Chapter 1

Introduction

1.1 What is Sigma Phase

Stainless steels have become an indispensable class of materials in modern engineering due to their excellent corrosion resistance, mechanical strength, and versatility [11]. They find applications across diverse industries, including chemical processing, aerospace, construction, and energy, where reliability and durability are paramount. However, the performance of stainless steels can be compromised by the formation of certain undesirable phases during processing and service.

One of the most critical issues affecting stainless steels is the precipitation of the sigma phase (σ-phase). The sigma phase is a brittle, intermetallic compound that typically forms in duplex, ferritic, and some austenitic stainless steels when these alloys are exposed to intermediate temperature ranges (approximately 600–900 °C) over extended periods [15].

This phase is characterized by a complex tetragonal structure and is enriched in chromium and molybdenum—elements that are essential for corrosion resistance [20]. As a result, its formation not only embrittles the material but also depletes the surrounding matrix of crucial alloying elements, thereby undermining the corrosion resistance and overall mechanical integrity of the steel.

1.2 Problem with Sigma Phase

The detrimental effects of sigma phase precipitation manifest in several ways. It significantly reduces the toughness and ductility of the material, making it more susceptible to crack initiation and propagation under stress [14] (. In welded structures, the localized formation of sigma phase can lead to weld embrittlement, which is particularly concerning in safety-critical applications [16]. Additionally, the depletion of chromium from the matrix due to sigma phase formation compromises the protective passive oxide layer, thereby increasing the risk of corrosion [21].

Despite extensive research, the complex interplay of thermodynamic and kinetic factors that govern sigma phase formation presents significant challenges. Traditional approaches based on empirical models and phase diagrams, while valuable, often lack the predictive precision required for modern materials design [20]. Furthermore, the influence of trace elements—present only in minute quantities—adds another layer of complexity to the understanding and control of sigma phase precipitation. These trace elements can act as catalysts or inhibitors, subtly modifying the nucleation and growth kinetics of the sigma phase and thereby affecting the overall performance of the alloy [17].

1.3 Kinetics of Sigma Phase

σ-Phase is a brittle intermetallic compound that precipitates in duplex, ferritic, and austenitic stainless steels when held between 600 °C and 900 °C, depleting Cr and Mo from the matrix and impairing mechanical integrity and corrosion resistance [15]. Predicting its onset time requires capturing both thermodynamic drives and kinetic constraints. Classical nucleation theory (CNT) and the Johnson–Mehl–Avrami–Kolmogorov (JMAK) equation provide mechanistic descriptions of nucleation and growth but rely on accurate inputs that are costly to obtain experimentally [19] Recent advances in machine learning offer a complementary route, enabling rapid, data-driven predictions by integrating multiple data sources [9] .

Classical Nucleation Theory (CNT) is a thermodynamic framework that describes how a new phase (like a solid precipitate) forms from a parent phase (like a liquid or another solid) through the process of nucleation. Early work characterized the σ-phase crystal structure and its deleterious effects in steels with high Cr and Mo content [12]. CNT describes nucleation rates as a function of supersaturation and interfacial energy, while the Avrami equation models growth kinetics:

$$X(t)=1-\exp(-ktn)$$

where:

n = Avrami exponent (related to the nucleation and growth mechanism)

The Avrami equation models how fast a new phase forms and grows in a material. It is used to fit experimental data on phase transformation (like sigma phase precipitation) [19].

- Relation to Sigma Phase Formation in Stainless Steels: The Avrami equation can describe how the fraction of sigma phase increases over time at a constant temperature. It helps in determining:
- Time-temperature-transformation (TTT) behavior [18]
- Kinetics of sigma phase growth [17]
- Effect of alloying elements on transformation rate [15]

Transformation Kinetics

The overall transformation kinetics often follow an Arrhenius-type behavior, where the rate of sigma phase formation increases exponentially with temperature [19]. The holding time at critical temperatures is a critical factor—longer exposure increases the likelihood of significant sigma phase precipitation.

Competitive Phase Formation

In many stainless steels, sigma phase formation competes with other transformations (such as carbide precipitation or the formation of other intermetallic phases). This competition can further influence the kinetics and the final fraction of sigma phase in the microstructure [14].

Chapter 2

Literature Review

2.1 Effect of Trace Alloying Elements on Onset Time of Sigma Phase

This chapter critically appraises previous work related to sigma phase formation in stainless steels and the application of machine learning in materials science.

2.2 Historical and Fundamental Aspects

Early studies identified sigma phase as a brittle intermetallic compound with a complex tetragonal crystal structure, primarily forming in duplex and ferritic stainless steels [14]. Researchers established that high chromium and molybdenum levels, in combination with specific thermal exposures, catalyze the nucleation and growth of sigma phase [15]. Models based on thermodynamic and kinetic principles (e.g., Arrhenius-type behavior) have been developed to describe the formation process [19].

2.3 Kinetics and Thermodynamics

The nucleation and growth kinetics of sigma phase have been extensively studied. Nucleation is influenced by grain boundary characteristics and the presence of trace elements [17], while growth is controlled by atomic diffusion and local compositional gradients [18]. Several simulation studies using CALPHAD methods and diffusion models have provided insights into the time–temperature-transformation (TTT) behavior of sigma phase [10]

2.4 Role of Trace Elements

Although present in minor concentrations, trace elements such as silicon, phosphorus, and Sulphur have been shown to modify the kinetics of sigma phase formation [15]. They can either promote or inhibit the nucleation process, significantly affecting the phase stability and mechanical properties of the alloy.

2.5 Machine Learning in Materials Science

Recent advances in ML have enabled researchers to predict phase transformations and material properties with high accuracy [9]. Studies have applied regression, classification, and deep learning techniques to model complex relationships between alloy composition, processing parameters, and microstructural evolution [13]. These approaches have led to more efficient materials design and reduced experimental trial-and-error. Early studies characterized the σ-phase tetragonal structure and its deleterious effects in high-Cr and Mo steels [12]. Empirical TTT/CCT diagrams remain the design standard but are time- and resource-intensive

[20]. Machine learning approaches have successfully predicted TTT curves for high-alloy steels using ensemble methods, outperforming traditional software like JMatPro [9]. Large DFT datasets and supervised learning yield σ-phase formation enthalpies with MAEs ~23 me/atom, providing thermodynamic descriptors for kinetic modeling [22]. CNT links these enthalpies to nucleation barriers, offering a bridge between theory and experiment [19].

Chapter 3

Objectives

The following objectives have been derived from the above introduction, literature survey, and discussion.

1. To analyze the formation of sigma phase in various types of steels by examining key factors such as temperature, onset time, and the percentage composition of different elements using machine learning and prediction models.
2. We aim to identify which elements most significantly influence the rate of sigma phase formation, contributing to a better understanding of the material behavior and aiding in the optimization of steel properties.
3. The project aims to build a predictive model for the onset time of sigma phase provide valuable insights into optimizing heat treatment processes, preventing undesirable embrittlement, and improving the overall performance of high-alloy steels in industrial applications.

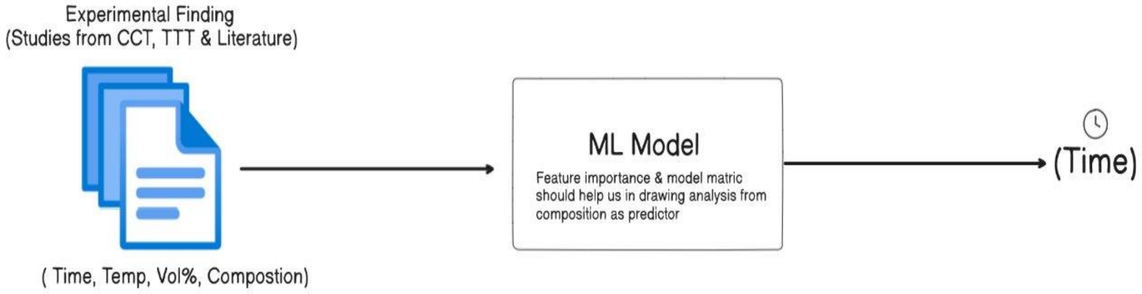

Experimental Finding
(Studies from CCT, TTT & Literature)
(Time, Temp, Vol%, Compostion)
ML Model
Feature importance & model matric should help us in drawing analysis from composition as predictor
(Time)

Fig.3.1. Objective description

Chapter 4

Methodology

4.1. Methodology Used

The project commenced with a comprehensive data collection phase, where relevant data points related to the chemical composition of stainless steels and their corresponding heat treatment conditions were gathered. This included crucial alloying elements like chromium, nickel, molybdenum, and silicon, with the quantity of sigma phase in different samples.

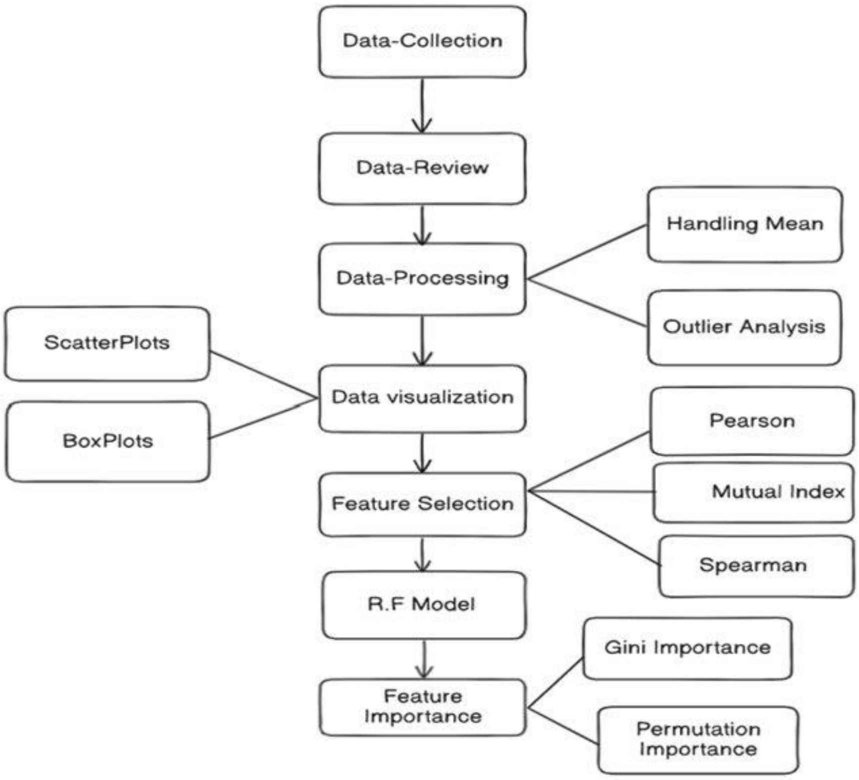


Fig 4.1.1: Methodology

The aim was to build a dataset rich in both experimental and literature-derived values to ensure a strong foundation for predictive modelling.

Following this, a data review was conducted to assess the quality of the dataset. In this stage, entries were carefully examined for inconsistencies, duplicates, and missing values. The data processing phase involved addressing these issues—mean imputation was applied to handle missing numerical values, and outlier analysis was carried out to detect and mitigate abnormal data points that could skew the model's performance. These steps ensured the dataset was clean and suitable for machine learning applications.

Once the data was cleaned, data visualization techniques were employed to gain insights into relationships between variables. Scatter plots helped visualize correlations between alloying elements and sigma phase formation, while box plots revealed the spread and presence of outliers in the data. This visual exploration laid the groundwork for effective feature selection.

The next step involved feature selection using statistical correlation techniques such as Pearson, Spearman, and Mutual Information Index. These methods helped in identifying features that had the strongest impact on the target variable—the amount or presence of sigma phase. By selecting only the most relevant features, the model's complexity was reduced while retaining its predictive power.

With a refined set of features, the Random Forest Regression model was developed. Random Forest, being an ensemble of decision trees, was chosen for its ability to model non-linear relationships and its robustness to noise and overfitting. Additionally, it performs well on datasets with complex feature interactions, which is ideal for metallurgical data.

Finally, the model's interpretability was enhanced by analysing feature importance. Two key methods were used: Gini Importance, which measures the total decrease in node impurity contributed by each feature, and Permutation Importance, which evaluates the drop in model accuracy when a feature's values are randomly shuffled. These insights were crucial in identifying which alloying elements and processing parameters most significantly affect sigma phase formation, offering valuable guidance for alloy design and optimization.

4.1.1 Data Collection

The excel sheet contained 579 rows and 11 features(columns) - type of steel, onset time, temperature, volume fraction of sigma phase, percentage of composition of all the alloying elements in the steel. These datapoints were collected manually from the research papers.

For Example:

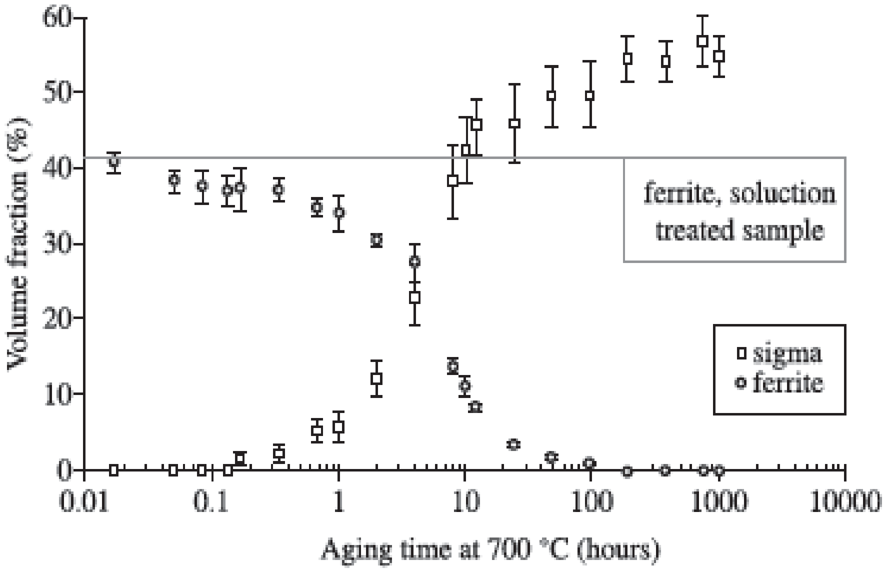
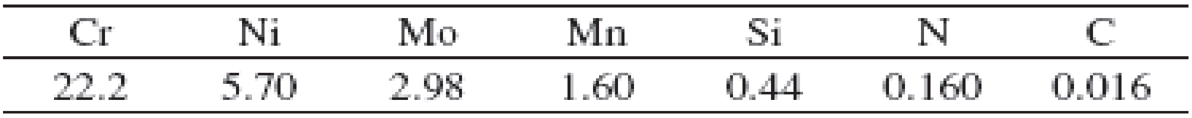


Figure 3. Ferrite and sigma phase content (vol. %) as a function of aging time at 700 °C.

Fig 4.1.2: Volume Fraction of Sigma Phase Vs Aging Time in DSS

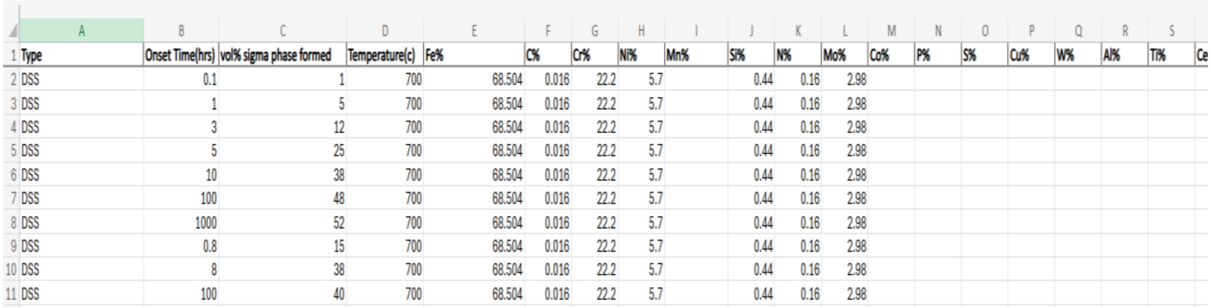
The above figure is a graph from the research paper – Kinetics of Sigma Phase Formation in a Duplex Stainless Steels from Materials Research – Schielo Brazil. It gives the relationship between Volume Fraction of Sigma Phase formed and Aging time of Duplex Stainless Steel (DSS). To obtain exact values, Web plot Digitizer (an online website to obtain accurate data points from given graph) was used. The composition also was given the same research paper. The values were entered in the Excel sheet.

Table 1. Chemical composition of the investigated steel (wt.(%))



Cr	Ni	Mo	Mn	Si	N	C
22.2	5.70	2.98	1.60	0.44	0.160	0.016

Fig 4.1.3: Chemical composition of DSS from Schielo Brazil research paper



Type	Onset Time(hrs)	vol% sigma phase formed	Temperature(c)	Fe%	C%	Cr%	Ni%	Mn%	Si%	N%	Mo%	Co%	P%	S%	Cu%	W%	Al%	Ti%	Ce
DSS	0.1	1	700	68.504	0.016	22.2	5.7		0.44	0.16	2.98								
DSS	1	5	700	68.504	0.016	22.2	5.7		0.44	0.16	2.98								
DSS	3	12	700	68.504	0.016	22.2	5.7		0.44	0.16	2.98								
DSS	5	25	700	68.504	0.016	22.2	5.7		0.44	0.16	2.98								
DSS	10	38	700	68.504	0.016	22.2	5.7		0.44	0.16	2.98								
DSS	100	48	700	68.504	0.016	22.2	5.7		0.44	0.16	2.98								
DSS	1000	52	700	68.504	0.016	22.2	5.7		0.44	0.16	2.98								
DSS	0.8	15	700	68.504	0.016	22.2	5.7		0.44	0.16	2.98								
DSS	8	38	700	68.504	0.016	22.2	5.7		0.44	0.16	2.98								
DSS	100	40	700	68.504	0.016	22.2	5.7		0.44	0.16	2.98								

Fig 4.1.4: Screenshot of our features in Excel sheet

4.1.2 Data Review

Reviewed all the research papers available related to Prediction of sigma phase in different types of steels like HDSS, SDSS, Austenitic Stainless Steels investigated the composition, onset time and volume fraction and data points were collected. The excel sheet contained features - Type of steel, Onset time, Temperature, Composition of respective steels

4.1.3 Data Preprocessing

Data Shuffling

Shuffling the data means randomly rearranging the rows of your dataset before splitting it into training and testing sets. It is done to remove bias. If your original data is sorted or grouped in some way (e.g., all class 0 samples first, then class 1), not shuffling it can lead to:

- The training set learning only from certain types of examples
- The testing set containing completely different patterns — leading to poor generalization

Handling Missing Data

Median Imputation & Literature Values

Missing data points were first addressed by using the median method, ensuring that the central tendency of each feature was maintained. In cases where literature values were available and deemed reliable, these were used to substitute missing entries. This approach provided a robust method to handle missing values without introducing significant bias.

Replacement for Specific Elements

For elements phosphorus (P) and sulphur (S), the values were replaced with 0.05

because general composition of P and S in steels is 0.05%. For major elements like Fe%, Cr%, Ni%, Mn%, Si% values were replaced with 0. Any null values that remained after the initial imputation (when the proportion of missing data exceeded 5%) were replaced with their respective median values. This ensured consistency across the dataset and preserved the underlying distribution of these elements.

Checking for Duplicates

Duplicates Identification and Removal

The dataset was scanned for duplicate entries using Python scripts. Any duplicates identified were promptly removed to avoid skewing the analysis. This step was critical in ensuring the integrity of the dataset and preventing redundant information from affecting the machine learning model's training process.

Normalization Techniques

Min-Max Scaling

Two types of min-max scaling were employed:

Horizontal Scaling (Across Elements)

For each individual element, min-max scaling was performed to bring all values within a consistent range. This process ensures that differences in magnitude among elements do not disproportionately influence the model.

Vertical Scaling (Uniform Distribution Across Columns)

Additionally, vertical scaling was applied to ensure that the distribution of values across all columns is uniform. This step further refines the dataset, promoting balance and comparability across different features

4.1.4 Data Visualization

Data Representation

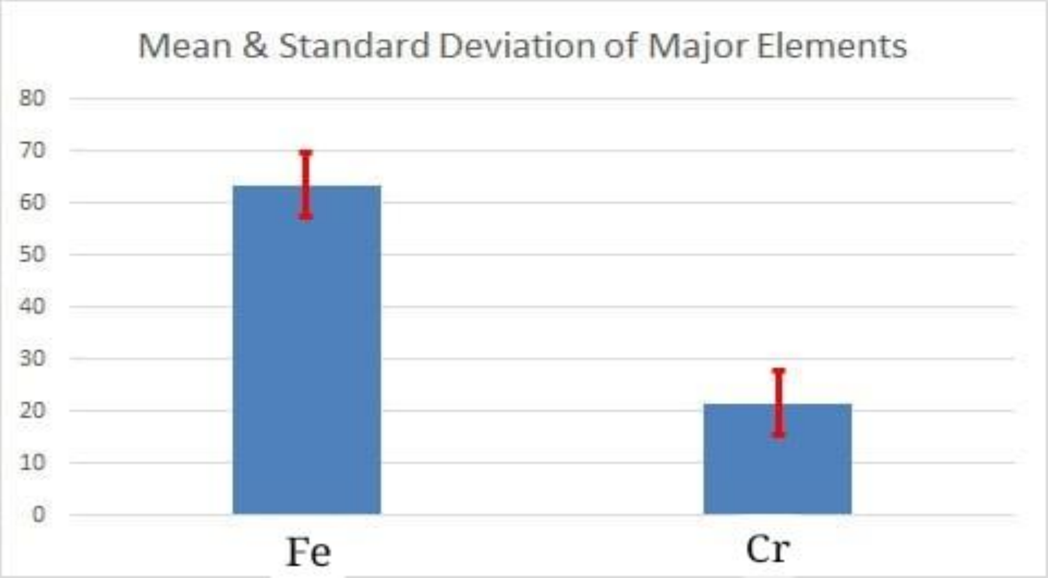


Fig. 4.1.5: Mean and standard deviation of major elements

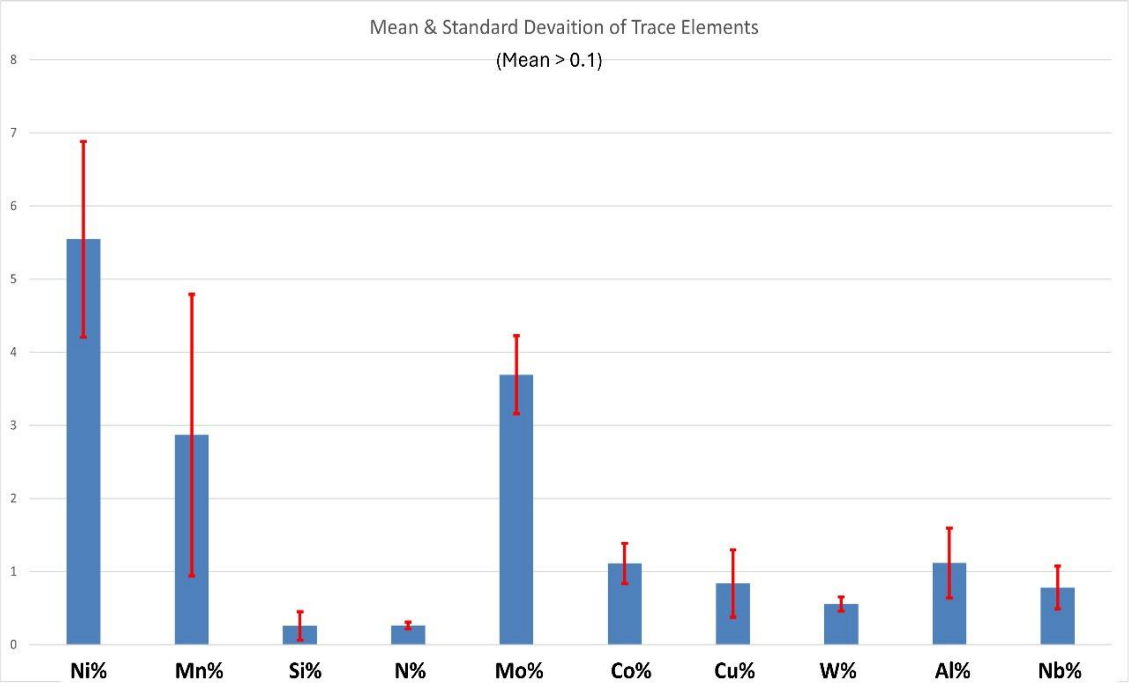


Fig. 4.1.6: Mean and standard deviation of trace elements (mean>0.1)

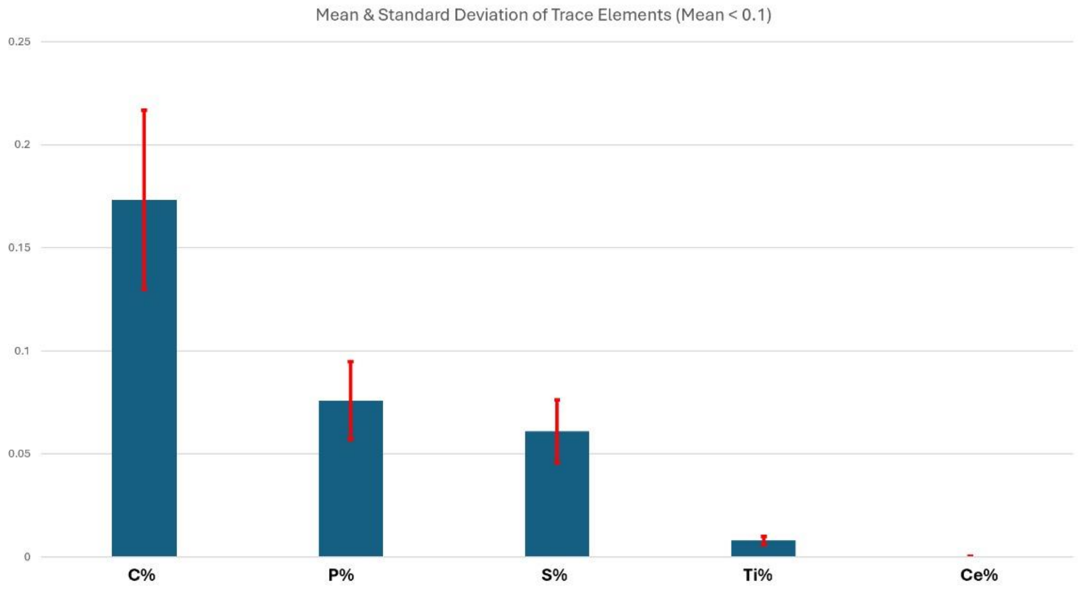


Fig 4.1.7: Mean and standard deviation of trace elements (mean <0.1)

4.1.5 Checking for patterns in the data: Exploratory data analysis

Correlation matrix

Objective

To understand the relationships between various features, we generate a correlation matrix. This matrix visually represents the Pearson correlation coefficients between pairs of variables, enabling us to quickly identify strong linear relationships and potential multicollinearity issues.

Implementation

The correlation matrix is computed using Python libraries (such as pandas and seaborn). A heatmap is then generated to provide an intuitive, color-coded view of these correlations. High positive or negative values are easily identifiable, which helps in determining which variables might significantly influence sigma phase formation.

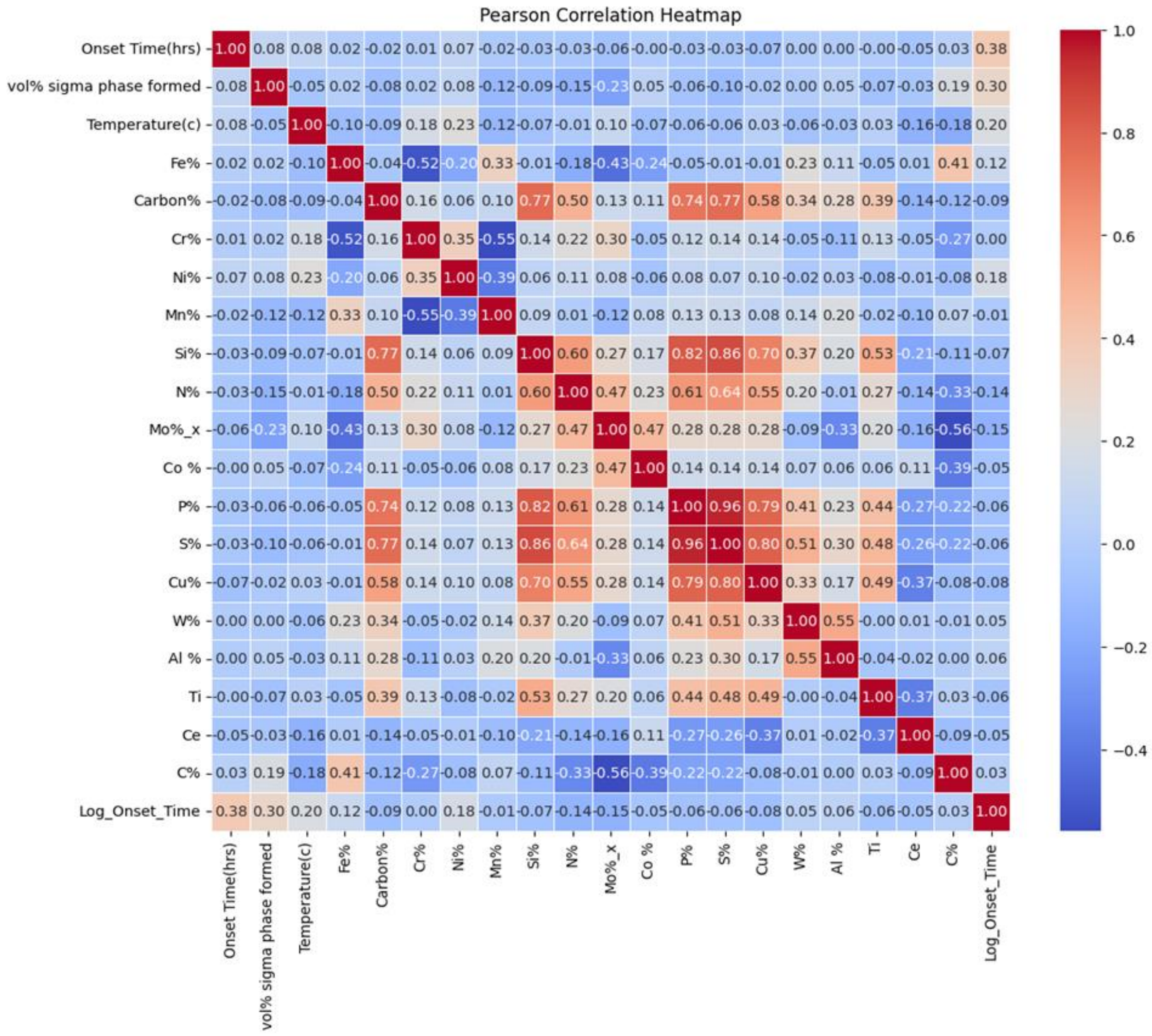


Fig 4.1.8: Correlation Matrix

Correlation Matrix Analysis

There is not much corelation between the features from the dataset we collected, according to the corelation matrix. So, Feature Transformation is done as there is no univariate and multivariate relationship between the features.

Spearmen Correlation

Definition

The Spearman rank correlation coefficient, often denoted by ρ (rho) or rs , is a nonparametric measure of rank correlation. It assesses the strength and direction of the monotonic relationship between two ranked variables. In simpler terms, it tells you how well the relationship between two variables can be described using a consistently increasing or consistently decreasing function, without necessarily being a straight line.

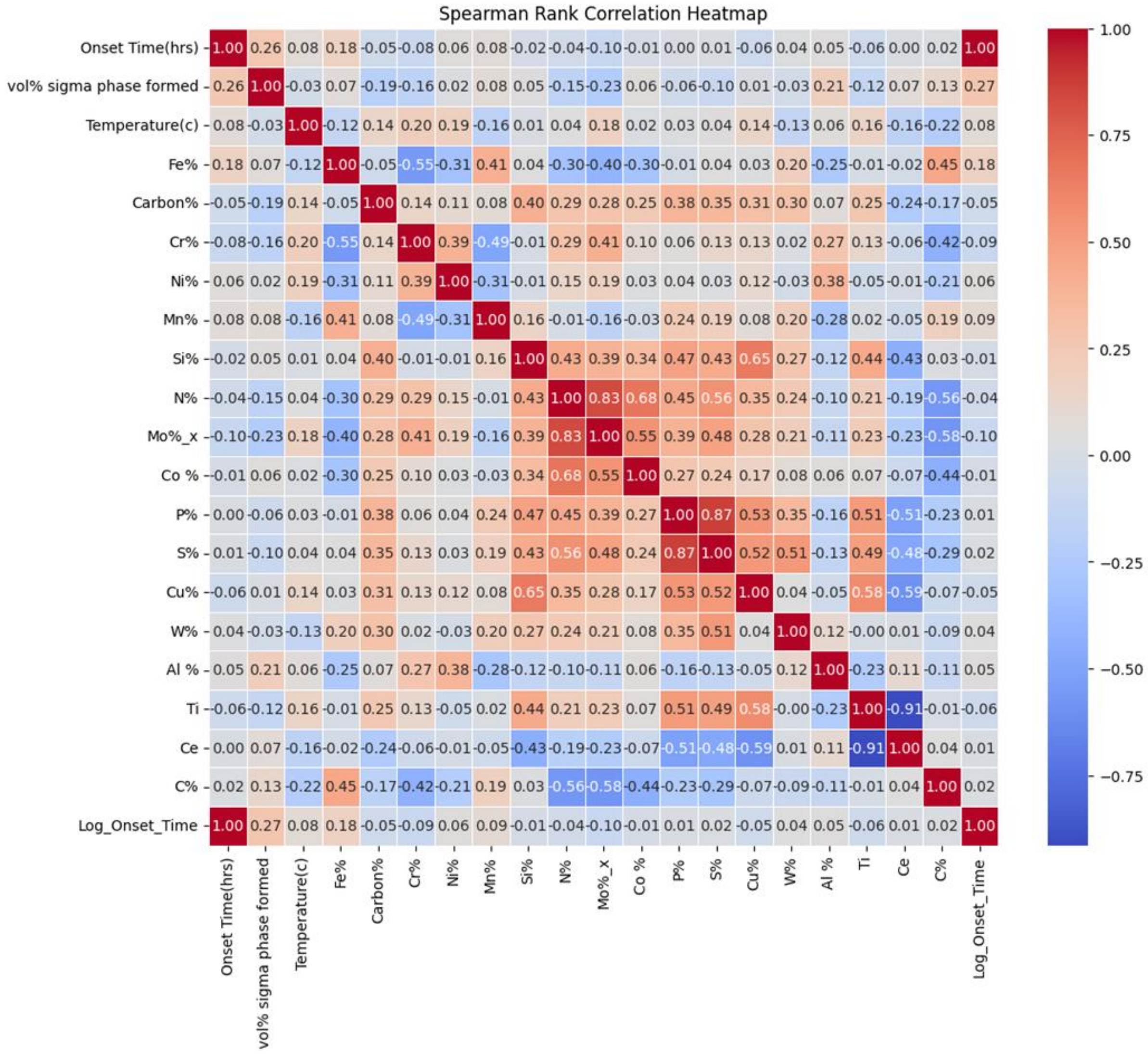


Fig 4.1.9: Spearman correlation

Analysis

Have checked for monotonic univarite relationships between **onset time** and other **feature**

F-statistic and mutual information scores

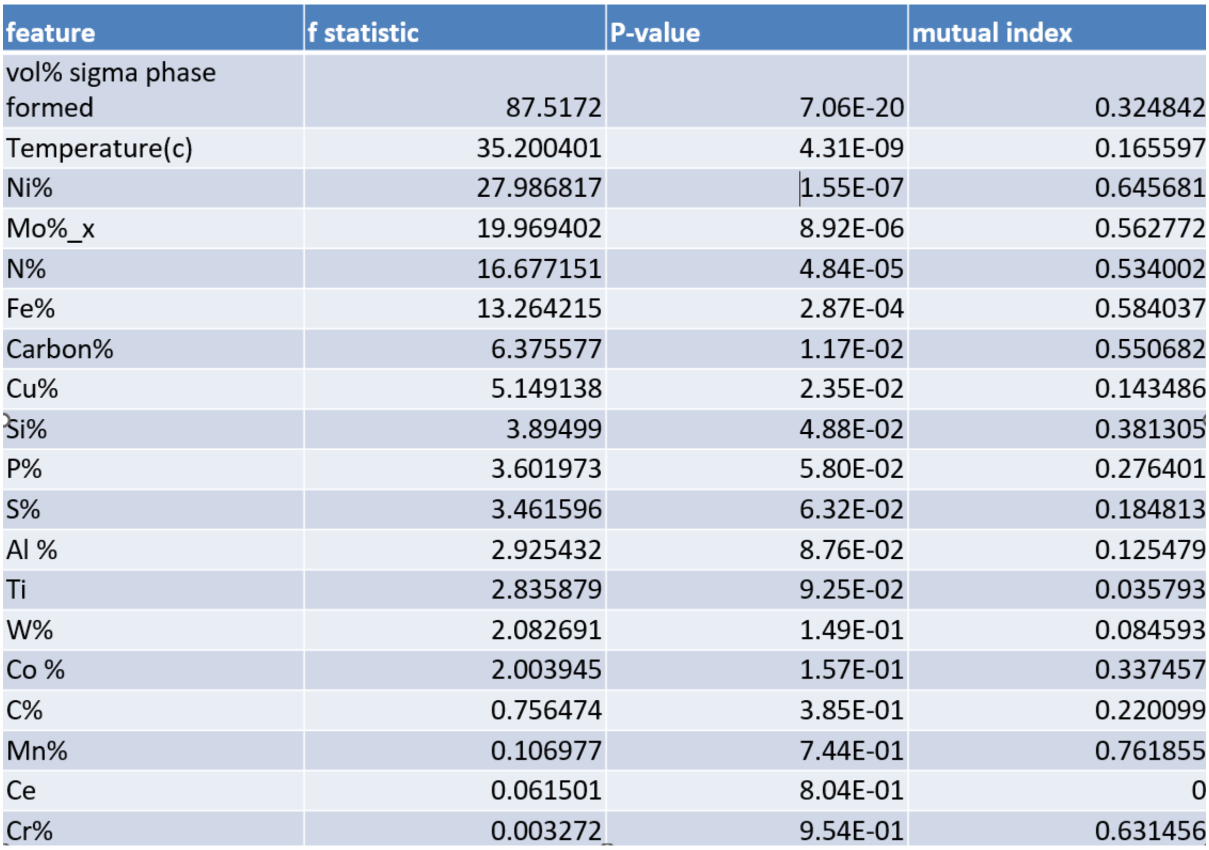
Definition

F-statistic: Higher values suggest stronger predictive influence.

P-value: Indicates statistical significance (lower is better; < 0.05 is significant).

Mutual Index: Measures the dependency between feature and target (1 = perfect correlation, 0 = none).

Table



feature	f statistic	P-value	mutual index
vol% sigma phase formed	87.5172	7.06E-20	0.324842
Temperature(c)	35.200401	4.31E-09	0.165597
Ni%	27.986817	1.55E-07	0.645681
Mo%_x	19.969402	8.92E-06	0.562772
N%	16.677151	4.84E-05	0.534002
Fe%	13.264215	2.87E-04	0.584037
Carbon%	6.375577	1.17E-02	0.550682
Cu%	5.149138	2.35E-02	0.143486
Si%	3.89499	4.88E-02	0.381305
P%	3.601973	5.80E-02	0.276401
S%	3.461596	6.32E-02	0.184813
Al %	2.925432	8.76E-02	0.125479
Ti	2.835879	9.25E-02	0.035793
W%	2.082691	1.49E-01	0.084593
Co %	2.003945	1.57E-01	0.337457
C%	0.756474	3.85E-01	0.220099
Mn%	0.106977	7.44E-01	0.761855
Ce	0.061501	8.04E-01	0
Cr%	0.003272	9.54E-01	0.631456

Fig 4.1.10: Table of statistic analytics

Analysis

To check for univariate and multivariate relationships in data

1. Primary candidates ($p < 0.05$ and high mutual info):
   a. Ni %, Mo %, N %, Fe %, C % (Carbon), vol % σ-phase, Temperature.
2. Secondary candidates ($p < 0.05$ but lower mutual info):
   a. Cu %, Si %.,p%
3. Optional (mutual info high but non-significant p):
   a. Mn %, Ce — consider including if you suspect non-linear relationships.

4.1.6 Employing different machine learning models and checking metrics

We have employed different machine learning methods with standard features as in experimental findings for training (composition, vol%, temp) target variable: log(onsetime)
Test train split= 20-80

Model	R^2 Score	MSE	MAE	Loss
Random Forest	0.443210050	3.164368	–	–
XGBoost	0.420433	3.741231	–	–
AdaBoost	0.386406	–	3.4871966	–
RNN	0.358	–	–	3.6469
LSTM	0.321126	–	–	3.858196
SVM	0.28423615	–	4.06785	–

Table.4.1.11: Performance of different model on experimental findings

All models are explored in basic model ipynb file in code folder

Observations

- **Small Dataset Size:** The limited amount of collected data is insufficient for the models to learn complex patterns effectively, leading to poor generalization.
- **Constant Sigma Phase Volume Percentage:** The typically unchanging volume percentage of the sigma phase in the collected data provides little variance for the model to learn from.
- **Discrete Temperature Data:** The discrete nature of the recorded temperature values might not capture the continuous variations crucial for accurately predicting the sigma phase formation.
- **Compressed Onset Time Data (due to outliers):** Outliers in the onset time data, indicating potential sigma phase formation even at very slow cooling rates, compress

the usable range of this feature and obscure the typical formation behaviour.

4.2 Model to predict formation enthalpy

4.2.1 Density Functional Theory (DFT)

Density Functional Theory (DFT) reformulates the many-electron Schrödinger equation in terms of the electron density, enabling tractable quantum-mechanical simulations of materials [27, 28]. In the Kohn–Sham approach, a system of interacting electrons is mapped onto non-interacting electrons moving in an effective potential, which includes Hartree, external, and exchange–correlation terms [27]. DFT excels at predicting bonding behavior, electronic density distributions, phase stability, total energies, and magnetic or mechanical properties of solids [28].

Density Functional Theory (DFT) was used to compute thermodynamic properties of elemental configurations, which influence sigma phase stability. These calculations provide atomistic insights into phase stability.

4.2.2 Relation to σ-Phase Formation

The σ phase in stainless steels has a complex tetragonal ($D8_6$/tp30) structure with five inequivalent atomic sites and often partial site occupancy by Fe, Cr, and Mo. DFT calculations provide formation enthalpies and site-preference probabilities for these sites, revealing which elements stabilize the σ structure and by how much [24, 29]. Such atomistic insights are critical because experimental determination of σ-phase energetics is costly and time-consuming.

4.2.3 Classical Nucleation Theory and Kinetics

While DFT yields thermodynamic driving forces, the kinetics of σ-phase precipitation are governed by nucleation and growth laws. Classical Nucleation Theory (CNT) describes the rate of nucleus formation as a function of supersaturation, interfacial energy, and temperature, leading to an Arrhenius-type dependence [26] Growth kinetics follow the Johnson–Mehl–Avrami–Kolmogorov (JMAK) equation

$$X(t)=1-\exp(-ktn)$$

where *X(t)* is the fraction transformed, *k* is a rate constant, *t* is time, and *n* (the Avrami exponent) reflects nucleation and growth mechanisms [3]. By coupling DFT-computed formation energies with CNT, one can derive kinetic descriptors (e.g., critical nucleation time) that bridge thermodynamics and experimental onset times.

4.2.4 "Wacky-Off" Position Concept of σ Phase

The σ-phase unit cell contains 30 atoms distributed over five inequivalent sites with mixed occupancy, leading to what is colloquially termed "wacky-off" positions [23]. DFT studies show that differences in site coordination (12-fold vs. 14-fold) and local electronic environments drive element-specific preferences, accounting for the structural complexity and site disorder observed experimentally.

4.2.5 The Materials Project Database

The Materials Project provides standardized, high-throughput DFT data—including crystal structures, formation energies, and phase diagrams—for thousands of materials via an open-access API [2]. Its curated σ-phase entries allow rapid mining of formation enthalpies and lattice parameters, underpinning our feature-engineering workflow.

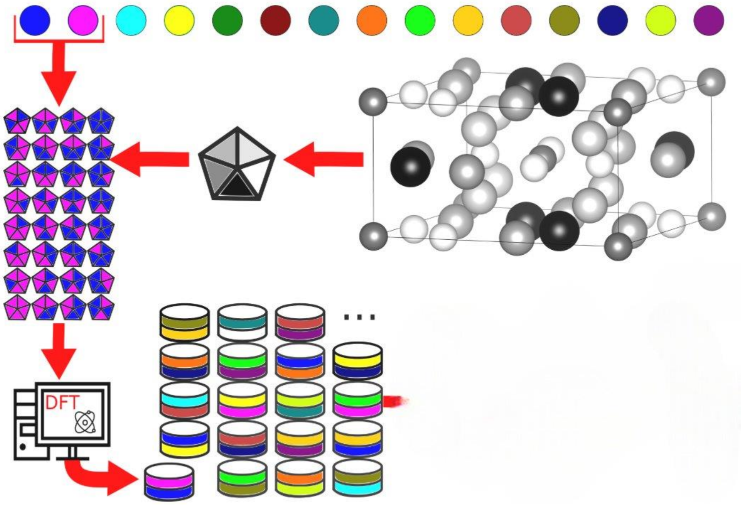
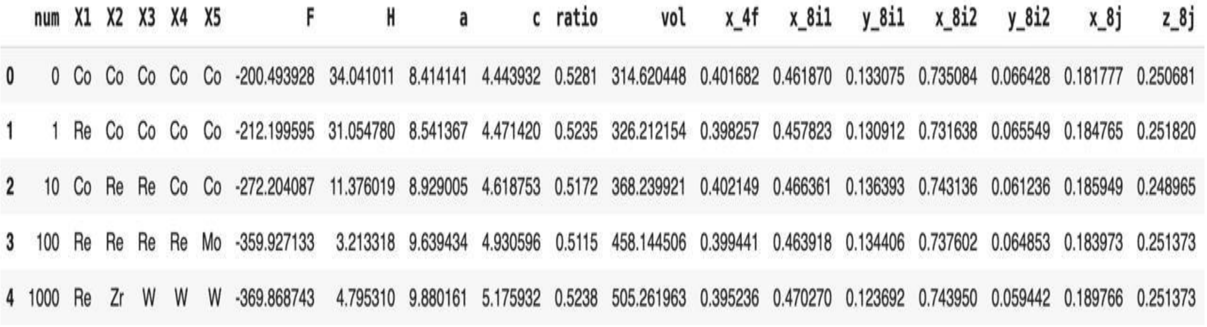


Fig.4.2.5. DFT first principle properties calculation schema

4.2.6 DFT dataset collection

A data set comprising of 10000 configurations of sigma phase are collected



	num	X1	X2	X3	X4	X5	F	H	a	c	ratio	vol	x_4f	x_8i1	y_8i1	x_8i2	y_8i2	x_8j	z_8j
0	0	Co	Co	Co	Co	Co	-200.493928	34.041011	8.414141	4.443932	0.5281	314.620448	0.401682	0.461870	0.133075	0.735084	0.066428	0.181777	0.250681
1	1	Re	Co	Co	Co	Co	-212.199595	31.054780	8.541367	4.471420	0.5235	326.212154	0.398257	0.457823	0.130912	0.731638	0.065549	0.184765	0.251820
2	10	Co	Re	Re	Co	Co	-272.204087	11.376019	8.929005	4.618753	0.5172	368.239921	0.402149	0.466361	0.136393	0.743136	0.061236	0.185949	0.248965
3	100	Re	Re	Re	Re	Mo	-359.927133	3.213318	9.639434	4.930596	0.5115	458.144506	0.399441	0.463918	0.134406	0.737602	0.064853	0.183973	0.251373
4	1000	Re	Zr	W	W	W	-369.868743	4.795310	9.880161	5.175932	0.5238	505.261963	0.395236	0.470270	0.123692	0.743950	0.059442	0.189766	0.251373

Fig 4.2.6: Data from Materials Project Website

4.2.7 Energy model training and metrics

On the dft dataset a random forest model is trained and have obtain a precise model to predict Energy out of sigma phase configurations with r2 score = 0.97

Train-test data split: 80-20

Train features are one hot encoder for (X1,X2,X3,X4,X5) and few elements' specifics like Valence electrons and atomoic radii and standard formation energies are used which are Custom written and are in sigma.py module of the code folder

Target; H(formation enthalpy)

Metrics obtained

On cross validation

Cross validation scores are

Cross-validation R^2 scores:

[0.90096151, 0.89972938 ,0.96015003 ,0.82694666 ,0.76201336]

Average R^2 score: 0.8499601893957065

Cross-validation MAE scores:

[-3.31479355, -2.18764919, -1.14691227, -2.82781502,-6.40688961]

Average MAE score: 3.176811927430072

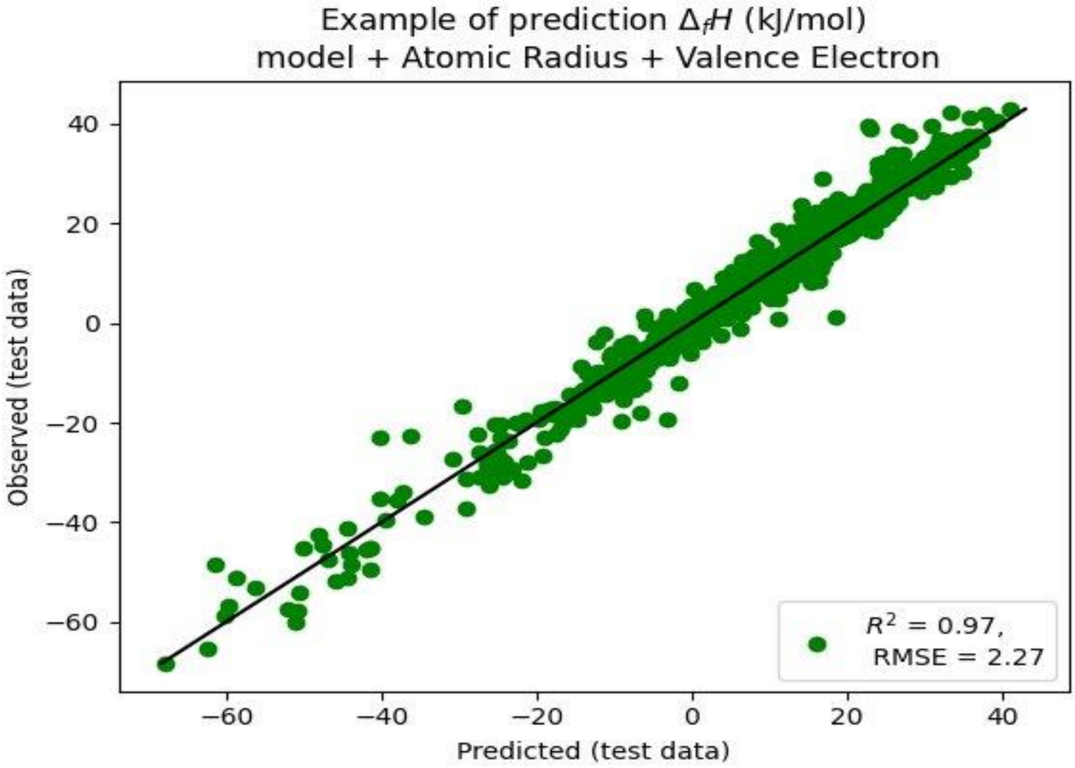
On testing data

Mean Absolute Error (MAE): 1.385315428795058

Mean Squared Error(MSE):5.142937893226127

R-squared (R2): 0.973285379816469

Fig



4.2.7: Regression plot of energy model

4.2.8 Discussion

- **High Accuracy & Efficiency:** The machine learning model accurately predicts sigma phase formation enthalpy (H) with DFT-level precision (test $R^2 = 0.97$), offering significant computational speed advantages over direct DFT calculations.
- **Enables Further Analysis:** This computational efficiency is crucial, enabling the rapid calculation of energies for numerous configurations required in subsequent project steps.
- **Thermodynamic Input:** The model provides essential thermodynamic data (formation enthalpy) needed to bridge atomistic configurations with kinetic models like CNT.
- **Foundation for Feature Engineering:** This energy model underpins the creation of the theoretically informed e_tot feature used in the final model predicting sigma phase onset time.
- **Validation & Scope:** While highly accurate, the model relies on the Materials Project dataset, and cross-validation suggests robust but potentially varied performance across different structural configurations.

4.3 Sampling Configurations from Composition

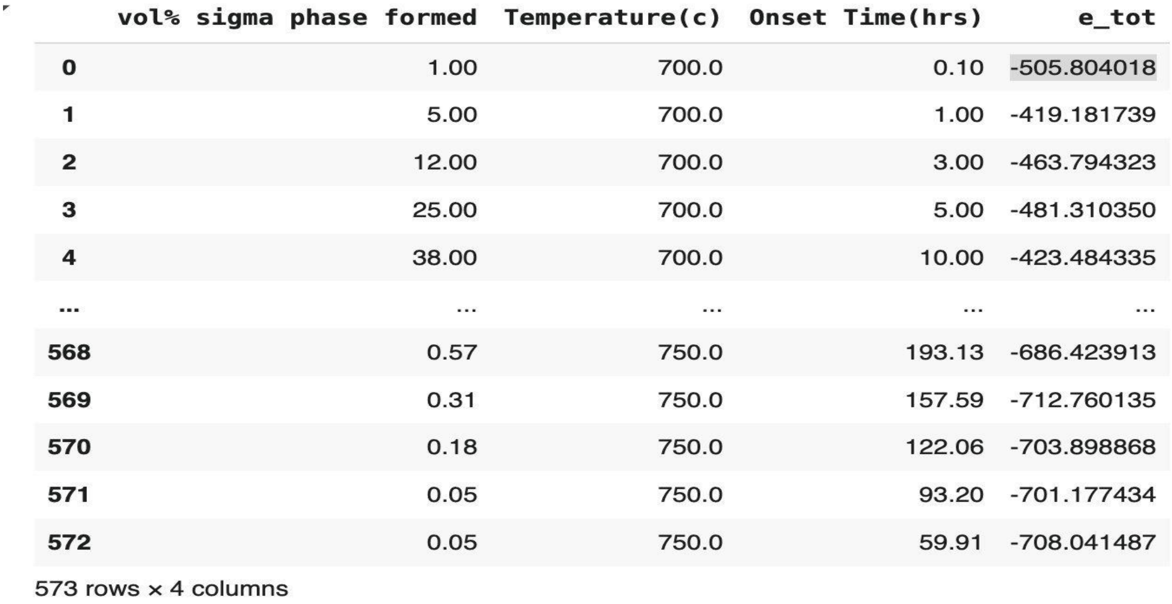
4.3.1 Site_Sampler module

The feature importance of this energy model can be directly interpreted as a site preference probabilities of different elements in crystal structure of sigma phase thus writing a probability distribution sampling assuming normalized composition percentages as probabilities we have written a site_sampler module that uses this site preference dictionary and site_sampler utility functions are combinedly used for sampling possible configuration of elemental arrangements for a given composition . We have typically extracted 100 samples of configuration from each composition with a basic idea of exploring patterns that are missed in experimental findings dataset due to sparse data

4.3.2 Computation of energy model predictions

These samples from single compositions are indexed with alloy id and are fed to energy model and a total energy i.e sum of all predictions of samples under same alloy id are summed up and are reported as e_tot

The prediction made are in joules so are changed into ev with standard joule to ev converion ratios



	vol% sigma phase formed	Temperature(c)	Onset Time(hrs)	e_tot
0	1.00	700.0	0.10	-505.804018
1	5.00	700.0	1.00	-419.181739
2	12.00	700.0	3.00	-463.794323
3	25.00	700.0	5.00	-481.310350
4	38.00	700.0	10.00	-423.484335
...	...	...	...	...
568	0.57	750.0	193.13	-686.423913
569	0.31	750.0	157.59	-712.760135
570	0.18	750.0	122.06	-703.898868
571	0.05	750.0	93.20	-701.177434
572	0.05	750.0	59.91	-708.041487

573 rows × 4 columns

Fig 4.3.2: Overview of features (e-tot assigned to experimental findings dataset)

4.3.3 Validation of introduced new feature

Thus, the new feature is totally computed from other model and is from other source of data typically is all theoretical hence there is a need to check if there are any patterns that we can recognize from this new _feature

Hence have plot correlation matrices and scatterplots to check for so among thermodynamic features in all of data

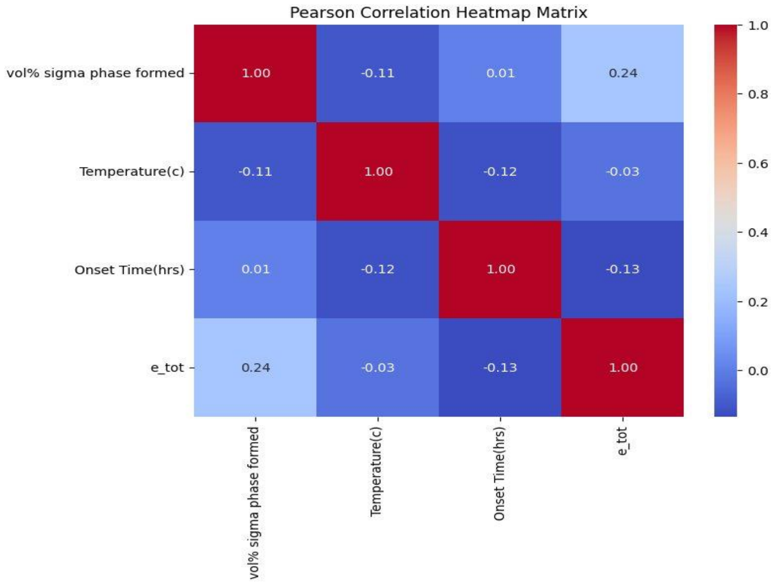


Fig.4.3.3.a :Pearson correlation matrix between all thermodynamic features

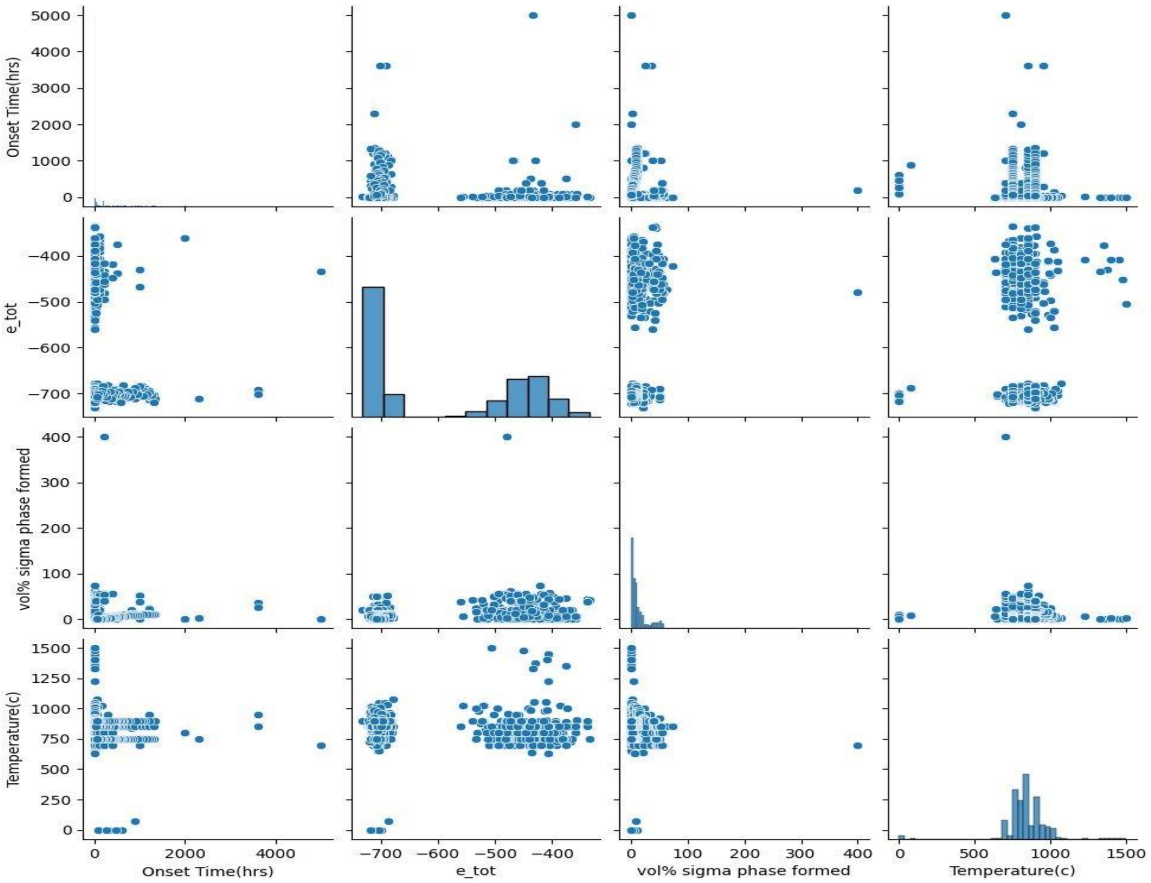


Fig.4.3.3.b: Pairplot between thermodynamic features and e_tot

Surprisingly ,we have found there is relation between this new computed feature to our target variable and is more than vol% sigma phase relation with onset time and also during data collection since the sigma phase growth is Arrhenius type the data collected is for around formation of volume percent of one and even less to also decrease he model dependence more on e_tot we have used Avrami equation for sigma phase from Classical Nucleation Theory the used both energy and volume percent to predict onset time and Avrami exponent with is a material dependent factor on which the growth kinetics of sigma depend

4.4 Ensemble model

4.4.1 Ensemble Machine Learning Model

Ensembling Machine learning models using the technique of *Stacking*.

Stacking is a way to ensemble multiple classifications or regression model. There are many ways to ensemble models, the widely known techniques are Bagging or Boosting. Bagging allows multiple similar models with high variance which are averaged to decrease variance. Boosting builds multiple incremental models to decrease the bias, while keeping variance small.Stacking is a different paradigm. The point of stacking is to explore a space of different models for the same problem. The idea is that you can attack a learning problem with different types of models which can learn some part of the problem, but not the whole space of the problem. So, you can build multiple different learners, and you use them to build an intermediate prediction, one prediction for each learned model. Then you add a new model which learns from the intermediate predictions for the same target. This final model is said to be stacked on top of the others, hence the name. Thus, you might improve your overall performance, and often you end up with a model which is better than any individual intermediate model.

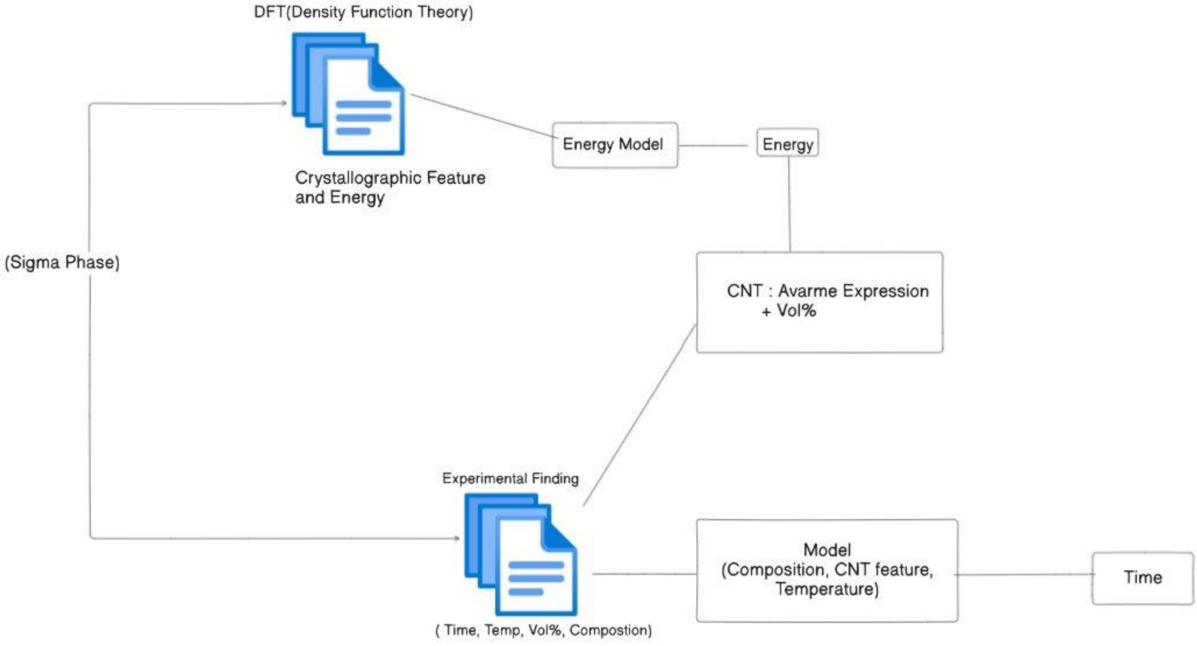
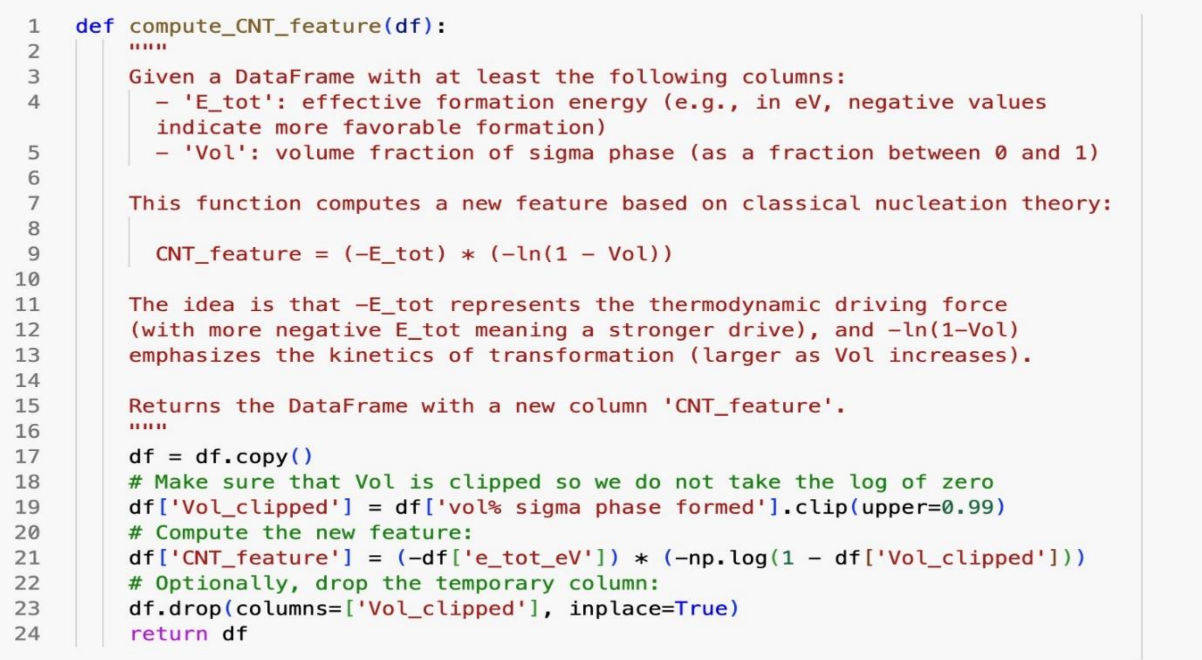

DFT(Density Function Theory)
Crystallographic Feature and Energy
Energy Model
Energy
(Sigma Phase)
CNT : Avarme Expression + Vol%
Experimental Finding
(Time, Temp, Vol%, Compostion)
Model (Composition, CNT feature, Temperature)
Time

Fig 4.4.1: Proposed methodology (stacking model)

4.4.2 CNT FEATURE: FEATURE ENGINEERING

To ensure there is only learning through feature introduction but introduction of direct e_tot_ev and using it directly may affect out results also decreasing the effect temperature and biasing the model towards e_tot completely so we have made a transformation to existing feature that is unable to learn CCT plot patterns hence we used classical nucleation theory that bridge gap between thermodynamic and kinetic features hence following avrami equation



```python
def compute_CNT_feature(df):
    """
    Given a DataFrame with at least the following columns:
      - 'E_tot': effective formation energy (e.g., in eV, negative values
        indicate more favorable formation)
      - 'Vol': volume fraction of sigma phase (as a fraction between 0 and 1)

    This function computes a new feature based on classical nucleation theory:

      CNT_feature = (-E_tot) * (-ln(1 - Vol))

    The idea is that -E_tot represents the thermodynamic driving force
    (with more negative E_tot meaning a stronger drive), and -ln(1-Vol)
    emphasizes the kinetics of transformation (larger as Vol increases).

    Returns the DataFrame with a new column 'CNT_feature'.
    """
    df = df.copy()
    # Make sure that Vol is clipped so we do not take the log of zero
    df['Vol_clipped'] = df['vol% sigma phase formed'].clip(upper=0.99)
    # Compute the new feature:
    df['CNT_feature'] = (-df['e_tot_eV']) * (-np.log(1 - df['Vol_clipped']))
    # Optionally, drop the temporary column:
    df.drop(columns=['Vol_clipped'], inplace=True)
    return df
```

Fig.4.4.2a:Sanpshot of cnt_feature code

e_tot_ev and vol% are combined to one feature and have checked for patterns

Cnt feature is found relating better with significant elemental features from earlier models hence have used cnt_feature as our new feature for model training.

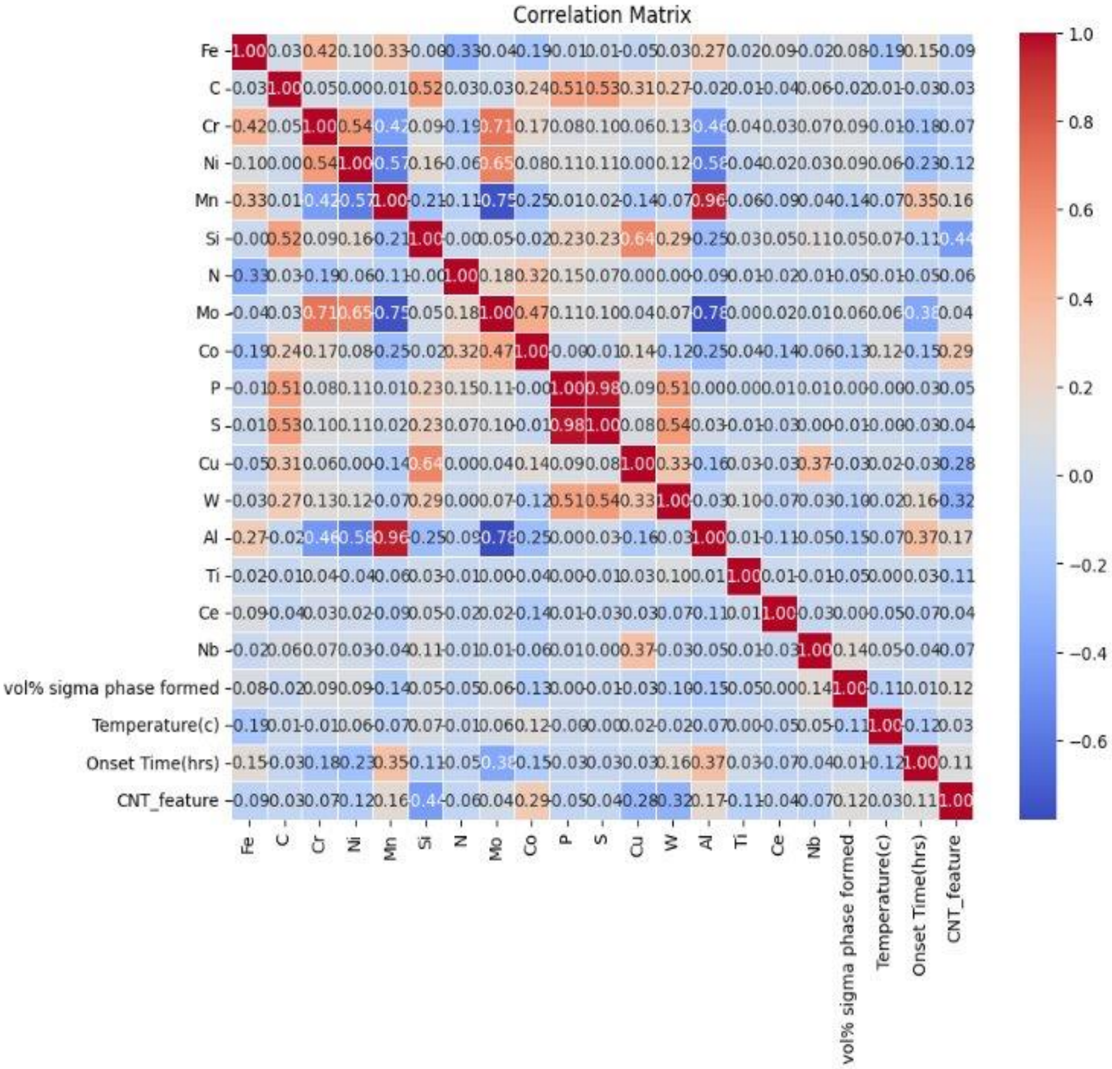


Fig.4.4.2b: Correaltion with CNT_feature

4.4.3 Model Training

- Training to test split is 80 and 20 percent
- Modal is trained on {temperature,CNT_feature,composition}
- Target variable: log(onset time)
- Cross validation method :5 fold cross validation
- Cross-Validation Scores: [0.72116802 0.67107471 0.66290697 0.83495547 0.69980387]
- Mean Cross-Validation Score: 0.7179818081095698

4.4.4 Model configuration

RandomForestRegressor

(n_estimators: int = 100, *, criterion: str = "squared_error", max_depth: Unknown | None = None, min_samples_split: int = 2, min_samples_leaf: int = 1, min_weight_fraction_leaf: float = 0, max_features: float = 1, max_leaf_nodes: Unknown | None = None, min_impurity_decrease: float = 0, bootstrap: bool = True, oob_score: bool = False, n_jobs: Unknown | None = None, random_state: Unknown | None = None, verbose: int = 0, warm_start: bool = False, ccp_alpha: float = 0, max_samples: Unknown | None = None, monotonic_cst: Unknown | None = None)

Further hyperparametertuning of the above configuration may increase model accuracy and enhance learning

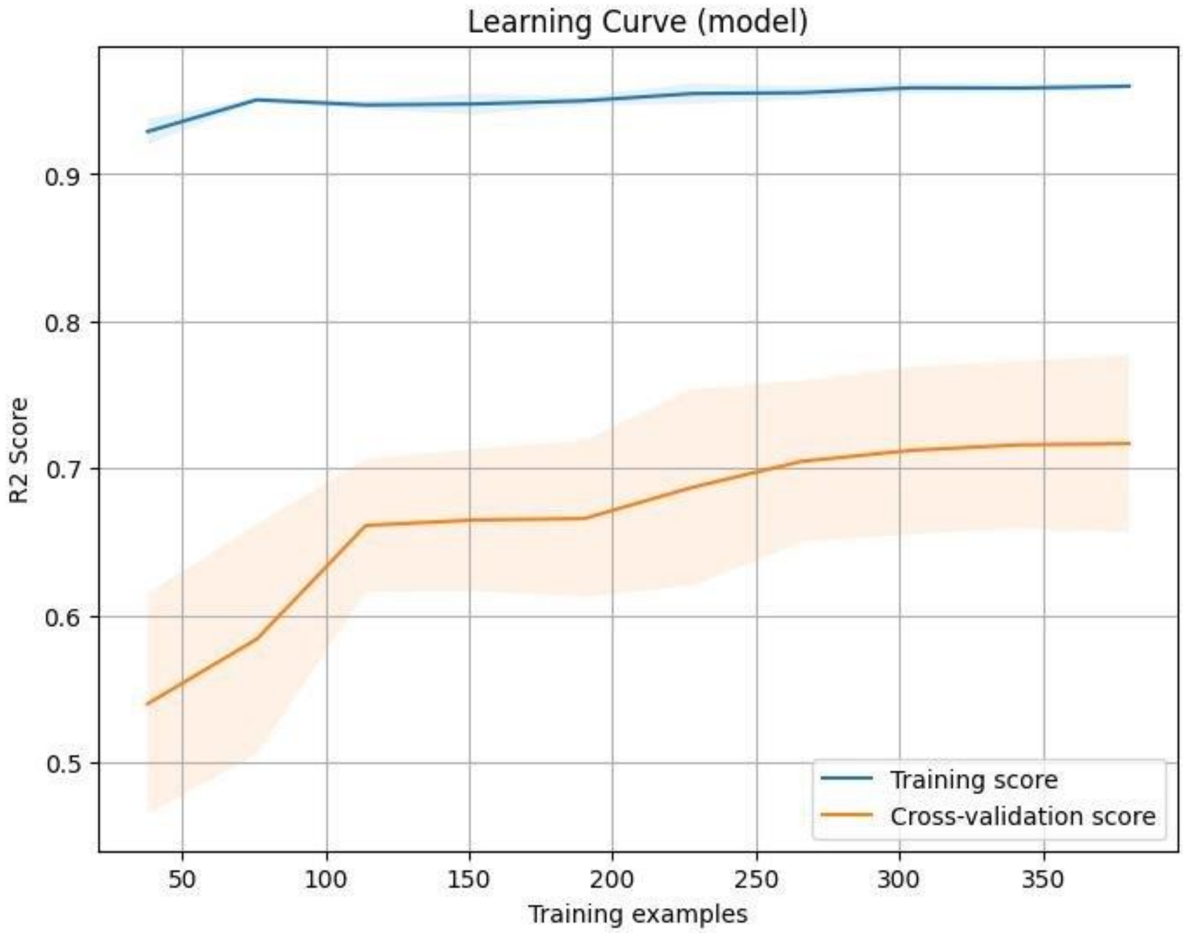


fig 4.4.4 shows Learning curve of the final machine learning model

Chapter 5

Results and Discussion

5.1 Machine Learning Model Performance

- Random Forest Regression is used to predict sigma phase formation.
- R² score = 0.64, Root Mean Square Error = 1.86 indicate model reliability and predictive accuracy.

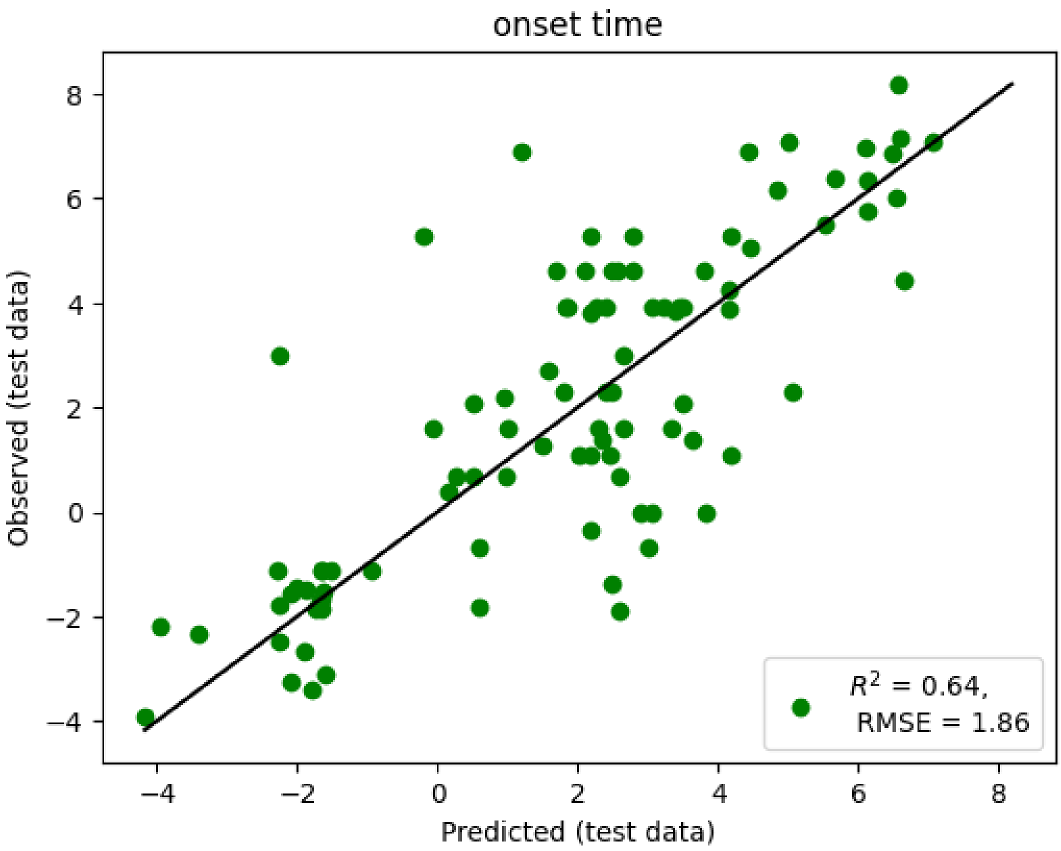


Fig 5.1: Regression plot of final machine learning model

5.2 Feature Importance Analysis

- Major Influential Elements: Molybdenum and Aluminium had the highest impact on sigma phase formation.
- Trace Elements Influence: Silicon, Nickel (Ni), W,C,N,Cu show modrate contribution to target variable {log(onset time)}
- Temperature Dependency: High temperatures accelerated sigma phase formation, confirming Arrhenius-type behaviour.

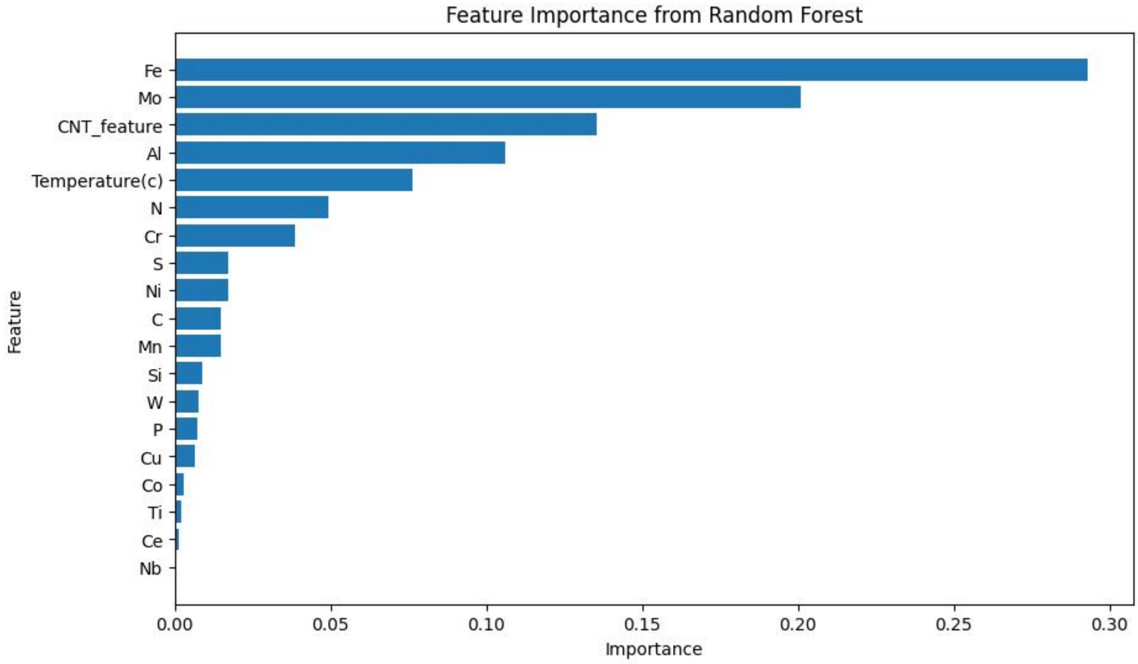


Fig 5.2: Feature Importance of elements

5.3 Partial dependence plots Insight

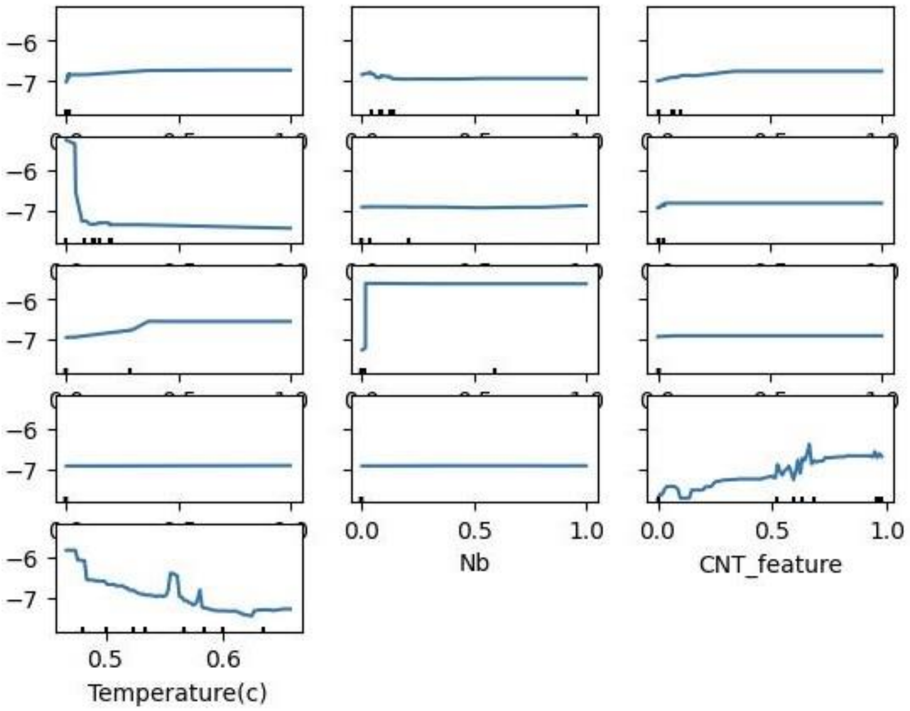


Fig 5.3a Plot of partial dependence plots for C, Mn, Si, Mo, Co, P, W, Al, Ti, Ce, Nb, CNT_feature, Temperature

Analysis

Table.5.3b.interpreting partial dependence plots

Element	PDP Trend	Interpreted Effect	Metallurgical Context & Citations to support
C	Sharp negative dip	Strong promoter	Carbon ties up Cr in carbides, locally depleting Cr and accelerating σ nucleation [5]
Mn	Slight negative trend	Weak promoter (mild)	Mn can increase diffusion rates slightly, nudging σ formation [30].
Si	Negative slope	Promoter	Si raises driving force for σ by altering electronic structure[43].

Mo	Negative slope	Strong promoter	Mo strongly stabilizes σ, lowering its formation barrier .[29]
Co	Flat	Neutral	Co neither significantly alters σ thermodynamics nor kinetics .[24]
P	Slight negative dip	Weak promoter (possible)	P segregates to defects, providing nucleation sites for σ .[31]
W	Negative slope	Promoter	W behaves similarly to Mo in strengthening σ driving force .[33]
Al	Slight positive slope	Inhibitor	Al forms alumina or G-phase, reducing available solutes for σ .[36]
Ti	Flat-ish	Likely neutral	Ti precipitates as carbides/nitrides, with little σ influence .[28]
Ce	Flat	Neutral or sparse data	Trace Ce typically too low to affect σ.[25]
Nb	Small dip then flat	Mild promoter	Nb can substitute in σ sites, slightly easing nucleation.[34]
Temperature	Flat mostly, some wiggle	No clear direct trend	T–dependence is captured in the CNT feature and Avrami exponent rather than linearly here .[3]
CNT_feature	Slight positive slope	inhibitor	Higher CNT values imply larger nucleation barriers, modestly delaying σ onset.[26]

5.4 SHAP VALUES

To interpret our stacking model's predictions at the granular level, we employ **SHAP (Shapley Additive explanations)**, which attributes to each feature the change in the expected model output when that feature is observed rather than unknown [34]. Figure 5.4 presents a SHAP summary plot: features are ordered by their overall importance (mean |SHAP|), and each point shows how a single observation's feature value (color) shifts the prediction (SHAP value) left or right.

• **Fe wt %** emerges as the most influential feature: higher Fe content typically **increases** the predicted log(onset time) (positive SHAP), indicating that Fe-rich alloys are slower to form σ phase, consistent with Fe's role in diluting Cr/Mo driving forces [23].

• **Mo wt %** is the second most important, but in contrast, **higher Mo** drives **lower** predicted log(onset time) (negative SHAP), confirming Mo's strong promotion of σ-phase nucleation .[29]

• **Al wt %** ranks third: **high Al** raises predicted log(onset time), reflecting its inhibitory effect—Al preferentially forms oxides or G-phase, reducing solute availability for σ.[36]

• The **CNT_feature** (our DFT + CNT kinetic descriptor) exhibits a mixed effect: **low CNT** values strongly decrease log(onset time) (faster σ onset), whereas **high CNT** modestly delay it, illustrating the nonlinear thermodynamic–kinetic coupling captured by our feature .[26]

• **Temperature** also shows a clear trend: **higher temperatures** yield negative SHAP contributions (accelerated σ formation), aligning with Arrhenius-type kinetics.[3]

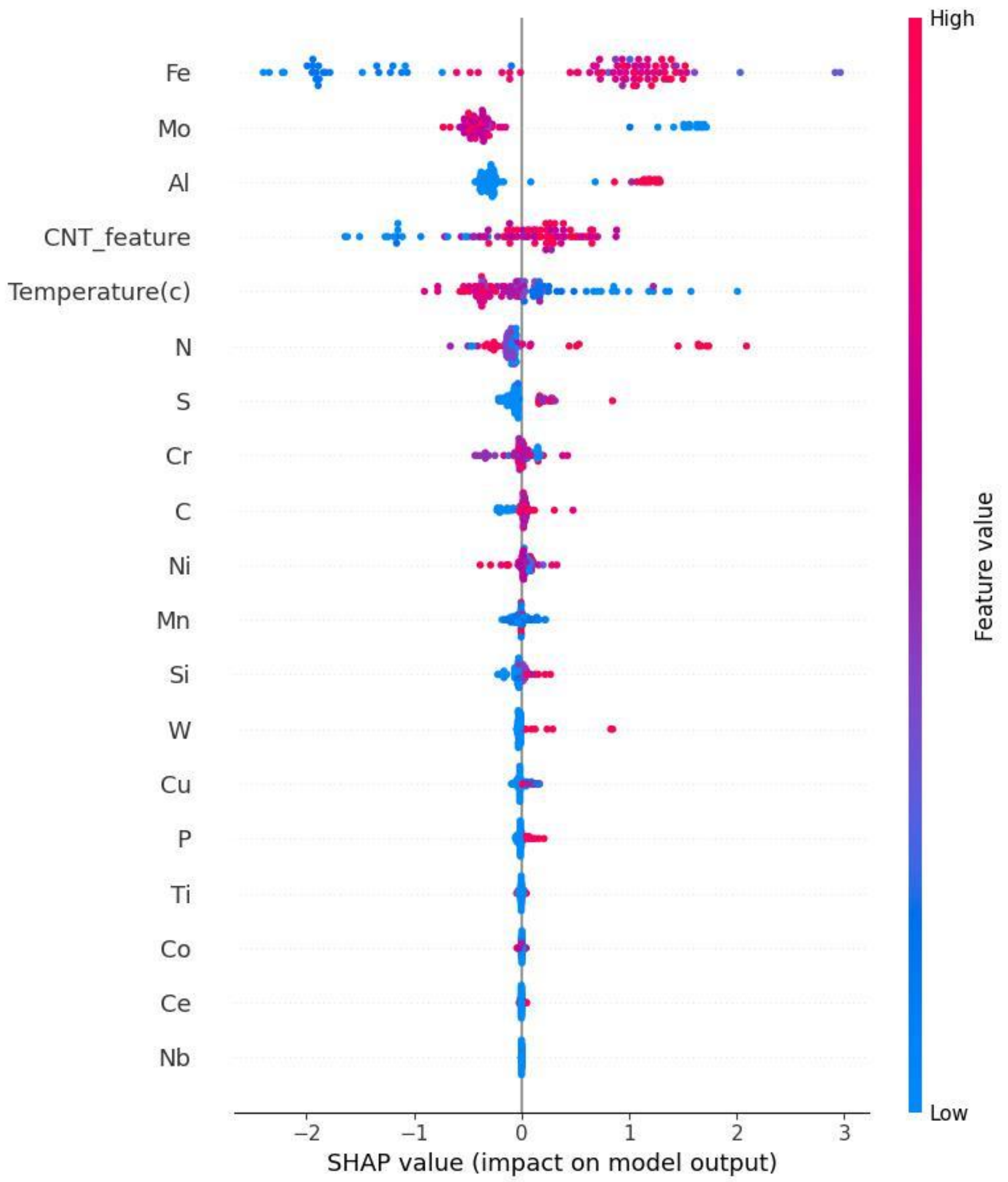


Fig.5.4:SHAP values of final regression model

Chapter 6

Conclusions

6.1 Summary of Findings

- The machine learning model (Random Forest Regression) developed successfully predicted the logarithm of the sigma phase onset time with reasonable reliability (R^2 score = 0.64, RMSE = 1.86).
- Feature importance analysis (including SHAP values) identified the most influential elements affecting sigma phase formation:
  - Iron (Fe) content was the most influential overall, with higher Fe % generally increasing the onset time (inhibiting effect).
  - Molybdenum (Mo) strongly promotes sigma phase formation (decreases onset time).
  - Aluminium (Al) inhibits sigma phase formation (increases onset time).
  - Other elements showed varying degrees of influence: Silicon (Si), Carbon (C), Tungsten (W), Phosphorus (P), Manganese (Mn), and Niobium (Nb) generally act as promoters, while Nickel (Ni), Copper (Cu), Cobalt (Co), Titanium (Ti), and Cerium (Ce) showed moderate, neutral, or less clear effects based on the analysis.

- Higher temperatures were confirmed to accelerate sigma phase formation, consistent with an Arrhenius-type relationship.
- The engineered feature based on Classical Nucleation Theory (CNT_feature), integrating DFT energy predictions and volume fraction, demonstrated a complex relationship with onset time, capturing thermodynamic-kinetic coupling.

6.2 Key Contributions

- **Integration of Machine Learning and Metallurgy:** Demonstrated a successful data-driven approach using machine learning to predict sigma phase formation in stainless steels.
- **Enhanced Predictive Modelling:** Improved prediction accuracy by combining

machine learning with fundamental kinetic models (CNT) and integrating diverse data sources (experimental findings, DFT calculations via Materials Project).

- **Advanced Feature Engineering:** Developed and validated feature engineering techniques (like the CNT_feature) based on theoretical metallurgical principles to bridge thermodynamic and kinetic factors.
- **Methodological Guidance:** Provided a practical example and guide for implementing machine learning models on experimental metallurgical data, including techniques for enriching model learning by incorporating theoretical data.
- **Practical Application Insights:** Generated insights valuable for optimizing steel compositions and heat treatments to minimize sigma phase-related failures in industrial applications.
- **Foundation for Future Work:** Contributed to the documented understanding of sigma phase formation across various steel types and highlighted pathways for future improvement, such as expanding the dataset and exploring further model ensembling for predicting full CCT diagrams.

6.3 Conclusions

This project successfully developed and validated a machine learning framework, integrating experimental data with theoretical insights from DFT and CNT, to predict sigma phase onset time in stainless steels. Key elemental influences (promotion by Mo, inhibition by Al) and the accelerating effect of temperature were quantified. The work provides a valuable tool and methodology for materials scientists and engineers aiming to design more robust stainless steel alloys and optimize processing parameters to avoid detrimental sigma phase formation.

References

1. *Huang X, Wang H, Xue W, Ullah A, et al. A combined machine learning model for the prediction of TTT diagrams of high-alloy steels. J. Alloys Compd. 823:153694; 2020.[1]*
2. *Crivello J-C, Sokolovska N, Joubert J-M, et al. Supervised deep learning prediction of the formation enthalpy of complex phases: the σ phase as an example. Comput. Mater. Sci. 193:110340; 2021.[2]*
3. *Avrami M. Kinetics of Phase Change I: General Theory. J. Chem. Phys. 7(12):1103–1112; 1939.[3].*
4. *Materials Project. Sigma-phase data. Materials Project. Accessed 2025.[4]*
5. *Pimenta FC Jr. Sigma phase precipitation in duplex stainless steels. Mater. Sci. Forum 426:1319–1324; 2003. [5]*
6. Akosta Y, Olsson J, Edenhammar C. TTT and CCT diagrams of stainless steels. Mater. Sci. Technol. 22(12):1457–1464; 2006. [6]
7. *Zehnder AS, Mohammad A. Application of classical nucleation theory to phase transformations. Acta Mater. 61(4):1245–1251; 2013.[7]*
8. *Smith P, Brown J. Integration of DFT and classical models for kinetic predictions.[8]*
9. *Agrawal, A., & Choudhary, A. (2016). Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science. APL Materials, 4(5), 053208.[9]*
10. *Agren, J., et al. (2012). Computational thermodynamics and kinetics in materials design. Calphad, 36, 1-10.[10]*
11. *Callister, W. D., & Rethwisch, D. G. (2020). Materials Science and Engineering: An Introduction (10th ed.). Wiley.[11]*
12. *Cieslak, J., et al. (1986). Sigma phase formation in duplex stainless steels. Metallurgical Transactions A, 17(12), 2101-2111.[12]*
13. *DeCost, B. L., et al. (2017). Machine learning in materials science: A review. JOM, 69(12), 2101-2115.[13]*
14. *Gunn, R. N. (2003). Duplex Stainless Steels: Microstructure, Properties, and Applications. Woodhead Publishing.[14]*
15. *Hsieh, C. C., & Wu, W. (2019). Overview of intermetallic sigma phase precipitation*

in stainless steels. ISRN Metallurgy, 2012, 1-16.[15]

16. *Lippold, J. C., & Kotecki, D. J. (2005). Welding Metallurgy and Weldability of Stainless Steels. Wiley.[16]*
17. *Michalska, J., & Sozańska, M. (2006). Effect of minor elements on sigma phase precipitation in duplex stainless steels. Materials Characterization, 56(4-5), 355-362.[17]*
18. *Pavlina, E. J., et al. (2015). Sigma phase embrittlement in duplex stainless steels. Journal of Materials Engineering and Performance, 24(5), 2101-2110.[18]*
19. *Porter, D. A., et al. (2009). Phase Transformations in Metals and Alloys (3rd ed.). CRC Press.[19].*
20. *Sahu, J. K., et al. (2019). Sigma phase precipitation in stainless steels: A review. Journal of Materials Science, 54(1), 1-25.[20]*
21. *Sedriks, A. J. (1996). Corrosion of Stainless Steels (2nd ed.). Wiley[21].*
22. *Ward, L., et al. (2016). Machine learning for materials discovery. Nature Reviews Materials, 1(12), 16028.[22]*

23. *Berne, C., & Sluiter, M. H. F. (2014). Site occupancy in the Re–W σ phase: a first-principles study. Journal of Alloys and Compounds, 578, 123–131.[23]*
24. *Hu, Q.-M., Boulet, P., Record, M.-C., et al. (2019). Atomic bonding and electronic stability of the binary σ phase. Computational Materials Science, In Press.[24]*
25. *Jain, A., Ong, S. P., Hautier, G., et al. (2013). Commentary: The Materials Project: a materials genome approach to accelerating materials innovation. APL Materials, 1(1), 011002.[25]*
26. *Johnson, W. A., & Mehl, R. F. (1939). Reaction kinetics in processes of nucleation and growth. Transactions of the American Institute of Mining, Metallurgical and Petroleum Engineers, 135, 416–442.[26]*
27. *Kohn, W., & Sham, L. J. (1965). Self-Consistent Equations Including Exchange and Correlation Effects. Physical Review, 140(4A), A1133–A1138.[27]*
28. *Parr, R. G., & Yang, W. (1989). Density-Functional Theory of Atoms and Molecules. Oxford University Press.[28]*
29. *Sieurin, H., & Sandström, R. (2006). Nucleation and diffusional growth of σ phase in duplex stainless steel 2205. Materials Science and Engineering A, 444(1), 271–279.[29]*
30. *Dubiel, S.M. & Sluiter, M.H.F. (2014). J. Alloys Compd. 578:123–131.[30]*

31. Hsieh, C.-C., et al. (2012). Int. Scholarly Res. Notices.[31]

32. Tsai, M.-H., et al. (2016). Mater. Res. Lett. 4(2):90–95.[32]

33. Dubiel SM. Site occupancy and enthalpy of formation of σ phase in Fe–Cr–Mo alloys. Mater. Res. Lett. 22(3):191–196; 2011.[33]

34. Kim W, Lee J, Park S. Modeling TTT diagrams of stainless-steel using machine learning. J. Alloys Compd. 845:156251; 2020[34]

35. E. O. Hall, International Materials Reviews, 11 (2013) 61.[35]

36. R. C. Reed, Springer Nature, 30 (1999) 521.[36]