Digital Chemical Engineering 12 (2024) 100165

Contents lists available at ScienceDirect

Digital Chemical Engineering

journal homepage: www.elsevier.com/locate/dche

Original article

Machine learning-enhanced optimal catalyst selection for water-gas shift reaction

Rahul Golder[1], Shraman Pal[1], Sathish Kumar C., Koustuv Ray *

Department of Chemical Engineering, Indian Institute of Technology, Kharagpur, Kharagpur 721302, West Bengal, India

ARTICLE INFO

Keywords:
Catalysis
Water-gas shift reaction
Machine learning
Bayesian optimization
Screening and prediction

ABSTRACT

The water-gas shift (WGS) reaction is pivotal in industries aiming to convert carbon monoxide, a byproduct of steam reforming of methane and other hydrocarbons, into carbon dioxide and hydrogen. Selecting an effective catalyst for this transformation poses a substantial challenge, as it requires a delicate balance between conversion, stability, and cost. We combine machine learning-driven prediction models with Bayesian optimization to explore and identify novel catalyst compositions. The proposed method efficiently explores the catalysis composition space for a predefined set of active metals, supports, and promoters to identify the most promising catalyst formulations. We assign weights to different performance metrics of catalysts, enabling tailored optimization according to specific industry needs. Our screening system streamlines catalyst discovery and facilitates the screening and selection of catalysts that balance conversion performance, stability, and cost-effectiveness. This approach holds significant promise for advancement in heterogeneous catalysis to meet the growing demands of efficient industrial processes.

1. Introduction

The water-gas shift (WGS) reaction is paramount for many industrial applications and environmental imperatives. At its core, the WGS reaction is the catalytic conversion of carbon monoxide (CO) and water vapour (H_2O) into carbon dioxide (CO_2) and hydrogen (H_2). The simple reaction bears profound implications that reverberate across diverse sectors, encompassing energy production, environmental sustainability, and chemical manufacturing.

$$CO + H_2O \rightleftharpoons CO_2 + H_2 \quad (1)$$

An immediate application of the WGS reaction is in industrial hydrogen production. The WGS reaction is vital in producing hydrogen from hydrocarbon feedstocks like methane, natural gas, and biomass-derived fuels. For example, it is used to remove carbon monoxide from synthetic gas to obtain high-purity hydrogen. At the other end of the spectrum, in numerous industrial processes, CO inevitably emerges as byproduct, typically from hydrocarbon reforming and gasification. However, realizing the full potential of the WGS reaction hinges on developing efficient and cost-effective catalysts. Catalyst selection is a multifaceted endeavour, requiring a delicate balance among conversion capability, catalyst stability, selectivity, and cost considerations. Conventional trial-and-error methods fall short due to the extensive composition space of catalyst materials and the corresponding time and cost involved to explore it satisfactorily. Our objective in this work is to propose a novel framework to examine the catalyst composition space in search of the optimal catalyst for the WGS reaction using machine learning (ML) models as catalyst performance predictors.

ML's ability to harness extensive data empowers engineers to make informed, data-driven decisions. Optimization of intricate chemical processes has been demonstrated by Zhou et al. (2017), Shokry et al. (2021), Nikita et al. (2023), while Reiser et al. (2022), Stanev et al. (2021) perform research to accelerate materials discovery. Additionally, it can enhance fault detection as illustrated by Heo and Lee (2018), Hu et al. (2021) and process control as shown by Wysotzki (1992), Faria et al. (2022). It also facilitates customized catalyst design, improves energy efficiency, bolster environmental sustainability, and ultimately elevates the precision and reliability of chemical engineering processes.

The integration of ML in the realm of heterogeneous catalysis has been a subject of recent exploration, offering innovative solutions to multifaceted challenges. Numerous studies have underscored the remarkable potential of ML in catalysis as shown in the works of Mou et al. (2023), Xu et al. (2021). These investigations have paved the way for harnessing the computational prowess of ML, providing avenues to alleviate the computational burden and bridge the gap between empirical observations and theoretical comprehension.

* Corresponding author.
E-mail address: koustuv@che.iitkgp.ac.in (K. Ray).
[1] Contributed equally.

https://doi.org/10.1016/j.dche.2024.100165
Received 3 March 2024; Received in revised form 5 June 2024; Accepted 8 June 2024
Available online 24 June 2024

An exemplary categorization of these methodologies emerges in the work of Goldsmith et al. (2018), where catalytic properties, descriptors, and the incorporation of Quantum Mechanical calculations are systematically elucidated. Esterhuizen et al. (2022) propose several explainable ML models which could help researchers interpret results from a scientific perspective. However, the study is limited to the proposed models without a specific application in heterogeneous catalysis. Gusmão et al. (2021) introduces a distinctive approach that combines ML with differential equations, particularly in modelling kinetic phenomena. This pioneering work hinges on the idea of constraining differential equations with differential–algebraic equations while also conducting parameters estimation from kinetic data. The novel framework serves as an essential tool for unravelling the intricate mechanisms governing catalytic reactions. Baumes et al. (2006) perform studies revolving around using Support Vector Machines (SVM) for predictive modelling in heterogeneous catalysis. They demonstrate their findings with two applications, one for olefin epoxidation and the other for light paraffin isomerization. They highlight the interpretable visualizations offered by the SVM model. An equally fascinating approach is presented by Lee et al. (2022) who tackles the synthesizability of catalysts as a ML task coupled with density functional theory. Furthermore, Fujinuma et al. (2022) offers a unique perspective, shedding light on the application of ML models in material science. Their work showcases the versatility of ML in addressing a diverse range of catalytic challenges.

While these foundational studies illuminate the broad prospects of ML in catalysis, the focus of this study is to the application in specific catalytic reactions. In this field, Madeira and Portela (2002) has extensively explored the predictive capabilities of various ML models in the oxidative dehydrogenation of n-butane. Their research highlights the effectiveness of SVMs compared to other techniques such as XGBoost, Random Forest, and K-Nearest Neighbours. Expanding this perspective, Khatamirad et al. (2023) delves into the data-driven design of In-based catalysts within the carbon-di-oxide to methanol reaction. Employing high-throughput simulations in tandem with Density Functional Theory (DFT), they efficiently generate a substantial dataset of oxygen vacancy formation energy. In a parallel effort, Denny et al. (2022) applies similar principles of density functional theory and ML to Pt-modified catalysts for ethanol reforming. This innovative fusion sheds light on the feasibility of synthesizing various catalyst compositions. In the context of acetylene semi-hydrogenation, Chen et al. (2022) leverages ML to guide the catalytic process.

Nevertheless, the prevailing challenge pertains to predicting catalyst performance based on descriptors and, more critically, to the actual design and development of novel catalyst combinations. As elucidated by Jones et al. (1998), optimizing black box functions forms a crucial aspect of this endeavour. In this context, Pandit et al. (2022) employs a combination of supervised ML and Density Functional Theory to design NiCoCu-based catalysts tailored for the hydrogen evolution reaction. Artrith et al. (2020) propose a combination of machine learning and first principles to predict the selectivity and activity of bimetallic catalysts for ethanol reforming using transition-state energies. They are able to predict four compositions of a specific Pt catalyst structure as a potential candidate for future experimental studies. Musa et al. (2022) performs acceleration of catalyst structure searches with ML. This innovative approach positions ML as an accelerator for streamlining the catalyst design process. Moreover, Deng et al. (2023) introduces iterative ML techniques for catalyst screening in the production of H2O2. Their groundbreaking work incorporates coordinate information from single-atom catalysts as an additional feature in the predictive modelling process. Roy et al. (2021), Roy et al. (2022) conduct screening of alloy-based catalyst for CO_2 hydrogenation to methanol by utilizing ML models as their basis. Ayodele et al. (2021) perform extensive comparisons of different neural network structures to model the effect of process parameters on the conversion of CO and methane during reforming. However, their study does not involve working on the optimal catalyst but rather the optimal operating conditions. Nevertheless, their ordering of operating parameters in relative importance provides a soft sanity check for our work. Furthermore, current literature demonstrates highly outdated ML models or basic neural networks. Most of the work is limited to contributing only ML models as predictors while disregarding the impact of such models.

The work of de Oliveira and Pacheco (2022) introduce a novel framework utilizing Convolutional Neural Networks (CNN) to screen catalysts. They demonstrate their method on several reactions like the Water-gas shift reaction, selective CO oxidation and steam reforming. However, their results have not been validated through experimental work. In the specific context of the WGS reaction, which is the main focus of our work, Kim and Kim (2022b), Kim and Kim (2022a) propose an ML-based screening, design of a catalyst with variable composition of a fixed set of elements. They study the utilization of several advanced neural networks to predict the optimal structure for Pt-based catalysts with different supports, including Cerium (Ce) and Titanium (Ti). As cited previously, most of the work in the sub-domain of WGS reaction utilizes fixed components/elements while optimizing the fraction of each used. This does not readily generalize to catalysts with other elements. As ML models thrive on diverse data, limiting the dataset from the literature to specific catalysts hinders the full potential of ML predictions. Aggregating data from the literature of different catalysts results in a diverse and large dataset. Odabaşı et al. (2014) aggregated catalyst data for water gas shift reaction from 2002–2012. Searching and optimizing models using this dataset would result in a more generalizable outcome, as shown by Chattoraj et al. (2022). However, they only utilize older ML models and provide no measure of stability. Their research is constrained to CO conversion quality while not providing any result that chemical engineers and researchers can utilize.

Our approach aims to efficiently explore and identify the most promising catalyst combination and compositions from the infinite composition space based on user-specified importance to conversion quality, catalyst stability and cost. We evaluate several advanced ML models to predict these vital catalyst performance metrics. We follow it up with a Bayesian optimization search within catalyst composition space utilizing our trained models as approximate predictors to ultimately find the optimal catalyst. Therefore, in the present study we focused on:

- Machine learning-based catalytic performance metric predictors based on catalyst composition elemental properties and operating conditions.
- Bayesian optimization search for examining catalyst composition space to find the optimal catalyst composition.
- A flexible framework to allow users to configure importance weightages to catalyst performance metrics: conversion, stability and cost as required.

As a caveat, we cannot provide any experimental validation of our computational results. We leave this to other researchers better equipped to test it using their laboratory resources. The following section provides the details of the method employed.

2. Computational methodology

This section outlines the methods employed to develop our framework. Our framework can be divided into two major categories: one for ML model training and second for optimal catalyst search. In our investigation, we conducted ML regression tasks utilizing the dataset provided by Chattoraj et al. (2022) comprising 4190 samples and 25 features. We benchmark our results against theirs in the subsequent sections to justify the significance of our research (see Fig. 1).

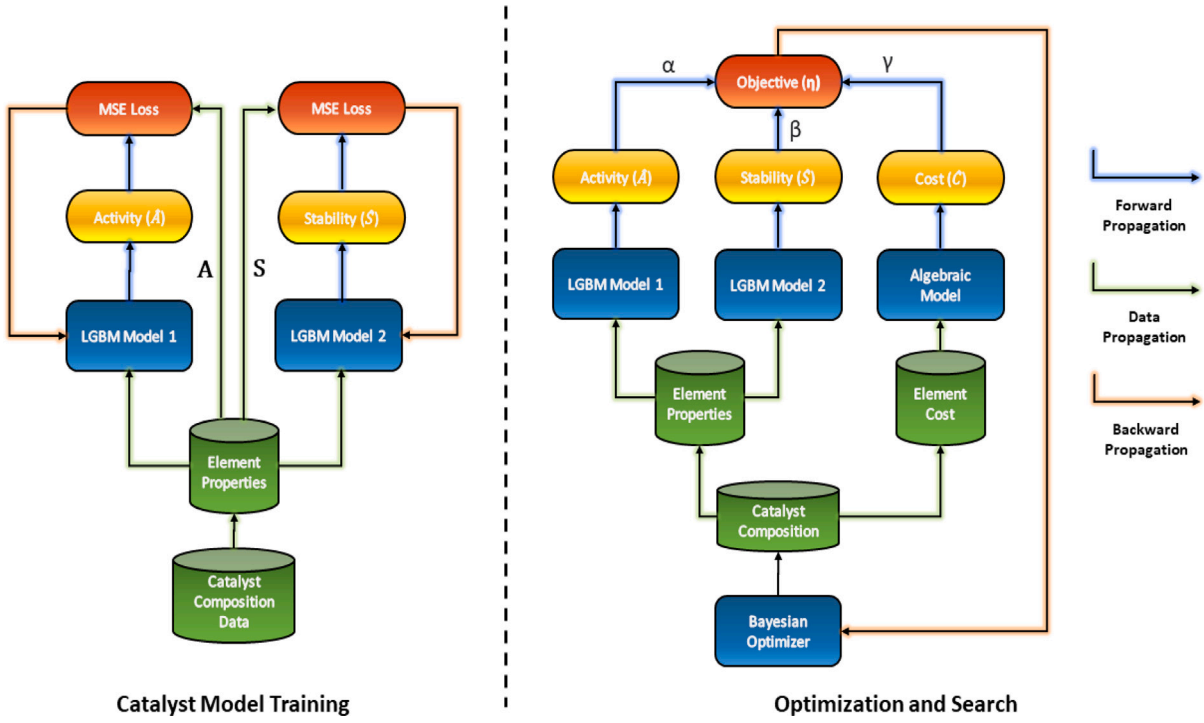


Fig. 1. Catalyst screening process.

2.1. Machine learning model for catalyst performance prediction

Several catalytic properties serve as crucial performance metrics in our study. These encompass activity, selectivity, stability, and cost, significantly influencing feasibility in large-scale and laboratory-scale experiments. Activity is approximated as the percentage conversion, while selectivity is determined based on the weight or mole distribution of products. Stability, on the other hand, is estimated using time-on-stream (TOS) measurements i.e the hours of operation. The cost factor encompasses ingredient or element prices, production expenses, and storage costs. These targets are numeric in nature, framing our study as a regression task.

Our study considers a range of operating parameters, including reaction temperature, feed composition, contact time, calcination temperature & time. Additionally, we examine parameters related to the properties of the elements composing the catalyst. These properties, such as electronegativity, surface energy, atomic radius, melting point, and boiling point, serve as features to the ML model. Leveraging these numeric input features, we formulate our ML objective centred around a regression task. This task involves predicting the CO conversion as a percentage (0–100), Time on Stream TOS (min) scaled between 0–1 using ML models. We assess various ML models, including Linear Regression (Su et al., 2012), Support Vector Machine (Noble, 2006), Artificial Neural Network (ANN) (Jain et al., 1996), Decision Tree (Myles et al., 2004), Random Forest (Breiman, 2001), as well as gradient boosting frameworks such as XGBoost (Chen and Guestrin, 2016), and LightGBM (Ke et al., 2017). The catalyst cost has been modelled using an approximate algebraic model involving element costs, which we procured from the literature. Other factors which include storage and production costs can vary widely. Therefore for simplicity, we do not consider these though they can be incorporated along with another ML model for cost prediction.

Scientific tasks like this where a physical understanding holds paramount importance, benefit significantly from tree-based models due to their simple explainability. LGBM operates through gradient boosting with tree-based algorithms, sequentially developing multiple weak learners (decision trees). Each successive tree focuses on rectifying the errors or the residuals of the preceding trees, creating a highly efficient ensemble. Employing specific features for dataset splitting provides an ordered list of essential features, offering invaluable physical insights into the dominant factors. Like many advanced machine learning models, LGBM involves several hyperparameters. The large diversity of hyperparameters hold varying importance. Some can ensure that the model does not overfit to the data while others can control how quickly the model converges. Determining the optimal set of hyperparameters is a non-trivial task. Instead of relying on a brute-force approach with numerous trial-and-error iterations, we implement a Bayesian optimizer. This optimizer samples sets of hyperparameters to identify the most optimal configuration, reducing computation time significantly. We employ root mean squared error loss which is a standard in literature as our objective function. The culmination of our work involves training two ML models, each for activity and stability,. Selectivity is not considered in this study due to a lack of data. With these predictive models developed and trained, we advance to our second objective of catalyst screening.

2.2. Catalyst screening using Bayesian optimization

Our present study also involves a catalyst screening procedure in which we harness the fast predictive power of our trained models to explore the catalyst composition space. Within this space, defined by active metal, support, and promoter components there exist an infinite number of potential compositions, forming a continuous 3-dimensional domain. The practicality of comprehensive experimental exploration is hindered by the considerable time and cost involved. This challenge arises from the non-linearity of the composition space with respect to performance metrics of the catalysts. In light of these constraints, computational simulations offer a viable alternative. However, traditional simulations cannot predict conversion percentages accurately. Herein, ML-driven prediction models emerge as a valuable solution. A robust prediction model effectively acts as a simulator, enabling efficient exploration

Even with enhanced sampling efficiency, the catalyst composition space remains infinitely vast. To structure our search across this expansive domain, we employ Bayesian optimization. This methodology

facilitates focused exploration by disregarding regions likely to yield inferior results compared to existing samples. Consequently, we can efficiently pinpoint optimal compositions while also considering unexplored compositions with the potential for superior performance. Our study also identifies the ideal combination of elements for forming an optimal catalyst. Given that not all combinations might be physically feasible, we provide the flexibility for researchers with domain-specific knowledge to specify the set of components in advance. Our implementation can subsequently identify the optimal composition within the constraints of the specified elements. An additional feature of our work centres on the adaptability of each catalyst performance metric's importance in the optimal search. The objective function, designed to maximize the catalyst's performance, comprises a non-linear combination of activity, stability, and cost. Researchers can dynamically adjust the weights associated with each metric to reflect their specific priorities. For instance, in small-scale laboratory settings where cost considerations are paramount, the weight assigned to cost can be increased. Conversely, the corresponding weight can be augmented in industrial applications where stability takes precedence. This inherent flexibility caters to the diverse requirements within the context of the setting of the water-gas shift reaction.

2.3. Catalyst coefficient

To effectively compare and model the importance placed on each catalyst performance metric — activity, stability and cost, we introduce a coefficient η. The coefficient is a weighted sum of the exponential of the catalyst performance metrics where the weights can be tuned as required. The coefficient acts as the optimization objective for the Bayesian optimization algorithm. For activity or normalized CO conversion A, stability or normalized Time on Stream S and normalized catalyst cost C all scaled between 0 to 1, the coefficient is defined as follows:

$$\eta = Sig(\alpha \cdot e^{A} + \beta \cdot e^{S} + \gamma \cdot e^{-C}) \tag{2}$$

where α, β and γ are the respective weightages assigned by the user and Sig is the sigmoid function. As represented by Eq. (2), incorporating exponential functions in the coefficient design is a strategic choice. For catalyst screening, where catalyst performance metrics can be close for multiple samples, the use of exponential increases the sensitivity to the three catalyst performance metrics. We also expect to achieve better numerical stability with the exponential function involved. Additionally, the exponential of cost which is modelled as e^{-C} keeps the value positive.

Furthermore, the sigmoid function has been added to scale the output into 0 and 1. The coefficient expression can, therefore, be interpreted quite easily. Higher coefficient values closer to 1 symbolize a more optimal catalyst. A higher activity, higher stability and lower cost characterize it. Theoretically the highest value that η can attain is nearly 0.94. This is for a catalyst with highest activity, highest stability based on the dataset and no cost associated with it. Therefore, it can act as guide to how good a catalyst is based on its η value.

3. Results and discussion

This section presents the outcomes of our comprehensive study, structured into multiple subsections to provide a detailed analysis of our findings. We begin in Section 3.1 by elucidating the simulation settings employed in our research. Subsequently, in Section 3.2, we present the performance metrics of both the conversion and stability models. In Section 3.3, we delve into an in-depth examination of our proposed catalyst coefficient, exploring its various property-based attributes. Building on these insights, we proceed to Section 3.4, where we showcase four distinctive case studies, each representing distinct requirements for catalyst selection. These case studies underscore the adaptability and efficacy of our approach in addressing diverse needs suitable for industry. Finally, in Section 3.5, we introduce an array of interpretable analyses that enhance our understanding of the factors influencing our results, providing valuable insights into the outcomes derived from our analysis.

Table 1
Comparison of different applied models for Conversion Modelling.

Model	Activity		Stability	
	RMSE Loss	R2 Score	RMSE Loss	R2 Score
LGBM	8.65	0.98	0.07	0.88
LGBM + Hyperparameter Tuning	6.63	0.98	0.04	0.91
LGBM + Hyperparameter Tuning + Cross Validation	**3.52**	**0.99**	**0.004**	**0.94**
Chattoraj et al. (2022) ANN + XGBoost Ensemble	6.87	0.95	–	–

3.1. Simulation setting

The simulations were executed on a computing platform with $8GB$ of RAM and 8 processing threads. Python 3.10 served as the programming environment for code implementation, leveraging essential libraries such as Numpy and Pandas for data collection and preprocessing. Furthermore, for data visualization, we employed Matplotlib and Seaborn. Implementing the models, integral to assessing performance metrics, relied on a selection of libraries, including Scikit-Learn, LGBM, and Bayesian-Optimization. Notably, each model was meticulously trained for 10,000 iterations to ensure a robust performance evaluation. Bayesian optimization played a pivotal role for catalyst screening, meticulously running for 1000 iterations for each case considered in this extensive study.

3.2. Model performance analysis

In our study, we have applied our proposed ML-based algorithm to model two pivotal parameters within the dataset: CO Conversion and TOS (min). To gauge the efficacy of our methodology, we have undertaken a comprehensive comparative analysis, benchmarking our proposed model against other cited work. The ensuing sections, Sections 3.2.1 and 3.2.2, present a detailed exposition of the results obtained for the conversion model and the stability model, respectively.

3.2.1. Conversion model results

The models under consideration encompass linear regression, decision trees, artificial neural networks (ANNs), and support vector machines (SVM), Light Gradient Boosting Method (LGBM) and Extreme Gradient Boosting (XGBoost) The results of this comparison are presented in Table 1. Our label data ranges from 0 to 100 for the modelling task, representing percentage conversion. We employ Root Mean Square Error (RMSE) as our regression error metric, along with the R^2 Score Ash and Shwartz (1999) for further evaluation. Our analysis reveals that traditional statistical models such as Linear Regression, Support Vector Machines, and Decision Trees exhibit suboptimal performance in this regression task. This finding underscores the inherent complexity and non-linearity of the conversion model, necessitating more advanced approaches.

Our results indicate that gradient boosting algorithms, specifically XGBoost and LGBM, outperform traditional techniques, highlighting the advantage of leveraging sophisticated ML methods. Through hyperparameter tuning, we substantially reduce RMSE, reducing it from 8.648 (LGBM) to 6.633. To further enhance performance, we employ cross-validation to optimize the training and testing set distribution, resulting in an even lower RMSE of 3.523. This analysis demonstrates the potential for our proposed model to predict CO conversion with a marginal error of 3%. These outcomes underline our approach's efficacy in addressing the conversion model's intricacies, showcasing its ability to improve predictive accuracy significantly.

Table 2
Maximum coefficient value for different α, β, γ & Catalysts element combinations.

α	β	γ	η	Element combination	Remarks
0.1	0.1	0.8	0.71	(Mg, Pt)	Economically Feasible Catalyst
0.1	0.8	0.1	0.91	(Ti, Ce, Pt)	Highly Stable Catalysts
0.8	0.1	0.1	0.91	(Cu, Pd, Ce)	Catalysts with High Conversion
0.4	0.3	0.3	0.85	(Cu, Pd, Ce)	Balanced Catalysts with good conversion
0.3	0.4	0.3	0.85	(Ti, Ce, Pt)	Balanced Catalysts with good stability
0.3	0.3	0.4	0.80	(Cu, Pd, Ce)	Balanced Catalysts with good economic value

3.2.2. Stability model results

In the context of stability modelling, we extended the evaluation to compare the model's performance mentioned in the previous Section 3.2.1. The comparative results are presented in Table 1. For the regression task, we utilized the Root Mean Square Error (RMSE) as the error function, and the R^2 score served as our chosen regression metric, enabling efficient comparisons between the models under consideration.

The results echo similar trends observed during the conversion modelling task. Conventional techniques such as Linear Regression, Support Vector Machines (SVM), Decision Trees, and Artificial Neural Networks (ANN) demonstrated limited performance in regression analysis. In contrast, ML approaches, including Random Forest and Gradient Boosting algorithms such as XGBoost and LGBM, exhibited relatively robust performance on the testing dataset. Notably, by applying hyperparameter tuning, we achieved a noteworthy reduction in RMSE error, reaching an impressive value of 0.039. Furthermore, a meticulous 100-fold cross-validation process was executed to identify the optimal training and testing sets. This extensive cross-validation exercise further improved, lowering the error to an order of 1×10^{-3}, equivalent to an RMSE value of 0.00419. Notably, the R^2 score for the LightGBM model, combined with hyperparameter tuning and cross-validation, excelled at approximately 0.94, surpassing the performance of previous models. Our observations conclude that our approach outperforms not only traditional methods but also exhibits superior performance when compared to state-of-the-art techniques in stability modelling.

3.3. Coefficient analysis results

To optimize catalyst selection, we have introduced a novel catalyst activity coefficient. In Section 3.4, we thoroughly elucidate the catalyst's requisites and our proposed approach to engineering them. The coefficient's range spans from 0 to 1, and its value is contingent upon the significance coefficients, denoted as α, β, and γ, for the conversion (activity), time on stream (TOS, stability), and cost (economic) parameters, respectively. Our analysis reveals a crucial interplay between these coefficients. Maintaining α as a constant, augmenting γ and correspondingly diminishing β lead to a reduction in the maximum coefficient value. We further investigate the influence of α and β, demonstrating a direct proportionality between these coefficients and the catalyst activity, as evidenced in Figs. 2(a) and 2(b), respectively. Conversely, the coefficient shows an inverse relationship with γ, as substantiated by the qualitative trend diagram in Fig. 2(c). This observation aligns with the coefficient's design equation, as indicated in Eq. (2). These findings deepen our understanding of the catalyst activity coefficient and offer valuable insights for optimizing catalyst selection in diverse industrial applications.

3.3.1. Catalysts combination based on coefficient value

In this subsection, we present the results of our exhaustive investigation into diverse combinations of catalysts, each characterized by distinct values of α, β, and γ, to identify the catalyst combination that maximizes the coefficient value. The findings reveal several noteworthy insights as summarized in Table 2. Firstly, **Pt with a MgO support structure** emerges as a compelling catalyst composition choice for small scale labs as it represents a situation where cost is an important factor. Secondly, Pt catalysts in conjunction with Ce support and Ti promoters, offer a balanced solution, demonstrating notable stability and activity. This makes them particularly well-suited for applications demanding prolonged operational duration with a high conversion requirement. Lastly, the Pd-Cu catalysts with Ce support present another balanced option, featuring an attractive economic choice and exceptional conversion efficiency. These findings provide invaluable guidance for strategically selecting catalysts tailored to the specific requirements of diverse industrial or academic contexts.

Table 3
Screening for high conversion catalysts.

Cu	Pd	Ce	Reaction temperature °C	A	S	C	η
43.47	1.15	55.36	278.24	0.7212	0.36	2.85	**0.8686**
21.53	1.03	77.4	282.83	0.72	0.34	2.86	0.8681
24.42	0.99	74.57	282.61	0.71	0.36	2.76	0.8680
29.30	1.22	69.47	277.34	0.71	0.37	3.46	0.8668
38.33	1.43	60.22	283.81	0.71	0.36	3.11	0.8666

3.4. Screening analysis

We have performed four case studies for screening analysis using our optimization framework. These case studies include screening for different types of catalysts for WGS reactions like Catalysts with high conversion [3.4.1], highly stable catalysts [3.4.2], Catalysts with low economic cost [3.4.3], and balanced catalysts [3.4.4] considering all the parameters.

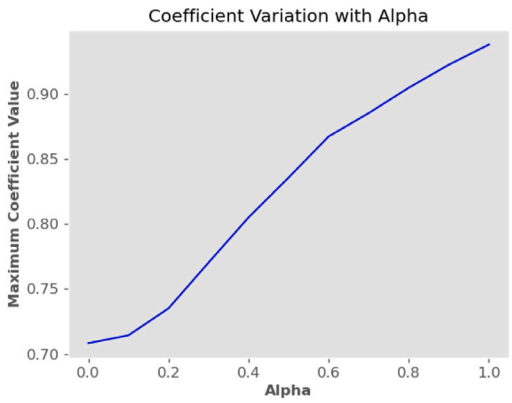
3.4.1. High conversion catalysts

As depicted in Table 2, we initially identified the Pd-Cu catalysts with Ce support as a balanced catalyst combination demonstrating high conversion. Utilizing this composition as a starting point, we embarked on a more intricate screening process to discover an optimized composition of each component to yield a higher coefficient value. Bayesian Optimization was employed to navigate the complex input space, with selected values for α, β, and γ set at 0.8, 0.1, and 0.1, respectively. Experimental data revealed a maximum coefficient value of 0.9146 for this configuration, signifying its promising potential.

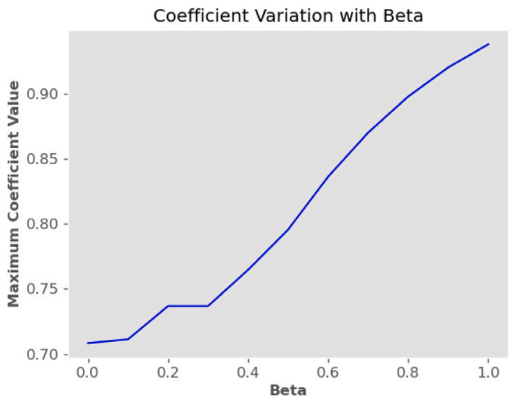
Table 3 contains the top five outcomes from an extensive screening process of over 1000 iterations. Notably, the Cu-Pd-based catalysts tended to lower reaction temperatures than other catalyst systems, indicating energy efficiency. The highest coefficient value obtained from the screening process reached 0.8686, laying the foundation for further investigations to optimize this coefficient value. These findings underscore the capacity for refinement and improvement in catalyst design, with the potential for even higher conversion rates in subsequent iterations.

3.4.2. Highly stable catalysts

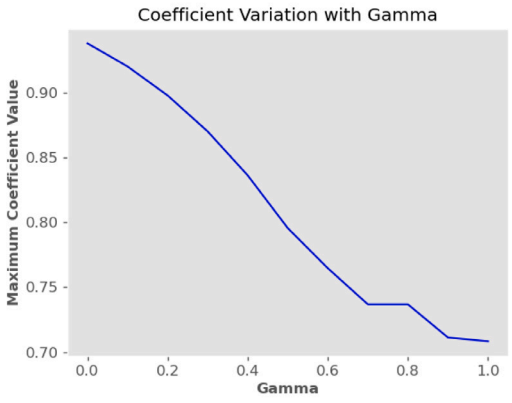
In Table 4 we have showed the results of screening for a stable catalyst. From Table 2 we can observe that a catalyst combination of Ti, Pt, Ce provides the highest coefficient values for a stable catalyst. Investigating the literature it is found that the most efficient catalyst set is "Pt catalyst doped with Ce and supported on TiO2". To further improve this composition, we conducted a screening analysis within



(a) Coefficient variation with α



(b) Coefficient variation with β



(c) Coefficient variation with γ

Fig. 2. Coefficient variation with different parameters.

the extensive space of catalyst properties. Similar to our previous case study, Bayesian optimization played a crucial role in the screening process. For this investigation, we set the values of α, β, and γ at 0.1, 0.8, and 0.1, respectively. The maximum coefficient value for this optimized combination is 0.9146, based on experimental data. Analysing the results in Table 4, we observe that the highest coefficient value achieved through screening is approximately 0.7912. Notably, the analysis indicates that the cost metric significantly impacts the coefficient value, contributing to its reduction compared to the maximum value reported in the literature. Additionally, the regression R^2 score of about 0.93 in our stability model suggests that, in some cases, the regression error may contribute to this unexpected reduction in the coefficient value. From the literature presented by Ammal and Heyden (2013) we have found that, the low-temperature Water-Gas Shift (WGS) reaction on TiO2(110)-supported Pt clusters with Ce as a doping material follows a novel CO-promoted redox mechanism, whereas the high-temperature WGS reaction adheres to a classical redox mechanism. The Pt–TiO2 interface sites demonstrate activity that is two orders of magnitude higher than that of the Pt(111) surface sites. Elementary processes occurring on the TiO2 surface, such as H2O dissociation and H diffusion, play a significant role in controlling the overall reaction rate. This is why the catalyst system is a stable one. A first-principles-based microkinetic model provides deep insights into the unique activity of Pt/TiO2 catalysts, enhancing our understanding of their catalytic performance.

3.4.3. Economic catalysts

Our analysis from Table 2 shows that a catalyst combination of Mg, Pt and Ce provides the cost effective catalyst. Upon performing the screening on the input space of catalyst compositions of the mentioned

Table 4
Screening for high stability catalysts.

Ti	Ce	Pt	Reaction Temperature °C	A	S	C	η
85.00	0.40	13.00	279.16	0.62	0.27	33.43	**0.7913**
85.00	0.40	13.00	271.13	0.59	0.27	33.43	0.7910
85.00	0.40	13.00	264.12	0.61	0.26	33.43	0.7907
98.72	0.40	0.88	290.53	0.62	0.26	37.69	0.7906
99.25	0.40	0.35	283.42	0.62	0.26	28.66	0.7904

elements and the reaction environment. Upon going through the literature we discovered that the best combination with these catalyst set is “Mg doped with Pt and supported on CeO_2.

Table 5 shows the results of screening analysis. By taking α, β, and γ as 0.1, 0.1, and 0.8 respectively, the highest experimental coefficient value achieved is 0.71. To further investigate the composition and property space of this catalyst, we utilized Bayesian optimization, conducting 1000 iterations in the screening process. Detailed results are shown in Table 5. Remarkably, the maximum coefficient value obtained through the screening process is approximately 0.7571, which is higher than the maximum coefficient value for the α, β, γ combination from Table 2. This suggests that the screened catalyst composition in this case study outperforms the previously reported catalyst in terms of cost-effectiveness. Furthermore, each catalyst within this Mg-Pt-based system requires a relatively high reaction temperature, consistent with the characteristics of this low-cost catalyst configuration.

From the works of Molinet-Chinaglia et al. (2024) we can observe more information on the specified catalysts combination. It is inferred that The Pt molar activity increases by a factor of 2.5 when the Pt content rises from 0.1 to 0.6 wt%. Notably, Pt0 nanoparticles exhibit

Table 5
Screening for low-cost catalysts.

Pt	Mg	Ce	Reaction Temperature °C	A	S	C	η
0.90	1.80	97.30	400.09	0.70	0.29	0.64	**0.7571**
0.10	1.80	98.10	402.04	0.70	0.29	0.07	0.7571
0.90	1.80	97.30	399.09	0.67	0.33	0.64	0.7570
0.90	27.08	72.02	400.71	0.69	0.30	0.66	0.7569
0.90	1.80	97.30	403.69	0.70	0.29	0.64	0.7568

Table 6
Screening for balanced catalysts.

Pd	Cu	Ce	Reaction Temperature °C	A	S	C	η
1.40	27.23	71.36	262.05	0.69	0.38	2.61	**0.8229**
0.93	38.80	60.25	280.37	0.70	0.36	4.32	0.8229
1.03	31.19	67.77	262.11	0.68	0.37	3.93	0.8223
1.53	50.50	47.96	261.72	0.68	0.38	3.23	0.8219
0.82	37.24	61.93	278.41	0.69	0.36	4.25	0.8216

greater activity compared to PtOx single atoms and small clusters. Beyond 0.6 wt% and up to 1.7 wt% of Pt, the Pt molar activity remains fairly constant. This plateau is attributed to the increase in the number of Pt0 nanoparticles without any sintering occurring. Also it is shown that CeO_2 acts as a very stable support materials. Overally the following catalyst composition is a cost effective catalyst combination.

3.4.4. Balanced catalysts

In continuation of our exploration, this case study is geared towards screening balanced catalysts, encompassing a synthesis of critical factors — conversion, stability, and cost. For this analysis, we set the values of α, β, and γ to 0.4, 0.3, and 0.3, respectively, based on our catalyst coefficient analysis, as presented in Table 2. The screening process unveiled notable findings in Table 6. Among these, a prominent observation is the maximum coefficient value achieved, which approaches approximately 0.83. It is close to the maximum coefficient value for the α, β, γ combination from Table 2, reinforcing the reliability of our approach. Notably, the catalyst in this scenario is Cu-Pd based, characterized by a relatively lower reaction temperature. Ce is the support material in this configuration, leading to a high composition ratio.

This screening methodology enables the exploration of catalyst combinations documented in the literature and opens the door to uncharted territory by allowing for the customization of catalysts. This flexibility presents a promising avenue for discovering novel catalysts tailored to the specific needs of the water gas shift reaction.

3.5. Explainable analysis

This section delves into the comprehensive interpretation of the results obtained from our modelling efforts. This detailed analysis significantly enhances the interpretability and reliability of our models. Our approach extends to both the conversion and stability models, where we have conducted an extensive explainable analysis to shed light on the intricacies of our models. These analyses encompass various aspects, primarily focusing on the importance of features in influencing model performance. By scrutinizing the significance of each feature, we gain a deeper understanding of their contributions to the overall model output. Moreover, our analysis also explores the dependencies between feature values and the model's predictions. This in-depth examination of value dependencies adds another layer of insight to our modelling results. The outcome of this interpretative analysis not only bolsters the trustworthiness of our models but also provides valuable insights for researchers and practitioners in the field. This newfound knowledge can be leveraged to make informed decisions, optimize processes, and advance the state-of-the-art in heterogeneous catalysis.

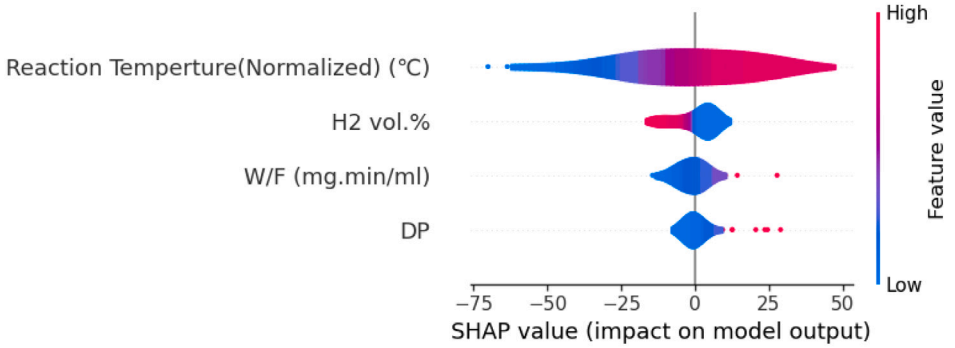
3.5.1. Conversion model explanation

To gain insight into the predictive factors driving our conversion model, we employed SHAP (SHapley Additive exPlanations) for explainable analysis. As depicted in Fig. 3(a), our investigation into feature importance revealed critical insights into the model's operation. Notably, the Reaction Temperature emerged as a dominant factor, exerting a pronounced influence on Carbon Monoxide (CO) conversion prediction. This observation underscores the sensitivity of the conversion process to variations in temperature. Furthermore, our analysis identified several other vital features, including the percentage volume of Hydrogen in the feed system, the contact time of catalysts and input materials, catalyst preparation methods such as deposition–precipitation (DP), and the percentage volume of water in the feed stream. These attributes ranked among the most influential in shaping the conversion model's predictions. The pronounced role of these features in CO conversion underscores their significance in catalytic processes and provides valuable guidance for catalyst design and optimization.

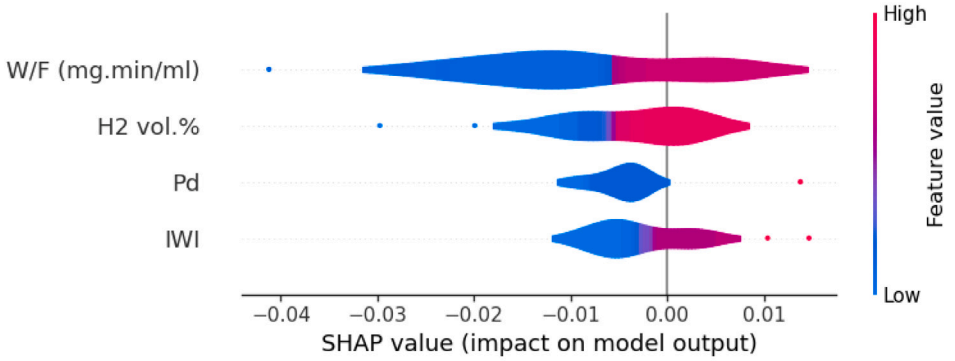
In Fig. 4, we present a detailed exploration of the relationship between reaction temperature and SHAP importance values, illustrating the gradient of contact time (W/F). The figure reveals a compelling trend where an increase in reaction temperature corresponds to an escalation in SHAP values for this specific feature. This observation signifies the influence of reaction temperature on the model's predictions. We obtained Fig. 6, delineating distinct boundaries for each critical feature to provide a more comprehensive view of the feature dependencies. In Sub Fig. 6(a), a clear transition between positive and negative effects of the reaction temperature is discernible, indicating the positive correlation reaction temperature has with the CO conversion. As the WGS reaction is endothermic, the reaction temperature increases the kinetics and shifts the thermodynamic equilibrium. The model's explainable analysis reinforces the theoretical understanding, thus reducing the model's opaqueness. Similarly, Sub Figs. 6(b) delve into the impacts of percentage H_2 volume, contact time, and percentage H_2O volume, respectively. These visualizations further act as sanity checks for our model's working. Increasing the H_2 volume in the feed has an inverse effect on the equilibrium, as illustrated by the decreasing SHAP values, while increasing the H_2O percentage in the feed has a positive impact. Increasing the contact time provides a larger catalyst surface area for the reaction, improving the conversion or activity. The model's correlations can be understood and clarified through these illustrations with our theoretical understanding.

3.5.2. Stability model explanation

This section presents a comprehensive explanatory analysis of the Time on Stream (TOS) prediction model. Our findings shed light on the factors influencing the model's predictions, offering valuable insights into the TOS modelling process. As illustrated in Fig. 3(b), our analysis reveals that contact time is the most critical feature in predicting TOS. This is evident from the contact time feature's significant and inversely proportional impact on its associated SHAP values. Notably, variations in contact time have a substantial effect on the TOS predictions, underlining its pivotal role in the catalytic process. Furthermore, our analysis identifies several other features as key contributors to the TOS prediction model's performance. Among these, the percentage volume of H_2, the presence of Pd, the use of Incipient Wetness Impregnation (IWI) as a preparation method, and the presence of La in the catalyst composition emerge as highly influential factors. These features collectively shape the TOS predictions and play a vital role in determining the catalyst's lifetime. Interestingly, our analysis reveals that reaction temperature does not exert a significantly profound influence on TOS predictions. While temperature remains an important parameter in catalytic reactions, our model appears to emphasize other features more when forecasting TOS. This explanatory analysis provides a deeper understanding of our TOS prediction model's inner workings and highlights the key features driving its predictions. These insights



(a) Conversion Model



(b) TOS Model

Fig. 3. Important features for modelling.

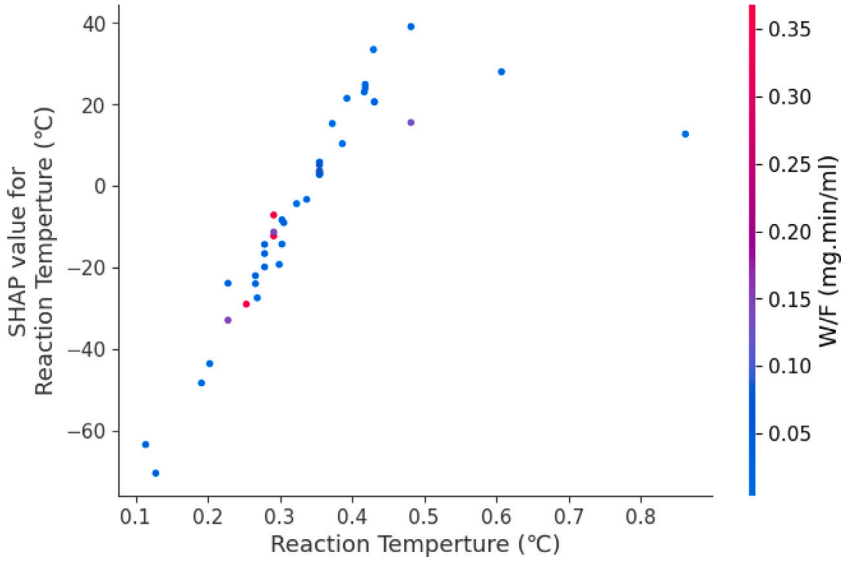


Fig. 4. Reaction temperature dependence for conversion model.

are can be quite valuable to researchers and practitioners in the field of heterogeneous catalysis, offering a clearer perspective on the factors governing catalyst performance and longevity.

In Fig. 5, we present the dependence plot for the model concerning the reaction temperature. The scatter plot reveals a trend: as the reaction temperature increases, the catalyst TOS predictions experience a decrease. However, this effect is not linear, suggesting a non-monotonic relationship. Our analysis infers that elevated reaction temperatures have a discernibly negative impact on the stability of the catalyst, which matches theoretical understanding. Additionally, in Fig. 7, we present force plots for various features, including contact time, the percentage of H_2 volume, Pd content, and La composition. In Fig. 7(a), it is apparent that prolonged contact time positively affects TOS and, hence, stability. This closely matches theoretical understanding as a large amount of the catalyst would take longer to deactivate for a given feed flow rate. The force plot in Fig. 7(b) suggests that an increased concentration of H_2 positively influences stability, albeit not to a significant degree.

Figs. 7(c) and 7(d) provide insights into the impact of Pd on the model's stability. In the case of Pd, we observe a fuzzy nature of the

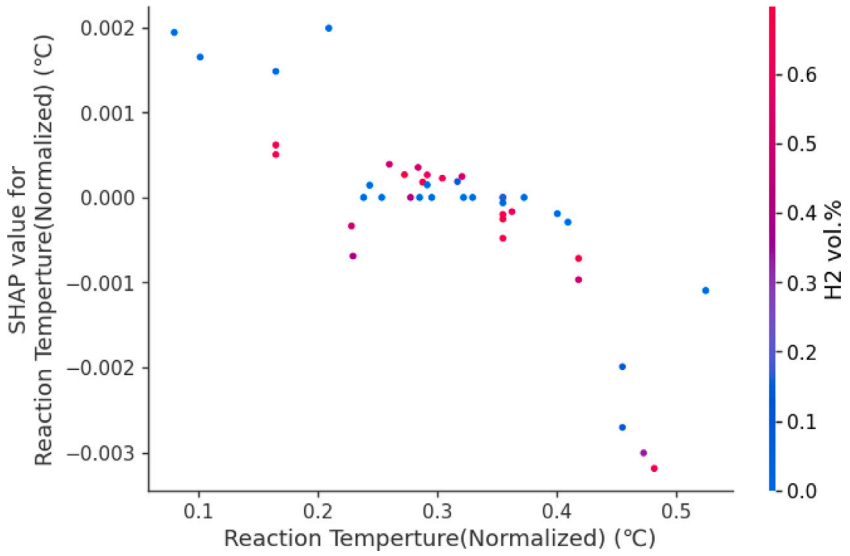


Fig. 5. Reaction temperature dependence for stability model.

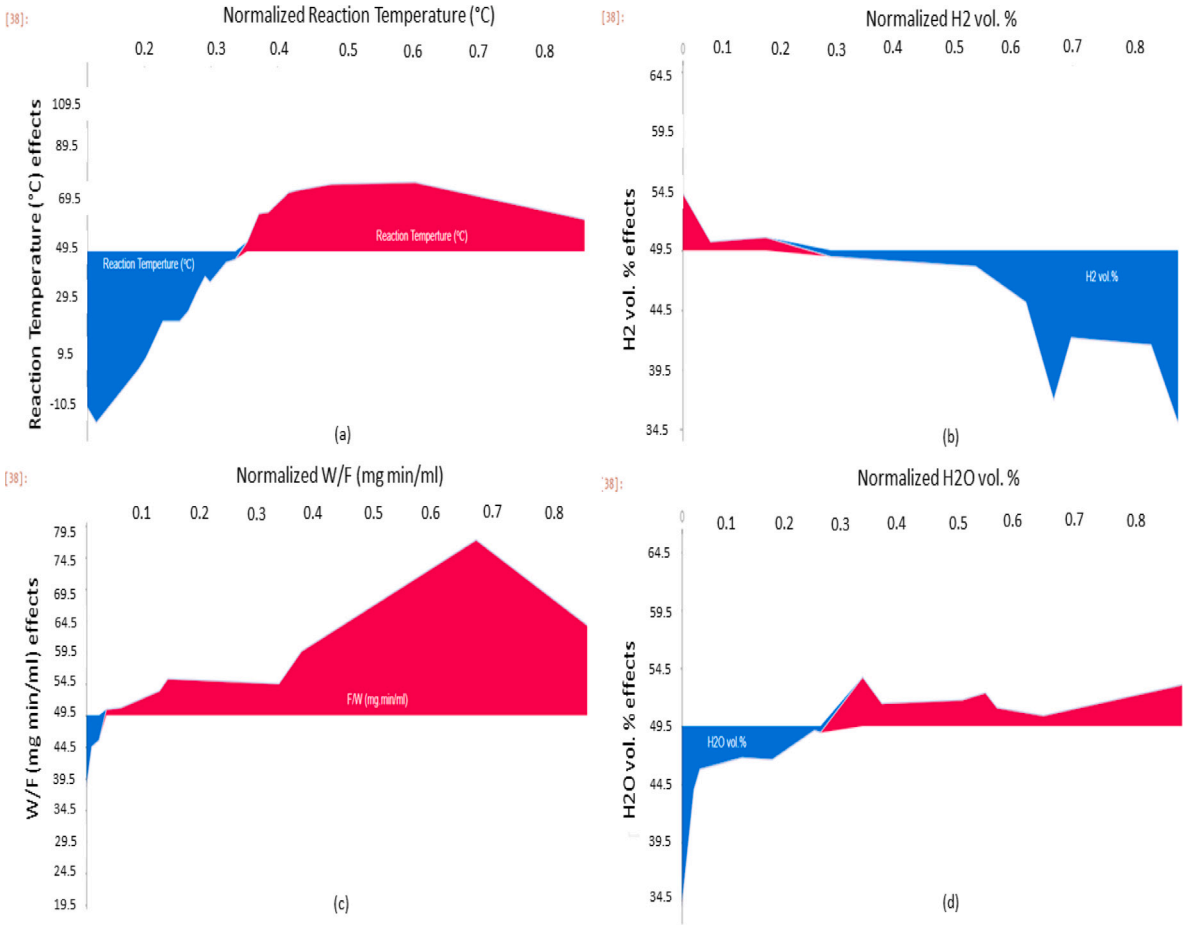


Fig. 6. SHAP on conversion model for (a) reaction temperature (b) H2 vol. % (c) contact time (d) H20 vol. %.

feature effect characterized by a positive influence with a degree of fuzziness. Conversely, for IWI, the feature effect is fuzzy in the opposite direction, suggesting a negative influence on model stability. The effect of these features is not strong, as the presence of Pd or a specific preparation method does not necessarily justify stability changes to the catalyst. These analyses offer valuable insights into the intricate interplay between various features and the TOS of the model, contributing to a deeper understanding of the factors influencing catalytic stability.

4. Conclusion

Our research has contributed significantly to catalyst screening and design theory and practice. Theoretically, we introduced a versatile pipeline-based architecture with the potential to customize the screening of datasets across diverse domains. This adaptable pipeline enables the accelerated discovery of effective catalysts for various applications, providing a crucial tool for future catalyst development. We have also

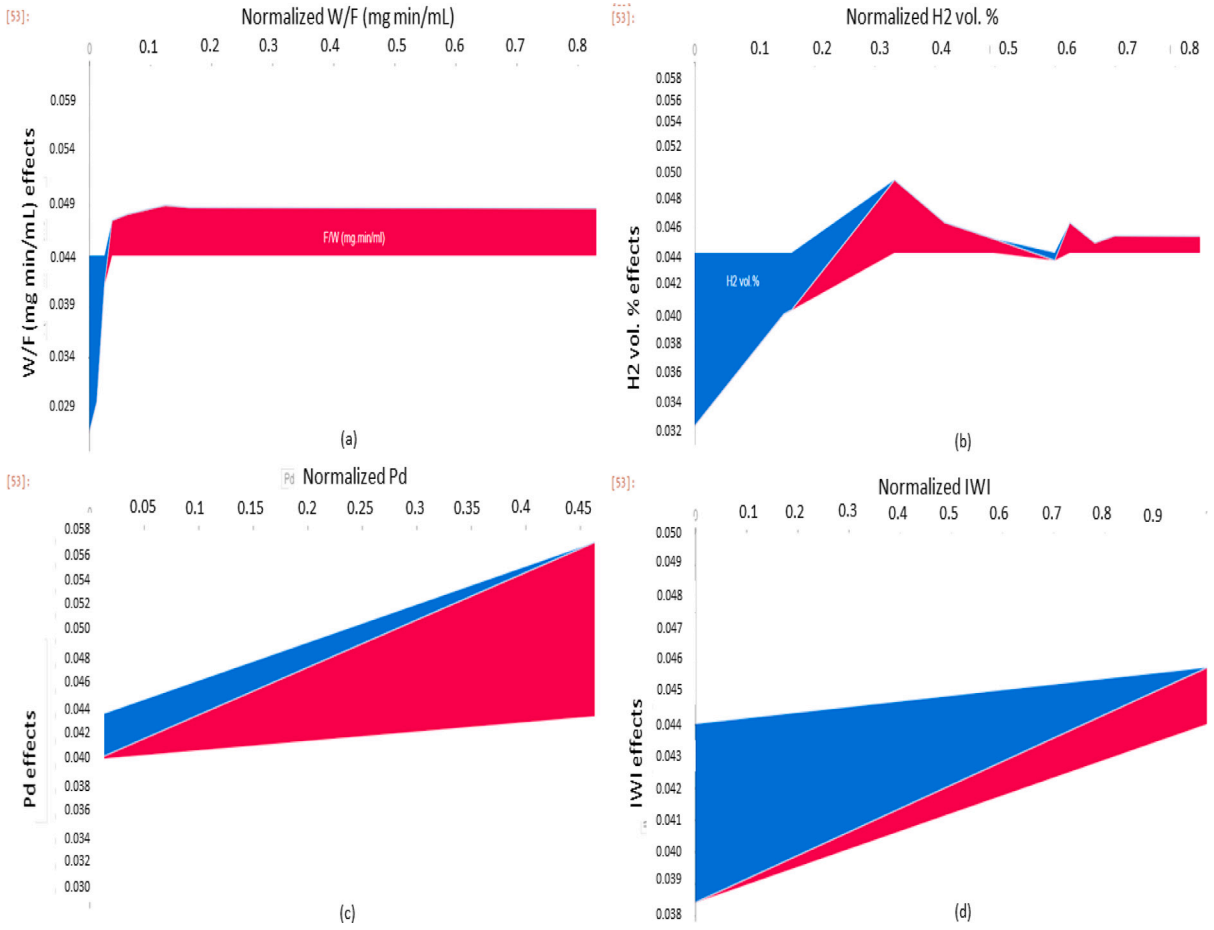


Fig. 7. SHAP on stability model for (a) contact time (b) H2 vol. % (c) Pd (d) IWI.

presented a novel catalysts coefficient, which combines critical catalyst attributes, including stability, activity, and economic feasibility, into a comprehensive measure of catalyst quality. This coefficient facilitates the evaluation and selection of catalysts for various reaction systems, marking a substantial advancement in the field. In practical terms, our research promises to optimize catalyst compositions to meet specific industrial requirements effectively. By tailoring catalysts to industrial needs, our approach has the potential to enhance the performance of industrial setups significantly.

Moreover, our research can evolve into a dedicated software system for systematic catalyst screening, reducing the need for extensive manual work and saving valuable time and resources. Our work aims to streamline catalyst selection in industrial settings by automating the screening process and driving operational efficiency. Furthermore, our research emphasizes the practice of explanatory analysis to demystify the screening process results. This increased interpretability transforms the 'black box' system into a tool that bridges scientific findings with industrial decision-makers. Enhanced communication and collaboration between academia and industry professionals are fostered, contributing to more informed decision-making and advancing the catalysis field. However, it is crucial to acknowledge limitations, including refining our time modelling system for improved predictive accuracy and the potential to enhance catalyst system stability screening. These limitations create opportunities for future research to build upon our work and further refine our methodologies. Looking ahead, we are committed to advancing our optimization framework to address sensitive optimization tasks within the heterogeneous catalysis industry. We aim to refine techniques and algorithms for more precise catalyst screening and develop a software system to expand our methodology to a broader spectrum of catalyst reaction systems. The convergence of advanced computational methods and domain-specific expertise will propel catalyst discovery and design into an era of efficiency and innovation.

Funding

No project fund was involved.

CRediT authorship contribution statement

Rahul Golder: Writing – original draft, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Shraman Pal:** Writing – original draft, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Sathish Kumar C.:** Validation, Investigation, Conceptualization. **Koustuv Ray:** Writing – review & editing, Supervision, Investigation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.dche.2024.100165.

References

Ammal, S.C., Heyden, A., 2013. Origin of the unique activity of Pt/TiO2 catalysts for the water–gas shift reaction. J. Catal. 306, 78–90.

Artrith, N., Lin, Z., Chen, J.G., 2020. Predicting the activity and selectivity of bimetallic metal catalysts for ethanol reforming using machine learning. ACS Catal. 10 (16), 9438–9444. http://dx.doi.org/10.1021/acscatal.0c02089.

Ash, A., Shwartz, M., 1999. R2: a useful measure of model performance when predicting a dichotomous outcome. Statist. Med. 18 (4), 375–384.

Ayodele, B.V., Alsaffar, M.A., Mustapa, S.I., Kanthasamy, R., Wongsakulphasatch, S., Cheng, C.K., 2021. Carbon dioxide reforming of methane over Ni-based catalysts: Modeling the effect of process parameters on greenhouse gasses conversion using supervised machine learning algorithms. Chem. Eng. Process. 166, 108484. http://dx.doi.org/10.1016/j.cep.2021.108484, URL https://www.sciencedirect.com/science/article/pii/S0255270121001847.

Baumes, L.A., Serra, J.M., Serna, P., Corma, A., 2006. Support vector machines for predictive modeling in heterogeneous catalysis: a comprehensive introduction and overfitting investigation based on two real applications. J. Comb. Chem. 8 (4), 583–596.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Chattoraj, J., Hamadicharef, B., Kong, J.F., Pargi, M.K., Zeng, Y., Poh, C.K., Chen, L., Gao, F., Tan, T.L., 2022. Theory-guided machine learning to predict the performance of noble metal catalysts in the water-gas shift reaction. ChemCatChem 14 (16), e202200355. http://dx.doi.org/10.1002/cctc.202200355, arXiv:https://chemistry-europe.onlinelibrary.wiley.com/doi/pdf/10.1002/cctc.202200355 URL https://chemistry-europe.onlinelibrary.wiley.com/doi/abs/10.1002/cctc.202200355.

Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining. pp. 785–794.

Chen, L., Li, X.-T., Ma, S., Hu, Y.-F., Shang, C., Liu, Z.-P., 2022. Highly selective low-temperature acetylene semihydrogenation guided by multiscale machine learning. ACS Catal. 12 (24), 14872–14881. http://dx.doi.org/10.1021/acscatal.2c04379.

de Oliveira, G.S., Pacheco, H.P., 2022. CatS: A predictive and user-friendly framework based on machine learning models for the screening of heterogeneous catalysts. Mol. Catal. 527, 112430. http://dx.doi.org/10.1016/j.mcat.2022.112430, URL https://www.sciencedirect.com/science/article/pii/S2468823122003169.

Deng, B., Chen, P., Xie, P., Wei, Z., Zhao, S., 2023. Iterative machine learning method for screening high-performance catalysts for H2o2 production. Chem. Eng. Sci. 267, 118368. http://dx.doi.org/10.1016/j.ces.2022.118368, URL https://www.sciencedirect.com/science/article/pii/S0009250922009538.

Denny, S.R., Lin, Z., Porter, W.N., Artrith, N., Chen, J.G., 2022. Machine learning prediction and experimental verification of Pt-modified nitride catalysts for ethanol reforming with reduced precious metal loading. Appl. Catal. B 312, 121380. http://dx.doi.org/10.1016/j.apcatb.2022.121380, URL https://www.sciencedirect.com/science/article/pii/S0926337322003216.

Esterhuizen, J.A., Goldsmith, B.R., Linic, S., 2022. Interpretable machine learning for knowledge generation in heterogeneous catalysis. Nat. Catal. 5 (3), 175–184. http://dx.doi.org/10.1038/s41929-022-00744-z.

Faria, R.d., Capron, B.D.O., Secchi, A.R., de Souza, M.B., 2022. Where reinforcement learning meets process control: Review and guidelines. Processes 10 (11), http://dx.doi.org/10.3390/pr10112311, URL https://www.mdpi.com/2227-9717/10/11/2311.

Fujinuma, N., DeCost, B., Hattrick-Simpers, J., Lofland, S.E., 2022. Why big data and compute are not necessarily the path to big materials science. Comm. Mater. 3 (1), 59. http://dx.doi.org/10.1038/s43246-022-00283-x.

Goldsmith, B., Esterhuizen, J., Bartel, C., Sutton, C., Jin-Xun, L., 2018. Machine learning for heterogeneous catalyst design and discovery. AIChE J. 64, http://dx.doi.org/10.1002/aic.16198.

Gusmão, G.S., Retnanto, A.P., da Cunha, S.C., Medford, A.J., 2021. Kinetics-informed neural networks. arXiv:2011.14473.

Heo, S., Lee, J.H., 2018. Fault detection and classification using artificial neural networks. IFAC-PapersOnLine 51 (18), 470–475. http://dx.doi.org/10.1016/j.ifacol.2018.09.380, URL https://www.sciencedirect.com/science/article/pii/S2405896318320664 10th IFAC Symposium on Advanced Control of Chemical Processes ADCHEM 2018.

Hu, G., Zhou, T., Liu, Q., 2021. Data-driven machine learning for fault detection and diagnosis in nuclear power plants: A review. Front. Energy Res. 9, http://dx.doi.org/10.3389/fenrg.2021.663296, URL https://www.frontiersin.org/articles/10.3389/fenrg.2021.663296.

Jain, A.K., Mao, J., Mohiuddin, K.M., 1996. Artificial neural networks: A tutorial. Computer 29 (3), 31–44.

Jones, D.R., Schonlau, M., Welch, W.J., 1998. Efficient global optimization of expensive black-box functions. J. Global Optim. 13 (4), 455–492. http://dx.doi.org/10.1023/A:1008306431147.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y., 2017. Lightgbm: A highly efficient gradient boosting decision tree. Adv. Neural Inf. Process. Syst. 30.

Khatamirad, M., Fako, E., De, S., Müller, M., Boscagli, C., Baumgarten, R., Ingale, P., d'Alnoncourt, R.N., Rosowski, F., Schunk, S.A., 2023. Data-driven design of enhanced in-based catalyst for CO2 to methanol reaction. ChemCatChem 15 (16), e202300570. http://dx.doi.org/10.1002/cctc.202300570, arXiv:https://chemistry-europe.onlinelibrary.wiley.com/doi/pdf/10.1002/cctc.202300570 URL https://chemistry-europe.onlinelibrary.wiley.com/doi/abs/10.1002/cctc.202300570.

Kim, C., Kim, J., 2022a. Comparative evaluation of artificial neural networks for the performance prediction of pt-based catalysts in water gas shift reaction. Int. J. Energy Res. 46 (7), 9602–9620. http://dx.doi.org/10.1002/er.7829, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/er.7829 URL https://onlinelibrary.wiley.com/doi/abs/10.1002/er.7829.

Kim, C., Kim, J., 2022b. Machine learning-based high-throughput screening, strategical design and knowledge extraction of Pt/CexZr1-xO2 catalysts for water gas shift reaction. Int. J. Energy Res. 46 (15), 21293–21308. http://dx.doi.org/10.1002/er.8488.

Lee, A., Sarker, S., Saal, J.E., Ward, L., Borg, C., Mehta, A., Wolverton, C., 2022. Machine learned synthesizability predictions aided by density functional theory. Comm. Mater. 3 (1), 73. http://dx.doi.org/10.1038/s43246-022-00295-7.

Madeira, L., Portela, M., 2002. Catalytic oxidative dehydrogenation of n-butane. Catal. Rev.-Sci. Eng. 44, 247–286. http://dx.doi.org/10.1081/CR-120001461.

Molinet-Chinaglia, C., Cardenas, L., Vernoux, P., Piccolo, L., Loridant, S., 2024. Tuning the metal loading of Pt/CeO2 catalysts for the water-gas shift reaction. Mater. Today Catal. 100046.

Mou, T., Pillai, H.S., Wang, S., Wan, M., Han, X., Schweitzer, N.M., Che, F., Xin, H., 2023. Bridging the complexity gap in computational heterogeneous catalysis with machine learning. Nat. Catal. 6 (2), 122–136. http://dx.doi.org/10.1038/s41929-023-00911-w.

Musa, E., Doherty, F., Goldsmith, B.R., 2022. Accelerating the structure search of catalysts with machine learning. Curr. Opin. Chem. Eng. 35, 100771. http://dx.doi.org/10.1016/j.coche.2021.100771, URL https://www.sciencedirect.com/science/article/pii/S2211339821001039.

Myles, A.J., Feudale, R.N., Liu, Y., Woody, N.A., Brown, S.D., 2004. An introduction to decision tree modeling. J. Chemom. J. Chemom. Soc. 18 (6), 275–285.

Nikita, S., Sharma, R., Fahmi, J., Rathore, A.S., 2023. Process optimization using machine learning enhanced design of experiments (DOE): ranibizumab refolding as a case study. React. Chem. Eng. 8, 592–603. http://dx.doi.org/10.1039/D2RE00440B.

Noble, W.S., 2006. What is a support vector machine? Nat. Biotechnol. 24 (12), 1565–1567.

Odabaşı, Ç., Günay, M.E., Yıldırım, R., 2014. Knowledge extraction for water gas shift reaction over noble metal catalysts from publications in the literature between 2002 and 2012. Int. J. Hydrog. Energy 39 (11), 5733–5746. http://dx.doi.org/10.1016/j.ijhydene.2014.01.160, URL https://www.sciencedirect.com/science/article/pii/S0360319914002407.

Pandit, N.K., Roy, D., Mandal, S.C., Pathak, B., 2022. Rational designing of bimetallic/trimetallic hydrogen evolution reaction catalysts using supervised machine learning. J. Phys. Chem. Lett. 13 (32), 7583–7593. http://dx.doi.org/10.1021/acs.jpclett.2c01401, PMID: 35950905.

Reiser, P., Neubert, M., Eberhard, A., Torresi, L., Zhou, C., Shao, C., Metni, H., van Hoesel, C., Schopmans, H., Sommer, T., Friederich, P., 2022. Graph neural networks for materials science and chemistry. Comm. Mater. 3 (1), 93. http://dx.doi.org/10.1038/s43246-022-00315-6.

Roy, D., Mandal, S.C., Pathak, B., 2021. Machine learning-driven high-throughput screening of alloy-based catalysts for selective CO2 hydrogenation to methanol. ACS Appl. Mater. Interfaces 13 (47), 56151–56163. http://dx.doi.org/10.1021/acsami.1c16696, PMID: 34787997.

Roy, D., Mandal, S.C., Pathak, B., 2022. Machine learning assisted exploration of high entropy alloy-based catalysts for selective CO2 reduction to methanol. J. Phys. Chem. Lett. 13 (25), 5991–6002. http://dx.doi.org/10.1021/acs.jpclett.2c00929, PMID: 35737450.

Shokry, A., Medina-González, S., Baraldi, P., Zio, E., Moulines, E., Espuña, A., 2021. A machine learning-based methodology for multi-parametric solution of chemical processes operation optimization under uncertainty. Chem. Eng. J. 425, 131632. http://dx.doi.org/10.1016/j.cej.2021.131632, URL https://www.sciencedirect.com/science/article/pii/S1385894721032137.

Stanev, V., Choudhary, K., Kusne, A.G., Paglione, J., Takeuchi, I., 2021. Artificial intelligence for search and discovery of quantum materials. Comm. Mater. 2 (1), 105. http://dx.doi.org/10.1038/s43246-021-00209-z.

Su, X., Yan, X., Tsai, C.-L., 2012. Linear regression. Wiley Interdiscip. Rev. Comput. Stat. 4 (3), 275–294.

Wysotzki, F., 1992. Machine learning and its application to process control. In: Gritzmann, P., Hettich, R., Horst, R., Sachs, E. (Eds.), Operations Research '91. Physica-Verlag HD, Heidelberg, pp. 571–574.

Xu, J., Cao, X.-M., Hu, P., 2021. Perspective on computational reaction prediction using machine learning methods in heterogeneous catalysis. Phys. Chem. Chem. Phys. 23, 11155–11179. http://dx.doi.org/10.1039/D1CP01349A.

Zhou, Z., Li, X., Zare, R.N., 2017. Optimizing chemical reactions with deep reinforcement learning. ACS Central Sci 3 (12), 1337–1344. http://dx.doi.org/10.1021/acscentsci.7b00492, PMID: 29296675.