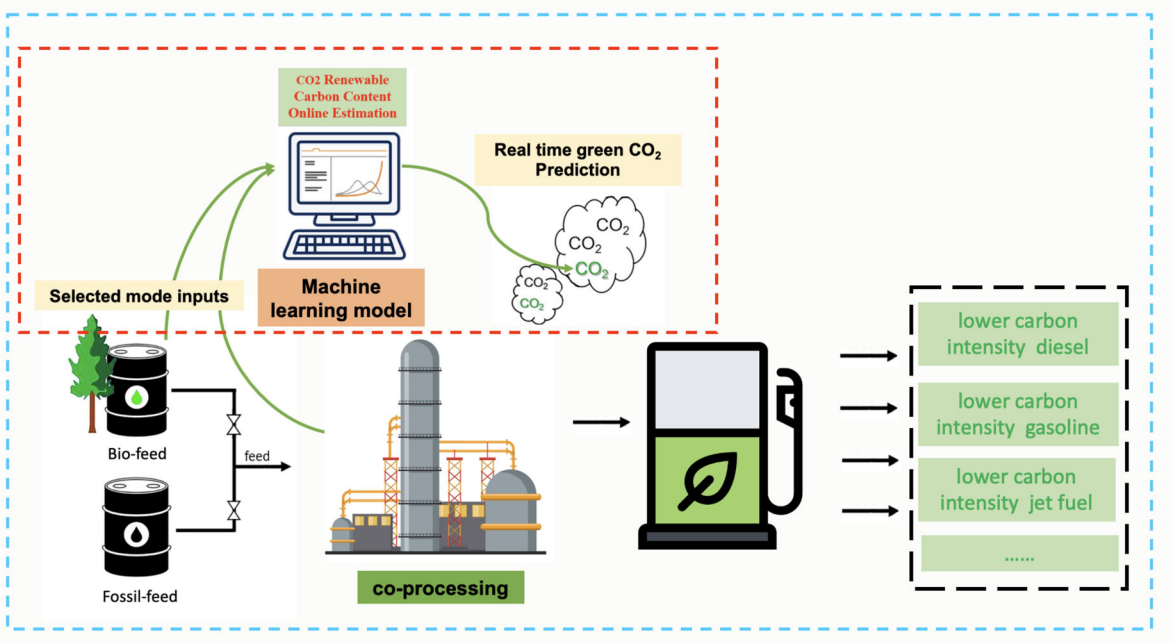
Graphical Abstract

Machine Learning for Real-Time Green Carbon Dioxide Tracking in Refinery Processes

Liang Cao,Jianping Su,Jack Saddler,Yankai Cao,Yixiu Wang,Gary Lee,Lim C. Siang,Yi Luo,Robert Pinchuk,Jin Li,R. Bhushan Gopaluni



Highlights

Machine Learning for Real-Time Green Carbon Dioxide Tracking in Refinery Processes

Liang Cao,Jianping Su,Jack Saddler,Yankai Cao,Yixiu Wang,Gary Lee,Lim C. Siang,Yi Luo,Robert Pinchuk,Jin Li,R. Bhushan Gopaluni

- Artificial intelligence was introduced for real-time tracking of green CO_2 in refineries.
- Model accuracy within 2.66% error was demonstrated using a commercial dataset.
- A cost-effective alternative to lab-based renewable carbon measurements is provided.
- Oil refining sustainability is enhanced through AI-driven emission tracking.

Machine Learning for Real-Time Green Carbon Dioxide Tracking in Refinery Processes

Liang Cao[a,1], Jianping Su[b,*,1], Jack Saddler[c], Yankai Cao[a], Yixiu Wang[a], Gary Lee[d], Lim C. Siang[d], Yi Luo[a], Robert Pinchuk[d], Jin Li[d] and R. Bhushan Gopaluni[a,**]

[a]*Department of Chemical and Biological Engineering, University of British Columbia, Vancouver, BC, V6T 1Z3, Canada*
[b]*China University of Petroleum, Beijing, 102200, China*
[c]*Forest Products Biotechnology/Bioenergy Group, University of British Columbia, Vancouver, British Columbia, V6T 1Z4, Canada*
[d]*Parkland Refining (B.C.) Ltd, Department of Low Carbon Strategy, Burnaby Refinery, Burnaby, BC V5C 1L7, Canada*

ARTICLE INFO

Keywords:
Biogenic feedstocks
Co-processing
Machine learning
Real-time monitoring
Green Carbon Dioxide Tracking

ABSTRACT

The global increase in greenhouse gas emissions presents an urgent environmental challenge, demanding innovative strategies for emission reduction and a fundamental shift in energy consumption practices. Co-processing biogenic feedstocks, such as used cooking oils and biocrudes derived from forest and agricultural residues, within existing oil refineries has been demonstrated as a cost-effective, scalable approach to producing low-carbon fuels, quickly helping the oil refiners to mitigate carbon dioxide emissions, leveraging the existing infrastructures. Despite its potential, monitoring the "green" CO_2 emissions originating from biogenic feedstocks during co-processing remains challenging. The molecular structure of biogenic components becomes indistinguishable from fossil-based molecules, necessitating costly, labor-intensive, and time-consuming sample collection and testing procedures, often involving isotope carbon analysis. This work proposes a new approach by applying artificial intelligence to model green CO_2 emissions in real-time. By analyzing over 102,000 samples of industrial data from a commercial FCC unit, a robust machine learning framework is developed to provide continuous, cost-effective, and accurate green CO_2 monitoring. The methodology encompasses a comparative analysis of ten input analysis techniques and five regression models to model emissions, achieving an average error margin of just 2.66% compared to traditional laboratory measurements. This AI-driven approach offers refiners and policymakers a practical tool for assessing the environmental performance of biogenic feedstock co-processing, facilitating informed decision-making in renewable fuel production.

Nomenclature

AI	Artificial Intelligence
AMS	Accelerator Mass Spectrometry
CatBoost	Categorical Boosting
CO_2	Carbon Dioxide
EWMA	Exponential Weighted Moving Average
FCC	Fluid Catalytic Cracking
GHG	Greenhouse Gas
IoT	Internet of Things

*Corresponding author
**Corresponding author
jianping.su@cup.edu.cn (J. Su); bhushan.gopaluni@ubc.ca (R.B. Gopaluni)
ORCID(s): 0000-0002-2880-3097 (L. Cao)
[1]Equal contribution author

LARS	Least Angle Regression
LASSO	Least Absolute Shrinkage and Selection Operator
LightGBM	Light Gradient Boosting Machine
LSC	Liquid Scintillation Counting
OLS	Ordinary Least Squares
SGDR	Stochastic Gradient Descent Regressor
XGBoost	Extreme Gradient Boosting

1. Introduction

Climate change has become an urgent global issue that requires immediate and strategic interventions. As part of this effort, societies around the world are striving to reduce carbon emissions, a challenge that is particularly complicated for the transportation and industrial sectors [1–3]. Technological constraints in long-distance transport, the maturity of alternative industrial solutions, and the heavy dependence on fossil fuels due to their cost effectiveness have hindered decarbonization in these sectors [4, 5].

Initiating policy interventions is necessary to drive change in these sectors [1, 6, 7]. Market-oriented policies, such as low carbon fuel standards, have been implemented on the West Coast of North America, notably in regions such as California, Oregon, and British Columbia [6, 8]. These policies require fuel suppliers to decrease the carbon intensity of their products. Furthermore, policies in North America and Europe also mandate specific percentages of renewable fuels [9]. Canada has also implemented a national clean fuel standard [10]. Initially, fuel suppliers adhered to the regulations by purchasing and blending bioethanol and biodiesel. However, as policy stringency has ramped up over time, these suppliers are now motivated to produce fuels with lower carbon intensity to maintain profitability.

Fuel suppliers in Canada, including oil refineries, are now under growing pressure to decarbonize their operations and products due to various policies such as carbon taxes, Canada's clean fuel standard, and environmental regulations that make CO_2 emissions costly [10]. However, it should be noted that these policies also guide industry adaptation to future needs. These policies not only require a shift towards greener operations, but also offer a unique opportunity for the fossil fuel industry to lead the transition to sustainable energy sources. By aligning their vast resources and technical expertise with environmental objectives, these entities can significantly accelerate global efforts to mitigate climate change, demonstrating a proactive commitment to a sustainable future. Instead of being viewed as an adversary to decarbonization, the fossil fuel industry, with its experience in energy production and vast infrastructure, should be seen as a partner in decarbonization efforts.

Collaboration with the fossil fuel industry is crucial to accelerate decarbonization. Co-processing refers to the simultaneous processing of biomass with fossil fuels in existing refinery infrastructures, which allows for the integration of renewable feedstocks without the need for substantial changes to existing operational systems. Fig. 1 shows a diagram of co-processing. Biogenic feedstock is low-carbon-intensive feedstocks, such as bio-crudes made from forest, mill residues [11, 12], microalgae, municipal sludge, and municipal waste [13]. However, in the processing of biogenic feedstocks, such as municipal waste, potential pollutants can be generated, requiring careful management and treatment strategies to mitigate environmental impacts [14, 15]. These materials can undergo processes such as hydrothermal liquefaction to produce biocrude, which is then suitable for further processing [16]. Co-processing biogenic feedstocks offers a promising pathway to reduce the carbon intensity of products and mitigate process emissions. The widespread adoption of co-processing feedstocks, such as biocrudes derived from forest/agricultural residues, is expected to occur when a stable and sufficient supply becomes available. Co-processing represents one of the few methods available for traditional oil companies to reduce their greenhouse gas emissions, as burning fossil fuels contributes to 70% of the life-cycle emissions of petroleum products.

The commercialization of co-processing biogenic feedstocks has been successfully implemented in both hydrotreaters and fluid catalytic crackers (FCCs) [17–19]. To enable effective co-processing in

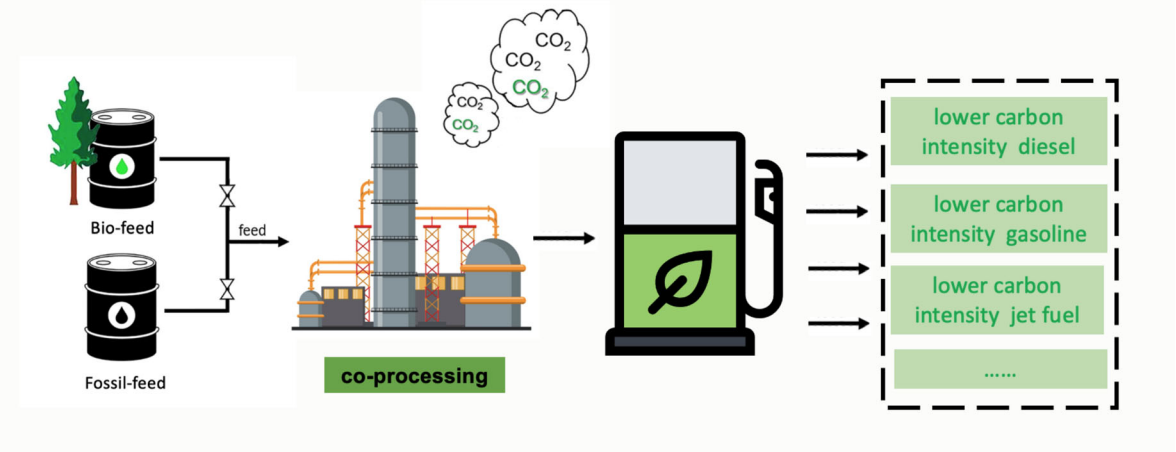


Figure 1. A diagram of co-processing

hydrotreaters, substantial modifications are necessary to enhance their capacity for biogenic feedstock. However, the upgrading of hydrotreaters to co-process biocrudes has proven to be a considerable technical challenge [20].

FCC co-processing is more robust compared to co-processing at the hydrotreater, allowing for a higher co-processing capacity [21–23]. The FCC unit, which is used primarily to convert heavy fractions of crude oil into lighter components, is also the largest source of greenhouse gas (GHG) emissions within oil refineries [24]. This is primarily due to constant regeneration of the catalyst, where the deposited coke is burned, subsequently releasing CO_2 into the atmosphere. The schematic representation of the FCC is shown in Fig. 2.

With the introduction of biogenic feedstocks in the co-processing method, renewable biogenic feedstocks also contribute to CO_2 emissions. Green CO_2 refers to CO_2 emissions originating from renewable biogenic sources, rather than from fossil fuels. In the context of industrial processes, particularly in the co-processing of biogenic feedstocks within oil refineries, green CO_2 generally denotes the emissions resulting from the combustion or processing of renewable materials such as plant-based biomass or bio-wastes. Unlike CO_2 produced by fossil fuels, which increases the net amount of carbon in the atmosphere, green CO_2 is regarded as part of the carbon cycle, since the carbon emitted is balanced by the carbon absorbed during the growth of biomass. Therefore, while still contributing to total CO_2 emissions, green CO_2 is considered less impactful on the overall carbon footprint and the potential for global warming. Green CO_2 emissions from this process must be rigorously measured for multiple reasons. First, the refinery needs to pay carbon taxes based on the amount of CO_2 emissions. Therefore, an accurate assessment of green CO_2 is crucial to determine the financial obligations of the refinery under environmental taxation regulations. Second, it helps in the formulation of effective policies, providing the government with accurate data to guide efforts to reduce the overall carbon footprint of the industrial sector.

In this work, artificial intelligence (AI) is adopted for the first time in the field of biogenic feedstocks co-processing for green CO_2 modeling, offering a promising avenue to address these challenges. The industrial partner, Parkland Refining Ltd., is actively involved in coprocessing oleochemical / lipid feedstocks such as tallow, canola oil and tall oil, thus reducing the carbon intensity (CI) of the various fuels manufactured. The partner has provided a wealth of valuable data from multiple operational scenarios, laying the crucial foundation for the development of AI models in this study. Using this extensive dataset, along with internet of things (IoT) technology and sensor networks, this study offers large-scale real-time monitoring of green CO_2 emissions in an industrial setting. To ensure the validity and quality of the results, this study cross-validated them with quarterly experimental results sampled from Parkland Refining Ltd.

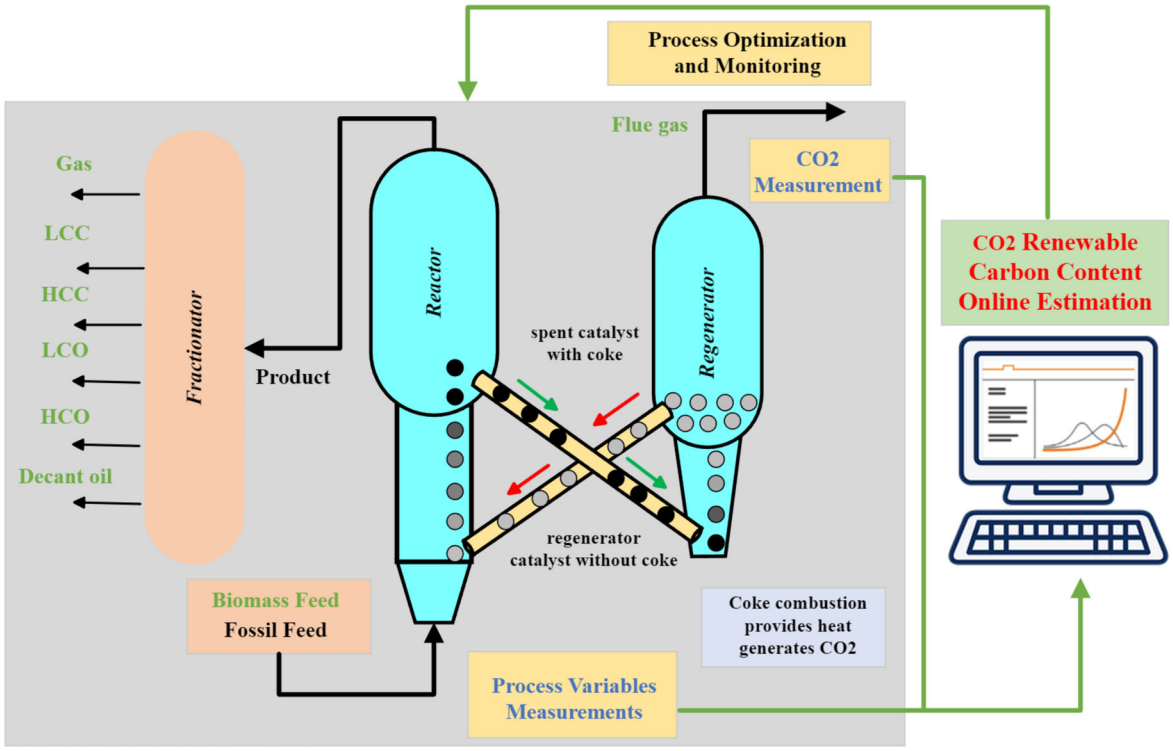


Figure 2. A flow diagram of a Fluid Catalytic Cracking (FCC) unit with online renewable content monitor.LCC: Light Catalytic Cracking gasoline; HCC: Heavy Catalytic Cracking gasoline; LCO: Light Cycle Oil; HCO: Heavy Cycle Oil; CO_2: Carbon Dioxide.

In this study, a novel approach is introduced to monitor green CO_2 in the refining industry by integrating artificial intelligence with the co-processing of biogenic feedstocks. This novel contribution is underscored by several key advances:

1. **First to apply AI in co-processing:** To the best of the author's understanding, this study is the first to leverage AI for real-time green CO_2 tracking in this context, offering a scalable and efficient alternative to conventional methods such as AMS ^{14}C.
2. **Utilization of a large commercial dataset:** This research is uniquely supported by a large dataset of 102,000 samples from Parkland Refining Ltd., significantly enhancing the precision and reliability of findings.
3. **Real-time monitoring and cost efficiency:** This data-driven approach not only facilitates real-time green CO_2 monitoring but also represents a cost-effective alternative to ^{14}C measurements, making it a practical solution for the industry.
4. **Sustainability impact:** This method empowers refineries to accurately quantify and reduce their green CO_2 emissions, contributing to more accurate carbon accounting practices.

2. Review of the Green CO_2 Tracking in Co-processing

Climate change and global warming require industries, particularly those with significant carbon outputs such as oil refineries, transition towards more environmentally sustainable practices. This shift is compelled not merely by the need to protect the environment, but also by strict policies and regulations that demand the reduction of greenhouse gas emissions [1, 6, 7]. In this context, accurate quantification of green CO_2 emissions becomes crucial, especially in processes such as co-processing, where biogenic feedstocks are used along with fossil feedstocks in refineries [4, 5, 8].

Distinguishing between biogenic and fossil-derived CO_2 is crucial to evaluate the environmental impact of fuels produced through co-processing. Tracking green CO_2 involves measuring the amount of renewable content in co-processed fuels. The goal is to create affordable, efficient and fast methods that not only stimulate scientific development but also aid industries in meeting environmental regulations and sustainability goals [17, 18]. Moreover, these improvements are crucial to giving policymakers and stakeholders the right tools for informed decision making and effective environmental management [9, 10, 19, 20]. These methodologies are broadly classified into two categories: direct and indirect measurement ([25]). This review focuses on these classifications, exploring the merits and limitations inherent in the different methodologies within each category.

Direct measurement means collecting samples and subsequently analyzing the renewable content within co-processed fuels. This is the most straightforward way to know the renewable content, similar to how oil refineries measure the other properties of fuels such as density, energy content, elemental composition, etc. Accelerator mass spectrometry (AMS) ^{14} C is considered the benchmark to distinguish between biogenic and fossil-derived CO_2, attributed to its exceptional precision and reliability [23, 26]. However, despite its analytical precision, the AMS ^{14}C technique faces difficulties that include high costs (with testing alone priced around $300-$500 for each sample, which does not include sample collection and other associated expenses), long processing times, and operational complexity, making frequent tests impractical for many facilities [23, 26].

Alternatives to AMS ^{14}C, including liquid scintillation counting (LSC) ^{14}C and Carbon-13 (^{13}C) analysis, have been explored to address these drawbacks. Although these methods provide additional means to measure biogenic CO_2, they share some limitations with AMS ^{14}C, such as the need for stable conditions to obtain accurate readings, the difficulty in detecting low biogenic CO_2 concentrations, and potential risks of sample contamination, which may lead to inaccurate results [27, 28].

In response to these limitations, both the industrial and academic sectors are advancing measurement technologies to improve accuracy, lower costs, and streamline operational procedures for easier adoption in industrial settings. Indirect measurement methods represent an important complement to direct methods to track green CO_2 in co-processing. They offer the potential for continuous, real-time monitoring and can reduce the need for expensive and time-consuming laboratory analyzes. Unlike direct measurements that evaluate the biogenic carbon content through physical or chemical analysis of fuel samples, indirect methods estimate renewable content based on operational data, mathematical models, or surrogate measurements. This approach can be beneficial when direct measurements are impractical or when processes involve complex mixtures of feedstocks and products, as is common in co-processing scenarios [29, 30].

One prevalent method within this category is mass balance calculation, where the input and output flows of carbon in the refinery processes are analyzed to estimate the renewable content of produced fuels. However, this method requires accurate and comprehensive data on all materials entering and leaving the system, which can be challenging to obtain in real-world operations. Additionally, assumptions made during the calculations, such as constant process efficiency or identical behavior of biogenic and fossil carbon in the process, can significantly affect the accuracy of the results [25].

Another indirect measurement approach is the use of process simulation and modeling techniques. These methods involve developing mechanistic models [31] or soft sensor models [32, 33], which is very popular and effective in many fields. In this work,to describe how different types of feedstock behave within the co-processing environment [34, 35], data from process monitoring systems—such as flow rates, temperatures, and pressures—are integrated to estimate the renewable content of fuels without the need for extensive laboratory testing.

Despite their potential, indirect methods are not without challenges. The quality and availability of operational data, the complexity of refinery processes, and the dynamic nature of production can all complicate the application and reliability of these methods. In addition, the regulatory acceptance of indirect measurement results can vary as standards and verification procedures are still evolving. As such, although indirect methods have the promise of simplifying the tracking of green CO_2, they must be carefully developed, validated and implemented to ensure their precision and reliability [30]. As the field evolves, the integration of direct and indirect measurement approaches could provide a more comprehensive and practical solution for assessing the renewable content of co-processed fuels.

In conclusion, as the industry moves forward, the development of new, efficient, and cost-effective methods for tracking green CO_2 will be paramount. The advancements in this field will aid not only in fulfilling regulatory requirements but also in actualizing the potential environmental benefits of using biogenic feedstocks in co-processing. Collaborative efforts of researchers, industry professionals, and policy makers will be essential to overcome existing challenges and achieve significant strides in reducing carbon emissions from refinery operations [23, 36].

3. Methods

3.1. Online Green CO_2 Monitoring

Continuous processing of biogenic feedstocks by refineries generates an abundance of data. In particular, refineries already have the ability to measure total CO_2 emissions in real time, but the amount of green CO_2 can only be determined by sampling through AMS^{14}C. This situation creates an opportunity: this work proposes the integration of machine learning methodologies to first establish a reliable and accurate real-time model to predict total CO_2 emissions. Once this model is established, this study can proceed with a detailed analysis to approximate the green CO_2 emissions. Total CO_2 emissions are considered to be a linear combination of input variables, as follows:

$$\text{Total } CO_2 = \underbrace{\boldsymbol{a} \cdot fossil\ feed + \boldsymbol{b} \cdot bio\ feed}_{CO_2(fossil,\ bio)} + \varepsilon(fossil,\ bio) \tag{1}$$

The predicted CO_2 is further decomposed into two components. $CO_2(fossil, bio)$ represents the main contribution of fossil feed and bio feed to the CO_2 production, while $\varepsilon(fossil, bio)$ accounts for the additional contribution or adjustment to the CO_2 production that is not directly explained by the fossil feed and bio feed. For instance, changes in reaction conditions, such as variations in temperature and pressure, can affect the rate of reaction and the products, ultimately affecting the production of CO_2. In practical operations, the precision and reliability of the model are enhanced by optimizing the ε term, ensuring that the predicted values are as close as possible to the actual emissions. $\varepsilon(fossil, bio)$ is defined as follows:

$$\varepsilon(fossil,\ bio) = \boldsymbol{c} \cdot feature\ I + \boldsymbol{d} \cdot feature\ II + \boldsymbol{e} \cdot feature\ III + \cdots \tag{2}$$

Here, *fossil feed, bio feed, feature I, feature II, feature III,...* represent the selected variables in the co-processing, while ***a,b,c,d,e,...*** stand for the coefficients in the linear regression model that are determined through data analysis.

Defining the co-processing ratio, r_{co}, as the proportion of bio feed to the combined amount of bio feed and fossil feed (bio feed / (bio feed + fossil feed)), it is noteworthy that green CO_2 is not only related to the bio feed, but also the corresponding $\varepsilon(fossil,\ bio)$ of the co-processing ratio. Consequently, the formula for green CO_2 can be further derived as shown below:

$$Green\ CO_2 = \boldsymbol{b} \cdot bio\ feed\ +\ r_{co} \cdot \varepsilon(fossil,\ bio) \tag{3}$$

Using machine learning methodologies to analyze and predict total CO_2 emissions, this work is able to further estimate green CO_2 emissions. The proposed approach represents a notable advancement in the monitoring accuracy of green CO_2. This innovative solution effectively tackles a long-standing challenge in the refining industry, allowing real-time monitoring of the biogenic fraction in various products and outputs.

3.2. Data Preprocessing and Feature Selection of Green CO_2 in FCC Co-processing

The data for this study were collected from a commercial refinery between May 2020 and March 2023, resulting in over 102,000 samples recorded every 10 minutes. Outliers were removed using the three-sigma rule to ensure data quality. After cleaning, the dataset was split into training (80%) and testing (20%) sets.

The FCC co-processing has a substantial array of variables, utilizing all these variables without a careful selection strategy inevitably risks model overfitting. Such overfitting undermines the model's versatility and impairs its ability to accurately predict unseen data, thereby diminishing the effectiveness of the model [37]. Therefore, the identification of key features for the modeling phase is a crucial step. The initial selection of characteristics is performed with expert knowledge, as shown in Fig. 3. The selection of these features is driven by their essential role in the overall process.

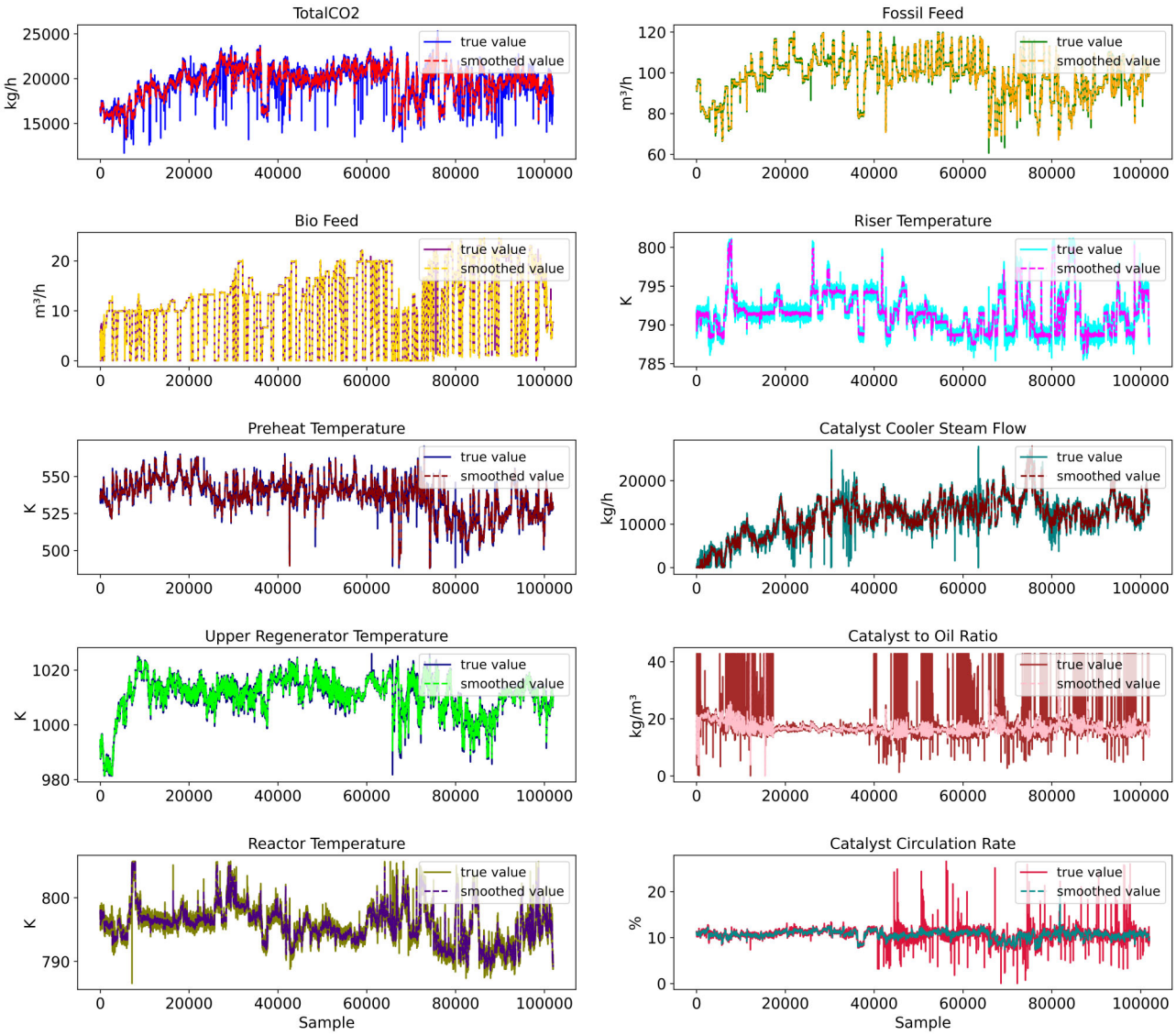


Figure 3: Initial feature selection and data smoothing

Following initial selection, data smoothing is used to reduce noise and highlight underlying trends. This study uses the exponential weighted moving average (EWMA) method [38] for smoothing. This method has proven effective in reducing noise, thereby providing a cleaner dataset, and plays an instrumental role in enhancing the robustness of feature selection. To further assess the importance of these features, this work uses a variety of machine learning algorithms. The principle of selection criteria is consistency, meaning that a feature had to be identified as important across a majority of the models to be considered.

Given the large number of features in the dataset, this work employs ten different feature selection methods to analyze the importance of each feature, selecting those consistently significant across all methods as model input. These methods are carefully chosen for their unique strengths in feature selection and encompass all current feature selection methodologies.

Mutual information based feature selection is used for its ability to capture any kind of statistical dependency, especially useful for nonlinear relationships [39]. The least absolute shrinkage and selection operator (LASSO) [40] is used for its ability to perform both variable selection and regularization to improve the precision of the prediction and the interpretability of the statistical model. Elastic net models were used for their ability to handle multicollinearity, offering stable estimates even when variables are highly correlated [41].

Boruta feature selection [42], a wrapper built around the random forest classification algorithm, is used for its ability to efficiently handle high-dimensional data, ensuring robustness against overfitting. Other tree-based models such as the random forest regressor, the gradient boosting regressor, the extreme gradient boosting (XGBoost), the light gradient boosting machine (LightGBM) and the categorical boosting (CatBoost) [43] were utilized for their strengths in handling the skewed error distribution and minimizing the influence of outliers.

Each of these models is trained on the same dataset, ensuring a fair comparison of the importance of the features. To facilitate a direct comparison and highlight universally significant features, all feature importance was normalized using the min-max scaler. Ultimately, this work will select the most important features for further modeling. This comprehensive approach, considering the potential biases and uncertainties of single model applications, establishes a robust framework for feature importance analysis in biomass co-processing.

Fig. 4 shows the feature importance as identified by different algorithms. Among the top-ranked features, variables such as the "fossil feed", the "catalyst circulation rate," and the "bio feed," which are significant contributors to the coke generation in the refining process. Interestingly, 'catalyst cooler steam flow' and 'upper regenerator temperature' also emerged as crucial features. These features are consequential to the coke combustion process, reflecting the amount of coke that has been burned.

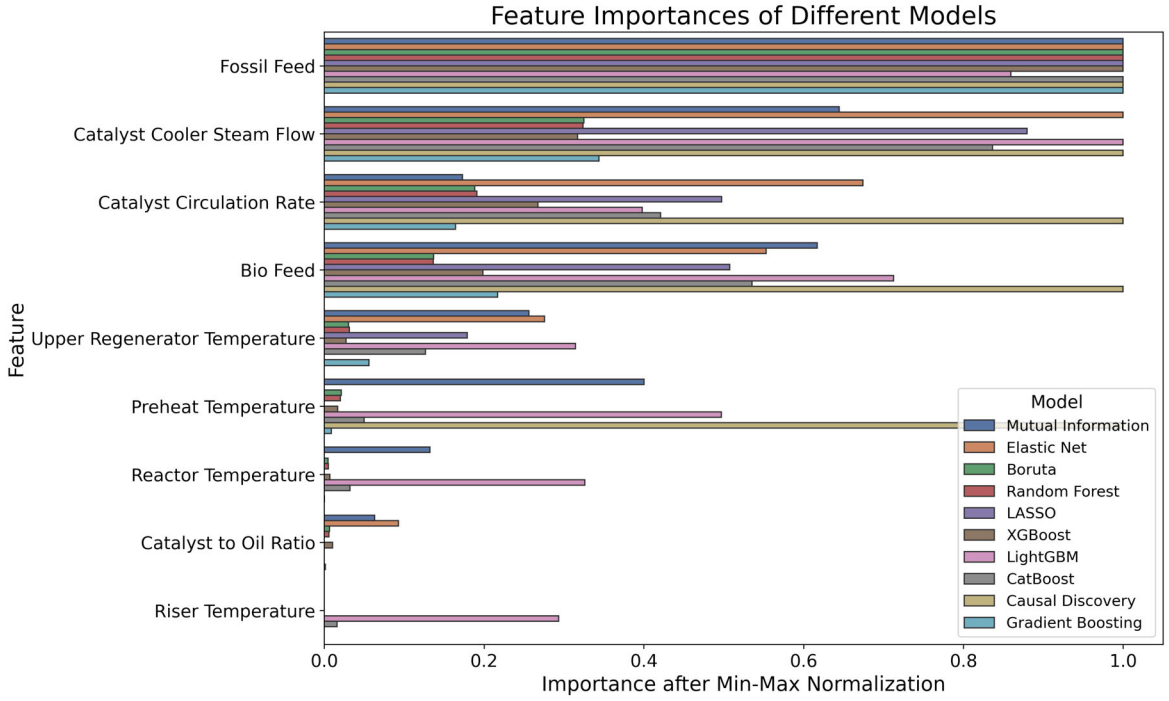


Figure 4. Feature selection using feature importance. LASSO: Least Absolute Shrinkage and Selection Operator; XGBoost: Extreme Gradient Boosting; LightGBM: Light Gradient Boosting Machine; CatBoost: Categorical Boosting.

In previous studies, 'preheat temperature' was recognized as a critical feature, guided by its identification through causal discovery methods [44]. However, in current research, a consistency-based principle for selection criteria is adhered to, meaning that a feature is considered significant if and only if it is deemed important by a majority of algorithms used. As a consequence, although

the "preheat temperature" had potential significance in the causal discovery method, it is not included because it did not meet the selection criteria in this work. The 'riser temperature', theoretically related to coke generation, is not chosen because of the existence of intricate feedback loops in the system. Ultimately, this work selected fossil feed, catalyst cooler steam flow, catalyst circulation rate, bio feed, and upper regenerator temperature for green CO_2 modeling.

3.3. Modeling of Green CO_2 in FCC Co-processing

When developing an online system for monitoring green CO_2 emissions, careful attention was given to designing a model that is easy to interpret, ensuring it can be readily understood by both engineers and government officials. In order to achieve a reliable, interpretable and simple online monitoring system for green CO_2, a variety of linear regression models are employed in this study. The ordinary least squares (OLS) model is used for its ability to minimize the sum of squared residuals and aim for the least mean squared error. For high-dimensional data analysis, the least angle regression model (LARS) [45] is used, praised for its robustness against overfitting. This work also incorporated the Bayesian Ridge model, known for its proficiency in managing multicollinearity and offering stable estimates through a probabilistic model.

The stochastic gradient descent regressor (SGDR) [46] is also adopted. Stochastic gradient descent is a popular optimization method that is used in various machine learning algorithms. SGDR efficiently fits linear models on large data sets, making it ideal for managing a wide range of features and samples. For dealing with skewed error distributions, the Theil-Sen model [47] is used, while the Huber regression [48] is used to minimize the influence of outliers.

Fig. 5 illustrates the steps of the proposed algorithm for the online monitoring of green CO_2 emissions. The process begins with co-processing data after an initial selection of features such as fossil feed, bio feed, preheat temperature, and catalyst circulation rate. These features undergo data smoothing to reduce noise and emphasize trends. Subsequently, the final inputs are selected, including fossil feed, catalyst cooler steam flow, bio feed, and upper regenerator temperature, to establish the machine learning model for predicting total CO_2 and green CO_2 emissions.

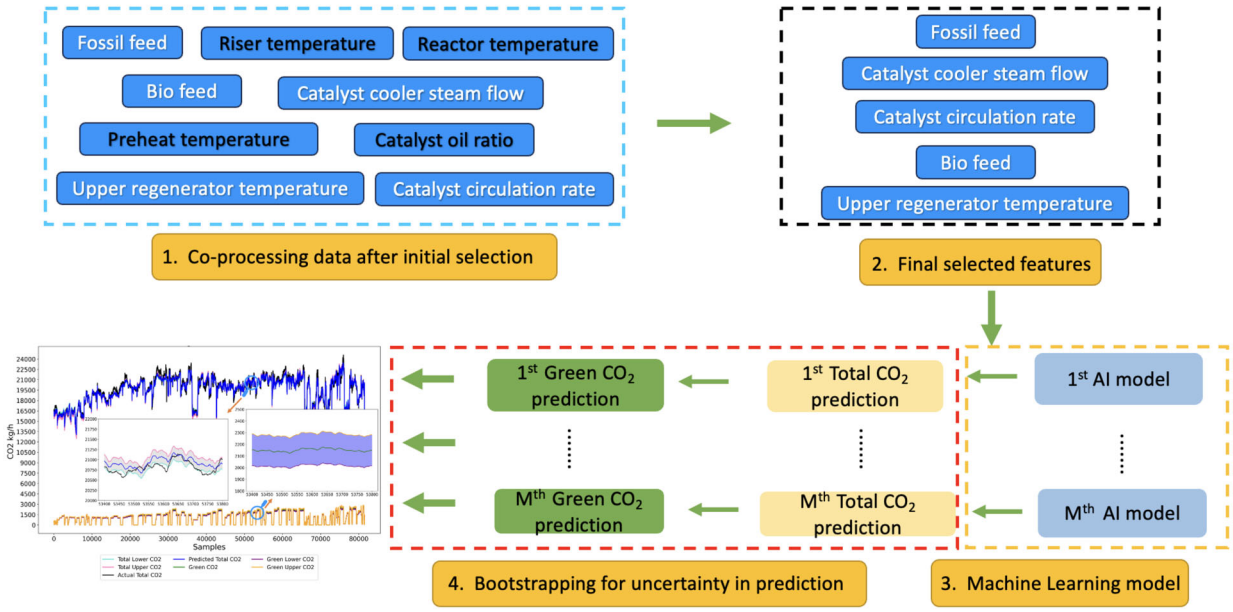


Figure 5. Schematic diagram of the proposed green CO_2 online monitoring algorithm

Specifically, this work performed nonreplacement sampling on the original data, with each subset constituting 30% of the original dataset. This process is repeated 10,000 times, resulting in 10,000 distinct subsets. To ensure robustness and mitigate potential bias, a comparison of multiple regression models is carried out, accompanied by an in-depth further analysis using the bootstrapping method.

Subsequent modeling analysis is performed on these 10,000 subsets to derive a distribution for bio feed coefficients $\boldsymbol{b}$, fossil feed coefficients $\boldsymbol{a}$, and their corresponding ratio, accompanied by the 95% and 99% confidence intervals. By using these sub-samples to create several models, this work can improve the robustness of models and get a better sense of the prediction uncertainty.

4. Results

4.1. Green CO_2 Online Monitoring

As represented in Fig. 6, the ratio of biofeed coefficient to fossil feed coefficient $\boldsymbol{b/a}$ is approximately 0.62. This ratio is crucial because it clarifies the relative impact of biofeed and fossil feed on total CO_2 emissions under identical conditions in the co-processing of FCC units. Specifically, when biofeed and fossil feed are processed at the same flow rates, the CO_2 emissions from biofeed amount to only 62% of those produced by fossil feed. It is worth mentioning that the analysis, leveraging big data and machine learning in a commercial-scale refinery, supports this conclusion with broader applicability and practicality compared to traditional laboratory-scale studies. The only deviation is observed with the Theil-Sen method, which produced a slightly lower coefficient, below 0.6.

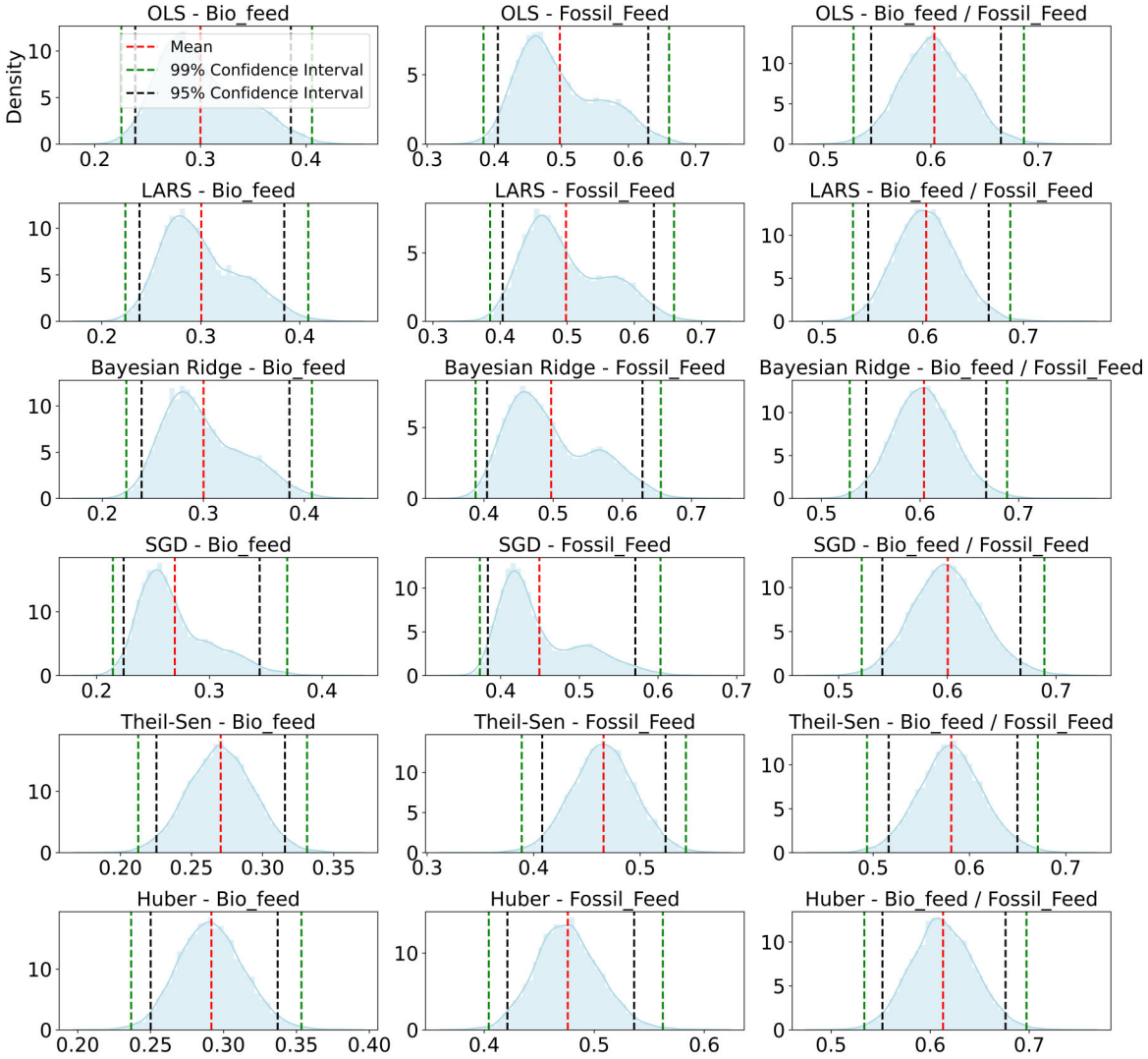


Figure 6. Bio feed coefficient and fossil feed coefficient under different methods. OLS: Ordinary Least Squares; LARS: Least Angle Regression; SGD: Stochastic Gradient Descent.

Following the comparison of multiple regression models and the bootstrapping analysis, this work used the coefficients derived to establish online green CO_2 monitoring. As an illustration, Fig. 7 and

Fig. 8 presents the results of the Huber regression. This monitor not only offers continuous tracking of total CO_2 emissions with a 95% confidence interval, but also provides refiners with dynamic, real-time tracking of green CO_2 emissions with the same confidence level.

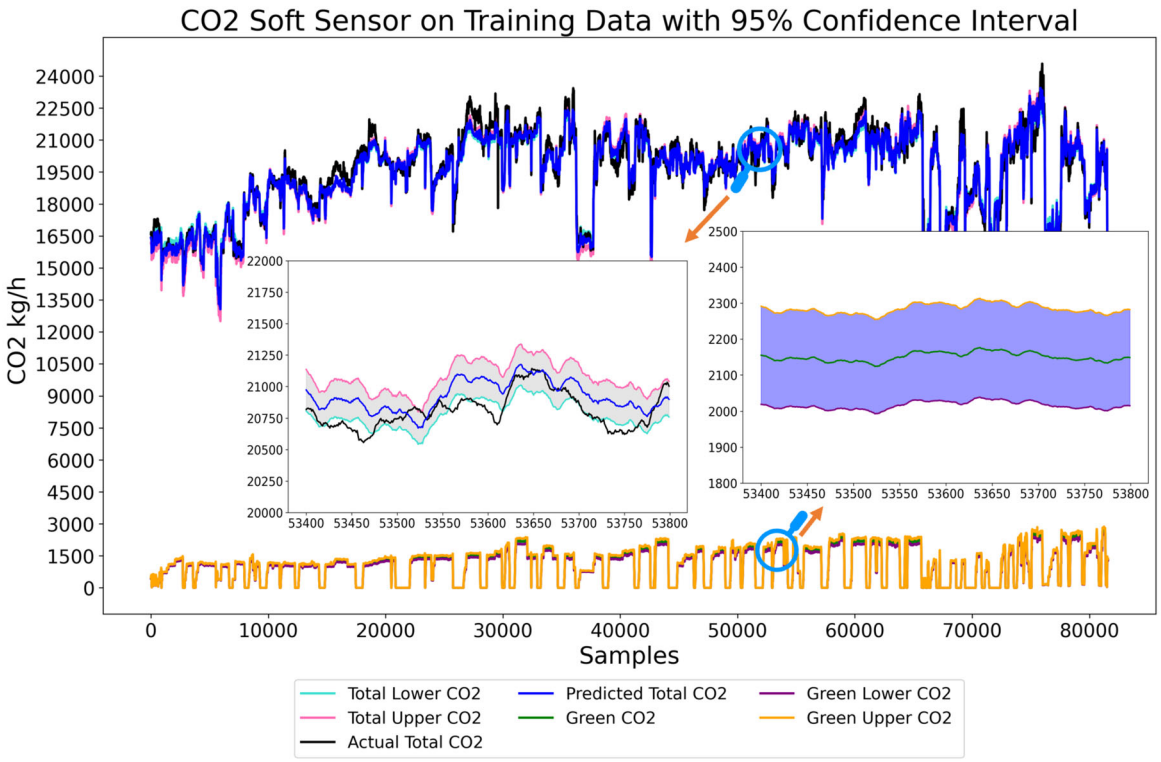


Figure 7. Total CO_2 emissions and green CO_2 emissions on training data for the Huber regression

Take the 53,500th training sample as an example, where the actual CO_2 emission value is 20,750 kg/h. This work predicted that the CO_2 emission would be 20,780 kg/h, with a lower limit of 20,600 kg/h and an upper limit of 20,900 kg/h. This work also predicted that the green CO_2 emissions will be 2,130 kg / hour, with a lower limit of 2,000 kg/h and an upper limit of 2,280 kg/h. For the 3,300th test sample, where the actual CO_2 emission value is 18,790 kg/h, the model predicted the CO_2 emission to be 18,780 kg/h, with a lower limit of 18,660 kg/h and an upper limit of 18,860 kg/h. It also predicted green CO_2 emissions to be 2,420 kg/h, with a lower limit of 2,295 kg/h and an upper limit of 2,540 kg/h.

4.2. Verification by AMS ^{14}C

From the third quarter of 2021 to the first quarter of 2023, Parkland refinery collected 13 samples and sent them for third-party testing. Each sample, with a significant investment in terms of human resources and finances, is of extreme value. The ^{14}C results from these tests served as a means to verify the accuracy of the proposed green CO_2 online monitor. However, the reliability of ^{14}C results hinges on several conditions, such as stable operating conditions. Some samples might have been collected under circumstances where these conditions were not met, making their results less reliable. This work used real-time monitors to validate and analyze these sampling results, scrutinizing their reliability. This not only affirmed the accuracy of proposed algorithm in providing predictive results, but also demonstrated its practicality as a third-party tool for validating the effectiveness of sampling methods.

Using Huber regression as an example, this work generated predictions for the proportion of biogenic carbon content, calculated as the ratio of green CO_2 to total CO_2. Fig. 9 presents the predicted ^{14}C results and their 95% and 99% confidence intervals under different co-processing ratios. The 13 samples collected by the Parkland refinery are indicated by the × mark on the graph. The results

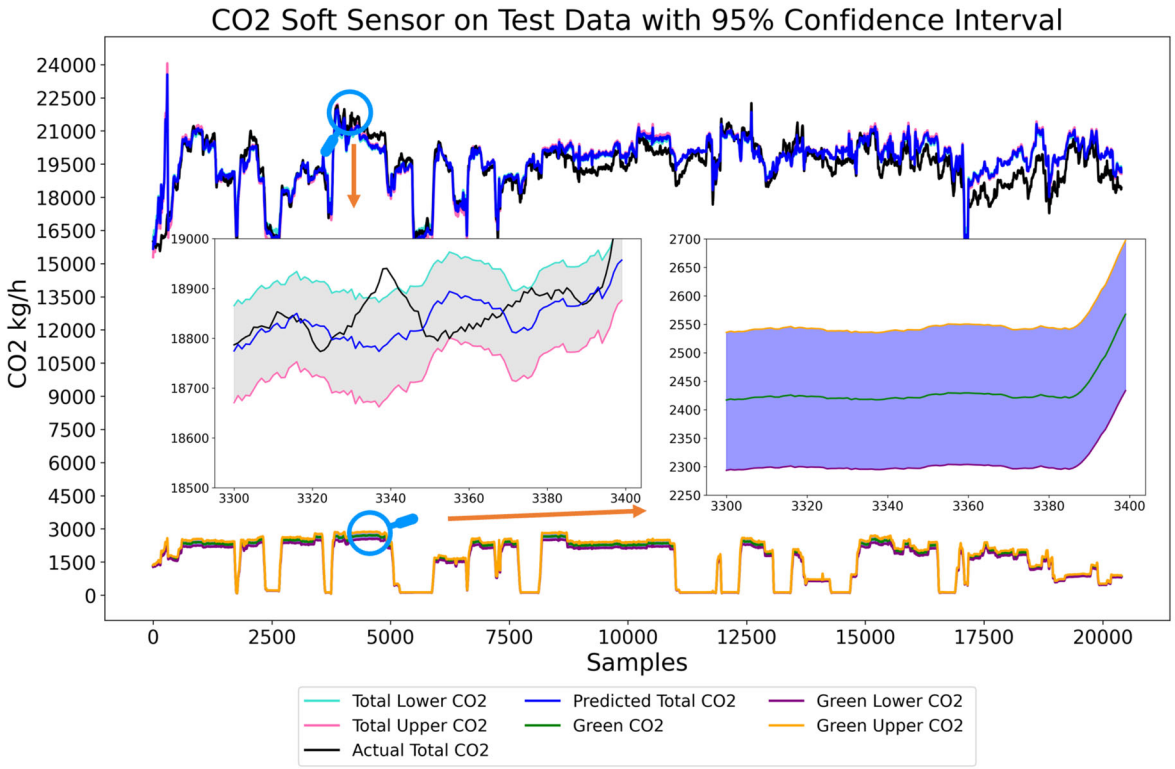


Figure 8. Total CO_2 emissions and green CO_2 emissions on testing data for the Huber regression

revealed that six samples were within the 95% confidence interval, eight samples were within the 99% confidence interval, while five samples were beyond the 99% confidence interval.

When examining these five samples that fell outside the 99% confidence interval, this work found that operational factors during the sampling phase were influencing the measurements. Four of these samples, represented on the left in Fig. 9, were collected during periods of flux, where the feedstock is transitioning between biomass, fossil fuels, or a mixture of both. These instances of transition could create an unstable environment within the processing unit, which likely impacted the accuracy of these measurements.

Additionally, the sample represented in the upper right corner of the Fig. 9 demonstrates an unusual spike. This sample is collected four minutes before a significant, although temporary, increase in feed rate is observed. This temporary fluctuation may have introduced further inconsistencies in the sample, indicating that subtle changes in operating conditions could significantly affect the accuracy of the measurements. This work hypothesizes that such sudden changes in feed rate may have influenced the concentration of biogenic carbon, thus accounting for the observed discrepancy.

After investigating the six samples that fall within the 95% confidence interval presented in Table 1, additional examination of the comparison between the predictions of the AI model and laboratory measurements ^{14}C revealed low error margins, ranging from 0. 89% to 4. 28%, with a mean error of 2.66%. These findings verify the precision of the AI model in predicting green CO_2 content.

To ensure a robust analysis, the study is not limited to a single model. Instead, four additional methods for monitoring green CO_2 were explored, as demonstrated in Fig. 10 and Fig. 11. The results were consistent across all the methods, strengthening the previous observations. Regardless of the algorithm used, the same five ^{14}C ample results deviated from the predictions, while the rest fell within the predicted confidence interval of the proposed model, indicating that they are likely to be representative and accurate.

From a policymaker's perspective, it might be reasonable to require comprehensive data from refineries. However, such a strategy may pose some challenges. For example, collecting various data (including tasks such as 24-hour sampling) can be technically challenging due to complexities such as

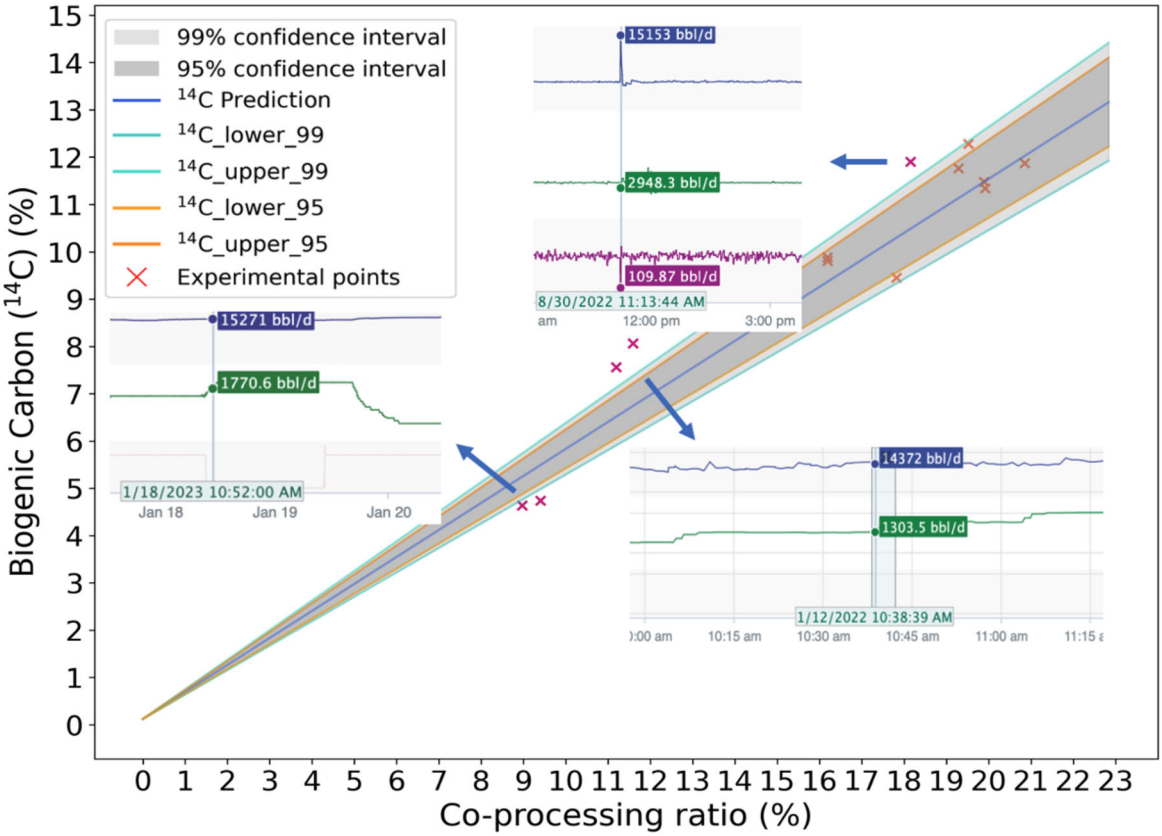


Figure 9. Validation of Accelerator Mass Spectrometry ^{14}C results with modeled 95%,99% confidence interval using Huber regression

Table 1
Experimental and AI model results of ^{14}C (green CO_2) within 95% confidence interval during co-processing

	Co-processing Ratio	Experiment	Huber regression model	Error
Sample 1	16.21%	10.51%	10.12%	3.71%
Sample 2	16.22%	10.59%	10.24%	3.31%
Sample 3	19.31%	12.63%	12.09%	4.28%
Sample 4	19.92%	12.32%	12.43%	0.89%
Sample 5	19.95%	12.17%	12.41%	1.97%
Sample 6	20.89%	12.75%	12.98%	1.81%

blockages, and can result in substantial costs, as these tasks often require the involvement of third-party entities. Furthermore, because of the lack of detailed instructions or guidance on the sampling methods that refineries should follow, the large amount of data requested may be classified as unnecessary or not representative. This could create an administrative burden on refiners, government agencies, and auditors in the future. Based on an extensive dataset and using different machine learning algorithms, the study demonstrates a reliable method for determining renewable content, which may substantially mitigate these operational and administrative difficulties.

5. Discussion

This study demonstrates how AI can bridge the gap between complex operational data and real-time green CO_2 estimation. Despite the contributions of this study, there are limitations that should be acknowledged. One such limitation involves the representativeness of the dataset. Although this study uses a large dataset from Parkland Refining Ltd., it represents data from a single refinery. Therefore,

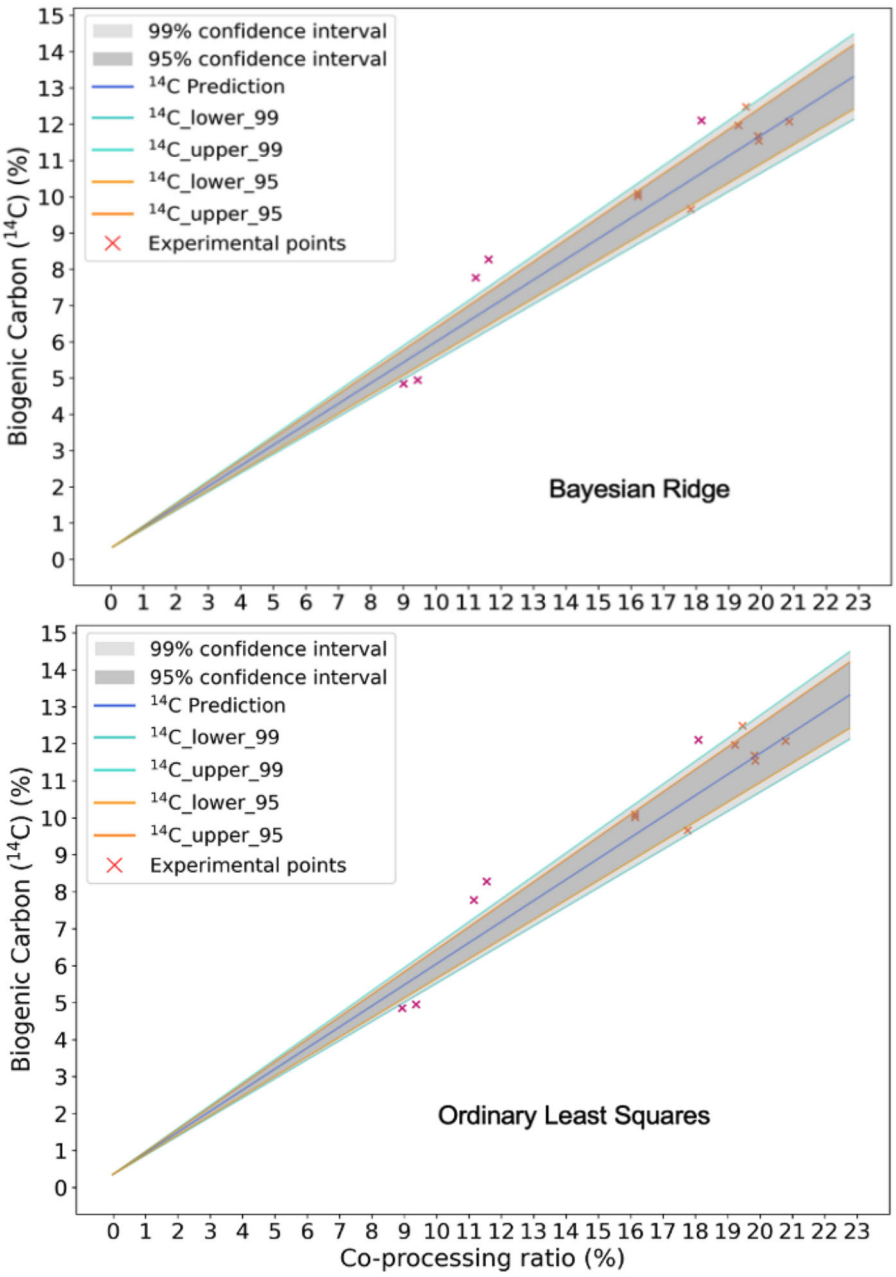


Figure 10. Validation of Accelerator Mass Spectrometry ^{14}C results with modeled 95%, 99% confidence interval using Bayesian ridge and ordinary least squares

the generalizability of the findings to other refineries may be limited. Additionally, while machine learning models offer powerful tools for prediction and analysis, they inherently depend on the quality and range of the data provided. Any bias or anomaly in the dataset could influence the predictions and interpretations of the models. Moreover, transitioning from traditional measurement methods to AI-based approaches for green CO_2 monitoring introduces challenges, particularly in gaining regulatory acceptance and ensuring seamless integration into existing frameworks.

Given these limitations, future research should explore integrating additional data sources, refining model architectures, and conducting multi-site validations. Further, aligning these AI-based estimates with evolving regulatory frameworks and stakeholder needs can ensure that this approach contributes meaningfully to the global decarbonization agenda.

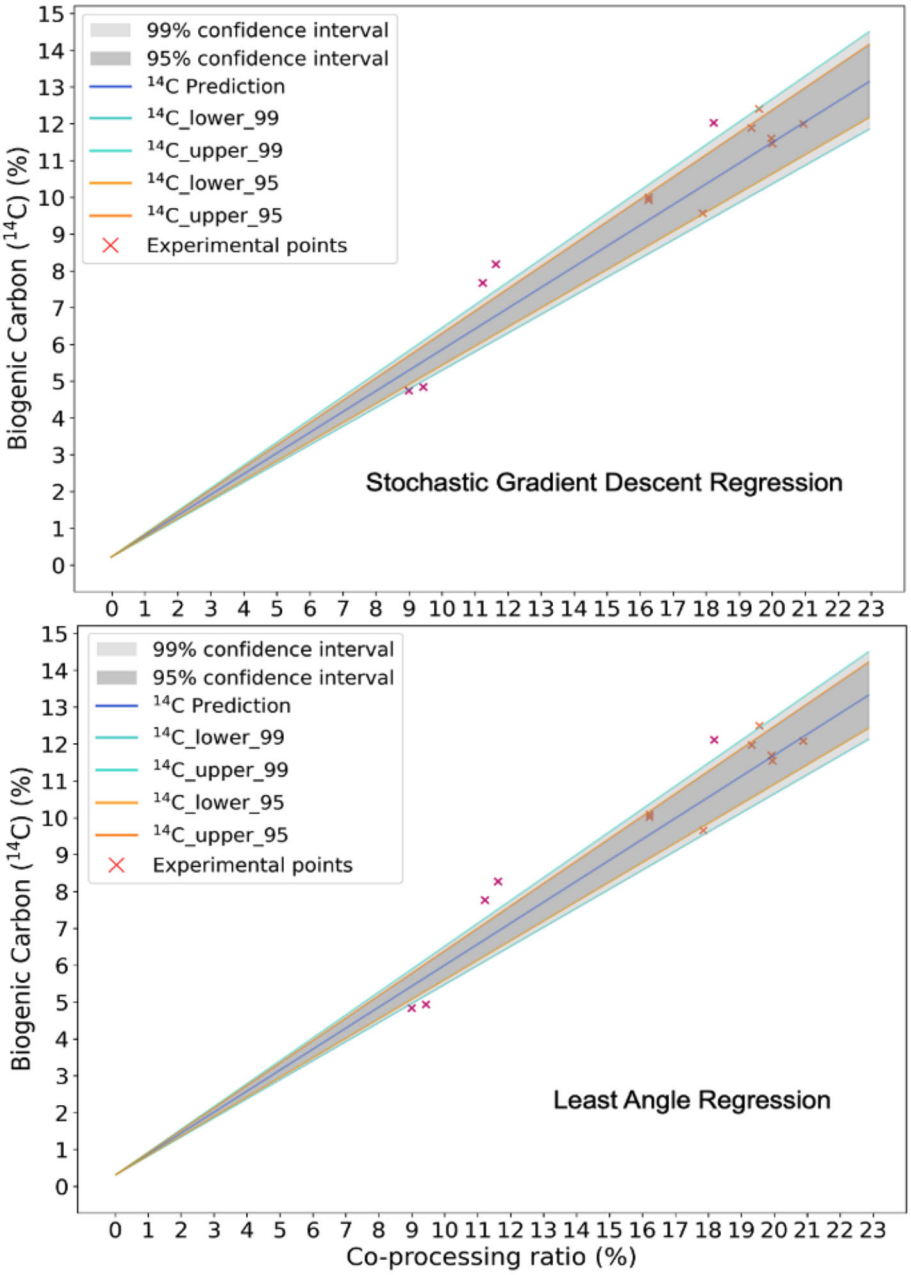


Figure 11. Validation of Accelerator Mass Spectrometry ^{14}C results with modeled 95%, 99% confidence interval using stochastic gradient descent regression and least angle regression

6. Conclusion

In summary, this study introduces an innovative framework for accurately modeling green CO_2 emissions during the co-processing of biogenic feedstocks. Leveraging the power of big data and artificial intelligence, models have been developed to track green CO_2 emissions. This method represents a significant advance over traditional direct measurement techniques, allowing continuous real-time monitoring with increased efficiency and accuracy, demonstrated by an average error margin of just 2.66% compared to conventional laboratory measurements. This precision underscores the efficacy of the machine learning model and marks a substantial improvement in the field. The effectiveness of the approach has been confirmed through practical tests using ^{14}C measurements. This research not only propels the use of artificial intelligence within the oil refining sector, but also has the potential to guide the industry towards more sustainable practices.

References

[1] International Energy Agency. World Energy Outlook 2022, https://www.iea.org/reports/world-energy-outlook-2022; 2022 [accessed 4 March 2024].

[2] International Energy Agency. Net Zero by 2050. https://www.iea.org/reports/net-zero-by-2050; 2021 [accessed 4 December 2024].

[3] International Energy Agency. Renewables 2022. https://www.iea.org/reports/renewables-2022; 2022 [accessed 4 December 2024].

[4] International Energy Agency. Transport. https://www.iea.org/reports/transport; 2022 [accessed 4 December 2024].

[5] International Energy Agency. The challenge of reaching zero emissions in heavy industry. https://www.iea.org/articles/the-challenge-of-reaching-zero-emissions-in-heavy-industry; 2020 [accessed 4 December 2024].

[6] Ebadian M, van Dyk S, McMillan JD, Saddler J. Biofuels policies that have encouraged their production and use: An international perspective. Energy Policy 2020;147:111906. https://doi.org/10.1016/j.enpol.2020.111906.

[7] Aquila G, de Oliveira Pamplona E, de Queiroz AR, Junior PR, Fonseca MN. An overview of incentive policies for the expansion of renewable energy generation in electricity power systems and the Brazilian experience. Renew Sustain Energy Rev 2017;70:1090-8. https://doi.org/10.1016/j.rser.2016.12.013.

[8] Yeh S, Witcover J, Lade GE, Sperling D. A review of low carbon fuel policies: Principles, program status and future directions. Energy Policy 2016;97:220-34. https://doi.org/10.1016/j.enpol.2016.07.029.

[9] Axsen J, Wolinetz M. What does a low-carbon fuel standard contribute to a policy mix? An interdisciplinary review of evidence and research gaps. Transport Policy 2023;133:54-63. https://doi.org/10.1016/j.tranpol.2023.01.008.

[10] Conigrave B. Canada's transition to net zero emissions. OECD Economics Department Working Papers No. 1760. Paris: OECD Publishing; 2023. https://doi.org/10.1787/efc1f36a-en.

[11] Su J, Cao L, Lee G, Tyler J, Ringsred A, Rensing M, et al. Challenges in determining the renewable content of the final fuels after co-processing biogenic feedstocks in the fluid catalytic cracker (FCC) of a commercial oil refinery. Fuel 2021;294:120526. https://doi.org/10.1016/j.fuel.2021.120526.

[12] Fogassy G, Thegarid N, Schuurman Y, Mirodatos C. From biomass to bio-gasoline by FCC co-processing: effect of feed composition and catalyst structure on product quality. Energy Environ Sci 2011;4:5068-76. https://doi.org/10.1039/C1EE02012A.

[13] Wei Y, Xu D, Xu M, Zheng P, Fan L, Leng L, Kapusta K. Hydrothermal liquefaction of municipal sludge and its products applications. Sci Total Environ 2024;908:168177. https://doi.org/10.1016/j.scitotenv.2023.168177.

[14] Xia H, Tang J, Aljerf L, Wang T, Gao B, Xu Q, et al. Assessment of PCDD/Fs formation and emission characteristics at a municipal solid waste incinerator for one year. Sci Total Environ 2023;883:163705. https://doi.org/10.1016/j.scitotenv.2023.163705.

[15] Xia H, Tang J, Aljerf L, Wang T, Qiao J, Xu Q, et al. Investigation on dioxins emission characteristic during complete maintenance operating period of municipal solid waste incineration. Environ Pollut 2023;318:120949. https://doi.org/10.1016/j.envpol.2022.120949.

[16] Elliott DC, Biller P, Ross AB, Schmidt AJ, Jones SB. Hydrothermal liquefaction of biomass: Developments from batch to continuous process. Bioresour Technol 2015;178:147-56. https://doi.org/10.1016/j.biortech.2014.09.132.

[17] van Dyk S, Su J, McMillan JD, Saddler J. Potential synergies of drop-in biofuel production with further co-processing at oil refineries. Biofuels Bioprod Biorefin 2019;13(3):760-75. https://doi.org/10.1002/bbb.1974.

[18] Han X, Wang H, Zeng Y, Liu J. Advancing the application of bio-oils by co-processing with petroleum intermediates: a review. Energy Convers Manag X 2021;10:100069. https://doi.org/10.1016/j.ecmx.2020.100069.

[19] Bezergianni S, Dimitriadis A, Kikhtyanin O, Kubička D. Refinery co-processing of renewable feeds. Prog Energy Combust Sci 2018;68:29-64. https://doi.org/10.1016/j.pecs.2018.03.003.

[20] Zacher AH, Elliott DC, Olarte MV, Wang H, Jones SB, Meyer PA. Technology advancements in hydroprocessing of bio-oils. Biomass Bioenergy 2019;125:151-68. https://doi.org/10.1016/j.biombioe.2019.05.004.

[21] Stefanidis SD, Kalogiannis KG, Lappas AA. Co-processing bio-oil in the refinery for drop-in biofuels via fluid catalytic cracking. WIREs Energy Environ 2018;7(3):e281. https://doi.org/10.1002/wene.281.

[22] Talmadge MS, Baldwin RM, Biddy MJ, McCormick RL, Beckham GT, Ferguson GA, et al. A perspective on oxygenated species in the refinery integration of pyrolysis oil. Green Chem 2014;16:407-53. https://doi.org/10.1039/c3gc41951g.

[23] Lammens TM. Effect of Various Green Carbon Tracking Methods on Life Cycle Assessment Results for Fluid Catalytic Cracker Co-processing of Fast Pyrolysis Bio-oil. Energy Fuels 2022;36(20):12617-27. https://doi.org/10.1021/acs.energyfuels.2c01676.

[24] de Mello LF, Pimenta RDM, Moure GT, Pravia ORC, Gearhart L, Milios PB, et al. A technical and economical evaluation of CO2 capture from FCC units. Energy Procedia 2009;1(1):117-24. https://doi.org/10.1016/j.egypro.2009.01.018.

[25] Su J, van Dyk S, O'Connor D, Saddler J. A comparison of methods used to track the 'green molecules' and determine the carbon intensities of co-processed fuels. Biofuels Bioprod Biorefin 2023;17(4):775-85.

[26] Dell'Orco S, Christensen ED, Iisa K, Starace AK, Dutta A, Talmadge MS, Magrini KA, Mukarakate C. Online biogenic carbon analysis enables refineries to reduce carbon footprint during coprocessing biomass-and petroleum-derived liquids. Anal Chem 2021;93(10):4351-60. https://doi.org/10.1021/acs.analchem.0c04108.

[27] Li ZH, Magrini-Bair K, Wang H, Maltsev OV, Geeza TJ, Mora CI, Lee JE. Tracking renewable carbon in bio-oil/crude co-processing with VGO through 13C/12C ratio analysis. Fuel 2020;275:117770.

[28] O'Connell A, Su J, Ringsred A, Prussi M, Saddler J, Scarlat N. Tracking the Biogenic Component of Lower-Carbon Intensive, Co-Processed Fuels—An Overview of Existing Approaches. Appl Sci 2022;12(24):12753. https://doi.org/10.3390/app122412753.

[29] Su J, van Dyk S, Saddler J. Repurposing oil refineries to "stand-alone units" that refine lipids/oleochemicals to produce low-carbon intensive, drop-in biofuels. J Clean Prod 2022;376:134335. https://doi.org/10.1016/j.jclepro.2022.134335.

[30] Su J, Cao L, Lee G, Tyler J, Ringsred A, Rensing M, van Dyk S, O'Connor D, Pinchuk R, Saddler J. Challenges in determining the renewable content of the final fuels after co-processing biogenic feedstocks in the fluid catalytic cracker (FCC) of a commercial oil refinery. Fuel 2021;294:120526. https://doi.org/10.1016/j.fuel.2021.120526.

[31] Tang J, Zhuang J, Aljerf L, Xia H, Wang T, Gao B. Numerical simulation modelling on whole municipal solid waste incineration process by coupling multiple software for the analysis of grate speed and air volume ratio. Process Saf Environ Prot 2023;176:506-27. https://doi.org/10.1016/j.psep.2023.05.101.

[32] Tang J, Xia H, Aljerf L, Wang D, Ukaogo PO. Prediction of dioxin emission from municipal solid waste incineration based on expansion, interpolation, and selection for small samples. J Environ Chem Eng 2022;10(5):108314. https://doi.org/10.1016/j.jece.2022.108314.

[33] Zhuang J, Tang J, Aljerf L. Comprehensive review on mechanism analysis and numerical simulation of municipal solid waste incineration process based on mechanical grate. Fuel 2022;320:123826. https://doi.org/10.1016/j.fuel.2022.123826.

[34] Cao L, Su J, Wang Y, Cao Y, Siang LC, Li J, Saddler J, Gopaluni B. Causal discovery based on observational data and process knowledge in industrial processes. Ind Eng Chem Res 2022;61(38):14272-83. https://doi.org/10.1021/acs.iecr.2c01326.

[35] Cao L, Su J, Saddler J, Cao Y, Wang Y, Lee G, Siang L, Pinchuk R, Li J, Gopaluni B. Real-time tracking of renewable carbon content with AI-aided approaches during co-processing of biofeedstocks. Appl Energy 2024;360:122815. https://doi.org/10.1016/j.apenergy.2024.122815.

[36] Cruz PL, Montero E, Dufour J. Modelling of co-processing of HDO-oil with VGO in a FCC unit. Fuel 2017;196:362-70. https://doi.org/10.1016/j.fuel.2017.01.112

[37] Xia H, Tang J, Aljerf L, Cui C, Gao B, Ukaogo PO. Dioxin emission modeling using feature selection and simplified DFR with residual error fitting for the grate-based MSWI process. Waste Manag 2023;168:256-71. https://doi.org/10.1016/j.wasman.2023.05.056.

[38] Hunter JS. The Exponentially Weighted Moving Average. J Quality Technol 1986;18(4):203-10. https://doi.org/10.1080/00224065.1986.11979014.

[39] Estevez PA, Tesmer M, Perez CA, Zurada JM. Normalized Mutual Information Feature Selection. IEEE Trans Neural Netw 2009;20(2):189-201. https://doi.org/10.1109/TNN.2008.2005601.

[40] Meinshausen N, Bühlmann P. High-dimensional Graphs and Variable Selection with the Lasso. Ann Stat 2006;34(3):1436-62. https://doi.org/10.1214/009053606000000281.

[41] Sun W, Braatz RD. ALVEN: Algebraic learning via elastic net for static and dynamic nonlinear model identification. Comput Chem Eng 2020;143:107103. https://doi.org/10.1016/j.compchemeng.2020.107103.

[42] Kursa MB, Rudnicki WR. Feature selection with the Boruta package. J Stat Softw 2010;36:1-13. https://doi.org/10.18637/jss.v036.i11.

[43] Yang Y, Lv H, Chen N. A survey on ensemble learning under the era of deep learning. Artif Intell Rev 2023;56:5545-5589. https://doi.org/10.1007/s10462-022-10283-5.

[44] Su J, Cao L, Lee G, Gopaluni B, Siang LC, Cao Y, van Dyk S, Pinchuk R, Saddler J. Tracking the green coke production when co-processing lipids at a commercial fluid catalytic cracker (FCC): combining isotope 14C and causal discovery analysis. Sustain Energy Fuels 2022;6(24):5600-7. https://doi.org/10.1039/D2SE01373H.

[45] Fraley C, Hesterberg T. Least Angle Regression and LASSO for Large Datasets. Stat Anal Data Min 2009;1(4):251-9. https://doi.org/10.1002/sam.10021.

[46] Ighalo JO, Adeniyi AG, Marques G. Application of Linear Regression Algorithm and Stochastic Gradient Descent in a Machine-learning Environment for Predicting Biomass Higher Heating Value. Biofuels Bioprod Biorefin 2020;14(6):1286-95. https://doi.org/10.1002/bbb.2140.

[47] Ohlson JA, Kim S. Linear valuation without OLS: the Theil-Sen estimation approach. Rev Account Stud 2015;20:395-435. https://doi.org/10.1007/s11142-014-9300-0.

[48] Huber PJ. Robust Estimation of a Location Parameter. Ann Math Stat 1964;35(1):73-101. 10.1214/aoms/1177703732.