

**Comparative Analysis and Optimization of Machine Learning  
Algorithms for Prediction of Adsorption Energy of Methane  
Related Species on Cu-based Alloys**



By

Haseeb Ahmad Khan Akhunzada

(Registration No: 00000329273)

Department of Chemical Engineering  
School of Chemical and Materials Engineering  
National University of Sciences & Technology (NUST)  
Islamabad, Pakistan

(2024)

**Comparative Analysis and Optimization of Machine Learning  
Algorithms for Prediction of Adsorption Energy of Methane  
Related Species on Cu-based Alloys**



By

Haseeb Ahmad Khan Akhunzada

(Registration No: 00000329273)

A thesis submitted to the National University of Sciences and Technology, Islamabad,

in partial fulfillment of the requirements for the degree of

Master of Science in  
Process Systems Engineering

Supervisor: Dr. Iftikhar Ahmad

School of Chemical and Materials Engineering

National University of Sciences & Technology (NUST)

Islamabad, Pakistan

(2024)




**THESIS ACCEPTANCE CERTIFICATE**

Certified that final copy of MS thesis written by Mr Haseeb Ahmad Khan Akhonzada (Registration No 00000329273), of School of Chemical & Materials Engineering (SCME) has been vetted by undersigned, found complete in all respects as per NUST Statues/Regulations, is free of plagiarism, errors, and mistakes and is accepted as partial fulfillment for award of MS degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis.


Signature: 

Name of Supervisor: Dr Iftikhar Ahmad

Date: 13-09-2024

Signature (HOD): 

Date: 13-09-2024

Signature (Dean/Principal): 

Date: 13/9/24

National University of Sciences & Technology (NUST)

## MASTER'S THESIS WORK

Formulation of Guidance and Examination Committee (GEC)

Name: Haseeb ahmad khan Akhuzada NUST Reg No: 00000329273  
 Department: Department of Chemical Engineering Specialization: Master of Science in Process System Engineering  
 Credit Hour Completed: 24.0 CGPA: 3.19

## Course Work Completed

S/No:	Code:	Title:	Core/Elective:	CH:	Grade:
1.	PSE-801	Process Systems Theory	Compulsory	3.0	B
2.	PSE-852	Process Modelling and Simulation	Compulsory	3.0	B
3.	TEE-820	Process Intensification	Compulsory	3.0	B+
4.	PSE-802	Optimization and Decision Analysis	Compulsory	3.0	C+
5.	PSE-823	Advanced Process Dynamics and Control	Compulsory	3.0	B
6.	CSE-801	Computational Fluid Dynamics	Elective	3.0	B
7.	ENE-809	Waste Water Treatment & Design	Elective	3.0	B+
8.	CHE-814	Product Technology	Elective	3.0	B+
9.	PSE-802	Optimization and Decision Analysis	Repeat	3.0	B
10.	RM-898	Research Methodology	Additional	2.0	Q

Date 21 - Aug - 2023

Student's Signature

## Thesis Committee

- Name: Iftikhar Ahmad (Supervisor)  
Department: Department of Chemical Engineering
- Name: Erum Pervaiz (Internal)  
Department: Department of Chemical Engineering
- Name: Zakir Hussain (Internal)  
Department: Department of Materials Engineering

Signature

Signature

Signature

Date: 21 - Aug - 2023

Signature of Head of Department:

APPROVAL

Signature of Dean/Principal:

Date: 21 - Aug - 2023

School of Chemical &amp; Materials Engineering (SCME) (SCME) H-12 Campus,



Form: TH-04

National University of Sciences & Technology (NUST)

MASTER'S THESIS WORK

We hereby recommend that the dissertation prepared under our supervision by

Regn No & Name: 00000329273 Haseeb Ahmad Khan Akhunzada

Title: Comparative Analysis and Optimization of Machine Learning Algorithms for prediction of adsorption energy of Methane Related Specie of Cu-based Alloys.

Presented on: 22 Aug 2024 at: 1500 hrs in SCME

Be accepted in partial fulfillment of the requirements for the award of Master of Science degree in Process Systems Engineering.

Guidance & Examination Committee Members

Name: Dr Zakir Hussain

Signature: [Signature]

Name: Dr Erum Pervaiz

Signature: [Signature]

Supervisor's Name: Dr Iftikhar Ahmad

Signature: [Signature]

Dated: \_\_\_\_\_

[Signature]  
Head of Department

Date 5/9/24

[Signature]  
Dean/Principal

Date 5/9/24

School of Chemical & Materials Engineering (SCME)

## **AUTHOR'S DECLARATION**

I **Haseeb Ahmad Khan Akhunzada** hereby state that my MS thesis titled “**Comparative Analysis and Optimization of Machine Learning Algorithms for prediction of adsorption energy of Methane Related Species on Cu-based Alloys**” is my own work and has not been submitted previously by me for taking any degree from National University of Sciences and Technology, Islamabad or anywhere else in the country/ world.

At any time, if my statement is found to be incorrect even after I graduate, the university has the right to withdraw my MS degree.

Name of Student: Haseeb Ahmad Khan Akhunzada

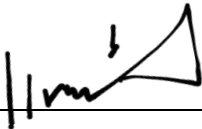
Date: 16<sup>th</sup> Sep 2024

## PLAGIARISM UNDERTAKING

I solemnly declare that the research work presented in the thesis titled “**Comparative Analysis and Optimization of Machine Learning Algorithms for prediction of adsorption energy of Methane Related Species on Cu-based Alloys**” is solely my research work with no significant contribution from any other person. Small contribution/ help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero-tolerance policy of the HEC and the National University of Sciences and Technology (NUST), Islamabad towards plagiarism. Therefore, I as an author of the above-titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/cited.

I undertake that if I am found guilty of any formal plagiarism in the above-titled thesis even after the award of MS degree, the University reserves the right to withdraw/revoke my MS degree and that HEC and NUST, Islamabad have the right to publish my name on the HEC/University website on which names of students are placed who submitted plagiarized thesis.

Student Signature:  \_\_\_\_\_

Name: Haseeb Ahmad Khan Akhunzada

## **DEDICATION**

*I dedicate this thesis to my parents, who have provided unwavering support, motivation, and love throughout my academic journey.*



## **ACKNOWLEDGEMENTS**

I extend my most profound appreciation to the Almighty Allah for providing me with the strength, guidance, and knowledge to complete this thesis. Furthermore, I'm eternally grateful for the boundless resources and wisdom bestowed upon me.

I want to express my sincere gratitude to my research supervisor, Dr. Iftikhar Ahmad, for their unwavering support, guidance, and compassionate counsel throughout my research journey. I am also grateful to my thesis committee members, Dr. Nouman Ahmed and Dr. Erum Pervaiz, for their insightful recommendations and valuable contributions.

I am also grateful to the School of Chemical and Materials Engineering leadership, Dr. Erum Pervaiz, for fostering a research-oriented environment that allowed me to develop my skills and complete this work entirely.

Last, I want to express my most profound appreciation to my parents, family, and friends for their constant support and encouragement throughout my academic journey. Without their love and support, this accomplishment would not have been possible.

Haseeb Ahmad Khan Akhunzada

# TABLE OF CONTENTS

ACKNOWLEDGEMENTS.....	ix
TABLE OF CONTENTS.....	x
LIST OF TABLES.....	xii
LIST OF FIGURES.....	xiii
LIST OF SYMBOLS, ABBREVIATIONS AND ACRONYMS.....	xiv
ABSTRACT.....	xv
CHAPTER 1: INTRODUCTION.....	1
1.1 Background.....	1
1.2 Objectives.....	6
1.3 Thesis outlines.....	7
CHAPTER 2: LITERATURE REVIEW.....	8
2.1 Literature review.....	8
CHAPTER 3: METHODOLOGY AND MODELLING OF ALOGRITHMS.....	13
3.1 Methodology.....	13
3.2 Dataset and existing research.....	15
3.2.1 Data visualization.....	15
3.2.2: Machine learning algorithms.....	17
3.2.3: Gradient Boosting Algorithms.....	17
3.2.4: Light Gradient Boosting Model.....	20
3.3 Genetic Algorithm.....	21
3.3.1 Population.....	21
3.3.2 Determination of Parents and Offspring.....	22
3.3.3 Crossover.....	24
3.3.4 Mutation.....	26

3.4	Modelling Parameters and Hyperparameter tuning .....	26
CHAPTER 04	RESULTS AND DISCUSSION.....	29
4.1	Models evaluation and comparison .....	29
CHAPTER 5:	CONCLUSIONS AND RECOMMENDATIONS.....	41
REFERENCES.....		42

## LIST OF TABLES

<b>Table 3.1:</b> Input features (descriptors) used for prediction of adsorption energies.....	15
<b>Table 3.2:</b> Chromosomes.....	21
<b>Table 3.3:</b> Hyperparameters utilized in the various boosting models.....	27
<b>Table 3.4:</b> Genetic algorithm parameters used to optimize the hyperparameters of Catboost Model.....	28
<b>Table 4.1:</b> Comparison of ML methods .....	29

## LIST OF FIGURES

<b>Figure 1.1:</b> Schematic overview of oxidative coupling of methane (OCM) process.[12].....	2
<b>Figure 1.2:</b> Schematic illustration of active site renewal on the oxide surface through replacing the oxygen vacancy formed by methane activation.....	4
<b>Figure 1.3:</b> Impact of surface reducibility on the activity of the catalyst and it impacts on C2 yield as well as showing the increased selectivity of Cox [12] .....	5
<b>Figure 3.1:</b> Methodology .....	13
<b>Figure 3.2:</b> Boxplot of the feature variables .....	17
<b>Figure 3.3:</b> Schematic representation of Genetic Algorithm .....	23
<b>Figure 3.4:</b> Roulette wheel selection.....	24
<b>Figure 3.5:</b> Single point crossover .....	25
<b>Figure 3.6:</b> Double points crossover .....	25
<b>Figure 3.7:</b> Uniform crossover.....	25
<b>Figure 3.8:</b> After the crossover phase, the mutation operator changes one or more genes in the children's solutions.....	26
<b>Figure 4.1:</b> DFT-calculated adsorption energies of CH <sub>3</sub> on Cu-based alloys and their predicted values (a) Catboost (b) LightGBM (c) XGBoost.....	30
<b>Figure 4.2:</b> Permutation importance scores of the feature variables.....	31
<b>Figure 4.3:</b> Partial Dependence Plots of the feature variables to the prediction of the adsorption energies .....	38
<b>Figure 4.4:</b> SHAP beeswarm summary plot.....	40

## LIST OF SYMBOLS, ABBREVIATIONS AND ACRONYMS

$Y_i$	Predicted value
$Y_i^{\text{exp}}$	Actual value
AI	Artificial intelligence
GA	Genetic algorithm
XGBoost	Extreme Gradient Boosting
LGBM	Light Gradient Boosting Model
Catboost	Categorical Boosting
LM	Levenberg-Marquardt (an optimization algorithm)
ML	Machine learning
CC	Correlation coefficient
RMSE	Root means squared error
AN	Atomic number
G	Group
R	Atomic Radius
AM	Atomic Mass
EN	Electronegativity
IE	Ionization Energy
SE	Surface Energy

## ABSTRACT

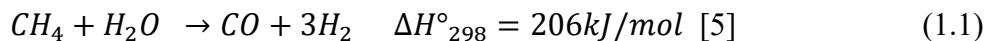
Demand of ethylene is surging at unprecedented rate and therefore conversion of methane into these important value-added chemicals using one-step processes such oxidative coupling of methane is of paramount importance. It would be beneficial and significant to design a catalyst which can suppress undesired over reactions which lead to the production of CO<sub>x</sub> gases resulting in issues of yield and selectivity. To this end, in this study, Machine learning (ML) models integrated with genetic algorithm (GA) have been developed to predict, evaluate, and analyze adsorption energies of methane related species on Cu-based Alloys. Comparative study of different ML algorithms integrated with Genetic Algorithm (GA) were performed to improve the ML model's architecture and parameters selection. The results proposed that Categorical boosting (Catboost) model outperformed all other models and effectively predicts adsorption energies compared to other models (RMSE = 0.0977, CC = 96.5 %). Permutation importance score was utilized to assess prediction performance and accuracy which revealed that group number and surface energy contributed the highest to the model. The partial dependence plots (PDPs) analysis shows the potential effects of each influencing parameter impact on the prediction of the respective adsorption energies and as well as shows that how these factors will interact during oxidative coupling of methane (OCM). Finally, in order to analyze the distribution of the data points of the features and how they affect the model's predictions, bee swarm plot was employed.

**Keywords:** Heterogenous Catalysis, Machine Learning, Genetic Algorithm (GA), Boosting, Artificial Intelligence, Oxidative Coupling of methane (OCM),

# CHAPTER 1: INTRODUCTION

## 1.1 Background

Methane can be extracted from natural gas and as the reserves of petroleum diminish, it will eventually become an essential and significant raw material for manufacture of fuels and chemicals [1]. Research has shown that methane will last us 60 years once we exhaust oil reserves [2]. Its potential as a feedstock has not been tapped to its extent. Methane can either be converted directly or indirectly to fuels and chemicals. The indirect pathway utilizes synthesis gas while directly it can be converted methanol and C<sub>2</sub> hydrocarbons [3]. As of today, the economically viable pathway is via the indirect route. Synthesis gas can be produced using methane through steam reforming, dry reforming and partial oxidation. Commercially, mass production of synthesis gas is done using steam reforming which is endothermic in nature [4]:



However, this process has its drawbacks. Depending on the catalyst used, the temperatures can range from 700K to 1050K requiring increased heat fluxes [6]. Elevated pressure is also needed which may reach 25 bars [7]. Therefore, the reactor must be able to withstand these demanding conditions.

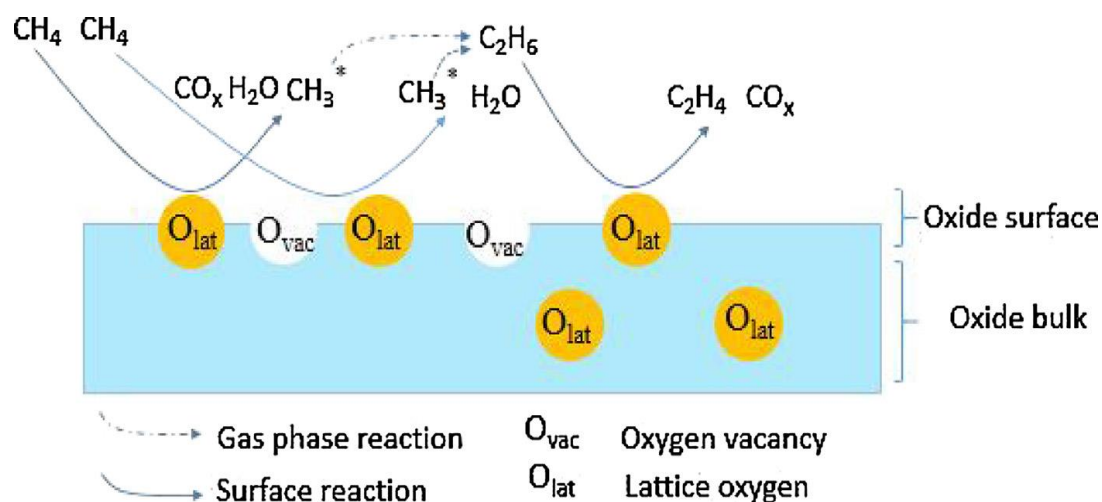
Stringent heat regulation is also paramount as huge quantities of heat are required to carry out this process [8]. This leads to greater energy requirements and higher costs [9]. Furthermore, the efficiency of this reaction is low, and the process suffers from stability issues [10].

In contrast, the direct conversion of Methane in the presence of an oxidant into value-added chemicals such as methanol and ethylene present a greater opportunity [11]. The reason being that ethylene is one of the most sought-after chemicals and it can be transported with ease in its liquid state. Additionally, it is an environment-friendly process which will contribute to the decarbonization of the industry [12]. Extensive research was carried out in the 1980s on oxidative coupling of methane (OCM), but certain drawbacks relating to catalysis of the process led to its decline.

OCM has 2 components. Heterogeneously, CH<sub>4</sub> is first activated on the metal-oxide surface. Conversely, in the homogenous gas-phase coupling of free radical occurs [13]. Methyl radicals (CH<sub>3</sub><sup>\*</sup>) are formed when hydrogen is ejected from Methane by the activated oxygen molecules present on the catalyst surface. Resultantly, Ethane (C<sub>2</sub>H<sub>6</sub>) is produced as a result of coupling

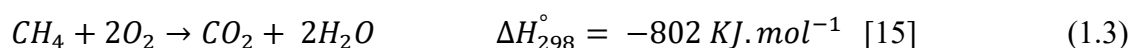
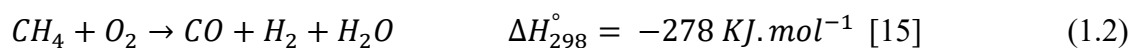


of  $\text{CH}_3^*$  radicals in the gaseous phase. In the final step, ethane is dehydrogenated to form ethylene ( $\text{C}_2\text{H}_4$ ). OCM reactions are highly exothermic requiring temperatures of 950-1200 K as vast quantities of energy is required to cleave the C-H bond ( $429 \text{ KJ mol}^{-1}$ ) [14, 15]. Figure 1.1 represents a schematic overview of oxidative coupling of methane (OCM) process.

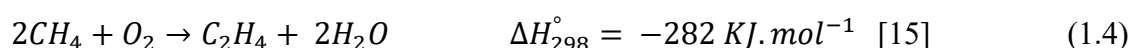


**Figure 1.1:** Schematic overview of oxidative coupling of methane (OCM) process.[12]

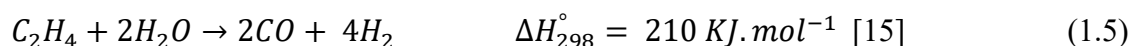
However, OCM suffers from serious shortcoming of low  $\text{C}_2$  yield. This is primarily due to secondary reactions of  $\text{CH}_3^*$  radicals. OCM is a fairly complex reaction as vigorously reactive radical species along with oxygen occupy the reactor [12]. Oxygen acts as an oxidant and a multitude of oxidation and dehydrogenation steps occur. Catalyst capable of activating  $\text{CH}_4$  also activates the Ethane at the same rate which causes production of  $\text{CO}_x$  gases which are undesirable and particularly stable. Furthermore, vacant surface oxygen sites present on the catalyst are refilled by gas-phase oxygen and adsorption of oxygen occurs which enhances  $\text{CO}_x$  formation [11]. According to thermodynamic standpoint the formation of partial and total oxidation products ( $\text{CO}_x$ ) is favorable which restricts  $\text{C}_2$  yield. Complete and partial oxidation occurs according to the following reaction respectively:



It can be seen that both oxidation pathways are exothermic with the former take precedence over the latter at similar conditions. Coupling those results in ethylene forms by the following route:



The enthalpy of the above reaction is near to partial oxidation of methane but dwarfs in comparison to full oxidation of ethylene:



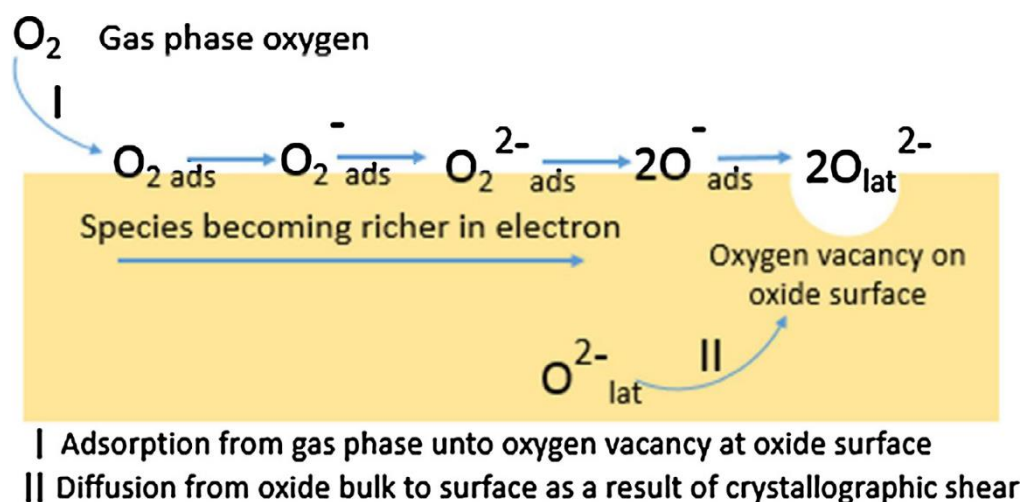
This points to the fact that the conditions that favor formation of ethylene at these temperatures are not favorable in for the activation of methane using oxygen [15]. Nonetheless, when viewed through a thermodynamic lens, partial oxidation and total oxidation are significantly more thermodynamically favorable compared to the oxidative coupling of methane.

The principal and rate-limiting step in OCM reaction is the splitting of C-H bond in in CH<sub>4</sub> endothermically. A study was conducted regarding the kinetics of the reaction, and it revealed that an increase in C<sub>2</sub> yield was the result of an increase in the initial formation rate of methyl radical. Additionally, they discovered that the highest number of methyl radicals were present due to oxygen-attacked hydrogen abstraction of CH<sub>4</sub> rather than from hydrogen abstraction occurring in the gas phase. This shows the paramount importance of catalyst's ability to generating surface oxygen species which are crucial for formation of methyl radicals in OCM reaction [16].

Consequently, it is of utmost importance to stabilize CH<sub>3</sub> so that dehydrogenation to reactive intermediates such as CH<sub>2</sub> and CH is prevented which end up as oxides of carbon impacting the yield significantly[17] . On the surface of the catalysts, Methyl radical forms Methoxide ion by electron acceptance which acts as surface intermediates in the production of CO<sub>x</sub>. The methylene radical (CH<sub>2</sub>) also couples with CO in gas-phase to form Ketene (CH<sub>2</sub>CO). Eventually, these reactions affect yield and selectivity. Therefore, an optimum catalyst should aid in the formation of methyl radicals while suppressing its secondary oxidations reactions which are a source of CO<sub>x</sub>.

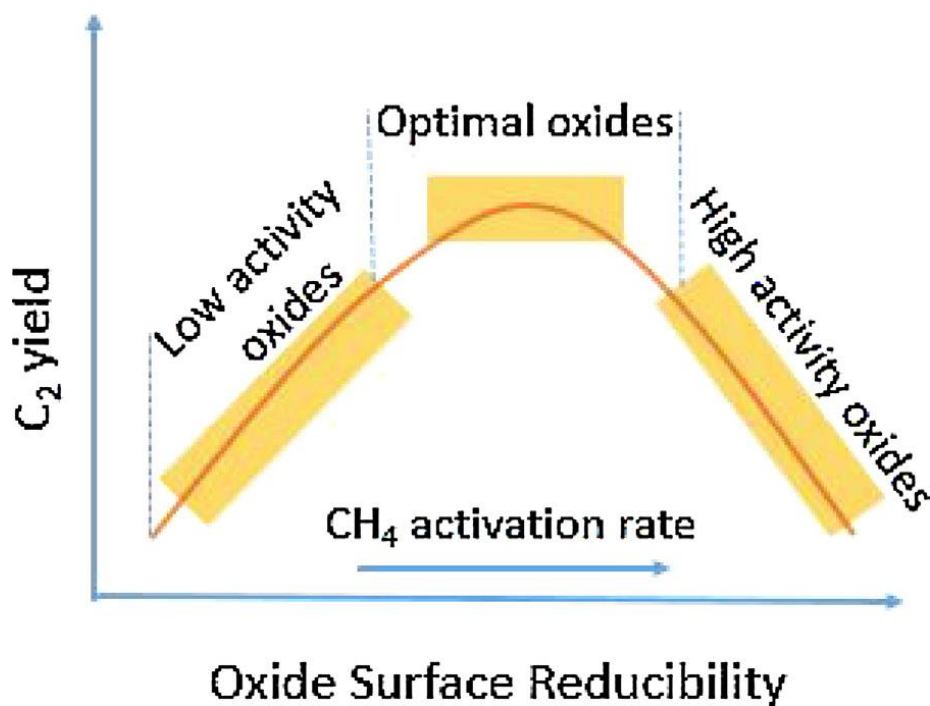
Furthermore, there are other crucial aspects that are required to be taken into consideration in order to understand OCM reaction and design novel catalysts for this purpose. First and foremost are the active sites that activate C-H bond of methane which are widely accepted to be reactive oxygen species on the catalyst surface. The nature and impact of these different oxygen species is subject to opinion. In recent times, the development of DFT calculations is an endeavor to elucidate the discrete functions of the different oxygen species in the OCM reaction. This is crucial for assisting in the development of an effective OCM catalyst that suppresses COX production and improves C2 product selectivity [18].

After the activation of CH<sub>4</sub> on these active sites and the release of H<sub>2</sub>O or CO<sub>x</sub> oxygen vacancies evolve on the catalyst. Oxygen vacancy concentration in oxide catalysts, especially at high temperatures, is strongly related to the abundance of lattice oxygen and its availability for reactions. The energy needed for hydrogen removal to generate a methyl radical drastically reduces with the concentration of oxygen vacancies on the catalyst. Additionally, depending on the location and ability of the oxygen vacancies, they could also have an effect on the refilling of these vacancies by oxygen from the gas phase. Figure 1.2 depicts the conversion of gas phase oxygen into lattice oxygen species on the catalyst.[19]



**Figure 1.2:** Schematic illustration of active site renewal on the oxide surface through replacing the oxygen vacancy formed by methane activation

The oxide catalysts' surface reducibility is a crucial component that influences OCM reaction. The availability of a particular lattice oxygen species is determined by the oxide catalyst's extent and likelihood of reduction. Increased surface reducibility resulted in higher OCM activity, whereas C<sub>2</sub> selectivity eventually declined. Furthermore, methane oxidation becomes disturbed at a certain point, leading to increased amounts of CO<sub>x</sub>. The particular type of lattice oxygen species that are present on the catalyst has a significant impact on the oxide surface reducibility. It's likely that a more electrophilic lattice oxygen will result in less surface reducibility. Figure 1.3 depicts the impact of surface reducibility on the activity of the catalyst and it impacts on C<sub>2</sub> yield as well as showing the increased selectivity of CO<sub>x</sub> beyond a certain point.



**Figure 1.3:** Impact of surface reducibility on the activity of the catalyst and its impact on C<sub>2</sub> yield as well as showing the increased selectivity of Cox [12]

In heterogeneous catalysis, the essential condition for it to take place typically involves the adsorption of molecules from the reacting substances onto the inner or outer surface of the catalyst [20]. Now ML is used in the study because the industry is undergoing a transformation towards AI-driven smart manufacturing, commonly referred to as Industry 4.0, machines are now capable of autonomous communication and collaboration.

Machine learning is at the forefront of AI, allowing systems to create mathematical models from training data. This helps them learn and make predictions, even in unfamiliar situations [21]. The precision comes from carefully selecting features and effective training, making ML a valuable tool for understanding complex environmental phenomena with variations over time and space [22-24].

ML has a wide impact, touching all aspects of life and being a key part of Industry 4.0. Inspired by how the human brain learns, ML gives computers the ability to handle complex problems by learning and adapting to different inputs. It has many advantages, like dealing with complexity, improving computational efficiency, handling uncertainty, and aiding decision-making [24]. ML's progress has benefited various scientific and engineering fields, including materials science, bioengineering, construction management, and transportation engineering [25].

Numerous studies have explored the diverse applications of AI and machine learning in various industries. These encompass real-time monitoring [26, 27], predictive maintenance [28, 29], quality control [30, 31], energy efficiency enhancements [32, 33], data mining and analytics [34], drug discovery and development [35], streamlining industrial processes [36], defending against process-aware assaults on industrial control systems [37], optimization [38], monitoring and diagnosing faults in supervised processes [39], IoT-based Smart Agriculture systems [40], industrial tomography [41], predicting Heart Failure Disease [42], earthquake engineering [25], and more.

This study aims to develop a machine learning model that has the ability to accurately predict the adsorption energies of CH<sub>4</sub> related species which are CH<sub>3</sub>, CH<sub>2</sub>, CH on Cu-based alloys. The effect of doping of copper with various transition metals for use as a catalyst in OCM is evaluated.

The electric charge distribution of the individual components within the oxide significantly affects their catalytic behavior. If a dopant is integrated within the crystal structure of the metal, the morphology is modified leading to defects in the lattice and electronic structure of the oxide is also changed [43]. Charge transfer would be facilitated, and an improved energy flow would take place. The characteristics and properties of the different components synergize so as to contribute to the aforementioned functionalities [44].

To this end gradient boosting algorithms were employed which are eXtreme Gradient Boosting (XGboost), CatBoost (Categorical Boosting) and light gradient boosting. The goal of conducting a comparative analysis of the models was to avoid pitfalls of overfitting and underfitting, with the ultimate aim of improving predictive model performance. The hyperparameters were tuned using genetic algorithm. The results achieved will provide insight into the role of catalysis and its ability to suppress undesired over-reactions.

## 1.2 Objectives

The objectives of the thesis are given below,

- Development of machine learning models for calculating adsorption energies of Methane on Cu-based alloys.
- Optimization of the model's hyperparameters using Genetic Algorithm for accurate predictions of adsorption energies of Methane.

- Comparison of ML models to assess advantages and drawbacks based using different metrics.

### **1.3 Thesis outlines**

The Thesis follows the following pattern: In Chapter 1, we provide an Introduction and background to the research topic. This is followed by Chapter 2, where we give a detailed literature for the assessment of adsorption energies of Methane on Cu-based alloys. Chapter 3 delves into the research methodology employed in the development of optimization framework for adsorption energies of Methane on Cu-based alloys. Moving on to Chapter 4, we present the outcomes of our research, including different machine learning model comparison based on the advantages and drawback of different metrics.

## CHAPTER 2: LITERATURE REVIEW

### 2.1 Literature review

With a global natural gas reserves of approximately  $208 \times 10^{12}$  cubic feet [45, 46] and the exhaustion of fossil fuels, research on converting methane into useful compounds using innovative technologies is rapidly rising. Currently, the industrial-scale conversion of methane to olefins involves reforming  $\text{CH}_4$  into  $\text{CO}$  and  $\text{H}_2$ , followed by Fischer-Tropsch synthesis [47, 48] or direct conversion of syngas to light olefins using bifunctional catalysts like  $\text{ZnCrOx-SAPO}$  [47]. However, this process consumes huge amounts of energy as the cleavage of all 4 C-H bonds in the first step doesn't lead to yielding of  $\text{CH}_2$  or  $\text{CH}_3$  radicals which is integral to the production of  $\text{C}_2+$  species. Furthermore, it doesn't utilize carbon properly due to conversion to  $\text{CO}_2$ .

This has led to an enhanced interest in the recent decades for direct conversion of methane into value added chemicals such as ethylene, formaldehyde etc. This is due to its lower costs and greener process. Initially, Methane was converted into these compounds by passing it over iron-embedded silica matrix without the presence of oxygen. It resulted in high selectivity to the required  $\text{C}_2$  and  $\text{C}_3$  hydrocarbons however, the catalyst would quickly deactivate leading to decline in its application to commercial use [49].

Therefore, oxidative coupling of methane (OCM) has sparked the highest interest. This was pioneered by Keller and Bhasin in 1982 [50]. Beyond that, extensive research has been carried out to optimize this process and increased its yield and selectivity. With the research of Baerns in 1984, it was found that suitable  $\text{C}_2$  yields can be attained through this process. He attained 58 % selectivity at 5 % conversion of Methane by utilizing a  $\text{PbO}/\text{Al}_2\text{O}_3$  catalyst [13]. Following this, Ito et al. achieved 50 %  $\text{C}_2$  selectivity at 28 % conversion by using a lithium promoted  $\text{MgO}$  catalyst ( $\text{LiMgO}$ ) [13]. Initial research was aimed at maximizing  $\text{C}_2$  yield by modifying catalyst properties and composition however, economic considerations have shown that  $\text{C}_2$  selectivity is more integral than yield given that there is acceptable  $\text{CH}_4$  conversion.

Since then, multitude of catalysts which may include pure oxides, alkaline earth, or other metals have been subject to experiment to achieve the desired results for OCM reaction. With the advent of 1990s, alkaline earth metal oxides were tested to find suitable candidates for the OCM reaction with focus on either basis of  $\text{MgO}$  or  $\text{CaO}$ . These showed improved  $\text{C}_2+$

selectivity according to their basicity. Li/MgO was found to be a good catalyst. Gao et al. integration of 3 wt.% Li onto MgO resulted in catalyst that gave substantial CH<sub>4</sub> conversion and C<sub>2+</sub> selectivity at 30% and 62% respectively [51]. This catalyst faced a drawback of losing its metal ions at its reaction conditions of 780 °C so it had an inherent weakness of catalytic stability [52]. Adversely, pure CaO powder showed greater CH<sub>4</sub> conversion of 40 % and C<sub>2+</sub> selectivity of 50 %. In comparison, CaO- fully coated silica particles exhibited a more reliable and stable catalytic activity [53]. Lastly, alkali metal-doped CaO catalysts, Na-CaO catalysts showed best results with 24.7 methane conversion and 68.8 % C<sub>2+</sub> selectivity.

Rare-earth metal oxides were also researched to find potential catalyst among them. It was found that La<sub>2</sub>O<sub>3</sub>- based catalysts were best performing as they showed very good activity of catalyst as well as stability while at low temperature reaction condition [54]. Hou et al, studied La-based catalysts and achieved C<sub>2+</sub> yield of around 18 % over a Li-ZnO/ La<sub>2</sub>O<sub>3</sub> catalyst [55]. Some recent studies involving doped and undoped lanthanide catalysts have shown that Sr-doped La<sub>2</sub>O<sub>3</sub> is the most suitable catalyst among the rare-earth metal oxides. They argued that as oxygen concentration is essential for enhanced C<sub>2+</sub> selectivity, La augmented the surface oxygen species during the reaction [56]. The doping of La<sub>2</sub>O<sub>3</sub> with an ion that has lower valency like Sr produced a p-type semiconductor that has vacancies for oxygen resulting in easy diffusion of oxygen through these vacancies to create oxygen ions that activated methane in a very reactive manner at lower temperatures of around 400 °C [57].

Extensive studies have been performed on Mn-Na-W-SiO<sub>2</sub> catalyst due to its good C<sub>2+</sub> selectivity. Wu et al. put forward that W-O-Si bond caused the breaking of C-H bond in CH<sub>4</sub> as the lattice oxygen associated with H in CH<sub>4</sub> [58]. Furthermore, Ji et al. concluded that the methane conversion is directly proportional to W concentration by experimenting on different W concentration using Mn-Na-W catalyst. The catalytic activity enhanced with increase in W content [59]. Even though increased C<sub>2+</sub> selectivity has been attributed to Na-W-Mn-SiO<sub>2</sub> catalysts, they suffer from significant deactivation due to reduction in their surface area, displacement of Na<sub>2</sub>WO<sub>4</sub> from the catalyst and the formation of inactive phases [60, 61].

Another catalyst widely prevalent are perovskites (ABO<sub>3</sub>) due to their redox properties which are controllable. As there is a huge selection of elements that can be introduced in either A or B position to create different compositions with various redox potentials for the OCM reaction. Sim et al. considered main factors that affected the C<sub>2+</sub> yield in OCM by testing ten ABO<sub>3</sub> catalysts. He concluded that conductivity of oxygen ion can be used as a benchmark for



deducing performance of the catalyst. Those perovskites that had greater oxygen ion conductivity resulted in greater methane conversion. On the contrary, higher binding energies of some perovskites are more favorable to the formation of CO and further oxidation to CO<sub>2</sub> by adsorbed oxygen species [61].

Recently, collaborated research efforts have been made to explore other aspects of OCM including nature of active sites, oxygen vacancy, acid-base properties and surface oxide reducibility. Active sites that activate C-H bond of CH<sub>4</sub> are generally accepted as oxygen species present on the catalyst surface. Although, the exact nature of these active is debated as there are different types of oxygen species present on the surface of the catalyst which play their unique role in oxidation [62]. Gordienko et al. studied the properties and the role of these oxygen species using NaWMn/ SiO<sub>2</sub> catalyst. He showed that bounded oxygen species were present on the catalyst which were both strong and weakly bounded in nature. They also found that selectivity was dependent on the reduction temperature therefore different species of oxygen aided the reaction using different routes. Weakly bounded oxygen specie were concluded to be the main contributor to the overall stability and activity of the OCM reaction [63]. Apart from the oxygen species, oxygen itself considerably enhanced CH<sub>4</sub> conversion but with a serious drawback of greater CO<sub>x</sub> formation [64].

Lately, focus of the research has also been on oxygen vacancy which is due to the hydrogen abstraction of methane resulting in release of H<sub>2</sub>O or CO<sub>x</sub>. The catalyst's oxygen vacancy concentration, particularly at high temperatures, is closely related to the amount of lattice oxygen and their availability for reaction. Efficient OCM catalysts should have a high concentration of oxygen vacancies at normal OCM temperatures. Cheng et al. experimented on systems of iron oxide to discover the role of these vacancies by utilizing DFT computation. It was revealed that there was a notable decrease in the energy required to form methyl radicals via hydrogen abstraction if there were a higher number of oxygen vacancies [65]. In addition, the relative position along with stability of the oxygen vacancies could impact their replenishment by the gas phase oxygen or through migration of the vacancies [25]. This refilling of the oxygen vacancies causes the re-oxidation of the reduced cations. As a consequence, the active sites get restored in order for the next catalytic cycle to proceed. Research has shown that vacancies present on at the outer surface have been found to be more stable relative to those that are at the subsurface therefore, the more energetically favorable pathway is from the later to former [66].

It is widely known that methane is activated heterogeneously at the oxide catalyst surface. The behaviors of the component species of the oxide in catalyzing processes are greatly influenced by their distribution of electric charge, with the surface oxide anions (oxygen and hydroxyl groups) and surface metal cations being regarded as basic and acidic centers, respectively. Various studies have shown that acid-base property is central to the improved performance of the OCM reaction [67]. Zavyalova et al. performed an in-depth analysis of the data available on OCM catalysts revealing that in cases where the basic character was strong and greater, it led to a good C<sub>2</sub> selectivity [43]. When a dopant metal is introduced into the crystal structure, the lattice as well as the electronic configuration is altered of the host oxide creating a hybrid material. Electronic or charge transfer among the constituents of the hybrid could become more facilitated via a synergy of properties of the individual components [68]. A notable finding by Song et al. showed that basicity of catalysts used in OCM could be modified by proper doping. By utilizing Sr-doped La<sub>2</sub>O<sub>3</sub> nanofibers, they discovered that sites with strong basic character are essential for heightened OCM performance specifically at temperatures beyond 600 °C. Furthermore, with the introduction of greater Sr, the quantity along with their robustness of the basic sites was also enhanced [44].

Another pivotal factor that impacts the catalytic activity as well as C<sub>2</sub> selectivity is the surface reducibility of the oxide catalysts. The degree and ease of reduction of the oxide catalyst determines the availability of the specific lattice oxygen species. By utilizing DFT technique, Kumar et al. attempted to investigate doped and undoped catalysts and figure out the impact their surface reducibility had on the performance of OCM reaction [69]. It was revealed that greater surface reducibility led to enhanced OCM activity although the selectivity of C<sub>2</sub> gradually decreased. Moreover, after a certain point, the oxidation of methane is disrupted resulting in greater quantities of CO<sub>x</sub>. The oxide surface reducibility is highly dependent on the type of lattice oxygen species present on the catalyst. A more electrophilic lattice oxygen is probably going to cause surface reducibility to be lower. According to a study, the oxygen present in the lattice of LaAlO<sub>3</sub> showed better C<sub>2</sub> selectivity as it had more electrophilic character [64]. Those catalysts whose reduction temperature is low and simultaneously surface reducibility is high are supposed to show improved and enhanced oxidation and vice versa [70]. This indicates that there may be a limit to just how low the reduction temperature can be set for a specific oxide catalyst in order to increase C<sub>2</sub> selectivity. Therefore, determining how to properly optimize the surface reducibility is essential to constructing effective OCM oxide catalysts.

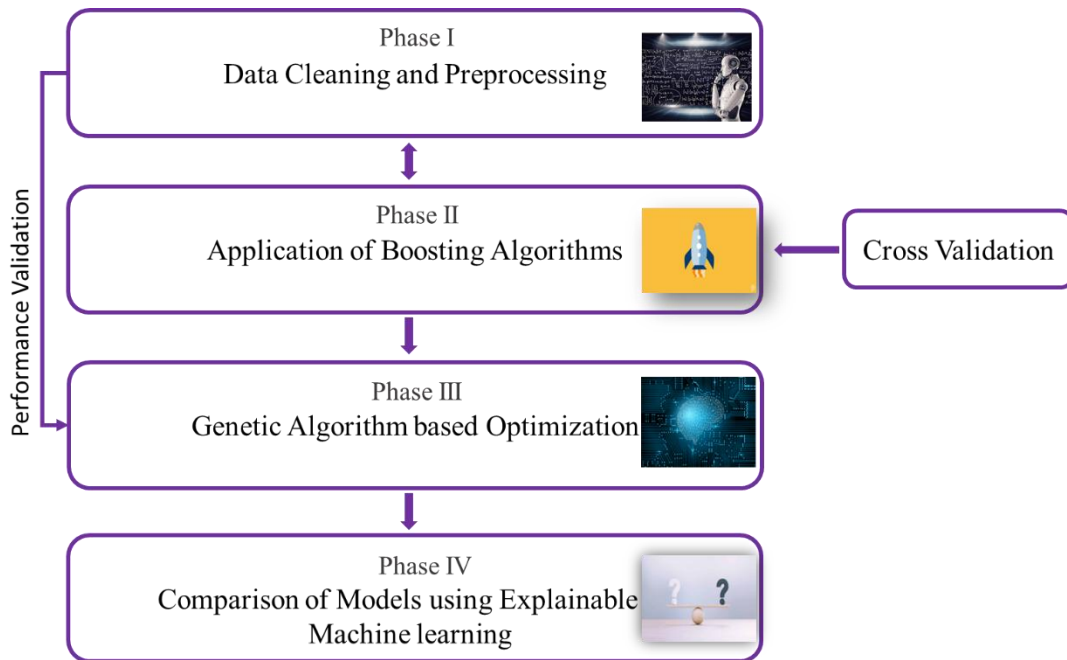
OCM reaction has immense potential to convert methane directly into value-added chemicals therefore, in order to design and synthesize novel catalysts that can bring the process to an industrially sustainable scale in terms of C<sub>2</sub> yield, more innovative and different approaches must be pursued. Hence, to this end, different novel approaches have been employed such as analyzing reaction kinetics of the various OCM pathways, utilizing computational techniques to optimize catalyst composition along with experimentation focusing on fundamentals of the reaction [56, 71, 72].

Most recently, as computational power has significantly advanced, computational tools have been employed to design new and innovative catalysts by varying compositions of the catalysts materials as well as other crucial factors and parameters influencing OCM. Kondratenko et al. used an innovative technique where they researched on the role of individual components of the catalyst utilizing well planned and developed experiments thereafter optimizing the catalyst composition [73]. Nonetheless, for a more realistic catalyst design, common conditions of the OCM reaction need to be considered in both experimental or computational approach.

# CHAPTER 3: METHODOLOGY AND MODELLING OF ALGORITHMS

## 3.1 Methodology

The figure 3.1 represent the overall methodology of this study, and is briefly explained below



**Figure 3.1:** Methodology

### Phase I: Data Cleaning and Preprocessing

In order to remove inefficiencies and standardize the data, preprocessing was carried out. Feature scaling was carried out to transform all value to similar scale so that all features contribute equally to the model. The dataset was divided into 90% training and 10% testing dataset so that more data points are available for enhanced training and validation.

### Phase II: Application of boosting algorithms

In this phase, application of 3 boosting models was carried out. The parameters of the models were defined and given specific ranges. Furthermore, to avoid the pitfalls of overfitting 5-fold cross validation was carried out. This will also prevent premature generalization of results as each time different set of data points would be taken for training and testing. Hyperparameters were also initiated and defined so as to optimize them in the next phase. For evaluating the

results, Root Mean Squared Error (RMSE) and Correlation Coefficient (CC) were utilized which will quantify the accuracy of the model.

### **Phase III: Genetic Algorithm based Optimization**

GA framework was used in optimization of hyperparameters of different Machine learning models. Optimization was enhanced by utilizing different functionalities such as crossover and mutation probability allowing us to reach at the optimal solution over successive generations. For more stringent optimization, 5-fold cross validation was carried out again. Elitism was chosen so that the best solutions are preserved for next generation.

### **Phase IV: Comparison of Models using Explainable Machine learning**

In this phase comparison of different Machine learning models were carried out. ML model were assessed on the bases of advantages and drawbacks of different metrics. It was done in order to compare the advantages and disadvantages of different models and to select the appropriate model for the process. For this purpose, partial dependence plots, permutation importance score and bee swarm plot were chosen to comprehend a model's operations and provide complete understanding.

### **Model's performance and validation**

The model's performance was calculated by two measures: root-mean-squared error (RMSE) and correlation coefficient (CC/R). These values were determined using the following equation (3.1) and (3.2).

$$R^2 = 1 - \frac{\sum_{i=0}^n (Y_i^{exp} - Y_i)}{\sum_i^n (Y_i^{exp} - Y_{avg}^{exp})} \quad (3.1)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_i^n (Y_i^{exp} - Y_i)^2} \quad (3.2)$$

$Y_i^{exp}$  and  $Y_i$  stand for the actual and the expected outcomes results, and n indicates the total test samples.

The coefficient of determination, typically denoted as R-squared, falls within the range of 0 to 1. A value of 0 suggests that the output variable cannot be effectively predicted from the regressor variables, while a value of 1 signifies that the response variable is entirely predictable

from the regressor variables. The RMSE (Root Mean Square Error) is always a non-negative value, and a smaller RMSE indicates more accurate model predictions.

### 3.2 Dataset and existing research

The dataset consisted of 46 transition metal and their respective physico-chemical and electrical properties were considered which makes 12 as the total number of descriptors. Table 3.1 shows the values of the 12 features of each element used for prediction of the adsorption energies. The data was taken from [17]. They performed Density Functional Theory (DFT) using Vienna ab initio simulation package (VASP) to calculate the adsorption energies of the various Cu-based alloys. The DFT modelling and parameters can be found in [17]. As for the model construction of the Cu-based alloys, these 46 elements replaced the Cu atom present at the center of the surface layer of the slab model. Toyao et al. [17] modelled 9 different algorithms and concluded that extra tree regression (ETR) gave superior results with the root mean squared error (RMSE) in the range of 0.24-0.27 eV. Moreover, Zhang et al. (2021). [74] applied Gaussian process regression in conjunction with Bayesian optimization for hyperparameter tuning. They restricted the RMSE in the range of 0.12-0.154 eV.

#### 3.2.1 Data visualization

Data visualization can be defined as presenting information in form of graph or pictures making it understandable and interpretable [75]. For this purpose, boxplots were constructed for the descriptors as shown in Fig 3.2. They allow us to visualize the measures of central tendency as well as range and quartiles. Comprehending the structure of boxplot allows improved evaluation and assessment of the data [76]. We can clearly see the interquartile range as well as outliers. Skewness in the data can also be noticed in some of the parameters.

**Table 3.1:** Input features (descriptors) used for prediction of adsorption energies

Element	AN	AM / g mol <sup>-1</sup>	G	P	R / Å	EN	m. p. / K	b. p. / K	$\Delta_{\text{fus}}H$ / kJ mol <sup>-1</sup>	$\rho$ / g cm <sup>-3</sup>	IE / eV	SE / J m <sup>-2</sup>
Mg	12	24.31	2	3	1.45	1.23	923	1363	8.5	1.74	7.65	0.54
Al	13	26.98	13	3	1.18	1.47	933	2792	10.8	2.7	5.99	0.8
Si	14	28.09	14	3	1.11	1.74	1687	3538	50.2	2.33	8.15	1.28
Ca	20	40.08	2	4	1.94	1.04	1115	1757	8.5	1.54	6.11	0.46
Sc	21	44.96	3	4	1.84	1.2	1814	3109	14.1	2.99	6.56	1.2
Ti	22	47.87	4	4	1.76	1.32	1941	3560	14.2	4.51	6.83	1.93
V	23	50.94	5	4	1.71	1.45	2183	3680	21.5	6	6.75	2.38
Cr	24	52	6	4	1.66	1.56	2180	2944	21	7.15	6.77	3.2
Mn	25	54.94	7	4	1.61	1.6	1519	2334	12.9	7.3	7.43	3.39

Fe	26	55.85	8	4	1.56	1.64	1811	3134	13.8	7.87	7.9	2.45
Co	27	58.93	9	4	1.52	1.7	1768	3200	16.1	8.86	7.88	2.11
Ni	28	58.69	10	4	1.49	1.75	1728	3186	17	8.9	7.64	1.92
Cu	29	63.55	11	4	1.45	1.75	1358	2835	12.9	8.96	7.73	1.31
Zn	30	65.38	12	4	1.42	1.66	693	1180	7.1	7.14	9.39	0.35
Ga	31	69.72	13	4	1.36	1.82	303	2477	5.6	5.91	6	0.48
Ge	32	72.64	14	4	1.25	2.02	1211	3106	36.9	5.32	7.9	0.87
Se	34	78.96	16	4	1.03	2.48	494	958	6.7	4.81	9.75	0.05
Sr	38	87.62	2	5	2.19	0.99	1050	1650	7.4	2.64	5.69	0.34
Y	39	88.91	3	5	2.12	1.11	1795	3618	11.4	4.47	6.22	0.96
Zr	40	91.22	4	5	2.06	1.22	2128	4682	21	6.52	6.63	1.57
Nb	41	92.91	5	5	1.98	1.23	2750	5017	30	8.57	6.76	2.07
Mo	42	95.96	6	5	1.9	1.3	2896	4912	37.5	10.2	7.09	2.8
Tc	43	99	7	5	1.83	1.36	2430	4538	33.3	11	7.28	2.23
Ru	44	101.07	8	5	1.78	1.42	2607	4423	38.6	12.1	7.36	2.58
Rh	45	102.91	9	5	1.73	1.45	2237	3968	26.6	12.4	7.46	1.99
Pd	46	106.42	10	5	1.69	1.35	1828	3236	16.7	12	8.34	1.34
Ag	47	107.87	11	5	1.65	1.42	1235	2435	11.3	10.5	7.58	0.77
Cd	48	112.41	12	5	1.61	1.46	594	1040	6.2	8.69	8.99	0.2
In	49	114.82	13	5	1.56	1.49	430	2345	3.3	7.31	5.79	0.3
Sn	50	118.71	14	5	1.45	1.72	505	2875	7.2	7.26	7.35	0.54
Sb	51	121.76	15	5	1.33	1.82	904	1860	19.8	6.68	8.61	0.16
Te	52	127.6	16	5	1.23	2.01	723	1261	17.5	6.24	9.01	0.08
La	57	138.91	4	6	1.95	1.08	1191	3737	6.2	6.15	5.58	0.7
Ce	58	140.12	5	6	1.85	1.08	1071	3716	5.5	6.77	5.54	1.02
Pr	59	140.91	6	6	2.47	1.07	1204	3793	6.9	6.77	5.47	0.78
Nd	60	144.24	7	6	2.06	1.07	1294	3347	7.1	7.01	5.53	0.82
Hf	72	178.49	4	6	2.08	1.23	2506	4876	27.2	13.3	6.83	1.71
Ta	73	180.95	5	6	2	1.33	3290	5731	36.6	16.4	7.55	2.34
W	74	183.84	6	6	1.93	1.4	3695	5828	52.3	19.3	7.86	3.23
Re	75	186.21	7	6	1.88	1.46	3459	5869	60.4	20.8	7.83	2.58
Os	76	190.23	8	6	1.85	1.52	3306	5285	57.9	22.59	8.44	2.92
Ir	77	192.22	9	6	1.8	1.55	2719	4701	41.1	22.5	8.97	2.28
Pt	78	195.08	10	6	1.77	1.44	2042	4098	22.2	21.5	8.96	1.48
Au	79	196.97	11	6	1.74	1.42	1337	3129	12.7	19.3	9.23	0.74
Tl	81	204.38	13	6	1.56	1.44	577	1746	4.1	11.8	6.11	0.22
Pb	82	207.2	14	6	1.54	1.55	601	2022	4.8	11.3	7.42	0.25

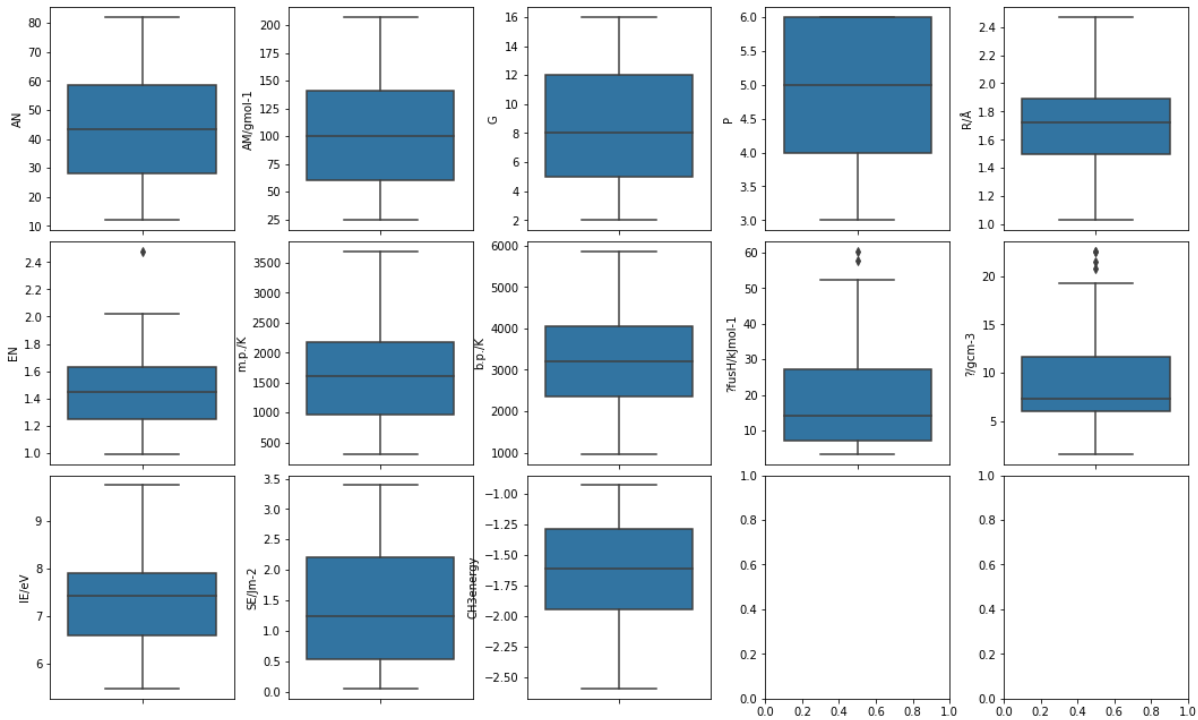
Atomic number (AN), group (G), period (P), atomic radius (R) in Å, atomic mass (AM) in g mol<sup>-1</sup>, electronegativity (EN), melting point (m.p.) in K, boiling point (b.p.) in K, enthalpy of fusion ( $\Delta_{\text{fus}}H$ ) in kJ mol<sup>-1</sup>, density at 25 °C ( $\rho$ ) in g cm<sup>-3</sup>, ionization energy (IE) in eV, surface energy (SE) in eV

### 3.2.2: Machine learning algorithms

In this study, we employed eXtreme Gradient Boosting (XGboost), CatBoost (Categorical Boosting) Regression, and Light Gradient Boosting Model (LGBM) to accurately and precisely estimate the adsorption energies. of the methane related species-CH<sub>3</sub>, CH and CH on the 46 unique Cu-based alloys. The design, training and testing of these models were carried out using Python software. For the purpose of hyperparameter optimization, we utilized Genetic Algorithm.

### 3.2.3: Gradient Boosting Algorithms

In recent decades, boosting algorithms have emerged as a propitious and promising technique for constructing machine learning models. The theoretical basis behind this approach is to merge or combine the solutions of weak learners to form a strong learner which has enhanced prediction capability and superior accuracy [77]. This strong learner can be achieved by iteratively improving (boosting) the weak base-learners. According to research, gradient boosting decision tree (GBDT) is a popular ML technique because of its effectiveness, efficacy, precision, and interpretability [78]. An analogy can be constructing an adequate and satisfactory hypothesis from a comparatively mediocre hypothesis [79].



**Figure 3.2:** Boxplot of the feature variables



The first boosting algorithms were developed by Schapire and Freund who labelled its concept as ‘‘gathering wisdom from a council of fools’’ [80, 81]. The uncomplicated base learners are the fools; however, these simple, inaccurate base learners have some useful information regarding the structure and framework of the problem. With every iteration, the base learner will gradually move towards the improved solution. In the case of regression, each new learner will attain an importance according to its contribution to reaching the optimum solution and it will be represented accordingly. For example, the weak learner will be given a specific weight (AdaBoost) and the higher weight learners will be given more importance in the final solution.

In contrary, scaling will be employed according to some parameter such as learning rate (Gradient boosting, XGboost) and each weak learner will contribute equally to the solution with each basic learner taking us to the required results in an iterative way. Finally, the predictions of the individual learners are combined into more precise and accurate estimations. With regards to boosting in regression, the relationship between the descriptors ( $x$ ) and the expectation of the response  $E(Y)$  is quantified using an interpretable function  $E(Y | X = x) = f(x)$ . if there are multiple descriptors, an additive model is formed by adding the effects of the individual predictors:

$$f(\mathbf{x}) = \beta_0 + h_1(x_1) + \dots + h_p(x_p) \quad [77] \quad (3.3)$$

In this equation,  $\beta_0$  is the intercept. The descriptors are  $x_1, \dots, x_p$  forming component  $X$ .  $h_1(\cdot), \dots, h_p(\cdot)$  perform the function of integrating and incorporating the effect of the predictors.

### **eXtreme Gradient Boosting (XGboost)**

Xgboost is a scalable ensemble algorithm, which is based on gradient boosting. It is an effective and reliable approach to addressing problems involving machine learning [82]. XGBoost constructs an additive extension of the objective function by mitigating the loss function similar to gradient boosting. A different loss function is utilized to regulate the complexity and intricacy of the trees as XGBoost utilizes decision trees as its basic estimator.

$$L_{XgB} = \sum_{i=1}^N L(y_i, F(x_i)) + \sum_{j=1}^M \Omega(h_m) \quad [83] \quad (3.4)$$

$$\Omega(h) = \gamma B + \frac{1}{2} \lambda \|W\|^2 \quad [84] \quad (3.5)$$

In the above equations,  $B$  is total leaves in the tree *and*  $W$  refers to the scores of the output of the leaves. Pruning can also be utilized. Gamma which is denoted by  $\gamma$  is the minimum loss reduction that will create a new split. The trees complexity can be reduced by adjusting depth of the trees, using L1 and L2 regularization and tuning the learning rate to prevent overfitting. Randomization is another feature of Xgboost which further mollifies overfitting and enhance computational speeds. The parameters for this purpose are random subsampling and column subsampling.

The success of XGBoost is mostly due to its scalability in all situations. It operates more than 10 times quicker than currently used solutions. Furthermore, in scenarios where there are memory constraints, it scalable to billions of samples. The scalability of XGBoost may be directly attributed to a number of significant algorithmic and systemic enhancements [83].

One of these enhancements is a unique tree learning approach for managing sparse data. Learning is accelerated by parallel and distributed computing, which speeds up model search. Moreover, XGBoost uses out-of-core processing to process extremely large datasets.

### **CatBoost (Categorical Boosting) Regression**

Catboost Regression is a relatively new gradient boosting algorithm. This algorithm has the ability to handle categorical descriptors in a novel way tackling them at the training stage rather than during preprocessing [10]. Furthermore, it introduces a new framework in order to approximate leaf values while choosing structure of the tree that avoids or combats overfitting. It has also integrated GPU implementation that allows swift and accelerated training than XGBoost and LightGBM on ensemble of similar sizes.

Categorical features are characterized by a discrete collection of values known as categories, which are not always similar with one another. One-hot encoding is the commonly used technique which is used for categorical features which have low cardinality. CatBoost employs a more effective method that minimizes overfitting and permits the utilization of the entire dataset for training purposes. Specifically, we randomly permutate the dataset and calculate the average label value for the instance with the identical category value placed before the provided one in the permutation for each example.

Taking a dataset  $D = \{(X_i, Y_i)\}_{i=1..n}$ , in which  $X_i = (x_{i,1}, \dots, x_{i,s})$  is a vector with  $s$  features, and  $Y_i \in \mathbb{R}$  is the value of the label. If the permutation is  $\sigma = (\sigma_1, \dots, \sigma_n)$ , then substitution of  $x_{\sigma_j,k}$  occurs with

$$\frac{\sum_{j=1}^{p-1} [x_{\sigma_j,k} = x_{\sigma_p,k}] Y_{\sigma_j} + a.P}{\sum_{j=1}^{p-1} [x_{\sigma_j,k} = x_{\sigma_p,k}] + a} \quad [85] \quad (3.6)$$

where we additionally add the previous value  $P$  and a parameter  $a > 0$  with a value greater than zero that represents the prior's weight. The technique of adding prior is rather common, and it contributes to the reduction of noise that is produced from low-frequency categories [86].

Oblivious trees are the foundation of CatBoost's base predictors. When it comes to trees like this, the criteria for splitting are consistent throughout the whole level of the tree. These trees are finely balanced and are less likely to become overfit.

#### 3.2.4: Light Gradient Boosting Model

The efficiency and scalability of models when dealing with high dimensionality and large datasets is still subpar. This is due to the reason that for every feature, all the data points have to be examined in order to calculate the information gain caused by splitting. To counter this issue, an improved gradient boosting decision tree (GBDT) called 'LightGBM' was created. It has integrated two novel methods called Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB).

When GOSS is utilized, a substantial portion of the data instances that have modest gradients are left out and the remaining are used to calculate the information gain. GOSS is able to produce a sufficiently accurate estimate of the information gain with the smaller dataset by exploiting the fact that data points that have greater gradients play a more significant part in the calculation of the information gain. However, this will cause a bias issue in favor of the sample with bigger gradients and alter the initial data distribution. GOSS does a random sampling on the data with low gradients while maintaining all the samples with higher gradients in order to address this problem. When calculating the information gain, GOSS boosts the weights (i.e., by applying a constant multiplier) of the data points with small gradients because the selection would still be skewed toward the data with higher gradients [87]. In order to decrease the features, EFB groups together features that cannot normally take non-zero values

concurrently (that is, they are mutually incompatible). This greedy algorithm can achieve a reliable and accurate approximation ratio and, as a result, can successfully reduce the number of features without significantly compromising the accuracy of split point determination. The integration of these techniques substantially decreases the memory consumption and enhances computational speed [88]. Another reason behind these improvements is that LightGBM transforms continuous values of features into discrete bins which boost the training process [89].

### 3.3 Genetic Algorithm

Established by John Holland in the 1970s, genetic algorithm (GA) is an adaptive heuristic search method which derives its basis from genetics of population [90]. Under the hood, it is following principle of “*Survival of the fittest*” introduced by Charles Darwin [91]. Natural selection is a key idea behind GA.

The functions of all the genetic operators are as follows:

#### 3.3.1 Population

The initial population was randomly generated, and each conceivable solution is referred to as a chromosome, as demonstrated in Table 1.

$$P = \{p_1, p_2, \dots, p_{\text{pop\_size}}\} \quad (3.7)$$

$$p_i = [p_{i_1} \ p_{i_2} \ \dots \ p_{i_j} \ \dots \ p_{i_{\text{no\_vars}}}] \quad (3.8)$$

$$\text{para}_{\min}^j \leq p_{i_j} \leq \text{para}_{\max}^j \quad (3.9)$$

**Table 3.2:** Chromosomes

Chromosome No. 1	1011000101110010
Chromosome No. 2	1001010110111001

In equation (3.7), `pop_size` represents the total population size, while in equation (3.8), `no_vars` denotes the number of variables to be optimized. The symbols  $\text{para}_{\min}^j$  and  $\text{para}_{\max}^j$  correspond to the minimum and maximum values of the parameter  $p_{i_j}$  in equation (3.9).

It is an evolutionary algorithm which is initiated by providing solutions which form the population. A solution is an individual and the individual solutions comprise to form chromosomes. Every chromosome is defined by a set of genes. Each chromosome can be a probable solution in the search pool [92]. The optimum solution is selected after multiple generations. At a particular instant, a population is defined as a generation. For each generation, the chromosome is assessed according to their fitness value. As for the next generation, the chromosomes are chosen probabilistically in accordance with their respective fitness values [93]. Then GA employs the objective function to evaluate the population. If the benchmark or specifications of the objective function is met, the model stops. Conversely, offspring may be produced. The chromosome's fitness is the criteria which is used for reproduction of offspring [94]. The individuals which have adapted appropriately are retained while others are removed [95].

Individuals that comprise the mating pool are known as parents. The parents are selected sequentially, or random selection may take place. For a pair of parents chosen, in order to introduce diversity, variation operators are applied.

Firstly, crossover operator pairs parental features and produces newer generations that carry genes (features) from both parents. As GA is based on randomness, the genes are selected randomly. The crossover can be one-point, two-point, or homologous in order to exchange genes. As the next step, the mutation operator takes over. This variation operator mutates some genes and changes their value. This way, new individuals are introduced into the population. This will introduce diversity which is essential for reaching the optimum solution.

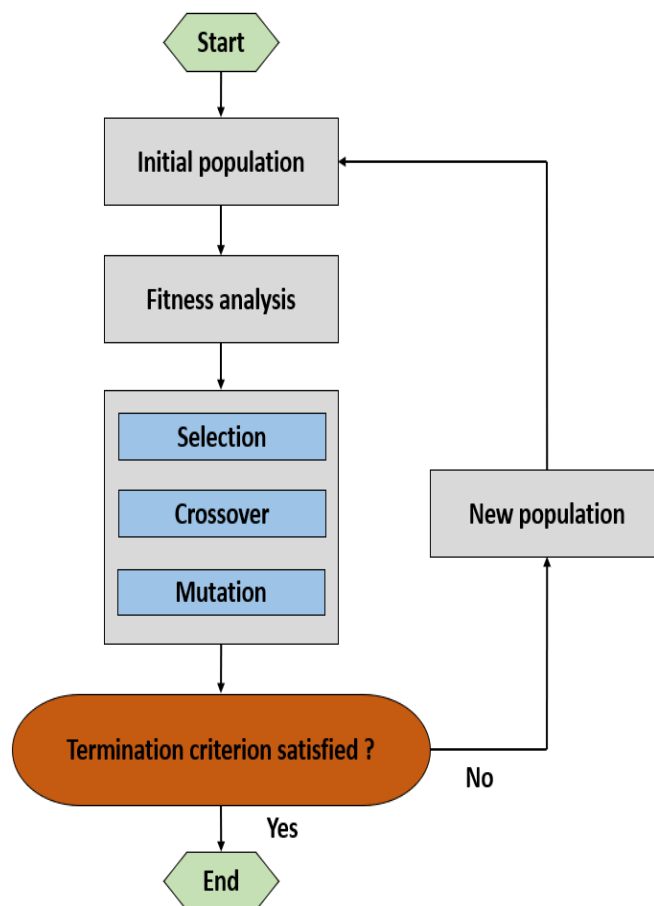
### *3.3.2 Determination of Parents and Offspring*

During the selection process, the algorithm identifies which chromosomes will serve as parents for reproduction and mating. Additionally, it determines the number of offspring that each selected chromosome will generate.

**Objective of Selection:** The primary objective of the selection process is to give preference to individuals with higher fitness levels. This is often summarized by the principle that "the better an individual's fitness, the greater its likelihood of becoming a parent [96]. Several well-known selection methods are as follows:

**Tournament Selection:** This method is widely regarded as one of the most common and efficient techniques in the field of genetic algorithms due to its simplicity and effectiveness

[97]. The tournament selection process involves randomly selecting individuals from the broader population. These selected individuals then engage in a competition to determine which one possesses the highest fitness value. The victor of this competition is chosen as a parent for the next generation. Typically, individuals compete in pairs, forming binary tournaments or "tournament size." Tournament selection ensures diversity by offering an equal opportunity to all individuals, although it may slightly slow down convergence. Notably, tournament selection is adept at utilizing computational resources, particularly when implemented in parallel. It also demonstrates resilience against domination by a few individuals, thus enhancing its robustness. Moreover, it eliminates the need for fitness scaling or sorting procedures [98].



**Figure 3.3:** Schematic representation of Genetic Algorithm

Proportional Roulette Wheel Selection: In this approach, potential solutions are depicted as segments on a roulette wheel, where the size of each segment corresponds to its fitness value. Random spinning of the wheel is then employed to select the solutions that will contribute to generating the next generation, as illustrated in Figure 3.4. Rank-based selection is a variation

of this approach where individuals are assessed based on their ranks rather than their absolute fitness values, ensuring that every individual has a chance of being chosen [99].

Rank Selection: It is employed for parent selection, involving the utilization of a ranking mechanism. Within this context, the fitness value is utilized to assign rankings to individuals within the population. The individual with the highest fitness receives the highest rank (n), while the lowest-ranked individual is assigned a rank of 1. Each chromosome's ranking is determined based on its expected value [100].

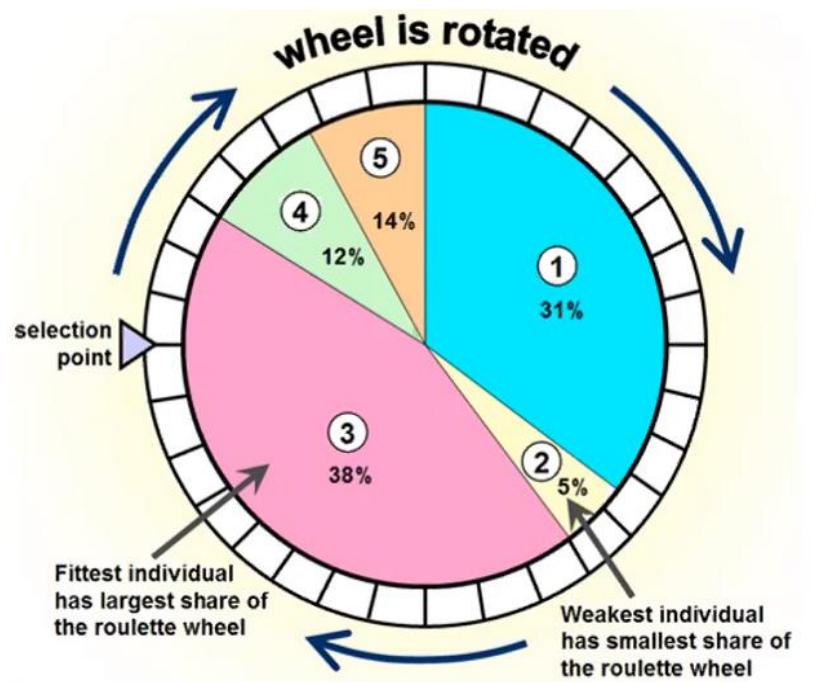
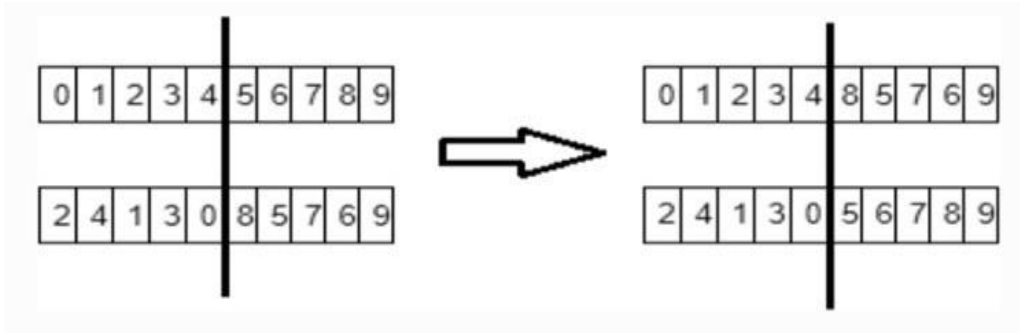


Figure 3.4: Roulette wheel selection

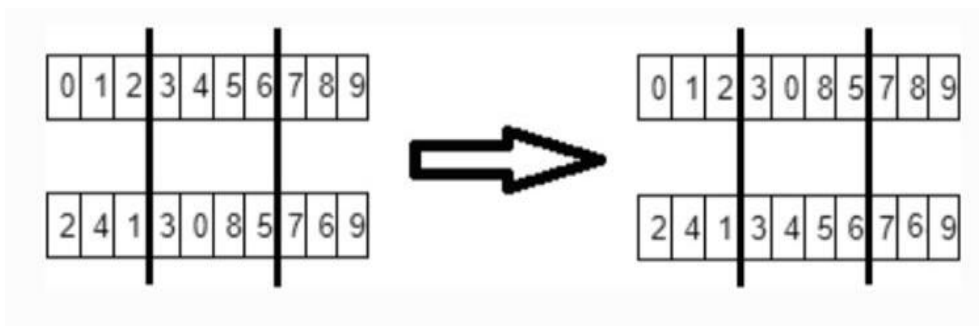
### 3.3.3 Crossover

This process involves merging the genetic data from two or more parents to generate offspring. Genetic algorithms often utilize crossover operators like single-point, double-point, and uniform. In the Single-Point Crossover method, a random crossover point is selected, and genetic data exchange between two parents occurs beyond that particular point, as illustrated in Figure 3.5 [101].



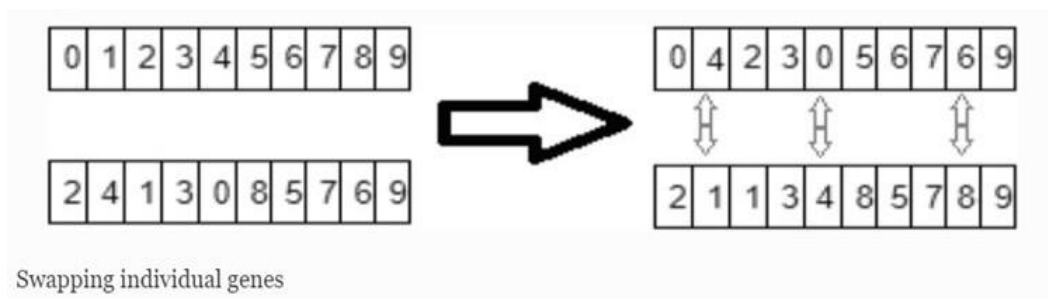
**Figure 3.5:** Single point crossover

Double Points Crossover: In this method, entails the random selection of two or more crossover points. The exchange of genetic information between parents occurs based on the segments created, as exemplified in Figure 3.6 [101].



**Figure 3.6:** Double points crossover

Uniform crossover: The parental individual cannot be divided into separate segments. Instead, each parent is considered to represent each gene independently. The choice of whether to swap a gene with its corresponding gene at the same position in another chromosome is determined through a random process, as depicted in Figure 3.7 [101].



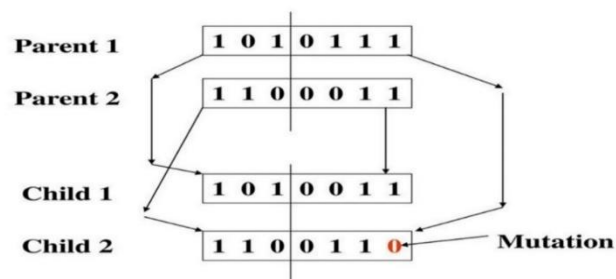
**Figure 3.7:** Uniform crossover



### 3.3.4 Mutation

Mutation plays a vital role in preserving genetic diversity from one generation to the next. During the mutation process, genes within the chromosomes undergo changes. This alteration can lead to variations in the characteristics of chromosomes inherited from their parents. Notably, the mutation process generates three additional offspring [93]. In practice, within the Genetic Algorithm (GA), this operator prevents solutions from becoming identical and enhances the likelihood of avoiding local optima. Refer to Figure 3.8 for a conceptual depiction of this operator. After the crossover (replication) stage, minor modifications in some randomly selected genes can be observed in the diagram [102].

#### Genetic Algorithm Mutation



**Figure 3.8:** After the crossover phase, the mutation operator changes one or more genes in the children's solutions.

## 3.4 Modelling Parameters and Hyperparameter tuning

Preprocessing techniques were applied to the different algorithms. As a first step, the data was standardized. As different features have value ranges which vary relative to each other, therefore feature scaling should be performed to learn a precise regressor [93]. The dataset was divided into training (90%) and testing (10%) datasets. Due to the small size of the dataset, 5-fold cross validation was utilized. This will also avoid the issue of overfitting. The various hyperparameters utilized in the models can be seen in table 3.3.

**Table 3.3:** Hyperparameters utilized in the various boosting models

<b>Algorithm</b>	<b>Parameters</b>	<b>Optimized Values</b>
<b>Xgboost</b>	max depth	2
	n_estimators	116
	learning rate	0.054
	gamma	0.064
	Iterations	123
<b>Catboost</b>	learning rate	0.039
	depth	2
	l2_leaf_reg	0.26
	num_leaves	2
<b>LightGBM</b>	max depth	4
	min_data_in_leaf	3
	num_iterations	55

As the various models have unique hyperparameters, therefore for the algorithms, ranges were specified for each and then fed to genetic algorithm for optimization. Finally, these hyperparameters were used for testing of the models. The evaluation was done using Root Mean Squared Error (RMSE) and Correlation Coefficient (CC).

**Table 3.4:** Genetic algorithm parameters used to optimize the hyperparameters of Catboost Model.

<b>Parameter</b>	<b>Value</b>
<b>Number of Generations</b>	20
<b>Population Size</b>	10
<b>Crossover Probability</b>	0.7-0.2
<b>Mutation Probability</b>	0.3-0.7
<b>Criteria</b>	Max Fitness
<b>Cross Validation</b>	5-fold
<b>Elitism</b>	FALSE

## CHAPTER 04 RESULTS AND DISCUSSION

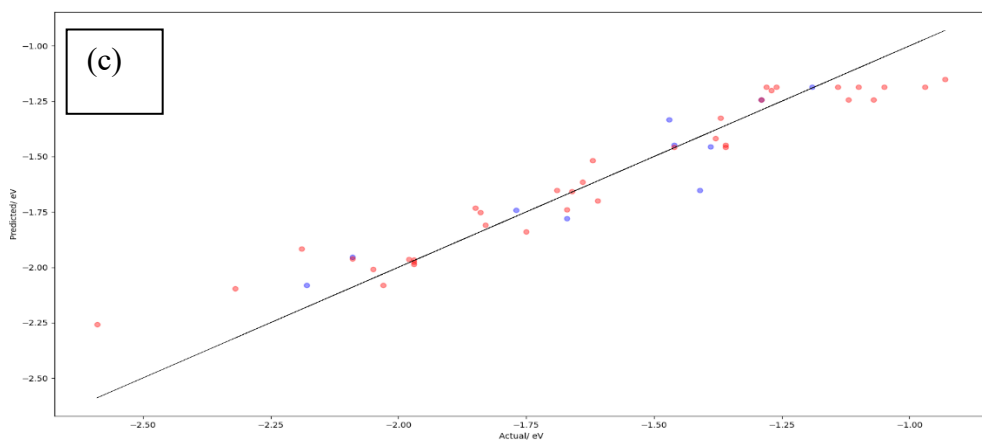
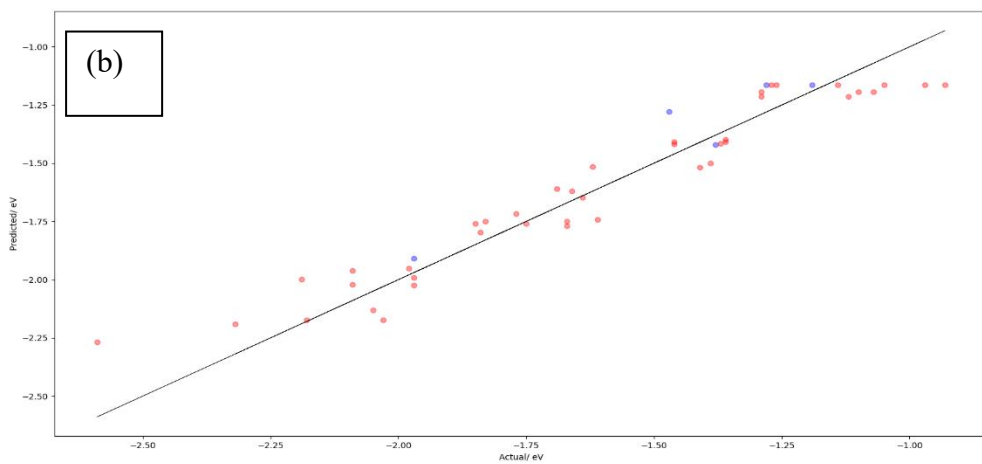
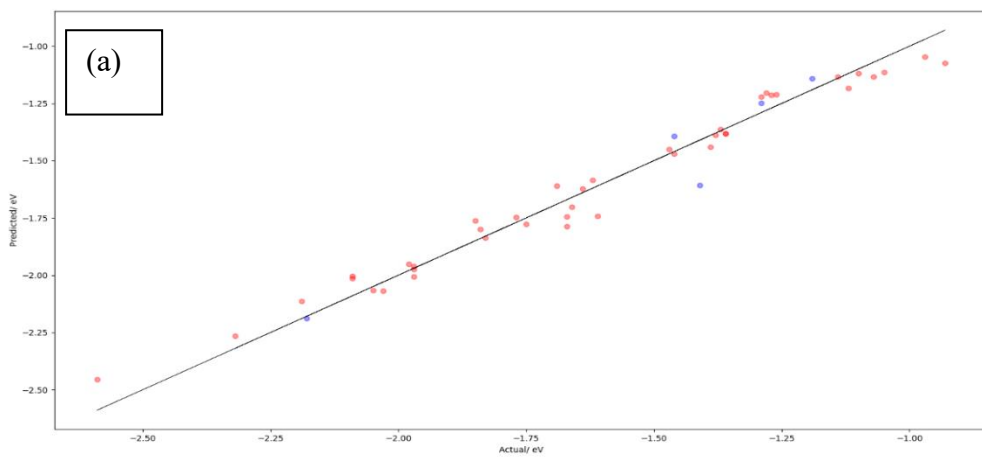
### 4.1 Models evaluation and comparison

Various machine learning (ML) models, including XGBoost, Catboost, LightGBM were trained in Python programming language in order to predict adsorption energies of CH<sub>3</sub>, CH<sub>2</sub>, CH on Cu-based alloys. Other than CH<sub>3</sub>, remaining in supplementary files.

For the purpose of assessing the performance of the ML models, the predicted adsorption energies were plotted against the actual adsorption energies as shown in fig 4.1. The line in the figure represents the actual values of the adsorption energies. Therefore, points that lie close to the line indicate enhanced and superior prediction capability. The model prediction performance for predicting adsorption energies of CH<sub>3</sub> on Cu-base alloys was evaluated using CC and RMSE as shown in table 4.1. It is evident that Catboost outperformed LightGBM and Xgboost with an RMSE of 0.09772 and CC of 96.5%. It was followed by LightGBM and Xgboost whose RMSE and CC values are 95.4%, 93.5% and 0.1071, 0.116 respectively. CatBoost has both CPU and GPU implementations. The GPU implementation allows for much faster training. It uses an efficient method called ordered boosting, which is specifically designed to deal with categorical features directly. Another advantage of this algorithm is that it uses a new schema for calculating leaf values when selecting the tree structure, which helps to reduce overfitting. It also includes built-in mechanisms to handle outliers and missing values. It is evident that incorporating boosting with optimization algorithms allows for enhanced accuracy and substantial decrease in errors such as RMSE. Rather than conducting experiments, adsorption energy of CH<sub>3</sub> and other compounds can be predicted using the various boosting algorithms which will effectively reduce the elapsed time and costs associated with experiments.

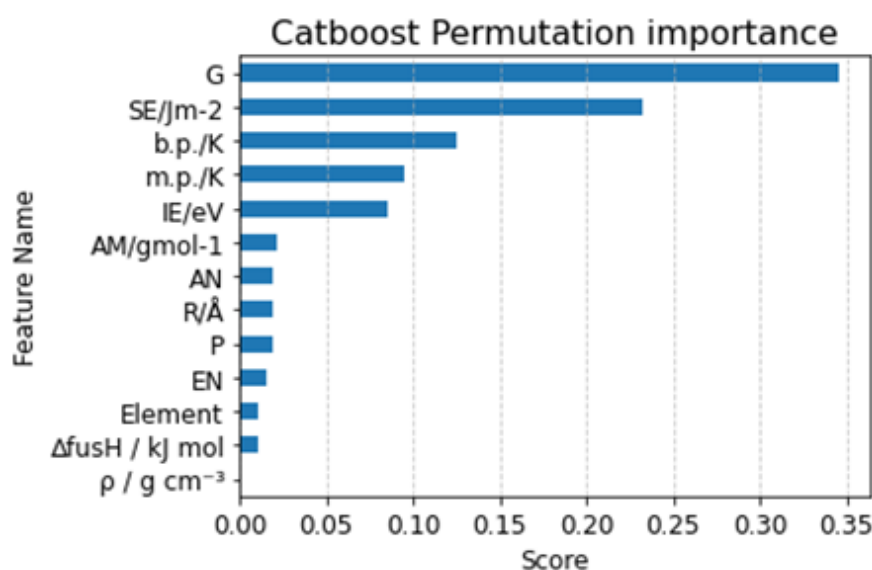
**Table 4.1:** Comparison of ML methods

Model	CC	RMSE
LGBM	95.4%	0.1071
Catboost	96.5%	0.0977
Xgboost	93.5%	0.1116



**Figure 4.1:** DFT-calculated adsorption energies of CH<sub>3</sub> on Cu-based alloys and their predicted values  
(a) Catboost (b) LightGBM (c) XGBoost

Opacity lies at the core of black box problem, therefore, as a consequence, it is difficult to interpret how they function and what actually goes on [103]. Gradient boosting methods are black box models and their lack of explainability poses a challenging task. The goal of explainability in ML is to provide humans with a thorough comprehension of a model's operation and decision-making process, without requiring them to fully comprehend every nuance of the algorithm [104]. Furthermore, it provides comprehensive understanding to the users, offering them information they need to assess if they should act and enforce a model's advice [105]. Different techniques are employed in explainable ML.



**Figure 4.2:** Permutation importance scores of the feature variables

The statistical significance and relevance of each variable to a model's performance is calculated by variable importance methods (also known as "feature importance" in the ML domain). During model construction, it is common practice to utilize variable importance approaches to evaluate whether the model is learning appropriately and what variables have the most impact on the target variable. Permutation importance assesses a black-box model's prediction performance after shuffling a single variable. Permutation importance retests the model with variable values repeatedly which are random and different. If rearranging a variable's values doesn't affect prediction accuracy, the variable's contribution isn't significant to the model's output. Adversely, if randomizing a variable's values reduces the ability to generalize, that variable is more essential to the model's predictions [106].

Permutation importance scores in the case of CH<sub>3</sub> can be seen in fig 4.2. It is evident that the group number has the highest impact with a score of 0.346. It is followed by surface energy (0.23) and boiling point (0.12).

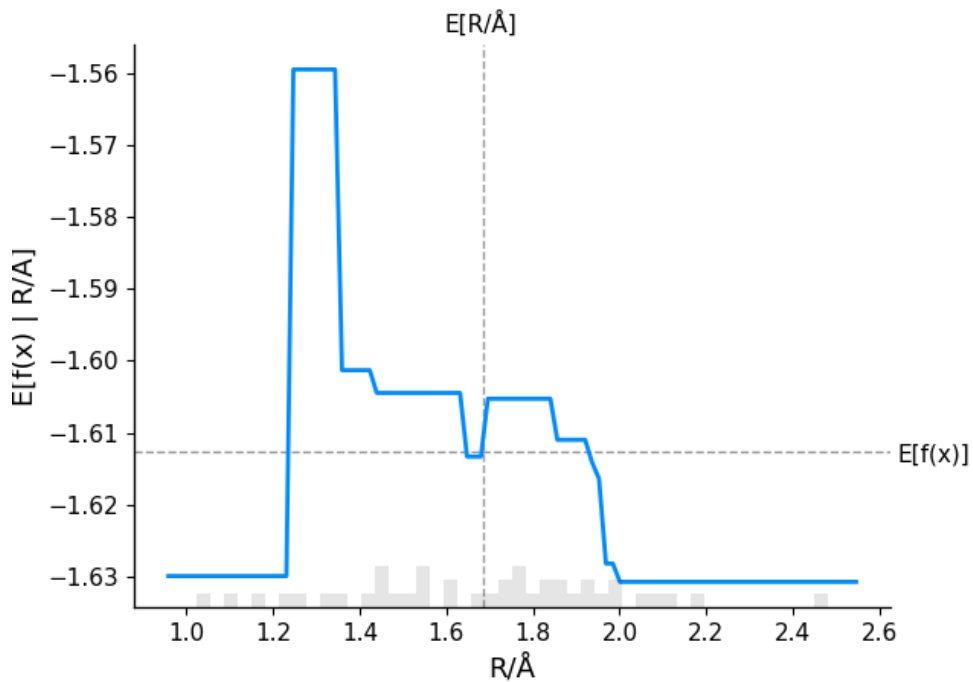
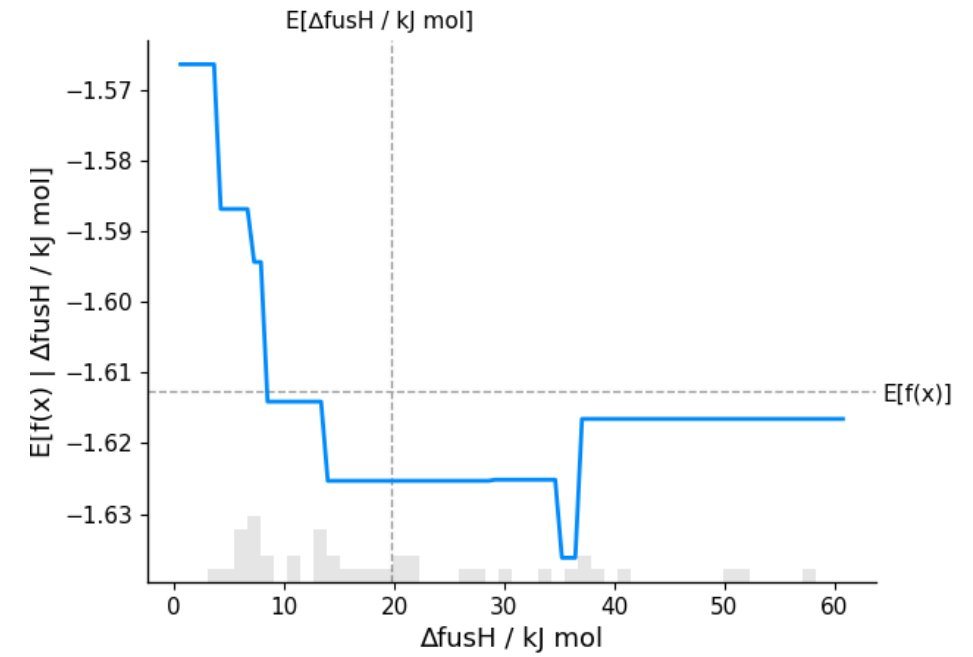
Partial dependency is the response of the target function when one or two features are varied. Partial dependence plots (PDPs) show how a model's projected result shifts in response to a change in a set of features. PDPs have a significant benefit over other explainability methods since they can show the interaction between the variables and the predictions, even if the relationship is nonlinear. Fig.4.3 shows the PDPs.

The PDP that relates group number to the output shows a continuous and steep decline in adsorption energy. This can be attributed to the acid-base properties of the catalyst. As we go across the periodic table, the basicity of the element decreases. Dopants with high electropositivity can modify the morphology and create defects in the structure of the host oxides [43]. Zavyalova et al. research on an abundance of data on OCM catalysts showed that high basicity is paramount for improved C<sub>2</sub> selectivity as it directly influences the bond formed between the anions and cations [107]. Smaller atomic size and higher electropositivity contribute to basicity.

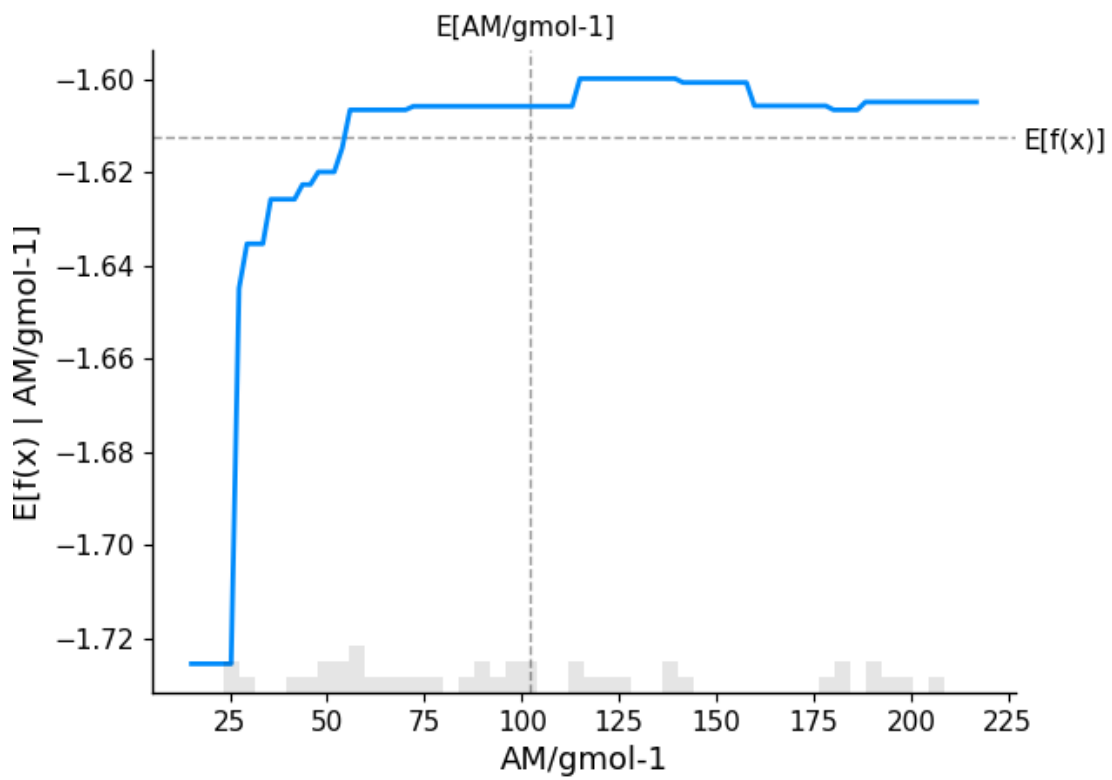
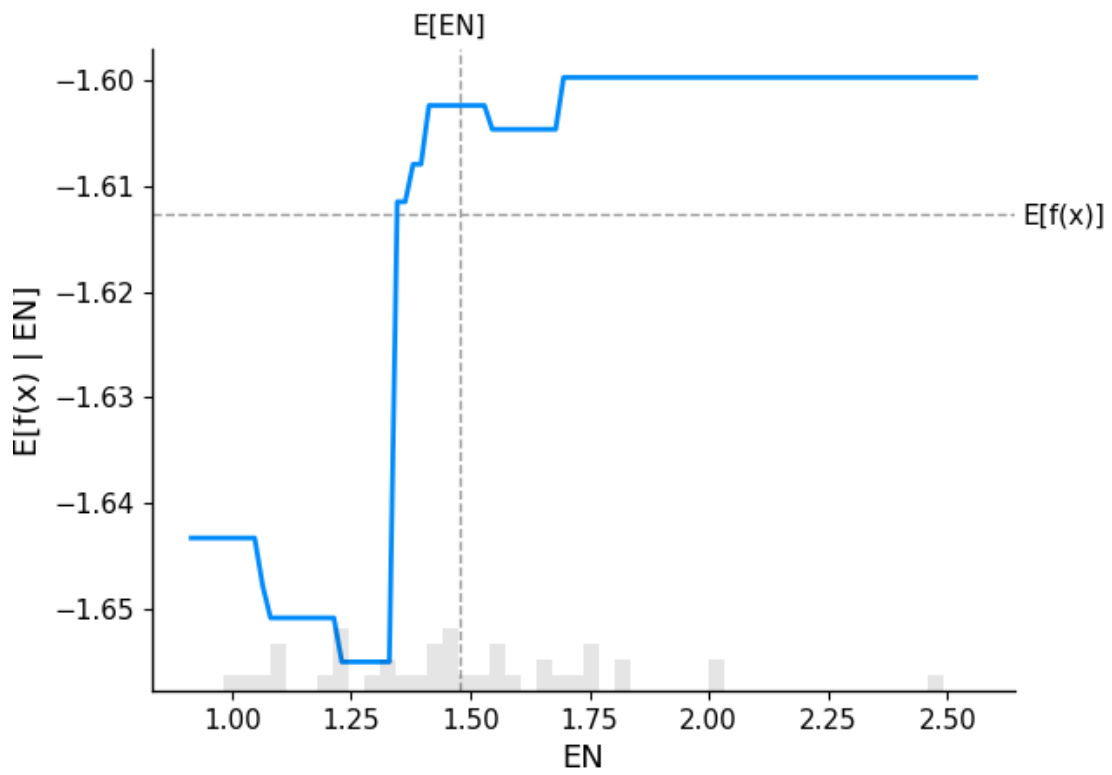
With respect to the atomic radius, the PDP show reveals a fairly constant value of the output till 1.6 R/Å and then it falls and rises again. At 1.9 R/Å, a sharp downward descent continues till 2.0 R/Å where it eventually becomes constant. In the case of AN, there is sudden increase in the adsorption energy at 12 atomic number. Then it rises gradually and becomes fairly constant at 43 AN. The atomic mass PDP shows a rapid rise at approximately 25 g/mol<sup>-1</sup>. Beyond this, the output value remains between -1.61 and -1.62. The change in adsorption energy with respect to atomic radius and AN can be explained by surface reducibility which is a critical factor in OCM. The presence of selective lattice oxygen species is contingent upon both the degree to which the oxide catalyst can be reduced and the ease with which this reduction process occur. Kumar et al. discovered that improved OCM activity takes place at higher surface reducibility [43].

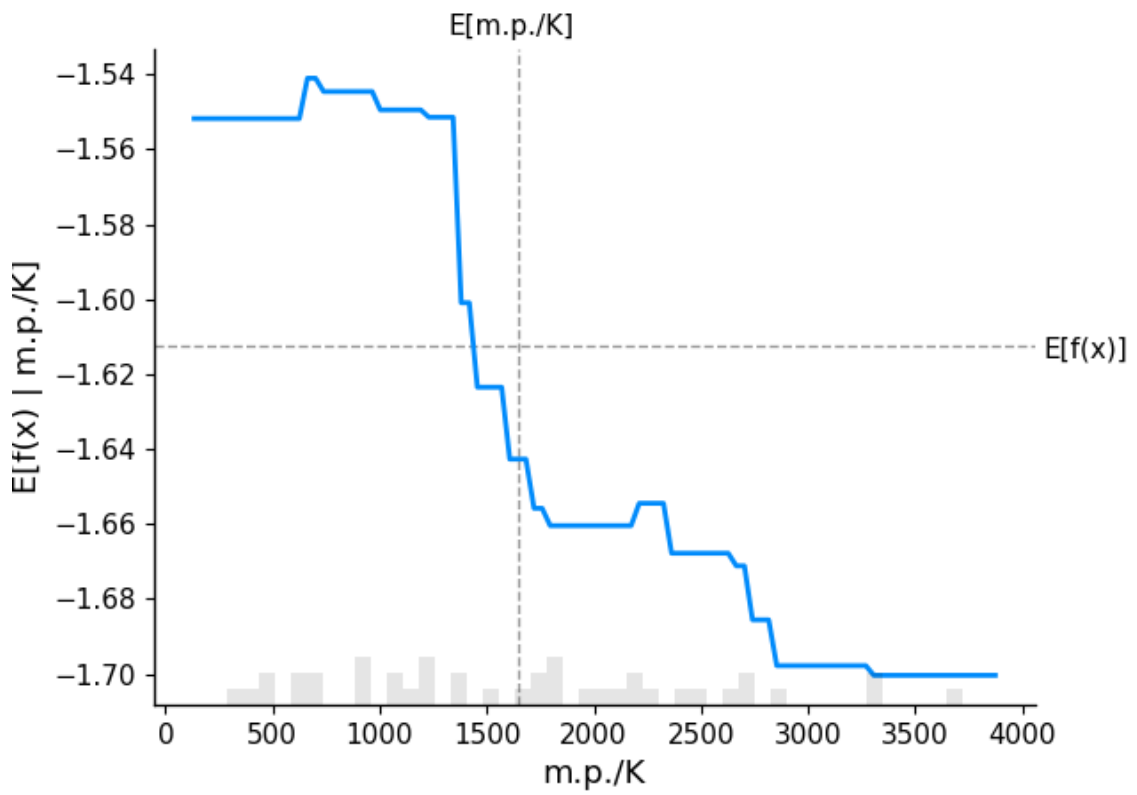
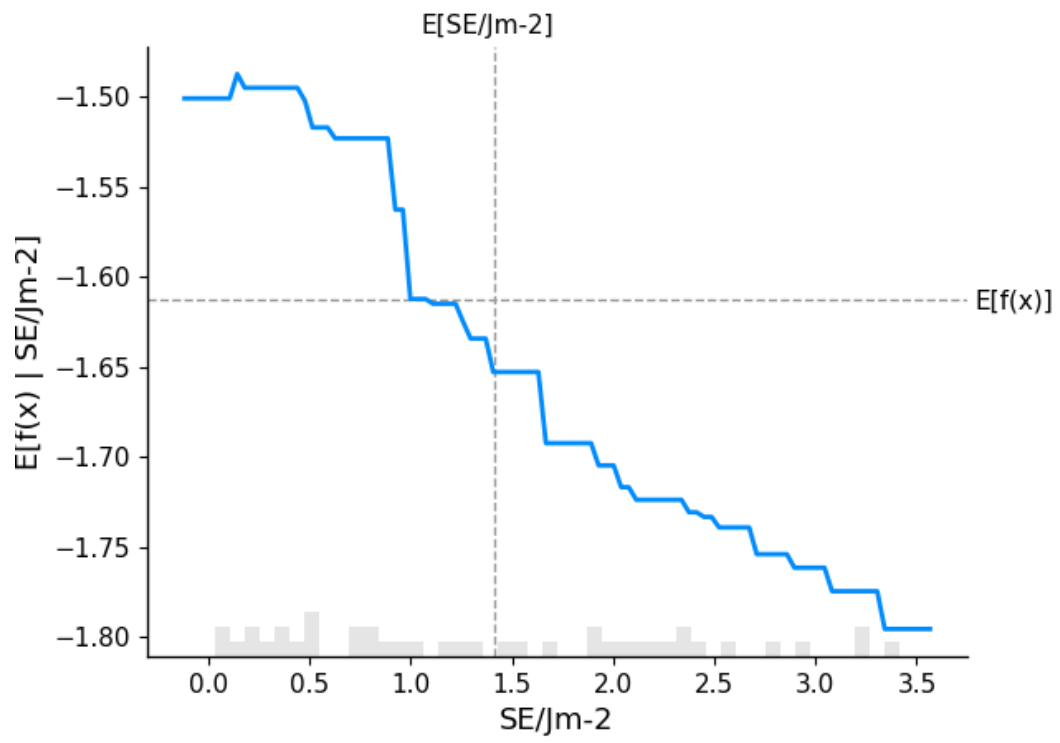
The electronegativity PDP shows a downward trajectory till 1.25 and then a steep vertical ascent till 1.75. Beyond this, it becomes flat. This was expected as weak electronegative cation will result in a high partial charge on oxygen and therefore strong basic character. Therefore, the probability of finding oxygen vacancies containing at least one electron increases with

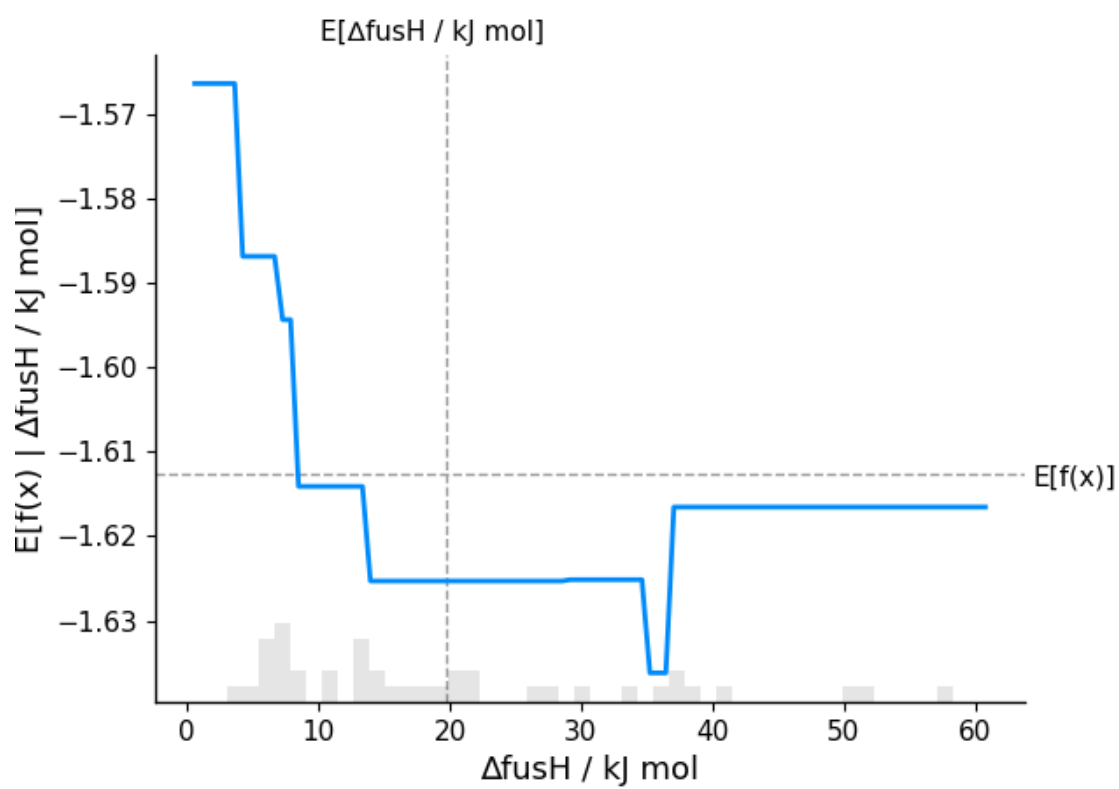
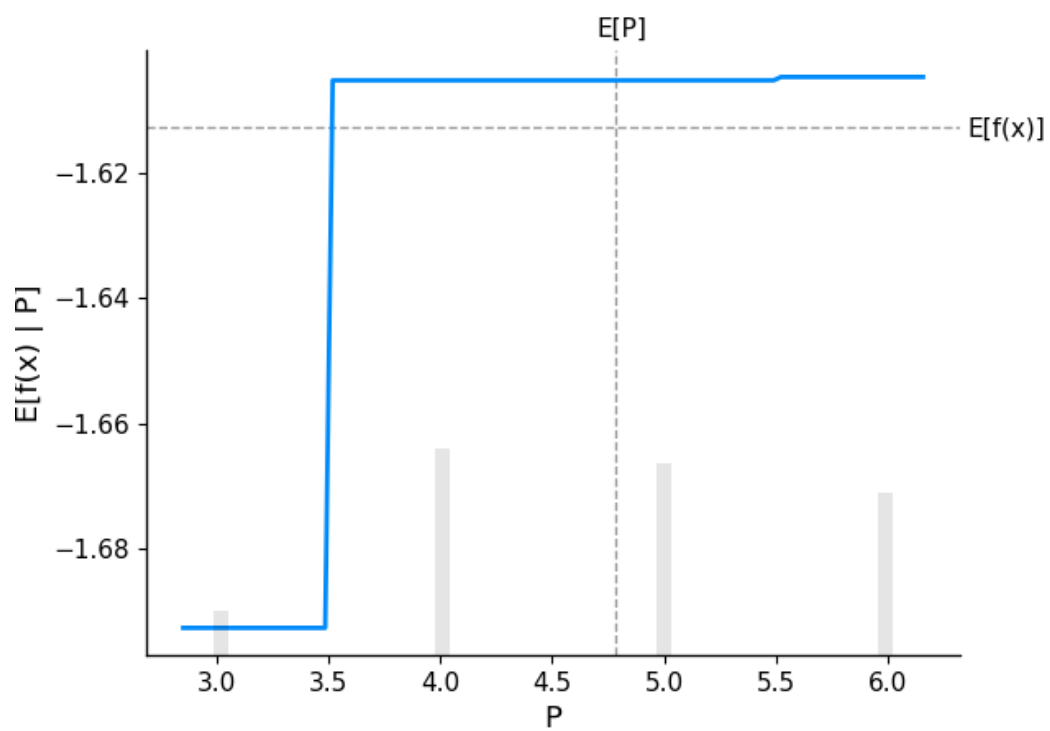
increasing electropositive character of the cation, creating an adsorption site for gaseous oxygen. The higher the difference in electronegativity between cation and oxygen, the more basic is the oxide [108]. The PDP of surface energy show a consistent increase in adsorption energy as the surface energy of the doped element increases.

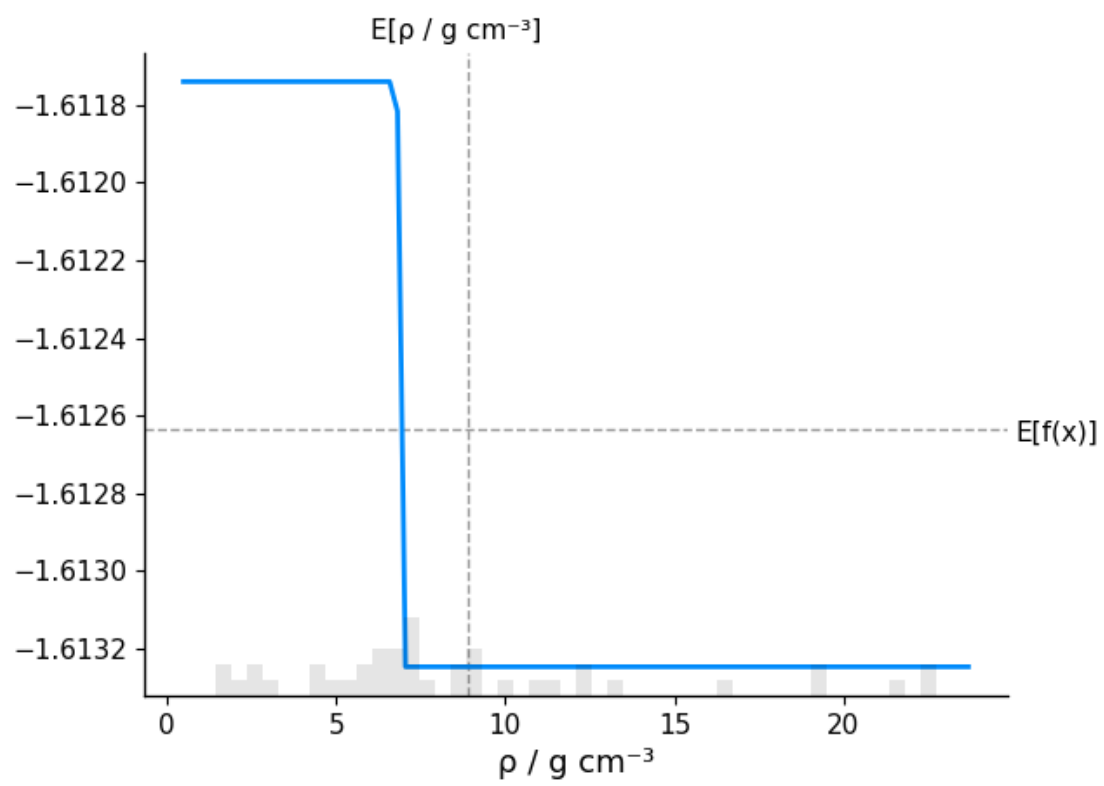
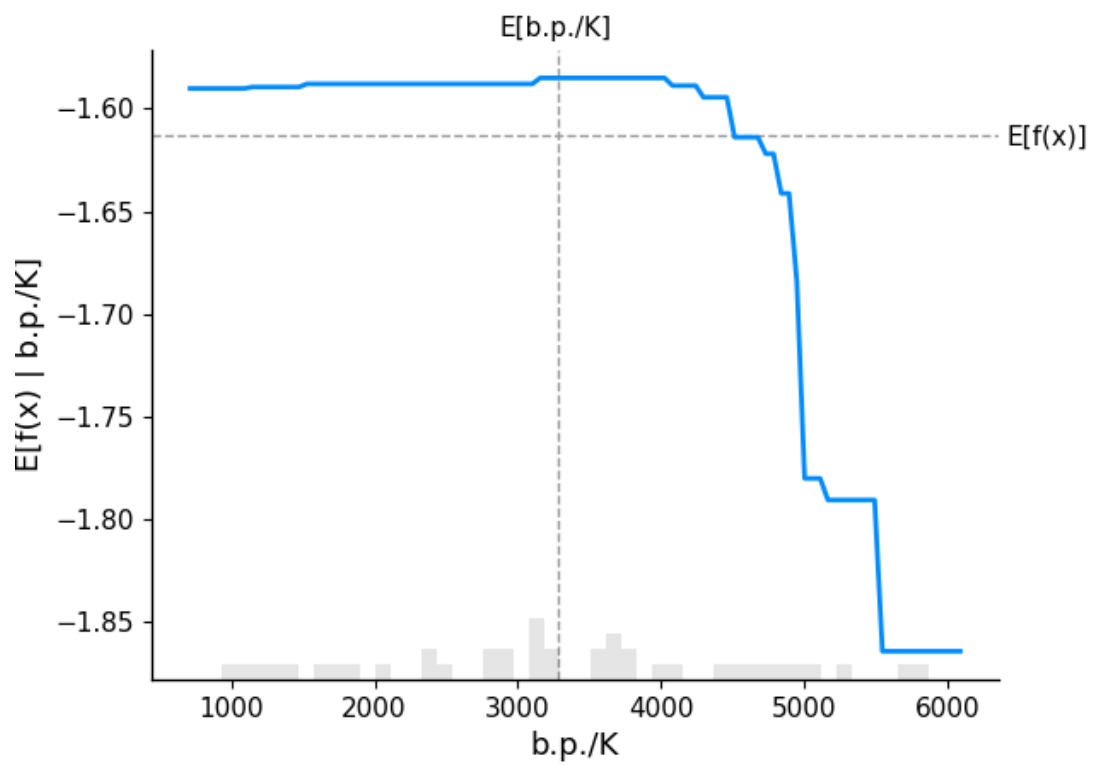


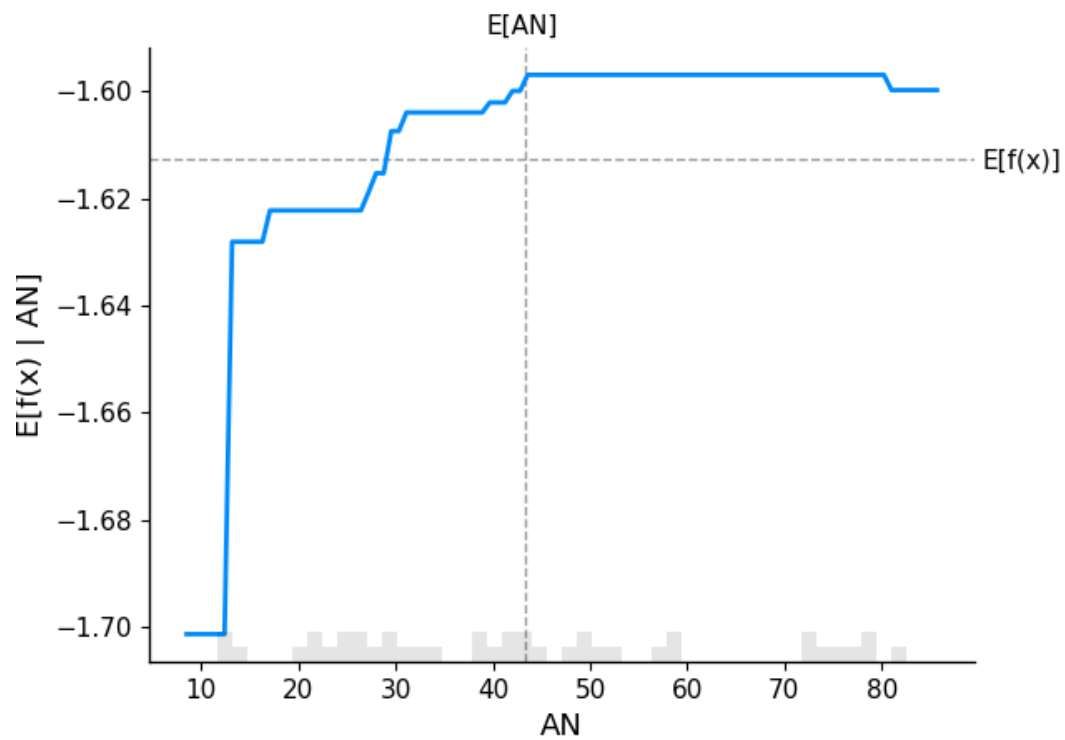
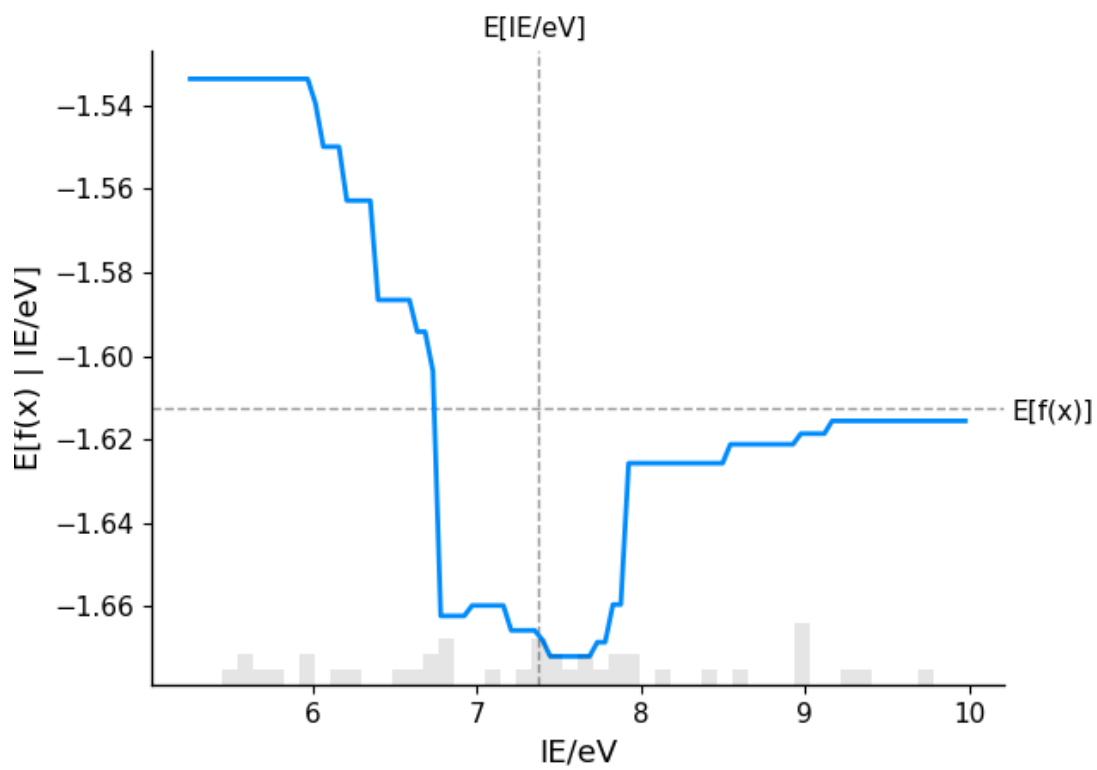












**Figure 4.3:** Partial Dependence Plots of the feature variables to the prediction of the adsorption energies

This is anticipated because higher surface energy is a precursor to stronger adsorption. As high surface energy is conducive for wettability, hence the gas molecules will spread on the catalyst surface, adhering to the surface resulting in improved adsorption [109]. The adsorption of methane on the catalyst surface is a critical, rate-determining step which is activated by binding to the catalyst's surface. Furthermore, adsorption of oxygen will be facilitated by it leading to improved  $\text{CH}_3$  radical formation. The decrease in the overall energy of the system leads to better adsorption.

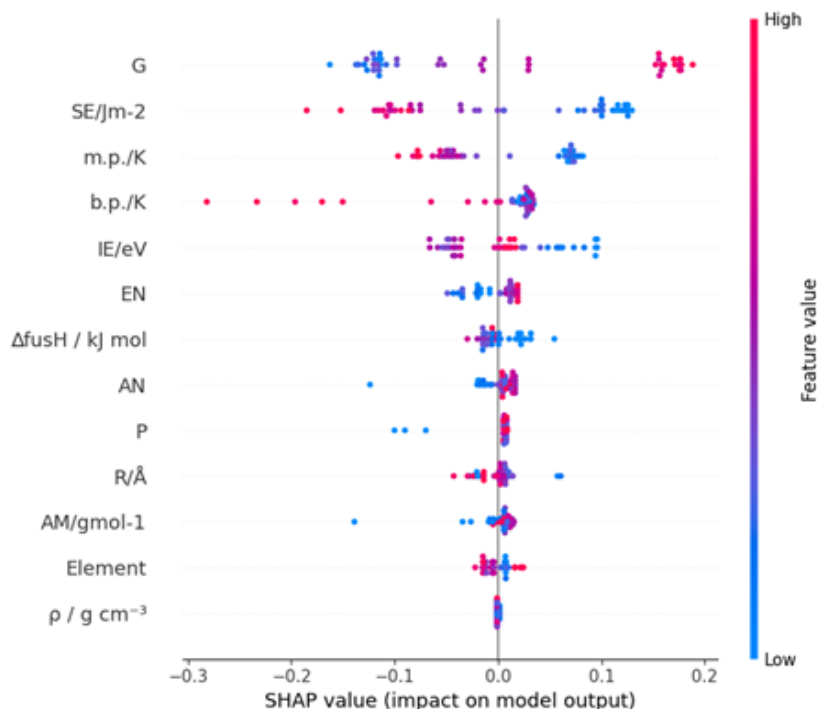
The PDP of ionization energy shows better adsorption at low ionization energies. As it is apparent that metal with low ionization energies are more beneficial for forming reactive species. High ionization energy will inhibit electron transfer affecting its interplay with adsorbates. Active sites that activates and cleaves C-H bond are reactive oxygen species present on the catalyst surface. These oxygen species include chemisorbed oxygen ( $\text{O}^{-1}$ ), dissociative adsorbed oxygen ( $\text{O}^{-1}$ ), adsorbed oxygen ions ( $\text{O}^{-2}$ ), and lattice oxygen ( $\text{O}^{-2}$ ) [69, 110]. Both strongly and weakly bounded oxygen species are present in the catalyst. However, a study conducted by Gordienko et al revealed that weakly bounded oxygen species were responsible for the activity and stability of OCM [43]. The  $\text{CH}_4$  related species will readily react with oxygen too. Therefore, low ionization energy is paramount to increase in the performance of the process. However, after 7.5 eV, it starts climbing again.

The PDP of melting point shows an exponential decrease after approximately 1250K. It reaches its lowest point at 3300K. The boiling point chart is similar to that of melting point as initially it remains constant and then a rapid decrease starts at around 4000K and stops at 5500K. As the enthalpy of fusion increases the adsorption energy falls steeply and reaches its lowest point at  $28 \text{ KJ/mol}^{-1}$ . The PDP of density shows an interesting relation with the output. At  $2.5 \text{ g/cm}^{-3}$  it starts a nearly vertical climb. Moving on, it starts falling and becomes constant at  $7.5 \text{ g/cm}^{-3}$ .

The relationship between features and the target variable can be further explored using two-way partial dependence plot which will allow us to analyze how the adsorption energy varies with the operational variables. We can extract complex or convoluted patterns from the data. Furthermore, it will be crucial during the optimization of the process.

A game theory-based extension of Shapely values, SHAP (SHapley Additive exPlanations) is a pragmatic approach for explaining the outcomes of machine learning predictions. Shapley

values provide an approach that is both mathematically just and distinctively different for attributing the payoff of a cooperative game to the individuals who participated in the game [111]. In order to successfully apply the Shapley values approach to ML models, the most important step is to establish a cooperative game in which the players represent the input parameters and the outcome is the model prediction [112]. In the Figure 4.4, a beeswarm summary plot is visible which shows feature importance and the effect of the various features on the outcome of prediction. On the y-axis, the characteristics are ordered from highest to lowest based on the aggregate SHAP value magnitudes of all of the occurrences. You can also view the distribution of the affects that each feature has on the model's predictions along the x-axis for each feature. The values of the parameters are signified by the colors of the dots as follows: High in red, and low in blue. However, as higher adsorption energies tend be more negative, therefore red will be interpreted as lower while blue occurrences are to be interpreted as higher. For example, it is clearly visible in figure 4.4 surface energy ranks second in the importance of features and it relevance to the model predictions. Values of higher magnitude of surface energy (blue dots) contribute in positive manner to the output while the lower values (red dots) contribute negatively. This is accurate as adsorption has a direct relation with surface energy.



**Figure 4.4:** SHAP beeswarm summary plot

## CHAPTER 5: CONCLUSIONS AND RECOMMENDATIONS

In this research, three distinct boosting machine learning models were employed. These models were fine-tuned through the application of a genetic algorithm, aiming to optimize and forecast the adsorption energy of CH<sub>4</sub>-related species namely CH<sub>3</sub>. Catboost model outperformed other models with CC and RMSE of 96.5% and 0.0977 respectively. By utilizing advanced and state-of-the-art machine learning algorithms, the accuracy and precision of the results can be enhanced with low computational cost. The results were interpreted using permutation importance, partial dependence plots and bee swarm plot, valuable methods for gaining insights into individual feature significance and illustrating variable-outcome relationships. Analysis of the features revealed that group number and surface energy were the highest contributors to predicting the adsorption energies. By leveraging machine learning models, one can potentially enhance the understanding of adsorption processes on catalyst surfaces, which is crucial for optimizing OCM reactions. Accurate predictions of adsorption energies can contribute in identifying optimal catalyst materials with favorable adsorption properties, leading to improved catalytic performance in OCM reactions. Predicting adsorption energies helps in understanding the thermodynamics of the adsorption process, guiding the selection of reaction pathways that lead to higher yields of desired products, such as ethylene and ethane. By accurately predicting adsorption energies, machine learning models can assist in designing catalysts that efficiently adsorb and activate methane and suppress undesired over reactions of other species, enhancing the overall efficiency of the OCM process. Predictive models can potentially reduce the need for extensive experimental testing by providing valuable insights into the adsorption behavior of different catalysts, allowing researchers to focus on the most promising candidates. Machine learning can facilitate the screening of a large number of potential catalysts, accelerating the discovery of materials with superior adsorption characteristics for OCM applications. Understanding adsorption energies provides insights into the underlying mechanisms of OCM reactions, helping researchers optimize conditions for higher selectivity and yield of desired products.



## REFERENCES

- [1]. Tang, P., et al., *Methane activation: the past and future*. Energy & Environmental Science, 2014. **7**(8): p. 2580-2591.
- [2]. Cornot-Gandolphe, S., *Changes in world gas reserves and resources*. Energy exploration & exploitation, 1995. **13**(1): p. 3-16.
- [3]. Holmen, A., *Direct conversion of methane to fuels and chemicals*. Catalysis today, 2009. **142**(1-2): p. 2-8.
- [4]. Rostrup-Nielsen, J.R., *Catalysis science and technology*. by Anderson, JR, Boudart, M, 1984. **5**: p. 1.
- [5]. Li, G., et al., *Oxidative dehydrogenation of light alkanes with carbon dioxide*. Green Chemistry, 2021. **23**(2): p. 689-707.
- [6]. Zhang, H., Z. Sun, and Y.H. Hu, *Steam reforming of methane: Current states of catalyst design and process upgrading*. Renewable and Sustainable Energy Reviews, 2021. **149**: p. 111330.
- [7]. Xu, J. and G.F. Froment, *Methane steam reforming, methanation and water-gas shift: I. Intrinsic kinetics*. AIChE journal, 1989. **35**(1): p. 88-96.
- [8]. Lunsford, J.H., *Catalytic conversion of methane to more useful chemicals and fuels: a challenge for the 21st century*. Catalysis today, 2000. **63**(2-4): p. 165-174.
- [9]. LeValley, T.L., A.R. Richard, and M. Fan, *The progress in water gas shift and steam reforming hydrogen production technologies—A review*. International Journal of Hydrogen Energy, 2014. **39**(30): p. 16983-17000.
- [10]. Parthasarathy, P. and K.S. Narayanan, *Hydrogen production from steam gasification of biomass: influence of process parameters on hydrogen yield—a review*. Renewable energy, 2014. **66**: p. 570-579.
- [11]. Tang, H., J. Yao, and Y. Zhu, *Recent developments and future prospects for zinc-ion hybrid capacitors: a review*. Advanced Energy Materials, 2021. **11**(14): p. 2003994.
- [12]. Gambo, Y., et al., *Recent advances and future prospect in catalysts for oxidative coupling of methane to ethylene: A review*. Journal of industrial and engineering chemistry, 2018. **59**: p. 218-229.
- [13]. Lunsford, J.H., *The catalytic oxidative coupling of methane*. Angewandte Chemie International Edition in English, 1995. **34**(9): p. 970-980.
- [14]. Blanksby, S.J. and G.B. Ellison, *Bond dissociation energies of organic molecules*. Accounts of chemical research, 2003. **36**(4): p. 255-263.

- [15]. Takahashi, K., et al., *Unveiling hidden catalysts for the oxidative coupling of methane based on combining machine learning with literature data*. ChemCatChem, 2018. **10**(15): p. 3223-3228.
- [16]. Arutyunov, V.S., et al., *On the role of a catalyst in high-temperature reactions of methane oxidation*. KINETICS AND CATALYSIS C/C OF KINETIKA I KATALIZ, 1999. **40**: p. 382-387.
- [17]. Toyao, T., et al., *Toward effective utilization of methane: machine learning prediction of adsorption energies on metal alloys*. The Journal of Physical Chemistry C, 2018. **122**(15): p. 8315-8326.
- [18]. Varghese, J.J., Q.T. Trinh, and S.H. Mushrif, *Insights into the synergistic role of metal–lattice oxygen site pairs in four-centered C–H bond activation of methane: the case of CuO*. Catalysis Science & Technology, 2016. **6**(11): p. 3984-3996.
- [19]. Merino, N.A., et al., *Lal–xCaxCoO3 perovskite-type oxides: Identification of the surface oxygen species by XPS*. Applied Surface Science, 2006. **253**(3): p. 1489-1493.
- [20]. Dąbrowski, A., *Adsorption—from theory to practice*. Advances in colloid and interface science, 2001. **93**(1-3): p. 135-224.
- [21]. Ağbulut, Ü., et al., *Performance assessment of a V-trough photovoltaic system and prediction of power output with different machine learning algorithms*. Journal of Cleaner Production, 2020. **268**: p. 122269.
- [22]. Bakay, M.S. and Ü. Ağbulut, *Electricity production based forecasting of greenhouse gas emissions in Turkey with deep learning, support vector machine and artificial neural network algorithms*. Journal of Cleaner Production, 2021. **285**: p. 125324.
- [23]. Hamrani, A., A. Akbarzadeh, and C.A. Madramootoo, *Machine learning for predicting greenhouse gas emissions from agricultural soils*. Science of The Total Environment, 2020. **741**: p. 140338.
- [24]. Ahmad, I., et al., *Machine learning applications in biofuels' life cycle: Soil, feedstock, production, consumption, and emissions*. Energies, 2021. **14**(16): p. 5072.
- [25]. Xie, Y., et al., *The promise of implementing machine learning in earthquake engineering: A state-of-the-art review*. Earthquake Spectra, 2020. **36**(4): p. 1769-1801.
- [26]. Jadoon, U.K., et al., *An intelligent sensing system for estimation of efficiency of carbon-capturing unit in a cement plant*. Journal of Cleaner Production, 2022. **377**: p. 134359.
- [27]. Ahmad, I., et al., *Prediction of molten steel temperature in steel making process with uncertainty by integrating gray-box model and bootstrap filter*. Journal of Chemical Engineering of Japan, 2014. **47**(11): p. 827-834.

- [28]. Kumar, A., R. Shankar, and L.S. Thakur, *A big data driven sustainable manufacturing framework for condition-based maintenance prediction*. Journal of computational science, 2018. **27**: p. 428-439.
- [29]. Ahmad, I., et al., *Data-based fault diagnosis of power cable system: comparative study of k-NN, ANN, random forest, and CART*. IFAC Proceedings Volumes, 2011. **44**(1): p. 12880-12885.
- [30]. Ahmad, I., et al., *Data-based sensing and stochastic analysis of biodiesel production process*. Energies, 2018. **12**(1): p. 63.
- [31]. Ullah, H., et al., *Optimization based comparative study of machine learning methods for the prediction of bio-oil produced from microalgae via pyrolysis*. Journal of Analytical and Applied Pyrolysis, 2023. **170**: p. 105879.
- [32]. Shahzad, J. and I. Ahmad, *Estimation of cutpoint temperature under uncertain feed composition and process conditions using artificial intelligence methods*, in *Computer Aided Chemical Engineering*. 2021, Elsevier. p. 971-976.
- [33]. Ahmad, I., et al., *Gray-box modeling for prediction and control of molten steel temperature in tundish*. Journal of Process Control, 2014. **24**(4): p. 375-382.
- [34]. Ge, Z., et al., *Data mining and analytics in the process industry: The role of machine learning*. Ieee Access, 2017. **5**: p. 20590-20616.
- [35]. Gupta, R., et al., *Artificial intelligence to deep learning: machine intelligence approach for drug discovery*. Molecular diversity, 2021. **25**: p. 1315-1360.
- [36]. De, R., et al., *System and method for industrial process automation controller farm with flexible redundancy schema and dynamic resource management through machine learning*. 2019, Google Patents.
- [37]. Keliris, A., et al. *Machine learning-based defense against process-aware attacks on industrial control systems*. IEEE.
- [38]. Weichert, D., et al., *A review of machine learning for the optimization of production processes*. The International Journal of Advanced Manufacturing Technology, 2019. **104**(5-8): p. 1889-1902.
- [39]. Lahdhiri, H., et al., *Supervised process monitoring and fault diagnosis based on machine learning methods*. The International Journal of Advanced Manufacturing Technology, 2019. **102**: p. 2321-2337.
- [40]. Bhanu, K.N., H.J. Jasmine, and H.S. Mahadevaswamy. *Machine learning implementation in IoT based intelligent system for agriculture*. IEEE.

- [41]. Rymarczyk, T., et al., *Logistic regression for machine learning in process tomography*. Sensors, 2019. **19**(15): p. 3400.
- [42]. Alotaibi, F.S., *Implementation of machine learning model to predict heart failure disease*. International Journal of Advanced Computer Science and Applications, 2019. **10**(6).
- [43]. Zavyalova, U., et al., *Statistical analysis of past catalytic data on oxidative methane coupling for new insights into the composition of high-performance catalysts*. ChemCatChem, 2011. **3**(12): p. 1935-1947.
- [44]. Song, J., et al., *Monodisperse Sr–La 2 O 3 hybrid nanofibers for oxidative coupling of methane to synthesize C 2 hydrocarbons*. Nanoscale, 2015. **7**(6): p. 2260-2264.
- [45]. Hoek, A. and L.M. Kersten, *The Shell Middle Distillate Synthesis process: technology, products and perspective*. Studies in surface science and catalysis, 2004: p. 25-30.
- [46]. Suarez, A.I.O., E.J.M. SzØcsØnyi, and J. Hensen, *Ruiz-Martínez, EA Pidko, J. Gascon*, ACS Catal, 2016. **6**: p. 2965-2981.
- [47]. Gu, B., et al., *Effects of the promotion with bismuth and lead on direct synthesis of light olefins from syngas over carbon nanotube supported iron catalysts*. Applied Catalysis B: Environmental, 2018. **234**: p. 153-166.
- [48]. Meng, F., et al., *Highly active ternary oxide ZrCeZnOx combined with SAPO-34 zeolite for direct conversion of syngas into light olefins*. Catalysis Today, 2021. **368**: p. 118-125.
- [49]. Guo, X., et al., *Direct, nonoxidative conversion of methane to ethylene, aromatics, and hydrogen*. science, 2014. **344**(6184): p. 616-619.
- [50]. Keller, G.E. and M.M. Bhasin, *Synthesis of ethylene via oxidative coupling of methane: I. Determination of active catalysts*. Journal of Catalysis, 1982. **73**(1): p. 9-19.
- [51]. Gao, Z., J. Zhang, and R. Wang, *Formation of hydrogen in oxidative coupling of methane over BaCO3 and MgO catalysts*. Journal of natural gas chemistry, 2008. **17**(3): p. 238-241.
- [52]. Arndt, S., et al., *A critical assessment of Li/MgO-based catalysts for the oxidative coupling of methane*. Catalysis Reviews, 2011. **53**(4): p. 424-514.
- [53]. An, B.-I., et al., *Activation of methane to C2 hydrocarbons over unpromoted calcium oxide catalysts*. Bulletin of the Korean Chemical Society, 2007. **28**(6): p. 1049-1052.
- [54]. Otsuka, K., K. Jinno, and A. Morikawa, *Active and selective catalysts for the synthesis of C2H4 and C2H6 via oxidative coupling of methane*. Journal of Catalysis, 1986. **100**(2): p. 353-359.

- [55]. Hou, S., Y. Kou, and L. Liu, *Li-ZnO/La<sub>2</sub>O<sub>3</sub> Catalyst for the Oxidative Coupling of Methane at Low Temperature*. ACTA PHYSICOCHEMICA SINICA, 2006. **22**(8): p. 1040.
- [56]. Li, Z., et al., *Fast Optimization of LiMgMnO<sub>x</sub>/La<sub>2</sub>O<sub>3</sub> Catalysts for the Oxidative Coupling of Methane*. ACS Combinatorial Science, 2017. **19**(1): p. 15-24.
- [57]. Schucker, R.C., et al., *The effect of strontium content on the activity and selectivity of Sr-doped La<sub>2</sub>O<sub>3</sub> catalysts in oxidative coupling of methane*. Applied Catalysis A: General, 2020. **607**: p. 117827.
- [58]. Wu, J., et al., *Mechanistic study of oxidative coupling of methane over Mn<sub>2</sub>O<sub>3</sub>□Na<sub>2</sub>WO<sub>4</sub>SiO<sub>2</sub> catalyst*. Applied Catalysis A: General, 1995. **124**(1): p. 9-18.
- [59]. Ji, S.-f., et al., *The relationship between the structure and the performance of Na-W-Mn/SiO<sub>2</sub> catalysts for the oxidative coupling of methane*. Applied Catalysis A: General, 2002. **225**(1-2): p. 271-284.
- [60]. Zhu, Q., et al., *Sulfur as a selective 'soft' oxidant for catalytic methane conversion probed by experiment and theory*. Nature chemistry, 2013. **5**(2): p. 104-109.
- [61]. Shen, Q., et al., *Single chromium atoms supported on titanium dioxide nanoparticles for synergic catalytic methane conversion under mild conditions*. Angewandte Chemie International Edition, 2020. **59**(3): p. 1216-1219.
- [62]. Chen, S. and X. Ma, *The role of oxygen species in the selective oxidation of methanol to dimethoxymethane over VO<sub>x</sub>/TS-1 catalyst*. Journal of Industrial and Engineering Chemistry, 2017. **45**: p. 296-300.
- [63]. Gordienko, Y., et al., *Oxygen availability and catalytic performance of NaWMn/SiO<sub>2</sub> mixed oxide and its components in oxidative coupling of methane*. Catalysis today, 2016. **278**: p. 127-134.
- [64]. Kim, I., et al., *Selective oxygen species for the oxidative coupling of methane*. Molecular Catalysis, 2017. **435**: p. 13-23.
- [65]. Cheng, Z., et al., *Methane adsorption and dissociation on iron oxide oxygen carriers: the role of oxygen vacancies*. Physical Chemistry Chemical Physics, 2016. **18**(24): p. 16423-16435.
- [66]. Cheng, Z., et al., *Oxygen vacancy promoted methane partial oxidation over iron oxide oxygen carriers in the chemical looping process*. Physical Chemistry Chemical Physics, 2016. **18**(47): p. 32418-32428.

- [67]. Chu, C., et al., *Correlation between the acid–base properties of the La<sub>2</sub>O<sub>3</sub> catalyst and its methane reactivity*. Physical Chemistry Chemical Physics, 2016. **18**(24): p. 16509-16517.
- [68]. Ferrando, R., et al., *Interface-stabilized phases of metal-on-oxide nanodots*. ACS nano, 2008. **2**(9): p. 1849-1856.
- [69]. Kumar, G., et al., *Correlation of methane activation and oxide catalyst reducibility and its implications for oxidative coupling*. Acs Catalysis, 2016. **6**(3): p. 1812-1821.
- [70]. Derk, A.R., et al., *Methane oxidation by lanthanum oxide doped with Cu, Zn, Mg, Fe, Nb, Ti, Zr, or Ta: the connection between the activation energy and the energy of oxygen-vacancy formation*. Catalysis letters, 2013. **143**: p. 406-410.
- [71]. Thybaut, J.W., et al., *Catalyst design based on microkinetic models: Oxidative coupling of methane*. Catalysis today, 2011. **159**(1): p. 29-36.
- [72]. Ghose, R., H.T. Hwang, and A. Varma, *Oxidative coupling of methane using catalysts synthesized by solution combustion method: Catalyst optimization and kinetic studies*. Applied Catalysis A: General, 2014. **472**: p. 39-46.
- [73]. Kondratenko, E.V., et al., *Developing catalytic materials for the oxidative coupling of methane through statistical analysis of literature data*. Catalysis Science & Technology, 2015. **5**(3): p. 1668-1677.
- [74]. Zhang, Y. and X. Xu, *Predictions of adsorption energies of methane-related species on Cu-based alloys through machine learning*. Machine Learning with Applications, 2021. **3**: p. 100010.
- [75]. Sadiku, M., et al., *Data visualization*. International Journal of Engineering Research And Advanced Technology (IJERAT), 2016. **2**(12): p. 11-16.
- [76]. Potter, K., et al. *Methods for presenting statistical information: The box plot*.
- [77]. Mayr, A., et al., *The evolution of boosting algorithms*. Methods of information in medicine, 2014. **53**(06): p. 419-427.
- [78]. Friedman, J.H., *Greedy function approximation: a gradient boosting machine*. Annals of statistics, 2001: p. 1189-1232.
- [79]. Ramraj, S., et al., *Experimenting XGBoost algorithm for prediction and classification of different datasets*. International Journal of Control Theory and Applications, 2016. **9**(40): p. 651-662.
- [80]. Schapire, R.E., *The Strength of Weak Learnability*. Machine Learning, 1990. **5**(2): p. 197-227.

- [81]. Freund, Y., *Boosting a weak learning algorithm by majority*. Information and computation, 1995. **121**(2): p. 256-285.
- [82]. Bentéjac, C., A. Csörgő, and G. Martínez-Muñoz, *A comparative analysis of gradient boosting algorithms*. Artificial Intelligence Review, 2021. **54**: p. 1937-1967.
- [83]. Chen, T. and C. Guestrin. *Xgboost: A scalable tree boosting system*.
- [84]. Demšar, J., *Statistical comparisons of classifiers over multiple data sets*. The Journal of Machine learning research, 2006. **7**: p. 1-30.
- [85]. Breiman, L., *Out-of-bag estimation*, *ftp. stat. berkeley. edu/pub/users/breiman. OOBestimation. ps*, 1996. **199**(6).
- [86]. Cestnik, B. *Estimating probabilities: A crucial task in machine learning*.
- [87]. Petrakieva, S., O. Garasym, and I. Taralova, <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7038771>. 2014.
- [88]. Ke, G., et al., *Lightgbm: A highly efficient gradient boosting decision tree*. Advances in neural information processing systems, 2017. **30**.
- [89]. Minastireanu, E.-A. and G. Mesnita, *Light gbm machine learning algorithm to online click fraud detection*. J. Inform. Assur. Cybersecur, 2019. **2019**: p. 263928.
- [90]. Tsoukalas, L.H. and R.E. Uhrig, *Fuzzy and neural approaches in engineering*. 1996: John Wiley & Sons, Inc.
- [91]. Zhong, J., et al. *Comparison of performance between different selection strategies on simple genetic algorithms*. IEEE.
- [92]. Goldberg, D.E., *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Longman Publishing Co., Inc. 1989.
- [93]. Kumar, M., et al., *Genetic algorithm: Review and application*. Available at SSRN 3529843, 2010.
- [94]. Sharma, M., *Role and working of genetic algorithm in computer science*. International Journal of Computer Applications and Information Technology (IJCAIT), 2013. **2**(1): p. 27-32.
- [95]. Oliveira, A.L.I., et al., *GA-based method for feature selection and parameters optimization for machine learning regression applied to software effort estimation*. information and Software Technology, 2010. **52**(11): p. 1155-1166.
- [96]. Hassanat, A., et al., *Choosing mutation and crossover ratios for genetic algorithms—a review with a new dynamic approach*. Information, 2019. **10**(12): p. 390.
- [97]. Razali, N.M. and J. Geraghty. *Genetic algorithm performance with different selection strategies in solving TSP*. International Association of Engineers Hong Kong, China.

- [98]. Oladele, R.O. and J.S. Sadiku, *Genetic algorithm performance with different selection methods in solving multi-objective network design problem*. International Journal of Computer Applications, 2013. **70**(12).
- [99]. Jebari, K. and M. Madiafi, *Selection methods for genetic algorithms*. International Journal of Emerging Sciences, 2013. **3**(4): p. 333-344.
- [100]. Shukla, A., H.M. Pandey, and D. Mehrotra. *Comparative review of selection techniques in genetic algorithm*. IEEE.
- [101]. Katoch, S., S.S. Chauhan, and V. Kumar, *A review on genetic algorithm: past, present, and future*. Multimedia tools and applications, 2021. **80**: p. 8091-8126.
- [102]. Ahmed, Z.H. *An improved genetic algorithm using adaptive mutation operator for the quadratic assignment problem*. IEEE.
- [103]. Zednik, C., *Solving the black box problem: A normative framework for explainable artificial intelligence*. Philosophy & technology, 2021. **34**(2): p. 265-288.
- [104]. Goldstein, A., et al., *Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation*. journal of Computational and Graphical Statistics, 2015. **24**(1): p. 44-65.
- [105]. Vellido, A., *The importance of interpretability and visualization in machine learning for applications in medicine and health care*. Neural computing and applications, 2020. **32**(24): p. 18069-18083.
- [106]. Petch, J., S. Di, and W. Nelson, *Opening the black box: the promise and limitations of explainable machine learning in cardiology*. Canadian Journal of Cardiology, 2022. **38**(2): p. 204-213.
- [107]. Dorogush, A.V., V. Ershov, and A. Gulin, *CatBoost: gradient boosting with categorical features support*. arXiv preprint arXiv:1810.11363, 2018.
- [108]. Dubois, J.-L. and C.J. Cameron, *Common features of oxidative coupling of methane cofeered catalysts*. Applied catalysis, 1990. **67**(1): p. 49-71.
- [109]. Liu, Y.C. and D.N. Lu, *Surface energy and wettability of plasma-treated polyacrylonitrile fibers*. Plasma chemistry and plasma processing, 2006. **26**: p. 119-126.
- [110]. Fleischer, V., et al., *Investigation of the surface reaction network of the oxidative coupling of methane over Na<sub>2</sub>WO<sub>4</sub>/Mn/SiO<sub>2</sub> catalyst by temperature programmed and dynamic experiments*. Journal of catalysis, 2016. **341**: p. 91-103.
- [111]. Shapley, L.S., *A value for n-person games*. 1953.



- [112]. Merrick, L. and A. Taly. *The explanation game: Explaining machine learning models using shapley values*. Springer.