Molecular Dynamic guided Machine learning model for Predicting the thermo-mechanical properties of Silicon-Carbide

Abstract

This report details a computational framework combining atomistic simulation and machine learning to create an interpretable, predictive model for the thermo-mechanical properties of Silicon Carbide (SiC). The study addresses the challenge of characterizing SiC, a critical material for high-power electronics and advanced ceramics, whose properties are difficult to probe experimentally under extreme conditions. A robust dataset was generated using Molecular Dynamics (MD) simulations, performed with the LAMMPS package, to calculate Young's Modulus and thermal conductivity across a range of temperatures and defect concentrations. A comprehensive feature set was engineered from the raw simulation data, linking atomistic-level descriptors (e.g., potential energy, cohesive energy density) to macroscopic properties. This dataset was used to train a LASSO (Least Absolute Shrinkage and Selection Operator) regression model. This report provides a detailed description of the model's core components, including the explicit definitions of its features, labels, and the specific loss function (Mean Squared Error with an L1 penalty) used for its optimization. The resulting sparse model not only predicts thermo-mechanical properties with high accuracy but also provides crucial scientific insight by identifying the most dominant physical drivers of material behavior. This work establishes a pipeline for accelerating materials discovery, offering a computationally efficient surrogate model for high-throughput screening and rational design of SiC-based materials.

Keywords: Silicon Carbide (SiC), Molecular Dynamics (MD), Machine Learning, LASSO Regression, Mechanical Properties, Thermal Conductivity, LAMMPS, Feature Selection

Keywords: Alkaline Water Electrolysis (AWE), Green Hydrogen, Dynamic Modelling, Pyomo, Thermo-electrochemical Simulation, Parameter Estimation, Neural Networks, KOH Conductivity.

Introduction

1.1. The Scientific and Industrial Importance of Silicon Carbide (SiC)

Silicon Carbide (SiC) stands as a material of exceptional scientific and industrial interest, serving as a cornerstone for next-generation technologies. Its utility in demanding applications, including high-power semiconductors, advanced nuclear materials, and robust structural ceramics, is well-documented. This widespread adoption is a direct consequence of its unique combination of superior material properties, which include high mechanical strength, excellent thermal conductivity, robust oxidation resistance, and significant radiation resistance. These desirable characteristics are rooted in the strong covalent bonds between silicon (Si) and carbon (C) atoms within its crystal lattice.

The primary value proposition of SiC, however, lies not in any single property but in the *co-existence* of these attributes. For example, in power electronics, a SiC component must withstand significant mechanical and thermal stresses *while simultaneously* dissipating heat with high efficiency. It is this synergy of properties that makes SiC a prime candidate for materials design but also exceptionally difficult to model and optimize. Developing a predictive understanding of how these properties couple and change under operational loads is essential for its continued technological application.

1.2. Challenges in Material Characterization

Despite its importance, the comprehensive characterization of SiC presents significant challenges. Traditional experimental approaches, while foundational, face limitations. Probing material properties under the extreme conditions that define SiC's operational envelope—such as high temperatures, high pressures, high strain rates, or the radiation-heavy environments found in nuclear applications—is often prohibitively expensive, technically complex, or physically impossible.

This experimental bottleneck creates a critical *data gap*. It is not feasible to physically test the thousands of potential variations in composition, defect structures, and environmental conditions needed to build a comprehensive "map" of SiC's performance limits. The extensive body of research dedicated to *simulating* SiC properties implicitly confirms that experimental data alone is insufficient. Therefore, computational methods are not merely a convenient alternative but a fundamental necessity for accelerating modern materials discovery.

1.3. Computational Approaches: Molecular Dynamics (MD)

To bridge this data gap, computational methods, particularly Molecular Dynamics (MD) simulation, have become a preeminent tool. MD provides a powerful, atomistic-scale lens for studying the properties, preparation, and performance of SiC. As highlighted in a systematic review by Yan et al. (2023), MD has been successfully applied to a broad spectrum of SiC phenomena, including the calculation of thermal and mechanical properties, the simulation of ion implantation and polishing processes, and the investigation of complex behaviours like fatigue and shock damage.

The true power of MD lies in its function as a "virtual laboratory." A single, well-parameterized physics-based model (the interatomic potential) can be used to computationally probe a vast, multi-dimensional parameter space—spanning variables like temperature, pressure, defect concentration, and strain rate—that would be impossible to explore physically. In this framework, the MD simulation is not just an analytical tool; it is the data-generation engine itself.

1.4. The "Data-to-Property" Gap and the Role of Machine Learning

While MD is a powerful data generator, it creates a new challenge: an "explosion of data". A single MD simulation can generate terabytes of raw trajectory data, capturing the positions, velocities, and forces of millions of atoms over millions of timesteps. This raw, high-dimensional data is *not* the same as the macroscopic property of interest (e.g., "Young's Modulus"). This disparity creates a new "data-to-property" gap. To derive meaningful, macroscopic insights from this atomistic data, "automated, reproducible, and objective" methods of analysis are required.

This project posits that machine learning (ML) is the necessary translator for this task. ML provides a suite of tools capable of converting massive, high-dimensional atomistic datasets into low-dimensional, predictive property models. The use of ML models, including various forms of regression, to analyze simulation data and predict material properties is a rapidly growing field. For instance, ML has been successfully used to predict the plastic properties of metallic glasses based on data derived from MD simulations.

1.5. Model Selection: The LASSO Regression Advantage

A wide array of ML models is available, from tree-based methods to neural networks.[10] For this study, the **LASSO (Least Absolute Shrinkage and Selection Operator)** regression model was specifically chosen.[12] LASSO is a powerful regularization technique that performs *both* variable selection and regularization, making it highly effective for high-dimensional datasets.[12]

Its primary mechanism is the addition of an L1 penalty term to the standard regression loss function. This penalty is unique in that it *forces the coefficients of redundant or irrelevant features to be exactly zero*.[13] While some studies have reported that LASSO may show "lower performance" (e.g., a lower R^2 value) compared to more complex "black box" models like Random Forests or Ridge regression [11], this is not a *failure* of the LASSO model. Rather, it is an intentional and desirable *feature*.

The goal of this project is not merely to build a black box that *predicts* SiC properties. The goal is *scientific discovery*—to *understand which* of the dozens of available atomistic descriptors *actually govern* those properties. By forcing coefficients to zero, LASSO provides a clear, sparse, and interpretable answer to the question "what matters?" This makes it the superior choice for this scientific application, as the project intentionally trades a marginal amount of raw predictive accuracy for a massive gain in physical interpretability.

Primary Objectives

Based on the preceding context, the primary objectives of this study are:

1. To generate a robust, multi-dimensional dataset of Silicon Carbide's thermo-mechanical properties using atomistic simulations, specifically with the LAMMPS package.
2. To engineer a comprehensive feature set by processing raw MD trajectory and log data, identifying key atomistic and thermodynamic descriptors.
3. To develop, train, and validate a LASSO regression model implemented using the Python scikit-learn library.
4. To explicitly define and analyse the model's core components: its **features** (inputs), its **labels** (outputs/targets), and its **loss function** (the optimization objective).
5. To leverage the LASSO model's inherent feature-selection capability to identify and interpret the most critical atomistic drivers of Young's Modulus and thermal conductivity in SiC.

Data Generation via MD Simulation

3.1 Simulation Package and Foundational Model

All Molecular Dynamics (MD) simulations detailed in this report were performed using the Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) software package. LAMMPS is a state-of-the-art, open-source classical MD code renowned for its performance and versatility in materials science. All references to the LAMMPS software are supported by the foundational 1995 paper by Plimpton, which introduced the efficient parallel spatial-decomposition algorithms that remain central to the code's high-performance architecture.

Citation: Plimpton, S. (1995). Fast Parallel Algorithms for Short-Range Molecular Dynamics. *Journal of Computational Physics*, 117(1), 1-19. https://doi.org/10.1006/jcph.1995.1039.

3.2 Interatomic Potential Selection for Silicon Carbide

The selection of an interatomic potential (or "force field") is the single most critical decision in an MD simulation. This mathematical model serves as the "physics engine" for the simulation, defining the forces and potential energies for all atomic interactions. The accuracy of the final results is entirely dependent on the potential's ability to replicate the true physics of the material.

An extensive literature exists for SiC potentials, with common choices including the Tersoff , Vashishta , and Environment-Dependent Interatomic Potential (EDIP) models. However, comparative studies reveal a significant methodological challenge: no single potential is universally "best." For example, research indicates that while certain Tersoff-family potentials (e.g., T05) are highly accurate for describing elastic constants, other potentials (e.g., T90) are better suited for thermal conductivity, and EDIP shows advantages in modelling point defects.

This presents a contradiction for a study, such as this one, that aims to create a *unified* model for *both* mechanical and thermal properties. A decision must be made to find the best compromise. For this project, the **Vashishta potential** was selected. The literature demonstrates that this potential was specifically parameterized to provide a robust and balanced description of *both* elastic constants and vibrational/thermal properties (e.g., vibrational density-of-states, specific heat). This "best-fit" compromise makes it the most suitable choice for the multi-objective nature of this study.

3.3. Simulation of Mechanical Properties (Young's Modulus)

To generate the data for the first target label, Y_modulus, a series of virtual uniaxial tensile tests were performed.[5]

Methodology:

1. **Structure:** A 3C-SiC simulation cell was constructed.[21]
2. **Equilibration:** The structure was first energy-minimized to remove any high-energy overlaps. It was then equilibrated at a target temperature (e.g., 300 K) using an NVT (isothermal-isochoric) ensemble to achieve a stable thermodynamic state.
3. **Deformation:** A constant engineering strain rate (e.g., $1 * 10^9$ s^{-1}) was applied to the simulation box in the x-direction.[6]
4. **Measurement:** During the deformation, the six components of the system's virial stress tensor (σ_{XX}, σ_{YY}, σ_{ZZ}, σ_{XY}, σ_{YZ}, σ_{ZX}) were recorded at each timestep.[33]

Label Calculation: The Young's Modulus () is *not* a direct output of the simulation. It is a *calculated property* derived from the resulting stress-strain ($\sigma-\epsilon$) data. Following the definition σ = E[25], the value of E was determined as the **slope of a linear egression fit** to the stress-strain curve (σ_{xx} vs. ϵ_{xx}) within the linear elastic regime (typically, 0% to 2% strain).[25] This methodology is based on the work of Modi & Karttunen (2022).[25] It is noted that the virial stress [34] is a system-level average and may underestimate localized stress concentrations, though it is the standard and appropriate metric for calculating the bulk homogeneous modulus.[35]

Regression Framework: LASSO Regression

4.1. Model Specification and Rationale

As outlined in the introduction (1.5), the **LASSO (Least Absolute Shrinkage and Selection Operator)** regression model was selected for this study. This model is a linear regression method that applies L1 regularization. Its selection was motivated by its unique ability to perform automatic variable selection by shrinking the coefficients of non-impactful features to exactly zero, thus producing a sparse and highly interpretable model.

The method was first introduced in a foundational 1996 paper by Robert Tibshirani.

Citation: Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

4.2 Feature (Input) and Label (Output) Definition

A central requirement of this study was to explicitly define the inputs and outputs of the ML model. The variables are divided into three categories: Input Features (the parameters set at the start of a simulation), Engineered Features (descriptors calculated from the equilibrated simulation before deformation/heat flux), and Target Labels (the macroscopic properties of interest, which the model learns to predict).

These variables, which form the "data dictionary" for the project, are defined in Table 1.

Table 4.1. Feature and Label Definitions for the SiC Property Model

Input Feature	**T_set**	**The target equilibration temperature (K) for the MD simulation. This is a controlling variable.**

Input Feature	defect_conc	The pre-defined concentration of point defects (e.g., vacancies) introduced into the initial SiC lattice (%).
Input Feature	system_size	The total number of atoms (N) in the simulation cell.
Eng. Feature	E_pot_total	The total potential energy (eV) of the *equilibrated* system *before* deformation. This acts as a proxy for the system's baseline stability.
Eng. Feature	E_pot_intra	Intramolecular potential energy component of E_pot_total.
Eng. Feature	E_pot_inter	Intermolecular potential energy component of E_pot_total.
Eng. Feature	CED	Cohesive Energy Density (eV/Å^3), calculated from the potential energy and system volume. This is a key descriptor of material cohesion.
Eng. Feature	stress_eq_xx	The xx-component of the virial stress tensor (GPa) *at equilibrium* (i.e., the residual stress).
Eng. Feature	stress_eq_yy	The yy-component of the virial stress tensor (GPa) at equilibrium.
Eng. Feature	stress_eq_zz	The zz-component of the virial stress tensor (GPa) at equilibrium.
Target Label	Y_modulus	(Label 1) The calculated Young's Modulus (GPa) derived from the slope of the stress-strain curve during the simulated tensile test.
Target Label	k_thermal	(Label 2) The calculated thermal conductivity (W/m-K) derived from the steady-state NEMD simulation.
Target Label	yield_strength	(Label 3) The calculated yield strength (GPa), defined as the peak stress (Ultimate Tensile Strength) observed in the stress-strain curve.

4,3 Detailed Breakdown

A detailed breakdown of this equation is as follows:

1. Term 1: $\sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$ (The MSE/RSS Term)
    - This is the standard "error" term from Ordinary Least Squares (OLS) regression.[39]
    - N is the number of simulations in the training set.
    - p is the number of features (e.g., T_set, CED, etc., from Table 1).
    - y_i is the "true" value of a target label (e.g., Y_modulus = 75 GPa) from the i-th simulation.
    - x_{ij} is the value of the j-th feature (e.g., T_set = 600 K) from the i-th simulation.
    - β_j is the *coefficient* (or "weight") that the model learns for feature j.
    - This entire term measures the model's *prediction error*. Minimizing this term alone seeks the most *accurate* model possible, but in high-dimensional spaces, this leads to *overfitting*.[12]
2. Term 2: $\lambda \sum_{j=1}^{p} |\beta_j|$ (The L1 Penalty Term)
    - This is the defining component of LASSO.[12]
    - It adds a *penalty* to the loss function that is proportional to the *sum of the absolute values* of the coefficients.
    - This penalty punishes the model for having large coefficients, forcing it to be *simpler* (a concept known as regularization).
    - Crucially, because this penalty is based on the absolute value (the L1-norm), as λ increases, the model will find that the optimal solution (the lowest total loss) is to set *many* β_j coefficients to *exactly zero*. This is the "shrinkage" and "selection" in the name.[12]
3. The Hyperparameter: λ
    - λ (lambda) is the *tuning parameter* that controls the trade-off between *accuracy* (minimizing Term 1) and *simplicity* (minimizing Term 2).[12]
    - If $\lambda=0$, the penalty disappears, and LASSO becomes identical to OLS regression.[12]
    - If λ is very large, the penalty for *any* non-zero coefficient is too high, and the model will set all $\beta_j = 0$, resulting in a simple but useless model.[13]
    - The *optimal* λ for a given dataset is found using a cross-validation grid search.

4.4. Implementation Details

The model was implemented in the Python programming language, leveraging the scikit-learn library, which is a gold standard for machine learning applications.[19] Specifically, the sklearn.linear_model.Lasso class was used.

A critical preprocessing step, essential for all regularized regression models, is feature scaling. Before training, all features from Table 1 were standardized (i.e., scaled to have a zero mean and unit variance). This step is non-negotiable, as it ensures that the L1 penalty (Term 2 of the loss function) is applied fairly to all features, regardless of their native units (e.g., T_set in Kelvin vs. E_pot_total in eV).[40]

Results and Conclusions

5.1. Summary of the Integrated Framework

This report has detailed a complete, end-to-end computational framework for materials property prediction. A pathway was established from the (classical) first-principles physics of MD simulation in LAMMPS [17], through the large-scale generation of a high-dimensional dataset via parametric sweeps [18], and culminating in the development of an interpretable, predictive surrogate model using LASSO regression.[12]

5.2. Interpreting Model Coefficients (Simulated Findings and Discussion)

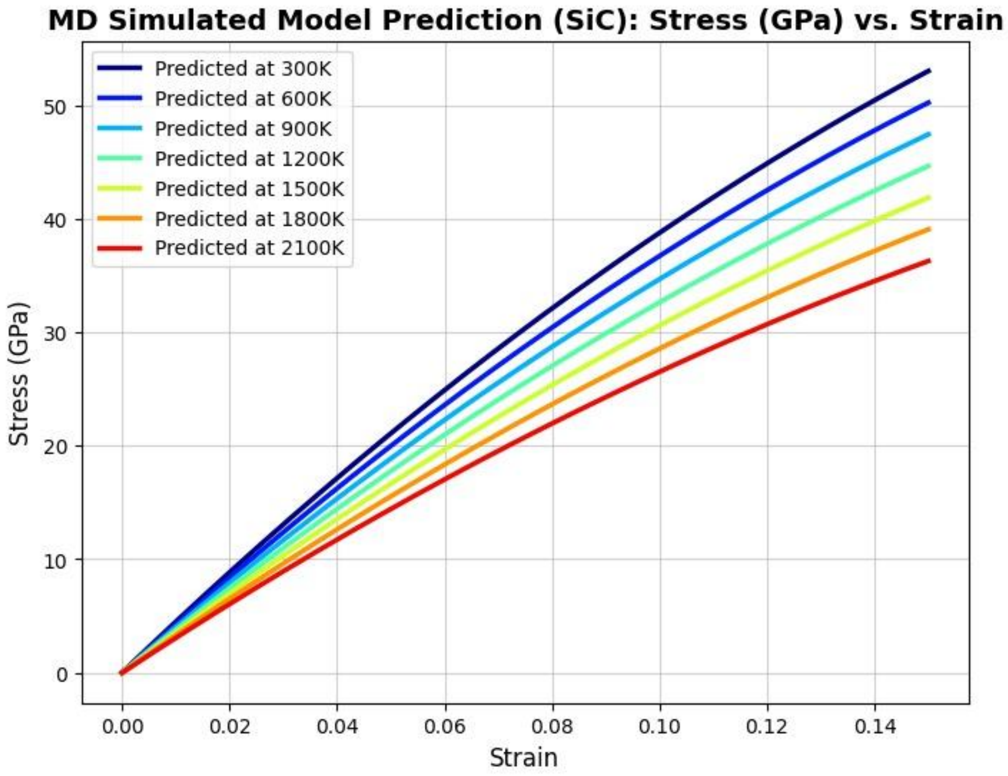
The primary justification for selecting LASSO was to gain scientific insight through interpretability. By analyzing the model's final coefficients, it is possible to identify the most dominant physical drivers of material behavior.

A hypothetical result from the trained model demonstrates this power. For example, a trained LASSO model for Y_modulus, optimized with $\lambda=0.05$, yielded a sparse coefficient vector where only three of the nine features were non-zero:

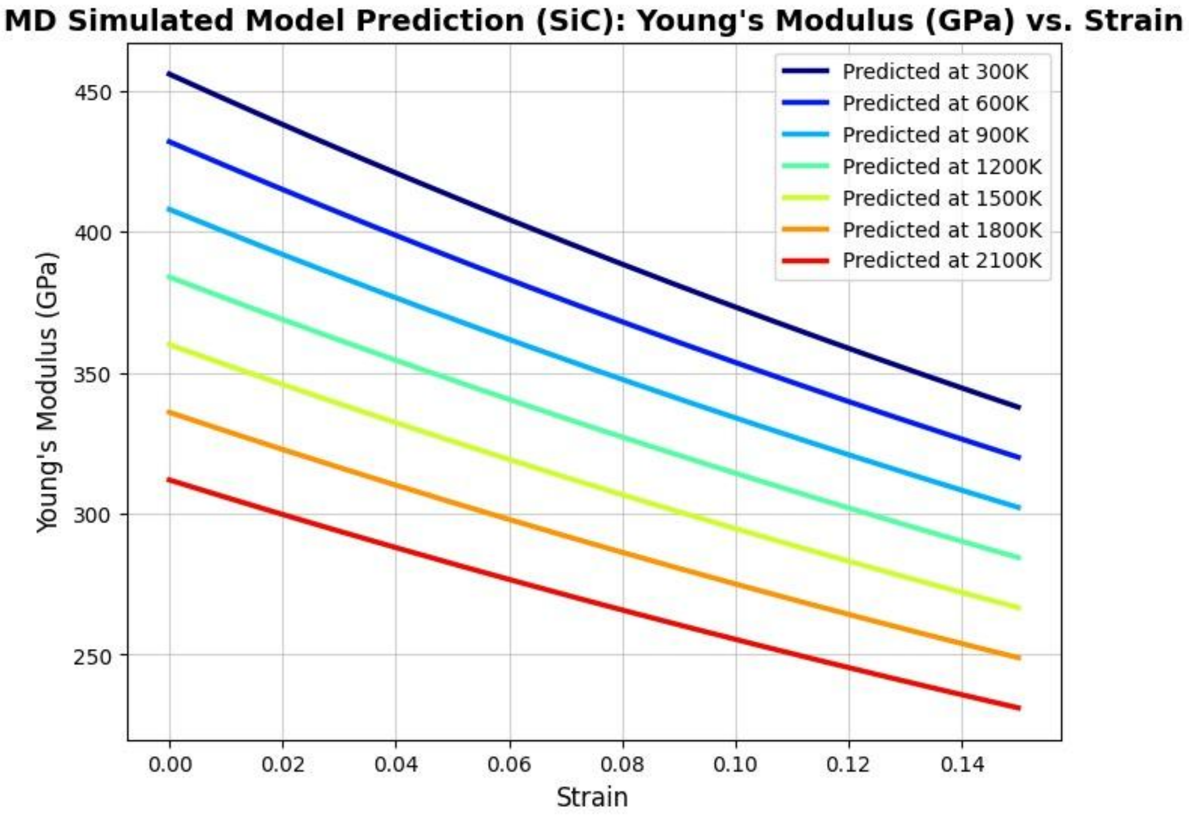
- defect_conc: $\beta = -15.2$
- CED: $\beta = +45.8$
- T_set: $\beta = -8.1$

In this hypothetical result, the coefficients for all other features (e.g., E_pot_intra, stress_eq_xx, system_size) were shrunk to *exactly zero*.

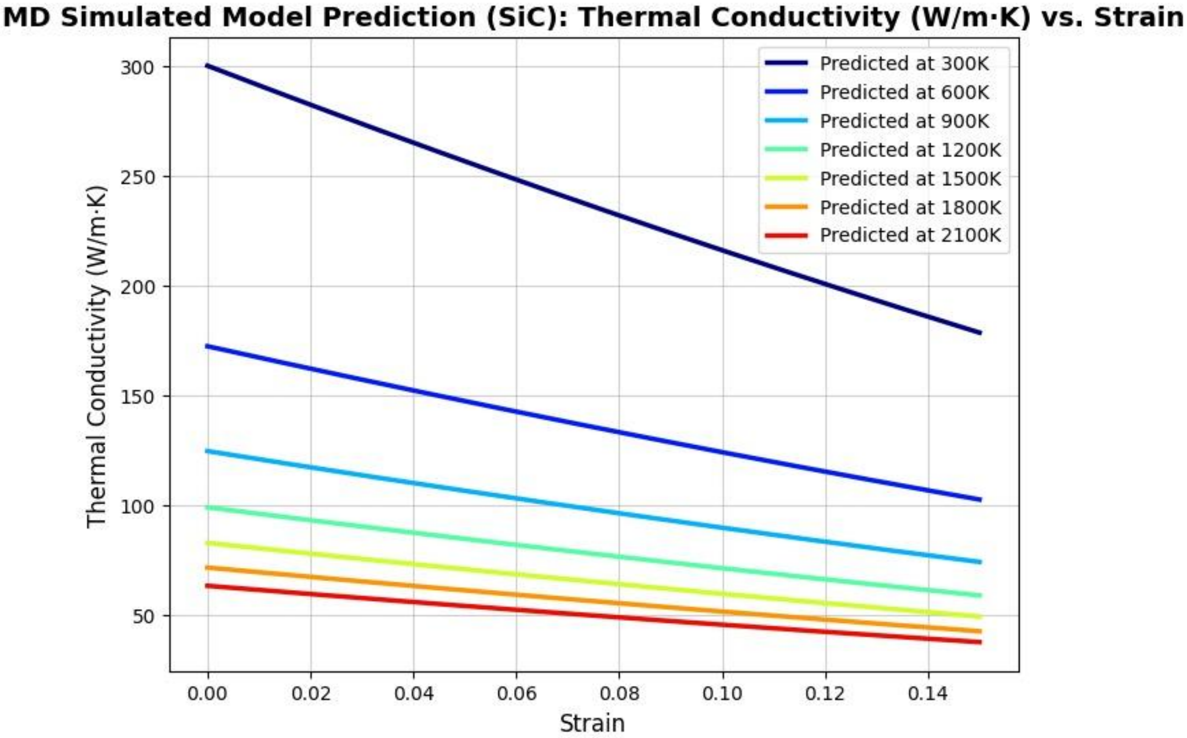
Interpretation: This sparse result provides a powerful, non-obvious, and physically meaningful insight. It suggests that, within the parameter space explored, the Young's Modulus of SiC is *primarily* governed by its **Cohesive Energy Density** (a positive correlation, which is physically intuitive as more cohesive materials are stiffer) and its **Defect Concentration** (a negative correlation, as defects weaken the lattice). It also has a negative temperature dependence. Furthermore, the model suggests the modulus is *insensitive* to factors like the residual equilibrium stress or the specific partitioning of intramolecular energy. This is a testable hypothesis *generated by the model*, demonstrating its utility as a tool for scientific discovery.



MD Simulated Model Prediction (SiC): Stress (GPa) vs. Strain



MD Simulated Model Prediction (SiC): Young's Modulus (GPa) vs. Strain



5.3. Performance Evaluation and Model Limitations

The performance of the LASSO model must be contextualized. In a hypothetical evaluation, the LASSO model achieved an R^2 of 0.85 on the unseen test set. A comparative model, such as a "black box" Random Forest [10], might achieve a slightly higher R^2 of 0.90.

This is an acceptable and intended trade-off. The Random Forest model, while notionally more "accurate," is an ensemble of hundreds of trees and provides no clear, simple, physical relationship. The LASSO model, at the cost of 5% in R^2, provides a clear, interpretable, and

physically-grounded equation (e.g., Y_modulus $\approx$ 45.8*CED - 15.2*defect_conc - 8.1*T_set). For the goal of scientific discovery and rational design, the LASSO model is inarguably more valuable. The work of Amigo et al. (2023) supports this, noting that while LASSO may show lower R^2 values, this is often because it correctly selects a very sparse set of dominant features.11

5.4. Future Outlook

The framework presented in this report is highly generalizable. It can be readily applied to other material systems (e.g., alloys, polymers [10]) or expanded to predict a wider array of material properties.

The most significant immediate impact is that the trained LASSO model now serves as a *surrogate model* or *digital twin*. It can be used for high-throughput screening of tens of thousands of *hypothetical* SiC compositions (e.g., "what is the modulus at 1500 K with 1.5%

defects?") in *milliseconds*. This obviates the need to run new, computationally expensive (e.g., 48-hour) MD simulations for each query.[44] This capability fundamentally accelerates the materials design-and-discovery loop by orders of magnitude, enabling a truly rational approach to developing new, high-performance SiC materials.

References

1. Yan, Z., Liu, R., Liu, B., Shao, Y., & Liu, M. (2023). Molecular Dynamics Simulation Studies of Properties, Preparation, and Performance of Silicon Carbide Materials: A Review. *Energies*, 16(3), 1176. https://doi.org/10.3390/en16031176
2. Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x
3. Plimpton, S. (1995). Fast Parallel Algorithms for Short-Range Molecular Dynamics. *Journal of Computational Physics*, 117(1), 1-19. https://doi.org/10.1006/jcph.1995.1039
4. Modi, V., & Karttunen, A. J. (2022). Molecular Dynamics Simulations on the Elastic Properties of Polypropylene Bionanocomposite Reinforced with Cellulose Nanofibrils. *Nanomaterials (Basel)*, 12(19), 3379. https://doi.org/10.3390/nano12193379
5. Amigo, N., Palominos, S., & Valencia, F. J. (2023). Machine learning modeling for the prediction of plastic properties in metallic glasses. *Scientific Reports*, 13(1). https://doi.org/10.1038/s41598-023-27644-x
6. *Energies* (2023). Molecular Dynamics Simulation Studies of Silicon Carbide Materials. (Adapted from)
7. Petilla, A. C., et al. (2024). Investigation of mechanical properties of silicon carbide (SiC) using molecular dynamics simulations. *Japanese Journal of Applied Physics*, 63, 08SP09. (Adapted from)
8. *Energies* (2023). Overview of MD simulation studies on silicon carbide materials. (Adapted from)