Cite this: DOI: 00.0000/xxxxxxxxxx

Tracking the green coke production when co-processing lipids at a commercial fluid catalytic cracker (FCC): combining isotope ^{14}C and causal discovery analysis

Jianping Su,[a] Liang Cao,[b‡] Gary Lee[c], Bhushan Gopaluni[b],Lim C. Siang[c], Yankai Cao[b], Susan van Dyk[a], Robert Pinchuk[c], Jack Saddler[a†]

Received Date
Accepted Date

DOI: 00.0000/xxxxxxxxxx

Co-processing biogenic feedstocks allows oil refiners to use their infrastructure while reducing the carbon intensity of the fuels they produce. Although policies such as British Columbia and California's low carbon fuel standards have incentivized refiners to make these lower carbon intensity fuels, tracking the "green molecules" has proven to be challenging, particularly if the biogenic feedstocks are inserted at the fluid catalytic cracker. Various models based on commercial fluid catalytic cracker co-processing data were used to predict the green component (the renewable part) of combusted coke with these values compared to the results obtained using ^{14}C analysis. As the complexity and cost of sampling the flue gas made frequent testing impractical, a model that could better predict the renewable content of the fuels was developed. A combination of process data assessment and causal discovery significantly minimized prediction errors and provided a more robust model. This approach, combined with regular ^{14}C validation, is the most practical way to quantify the renewable content of the fuels when following a co-procesing regime and will likely be needed by both refiners and policymakers.

Keywords: Tracking green molecules; Co-processing; Decarbonization; Causal discovery; Low carbon intensity fuels

1 Introduction

Decarbonizing the transport sector has proven to be challenging with renewable fuels only contributing around 4% of the world's fuels, even after several decades of development[1]. As "conventional biofuels" such as bioethanol and biodiesel are not "drop-in", they have limited potential to decarbonize long distant transport such as aviation, trucking and marine[2]. To date, drop-in biofuels such as renewable diesel are produced by dedicated "standalone" refineries, using lipid feedstocks, as exemplified by companies such as Neste[3]. An alternative approach to making lower carbon intensity (CI) fuels is to co-process biogenic feedstocks at existing oil refineries as this approach makes use of existing infrastructure, downstream supply chains and expertise in processing/selling liquid fuels[4,5]. Co-processing has been fully commercialized in various parts of the world, such as Europe and North America, with policies such as the low carbon fuels standard (LCFS) incentivizing the production and use of lower-CI fuels[6,7].

Currently, co-processing is achieved by feeding oleochemical/lipid feedstocks, such as fats, oils, and greases (FOG's), into various unit operations in existing oil refineries[4,8]. As the global availability of lipids is limited, in the future, it is anticipated that biomass-derived biocrudes will supplement lipid feedstocks[9,10]. For example, the Swedish oil company Preem initially co-processed tall oil at their Gothenburg refinery in 2010[11]. Currently, they have not only increased their co-processing ratio from 30% to 85%[12] and have also commercialized co-processing fast pyrolysis oils made from sawdust or agricultural residues[13].

Progress in this area has been primarily "incentivized" by policies that require fuel suppliers to reduce the carbon intensities of the fuels they produce. However, to generate credits, both the volume of the renewable fraction and its carbon intensity needs to be quantified. Earlier work has shown that tracking the green molecules can be quite challenging when biogenic feedstocks are blended with fossil fuels[9,14,15]. One problem is the unequal allocation of green molecules to each fraction combined with the limited ways in which the renewable content of each stream can be quantified. For example, ASTM-certified AMS ^{14}C quantification is costly, time-consuming and is typically carried out via a third party. Similarly, sampling the various streams, such as fuel gases with high H2S content, is challenging and expensive. Consequently, alternative, more user-friendly and cheaper methods are needed to supplement ^{14}C monitoring with the hope that "soft sensors" can be established at various points within a refinery so data can be collected and validated[16,17].

As reported previously[18], commercial refinery operation data combined with multiple linear regression and a bootstrap method was successfully used to establish a "soft sensor" to estimate the amount of green coke generated during co-processing. These values were further assessed by comparing the ratio between the bio-

[a] *Forest Products Biotechnology/Bioenergy Group, The University of British Columbia, Vancouver, British Columbia, V6T 1Z4, Canada*
[b] *Data Analytics and Intelligent Systems Lab, The University of British Columbia, Vancouver, British Columbia, V6T 1Z3, Canada*
[c] *Parkland Refining (B.C.) Ltd., 2025 Willingdon Ave, Burnaby, British Columbia, V5J 0J3, Canada*
‡ These authors contributed equally to this work.
† Corresponding author: Jack (John) Saddler. E-mail: Jack.Saddler@ubc.ca

genic and fossil feedstock flows. However, when the various components that were relevant to the amount of coke generated were also considered, it proved difficult to determine which features were the most important. For example, statistical features such as the t value only showed statistical significance as compared to whether the values made sense for the refining operations, plus, how many features are required for effective modeling. Although the model is likely to perform better with the incorporation of more variables derived from training data, it could be less effective at predicting unseen testing data due to factors such as overfitting.

In the work described here, causal discovery analysis was incorporated into the model to try to better identify key features as causal-and-effect variables should contain all the necessary information and should be invariant in all situations[19]. For example, causal discovery does not need to determine the number of selected features in advance. Thus, the interpretability and robustness of the proposed model should be enhanced due to the introduction of causality analysis. If process-based industries, such as refinery co-processing biogenic feedstocks, can successfully use causal discovery-based models to mitigate the limitation of correlation and traditional machine learning algorithms, this should provide a useful way of tracking the green molecules during and after co-processing.

2 Methods

To predict the green/renewable molecules present in a particular refinery stream, after co-processing biogenic feedstocks, a data-driven model was developed. The model made use of long-term operational data, with the determined values compared to values derived by ^{14}C analysis. As AMS isotope ^{14}C analysis is only available at a few, dedicated labs, combined with the time taken and the cost associated with analysis, this method is unlikely to be incorporated into routine refinery lab analysis. However, ^{14}C analysis is the only method certified by ASTM and it is currently the method used by refiners who are co-processing biogenic feedstocks to quantify the carbon intensities of their processes and fuels, with the driver that they generate credits under British Columbia's and California's low carbon fuel standard (LCFS)[20].

2.1 FCC co-processing and data retrieval using Seeq

As oleochemical/lipid feedstocks are being co-processed by Parkland at their refinery in Burnaby, British Columbia, Canada[14,18], hourly data has been continuously generated over the last 17 months (as compared to the 12 months of data reported in previous work) [18]. With the accumulation of more data, the work reported here also assessed whether the coefficients that were used previously to calculate the amount of green molecules had changed in any way. The Seeq datalab was used to connect the refinery's process data with open access python library (70% for training and 30% for testing)[21].

As there is no direct way of measuring the coke burn, the reported values were derived from the flue gas as all of the coke generated is combusted to CO_2 and the flow of CO_2 is monitored continuously[22]. Consequently, we monitored the amount of CO_2 in the flue gas and built models to predict the green/renewable fraction of the flue gas. Essentially, we were trying to build a reliable and interpretable "soft sensor" that could measure the renewable content of the coke stream.

As described earlier[18,22], the selection of parameters that impacted the amount of coke produced and burned were based on process knowledge and the FCC heat balance, which is at the core of the catalytic cracking reactions. For example, just enough coke is burnt to satisfy the heat demand for the reactions such as heating the feedstocks, providing heat for the endothermic catalytic cracking reactions, etc. As discussed below, causal discovery was used to select model components and they were compared to established methods such as correlations.

2.2 Data "pretreatment"

The filtering of the commercial process data included removing data that was outside of a defined threshold as well as identifying outliers. For example, outliers, such as a lower feed rate, might be due to a turnaround or an occasional unit upset. Consequently, this data is likely to be not stable/not representative. Thus, 3-σ limits were used to set the upper and lower threshold limits. This provided a statistical calculation that refers to data x within three standard deviations (3σ) of the mean μ. The values within three standard deviations account for about 99.73% ($P(\mu-3\sigma \leq x \leq \mu+3\sigma) \approx 99.73\%$) and can usually be considered as normal operating process data.

Data standardization involves re-scaling the range so that the standardized data x_s has a zero-mean and unit-variance, but without distorting the differences over the range of x ($x_s = (x-\mu)/\delta$).). It is important to standardize the data when comparing measurements that have different units since variables measured at different scales do not contribute equally. For example, if one of the features has a broad range of values, the model outputs may be largely impacted by this particular feature. In the work reported here, we first removed any outliers and any data that was beyond the 3-σ limits threshold, then compared the impact of standardization on the model.

2.3 Causal discovery feature selection from observational data

In machine learning, when one wants to infer a target variable Y with a set of variables F, a subset S of F is usually sufficient. Thus, other variables (subset V) are not needed. However, when all the process variables are included, the "soft sensor" will be complex, possibly leading to overfitting and reduced generalization ability. Consequently, this influences the accuracy of the online prediction. A subset S that contains all the useful information is called a Markov blanket $MB(Y)$[23] and it can be summarised in the following equation:

$$P(Y \mid F) = P(Y \mid S,V) = P(Y \mid MB(Y),V) = P(Y \mid MB(Y)) \quad (1)$$

The Markov blanket (yellow zone in Figure 1) contains the parents (the variable that points to the target variable), the children (the variable that the target variable point to) and the spouse of

the target variable (another variable that also points to the target variable). The definition of a Markov blanket comes from the structural causal model framework and it establishes the optimal feature subset for the best prediction[24]. In the work reported here, we assessed the use of causal discovery to refine feature selection while comparing it with traditional selection methods which are typically based on process knowledge or statistical values such as the t and p values.

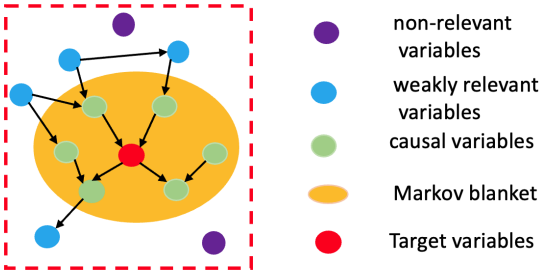

Fig. 1 An example of Markov blanket

A criticism of many correlation methodologies such as multiple linear regression is that correlation does not necessarily mean causation. Consequently, in the work reported here, we assessed if causal discovery could be used to select specific features. Past work has shown that causality can be used to obtain an interpretable stable model in complex industrial processes[17,25].

To discover causality or identify a Markov blanket from large amounts of process variables, causal discovery methods are typically divided into methods based on non-temporal[26–28] or temporal observation data[29,30]. Although the time dimension in the time series data contains important information that the effect cannot occur before the cause in time, the results of the time series data are sensitive to factors such as the frequency of data collection. As it is difficult to identify high time resolution causal relations from low time resolution data, causal discovery algorithms based on non-temporal observation data are typically used to provide a wider range of applications.

For causal discovery algorithms based on non-temporal observation data, the three predominant kinds of dedicated causal discovery algorithms are constraint-based[26], score-based[28] and casual-function-based[27]. The constraint-based algorithms construct the causal structure with the conditional independence constraint. The score-based algorithms construct the causal structure using scoring functions and search algorithms to select the best causal network. The causal function-based algorithms construct the causal structure using special assumptions regarding the data generation mechanisms used.

In the work reported here, we used a causal-function-based discovery algorithm called the Direct linear non-Gaussian acyclic model (DirectLiNGAM)[27]. This model algorithm contains three basic assumptions including a direct acyclic graph(DAG), the relationship between variables is linear, plus the disturbances are independent and non-Gaussian. For the observation data matrix $X = (x_1, \cdots, x_d)$, the structural causal model of DirectLiNGAM can be represented as follows:

$$x_i = \sum_{k(j)<k(i)} b_{ij} x_j + e_i \Leftrightarrow X = BX + e \Leftrightarrow X = (I - B)^{-1} e \quad (2)$$

Assuming the data in this equation is generated by $BX + e$, B is a lower triangular matrix, $k(i)$ denotes the causal order of x_i, e_i denotes the noise of x_i. The primary goal is to estimate connection matrix B, causal order k, and disturbance e from the observation data X. Define a variable with no parents as an exogenous variable, the following lemmas can be used to find the causal structure.

Lemma1[27]: If x_j and its residual $r_i^{(j)} = x_i - \frac{\mathrm{cov}(x_i, x_j)}{\mathrm{var}(x_j)} x_j$ are independent for all $i \neq j$, then x_j is exogenous variables.

Lemma2[27]: If x_j is exogenous variable and define $r^{(j)}$ is a vector that collects the residuals $r_i^{(j)}$ when all x_i of X are regressed on $x_j (i \neq j)$. Then we can prove that $r^{(j)}$ still satisfies the LiNGAM model, $r^{(j)} = B^{(j)} r^{(j)} + e^{(j)}$.

In the DirectLiNGAM model, we can iteratively use Lemma 2 to find the exogenous variables and place the exogenous variables at the top until all of the other variables are ordered.

One of the reasons the DirectLiNGAM model was used is that it guarantees the convergence of the right solution within a small fixed number of steps if all the model assumptions are met and the sample size is infinite. It is not based on an iterative search in the parameter space and needs no initial guesses. It should be noted that existing causality analysis inevitably introduces false associations in the process of causal topology modelling of industrial process data. Thus, causal discovery algorithms should contain as much process knowledge as possible as, process knowledge, like causal orders and connections, can narrow down the search and result in more efficient learning. Examples of process knowledge are experience with factors such as complex material flow, information (control loop) and energy flow, the physical connection of multiple devices, etc. The DirectLiNGAM model can integrate this knowledge into the algorithm, and consequently, provides a more accurate causal structure.

2.4 Regression models

As mentioned earlier, a soft sensor contains two main components which include the selection of process variables and the establishment of regression models. One goal of the work reported here was to determine the ratio between the coefficient of the biogenic flow and fossil flow in the coke model so that the input (the co-processing ratio) could be used to estimate the amount of green coke/green flue gas generated. Consequently, causality-based selection plus linear regression was desirable since it facilitated the interpretation while maintaining a high prediction performance. To evaluate the effectiveness of this method against other machine learning methods, we compared it with correlation-based selection plus linear regression, coefficient of decision tree-based selection plus linear regression[31], robust regression[32], partial least squares (PLS) regression[33] and light gradient boosting machine (LightGBM) regression[34].

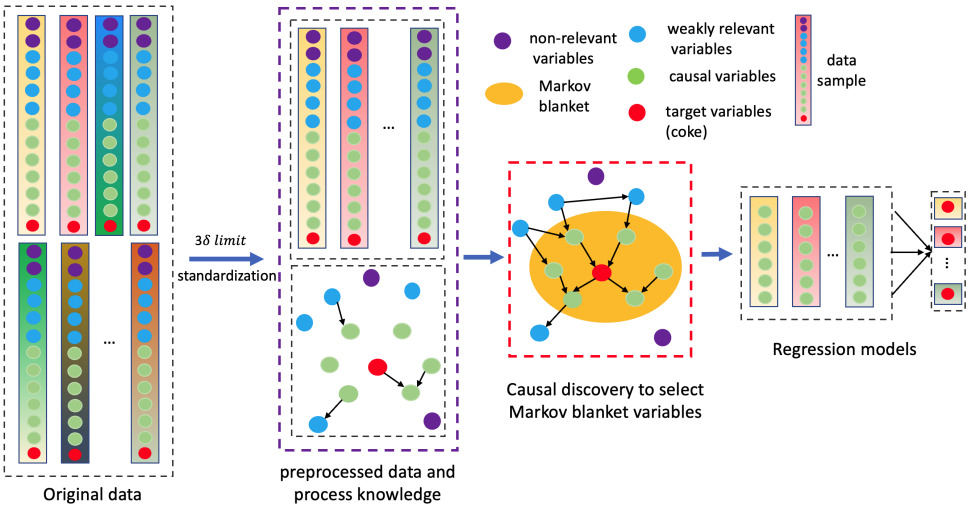

Fig. 2 Framework of causality-based soft sensor

2.5 Causality based green coke soft sensor

Figure 2 summarizes the framework of causality-based green coke, as determined by the soft sensor. Data "cleaning" involved the removal of outliers and any data that exceeded the 3σ limit, plus standardization of the cleaned data. Using this pretreated data and process knowledge regarding the causality among variables, the causal discovery was used to identify the causal graph of coke.

As the correlation scores and the coefficient of decision tree selection methods do not specify how many features should be selected, it first appears as though determining the best model is a trial-and-error process. Thus, to ensure an equitable comparison with causal discovery, we selected the same variables that were picked by the Markov blanket.

2.6 Estimation of the uncertainties by bootstrap method

To estimate a probable range, a model containing a bootstrap was used to determine the ratio between the coefficient of the biogenic flow and fossil flow in the coke model. As reported previously[18], a bootstrap method can be used to deal with the coefficients estimation uncertainties. This was done by repeatedly taking partial samples, calculating the coefficients, and taking the average and standard deviation of the coefficients. The average and standard deviation of the coefficients provided a range of ratios between the coefficient of the biogenic and fossil flow in the coke model.

2.7 Validation by isotope ^{14}C

One of the motivations of the reported work was that the renewable content of the coke stream is not routinely evaluated by the refinery, partly because this involves the use of ^{14}C, as defined by ASTM D6866. The few commercial labs able to perform this assay typically charge about 500 USD per sample[35,36]. As refiners increasingly co-process biogenic feedstocks as one way of decarbonizing their operation, quantifying the renewable content of the finished fuels becomes increasingly important, particularly if refiners are to obtain credits for their decarbonization efforts.

One way to validate the modeling results is to take flue gas samples and send them to a third party to quantify their renewable content via isotope ^{14}C. However, as well as being time-consuming and costly, another challenge is that the test results can sometimes be unrepresentative due to factors such as time delays, even though the coke generation reaction is rapid.

3 Results and discussion

3.1 Feature importance and selection - less is more

The initial features were selected based on knowledge of the process and knowing which factors would have the greatest impact on the amount of coke generated. To try to quantify the importance of these features, we used two components, the correlation scores and the coefficients of decision trees (Figures 3 and 4).

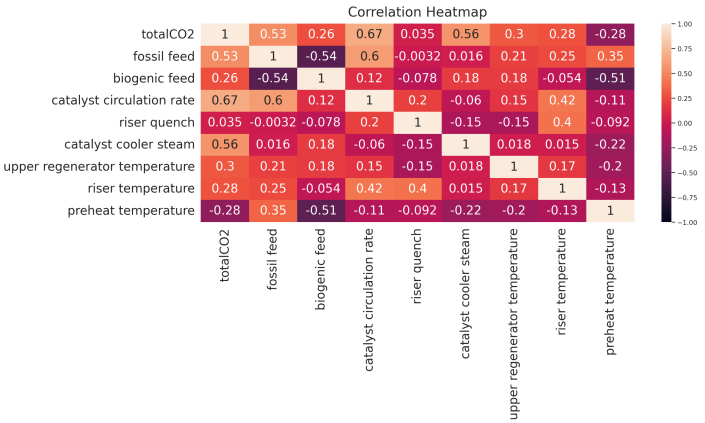

Fig. 3 Correlation of initial selected features

The correlation score is calculated with the Pearson correlation coefficient, which is used to measure the correlation (linear correlation) between two variables X and Y, with a value between -1 and 1. 1 means the total positive linear correlation, 0 means the no linear correlation, and -1 means the total negative linear

correlation. The definition of Pearson correlation can be given as follows:

$$r(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var[X]Var[Y]}} \tag{3}$$

where $r(X,Y)$ is the Pearson correlation coefficient between X and Y, $Var[X]$ is the variance of variable X, $Var[Y]$ is the variance of variable Y. $Cov(X,Y)$ is the covariance of X and Y. Based on the correlation score between input variables and total CO_2, top 5 varibles with high correlation with CO_2 are selected , including catalyst circulation rate (0.67), catalyst cooler steam (0.56),etc., to model total CO_2.

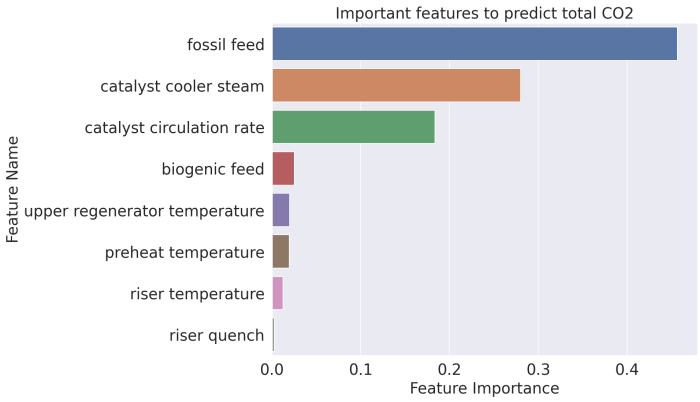

Fig. 4 Feature importance of initial selected features

The coefficient of decision trees is calculated based on entropy. Entropy is a measure of the uncertainty of a random variable. The definition of entropy can be given as follows:

$$\begin{aligned} &P(X = x_i) = p_i, i = 1,2,\cdots,n \\ &H(X) = -\sum_{i=1}^{n} p_i log p_i \end{aligned} \tag{4}$$

where $H(X)$ is the entropy of random variable X, n is the number of samples and p is the probability. Here the feature importance can be understood as the ability to provide reliable information (uncertainty reduction) in the output CO_2 prediction with knowing feature X. For example, if we know the value of fossil feed or catalyst cooler steam, we will be provided with lots of reliable information to predict CO_2, if we know the value of riser quench or riser temperature, little reliable information will be provided to predict CO_2.

It was apparent that the biogenic feed flow and preheat temperature showed a low correlation with the final coke/CO_2 generation, which contradicts the process itself as the flow of either stream (the fossil or the bio) will impact the amount of coke/CO_2 directly. It was also apparent that the preheat temperature, which is set by the operator, also impacts the coke yield of the unit as the amount of coke generate provides heat mainly to heat up the feeds. This highlighted the limitation of using correlation types of approaches for feature selection.

In contrast, the features selected by causal discovery, which were based on process knowledge, appeared to be more robust (Figure 5). Thus, we established the fossil feed flow rate and biogenic feed flow rate as exogenous variables, meaning no other features could influence the fossil and biogenic feed flow rates as they are controlled by the operators. Using the Markov blanket theory (Figure 1), five variables within the Markov blanket of total CO_2, including fossil feed, biogenic feed, preheat temperature, catalyst cooler steam, and catalyst circulation rate, were selected to model total CO_2. Surprisingly, the riser temperature was not covered within the Markov blanket, probably as, from a process point of view, changing the riser temperature likely changed the product profile. However, it should be noted that the riser temperature is usually set as a constant as it rarely changes.

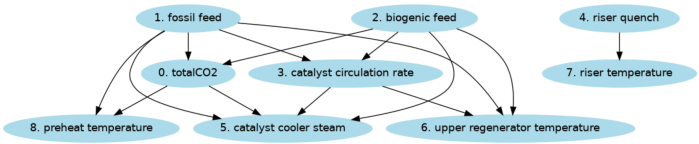

Fig. 5 Causal discovery of total CO_2 with process knowledge

It was also apparent that feature selection was tricky as, intuitively, more features should bring in more information and thus make better predictions. However, this is not true with real data which typically comes with considerable background noise. Thus, the more variables used, the more noise generated. It was hoped that by using the Markov blanket, just enough features were selected, containing just enough information and less noise.

3.2 Renewable coke soft sensors - simpler is superior

It should be noted that the number of features selected is critical when establishing a soft sensor. Theoretically, since the carbon source of CO_2 is either from fossil or biogenic feed, models based on these two features should be enough. However, the results summarised in Figure 6, using R^2,$RMSE$ of total CO_2 on the test data (RMSE means the distance between the predicted total CO_2 value made by different regression models and the actual total CO_2 value. R^2 means how well the predicted total CO_2 value made by the different regression models can explain the variation in the actual total CO_2 value.), show that the model is not reliable as, in real-world scenarios, where many parameters are changing without any control, relying on only two features can miss key information (Figure 6). Therefore, we next compared the models with the variables selected from a process point of view and compared them with the variables within the Markov blanket. It was apparent that, with data normalization, this improved the correlation in all three scenarios. However, the improvements were poorer when only two variables were selected. This showed that models based on causal discovery had the highest R^2 and lowest $RMSE$, were more effective, and resulted in a simpler model with fewer features (selected from the Markov blanket)(Figure 6).

Causal discovery based linear regression also outperformed when compared with other widely used machine learning algorithms such as correlation based, robust regression based, lightGBM regression based, PLS regression based and decision tree feature importance based soft sensors, as summarized in Figure 7. The causal discovery based linear regression soft sensor had the simplest structure, with the least features selected, and performed

better (lowest *RMSE*:494, and highest R^2:0.95) when compared to all of the other methods.

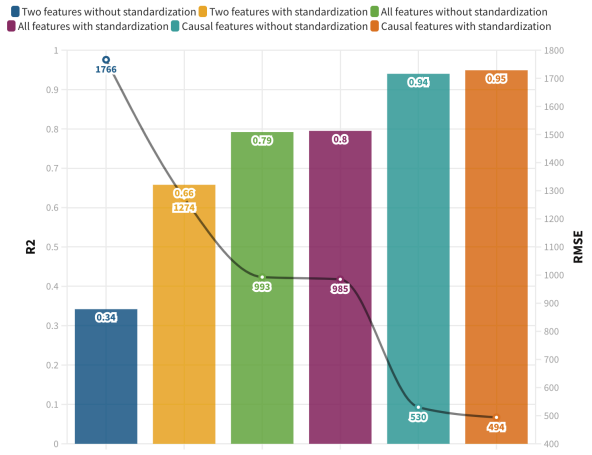

Fig. 6 Prediction performance index (R^2,RMSE) of different linear regression methods

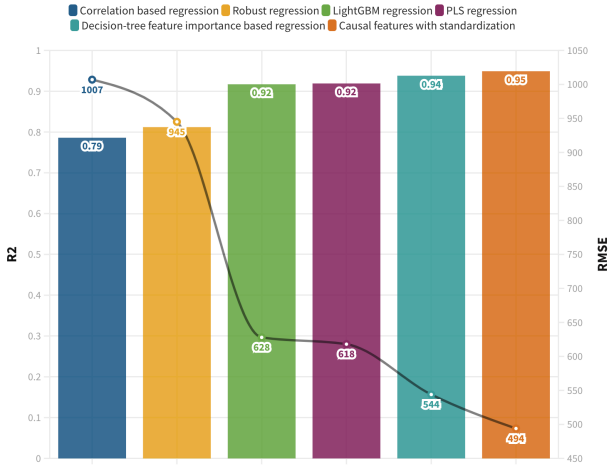

Fig. 7 Prediction performance index (R^2, RMSE) comparison of causal linear regression method with traditional machine learning methods

As discussed in the next section, we also hoped to use the ratio derived from the model to evaluate the green component of the streams.

3.3 Bootstrap regression coefficient analysis

As previously discussed, using causal discovery based regression methods can provide coefficients for both the fossil and biogenic feed and the ratio can be further used to infer the amount of green molecules in the products. However, the estimation of the coefficients can differ when the selection of training data varies. Therefore, using the bootstrap method and running each model 1 million times, generated a normal distribution for both the fossil and bio feed. This was used to calculate the 95% confidence interval (Figure 8).

The coefficient and 95% confidence interval of fossil and biogenic feed flow determined using the different methods are summarised in Table 1. Since the causal features using the standardization soft sensor showed the best performance, it is very likely that the coefficient and confidence interval used are optimum. Thus, by using this model, we could infer that increasing 1 unit of the fossil feed led to a 0.68 unit increase of burnt coke and, increasing 1 unit of biogenic feed resulted in a 0.43 unit increase of burnt coke. As the ratio between the coefficient of the biogenic and fossil feed flow was 0.63. Thus, within the 95% confidence interval, the lower bound ratio was 0.55 and the upper bound ratio was 0.72. Thus, theoretically, we should be able to determine the amount of green molecules when the feed ratio is known. To further validate the model the values obtained were next compared with the ^{14}C values that were determined for the samples.

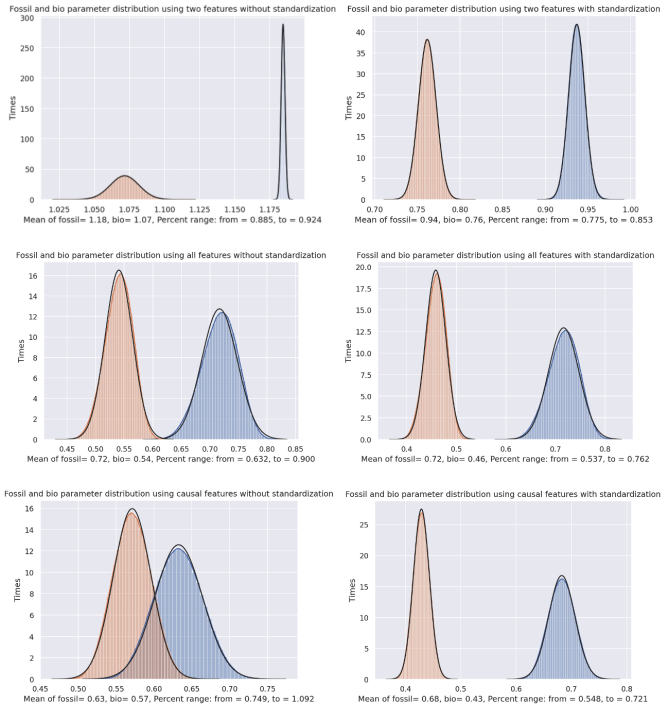

Fig. 8 Coefficient and 95% confidence interval of fossil and biogenic feed flow with bootstrap

3.4 Isotope ^{14}C validation and possible strategies for refiners and policymakers

^{14}C tests were conducted when a 16.2% co-processing (biogenic feed) ratio was used. As the biogenic carbon, as determined by this method, was 10.59% ± 0.06% and 10.51% ± 0.06% after duplicate assays, the ratio of the ^{14}C assays can be represented as:

$$test1 \quad ratio: \frac{\frac{10.59\%}{16.2\%}}{\frac{(100-10.59)\%}{(100-16.2)\%}} \approx \frac{0.65}{1.07} \approx 0.61$$

$$test2 \quad ratio: \frac{\frac{10.51\%}{16.2\%}}{\frac{(100-10.51)\%}{(100-16.2)\%}} \approx \frac{0.65}{1.07} \approx 0.61$$

These results indicated that the 95% confidence interval within the causal featured soft sensor (0.55-0.72) captured the results

Table 1 Coefficient and 95% confidence interval of fossil and biogenic feed flow in different methods

coefficient	fossil	bio	ratio (bio/fossil)	95% confidence lower bound	95% confidence upper bound
Two features without standardization	1.18	1.07	0.91	0.89	0.92
All features without standardization	0.72	0.54	0.75	0.63	0.90
Causal features without standardization	0.63	0.57	0.90	0.75	1.09
Two features with standardization	0.94	0.76	0.81	0.78	0.85
All features with standardization	0.72	0.46	0.64	0.54	0.76
Causal features with standardization	0.68	0.43	0.63	0.55	0.72
Isotope ^{14}C	1.07	0.65	0.61		

obtained by the two ^{14}C tests (0.61) with the error between the soft sensor (0.63) and ^{14}C assay (0.61) at about 3.3%.

In the future, as refineries will likely be asked to reduce their scope 1,2 and 3 emissions, quantifying the "green" component of flue gas will become increasingly important (scope 1 emissions). As there is currently no online monitoring equipment that provides quantification of the "green" component, intermittent ^{14}C analysis is the only way to quantify the "green" fractions with the assay only providing a "snapshot" of operations.

As refineries are increasingly under pressure to decarbonize, co-processing biogenic feedstocks provides one way to make use of the equipment and expertise within a refinery while producing lower carbon-intensive (CI) fuels. However, tracking and quantifying the amount of "green" molecules in the various streams is problematic with ^{14}C monitoring the only method accepted by policies such as California and BC's Low Carbon Fuels Standard (LCFS).

However, a combination of process data assessment and causal discovery significantly minimized prediction errors and could provide a more robust model. This approach, combined with regular ^{14}C validation, is likely to be the most practical way to quantify the carbon intensity of processes and fuels when following a co-processing regime with this method used by refiners and, hopefully, supported by policymakers.

4 Conclusions

Co-processing biogenic feedstocks will accelerate the decarbonization of transport fuels while leveraging global refining and their downstream supply chains. However, quantifying the renewable content of the various co-processed streams currently relies on sporadic ^{14}C monitoring which is expensive and only provides a "snapshot" of refinery operations and products. A combination of process data assessment and the use of the causal discovery model can be used to significantly minimize prediction errors and provide representative data. This approach, combined with regular ^{14}C validation, is likely to be the most practical way to quantify the carbon intensity of processes and fuels. The "fewer features based model" performed better than models with more features, likely due to the addition of less "noise" from more uncausal variables. The simple linear causal models performed better than other, typical, machine learning algorithms.

Author Contributions

Conceptualization, writing, model design, data analysis and original draft preparation were done by Jianping Su and Liang Cao. Jack Saddler, Bhushan Gopaluni, Yankai Cao, Gary Lee, Lim C. Siang, Susan van Dyk and Robert Pinchuk were involved in the review and editing of this manuscript, model analysis, and performing experiments. Jack Saddler, Bhushan Gopaluni and Yankai Cao supervised the work, discussed the results and contributed to the final manuscript. All authors provided critical feedback and helped shape the research, analysis and manuscript.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

Jianping Su and Liang Cao thank MITACS for financial support. We thank Jason Seeley for technical support and guidance on the flue gas. We also thank colleagues at Parkland for coordinating isotope ^{14}C testing/sampling.

Notes and references

1 IEA, Tracking transport 2021, 2021, https://www.iea.org/reports/tracking-transport-2021, Last accessed 16 March 2022.
2 S. Karatzos, J. D. McMillan and J. N. Saddler, Report for IEA Bioenergy Task, 2014, **39**, year.
3 A. Engman, T. Hartikka, M. Honkanen, U. Kiiski, L. Kuronen, K. Lehto, S. Mikkonen, J. Nortio, J. Nuottimäki and P. Saikkonen, Neste Proprietary Publication, Espoo, 2016.
4 S. van Dyk, J. Su, M. Ebadian and J. Saddler, Fuel, 2022, **324**, 124636.
5 S. Bezergianni, A. Dimitriadis, O. Kikhtyanin and D. Kubička, Progress in Energy and Combustion Science, 2018, **68**, 29–64.
6 M. Ebadian, S. van Dyk, J. D. McMillan and J. Saddler, Energy Policy, 2020, **147**, 111906.
7 S. van Dyk, J. Su, J. D. Mcmillan and J. Saddler, Biofuels, Bioproducts and Biorefining, 2019, **13**, 760–775.
8 S. van Dyk, J. Su, J. D. Mcmillan and J. Saddler, 2019.
9 A. de Rezende Pinho, M. B. de Almeida and P. R. Rochedo, Fuel Processing Technology, 2022, **229**, 107176.
10 A. de Rezende Pinho, M. B. de Almeida, F. L. Mendes, L. C. Casavechia, M. S. Talmadge, C. M. Kinchin and H. L. Chum, Fuel, 2017, **188**, 462–473.
11 R. Egeberg, K. Knudsen, S. Nyström, E. L. GRENNFELT and K. Efraimsson, Petroleum technology quarterly, 2011, **16**, year.
12 U. Frøhlke, HHaldor Topsoe and Preem achieve 85% co-processing of renewable feedstock, 2021,

https://blog.topsoe.com/haldor-topsoe-and-preem-achieve-85-co-processing-of-renewable-feedstock, Last accessed 05 September 2022.

13 UOP, Honeywell And Preem Conduct Commercial Co-Processing Trial to Produce Renewable Fuel, 2021, https://uop.honeywell.com/en/news-events/2021/september/honeywell-and-preem-conduct-commercial-co-processing-trial-to-produce-renewable-fuel, Last accessed 05 September 2022.

14 J. Su, L. Cao, G. Lee, J. Tyler, A. Ringsred, M. Rensing, S. van Dyk, D. O'Connor, R. Pinchuk and J. J. Saddler, Fuel, 2021, **294**, 120526.

15 M. Schimmel, G. Toop, S. Alberici and M. Koper, Final Report. Ecofys, 2018.

16 Z.-H. Li, K. Magrini-Bair, H. Wang, O. V. Maltsev, T. J. Geeza, C. I. Mora and J. E. Lee, Fuel, 2020, **275**, 117770.

17 L. Cao, F. Yu, F. Yang, Y. Cao and R. B. Gopaluni, Control Engineering Practice, 2020, **104**, 104626.

18 J. Su, L. Cao, G. Lee, B. Gopaluni, D. O'Connor, S. van Dyk, R. Pinchuk and J. Saddler, Biofuels, Bioproducts and Biorefining, 2022, **16**, 325–334.

19 P. Cui and S. Athey, Nature Machine Intelligence, 2022, **4**, 110–115.

20 C. A. R. Board, LCFS Pathways Requiring Public Comments, 2021, https://ww2.arb.ca.gov/resources/documents/lcfs-pathways-requiring-public-comments, Last accessed 05 September 2022.

21 S. Corporation, Seeq Data Lab, 2021, https://www.seeq.com/product/seeq-data-lab, Last accessed 05 September 2022.

22 R. Sadeghbeigi, Fluid catalytic cracking handbook: An expert guide to the practical operation, design, and optimization of FCC units, Butterworth-Heinemann, 2020.

23 X.-Q. Liu and X.-S. Liu, The Journal of Machine Learning Research, 2018, **19**, 1658–1707.

24 J. Pearl and D. Mackenzie, The Book of Why: The New Science of Cause and Effect, Basic Books, Inc., USA, 1st edn, 2018.

25 F. Yu, Q. Xiong, L. Cao and F. Yang, Control Engineering Practice, 2022, **122**, 105109.

26 P. Spirtes and C. Glymour, Social Science Computer Review, 1991, **9**, 62–72.

27 S. Shimizu, T. Inazumi, Y. Sogawa, A. Hyvärinen, Y. Kawahara, T. Washio, P. O. Hoyer and K. Bollen, J. Mach. Learn. Res., 2011, **12**, 1225–1248.

28 J. Ramsey, M. Glymour, R. Sanchez-Romero and C. Glymour, International Journal of Data Science and Analytics, 2017, **3**, 121–129.

29 J. Runge, S. Bathiany, E. Bollt, G. Camps-Valls and J. Zscheischler, Nature Communications, 2019, **10**, year.

30 J. Runge, Chaos: An Interdisciplinary Journal of Nonlinear Science, 2018, **28**, 075310.

31 S. M. Lundberg and S.-I. Lee, Advances in neural information processing systems, 2017, **30**, year.

32 J. Liu, P. C. Cosman and B. D. Rao, IEEE Transactions on Signal Processing, 2017, **66**, 698–713.

33 S. Wold, M. Sjöström and L. Eriksson, Chemometrics and intelligent laboratory systems, 2001, **58**, 109–130.

34 G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye and T.-Y. Liu, Advances in neural information processing systems, 2017, **30**, year.

35 M. Hurt, J. Martinez, A. Pradhan, M. Young and M. E. Moir, Energy & Fuels, 2020, **35**, 1503–1510.

36 M. R. Haverly, S. R. Fenwick, F. P. Patterson and D. A. Slade, Fuel, 2019, **237**, 1108–1111.