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

[1]Haseeb Ahmad, [1]Iftikhar Ahmad*,[1] Qazi Ahmed and [2]Meshal Shutaywi
[1]*National University of Sciences & Technology (NUST) School of Chemical & Materials Engineering (SCME).*
[2]*King Abdul Aziz University College of Science and Arts, Department of Mathematics*
iftikhar.salarzai@scme.nust.edu.pk*

(Received on 20th May 2025, accepted in revised form 23rd January 2026)

Summary: Machine Learning (ML) has helped to accelerate research and innovation in multifarious domains. In this study, ML models have been used to predict adsorption energies of methane related species on Cu-based Alloy. Comparative study of various ML algorithms integrated with GA were performed to improve the ML model's architecture and parameters selection. The results proposed that Categorical Boosting (Catboost) model with RMSE = 0.0977, CC = 96.5 % outperformed all other models and effectively predicted adsorption energies. 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). In addition, SHAP analysis was employed to further interpret the contribution of individual descriptors to adsorption energies, allowing for the identification of key factors such as electronegativity, atomic radius, ionization energy, and surface energy. These insights highlight how dopant selection alters catalytic performance, demonstrating the ability of ML not only to provide accurate predictions but also to generate design-relevant knowledge. Overall, this approach provides a reliable and efficient methodology for reducing experimental screening and accelerating the discovery of promising Cu-based alloy catalysts for methane conversion.

Keywords: Machine Learning Optimization; Adsorption Energy Prediction; Cu-Based Catalyst; Oxidative Coupling; Methane Reforming.

Introduction

Methane can be extracted from natural gas and as the other reserves of petroleum decline, it will eventually become an essential raw material for the manufacture of fuels and chemicals. Research predicts that methane will last 60 years once oil reserves are exhausted [1]. 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 to methanol and *other* $_{\text{hydrocarbons}}$ [2]. 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 [3]:

$$CH_4 + H_2OCO + 3H_2 \quad \Delta H^{\circ}_{298} = 206kJ/mol^{-1} \ (1)$$

However, this process has its drawbacks. Depending on the catalyst used, the temperatures can range from 700K to 1050K requiring increased heat fluxes [5]. Elevated pressure is also needed which may reach 25 bars [6]. 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 [6]. This leads to greater energy requirements and higher costs [7]. Furthermore, the efficiency of this reaction is low, and the process suffers from stability issues [8].

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 [9]. 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 [10]. Extensive research has been carried out in the 1980s on OCM, but certain drawbacks relating to catalysis of the process led to its decline. OCM involves two distinct components. In the heterogeneous pathway, methane (CH_4) is initially activated on the metal-oxide catalyst surface. In contrast, the homogeneous pathway proceeds via gas-phase coupling of free-radical species [11]. Methyl radicalsCH_3^* are formed when hydrogen is ejected from methane by the activated oxygen molecules present on the catalyst surface. Ethane (C_2H_6) is produced as a result of coupling of CH_3^*radicals in the gaseous phase. In the final step, ethane is dehydrogenated to form ethylene (C_2H_4). OCM reactions are highly exothermic requiring temperatures of 950-1200 K as vast

***To whom all correspondence should be addressed.**

quantities of energy is required to cleave the C-H bond (429 KJ mol^{-1}) [12, 13].

However, OCM suffers from serious shortcoming of low C_2 yield. This is primarily due to secondary reactions of CH_3^* radicals. OCM is a fairly complex reaction as vigorously reactive radical species along with oxygen occupy the reactor [10]. Oxygen acts as an oxidant and a multitude of oxidation and dehydrogenation steps occur. Catalyst capable of activating CH_4 also activates the ethane at the same rate which causes production of 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 CO_x formation [9]. According to thermodynamic standpoint the formation of partial and total oxidation products (CO_x) is favorable which restricts C_2 yield. Complete and partial oxidation occurs according to the following reaction respectively:

$$CH_4 + O_2 \rightarrow CO + H_2 + H_2O \; \Delta H_{298}{}^0 = 278 KJ/mol^{-1} \quad (2)$$

$$CH_4 + 2O_2 \rightarrow CO_2 + 2H_2O \; \Delta H_{298}{}^0 = 802 KJ/mol^{-1} \quad (3)$$

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:

$$2CH_4 + O_2 \rightarrow C_2H_4 + 2H_2O \quad \Delta H_{298}^{\circ} = 282 KJ/mol^{-1} \quad (4)$$

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

$$C_2H_4 + 2H_2O \rightarrow 2CO \quad \Delta H_{298}^{\circ} = 210 KJ/mol^{-1} \quad (5)$$

This points to the fact that the conditions that favor formation of ethylene at these temperatures are not favorable for the activation of methane using oxygen. Nonetheless, when viewed through a thermodynamic lens, partial oxidation and total oxidation are significantly more thermodynamically favorable compared to the OCM. The principal and rate-limiting step in OCM reaction is the splitting of C-H bond in CH_4 endothermically. A study was conducted regarding the kinetics of the reaction and it revealed that an increase in C_2 yield was the result of an increase in the initial formation rate of methyl radical. Additionally, it was observed that the majority of methyl radicals ($\cdot CH_3$) were formed through the interaction of methane with surface oxygen species on the catalyst. In this pathway, activated oxygen atoms attack CH_4 and abstract a hydrogen atom, generating $\cdot CH_3$. This mechanism was shown to be far more dominant than hydrogen abstraction occurring in the gas phase, where free radicals remove hydrogen directly from CH_4 without surface involvement. The result highlights the critical role of surface oxygen species in initiating radical formation and driving the OCM process. 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 [14]. Consequently, it is of utmost importance to stabilize $\cdot CH_3$ so that dehydrogenation to reactive intermediates such as CH_2 and CH is prevented which end up as oxides of carbon impacting the yield significantly [15]. On the surface of the catalysts, methyl radical forms methoxide ion by electron acceptance which acts as surface intermediates in the production of CO_x. The methylene radical (CH_2) also couples with CO in gas-phase to form Ketene (CH_2CO). Eventually, these reactions affect yield and selectivity. Therefore, an optimum catalyst should aid in the formation of methyl radicals while suppressing its secondary oxidation reactions which are a source of CO_X. 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 [16].

Previously various studies have investigated the adsorption energies of energy related species using machine learning algorithms. But, these research explorations have been general for instance, Yang *et al*. compared the catalytic effect of metals on energetic materials employing machine learning predictions Kennel Ridge Regression (KKR) [3]. Similarly, Usuga *et al* have used local descriptor-based ML refined by cluster analysis for predicting energies of bimetallic alloys, the CatBoost exhibited best performance with MAE of 0.019. Furthermore, the work of Takashi *et al*. is most related where for effective utilization of methane, machine learning predictions for adsorption energies on metal alloys is compared [13]. However, despite extensive research in analyzing the adsorption energies on catalyst surface, there is a significant gap in literature regarding the utilization of ML algorithms for the prediction of adsorption energies for compounds and catalysts behavior alike.

This paper aims to develop a machine learning model that has the ability to accurately predict the adsorption energies of $\cdot CH_3$ 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 [17]. Charge transfer would be facilitated, and an improved energy flow would take place. The characteristics and properties of the different components facilitate so as to contribute to the aforementioned functionalities [18].

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 hyper-parameters were tuned using GA. The results achieved will provide insight into the role of catalysis and its ability to suppress undesired over-reactions. 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 [17]. 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 [18]. 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.

Experimental

Dataset and existing research

The dataset consisted of 46 transition metal and their respective physio-chemical and electrical properties were considered which makes 12 as the total number of descriptors. These descriptors included atomic number, group number, atomic mass, atomic radius, density, melting point, boiling point, electronegativity, ionization energy, surface energy, and enthalpy of fusion, which collectively capture the essential structural and electronic characteristics of the dopant elements. The data was taken from Toyao *et al*. (2018) [15], who 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 the supplementary files of the publication. 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. Overall, this dataset provides a reliable benchmark, combining first-principles DFT-calculated adsorption energies with well-defined elemental descriptors, enabling both accurate prediction and meaningful interpretation of catalyst behavior when Cu is alloyed with different transition metals.

Toyao *et al*. (2018) 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) 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 [19]

Data visualization

Data visualization can be defined as presenting information in form of graph or pictures making it understandable and interpretable [20]. For this purpose, boxplots were constructed for the descriptors as shown in Fig. 1. 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 [21]. The interquartile range as well as outliers can be seen. Skewness in the data can also be noticed in some of the parameters.

Machine learning algorithms

In this study, Xgboost, CatBoost Regression, and Light Gradient Boosting Model were employed to estimate the adsorption energies of the methane related species- $\cdot CH_3$ on the 46 unique Cu-based alloys. The present study was implemented in Python, an open-source programming language widely adopted for machine learning and data-driven scientific research due to its simplicity, reproducibility, and extensive library ecosystem. Python provides integrated support for data preprocessing (via libraries such as pandas and scikit-learn), model construction (e.g., CatBoost, LightGBM, XGBoost), and hyperparameter optimization (sklearn-genetic for GA-based tuning). The principles underlying its use in this work involved encoding categorical variables, scaling physicochemical descriptors, and applying supervised regression models to predict adsorption energies. Python further enabled the systematic evaluation of different parameters, including RMSE, MAE, CC, and R^2 across cross-validation folds, thereby ensuring a rigorous assessment of model performance. In addition, explainability tools (SHAP and partial dependence plots) implemented in Python allowed visualization of how descriptors such as

electronegativity, ionization energy, surface energy, and melting/boiling points affect adsorption predictions.

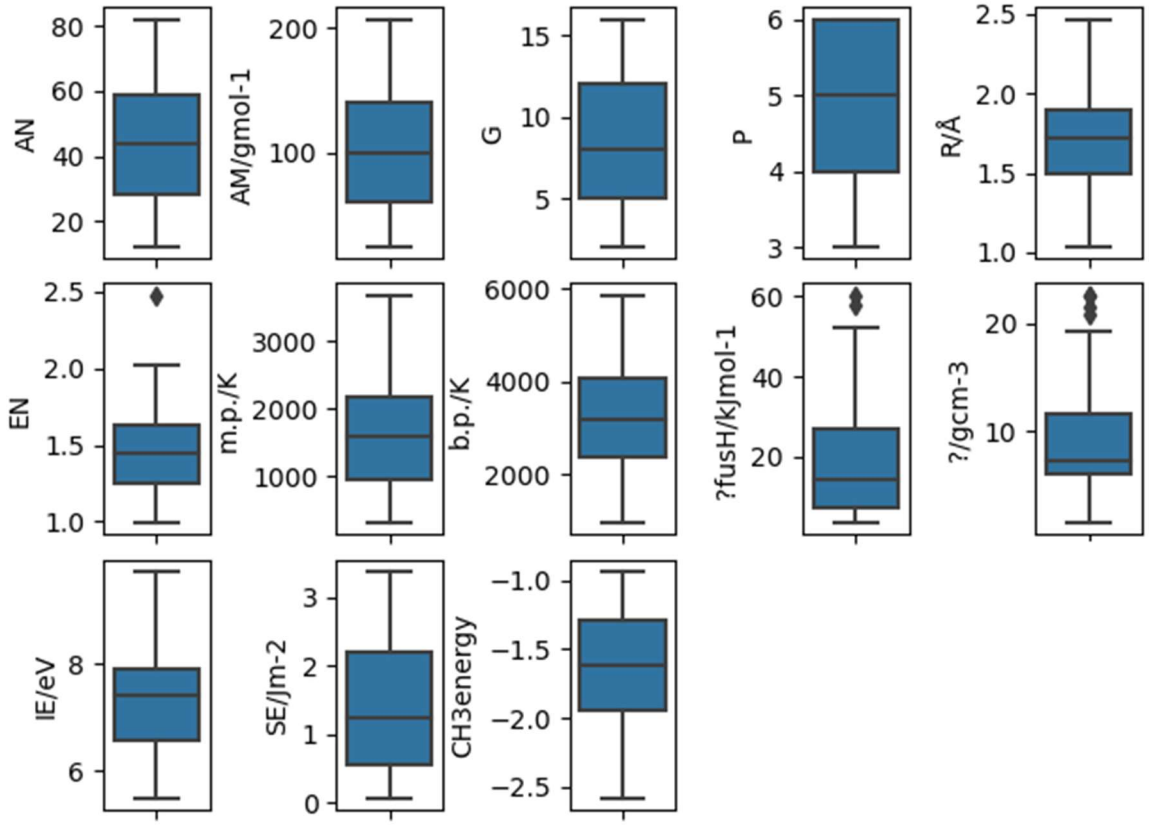

Fig. 1: Boxplot of the feature variables.

Gradient Boosting Algorithms

In recent decades, boosting algorithms have emerged as a promising technique for developing machine learning models. The theoretical basis behind this approach is to combine the solutions of weak learners to form a strong learner which has enhanced prediction capability and superior accuracy [22]. 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 and interpretability [23]. An analogy can constructs an adequate and satisfactory hypothesis from a comparatively mediocre hypothesis [24].

The first boosting algorithms were developed by Schapire and Fruend who labelled its concept as "garnering wisdom from a council of fools" [25–27]. 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

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 to the required results in an iterative way. Finally, the predictions of the individual learners are combined into more 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). When there are multiple descriptors, an additive model is

formed by adding the effects of the individual predictors:

$$f(x) = \beta_0 + h_1(x_1) + \cdot + h_p(x_p) \quad (6)$$

In this equation, β_0 is the intercept. The descriptors are x1, ..., xp forming component X. h1(·),...,hp(·) 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 approach to addressing problems involving machine learning [28]. 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 of the trees as XGBoost utilizes decision trees as its basic estimator.

$$L_{x_9B} = \sum_{i=1}^{N} L\big(y_i, F(x_i)\big) + \sum_{j=1}^{N} \Omega(h_m) \quad (7)$$

$$\Omega(\mathrm{h}) = \gamma\beta + \frac{1}{2}\boldsymbol{\lambda}\|\boldsymbol{W}\|^2 \quad (8)$$

In the above equations, β are total leaves in the tree and refers to the scores of the output of the leaves. γ 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 quicker than currently used solutions. The scalability of XGBoost may be directly attributed to a number of significant algorithmic and systemic enhancements [29]. 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 pre-processing [30]. Furthermore, it introduces a new framework in order to approximate leaf values while choosing structure of the tree that avoids or combats over-fitting. 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. CatBoost employs a more effective method that minimizes overfitting and permits the utilization of the entire dataset for training purposes. Specifically, the dataset are randomly permutated and the average label value for the instance are calculated with the identical category value placed before the provided one in the permutation for each example.

Taking a dataset $d_n = (X_i, Y_i)$i=1..n, in which $X_i = (i,1,....,i,s)$ is a vector with s features, and $Y_i \in$ R is the value of the label. If the permutation is $\sigma = (\sigma 1,,\sigma n)$, then substitution of occurs with

$$\frac{\sum_{j=1}^{p=1}[X_{\sigma j,k}=X_{\sigma p,k}]Y_{\sigma j}+a.P}{\sum_{j=1}^{p=1}[X_{\sigma j,k}=X_{\sigma p,k}]+a} \quad (9)$$

where we additionally add the previous value P and a parameter a¿0 $\alpha > 0$ 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 [32].

For the CatBoost regressor, four primary hyperparameters were tuned to optimize model performance. Iterations determine the total number of boosting rounds, where each new tree corrects the errors of the previous ensemble, directly influencing the learning capacity of the model. The learning rate controls the step size at each iteration, balancing convergence speed with generalization ability; smaller values improve stability but require more iterations. Depth specifies the maximum depth of each individual decision tree, governing the model's ability to capture complex feature interactions while also impacting the risk of overfitting. Finally, l2_leaf_reg is the L2 regularization coefficient applied to leaf values, which penalizes large weights and helps prevent over-complexity in the model. Together, these parameters define the trade-off between accuracy, robustness, and computational efficiency in CatBoost training.

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 [33]. In order to decrease the features, EFB groups together features that cannot normally take nonzero 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 [34]. Another reason behind these improvements is that LightGBM transforms continuous values of features into discrete bins which boost the training process [35].

Genetic Algorithm

Developed by John Holland in the 1970s, genetic algorithm (GA) is an adaptive heuristic search method which derives its basis from genetics of population [36]. Under the hood, it is following principle of "Survival of the fittest" introduced by Charles Darwin [37]. Natural selection is a key idea behind GA. It is an evolutionary algorithm which is initiated by providing solutions that form the population. In a genetic algorithm, a solution is represented as an individual, and a collection of individuals forms a chromosome. Every chromosome is defined by a set of genes. Each chromosome can be a probable solution in the search pool [38]. The optimum solution is selected after multiple generations. 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 [39]. 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 [4]. The individuals which have adapted appropriately are retained while others are removed [40].

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.

In the proposed framework, hyperparameter tuning of the boosting algorithms was carried out using GA. The GA was initialized with a predefined population size, which represents the number of candidate solutions evaluated in each generation. Through 20 successive generations, the algorithm iteratively evolved these candidate solutions to maximize model performance. New solutions were generated using two evolutionary operators: crossover probability, which defines the likelihood of combining two parent solutions to produce offspring, and mutation probability, which introduces random alterations to maintain diversity and avoid premature convergence. The fitness of each solution was assessed using the model's predictive accuracy, with the maximum fitness criterion serving as the optimization target. To ensure robustness and minimize overfitting, a 5-fold cross-validation strategy was applied at each evaluation step, providing

an unbiased estimate of the model's generalization ability.

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. 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.

As the various models have unique hyperparameters, therefore for the algorithms, ranges were specified for each and then fed to GA 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).

$$RMSE = \sqrt{\frac{\sum_{n=1}^{N}(\hat{r}_n - r_n)^2}{N}} \qquad (10)$$

$\hat{r}$ = prediction rate
r_n = true rating in testing data set
N = number of rating prediction pairs between the testing data and prediction result

$$R = r_{y\hat{y}} = \frac{\Sigma(\hat{y}_i - \bar{\hat{y}_i})(y_i - \bar{y})}{[\Sigma(y_i - \bar{y}_i)^2]^{\frac{1}{2}}[\Sigma(y_i - \bar{y}_i)^2]^{\frac{1}{2}}} \qquad (11)$$

$\hat{y}$ = predicted value of y: $\bar{y}$ = mean of y values

Results and Discussion

Models evaluation and comparison

Various ML models, including XGBoost, Catboost, and LightGBM were trained in Python programming language to predict adsorption energies of $\cdot CH_3$ on Cu-based alloys.

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. 2. The line in the Fig. 2 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 $\cdot CH_3$ on Cu-based alloys was evaluated using CC and RMSE as shown in Table-1. It is evident that Catboost outperformed LightGBM and Xgboost with an RMSE of 0.09772 ± 0.032 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 ±0.041, 0.116 ±0.049 respectively. In the case of coefficient of determination or R^2 Catboost emerged superior with a value of 0.92 followed by LightGBM 0.85 and 0.74 for XGBoost. CatBoost has both CPU and GPU implementations. The GPU implementation allows for much faster training. This gradient boosting algorithm successfully handles categorical features and takes advantage of dealing with them during training as opposed to preprocessing time [43]. It uses an efficient method called ordered boosting, which is specifically designed to deal with categorical features directly.

Table-1: Comparison of ML methods.

Model	CC	RMSE	R^2
LGBM	**95.4%**	**0.1071 ±0.032**	**0.85**
Catboost	**96.5%**	**0.0977 ±0.041**	**0.92**
Xgboost	**93.5%**	**0.1116 ±0.049**	**0.73**

Previously, Toyao *et al*. (2018) used Gaussian Process Regression with Bayesian optimization to obtain a low RMSE of roughly 0.24 eV ±0.06. Whereas the current work uses advanced gradient boosting methods in conjunction with GA optimization, resulting in a substantially reduced RMSE of 0.0977 eV with the CatBoost model and a CC of 96.5%. This improvement demonstrates the efficacy of integrating boosting-based machine learning algorithms with evolutionary hyperparameter tweaking.

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 over-fitting. 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 $\cdot CH_3$ and other compounds can be predicted using the various boosting algorithms which will effectively reduce the elapsed time and costs associated with experiments.

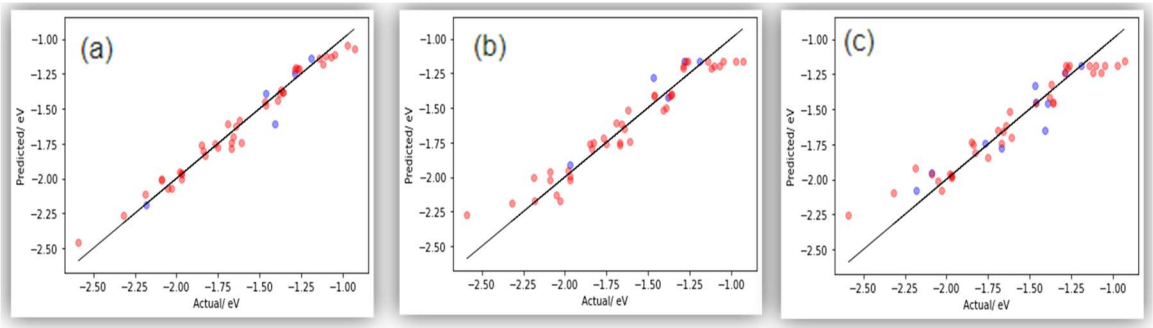

Fig. 2: DFT-calculated adsorption energies of ·CH_3 on Cu-based alloys and their predicted values. (a) Catboost (b) LightGBM (c) XGBoost. Color code: red: training set; blue: test set

Explainable Machine Learning

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 [44]. 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 [45]. 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 [46]. Different techniques are employed in explainable ML.

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 [47].

Permutation importance scores in the case of ·CH_3 can be seen in Fig. 3. 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 shows the PDPs that relate group number to the output show 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 electro-positivity can modify the morphology and create defects in the structure of the host oxides [48]. Zavyalova *et al*. research on an abundance of data on OCM catalysts showed that high basicity is paramount for improved C_2 selectivity as it directly influences the bond formed between the anions and cations [49].

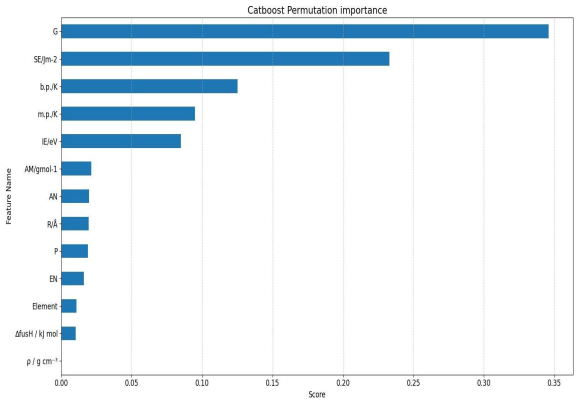

Fig. 3: Permutation importance scores of the feature variables.

Smaller atomic size and higher electro-positivity contribute to basicity. With respect to the atomic radius, the PDP show reveals a fairly constant value of the output till 1.6 R/˚A and then it falls and

rises again. At 1.9 R/°A, a sharp downward descent continues till 2.0 R/°A 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^{-1}. 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 occurs [9]. Kumar *et al*. discovered that improved OCM activity takes place at higher surface reducibility [48]. 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 [50]. The PDP of surface energy show a consistent increase in adsorption energy as the surface energy of the doped element increases. 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 [52]. 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 $\cdot CH_3$ 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 is more beneficial for forming reactive species. High ionization energy will inhibit electron transfer affecting its interplay with adsorbates. Active sites that activate and cleave C-H bond are reactive oxygen species present on the catalyst surface.

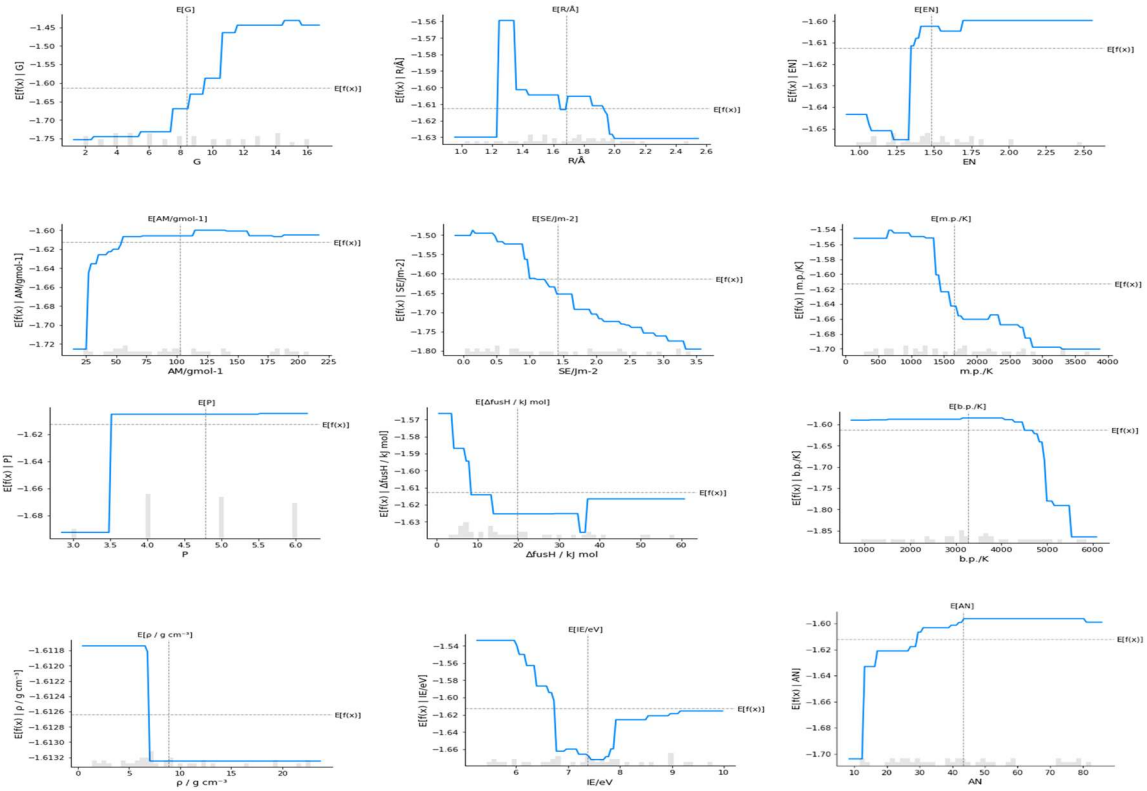

Fig. 4: PDPs of the feature variables to the prediction of the adsorption energies.

These oxygen species include chemisorbed oxygen (O^{-1}), dissociative adsorbed oxygen (O^{-1}), adsorbed oxygen ions (O^{-2}), and lattice oxygen (O^{-2}) [51, 17]. 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 [52]. The 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. Melting point and boiling point both exhibited secondary but consistent influence: higher values were associated with stronger adsorption, reflecting the role of cohesive energy and metallic bonding in stabilizing adsorbates. This observation is consistent with earlier catalytic studies, which emphasized the importance of thermodynamic stability in determining active site robustness and overall catalyst durability [5, 7, 8]. Density also contributed modestly, with higher density alloys showing sharper variations in adsorption strength. This may be linked to atomic packing effects and oxygen mobility, which directly influence surface reducibility and the formation of active oxygen species [9, 48]. Similarly, enthalpy of fusion (ΔfusH) was found to affect adsorption behavior, with lower values favoring weaker adsorption. This aligns with adsorption theory, where lower enthalpy of fusion correlates with reduced energetic requirements for surface modification, thereby influencing adsorbate–surface interactions [16, 17]. While these descriptors contributed less strongly than primary factors such as group number and surface energy, their secondary effects are important for fine-tuning adsorption energetics and improving the rational design of Cu-based alloys.

The relationship between features and the target variable can be further explored using two-way PDP 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 [50]. 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 [53]. In Fig. 5, a bee swarm 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. It shows the distribution of the effects that each feature has on the model's predictions along the x-axis for each feature. These distributions are shown for each feature individually. 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, in fig. 5 surface energy ranks second in the importance of features and its 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.

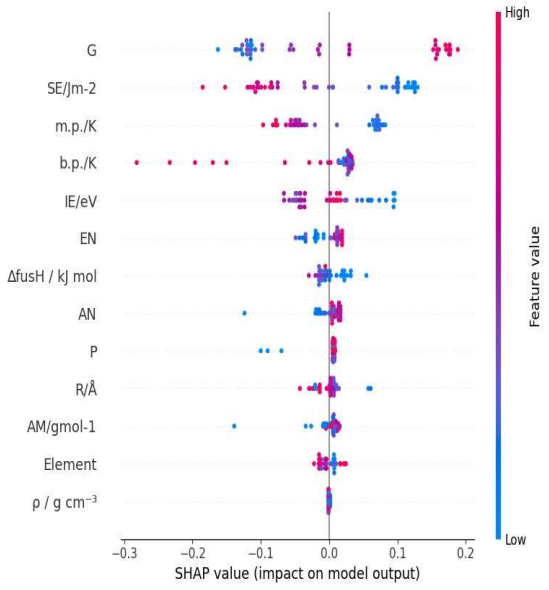

Fig. 5: SHAP beeswarm summary plot.

To provide a deeper interpretation and directly relate the explainability results to catalytic theory, SHAP local analyses were performed on representative Cu-based alloys. By focusing on specific cases such as Cu–Ni and Cu–Li, it becomes possible to elucidate how individual physicochemical descriptors influence adsorption energies and how these effects correspond to known adsorption mechanisms.

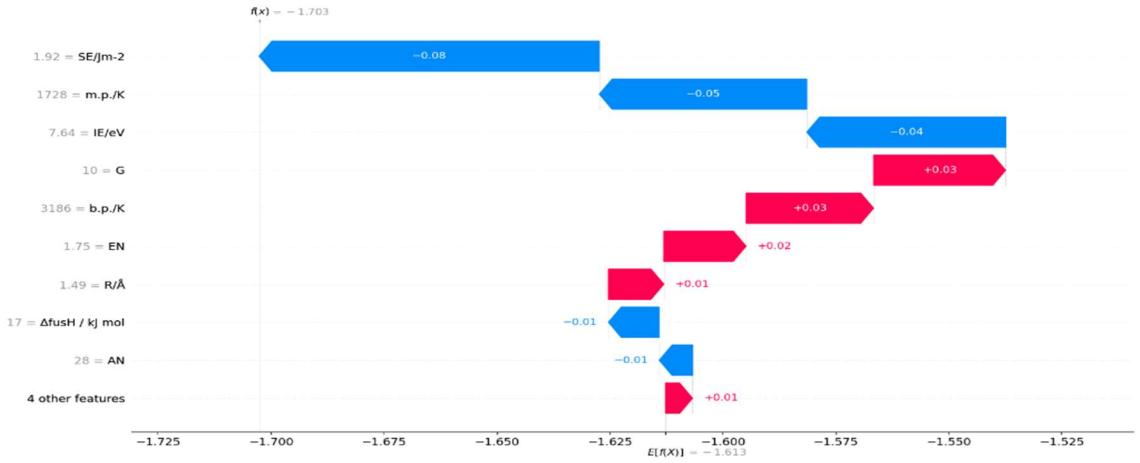

Fig. 6: SHAP Waterfall plot for Cu-Ni Alloy.

The SHAP waterfall analysis of Cu–Ni alloys **Fig. 6** reveals that surface energy and melting point exert the strongest influence, both driving adsorption energy towards more negative values. This is consistent with Ni's high surface reactivity and metallic bonding strength, which enhance adsorbate stabilization. At the same time, Ni's relatively high electronegativity (1.75) and small atomic radius counterbalance this effect, weakening CH_3 radical stabilization and favoring complete oxidation pathways. From a catalytic design perspective, this demonstrates that although Ni increases adsorption strength, it does so at the cost of selectivity, highlighting the trade-off between binding energy and C_2 yield in OCM.

The SHAP waterfall analysis of Cu–Sr (Fig. 6) reveals a distinct adsorption mechanism compared to Cu–Ni. Here, surface energy and ionization energy were the dominant descriptors, both driving adsorption energy towards less negative values (weaker binding). This is consistent with Sr's alkaline-earth character: its high electropositivity and low ionization energy promote the generation of surface oxygen vacancies while preventing excessive stabilization of CH_3 intermediates. As a result, CH_3 radicals are more likely to desorb and couple selectively to form C_2 products, rather than being over-oxidized. In contrast to Cu–Ni, which favors strong adsorption but poor selectivity, Cu–Sr demonstrates the catalytic principle that moderate adsorption strength enhances selectivity in OCM.

The comparative SHAP analyses of Cu–Ni and Cu–Sr highlight two contrasting adsorption mechanisms that reflect fundamental catalytic design principles. For Cu–Ni, strong contributions from surface energy and melting point resulted in more negative adsorption energies, indicating stronger CH_3 binding. However, Ni's higher electronegativity and smaller radius counteracted radical stabilization, leading to over-oxidation and reduced selectivity. In contrast, Cu–Sr displayed weaker adsorption energies, driven by Sr's high electropositivity, low ionization energy, and low surface energy. These characteristics favored the formation of surface oxygen vacancies and moderate CH_3 stabilization, which are critical for promoting C_2 product selectivity in OCM. Together, these case studies demonstrate that SHAP interpretability not only validates model predictions but also connects descriptor effects to catalytic theory, offering a mechanistic rationale for balancing activity and selectivity in catalyst design.

Research Benefits and Conclusion

In this research, three distinct boosting machine learning models were employed. These models were finetuned through the application of a GA, aiming to optimize and forecast the adsorption energy of *CH_4*-related species. The results were interpreted using permutation importance and PDPs, valuable methods for gaining insights into individual feature significance and illustrating variable outcome relationships. Primary influences included group number, electronegativity, ionization energy, and surface energy, which are well established as critical factors in methane activation and C–H bond cleavage. Taken together, these results reveal how both electronic descriptors (e.g., electronegativity, ionization energy) and thermodynamic descriptors (e.g., melting/boiling point, density, enthalpy of fusion) collectively govern adsorption energetics. From a catalytic design perspective, the findings indicate that alloys doped with electropositive elements possessing low ionization energy and large

atomic radii are promising candidates for enhancing C_2 selectivity in oxidative coupling of methane. 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 its related species molecules, 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. The predictive results provide clear guidance for Cu-based catalyst design in OCM. Alloys doped with highly electropositive elements such as Sr, Ca, and La emerged as promising candidates, as their predicted adsorption energies indicate moderate CH_3 stabilization that favors selective coupling to C_2 products. In contrast, transition-metal dopants such as Ni and Co displayed overly negative adsorption energies, consistent with excessive binding strength that promotes non-selective oxidation to COx. By highlighting these trends, the model narrows the experimental screening space from dozens of possible Cu-based alloys to a smaller set of promising candidates with favorable descriptor profiles. Moreover, the SHAP analysis suggests that low ionization energy, large atomic radius, and group number are critical design parameters for improving catalyst performance. These insights not only reduce the need for extensive trial-and-error experimentation but also point towards rational design strategies, such as combining electropositive dopants with transition metals to balance activity and selectivity. Ultimately, the integration of ML predictions with catalytic theory establishes a pathway for accelerated discovery and optimization of Cu-based OCM catalysts. Understanding adsorption energies provides insights into the underlying mechanisms of OCM reactions, helping researchers optimize conditions for higher selectivity and yield of desired products. Machine learning models can analyze complex datasets, extracting patterns and relationships that may not be immediately apparent through traditional approaches. This data-driven decision-making process can guide researchers toward more effective strategies for OCM.

References

1. A. P. York, T. C. Xiao, M. L. Green and J. B. Claridge, Methane oxyforming for synthesis gas production, *Catal. Rev.*, 49, 511 (2007).
2. A. Holmen, Direct conversion of methane to fuels and chemicals, *Catal. Today*, 142, 2 (2009).
3. J. R. Rostrup-Nielsen, Catalytic steam reforming, In *Catalysis: Science and Technology, Vol. 5*, Springer-Verlag, Berlin, Heidelberg, p. 1 (1984).
4. G. Li, C. Liu, X. Cui, Y. Yang and F. Shi, Oxidative dehydrogenation of light alkanes with carbon dioxide, *Green Chem.*, 23, 689 (2021).
5. H. Zhang, Z. Sun and Y. H. Hu, Steam reforming of methane: Current states of catalyst design and process upgrading, *Renew. Sustain. Energy Rev.*, 149, 111330 (2021).
6. J. Xu and G. F. Froment, Methane steam reforming, methanation and water-gas shift: I. Intrinsic kinetics, *AIChE J.*, 35, 88 (1989).
7. T. L. LeValley, A. R. Richard and M. Fan, The progress in water gas shift and steam reforming hydrogen production technologies — A review, *Int. J. Hydrogen Energy*, 39, 16983 (2014).
8. P. Parthasarathy and K. S. Narayanan, Hydrogen production from steam gasification of biomass: influence of process parameters on hydrogen yield — a review, *Renew. Energy*, 66, 570 (2014).
9. Y. Gambo, A. A. Jalil, S. Triwahyono and A. A. Abdulrasheed, Recent advances and future prospect in catalysts for oxidative coupling of methane to ethylene: A review, *J. Ind. Eng. Chem.*, 59, 218 (2018).
10. L. Hu, D. Pinto and A. Urakawa, Catalytic oxidative coupling of methane: heterogeneous or homogeneous reaction?, *ACS Sustain. Chem. Eng.*, 11, 10835 (2023).
11. J. H. Lunsford, The catalytic oxidative coupling of methane, *Angew. Chem. Int. Ed. Engl.*, 34, 970 (1995).
12. S. J. Blanksby and G. B. Ellison, Bond dissociation energies of organic molecules, *Acc. Chem. Res.*, 36, 255 (2003).
13. K. Takahashi, I. Miyazato, S. Nishimura and J. Ohyama, Unveiling hidden catalysts for the oxidative coupling of methane based on combining machine learning with literature data, *ChemCatChem*, 10, 3223 (2018).
14. V. S. Arutyunov, V. Y. Basevich, V. I. Vedeneev and O. V. Krylov, On the role of a catalyst in

high-temperature reactions of methane oxidation, *Kinet. Catal.*, 40, 382 (1999).
15. T. Toyao, K. Suzuki, S. Kikuchi, S. Takakusagi, K. I. Shimizu and I. Takigawa, Toward effective utilization of methane: machine learning prediction of adsorption energies on metal alloys, *J. Phys. Chem. C*, 122, 8315 (2018).
16. A. Dabrowski, Adsorption — from theory to practice, *Adv. Colloid Interface Sci.*, 93, 135 (2001).
17. U. Zavyalova, M. Holena, R. Schlögl and M. Baerns, Statistical analysis of past catalytic data on oxidative methane coupling for new insights into the composition of high-performance catalysts, *ChemCatChem*, 3, 1935 (2011).
18. J. Song, Y. Sun, R. Ba, S. Huang, Y. Zhao, J. Zhang, Y. Sun and Y. Zhu, Monodisperse Sr–La_2O_3 hybrid nanofibers for oxidative coupling of methane to synthesize C_2 hydrocarbons, *Nanoscale*, 7, 2260 (2015).
19. Y. Zhang and X. Xu, Predictions of adsorption energies of methane-related species on Cu-based alloys through machine learning, *Mach. Learn. Appl.*, 3, 100010 (2021).
20. M. Sadiku, A. E. Shadare, S. M. Musa, C. M. Akujuobi and R. Perry, Data visualization, *Int. J. Eng. Res. Adv. Technol.*, 2, 11 (2016).
21. K. Potter, H. Hagen, A. Kerren and P. Dannenmann, Methods for presenting statistical information: The box plot, In *Visualization of Large and Unstructured Data Sets (VLUDS)*, p. 97 (2006).
22. A. Mayr, H. Binder, O. Gefeller and M. Schmid, The evolution of boosting algorithms, *Methods Inf. Med.*, 53, 419 (2014).
23. J. H. Friedman, Greedy function approximation: a gradient boosting machine, *Ann. Stat.*, 29, 1189 (2001).
24. S. Ramraj, N. Uzir, R. Sunil and S. Banerjee, Experimenting XGBoost algorithm for prediction and classification of different datasets, *Int. J. Control Theory Appl.*, 9, 651 (2016).
25. R. E. Schapire, The strength of weak learnability, *Mach. Learn.*, 5, 197 (1990).
26. Y. Freund, Boosting a weak learning algorithm by majority, *Inf. Comput.*, 121, 256 (1995).
27. R. E. Schapire and Y. Freund, *Boosting: Foundations and Algorithms*, MIT Press, Cambridge, MA (2012).
28. C. Bentéjac, A. Csörgő and G. Martínez-Muñoz, A comparative analysis of gradient boosting algorithms, *Artif. Intell. Rev.*, 54, 1937 (2021).
29. T. Chen and C. Guestrin, XGBoost: A scalable tree boosting system, In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, San Francisco, p. 785 (2016).
30. A. V. Dorogush, V. Ershov and A. Gulin, CatBoost: gradient boosting with categorical features support, *arXiv preprint*, arXiv:1810.11363 (2018).
31. L. Breiman, Out-of-bag estimation, Technical Report, Statistics Department, University of California, Berkeley (1996).
32. B. Cestnik, Estimating probabilities: A crucial task in machine learning, In *Proceedings of the 9th European Conference on Artificial Intelligence*, Pitman Publishing, Stockholm, p. 147 (1990).
33. F. Alzamzami, M. Hoda and A. El Saddik, Light gradient boosting machine for general sentiment classification on short texts: a comparative evaluation, *IEEE Access*, 8, 101840 (2020).
34. G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye and T. Y. Liu, LightGBM: A highly efficient gradient boosting decision tree, *Adv. Neural Inf. Process. Syst.*, 30, 3146 (2017).
35. E. A. Minastireanu and G. Mesnita, Light GBM machine learning algorithm to online click fraud detection, *J. Inform. Assur. Cybersecur.*, 2019, 263928 (2019).
36. L. H. Tsoukalas and R. E. Uhrig, *Fuzzy and Neural Approaches in Engineering*, John Wiley & Sons, Inc., New York (1996).
37. J. Zhong, X. Hu, J. Zhang and M. Gu, Comparison of performance between different selection strategies on simple genetic algorithms, In *International Conference on Computational Intelligence for Modelling, Control and Automation (CIMCA-IAWTIC'06), Vol. 2*, IEEE, Vienna, p. 1115 (2005).
38. J. Shapiro, Genetic algorithms in machine learning, In *Advanced Course on Artificial Intelligence*, Springer-Verlag, Berlin, Heidelberg, p. 146 (1999).
39. M. Kumar, D. M. Husain, N. Upreti and D. Gupta, Genetic algorithm: Review and application, *SSRN Electron. J.*, SSRN 3529843 (2010).
40. M. Sharma, Role and working of genetic algorithm in computer science, *Int. J. Comput. Appl. Inf. Technol.*, 2, 27 (2013).
41. A. L. Oliveira, P. L. Braga, R. M. Lima and M. L. Cornélio, GA-based method for feature selection and parameters optimization for machine learning regression applied to software effort estimation, *Inf. Softw. Technol.*, 52, 1155 (2010).
42. W. Wang and Y. Lu, Analysis of the mean absolute error (MAE) and the root mean square

error (RMSE) in assessing rounding model, *IOP Conf. Ser.: Mater. Sci. Eng.*, 324, 012049 (2018).
43. A. G. Asuero, A. Sayago and A. G. González, The correlation coefficient: An overview, *Crit. Rev. Anal. Chem.*, 36, 41 (2006).
44. C. Zednik, Solving the black box problem: A normative framework for explainable artificial intelligence, *Philos. Technol.*, 34, 265 (2021).
45. A. Goldstein, A. Kapelner, J. Bleich and E. Pitkin, Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation, *J. Comput. Graph. Stat.*, 24, 44 (2015).
46. A. Vellido, The importance of interpretability and visualization in machine learning for applications in medicine and health care, *Neural Comput. Appl.*, 32, 18069 (2020).
47. J. Petch, S. Di and W. Nelson, Opening the black box: the promise and limitations of explainable machine learning in cardiology, *Can. J. Cardiol.*, 38, 204 (2022).
48. J. L. Dubois and C. J. Cameron, Common features of oxidative coupling of methane cofeed catalysts, *Appl. Catal.*, 67, 49 (1990).
49. Y. C. Liu and D. N. Lu, Surface energy and wettability of plasma-treated polyacrylonitrile fibers, *Plasma Chem. Plasma Process.*, 26, 119 (2006).
50. M. C. Cholewinski, M. Dixit and G. Mpourmpakis, Computational study of methane activation on γ-Al_2O_3, *ACS Omega*, 3, 18242 (2018).
51. V. I. Lomonosov, Y. A. Gordienko, M. Y. Sinev, V. A. Rogov and V. A. Sadykov, Thermochemical properties of the lattice oxygen in W, Mn-containing mixed oxide catalysts for the oxidative coupling of methane, *Russ. J. Phys. Chem. A*, 92, 430 (2018).
52. L. S. Shapley, A value for n-person games, In *Contributions to the Theory of Games, Vol. II*, Princeton University Press, Princeton, p. 307 (1953).
53. L. Merrick and A. Taly, The explanation game: Explaining machine learning models using Shapley values, In *Machine Learning and Knowledge Extraction (CD-MAKE 2020)*, Springer International Publishing, Cham, p. 17 (2020).