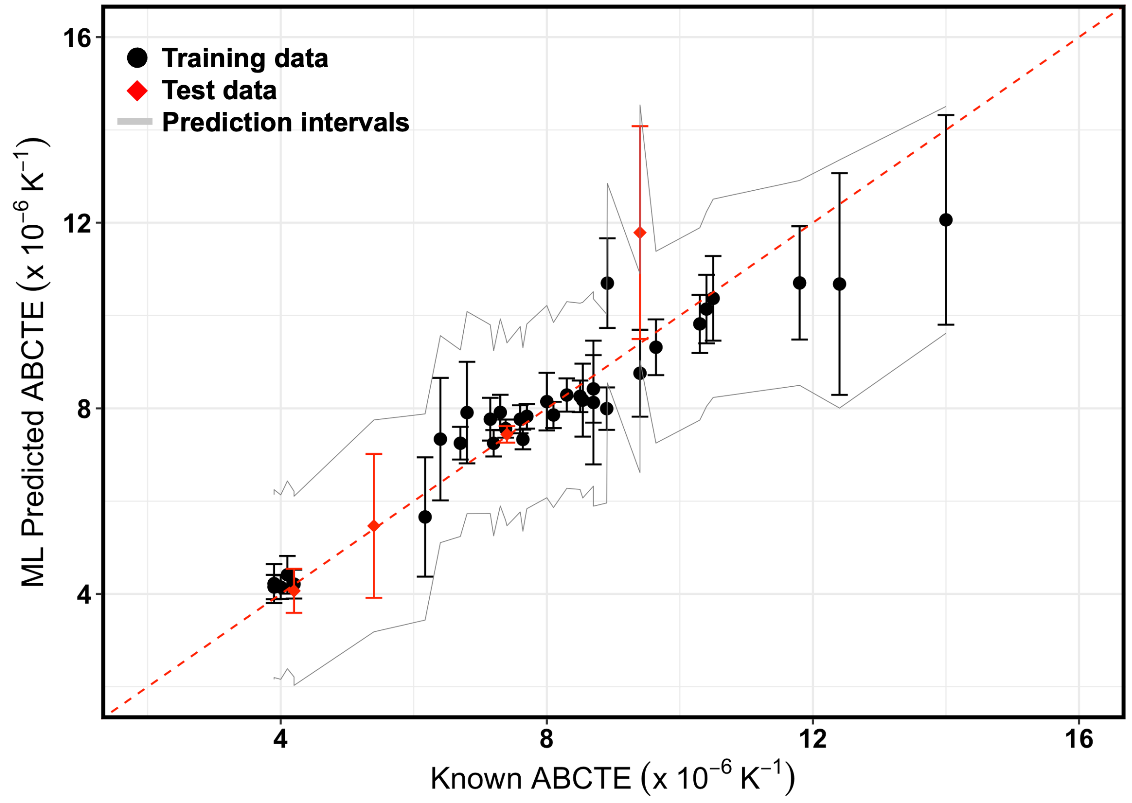


Figure 6.7: Performance of the eSVR model trained on the dataset including all three material classes in RE-Si-O space. The X- and Y-axes are the known and ML predicted ABCTE data, respectively. The error bar represent the standard error from the ensemble of eSVR model. The grey lines represent the prediction interval limits constructed using the doubt algorithm [113]. The red dashed line represents the X=Y line and the data points falling on this line indicate perfect agreement between the ML models and the known data. The black dots are the data points used to train the models. The red diamonds represent the test data that were not used to train the ML models.

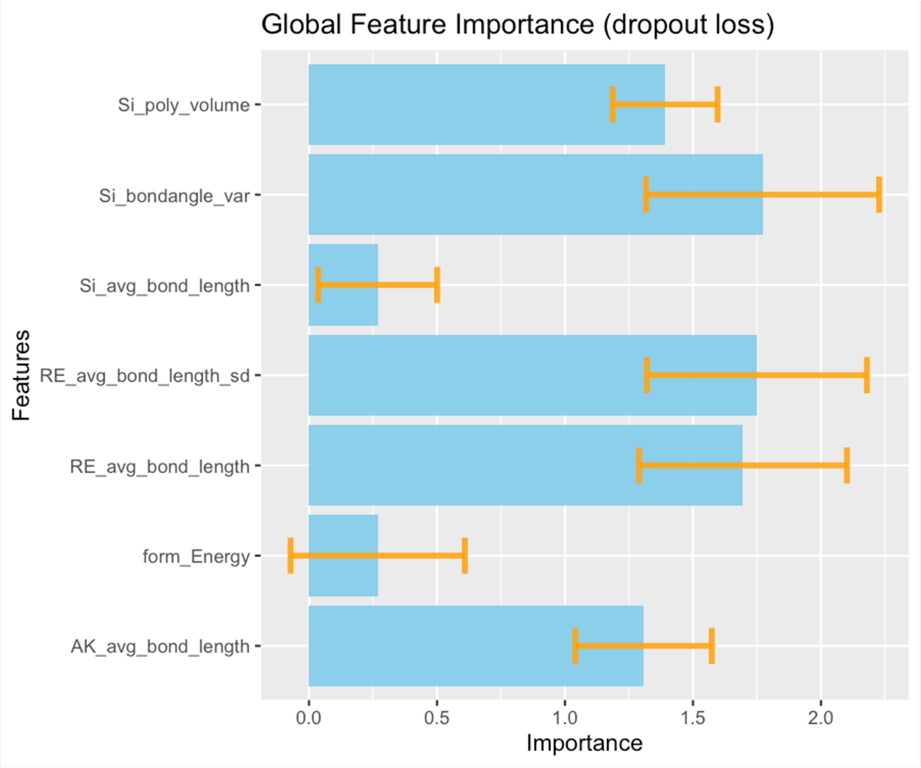


Figure 6.8: Results of global feature importance analysis for the eSVR model trained on the dataset including all three material classes in RE-Si-O space. Blue bar shows average impact, across all the models in the ensemble, on the dropout loss. Orange error bar depicts the standard deviation of the dropout loss across all the models.

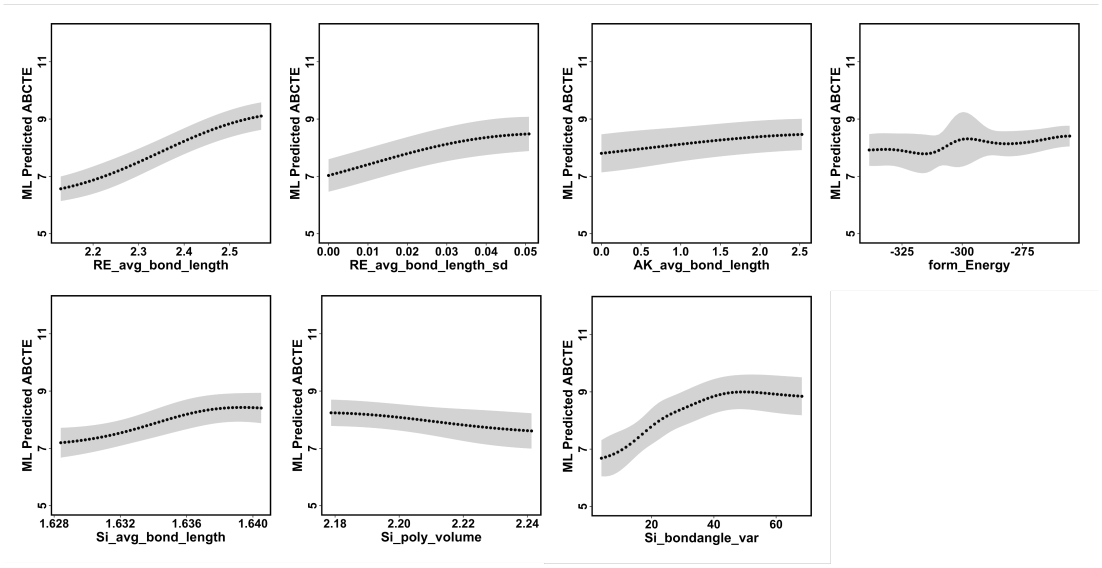


Figure 6.9: PDPs showing the relationship between the descriptors (x-axis) and CTE (y-axis). The shaded regions indicate the uncertainties based on standard deviation. The plots of representative of the training dataset including all three material classes in RE-Si-O space.

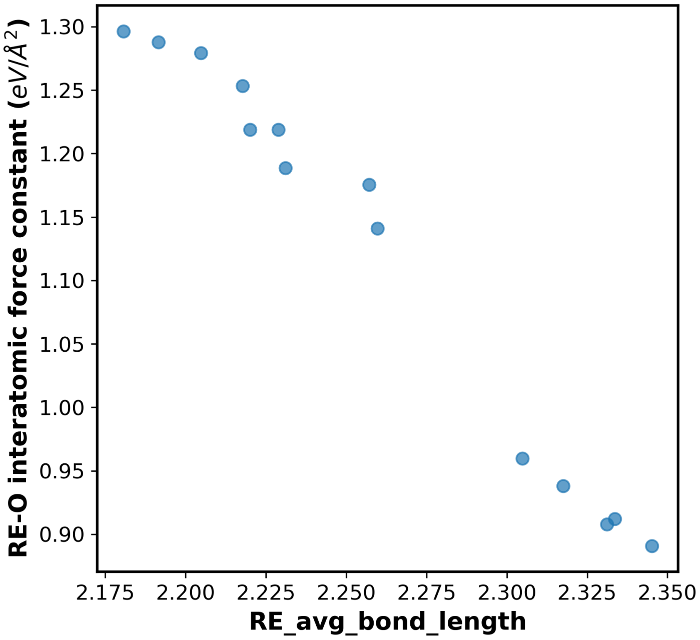


Figure 6.10: Plots showing inverse correlation between RE_avg_bond_length and interatomic force constants of RE-O polyhedron [77] obtained by DFT-QHA calculations for a select RE disilicate compounds.

Chapter 7

Conclusion

This dissertation has addressed the three goals initially stated:

1. Accelerate the design of novel compounds in the complex RE-Si-O chemical space with targeted CTE. This required me (in collaboration with several excellent researchers) to establish a hitherto unknown quantitative structure-property relationships in the RE-Si-O family of compounds using computational and ML approaches.

2. Ensure trust in the informatics approach by precisely knowing when/where a ML model can succeed and/or fail. I developed a prediction interval approach that provided an estimate of model confidence to make informed design decisions.

3. Understand why a black-box ML model makes a certain prediction. Structure-property relationships in materials can be better understood if the predictive understanding is backed up by science hidden in the model. The XAI approach was instrumental in helping me accomplish this objective.

Having accomplished these objectives, I want to reflect on the insights gained in this chapter. Although these goals were first stated separately, as I continued to make progress in my research I learned that they are closely linked. In fact, achieving each goal was part of a larger, unified effort to improve our knowledge of RE-Si-O crystal chemistry. The RE-Si-O chemical space, encompassing RE disilicates, RE monosilicates, and RE apatites, is both intricate and expansive. While traditional approaches have relied on ionic radii as the primary descriptor for structure-property relationships and energetics of RE silicates, this dissertation demonstrated the importance of employing DFT calculations and ML methods synergistically. These calculations described the ΔE trends across the RE-Si-O crystal chemistry, identifying energetically favorable polymorphs and providing optimized crystal structure data. This enabled the establishment of two novel sets of descriptors: unit cell parameters and polyhedral features. These descriptors offered deeper insights into the complex relationships between structure and CTE, across all three material classes.

The DFT-ML approach has been particularly impactful for RE disilicates, a crucial material class for EBC applications. DFT calculations yielded ΔE data that offered key insights into the energetics favoring polymorph formation. These findings were further validated by existing literature. The ML approach provided a fast-track method for predicting the CTE as a function of unit cell parameters. The uncertainties were quantified using bootstrap standard errors of the predictions. Importantly, experimental validations on the structure and CTE of $Sm_2Si_2O_7$ served to confirm the accuracy of our predictions, effectively closing the design loop.

Extending our approach of DFT and ML to RE monosilicates, similar results were achieved. Intrigued by a notable pattern in CTE anisotropy discovered in the recent research conducted by Ridley *et al.* I shifted focus to CTE anisotropy. I chose DFT to gain insights into the underlying mechanisms driving this phenomenon. The density of states calculations revealed a previously unidentified correlation between the d-orbital bandwidth and unit cell volume in the $C2/c$ structure for Sc_2SiO_5, Y_2SiO_5, and La_2SiO_5. To reduce the Y-4d bandwidth and, we conjecture, consequently lessen CTE anisotropy, Y_2SiO_5 should be synthesized in an open structure with reduced Y-O effective coordination numbers in both polyhedral units. Achieving this would likely necessitate novel, non-equilibrium materials synthesis and processing techniques.

Shifting the lens from unit cell parameters to polyhedral descriptors offers distinct advantages in describing structure-CTE relationships. First, polyhedral descriptors have the potential to carry mechanistic information that can be linked with phonon dispersion curves and phonon density of states. Second, they provide a unique set of descriptors applicable across the diverse RE-Si-O chemical space, including disilicates, monosilicates, and apatites. This versatility sets the stage for the development of a holistic machine learning model capable of mapping structure to CTE across the entire RE-Si-O domain. In this context, I introduced a novel representation scheme based on polyhedral descriptors and formation energy, aimed at uniquely fingerprinting both single and multi-component compounds, thereby enabling the ML model to establish a quantitative relationship between structure and CTE.

To compile the dataset for the holistic CTE model, I also explored the RE silicate apatites, an area where data has been lacking up to this point. The DFT calculations optimized non-stoichiometric ($RE_{9.33}Si_6O_{26}$) in P3 structure. In the case of $RE_9A_1Si_6O_{26}$ and $RE_8A_2Si_6O_{26}$, ΔE calculations revealed the interesting energetics trends. The ΔE calculations showed that in the $RE_8AE_2Si_6O_{26}$ series, AE ions, like Ba and Sr, majorly favor the $P\bar{3}$ structure, with $P6_3$ closely trailing. However, for Be and Mg, the $P6_3$ structure takes precedence. AE=Ca presents a mixed trend without a clear frontrunner. Contrarily, in the $RE_9A_1Si_6O_{26}$ series featuring alkali metals as A, a consistent preference emerges for the Pm structure. Notable exceptions include A=Rb and Na, aligning with $P\bar{6}$. This highlights AE's pronounced influence over structure compared to A, with specific elemental combinations adding nuance to the observed patterns.

I then built the holistic ML model for the dataset inluding all three classes of RE-Si-O space. In this iteration, I went beyond merely calculating the standard error and establish robust prediction intervals that do not rely on the assumption of normal distribution, taking into account both the uncertainties in the model and observational noise. One of the key outcomes of this model is the ability of the model trained solely on single-component RE silicates to predict the ABCTE of multi-component counterparts. I made predictions for β-$C2/m$ (Y, Yb, Lu, La)$_2$Si$_2$O$_7$ and β-$C2/m$ (Y, Yb, Er, Dy)$_2$Si$_2$O$_7$ with prediction uncertainties quantified. The agreement between ML prediction and experimental validation adds confidence to the approach.

Post hoc model interpretation of the validated eSVR model revealed the functional relationship between the polyhedral descriptors and the ABCTE. The PDP trends show that variables like Si_bondangle_var, RE_avg_bond_length_sd, and Si_avg_bond_length positively affect CTE, while Si_poly_volume negatively impacts it. These trends suggest a complex interplay of structural features affecting thermal behavior. Specifically, a more flexible structure is indicated by higher bond angle variance (Si_bondangle_var) and greater bond length variability (RE_avg_bond_length_sd and Si_avg_bond_length). Longer average bond lengths (RE_avg_bond_length) further point to weaker bonds, all contributing to a higher propensity for the material to expand when heated. Conversely, larger silicon-oxygen polyhedral volumes (Si_poly_volume) act as a counterbalance, making the structure more rigid and less prone to thermal expansion, thereby lowering the CTE. The inverse correlation between RE_avg_bond_length and the interatomic force constant for RE-O polyhedra confirms that longer RE-O bonds are weaker, reinforcing the trend that such bonds contribute to a higher CTE. These insights provide a nuanced understanding of how a microscopic parameter like RE_avg_bond_length can impact a macroscopic property like thermal expansion.

While post-hoc model explanations of ML models are insightful, generating an explicit equation goes a step further: it crystallizes these insights into a form that is both interpretable and generalizable. An equation reveals the mathematical relationships between variables, allowing for more direct scientific interpretation and offering a blueprint for targeted experimental design. To this end, in collaboration with my research group, we built a grammatical evolution model that is informed by the functional relationships of the variables with CTE that is captured by PDP. The equation captures relationships between polyhedral descriptors and the CTE in RE-Si-O crystal chemistry. Notably, the close agreement between the model's predictions and the experimental data provides a validation of our approach. This work represents a significant step forward in the field, offering a reliable, interpretable, and generalizable model that has the potential to significantly accelerate the design and discovery of new materials.

Chapter 8

Future Work

8.1 Equation-Informed Bayesian Models

This thesis yielded insights into polyhedral descriptors' intricate relationship with CTE in RE-Si-O crystal chemistry. It introduced an innovative approach, culminating in an explicit equation for CTE in RE-Si-O that enhances interpretability. One intriguing avenue for future work is leveraging the equation developed through grammatical evolution to inform Bayesian models. The equation, grounded in the functional relationships of polyhedral descriptors with CTE, could serve as a prior distribution for Bayesian models. This integration would merge the power of ML and the interpretability of mathematical equations, offering an innovative framework for materials design. Bayesian models inherently models uncertainties due to their probabilistic nature. The integration of the equation-derived prior distribution further augments this aspect, enhancing the model's ability to provide reliable prediction intervals.

8.2 Calibration for prediction interval algorithm

In the area of predictive modeling, I developed an algorithm for constructing prediction intervals that diverges from traditional convolution-based methods, such as those proposed by Mougan and Nielsen [113] and Srivatsava and Kumar [111]. The developed approach incorporates an additional sampling loop adjusted by Rademacher variables, aiming to address issues related to heavy-tailed or asymmetric error distributions. Despite these innovations, the current performance of our algorithm is subpar. A natural extension for improving our algorithm lies in calibration techniques aimed at achieving more reliable coverage probabilities. By post hoc adjusting the prediction intervals based on observed discrepancies, the algorithm could potentially yield more accurate intervals, especially when dealing with complex or limited data sets.

The following steps could be undertaken to achieve this:

1. After the prediction intervals are initially constructed, evaluate their actual coverage probabilities on validation datasets,
2. Identify the degree to which the observed coverage deviates from the expected confidence level,
3. Develop and implement an algorithm to adjust the initially constructed intervals based on the observed discrepancies. This could be inspired by existing work on calibrated forecasting, and
4. Re-evaluate the adjusted intervals on multiple datasets to ensure that the calibration process has effectively improved the coverage probabilities.

8.3 Influence of defects on electronic structure trends

Continuing of the effort on anisotropy in CTE, the exploration of point defects, specifically Y- and O-vacancies, emerges as a critical next step. Given our newfound understanding of the *d*-orbital bandwidth and its relation to unit cell volume, it will be interesting to probe how point defects can offer more nuanced control over these variables. Investigating the interplay between unit cell volume and these defects could potentially allow us to further modulate the Y-4*d* bandwidth and, by extension, the CTE anisotropy. This stands as a promising avenue for future research. Since experimental validation will be tricky, it will be worthwhile to check if the CTE anisotropy can be predicted using the quasi-harmonic approximations. This will validate the hypothesis and may motivate new experiments.

Appendix A

Trustworthy ML

A.1 Trustworthy ML algorithms

Distribution-free bootstrap based prediction interval algorithm. Glossary: α-confidence level, x_0-new observation, m_i-model error, o_i-observation error, t_i-training error

Algorithm A.1.

```
procedure PREDICTINTERVAL($\alpha, x_0$)
    Build $b$ bootstrap samples $B_i$ from $R(r)$
    Initialize bootstrap sample set $D = \phi$
    for each bootstrap sample $B_i$ do
        Build regression models $\bar{y}_r$
        Obtain centered samples $m_i = \hat{\mu}_r(x_i) - \bar{y}_r(x_i)$
        $D \rightarrow D \cap m_i$
    end for
    Initialize training error sample $E_1 = \phi$
    for each training sample $(x_i, y_i)$ do
        Compute error $t_i = y_i(x_i) - \hat{\mu}_r(x_i)$
        $E_1 \rightarrow E \cap t_i$
    end for
    Initialize OOB error sample set $E_2 = \phi$
    for each OOB sample $(x_i, y_i)$ do
```

```
        Compute error $o_i = y_i(x_i) - \hat{\mu}_r(x_i)$
        $E_2 \to E \cap o_i$
    end for
    Build the set $G = (1 - validationweight) \times (t_i) + (validationweight) \times (o_i)$
    Generate the prediction error distribution $\delta^{bm}{}_0$
    For a new datapoint $x_0$ :
    for $b = 1, ..., B$ do
        for $m = 1, ..., B/2$ do
            Generate Rademacher variable $\eta$
            Calculate $\epsilon$ by sampling a residual from $G$ adjusted by $\eta$
            $\delta^{bm}{}_0 = m_i + o_i$
        end for
    end for
    $PI(x_0) = \hat{\mu}_r(x_i) + \delta^{\alpha/2}{}_0, \delta^{\alpha/2}{}_0$
end procedure
```

Explainable ML algorithm for local global and intermediate levels using the SHAP and ICE methods along with k-means clustering

Algorithm A.2.

```
procedure SHAP_analysis(D, eSVM)    ▷ Procedure to construct the SHAP dataframe with the training dataset (D)
    RowLength ← Size(D)    ▷ Total number of instances of D
    for i ← 1 to RowLength do    ▷ Loops through each instance of D
        for j ← 1 to 50 do    ▷ Loops through 50 bootstrap samples
            M ← eSVM[j]
            Exp ← Model_explainer(M, D[j])    ▷ Generates a model explainer for a given bootstrap sample
            SHAP_pred[j] ← Predict_parts(Exp, new_observation=D[i])    ▷ Calculates the variable attributions to the prediction of a given instance
            Merged_SHAP_pred[i] ← Binding(SHAP_pred[j])    ▷ Merges the resulting variable attributions on every loop iteration
        end for
        Avg_Merged_SHAP_pred[i] ← Mean(Merged_SHAP_pred[i])    ▷ Averages the SHAP values of all the bootstrap samples for a given instance
        SHAP_dataframe ← Binding(Avg_Merged_SHAP_pred[i])
    end for
    return SHAP_dataframe
end procedure
procedure $k$-MEANS CLUSTERING(SHAP_dataframe)    ▷ Procedure for $k$-means clustering based on SHAP values
    $k$ ← 10    ▷ $k$: the number of clusters
    Cluster_info ← kmean(SHAP_dataframe, k)    ▷ Implements the $k$-means clustering algorithm
    return Cluster_info    ▷ Classifies each instance with a specific cluster label
end procedure
procedure ICE_ANALYSIS(D, eSVM, Cluster_info)    ▷ Procedure for ICE analysis based on cluster information
    idx ← cluster_label    ▷ Choose a cluster label of interest
    for i ← 1 to length(Cluster_info[idx]) do    ▷ Loops through all the instances with the given cluster label
        for j ← 1 to 50 do    ▷ Loops through 50 bootstrap samples
            M ← eSVM[j]
            Exp ← Model_explainer(M, D[j])
            ICE_pred[j] ← Predict_profile(Exp, new_observation=D[i])   ▷ Calculates individual ICE profiles
            Merged_ICEpred ← Binding(ICE_pred[j])   ▷ Merges the resulting ICE data on every iteration of the inner loop
        end for
        Merged_ICEdata ← Binding(Merged_ICEpred)   ▷ Merges the resulting ICE data on every iteration of the outer loop
    end for
    ICE_dataframe ← Mean(Merged_ICEdata)   ▷ Averages the ICE data across all the instances with the given cluster label
    return ICE_dataframe
end procedure
```

Appendix B

Crystal Chemistry of $RE_2Si_2O_7$

B.1 Total energy difference of $RE_2Si_2O_7$ from DFT

Table B.1: Total energy difference, in from DFT calculations with respect to the lowest energy structure in, where Space groups given in parentheses indicate the final converged structure when the tolerance is set at 0.0001 or lower in FINDSYM [157].

Space Group	**ΔE (meV/atom)**							
	$Dy_2Si_2O_7$	$Eu_2Si_2O_7$	$Ho_2Si_2O_7$	$Lu_2Si_2O_7$	$Tm_2Si_2O_7$	$Sc_2Si_2O_7$	$Er_2Si_2O_7$	$Tb_2Si_2O_7$
$C2/m$ (β)	**0.00**	143.66	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
$Pnma$ (δ)	11.53	146.69($P\bar{1}$)	16.10	34.01	24.92	69.06	20.21	6.79
$P2_1/c$ (η)	33.53	96.23	43.01	80.45	61.40	57.99	51.87	23.71
$P\bar{1}$ (α)	14.26	**0.00**	20.22	41.03	31.12	60.97	51.55 ($P2_1/c$)	7.88
$P2_1/c$ (G)	14.25 ($P\bar{1}$)	92.27 ($P\bar{1}$)	20.22 ($P\bar{1}$)	41.03 ($P\bar{1}$)	31.12 ($P\bar{1}$)	60.96 ($P\bar{1}$)	25.50 ($P\bar{1}$)	7.90 ($P\bar{1}$)
$P2_1/c$ (γ)	2.17	139.92	2.54	3.78	3.19	5.27	2.84	1.75
$P4_1$ (A)	29.58	73.70	38.50	72.51	55.73	122.94	46.54	20.34

Appendix C

RE Silicate Apatites

C.1 Correlation analysis of polyhedral descriptor set

The linear Pearson correlation coefficient (PCC) was then used to perform the pairwise statistical correlation analysis[213]. The inputs for the correlation analysis are the 25 descriptors (24 polyhedral descriptors and formation energy) extracted from the DFT optimized structures. The correlation plot is shown in Figure C.1, which indicates that most of the descriptors show strong statistical correlation and hence carry redundant information. We removed redundancy by only considering pairs whose PCC was less than 0.7. From a total of 25 descriptors, we down-selected only 7 descriptors as inputs for ML model building. They include: Si_avg_bond_length, Si_poly_volume, RE_avg_bond_length, AK_avg_bond_length, RE_avg_bond_length_sd, Si_bondangle_var and form_Energy.

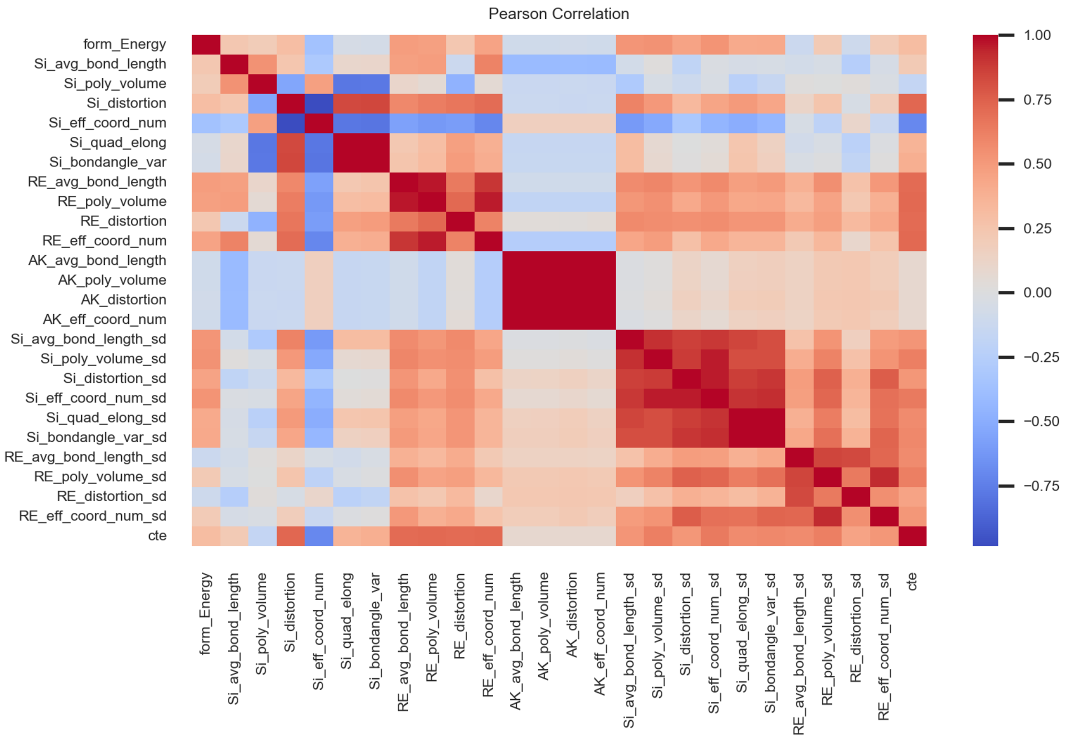


Figure C.1: Pair-wise statistical correlation analysis for the training data used for the holistic CTE model. Dark red and dark blue indicate strong positive and negative correlation, respectively. We remove redundancy by only considering variables whose PCC was less than 0.7 for building ML models.

C.2 DFT total energy data of RE silicate apatites

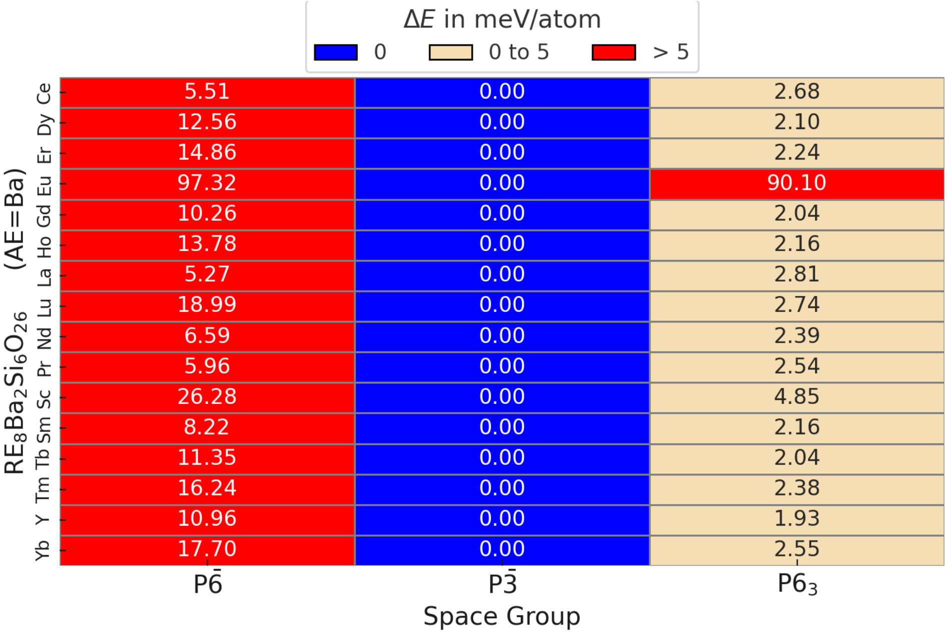


Figure C.2: Total energy difference, in from DFT calculations with respect to the lowest energy structure in RE$_8$AE$_2$Si$_6$O$_{26}$ for AE=Ba,where space groups given in parentheses indicate the final converged structure when the tolerance is set at 0.0001 or lower in FINDSYM. ΔE=0 signifies the ground structure. When ΔE is within a range of approximately 5 meV/atom, it suggests a close energetic competition with the ground state structure.

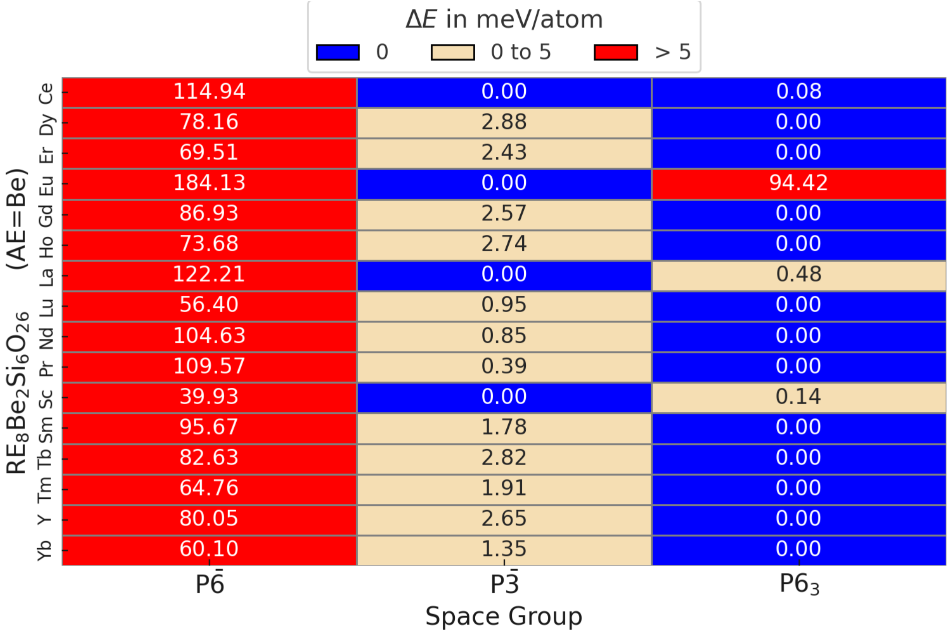


Figure C.3: Total energy difference, in from DFT calculations with respect to the lowest energy structure in $RE_8AE_2Si_6O_{26}$ for AE=Be,where space groups given in parentheses indicate the final converged structure when the tolerance is set at 0.0001 or lower in FINDSYM. ΔE=0 signifies the ground structure. When ΔE is within a range of approximately 5 meV/atom, it suggests a close energetic competition with the ground state structure.

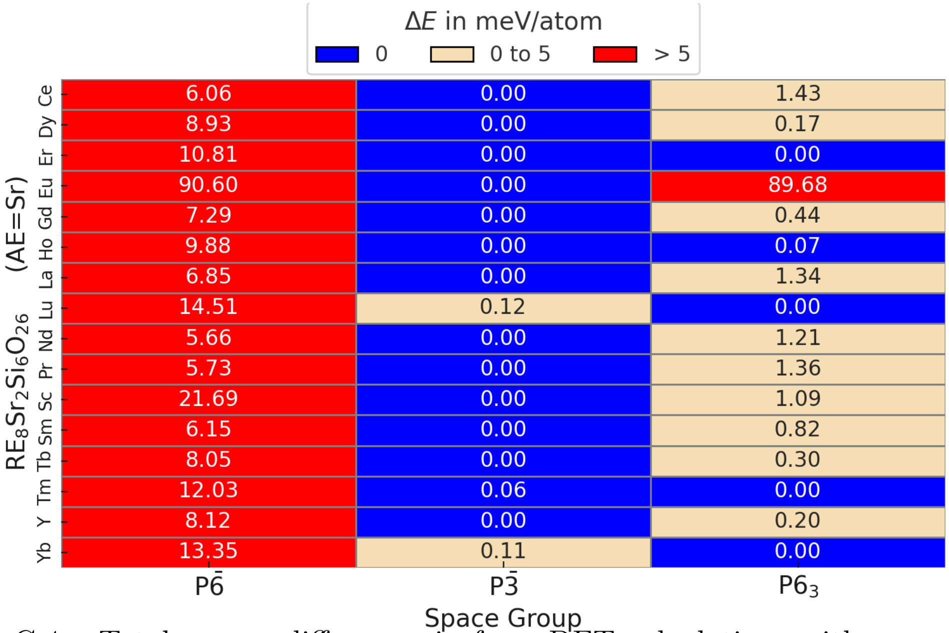


Figure C.4: Total energy difference, in from DFT calculations with respect to the lowest energy structure in RE$_8$AE$_2$Si$_6$O$_{26}$ for AE=Sr,where space groups given in parentheses indicate the final converged structure when the tolerance is set at 0.0001 or lower in FINDSYM. ΔE=0 signifies the ground structure. When ΔE is within a range of approximately 5 meV/atom, it suggests a close energetic competition with the ground state structure.

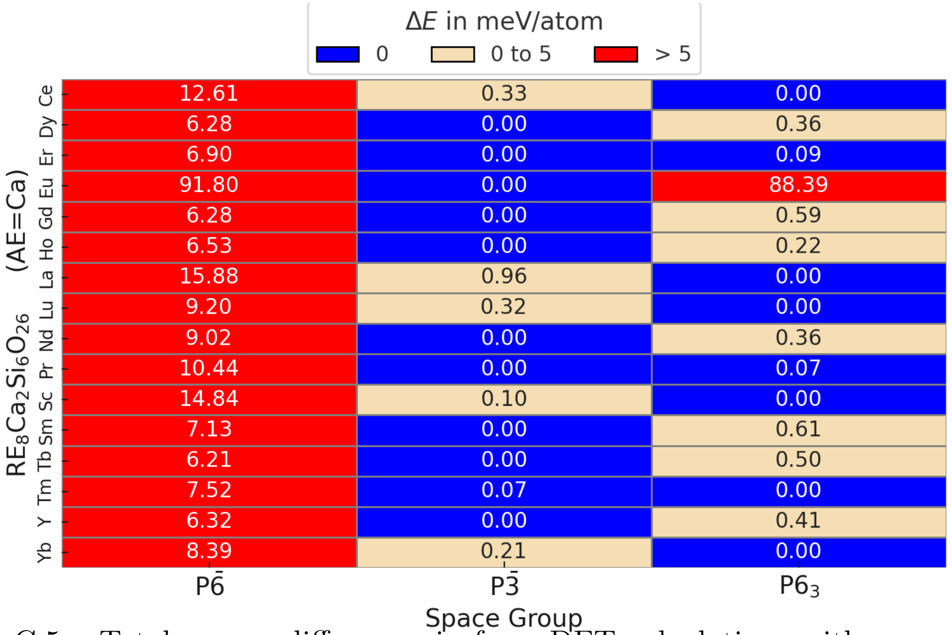
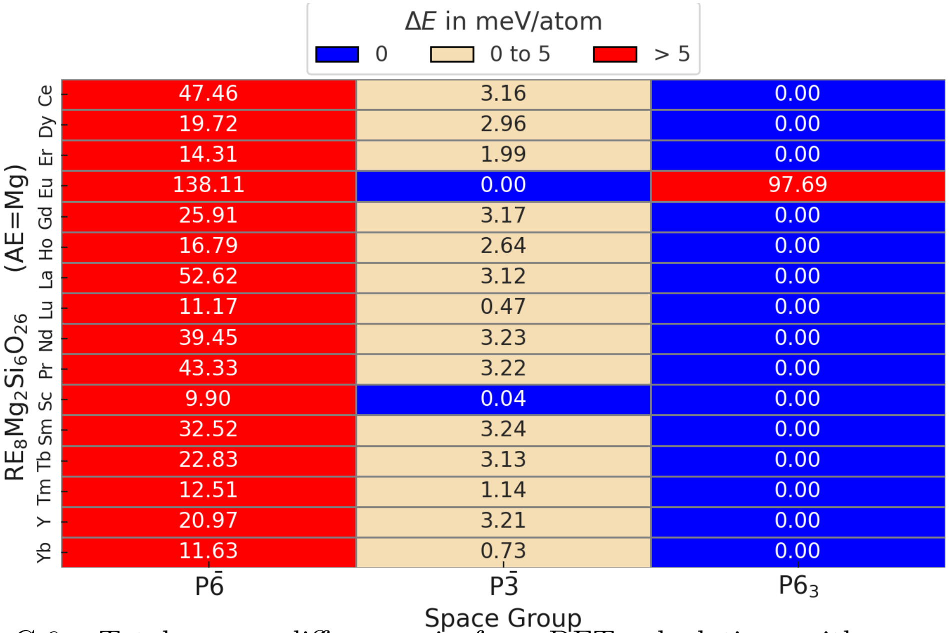


Figure C.5: Total energy difference, in from DFT calculations with respect to the lowest energy structure in $RE_8AE_2Si_6O_{26}$ for AE=Ca,where space groups given in parentheses indicate the final converged structure when the tolerance is set at 0.0001 or lower in FINDSYM. ΔE=0 signifies the ground structure. When ΔE is within a range of approximately 5 meV/atom, it suggests a close energetic competition with the ground state structure.



	$P\bar{6}$	$P\bar{3}$	$P6_3$
Ce	47.46	3.16	0.00
Dy	19.72	2.96	0.00
Er	14.31	1.99	0.00
Eu	138.11	0.00	97.69
Gd	25.91	3.17	0.00
Ho	16.79	2.64	0.00
La	52.62	3.12	0.00
Lu	11.17	0.47	0.00
Nd	39.45	3.23	0.00
Pr	43.33	3.22	0.00
Sc	9.90	0.04	0.00
Sm	32.52	3.24	0.00
Tb	22.83	3.13	0.00
Tm	12.51	1.14	0.00
Y	20.97	3.21	0.00
Yb	11.63	0.73	0.00

Figure C.6: Total energy difference, in from DFT calculations with respect to the lowest energy structure in $RE_8AE_2Si_6O_{26}$ for AE=Mg,where space groups given in parentheses indicate the final converged structure when the tolerance is set at 0.0001 or lower in FINDSYM. ΔE=0 signifies the ground structure. When ΔE is within a range of approximately 5 meV/atom, it suggests a close energetic competition with the ground state structure.

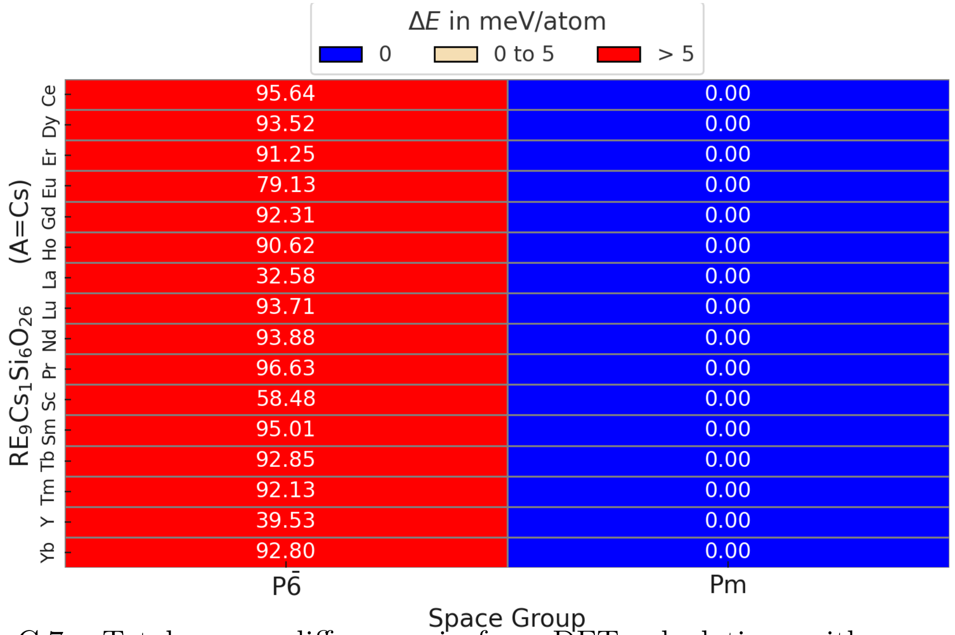
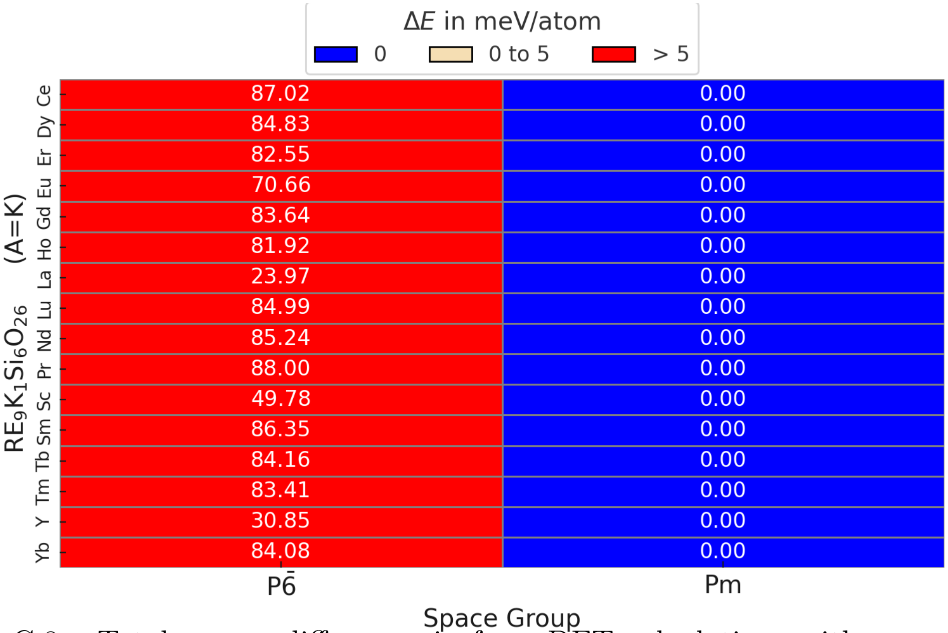


Figure C.7: Total energy difference, in from DFT calculations with respect to the lowest energy structure in RE$_9$A$_1$Si$_6$O$_{26}$ for A=Cs,where space groups given in parentheses indicate the final converged structure when the tolerance is set at 0.0001 or lower in FINDSYM. ΔE=0 signifies the ground structure. When ΔE is within a range of approximately 5 meV/atom, it suggests a close energetic competition with the ground state structure.



	P$\bar{6}$	Pm
Ce	87.02	0.00
Dy	84.83	0.00
Er	82.55	0.00
Eu	70.66	0.00
Gd	83.64	0.00
Ho	81.92	0.00
La	23.97	0.00
Lu	84.99	0.00
Nd	85.24	0.00
Pr	88.00	0.00
Sc	49.78	0.00
Sm	86.35	0.00
Tb	84.16	0.00
Tm	83.41	0.00
Y	30.85	0.00
Yb	84.08	0.00

Figure C.8: Total energy difference, in from DFT calculations with respect to the lowest energy structure in $RE_9A_1Si_6O_{26}$ for A=K,where space groups given in parentheses indicate the final converged structure when the tolerance is set at 0.0001 or lower in FINDSYM. ΔE=0 signifies the ground structure. When ΔE is within a range of approximately 5 meV/atom, it suggests a close energetic competition with the ground state structure.

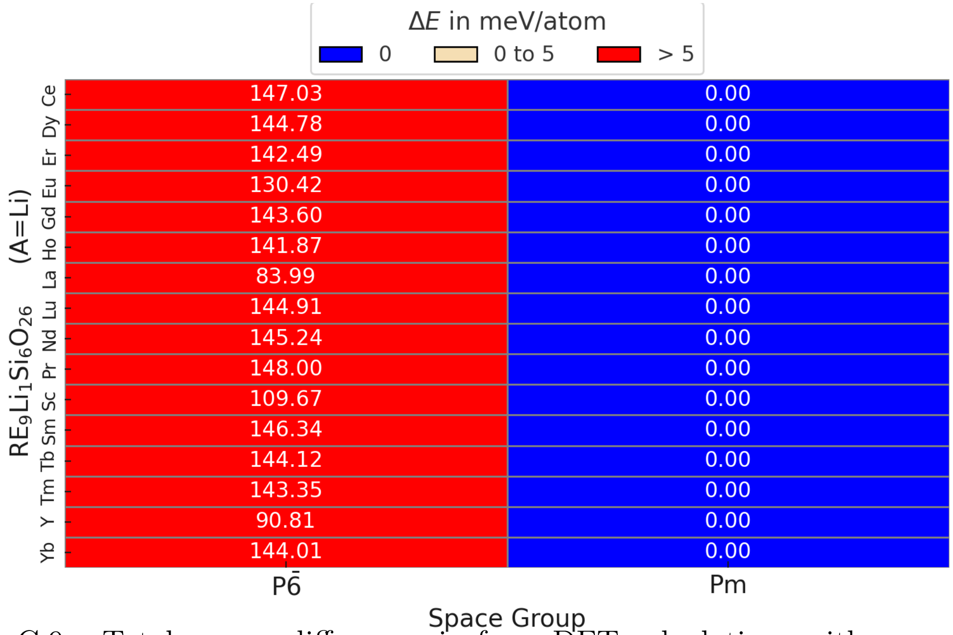


Figure C.9: Total energy difference, in from DFT calculations with respect to the lowest energy structure in $RE_9A_1Si_6O_{26}$ for A=Li,where space groups given in parentheses indicate the final converged structure when the tolerance is set at 0.0001 or lower in FINDSYM. ΔE=0 signifies the ground structure. When ΔE is within a range of approximately 5 meV/atom, it suggests a close energetic competition with the ground state structure.

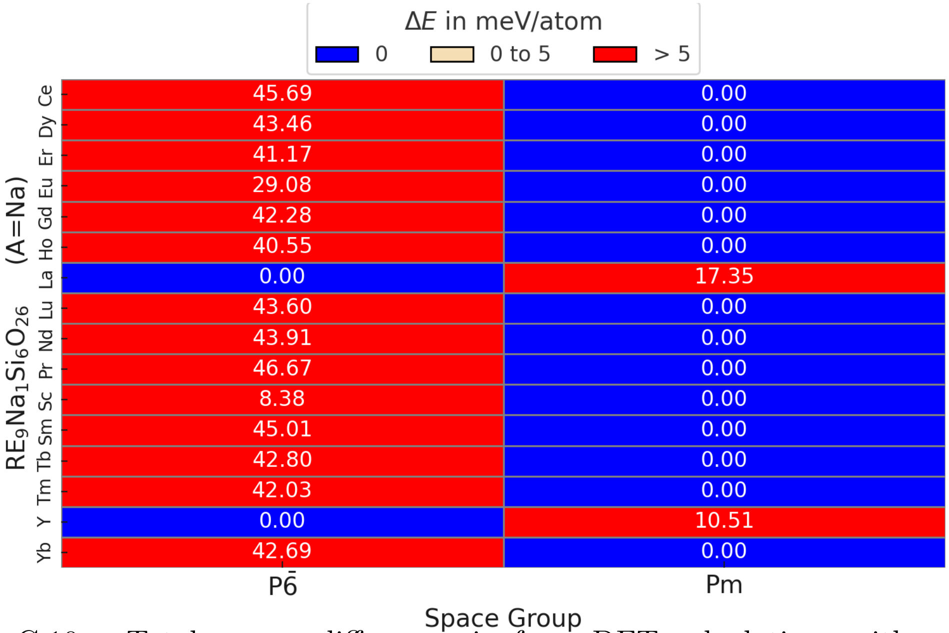


Figure C.10: Total energy difference, in from DFT calculations with respect to the lowest energy structure in $RE_9A_1Si_6O_{26}$ for A=Na,where space groups given in parentheses indicate the final converged structure when the tolerance is set at 0.0001 or lower in FINDSYM. ΔE=0 signifies the ground structure. When ΔE is within a range of approximately 5 meV/atom, it suggests a close energetic competition with the ground state structure.

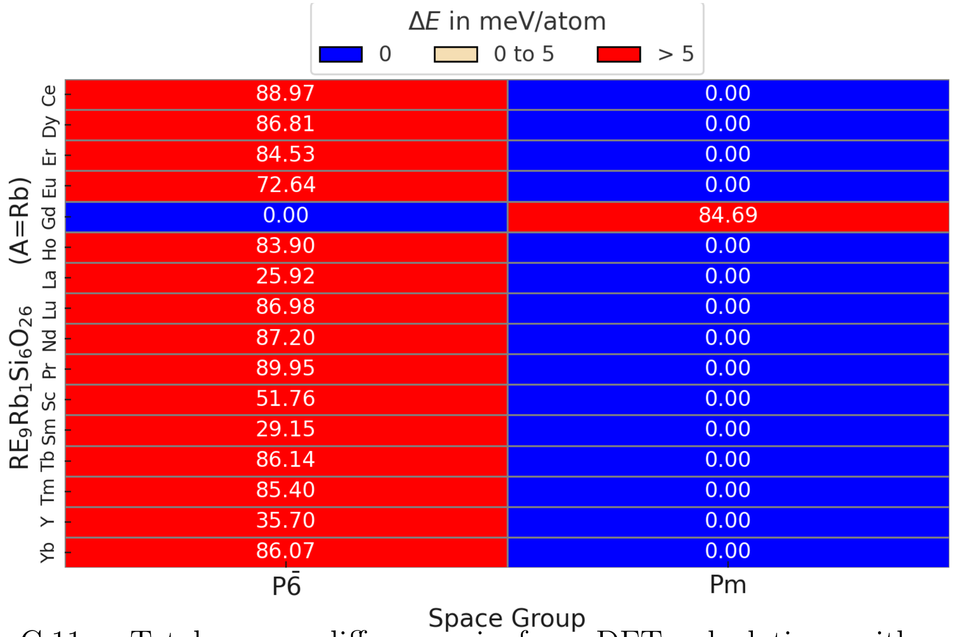


Figure C.11: Total energy difference, in from DFT calculations with respect to the lowest energy structure in $RE_9A_1Si_6O_{26}$ for A=Rb,where space groups given in parentheses indicate the final converged structure when the tolerance is set at 0.0001 or lower in FINDSYM. ΔE=0 signifies the ground structure. When ΔE is within a range of approximately 5 meV/atom, it suggests a close energetic competition with the ground state structure.

Appendix D

Publications

CTE

1. **Mukil V. Ayyasamy**, Prasanna V. Balachandran, "Correlation Between *d*-orbital Bandwidth and Coefficient of Thermal Expansion Anisotropy in RE_2SiO_5 Compounds (RE=Sc,Y,La)." *AIP Advances*, 12(4), 045012, (2022).

   Contribution: DFT paper. I contributed to the problem formulation, DFT calculation, analysis of the results, and writing of the manuscript

2. **Mukil V. Ayyasamy**, Jereon A. Deijkers, Haydn NG. Wadley, Prasanna V. Balachandran, "Density functional theory and machine learning guided search for RE2Si2O7 with targeted coefficient of thermal expansion." *Journal of the American Ceramic Society*, **103**, 4489-4497 (2021).

   Contribution: DFT + ML + Experimental paper. I contributed to the problem formulation, DFT calculation, ML, analysis of the results, and writing of the manuscript.

Model explanation (ME)

1. Kyungtae Lee, **Mukil Ayyasamy**, Paige Delsa, Timothy Q. Hartnett, and Prasanna V. Balachandran. "Phase Classification of Multi-Principal Element Alloys via Interpretable Machine Learning."*npj Comput Mater*, **8**, 25 (2022).

Contribution: ML paper. I contributed towards the conceptualization and development of the interpretable or explainable ML algorithm used in the paper.

2. Kyungtae Lee, **Mukil Ayyasamy**, and Prasanna V. Balachandran. "A Comparison of Explainable Artificial Intelligence Methods in the Phase Classification of Multi-Principal Element Alloys." *manuscript in prepraration* (2022).

   Contribution: ML paper. I contributed towards the conceptualization and development of the interpretable or explainable ML algorithm used in the paper.

Miscellaneous

1. Timothy Q. Hartnett, **Mukil V. Ayyasamy**, and Prasanna V. Balachandran, "Prediction of New Iodine-containing Apatites using Machine Learning and Density Functional Theory." *MRS Communications* **9**, 882-890 (2019).

   Contribution: DFT + Machine learning paper. Tim and I equally contributed to this paper. I was responsible for for the construction of one of the two classification algorithms, and equal portions of the DFT calculated formation energies. Writing was equally shared between Tim and myself.

2. Jon F Ihlefeld, Ting S Luk, Sean W Smith, Shelby S Fields, Samantha T Jaszewski, Daniel M Hirt, Will T Riffe, Scott Bender, Costel Constantin, **Mukil V. Ayyasamy**, Prasanna V. Balachandran, Ping Lu, M David Henry, Paul S Davids,"Compositional dependence of linear and nonlinear optical response in crystalline hafnium zirconium oxide thin films." *Journal of Applied Physics* **128**, 034101 (2019).

   Contribution: Experimental + DFT paper. I performed DFT calculations and calculated the band structures data.

Bibliography

[1] W Miller et al. "Negative thermal expansion: a review". In: *Journal of materials science* 44.20 (2009), pp. 5441–5451.

[2] Cho Yen Ho and Richard Erwin Taylor. *Thermal expansion of solids*. Vol. 4. ASM international, 1998.

[3] Rappal Sangameswara Krishnan, Ramachandran Srinivasan, and S Devanarayanan. *Thermal expansion of crystals: international series in the science of the solid state*. Elsevier, 2013.

[4] Brian Ralph. "The Iron-Nickel Alloys (A Hundred Years after the Discovery of Invar)". In: *Materials and Design* 5.17 (1996), p. 305.

[5] Charles Kittel. "Introduction to solid state physics Eighth edition". In: (2021).

[6] Eduard Grüneisen. "Theorie des festen Zustandes einatomiger Elemente". In: *Annalen der Physik* 344.12 (1912), pp. 257–306.

[7] Richard Car and Mark Parrinello. "Unified approach for molecular dynamics and density-functional theory". In: *Physical review letters* 55.22 (1985), p. 2471.

[8] RD King-Smith and RJ Needs. "A new and efficient scheme for first-principles calculations of phonon spectra". In: *Journal of Physics: Condensed Matter* 2.15 (1990), p. 3431.

[9] Charles Martinek and FA Hummel. "Linear thermal expansion of three tungstates". In: *Journal of the American Ceramic Society* 51.4 (1968), pp. 227–228.

[10] H Neumann. “Trends in the thermal expansion coefficients of the $A^{I}B^{III}C_2^{VI}$ and $A^{II}B^{IV}C_2^{V}$ chalcopyrite compounds”. In: *Kristall und Technik* 15.7 (1980), pp. 849–857.

[11] Rustum Roy, Dinesh K Agrawal, and Herbert A McKinstry. “Very low thermal expansion coefficient materials”. In: *Annual Review of Materials Science* 19.1 (1989), pp. 59–81.

[12] Arthur W Sleight. “Isotropic negative thermal expansion”. In: *Annual review of materials science* 28.1 (1998), pp. 29–43.

[13] Cora Lind. “Two decades of negative thermal expansion research: where do we stand?” In: *Materials* 5.6 (2012), pp. 1125–1154.

[14] Joseph T Schick and Andrew M Rappe. “Classical model of negative thermal expansion in solids with expanding bonds”. In: *Physical Review B* 93.21 (2016), p. 214304.

[15] Helen D Megaw. “Crystal structures and thermal expansion”. In: *Materials Research Bulletin* 6.10 (1971), pp. 1007–1018.

[16] Maryellen Cameron et al. “High-temperature crystal chemistry of acmite, diopside, hedenbergite jadeite, spodumene and ureyite”. In: *American Mineralogist: Journal of Earth and Planetary Materials* 58.7-8 (1973), pp. 594–618.

[17] Robert M Hazen and Ch T Prewitt. “Effects of temperature and pressure on interatomic distances in oxygen-based minerals”. In: *American Mineralogist* 62.3-4 (1977), pp. 309–315.

[18] Robert M Hazen. “Comparative crystal chemistry”. In: *Temperature, pressure, composition and the variation of crystal structure* (1982).

[19] Ethan T Ritz, Sabrina J Li, and Nicole A Benedek. “Thermal expansion in insulating solids from first principles”. In: *Journal of Applied Physics* 126.17 (2019), p. 171102.

[20] Pinku Nath et al. “High-throughput prediction of finite-temperature properties using the quasi-harmonic approximation”. In: *Computational Materials Science* 125 (2016), pp. 82–91.

[21] Stefano Baroni, Paolo Giannozzi, and Eyvaz Isaev. “Density-functional perturbation theory for quasi-harmonic calculations”. In: *Reviews in Mineralogy and Geochemistry* 71.1 (2010), pp. 39–57.

[22] John R Hardy. *The lattice dynamics and statics of alkali halide crystals.* Springer Science & Business Media, 2012.

[23] Atsushi Togo and Isao Tanaka. “First principles phonon calculations in materials science”. In: *Scripta Materialia* 108 (2015), pp. 1–5.

[24] Mackenzie Ridley et al. “Tailoring thermal properties of multi-component rare earth monosilicates”. In: *Acta Materialia* 195 (2020), pp. 698–707.

[25] Mukil V Ayyasamy and Prasanna V Balachandran. “Correlation between *d*-orbital bandwidth and local coordination environment in RE_2SiO_5 compounds with implications in minimizing the coefficient of thermal expansion anisotropy (RE= Sc, Y, La)”. In: *AIP Advances* 12.4 (2022).

[26] Phong CH Nguyen et al. “PARC: Physics-aware recurrent convolutional neural networks to assimilate meso scale reactive mechanics of energetic materials”. In: *Science advances* 9.17 (2023), eadd6868.

[27] Phong CH Nguyen et al. “Challenges and opportunities for machine learning in multiscale computational modeling”. In: *Journal of Computing and Information Science in Engineering* 23.6 (2023).

[28] Gui-Qin Liang and Jian Zhang. “A machine learning model for screening thermodynamic stable lead-free halide double perovskites”. In: *Computational Materials Science* 204 (2022), p. 111172.

[29] Santosh Behara, Taher Poonawala, and Tiju Thomas. “Crystal structure classification in ABO3 perovskites via machine learning”. In: *Computational Materials Science* 188 (2021), p. 110191.

[30] Francesca Tavazza, Brian DeCost, and Kamal Choudhary. “Uncertainty Prediction for Machine Learning Models of Material Properties”. In: *ACS omega* 6.48 (2021), pp. 32431–32440.

[31] Alex Zunger et al. “Special quasirandom structures”. In: *Phys. Rev. Lett.* 65 (3 July 1990), pp. 353–356. DOI: 10.1103/PhysRevLett.65.353.

[32] Susan Trolier-McKinstry and Robert E Newnham. *Materials engineering: bonding, structure, and structure-property relationships*. Cambridge University Press, 2018.

[33] George B Benedek and Felix MH Villars. *Physics with illustrative examples from medicine and biology: mechanics*. Springer Science & Business Media, 2000.

[34] Sebastiano Tosto. “Reappraising 1907 Einstein?s Model of Specific Heat”. In: *Open Journal of Physical Chemistry* 6.04 (2016), pp. 109–128.

[35] Charles Kittel and Paul McEuen. *Introduction to solid state physics*. John Wiley & Sons, 2018.

[36] Rappal Sangameswara Krishnan, Ramachandran Srinivasan, and S Devanarayanan. *Thermal expansion of crystals: international series in the science of the solid state*. Elsevier, 2013.

[37] Cho Yen Ho and Richard Erwin Taylor. *Thermal expansion of solids*. Vol. 4. ASM international, 1998.

[38] Randall F Barron and Brian R Barron. *Design for thermal stresses*. John Wiley & Sons, 2011.

[39] Sharon Ann Holgate. *Understanding solid state physics*. cRc Press, 2021.

[40] David M Jacobson and Giles Humpston. *Principles of brazing*. Asm International, 2005.

[41] Paul Kah et al. "TECHNIQUES FOR JOINING DISSIMILAR MATERIALS: METALS AND POLYMERS." In: *Reviews on Advanced Materials Science* 36.2 (2014).

[42] MB Uday et al. "Current issues and problems in the joining of ceramic to metal". In: *Joining technologies*. IntechOpen, 2016.

[43] Skyler R Hilburn. "Tailoring Thermal Expansion in Additively Manufactured Titanium Alloys to Enable Functional Grading". PhD thesis. The Pennsylvania State University, 2020.

[44] Alvin Levy. "Thermal Residual Stresses in Ceramic-to-Metal Brazed Joints". In: *Journal of the American Ceramic Society* 74.9 (1991), pp. 2141–2147.

[45] Irene Spitsberg and Jim Steibel. "Thermal and environmental barrier coatings for SiC/SiC CMCs in aircraft engine applications". In: *International Journal of Applied Ceramic Technology* 1.4 (2004), pp. 291–301.

[46] Nitin P Padture. “Advanced structural ceramics in aerospace propulsion”. In: *Nature materials* 15.8 (2016), pp. 804–809.

[47] Douglas C Hofmann et al. “Developing gradient metal alloys through radial deposition additive manufacturing”. In: *Scientific reports* 4.1 (2014), p. 5357.

[48] KS Ravichandran. “Thermal residual stresses in a functionally graded material system”. In: *Materials Science and Engineering: A* 201.1-2 (1995), pp. 269–276.

[49] TM Pollock, DM Lipkin, and KJ Hemker. “Multifunctional coating interlayers for thermal-barrier systems”. In: *MRS bulletin* 37.10 (2012), pp. 923–931.

[50] Douglas C Hofmann et al. “Compositionally graded metals: A new frontier of additive manufacturing”. In: *Journal of Materials Research* 29.17 (2014), pp. 1899–1910.

[51] Kang N Lee. “Current status of environmental barrier coatings for Si-based ceramics”. In: *Surface and Coatings Technology* 133 (2000), pp. 1–7.

[52] Kang N Lee, Nathan S Jacobson, and Robert A Miller. “Refractory oxide coatings on SiC ceramics”. In: *MRS bulletin* 19.10 (1994), pp. 35–38.

[53] Kang N Lee. “Key durability issues with mullite-based environmental barrier coatings for Si-based ceramics”. In: *J. Eng. Gas Turbines Power* 122.4 (2000), pp. 632–636.

[54] Kang N Lee and Robert A Miller. “Development and environmental durability of mullite and mullite/YSZ dual layer coatings for SiC and Si3N4 ceramics”. In: *Surface and Coatings Technology* 86 (1996), pp. 142–148.

[55] J Mesquita-Guimarães et al. "Mullite–YSZ multilayered environmental barrier coatings tested in cycling conditions under water vapor atmosphere". In: *Surface and Coatings Technology* 209 (2012), pp. 103–109.

[56] Kang N Lee et al. "Upper temperature limit of environmental barrier coatings based on mullite and BSAS". In: *Journal of the American Ceramic Society* 86.8 (2003), pp. 1299–1306.

[57] D Liu et al. "Residual stresses in environmental and thermal barrier coatings on curved superalloy substrates: Experimental measurements and modelling". In: *Materials Science and Engineering: A* 606 (2014), pp. 117–126.

[58] CV Cojocaru et al. "Performance of thermally sprayed Si/mullite/BSAS environmental barrier coatings exposed to thermal cycling in water vapor environment". In: *Surface and Coatings Technology* 216 (2013), pp. 215–223.

[59] Bradley T Richards, Hengbei Zhao, and Haydn NG Wadley. "Structure, composition, and defect control during plasma spray deposition of ytterbium silicate coatings". In: *Journal of materials science* 50 (2015), pp. 7939–7957.

[60] Yue Xu et al. "Rare earth silicate environmental barrier coatings: present status and prospective". In: *Ceramics International* 43.8 (2017), pp. 5847–5855.

[61] Y Xu and J Li. "Preparation and molten salt corrosion research of composite environmental barrier coatings of Lu2Si2O7 and Lu2SiO5". In: *Materials Research Innovations* 18.sup4 (2014), S4–958.

[62] Yan-Chun Zhou et al. "Theoretical Prediction and Experimental Investigation on the Thermal and Mechanical Properties of Bulk β-Yb 2 Si 2 O 7". In: *Journal of the American Ceramic Society* 96.12 (2013), pp. 3891–3900.

[63] XU Yue and YAN Zhaotong. “Investigation on the preparation of Si/mullite/Yb2Si2O7 environmental barrier coatings onto silicon carbide”. In: *Journal of Rare Earths* 28.3 (2010), pp. 399–402.

[64] IA Bondar. “Rare-earth silicates”. In: *ceramics international* 8.3 (1982), pp. 83–89.

[65] Robert Vaßen et al. “Overview on advanced thermal barrier coatings”. In: *Surface and Coatings Technology* 205.4 (2010), pp. 938–942. ISSN: 0257-8972. DOI: https://doi.org/10.1016/j.surfcoat.2010.08.151.

[66] Irene Spitsberg and Jim Steibel. “Thermal and Environmental Barrier Coatings for SiC/SiC CMCs in Aircraft Engine Applications”. In: *International Journal of Applied Ceramic Technology* 1.4 (2004), pp. 291–301. DOI: 10.1111/j.1744-7402.2004.tb00181.x.

[67] Nitin P. Padture. “Advanced Structural Ceramics in Aerospace Propulsion”. In: *Nature Materials* 15 (July 2016), p. 804. DOI: https://doi.org/10.1038/nmat4687.

[68] K.N. Lee. “Current status of environmental barrier coatings for Si-Based ceramics”. In: *Surface and Coatings Technology* 133-134 (2000), pp. 1–7. ISSN: 0257-8972. DOI: https://doi.org/10.1016/S0257-8972(00)00889-6.

[69] Nasrin Al Nasiri et al. “Thermal Properties of Rare-Earth Monosilicates for EBC on Si-Based Ceramic Composites”. In: *Journal of the American Ceramic Society* 99.2 (2016), pp. 589–596. DOI: 10.1111/jace.13982.

[70] Jun Ito. “Silicate apatites and oxyapatites”. In: *American Mineralogist: Journal of Earth and Planetary Materials* 53.5-6 (1968), pp. 890–907.

[71] Zhixue Qu et al. “Thermal conductivity of the gadolinium calcium silicate apatites: Effect of different point defect types”. In: *Acta Materialia* 59.10 (2011), pp. 3841–3850.

[72] Yuji Masubuchi, Mikio Higuchi, and Kohei Kodaira. “Reinvestigation of phase relations around the oxyapatite phase in the Nd2O3–SiO2 system”. In: *Journal of crystal growth* 247.1-2 (2003), pp. 207–212.

[73] Han Zhang et al. “A promising molten silicate resistant material: Rare-earth oxy-apatite RE9. 33 (SiO4) 6O2 (RE= Gd, Nd or La)”. In: *Journal of the European Ceramic Society* 40.12 (2020), pp. 4101–4110.

[74] Yiran Li, Jiemin Wang, and Jingyang Wang. “Theoretical investigation of phonon contributions to thermal expansion coefficients for rare earth monosilicates RE2SiO5 (RE= Dy, Ho, Er, Tm, Yb and Lu)”. In: *Journal of the European Ceramic Society* 40.7 (2020), pp. 2658–2666.

[75] F Lofaj et al. “Thermal expansion and glass transition temperature of the rare-earth doped oxynitride glasses”. In: *Journal of the European Ceramic Society* 24.12 (2004), pp. 3377–3385.

[76] Paul F Becher et al. “Compositional effects on the properties of Si-Al-RE-based oxynitride glasses (RE= La, Nd, Gd, Y, or Lu)”. In: *Journal of the American Ceramic Society* 85.4 (2002), pp. 897–902.

[77] Yixiu Luo et al. “Material-genome perspective towards tunable thermal expansion of rare-earth di-silicates”. In: *Journal of the European Ceramic Society* 38.10 (2018), pp. 3547–3554.

[78] Koichi Momma and Fujio Izumi. "VESTA: a three-dimensional visualization system for electronic and structural analysis". In: *Journal of Applied crystallography* 41.3 (2008), pp. 653–658.

[79] Koichi Momma and Fujio Izumi. "*VESTA*: a three-dimensional visualization system for electronic and structural analysis". In: *Journal of Applied Crystallography* 41.3 (June 2008), pp. 653–658. DOI: 10.1107/S0021889808012016.

[80] Robert M Hazen, Robert T Downs, and Charles T Prewitt. "Principles of comparative crystal chemistry". In: *Reviews in Mineralogy and Geochemistry* 41.1 (2000), pp. 1–33.

[81] Kristin A Denault et al. "Structure–composition relationships and optical properties in cerium-substituted (Sr, Ba)$_3$(Y, La)(BO$_3$)$_3$ borate phosphors". In: *Journal of Materials Chemistry C* 1.44 (2013), pp. 7339–7345.

[82] Xuelong Wang et al. "Quantitative structure-property relationship study of cathode volume changes in lithium ion batteries using ab-initio and partial least squares analysis". In: *Journal of Materiomics* 3.3 (2017), pp. 178–183.

[83] Ryosuke Uehara et al. "Effect of tin substitution on the chemical composition and thermal expansion properties of Zr2SP2O12". In: *Journal of Asian Ceramic Societies* 9.3 (2021), pp. 1194–1203.

[84] Siyuan Zhang et al. "Estimation thermal expansion coefficient from lattice energy for inorganic crystals". In: *Japanese journal of applied physics* 45.11R (2006), p. 8801.

[85] I David Brown. *The chemical bond in inorganic chemistry: the bond valence model*. Vol. 27. Oxford university press, 2016.

[86] Feliciano Giustino. *Materials modelling using density functional theory: properties and predictions*. Oxford University Press, 2014.

[87] Gabriel R Schleder et al. “From DFT to machine learning: recent approaches to materials science–a review”. In: *Journal of Physics: Materials* 2.3 (2019), p. 032001.

[88] Wenhao Sun et al. “The thermodynamic scale of inorganic crystalline metastability”. In: *Science advances* 2.11 (2016), e1600225.

[89] Alex Zunger et al. “Special quasirandom structures”. In: *Physical review letters* 65.3 (1990), p. 353.

[90] Tom M Mitchell and Tom M Mitchell. *Machine learning*. Vol. 1. 9. McGraw-hill New York, 1997.

[91] Stavros P Adam et al. “No free lunch theorem: A review”. In: *Approximation and optimization: Algorithms, complexity and applications* (2019), pp. 57–82.

[92] Sreerama K Murthy. “Automatic construction of decision trees from data: A multi-disciplinary survey”. In: *Data mining and knowledge discovery* 2 (1998), pp. 345–389.

[93] Madan Somvanshi et al. “A review of machine learning techniques using decision tree and support vector machine”. In: *2016 international conference on computing communication control and automation (ICCUBEA)*. IEEE. 2016, pp. 1–7.

[94] Leo Breiman. “Random forests”. In: *Machine learning* 45 (2001), pp. 5–32.

[95] Jerome H Friedman. “Greedy function approximation: a gradient boosting machine”. In: *Annals of statistics* (2001), pp. 1189–1232.

[96] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[97] Mostafa Sabzekar and Seyed Mohammad Hossein Hasheminejad. “Robust regression using support vector regressions”. In: *Chaos, Solitons & Fractals* 144 (2021), p. 110738.

[98] Prasanna V Balachandran et al. “Experimental search for high-temperature ferroelectric perovskites guided by two-step machine learning”. In: *Nature communications* 9.1 (2018), p. 1668.

[99] Laurens van der Maaten and Geoffrey Hinton. “Visualizing Data using t-SNE”. In: *Journal of Machine Learning Research* 9 (2008), pp. 2579–2605.

[100] Jize Zhang, Bhavya Kailkhura, and T Yong-Jin Han. “Leveraging uncertainty from deep learning for trustworthy material discovery workflows”. In: *ACS omega* 6.19 (2021), pp. 12711–12721.

[101] Julia Ling et al. “High-dimensional materials and process optimization using data-driven experimental design with well-calibrated uncertainty estimates”. In: *Integrating Materials and Manufacturing Innovation* 6 (2017), pp. 207–217.

[102] Fumiaki Tanaka et al. “Materials informatics for process and material co-optimization”. In: *2018 International Symposium on Semiconductor Manufacturing (ISSM)*. IEEE. 2018, pp. 1–3.

[103] Alexandre Abraham and Léo Dreyfus-Schmidt. “Sample noise impact on active learning”. In: *arXiv preprint arXiv:2109.01372* (2021).

[104] Chen Tao. “Applications of Bayesian Neural Networks in Outlier Detection”. In: *Big Data* (2023).

[105] Christian Fiedler, Carsten W Scherer, and Sebastian Trimpe. “Practical and rigorous uncertainty bounds for gaussian process regression”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 35. 8. 2021, pp. 7439–7447.

[106] Osman Mamun et al. “Uncertainty quantification for Bayesian active learning in rupture life prediction of ferritic steels”. In: *Scientific Reports* 12.1 (2022), p. 2083.

[107] Anastasios N Angelopoulos and Stephen Bates. “A gentle introduction to conformal prediction and distribution-free uncertainty quantification”. In: *arXiv preprint arXiv:2107.07511* (2021).

[108] Marc C. Kennedy and Anthony O’Hagan. “Bayesian calibration of computer models”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.3 (2001), pp. 425–464. ISSN: 1467-9868. DOI: 10.1111/1467-9868.00294.

[109] Benjamin Lu. “Constructing prediction intervals for random forests”. PhD thesis. Ph. D. thesis, Pomona College, 2017.

[110] Marie-Hélène Roy and Denis Larocque. “Prediction intervals with random forests”. In: *Statistical Methods in Medical Research* 29.1 (2020), pp. 205–229.

[111] Sricharan Kumar and Ashok N Srivistava. “Bootstrap prediction intervals in non-parametric regression with applications to anomaly detection”. In: *The 18th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ARC-E-DAA-TN6188. 2012.

[112] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Vol. 1. 10. Springer series in statistics New York, 2001.

[113] Carlos Mougan and Dan Saattrup Nielsen. “Monitoring model deterioration with explainable uncertainty estimation via non-parametric bootstrap”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. 12. 2023, pp. 15037–15045.

[114] Anthony Christopher Davison and David Victor Hinkley. *Bootstrap methods and their application*. 1. Cambridge university press, 1997.

[115] Isis Didier Lins et al. “Computing confidence and prediction intervals of industrial equipment degradation by bootstrapped support vector regression”. In: *Reliability Engineering & System Safety* 137 (2015), pp. 120–128.

[116] Ashley N Henderson, Steven K Kauwe, and Taylor D Sparks. “Benchmark datasets incorporating diverse tasks, sample sizes, material systems, and data heterogeneity for materials informatics”. In: *Data in Brief* 37 (2021), p. 107262.

[117] Alex J. Smola and Bernhard Schölkopf. “A tutorial on support vector regression”. In: *Statistics and Computing* 14 (2004), pp. 199–222.

[118] Przemysław Biecek. “DALEX: explainers for complex predictive models in R”. In: *The Journal of Machine Learning Research* 19.1 (2018), pp. 3245–3249.

[119] Lloyd S Shapley. *A value for n-person games*. Princeton University Press, 2016.

[120] Alex Goldstein et al. “Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation”. In: *journal of Computational and Graphical Statistics* 24.1 (2015), pp. 44–65.

[121] Mateusz Staniak and Przemysław Biecek. “Explanations of Model Predictions with live and breakDown Packages”. In: *The R Journal* 10.2 (2019), p. 395. ISSN: 2073-4859. DOI: 10.32614/rj-2018-072. URL: http://dx.doi.org/10.32614/RJ-2018-072.

[122] Ghanshyam Pilania. “Machine learning in materials science: From explainable predictions to autonomous design”. In: *Computational Materials Science* 193 (2021), p. 110360.

[123] Yihuang Xiong et al. “Data-driven analysis of the electronic-structure factors controlling the work functions of perovskite oxides”. In: *Physical Chemistry Chemical Physics* 23.11 (2021), pp. 6880–6887.

[124] Sanphawat Phromphithak, Thossaporn Onsree, and Nakorn Tippayawong. “Machine learning prediction of cellulose-rich materials from biomass pretreatment with ionic liquid solvents”. In: *Bioresource Technology* 323 (2021), p. 124642.

[125] Zhiliang Wang et al. “Machine Learning Guided Dopant Selection for Metal Oxide based Photoelectrochemical Water Splitting: The Case Study of Fe2O3 and CuO”. In: *Advanced Materials* (2021), p. 2106776.

[126] Przemysław Biecek, Szymon Maksymiuk, and Hubert Baniecki. *moDel Agnostic Language for Exploration and eXplanation*. R package version 2.2.0. 2021. URL: https://dalex.drwhy.ai,%20https://github.com/ModelOriented/DALEX.

[127] Scott M Lundberg and Su-In Lee. “A unified approach to interpreting model predictions”. In: *Proceedings of the 31st international conference on neural information processing systems*. 2017, pp. 4768–4777.

[128] Przemyslaw Biecek and Tomasz Burzykowski. *Explanatory model analysis: explore, explain, and examine predictive models*. CRC Press, 2021.

[129] Kyungtae Lee et al. “Phase classification of multi-principal element alloys via interpretable machine learning”. In: *npj Computational Materials* 8.1 (2022), pp. 1–12.

[130] Kyungtae Lee et al. “A comparison of explainable artificial intelligence methods in the phase classification of multi-principal element alloys”. In: *Scientific Reports* 12.1 (2022), p. 11591.

[131] Laura R. Turcer and Nitin P. Padture. “Towards multifunctional thermal environmental barrier coatings (TEBCs) based on rare-earth pyrosilicate solid-solution ceramics”. In: *Scripta Materialia* 154 (2018), pp. 111–117. ISSN: 1359-6462. DOI: https://doi.org/10.1016/j.scriptamat.2018.05.032. URL: http://www.sciencedirect.com/science/article/pii/S1359646218303269.

[132] Kang N. Lee, Dennis S. Fox, and Narottam P. Bansal. “Rare earth silicate environmental barrier coatings for SiC/SiC composites and Si_3N_4 ceramics”. In: *Journal of the European Ceramic Society* 25.10 (2005), pp. 1705–1715. ISSN: 0955-2219. DOI: https://doi.org/10.1016/j.jeurceramsoc.2004.12.013. URL: http://www.sciencedirect.com/science/article/pii/S095522190400531X.

[133] Kang N. Lee. “$Yb_2Si_2O_7$ Environmental barrier coatings with reduced bond coat oxidation rates via chemical modifications for long life”. In: *Journal of the American Ceramic Society* 102.3 (2019), pp. 1507–1521. DOI: 10.1111/jace.15978. URL: https://ceramics.onlinelibrary.wiley.com/doi/abs/10.1111/jace.15978.

[134] Nasrin Al Nasiri et al. “Thermal Properties of Rare-Earth Monosilicates for EBC on Si-Based Ceramic Composites”. In: *Journal of the American Ceramic Society* 99.2 (2016), pp. 589–596. DOI: 10.1111/jace.13982.

[135] Yue Xu et al. “Rare earth silicate environmental barrier coatings: Present status and prospective”. In: *Ceramics International* 43.8 (2017), pp. 5847–5855. ISSN: 0272-8842. DOI: https://doi.org/10.1016/j.ceramint.2017.01.153.

[136] Bradley T. Richards, Matthew R. Begley, and Haydn N.G. Wadley. “Mechanisms of Ytterbium Monosilicate/Mullite/Silicon Coating Failure During Thermal Cycling in Water Vapor”. In: *Journal of the American Ceramic Society* 98.12 (2015), pp. 4066–4075. DOI: 10.1111/jace.13792.

[137] Xiaomin Ren et al. “Equiatomic quaternary $(Y_{1/4}Ho_{1/4}Er_{1/4}Yb_{1/4})_2SiO_5$ silicate: A perspective multifunctional thermal and environmental barrier coating material”. In: *Scripta Materialia* 168 (2019), pp. 47–50. ISSN: 1359-6462. DOI: https://doi.org/10.1016/j.scriptamat.2019.04.018.

[138] Susumu Fujii et al. “Role of phonons on phase stabilization of $RE_2Si_2O_7$ over wide temperature range (RE = Yb, Gd)”. In: *Journal of the European Ceramic Society* (2019). ISSN: 0955-2219. DOI: https://doi.org/10.1016/j.jeurceramsoc.2019.10.060.

[139] J. Felsche. “Polymorphism and crystal data of the rare-earth disilicates of type $RE_2Si_2O_7$”. In: *Journal of the Less Common Metals* 21.1 (1970), pp. 1–14. ISSN: 0022-5088. DOI: https://doi.org/10.1016/0022-5088(70)90159-1. URL: http://www.sciencedirect.com/science/article/pii/0022508870901591.

[140] Alberto J. Fernández-Carrión, Mathieu Allix, and Ana I. Becerro. “Thermal Expansion of Rare-Earth Pyrosilicates”. In: *Journal of the American Ceramic Society* 96.7 (2013), pp. 2298–2305. DOI: 10.1111/jace.12388. URL: https://ceramics.onlinelibrary.wiley.com/doi/abs/10.1111/jace.12388.

[141] Yu. I. Smolin and Yu. F. Shepelev. “The crystal structures of the rare earth pyrosilicates”. In: *Acta Crystallographica Section B* 26.5 (May 1970), pp. 484–492. DOI: 10.1107/S0567740870002698.

[142] MD Dolan et al. “Structures and anisotropic thermal expansion of the α, β, γ, and δ polymorphs of $Y_2Si_2O_7$”. In: *Powder Diffraction* 23.1 (2008), pp. 20–25.

[143] J.Y. Wang, Y.C. Zhou, and Z.J. Lin. “Mechanical properties and atomistic deformation mechanism of γ-$Y_2Si_2O_7$ from first-principles investigations”. In: *Acta Materialia* 55.17 (2007), pp. 6019–6026. ISSN: 1359-6454. DOI: https://doi.org/10.1016/j.actamat.2007.07.010.

[144] Yan-Chun Zhou et al. “Theoretical Prediction and Experimental Investigation on the Thermal and Mechanical Properties of Bulk β-$Yb_2Si_2O_7$”. In: *Journal of the American Ceramic Society* 96.12 (2013), pp. 3891–3900. DOI: 10.1111/jace.12618.

[145] Yixiu Luo et al. “Theoretical Predictions on Elastic Stiffness and Intrinsic Thermal Conductivities of Yttrium Silicates”. In: *Journal of the American Ceramic Society* 97.3 (2014), pp. 945–951. DOI: 10.1111/jace.12764.

[146] Bin Liu et al. “Investigation of Native Point Defects and Nonstoichiometry Mechanisms of Two Yttrium Silicates by First-Principles Calculations”. In: *Journal of the American Ceramic Society* 96.10 (2013), pp. 3304–3311. DOI: 10.1111/jace.12474.

[147] Yixiu Luo et al. “Theoretical study on crystal structures, elastic stiffness, and intrinsic thermal conductivities of β-, γ-, and δ-$Y_2Si_2O_7$”. In: *Journal of Materials Research* 30.4 (2015), pp. 493–502. DOI: 10.1557/jmr.2015.1.

[148] J D James et al. “A review of measurement techniques for the thermal expansion coefficient of metals and alloys at elevated temperatures”. In: *Measurement Science and Technology* 12.3 (Feb. 2001), R1–R15. DOI: 10.1088/0957-0233/12/3/201.

[149] Vanessa Nilsen et al. “Prediction of concrete coefficient of thermal expansion and other properties using machine learning”. In: *Construction and Building Materials* 220 (2019), pp. 587–595. ISSN: 0950-0618. DOI: https://doi.org/10.1016/j.conbuildmat.2019.05.006.

[150] Nafisa Bano and Michel Nganbe. “Modeling of Thermal Expansion Coefficients of Ni-Based Superalloys Using Artificial Neural Network”. In: *Journal of Materials Engineering and Performance* 22.4 (2013), pp. 952–957.

[151] Akane Suzuki, Chen Shen, and Natarajan Chennimalai Kumar. “Application of computational tools in alloy design”. In: *MRS Bulletin* 44.4 (2019), pp. 247–251.

[152] Paolo Giannozzi et al. “QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials”. In: *Journal of Physics: Condensed Matter* 21.39 (2009), p. 395502.

[153] John P. Perdew et al. “Restoring the Density-Gradient Expansion for Exchange in Solids and Surfaces”. In: *Phys. Rev. Lett.* 100 (13 Apr. 2008), p. 136406. DOI: 10.1103/PhysRevLett.100.136406.

[154] David Vanderbilt. “Soft self-consistent pseudopotentials in a generalized eigenvalue formalism”. In: *Phys. Rev. B* 41 (11 Apr. 1990), pp. 7892–7895. DOI: 10.1103/PhysRevB.41.7892.

[155] Hendrik J. Monkhorst and James D. Pack. “Special points for Brillouin-zone integrations”. In: *Phys. Rev. B* 13 (12 June 1976), pp. 5188–5192. DOI: 10.1103/PhysRevB.13.5188.

[156] Andrea Dal Corso. “Pseudopotentials periodic table: From H to Pu ”. In: *Computational Materials Science* 95 (2014), pp. 337–350. ISSN: 0927-0256. DOI: http://dx.doi.org/10.1016/j.commatsci.2014.07.043.

[157] Harold T. Stokes and Dorian M. Hatch. “FINDSYM: program for identifying the space-group symmetry of a crystal”. In: *Journal of Applied Crystallography* 38.1 (Feb. 2005), pp. 237–238. DOI: 10.1107/S0021889804031528.

[158] S. Zec et al. “Low temperature $Ce_2Si_2O_7$ polymorph formed by mechanical activation”. In: *Materials Chemistry and Physics* 95.1 (2006), pp. 150–153. ISSN: 0254-0584. DOI: https://doi.org/10.1016/j.matchemphys.2005.05.036.

[159] A. Cuneyt Tas and Mufit Akinc. “Phase Relations in the System Ce_2O_3–$Ce_2Si_2O_7$ in the Temperature Range 1150° to 1970°C in Reducing and Inert Atmospheres”. In: *Journal of the American Ceramic Society* 77.11 (1994), pp. 2953–2960. DOI: 10.1111/j.1151-2916.1994.tb04530.x.

[160] Sankaran Nair Renjini et al. “Microwave Dielectric Properties and Low Temperature Sintering of $Sm_2Si_2O_7$ Ceramic for Substrate Applications”. In: *International Journal of Applied Ceramic Technology* 6.2 (2009), pp. 286–294. DOI: 10.1111/j.1744-7402.2008.02271.x.

[161] Damian M. Cupid and Hans J. Seifert. “Thermodynamic Calculations and Phase Stabilities in the Y-Si-C-O System”. In: *Journal of Phase Equilibria and Diffusion* 28.1 (2007), pp. 90–100.

[162] Huahai Mao, Malin Selleby, and Olga Fabrichnaya. "Thermodynamic reassessment of the Y_2O_3-Al_2O_3-SiO_2 system and its subsystems". In: *Calphad* 32.2 (2008), pp. 399–412. ISSN: 0364-5916. DOI: https://doi.org/10.1016/j.calphad.2008.03.003.

[163] O. Fabrichnaya et al. "Phase equilibria and thermodynamics in the Y_2O_3-Al_2O_3-SiO_2 system". In: *Z. Metallkd.* 92.9 (2001), pp. 1083–1097.

[164] Jean-Pierre Doucet et al. "Nonlinear SVM Approaches to QSPR/QSAR Studies and Drug Design". In: *Current Computer-Aided Drug Design* 3.4 (2007), pp. 263–289. DOI: 10.2174/157340907782799372.

[165] Chun-Hsin Wu, Jan-Ming Ho, and D. T. Lee. "Travel-time prediction with support vector regression". In: *IEEE Transactions on Intelligent Transportation Systems* 5.4 (2004), pp. 276–281. DOI: 10.1109/TITS.2004.837813.

[166] David Meyer et al. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. R package version 1.6-7. 2015. URL: http://CRAN.R-project.org/package=e1071.

[167] R Core Team. *R: A Language and Environment for Statistical Computing*. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria, 2012. URL: http://www.R-project.org/.

[168] David P. MacKinnon, Chondra M. Lockwood, and Jason Williams. "Confidence Limits for the Indirect Effect: Distribution of the Product and Resampling Methods". In: *Multivariate Behavioral Research* 39.1 (2004), pp. 99–128. DOI: 10.1207/s15327906mbr3901_4.

[169] James Carpenter and John Bithell. "Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians". In: *Statistics in*

Medicine 19.9 (2000), pp. 1141–1164. DOI: 10.1002/(SICI)1097-0258(20000515)19:9<1141::AID-SIM479>3.0.CO;2-F.

[170] R. Polikar. “Bootstrap - Inspired Techniques in Computation Intelligence”. In: *IEEE Signal Processing Magazine* 24.4 (July 2007), pp. 59–72. ISSN: 1558-0792. DOI: 10.1109/MSP.2007.4286565.

[171] Weimo Zhu. “Making Bootstrap Statistical Inferences: A Tutorial”. In: *Research Quarterly for Exercise and Sport* 68.1 (1997), pp. 44–55. DOI: 10.1080/02701367.1997.10608865.

[172] Prasanna V. Balachandran, Dezhen Xue, and Turab Lookman. “Structure–Curie temperature relationships in $BaTiO_3$-based ferroelectric perovskites: Anomalous behavior of $(Ba,Cd)TiO_3$ from DFT, statistical inference, and experiments”. In: *Phys. Rev. B* 93 (14 Apr. 2016), p. 144111. DOI: 10.1103/PhysRevB.93.144111.

[173] Prasanna V. Balachandran et al. “Predicting displacements of octahedral cations in ferroelectric perovskites using machine learning”. In: *Acta Crystallographica Section B* 73.5 (Oct. 2017), pp. 962–967. DOI: 10.1107/S2052520617011945. URL: https://doi.org/10.1107/S2052520617011945.

[174] Ludwig Holleis, B. S. Shivaram, and Prasanna V. Balachandran. “Machine learning guided design of single-molecule magnets for magnetocaloric applications”. In: *Applied Physics Letters* 114.22 (2019), p. 222404. DOI: 10.1063/1.5094553.

[175] R. C. Breneman and J. W. Halloran. “Hysteresis upon Repeated Cycling through the Beta-Alpha Cristobalite Transformation”. In: *Journal of Ceramic Science and Technology* 6.1 (2001), pp. 55–62. DOI: 10.4416/JCST2014-00048.

[176] Yirong He et al. “Development of refractory silicate-yttria-stabilized zirconia dual-layer thermal barrier coatings”. In: *Journal of Thermal Spray Technology* 9.1 (2000), pp. 59–67.

[177] Zhilin Tian et al. “Towards thermal barrier coating application for rare earth silicates RE_2SiO_5 (RE= La, Nd, Sm, Eu, and Gd)”. In: *Journal of the European Ceramic Society* 39.4 (2019), pp. 1463–1476. ISSN: 0955-2219. DOI: https://doi.org/10.1016/j.jeurceramsoc.2018.12.015.

[178] Jürgen Felsche. “The crystal chemistry of the rare-earth silicates”. In: *Rare earths*. Springer, 1973, pp. 99–197.

[179] Ziqi Sun, Meishuan Li, and Yanchun Zhou. “Thermal properties of single-phase Y_2SiO_5”. In: *Journal of the European Ceramic Society* 29.4 (2009), pp. 551–557.

[180] Nasrin Al Nasiri et al. “Thermal properties of rare-earth monosilicates for EBC on Si-based ceramic composites”. In: *Journal of the American Ceramic Society* 99.2 (2016), pp. 589–596.

[181] Mackenzie Ridley et al. “TAILORING THERMAL AND CHEMICAL PROPERTIES OF MULTI-FUNCTIONAL THERMAL/ENVIRONMENTAL BARRIER COATINGS”. In: ().

[182] Koichiro Fukuda and Hiroyuki Matsubara. “Anisotropic thermal expansion in yttrium silicate”. In: *Journal of materials research* 18 (2003), pp. 1715–1722.

[183] Zhilin Tian et al. “Theoretical and experimental determination of the major thermo-mechanical properties of RE2SiO5 (RE= Tb, Dy, Ho, Er, Tm, Yb, Lu, and Y) for environmental and thermal barrier coating applications”. In: *Journal of the European Ceramic Society* 36.1 (2016), pp. 189–202.

[184] Zuhair S Khan et al. "Synthesis and characterization of Yb and Er based monosilicate powders and durability of plasma sprayed Yb2SiO5 coatings on C/C–SiC composites". In: *Materials Science and Engineering: B* 177.2 (2012), pp. 184–189.

[185] Y Ogura, M Kondo, and T Morimoto. "Y sub 2 SiO sub 5 as oxidation resistant coating for C/C composites". In: *Tenth International Conference on Composite Materials. IV. Characterization and Ceramic Matrix Composites.* 1995, pp. 767–774.

[186] A.G. Evans. "Microfracture from thermal expansion anisotropy-I. Single phase systems". In: *Acta Metallurgica* 26.12 (1978), pp. 1845–1853. ISSN: 0001-6160. DOI: https://doi.org/10.1016/0001-6160(78)90097-4.

[187] C. M. Weyant et al. "Residual Stress and Microstructural Evolution in Environmental Barrier Coatings of Tantalum Oxide Alloyed with Aluminum Oxide and Lanthanum Oxide". In: *Journal of the American Ceramic Society* 89.3 (2006), pp. 971–978. DOI: 10.1111/j.1551-2916.2005.00830.x.

[188] A Seidl et al. "Generalized Kohn-Sham schemes and the band-gap problem". In: *Physical Review B* 53.7 (1996), p. 3764.

[189] MKY Chan and Gerbrand Ceder. "Efficient band gap prediction for solids". In: *Physical review letters* 105.19 (2010), p. 196403.

[190] Michael O'Neill and Conor Ryan. "Grammatical evolution". In: *IEEE Transactions on Evolutionary Computation* 5.4 (2001), pp. 349–358.

[191] Conor Ryan, John James Collins, and Michael O Neill. "Grammatical evolution: Evolving programs for an arbitrary language". In: *European conference on genetic programming*. Springer. 1998, pp. 83–96.

[192] Farzad Noorian, Anthony M. de Silva, and Philip H. W. Leong. “gramEvol: Grammatical Evolution in R”. In: *Journal of Statistical Software* 71.1 (2016), pp. 1–26. DOI: 10.18637/jss.v071.i01.

[193] Laura R. Turcer, Arundhati Sengupta, and Nitin P. Padture. “Low thermal conductivity in high-entropy rare-earth pyrosilicate solid-solutions for thermal environmental barrier coatings”. In: *Scripta Materialia* 191 (2021), pp. 40–45. ISSN: 1359-6462. DOI: https://doi.org/10.1016/j.scriptamat.2020.09.008.

[194] Luchao Sun et al. “A multicomponent γ-type $(Gd_{1/6}Tb_{1/6}Dy_{1/6}Tm_{1/6}Yb_{1/6}Lu_{1/6})_2Si_2O_7$ disilicate with outstanding thermal stability”. In: *Materials Research Letters* 8.11 (2020), pp. 424–430.

[195] Yu Dong et al. “High-entropy environmental barrier coating for the ceramic matrix composites”. In: *Journal of the European Ceramic Society* 39.7 (2019), pp. 2574–2579.

[196] AJ Fernández-Carrión et al. “Revealing structural detail in the high temperature $La_2Si_2O_7$–$Y_2Si_2O_7$ phase diagram by synchrotron powder diffraction and nuclear magnetic resonance spectroscopy”. In: *The Journal of Physical Chemistry C* 116.40 (2012), pp. 21523–21535.

[197] Mukil V Ayyasamy et al. “Density functional theory and machine learning guided search for $RE_2Si_2O_7$ with targeted coefficient of thermal expansion”. In: *Journal of the American Ceramic Society* 103.8 (2020), pp. 4489–4497.

[198] Andrew C Strzelecki et al. “High-Temperature Thermodynamics of Cerium Silicates, A-Ce2Si2O7, and Ce4. 67 (SiO4) 3O”. In: *ACS Earth and Space Chemistry* 4.11 (2020), pp. 2129–2143.

[199] Hiroki Okudera et al. "Temperature dependence of structural parameters in oxide-ion-conducting Nd9. 33 (SiO4) 6O2: single crystal X-ray studies from 295 to 900 K". In: *Journal of Solid State Chemistry* 177.12 (2004), pp. 4451–4458.

[200] ST Misture et al. "Synthesis, crystal structure, and anisotropic thermal expansion of Dy4. 67 (SiO4) 3O". In: *Journal of materials research* 19.8 (2004), pp. 2330–2335.

[201] Koichiro Fukuda, Toru Asaka, and Tomohiro Uchida. "Thermal expansion of lanthanum silicate oxyapatite (La9. 33+ 2x (SiO4) 6O2+ 3x), lanthanum oxyorthosilicate (La2SiO5) and lanthanum sorosilicate (La2Si2O7)". In: *Journal of Solid State Chemistry* 194 (2012), pp. 157–161.

[202] Jamesa L Stokes. *Thermal Expansion Coefficients of Ca2Y8 (SiO4) 6O2 and Ca2Yb8 (SiO4) 6O2 Apatite-type Silicates*. Tech. rep. 2021.

[203] Laurens Van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE." In: *Journal of machine learning research* 9.11 (2008).

[204] Fangfang Xia et al. "Generalized stacking fault energies of $Cr_{23}C_6$ carbide: a first-principles study". In: *Computational Materials Science* 158 (2019), pp. 20–25.

[205] P Boutinaud et al. "Spectroscopic investigations on $SrAl_2O_4$ polymorphs". In: *Journal of Luminescence* 159 (2015), pp. 158–165.

[206] Moureen C Kemei et al. "Crystal structures of spin-Jahn–Teller-ordered $MgCr_2O_4$ and $ZnCr_2O_4$". In: *Journal of Physics: Condensed Matter* 25.32 (2013), p. 326001.

[207] A. Walle and G. Ceder. "Automating first-principles phase diagram calculations". en. In: *Journal of Phase Equilibria* 23.4 (Aug. 2002), pp. 348–359. ISSN: 1054-9714, 1863-7345. DOI: 10.1361/105497102770331596. URL: http://link.springer.com/10.1361/105497102770331596 (visited on 07/02/2021).

[208] A van de Walle and M Asta. "Self-driven lattice-model Monte Carlo simulations of alloy thermodynamic properties and phase diagrams". In: *Modelling and Simulation in Materials Science and Engineering* 10.5 (Sept. 2002), pp. 521–538. ISSN: 0965-0393. DOI: 10.1088/0965-0393/10/5/304. URL: https://iopscience.iop.org/article/10.1088/0965-0393/10/5/304 (visited on 07/02/2021).

[209] A. van de Walle et al. "Efficient stochastic generation of special quasirandom structures". In: *Calphad* 42 (2013), pp. 13–18. ISSN: 0364-5916. DOI: https://doi.org/10.1016/j.calphad.2013.06.006. URL: https://www.sciencedirect.com/science/article/pii/S0364591613000540.

[210] Alejandro Salanova et al. "Phase stability and tensorial thermal expansion properties of single to high-entropy rare-earth disilicates". In: *Journal of the American Ceramic Society* 106.5 (2023), pp. 3228–3238.

[211] Brandon M. Greenwell. "pdp: An R Package for Constructing Partial Dependence Plots". In: *The R Journal* 9.1 (2017), pp. 421–436. URL: https://journal.r-project.org/archive/2017/RJ-2017-016/index.html.

[212] Farzad Noorian, Anthony M de Silva, and Philip HW Leong. "gramEvol: Grammatical evolution in R". In: *Journal of Statistical Software* 71 (2016), pp. 1–26.

[213] John D. Kelleher, Brian Mac Namee, and Aoife D'arcy. "Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies". In: The MIT Press, 2020.