Causal Discovery based on Observational Data and Process Knowledge in Industrial Processes

Liang Cao,[†] Jianping Su,[‡] Yixiu Wang,[†] Yankai Cao,[†] Lim C. Siang,[¶] Jin Li,[¶] Jack Nicholas Saddler,[‡] and Bhushan Gopaluni[*,†]

[†]*Department of Chemical and Biological Engineering, University of British Columbia, Vancouver, BC, V6T 1Z3, Canada*

[‡]*Forest Products Biotechnology/Bioenergy Group, The University of British Columbia, Vancouver, BC, V6T 1Z4, Canada*

[¶]*Department of Process Control Engineering, Burnaby Refinery, BC V5C 1L7, Canada*

E-mail: bhushan.gopaluni@ubc.ca

Abstract

Causal discovery approaches are gaining popularity in industrial processes. Existing causal discovery algorithms can indeed find some important causal relationships from industrial data, but at the same time, the algorithms may also give some incorrect causal relationships. In order to deal with this problem, we give four kinds of process knowledge definitions according to the special characteristics of complex industrial processes. Causal discovery algorithms will yield more accurate results and deeper insights if the process knowledge is properly addressed. Based on commercial-scale fluid catalytic cracker (FCC) unit data, we validate the effectiveness of the proposed methods with some state-of-the-art causal discovery algorithms.

1. Introduction

Modern industrial processes are often characterized by high dimensions, strong multicollinearity, nonlinear and high noise.[1,2] How to successfully adapt the data-driven approach to the industrial process with the above characteristics has become the focus of the industry. Machine learning and deep learning have gained significant interest and have been the dominant approaches in industrial processes.[3–6] Due to their excellent predictive accuracy, they are successfully applied to soft sensor and process monitoring. However, most machine learning models are black-box models, and as such it is difficult to interpret their behaviour in relation to the process variables. Industrial processes involve risk-sensitive tasks, any accidental situation can lead to disastrous consequences. In this case, the model must not only obtain accurate predictions but also provide interpretability and guarantee the stability of the prediction results.[7]

Causality assumes that data is generated based on the causal mechanism, so causality is interpretable and stable.[8] Now, the problem is how to discover causality or identify causal relationships from large amounts of process variables and provide guidance on obtaining better results.

1.1 Literature Review of Causal Discovery

In industrial processes, we divide causal discovery methods into three main categories (see Figure 1), namely randomized experiments,[9] computer simulation experiments,[10] and methods based on observational data, where methods based on observational data are further divided into temporal observational data and non-temporal observational data.[11–27]

Randomized experiments are the traditional way to find causality, but such active interventions are costly and time-consuming. It randomly assigned subjects to different groups with different interventions. In the case of a sufficient number of subjects, this method can offset the effects of known and unknown confounding factors on each group. However,

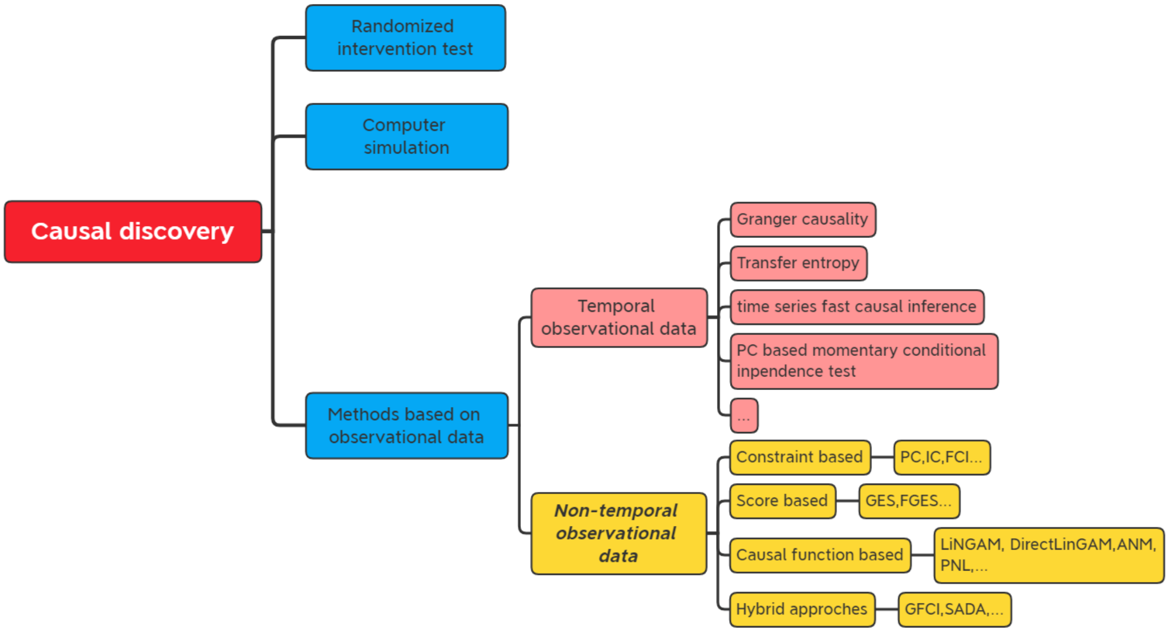
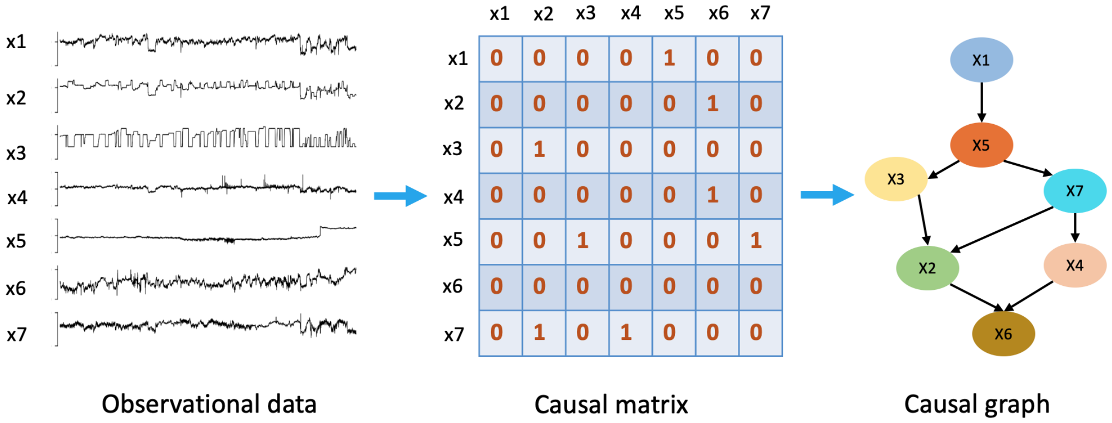

Causal discovery
Randomized intervention test
Computer simulation
Methods based on observational data
Temporal observational data
Granger causality
Transfer entropy
time series fast causal inference
PC based momentary conditional inpendence test
...
Non-temporal observational data
Constraint based
PC,IC,FCI...
Score based
GES,FGES...
Causal function based
LiNGAM, DirectLinGAM,ANM, PNL,...
Hybrid approches
GFCI,SADA,...

Figure 1: Casual discovery algorithms

in large-scale complex dynamical industrial processes, real experiments are rarely feasible. For the second method, computer simulation software like Aspen requires substantial expert knowledge and massive data during physical modelling. Therefore, it is difficult to find causal relationships in complex industrial processes without fully understanding the physical model and without big data on the process.

Causal discovery based on observational data avoids the above limitations and is currently a research hot spot in the field of causality. Furthermore, the causal discovery methods based on observational data can be divided into the methods based on non-temporal observational data[11–20] and the methods based on temporal observational data[23–27] . For temporal methods, Granger causality analysis[26] (GCA) and transfer entropy[25] (TE) are two of the most common methods for establishing pairwise causality of variables. Pairwise causality limits its application to indirect causation and confounders (common parents). Although temporal information can provide valid causal information, the results of temporal observational data are often sensitive to factors such as temporal resolution. In most cases, it is difficult to mine causal relationships that exist at high temporal resolution from low temporal resolution data.

In industrial processes, causal discovery algorithms based on non-temporal observational data have a wider range of applications, and we will focus on their applications in the following sections. Figure 2 shows a basic framework of casual discovery based on non-temporal observational data. There are mainly three kinds of dedicated causal discovery algorithms, constraint-based,[11,12,15,19] score-based,[20] and casual-function.[13,14,21,22]



	x1	x2	x3	x4	x5	x6	x7
x1	0	0	0	0	1	0	0
x2	0	0	0	0	0	1	0
x3	0	1	0	0	0	0	0
x4	0	0	0	0	0	1	0
x5	0	0	1	0	0	0	1
x6	0	0	0	0	0	0	0
x7	0	1	0	1	0	0	0

Figure 2: A basic framework of causal discovery based on non-temporal observational data

The constraint-based algorithms, like Peter-Clark (PC),[11] inductive causation (IC),[12] fast causal inference (FCI),[19] construct the causal structure based on conditional independence constraints. They mainly obtain the causal skeleton diagram through the conditional independence test and then learn the causal direction with the help of $d-$separation and V structure.[19] The constraint-based methods usually return partial undetermined causal directions (Markov equivalence class).

The score-based algorithms replace conditional independence tests with the scoring functions and search algorithms to select the best Bayesian causal network.[20] The problem is that this method involves a graph search process, which has high time complexity. It also assumes all confounders are observable, which is not practical. To deal with the Markov equivalence problems of constraint-based methods, many scholars and experts have proposed casual function approaches, like linear non-Gaussian acyclic model (LiNGAM),[13,14] additive noise model (ANM)[21] and post-nonLinear (PNL) model.[22] These approaches have special assumptions about the data generation mechanism. The disadvantage is that the assumptions usually have limited scope in real-world applications.

In recent years, some scholars have tried to combine the characteristics of different algorithms to design hybrid causal structure discovery algorithms, like Greedy Fast Causal Inference (GFCI)[17] and Split-and-Merge framework (SADA).[18] The hybrid approach tries to take full advantage of all approaches. It is mainly divided into two parts. First, use one approach to learn a basic skeleton with partial directions. Second, another approach is employed to further fine-tune the local causal structures.

1.2 Contribution and Organization

The existing causality analysis inevitably introduces false causal associations in the process of causal discovery since the industrial process data has the problems of high dimension, strong correlation and high noise. High dimensions increase the complexity of the causality analysis. High correlation means that redundant information is introduced into causality analysis, which makes it difficult to obtain accurate results. The presence of noise terms can also lead to inaccurate estimates of causal strength and even causal direction. Unfortunately, current causal discovery methods are unable to address the above-mentioned problems in industrial processes.

The good news is that there is a large amount of explicit or implicit process knowledge of industrial processes that can be used to overcome the above-mentioned shortcomings and improve the accuracy of causal discovery. This knowledge could be expert knowledge, transfer of complex material flow, information flow (control loop) and energy flow, the physical connection of multiple devices, etc. As an important contribution, we use process knowledge to narrow the search space and enable more efficient learning. In order to overcome the limitations of traditional causal discovery algorithms in industrial processes, we give several mathematical definitions of the process knowledge in the process of causal matrix learning and apply it to a commercial refinery for the first time. The commercial FCC co-processing unit demonstrates that the accuracy of current causal discovery algorithms is significantly improved when combined with process knowledge.

This work is organized as follows. Section 1 describes the motivation and reviews causal discovery algorithms. In section 2, detailed implementation procedures and algorithmic analysis of two prominent causal discovery algorithms are given. Section 3 verifies the effectiveness of proposed methodology with the commercial-scale FCC co-processing data. Section 4 discusses the current challenges and opportunities for causal discovery in industrial processes. Concluding remarks are presented in section 5.

2. Method

In this section, we introduce two powerful causal discovery algorithms, DirectLiNGAM (causal function) and GFCI (hybrid approach with constraint-based and score-based approach) , which cover the mainstream approaches in non-temporal causal discovery methods. It must be emphasized that the mentioned algorithms are only a small part of prominent causal discovery algorithms.

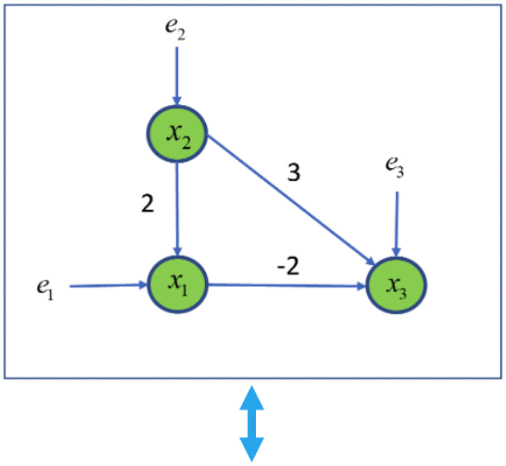
2.1 DirectLiNGAM

There are three basic assumptions in the DirectLiNGAM algorithm, i.e., direct acyclic graph(DAG), the relationship among variables is linear, disturbances are independent and non-Gaussian. Define observation data matrix $X = (x_1, \cdots, x_d)$, the structural equation model of DirectLiNGAM can be given as follows:

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

Here, data X is generated by $BX+e$, B is a lower triangular matrix, $k\left(i\right)$ denotes the causal order of x_i, e_i denotes the noise of x_i, our goal is to estimate connection matrix B, causal order k, and disturbance e from the observation data X.

Remark: For the acyclic graph assumption, we can interpret it from the perspective of stable processes. The industrial process is usually assumed to reach its equilibrium state. If the process is $X_t{=}BX_{t-1}{+}E_t$, then we have $X_t{=}X_{t-1}$ for each dynamic process at equilibrium state. Finally, the process will become $X_t{=}\mathrm{B}X_t{+}E_t \Rightarrow E_t{=}(I-B)X_t \Rightarrow E=(I-B)X$. The fundamental task of causal discovery is to find the causal matrix B, while B is the same in both cases if we have an equilibrium assumption.



$$\begin{cases} x_2 = e_2 \\ x_1 = 2x_2 + e_1 \\ x_3 = -2x_1 + 3x_2 + e_3 \end{cases} \Rightarrow \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & 2 & 0 \\ 0 & 0 & 0 \\ -2 & 3 & 0 \end{bmatrix}}_{\overline{B}} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \end{bmatrix}$$

Figure 3: A numerical example of DirectLiNGAM

Assuming there is a three-variable model and the structure of model is given in Figure 3. The causal order is $x_2 \to x_1 \to x_3$, then we have $k(2) = 1, k(1) = 2, k(3) = 3$, x_2 is equal to e_2 and therefore is an exogenous variable (variable with no parents). B (lower triangular matrix) can be obtained by permutation and scaling the matrix $\overline{B}$ in Figure 3. In DirectLiNGAM, we iteratively find exogenous variables and put exogenous variables at the top of the order until all the variables are ordered.[14] In practice, we can identify an exogenous variable by finding a variable that is the most independent of its residuals. The independence can be evaluated by taking the sum of mutual information or correlation between variable x_j and all the residuals. The procedures of identifying exogenous variables and finding the lower triangular matrix B can be given by the following two lemmas:

Lemma1:[14] Assume that the data X has infinite samples and satisfies the three basic assumptions in the DirectLiNGAM algorithm. 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:[14] Assume that the data X has infinite samples and satisfies the three basic assumptions in the DirectLiNGAM algorithm. 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, $r^{(j)} = B^{(j)}r^{(j)} + e^{(j)}$, which means $B^{(j)}$ can be permuted to be a lower triangular matrix, and elements of $e^{(j)}$ are non-Gaussian and mutually independent.

2.2 GFCI

GFCI is a hybrid approach which combines the GES and FCI. There are two assumptions for GFCI, i.e., Markov assumption and faithfulness assumption.[17,19] Figure 4 gives two examples to illustrate the definition of the two assumptions.

Markov Assumption: Define a joint probability distribution P as:

$$P\left(x_1, \cdots, x_d\right) = \prod_{i=1}^{d} P\left(x_i \left| Pa_i\right.\right) \tag{2}$$

Pa_i is the parent of x_i. Define G as an acyclic causal graph with vertex set V, then P satisfies the Markov assumption for the causal graph G if the following equation holds for all disjoint vertex sets x_i, x_j, x_k in V (the symbol $\perp_G$ denotes $d-separation$, and $\perp$ denotes independence):

$$x_i \perp_G x_j \left| x_k \right. \Rightarrow x_i \perp x_j \left| x_k \right. \tag{3}$$

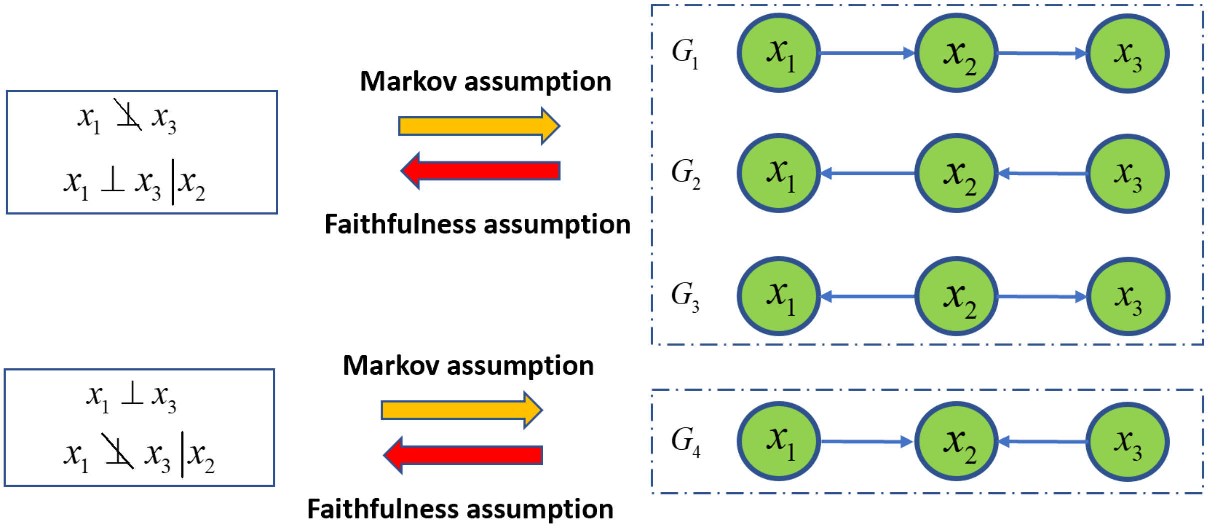


Figure 4: The examples of Markov and faithfulness assumption

Equivalently, x_i is dependent on x_j conditional on x_k in P only if x_i is $d-connected$ to x_j conditional on x_k in G.

Faithfulness Assumption: A joint probability distribution P satisfies the faithfulness assumption for causal graph G if the following equation holds for all disjoint vertex sets x_i, x_j, x_k:

$$x_i \perp_G x_j \,|x_k \Leftarrow x_i \perp x_j \,|x_k \tag{4}$$

Equivalently, x_i is $d-connected$ to x_j conditional on x_k in G only if x_i is dependent on x_j conditional on x_k in P.

2.2.1 FCI

FCI is a constraint-based method that could handle the problem of unmeasured confounders but often perform poorly when samples are small, or the causal graph is non-unique. There are two phases in the FCI, the first phase is to form a complete undirected graph by removing edges between variables that are unconditionally independent and conditionally independent. Then, it orients edges by identifying the collider structures (like G_4 in Figure 4) in the second phase.

2.2.2 GES

Unlike FCI which starts with a complete undirected graph, the GES algorithm starts with an empty graph and adds one directed edge to the graph at a time that most improves the score function. When the score cannot be improved any further, GES removes one edge at a time until no more edges can be removed to improve the scoring function. If there is an unmeasured latent confounder, then GES may include extra edges (spurious correlation) that are not in the true causal graph. The GES has great performance on small samples for the case without confounders.

2.2.3 GFCI

GFCI combines the advantages of original FCI and GES and has been shown to be more accurate in many situations. The detailed steps of GFCI are omitted for convenience, we refer the reader to the reference 17 to get more detailed descriptions. The main idea is to use GES first to improve the accuracy of both the adjacency and orientation phase of FCI by providing a more accurate initial graph. Then, for the output of GES, FCI is employed to further fine-tune the local structures and causal directions by removing the extra adjacencies and correcting the orientations.

2.3 Definition of Process Knowledge

Discovering causal relationships from observation data is bounded by different assumptions. Many prominent causal discovery algorithms, like DirectLiNGAM, and GFCI, have been proven effective to some extent in many cases. But for best results, causal discovery algorithms should be used by providing as much process knowledge as possible. Based on the characteristics of the industrial processes, we will propose four kinds of process knowledge in this work. Let us first define a process knowledge matrix $A^{kw} = \left[a_{ji}^{kw}\right]$ as follows:

$$a_{ji}^{kw} \doteq \begin{cases} 0 & \text{if } x_i \text{ does not have a directed edge to } x_j \\ 1 & \text{if } x_i \text{ has a directed edge to } x_j \\ \text{transpose equals 0,for all i} \neq j & \text{if } x_i \text{ is output variable} \\ \text{0,for all i} \neq j & \text{if } x_j \text{ is input variable} \\ -1 & \text{if no process knowledge is available} \end{cases} \tag{5}$$

If a_{ji}^{kw} is 0, which means a variable x_j does not receive the effect of x_i, there exists a forbidden edge between x_i and x_j. If a_{ji}^{kw} is 1, which means a variable x_j receives the effect of x_i, a edge is required between x_i and x_j. If a_{ji}^{kw} is 0 for all $i \neq j$, which means a variable x_j does not receive the effect of any other variables, variable x_j is an input variable. If $a_{ji}^{kw^T}$ is 0 for all $i \neq j$, which means a variable x_i does not produce effect on any other variables, variable x_i is output variable. Although the process knowledge definition in this work is similar to the process knowledge proposed in reference 14, this work focuses on the inputs, outputs and process connectivity information by considering the special characteristics of industrial processes.

2.4 Causal Discovery based on Observational Data and Process Knowledge

Figure 5 shows an example of the proposed method. At the top of Figure 5, we can find that, given the observational data, the result of causal discovery is quite different from the true graph when there is no process knowledge. At the bottom of Figure 5, when some process knowledge is provided, the result is greatly improved.

Figure 6 and Figure 7 give the flowchart of DirectLiNGAM and GFCI with process knowledge. The step-by-step implementation of the methodology is provided, which includes data preprocessing, how to use the process knowledge matrix, and how to get the causal

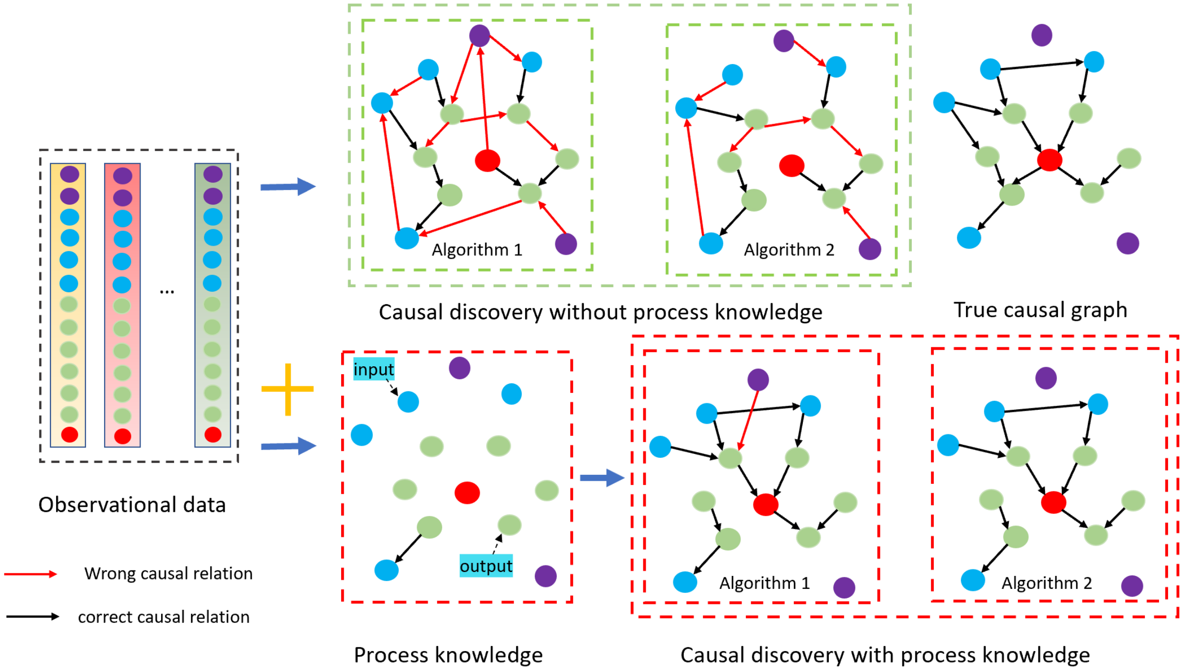


Figure 5: Framework of causal discovery based on observational data and process knowledge

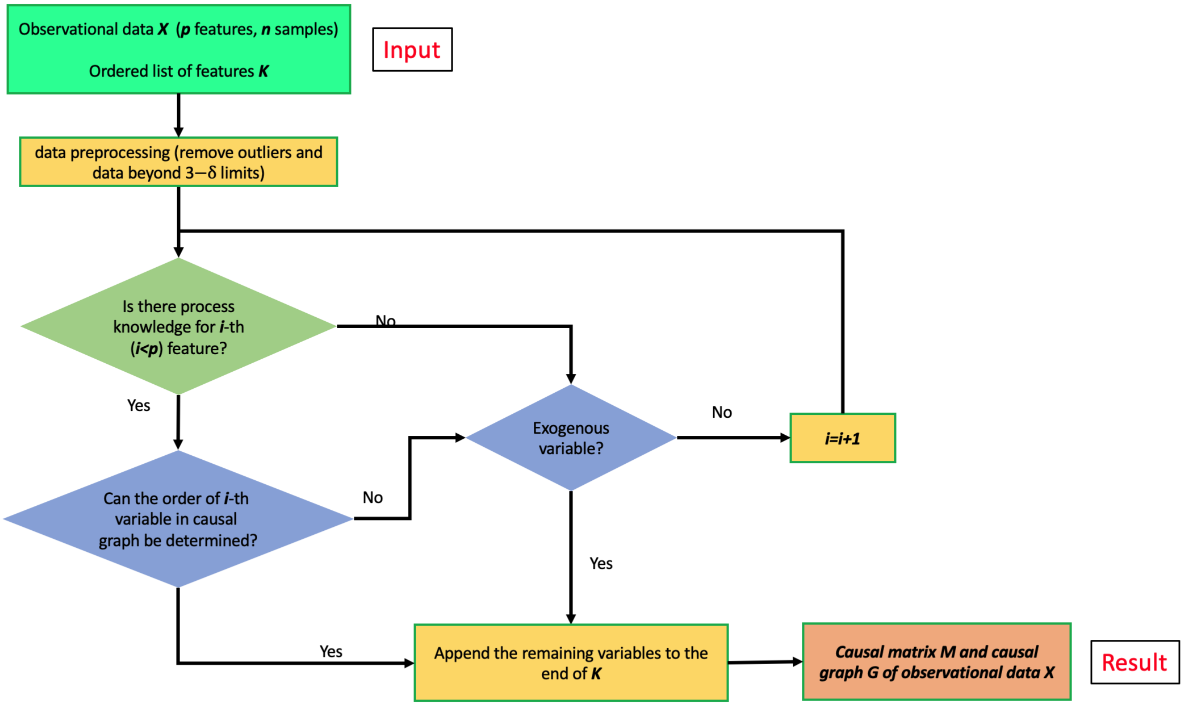

Observational data X (p features, n samples)
Ordered list of features K
Input
data preprocessing (remove outliers and data beyond 3−δ limits)
Is there process knowledge for i-th (i<p) feature?
No
Yes
Exogenous variable?
No
i=i+1
Can the order of i-th variable in causal graph be determined?
No
Yes
Yes
Append the remaining variables to the end of K
Causal matrix M and causal graph G of observational data X
Result

Figure 6: The flowchart of DirectLiNGAM with process knowledge

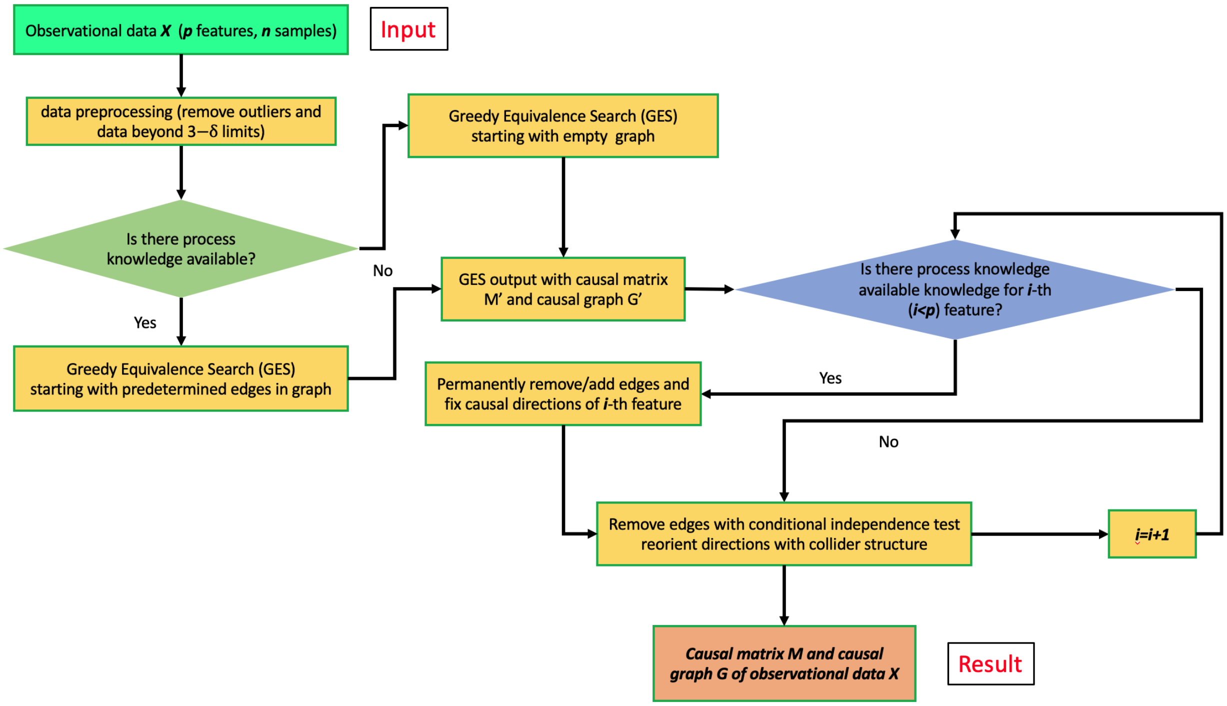

Observational data X (p features, n samples)
Input
data preprocessing (remove outliers and data beyond 3−δ limits)
Is there process knowledge available?
No
Yes
Greedy Equivalence Search (GES) starting with empty graph
Greedy Equivalence Search (GES) starting with predetermined edges in graph
GES output with causal matrix M' and causal graph G'
Is there process knowledge available knowledge for i-th (i<p) feature?
Yes
No
Permanently remove/add edges and fix causal directions of i-th feature
Remove edges with conditional independence test reorient directions with collider structure
i=i+1
Causal matrix M and causal graph G of observational data X
Result

Figure 7: The flowchart of GFCI with process knowledge

matrix and causal graph. It should be pointed out that process knowledge can be combined with various causal discovery algorithms, not just the two algorithms mentioned in this paper. We further demonstrate the details of the proposed method with real industrial process data in the next section.

3. Case Study

In this section, we focus on solving a causal structure learning problem in the fluid catalytic cracker (FCC) unit of Burnaby Refinery (British Columbia, Canada).[28] The FCC is an intermediate unit that processes the initial, heavy oil stream and “cracks” it to produce a wide variety of different products that will be further processed before blending into the final products. Fluid catalytic cracking results in a wide range of intermediate products which can be upgraded to gasoline, diesel, heavy fuel oil and liquified petroleum gas fractions. Figure 8 shows the FCC unit diagram with control loop.

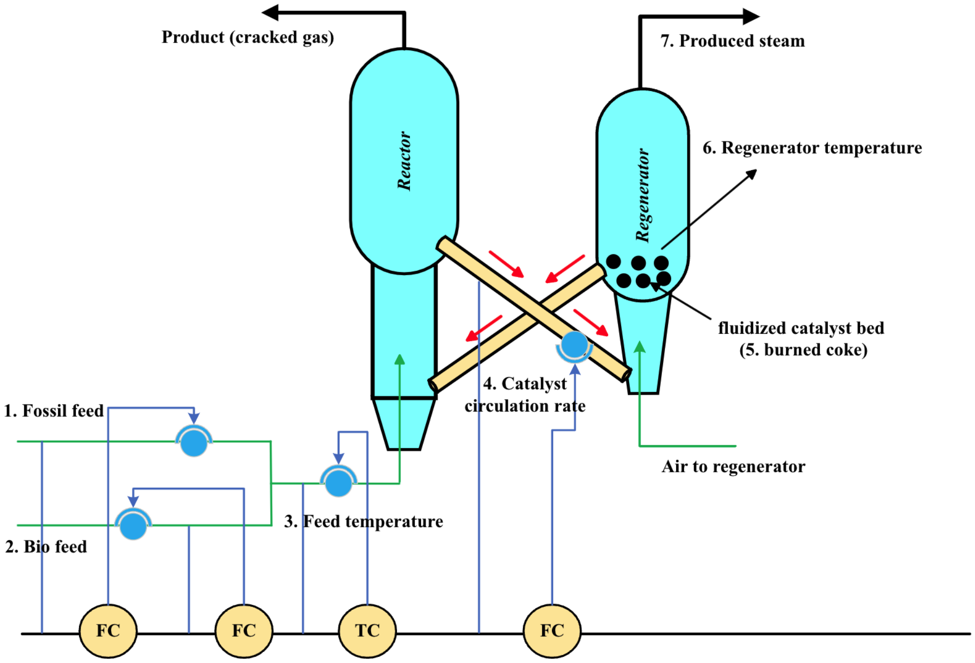


Figure 8: FCC unit with control loop

3.1 Data Preprocessing and Ground-truth Graph

In this study, one year's commercial process data (hourly data, 8955 samples) from the FCC unit is used for analysis. Basic filtering of the data includes removing data that is beyond a certain threshold and the outliers. Outliers, like the lower feed rate, might be caused due to turnaround or occasionally unit upset and therefore the data is not stable/representative. Thus, 3-σ limits are used to set the upper and lower threshold limits. It is a statistical calculation that refers to data x within three standard deviations (3σ) from the mean μ. The values within three standard deviations account for about 99.73% data ($P\left(\mu - 3\sigma \leq x \leq \mu + 3\sigma\right) \approx 99.73\%$) and usually are considered as normal operating process data. In this work, we first remove outliers and then remove samples that are beyond the 3-σ limits threshold.

Several variables that may impact the amount of burned coke are selected based on process knowledge and the FCC heat balance. These variables are catalyst circulation, fossil feed, bio feed, feed temperature, regenerator temperature, burned coke and produced steam.

The correlation ρ, data distribution and normalized samples are given in Figure 9.

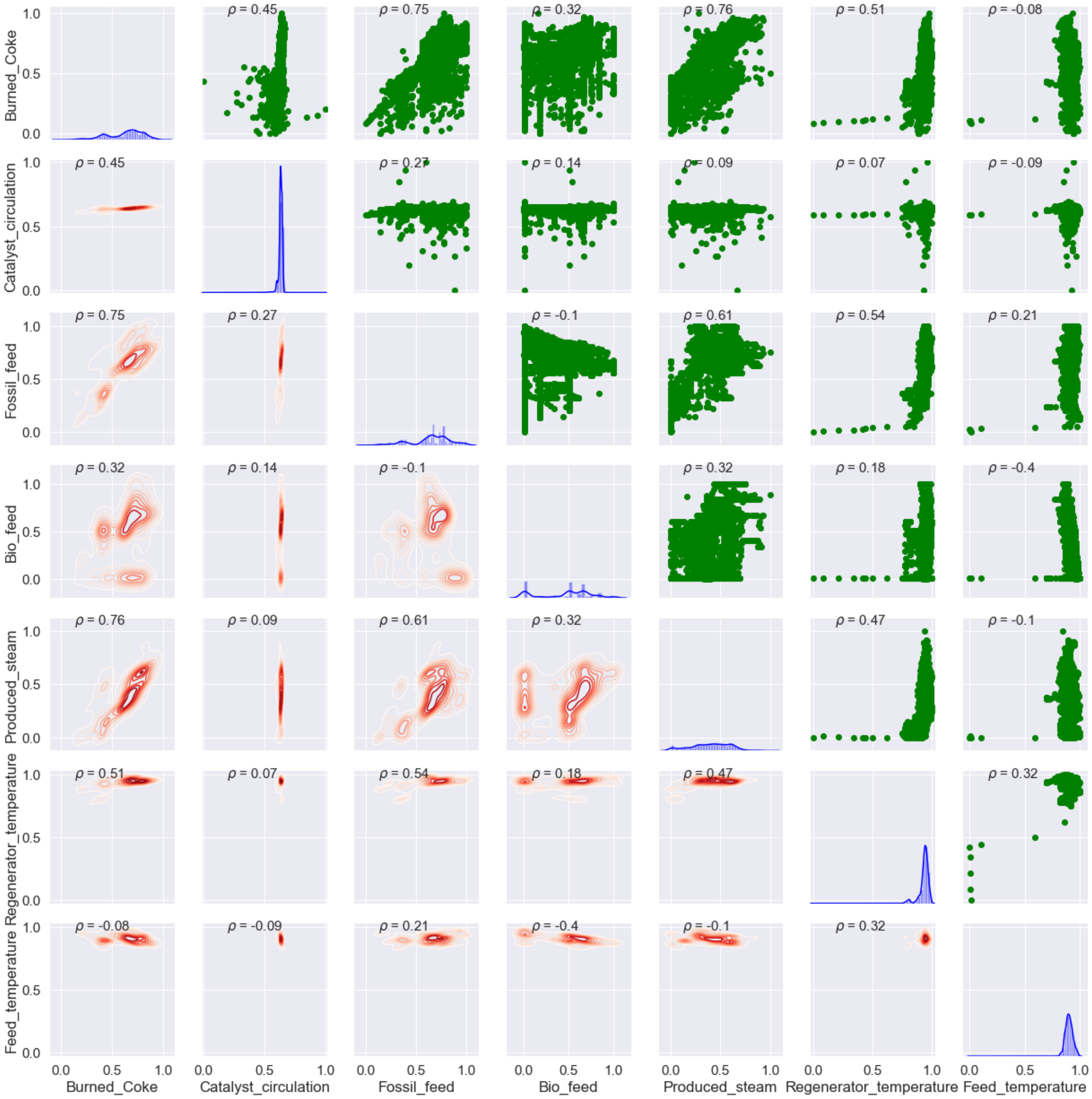
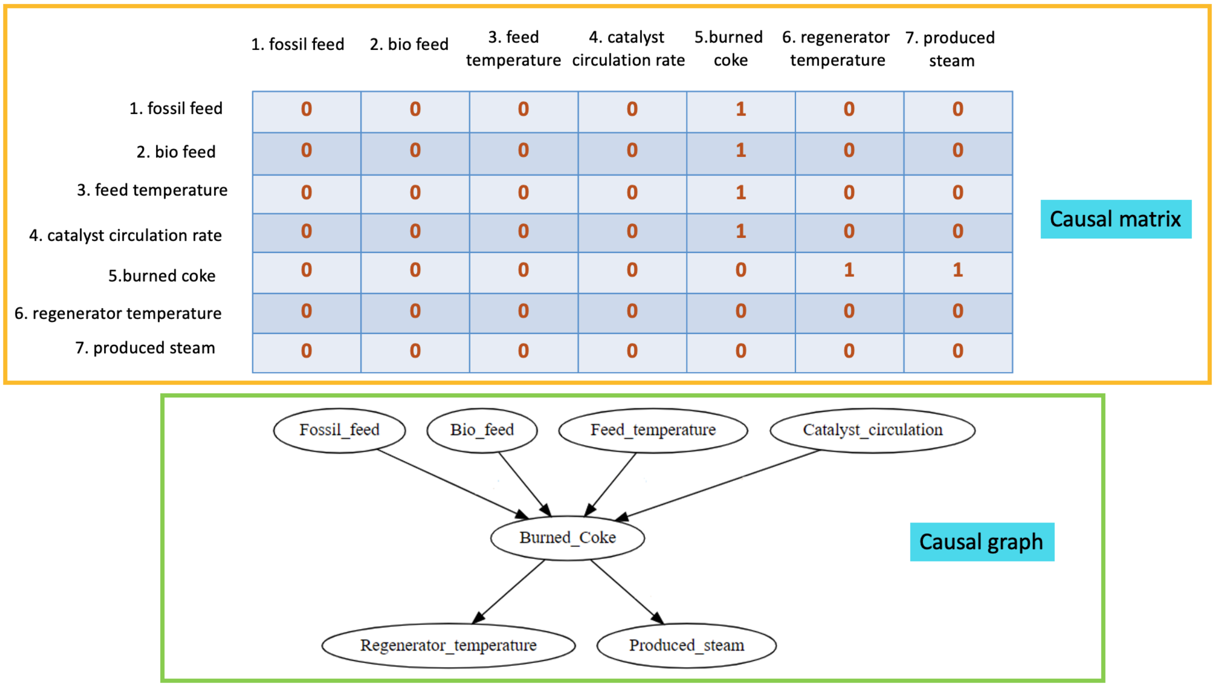


Figure 9: The correlation graph of FCC process variables

The goal of this work is to learn the causal graph from observable historical data. For the FCC process, we use the heat balance to understand the coke yield and get the ground-truth graph for the above-mentioned variables. Increasing the fossil feed and bio feed quantity will increase the burned coke. Increasing catalyst circulation will make more coke deposited in the unit. On the other hand, reducing the feed temperature will increase the coke yield since more coke is needed to heat the feedstocks to reach the pre-set temperature. Burning more coke will increase the regenerator temperature and produce more steam. The causal relationships described above are shown in Figure 10.



	1. fossil feed	2. bio feed	3. feed temperature	4. catalyst circulation rate	5.burned coke	6. regenerator temperature	7. produced steam
1. fossil feed	0	0	0	0	1	0	0
2. bio feed	0	0	0	0	1	0	0
3. feed temperature	0	0	0	0	1	0	0
4. catalyst circulation rate	0	0	0	0	1	0	0
5.burned coke	0	0	0	0	0	1	1
6. regenerator temperature	0	0	0	0	0	0	0
7. produced steam	0	0	0	0	0	0	0

Figure 10: The ground-truth causal matrix and causal graph of FCC process

3.3 Evaluation

We evaluate the performance of the proposed methods with the following metrics: precision, recall, and *g*-score. In this work, precision is defined as the proportion of correct or semi-correct (the edge exists in the ground-truth graph and its orientation does not contradict the true orientation) edges over all edges reported by algorithms; it is mainly used to measure the degree that edges are added by mistake. The recall is the proportion of edges in the ground truth graph that are correct or semi-correct; it is mainly used to measure the degree that edges have not been discovered. *g*-score is defined as the proportion of maximum net corrected edges over all edges reported by algorithms. The definition of these metrics is given as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{6}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{7}$$

$$g\text{-score} = \frac{\max\left(0, (TP - FP)\right)}{TP + FN} \tag{8}$$

where TP is short for true positive, which means the estimated directed edge is in the ground-truth graph; FP is short for false positive, which means the the estimated directed edge is not in the ground-truth graph. FN is short for false negative, which means the estimated directed edge is not in the estimated causal graph but in the ground-truth graph.

3.4 Results

The causal graphs estimated by DirectLiNGAM and GFCI algorithms are discussed in this section. Three cases, causal discovery without process knowledge, with partial process knowledge and with full process knowledge, are considered. In an estimated causal graph, the black directed edges are the estimated edges by algorithms; the circle on the edge means that the direction is not sure; the directed edges that are not in the ground-truth graph are labelled with the cross mark; the red and blue directed edges mean the edges are not in the estimated causal graph but in the ground-truth graph. The difference between red and blue edges is that the red edges contradict the true orientation, while the blue edges are missing edges. For DirectliNGAM, the values on the directed edges are estimated connection strengths. The difference in the output graph lies in that the assumptions of different approaches can change the results of the corresponding causal discovery algorithm.

3.4.1 Causal Discovery without Process Knowledge

The causal discovery results without process knowledge are presented in Figures 11-12 . For the case without process knowledge, DirectLiNGAM finds 3 out of 6 directed edges correctly, 2 directed edges are opposite and 1 directed edge is missing, the estimated causal graph contains 11 directed edges that are not in the true causal graph; GFCI finds 3 out of 6 directed edges correctly, 1 directed edge is opposite and 2 directed edges are missing, the

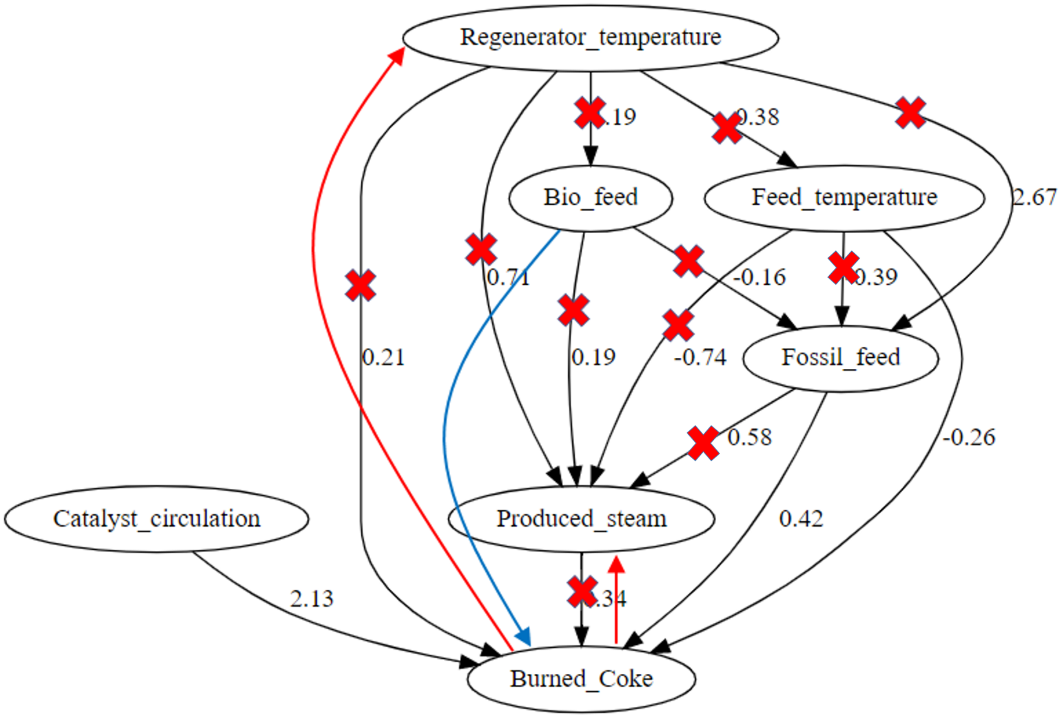


Figure 11: The DirectLiNGAM result without process knowledge

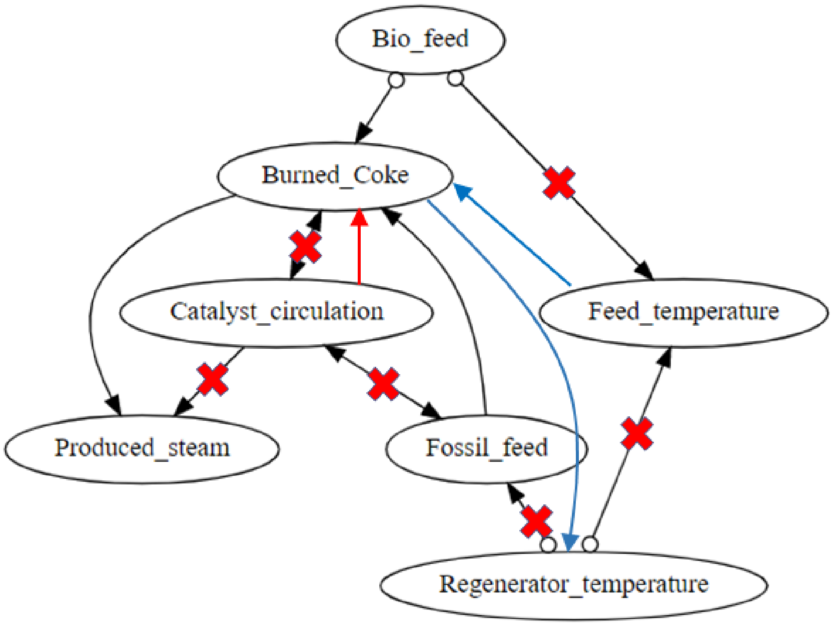


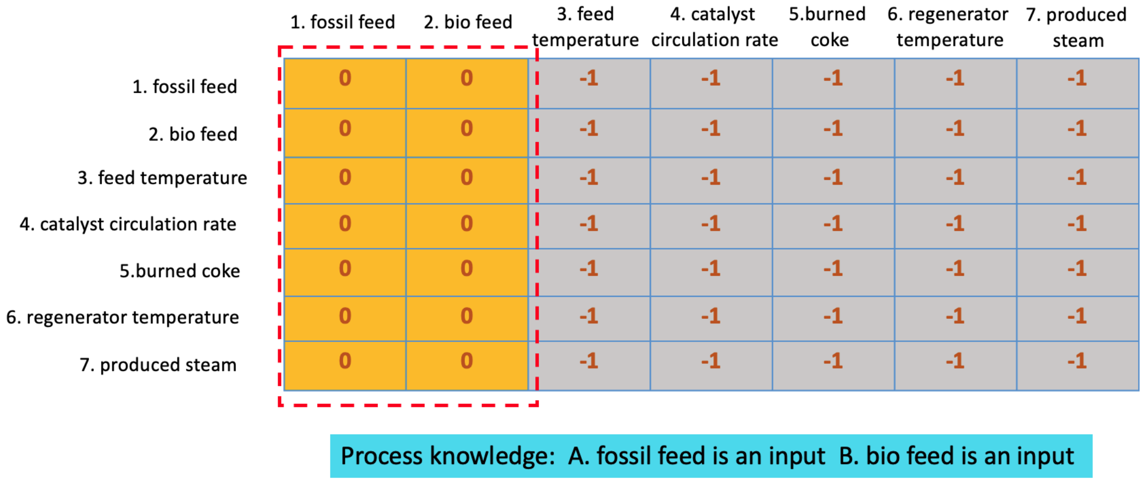
Figure 12: The GFCI result without process knowledge

estimated causal graph contains 6 directed edges that are not in the true causal graph. It is clear that both algorithms successfully discover several causal relations, but make lots of mistakes at the same time. The precision, recall and g-score of DirectLiNGAM (0.21, 0.5, 0, respectively) and GFCI (0.33, 0.5, 0, respectively) are very poor.

3.4.2 Process Knowledge of FCC process

As mentioned before, we can achieve better causal discovery results when combining causal discovery algorithms with physical modelling information. The principle of adding FCC process information to the causal discovery algorithms can be given as follows. To demonstrate the effectiveness of process knowledge, we choose to use two types of process knowledge, partial process knowledge and full process knowledge, to discover causal graphs of variables.

Figure 13 shows an example of partial process knowledge of FCC process. As inputs to the process, fossil feed and bio feed are determined by chemical engineers, and no other variables can affect these two variables, so we do not need to discover their parent nodes. Corresponding to the causal matrix, we can preset the value of the first column and the second column to be equal to 0.

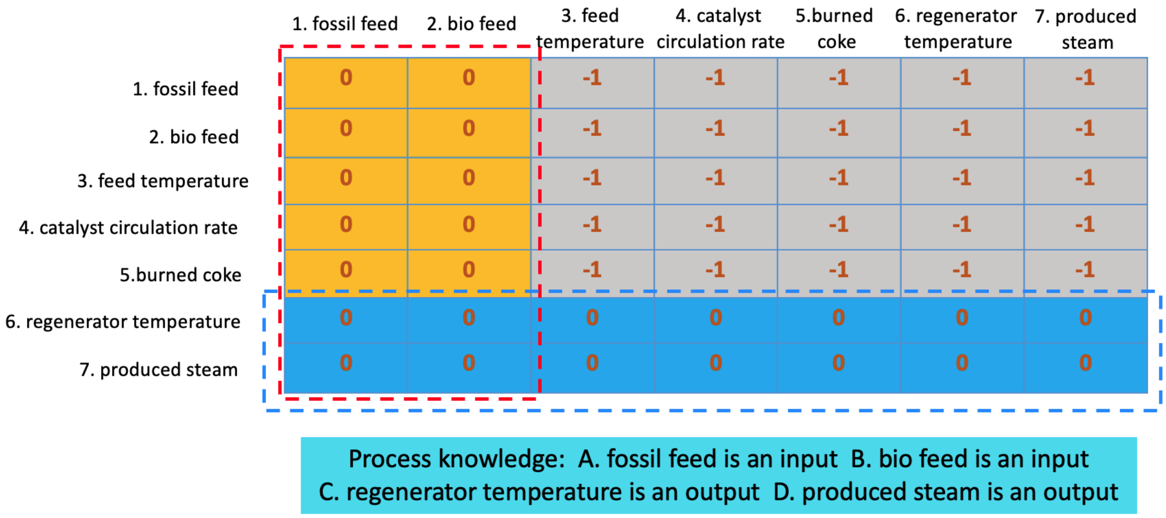


	1. fossil feed	2. bio feed	3. feed temperature	4. catalyst circulation rate	5.burned coke	6. regenerator temperature	7. produced steam
1. fossil feed	0	0	-1	-1	-1	-1	-1
2. bio feed	0	0	-1	-1	-1	-1	-1
3. feed temperature	0	0	-1	-1	-1	-1	-1
4. catalyst circulation rate	0	0	-1	-1	-1	-1	-1
5.burned coke	0	0	-1	-1	-1	-1	-1
6. regenerator temperature	0	0	-1	-1	-1	-1	-1
7. produced steam	0	0	-1	-1	-1	-1	-1

Process knowledge: A. fossil feed is an input B. bio feed is an input

Figure 13: Partial process knowledge of FCC process

In addition to the inputs mentioned above as process knowledge, we can also use outputs as process knowledge. Figure 14 shows an example of full process knowledge of the FCC process. As the outputs of the FCC process, produced steam and regenerator temperature do not produce any effect on any other variables, we do not need to discover their children nodes. Corresponding to the causal matrix, we can pre-set the value of the sixth row and the seventh row to be equal to 0.



	1. fossil feed	2. bio feed	3. feed temperature	4. catalyst circulation rate	5.burned coke	6. regenerator temperature	7. produced steam
1. fossil feed	0	0	-1	-1	-1	-1	-1
2. bio feed	0	0	-1	-1	-1	-1	-1
3. feed temperature	0	0	-1	-1	-1	-1	-1
4. catalyst circulation rate	0	0	-1	-1	-1	-1	-1
5.burned coke	0	0	-1	-1	-1	-1	-1
6. regenerator temperature	0	0	0	0	0	0	0
7. produced steam	0	0	0	0	0	0	0

Process knowledge: A. fossil feed is an input B. bio feed is an input
C. regenerator temperature is an output D. produced steam is an output

Figure 14: Full process knowledge of FCC process

It should be pointed out that, in addition to adding variable order (inputs/outputs) as process knowledge to improve the accuracy of causal discovery, we can further add forbidden and required edges to narrow down the search space according to the definition of process knowledge proposed in Equation 5.

3.4.3 Causal discovery with Process Knowledge

For the causal discovery with process knowledge, we will test causal discovery results with partial process knowledge and full process knowledge. In the case of partial process knowledge as shown in Figure 13, DirectLiNGAM finds 3 out of 6 directed edges correctly, 3 directed edge is missing, the estimated causal graph contains 9 directed edges that are not in the true causal graph; GFCI finds 3 out of 6 directed edges correctly, 1 directed edge is missing, the estimated causal graph contains 3 directed edges that are in the true causal graph. The precision, recall and g-score of DirectLiNGAM (0.25, 0.5, 0, respectively) and GFCI (0.5, 0.6, 0, respectively) get better compared to the case with no process knowledge.

In the case of full process knowledge as shown in Figure 14, DirectLiNGAM finds 5 out of 6 directed edges correctly, 1 directed edge is missing, the estimated causal graph contains 4 directed edges that are not in the true causal graph; GFCI finds 5 out of 6 directed edges correctly, 1 directed edge is missing, the estimated causal graph contains 2 directed edges that are not in the true causal graph. Figures 15-16 shows the causal graph with full process knowledge. The precision, recall and g-score of DirectLiNGAM (0.56, 0.83, 0.17, respectively) and GFCI (0.71, 0.83, 0.5, respectively) have significant improvements.

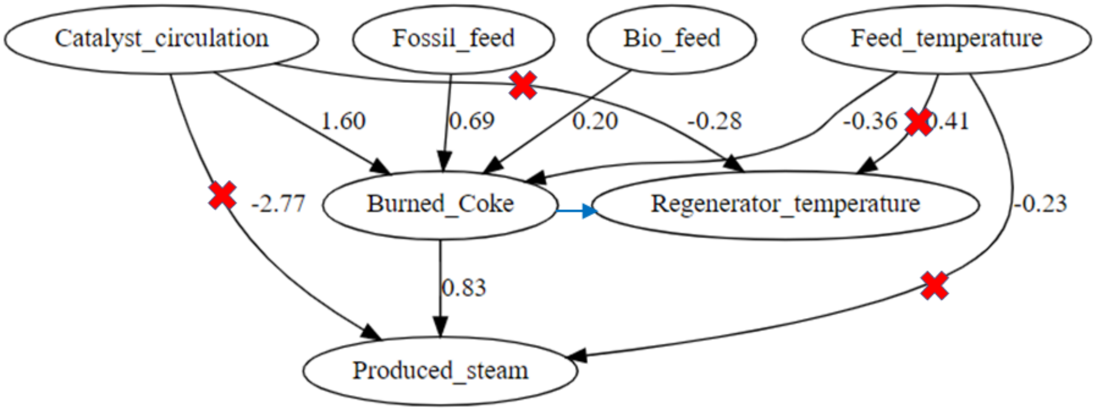


Figure 15: The DirectLiNGAM result with full process knowledge

Table 1 gives the causal discovery evaluation metrics of DirectLiNGAM and GFCI under different process knowledge situations. The number in bold black means the worst performance in all algorithms while the bold red means the best performance.

4. Discussion: Challenges and Opportunities

Discovering causal graphs from complex industrial process data faces many crucial challenges. Here we list 3 common challenges that prevent causal discovery methods from a much wider

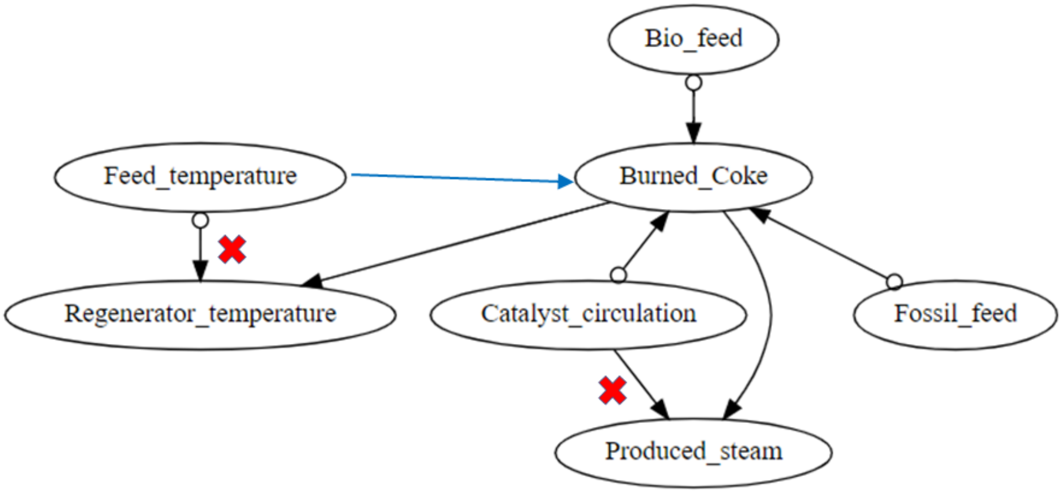


Figure 16: The GFCI result with full process knowledge

Table 1: Causal discovery evaluation metrics of DirectLiNGAM and GFCI

Algorithms	DirectLiNGAM			GFCI		
	TP	*FP*	*FN*	*TP*	*FP*	*FN*
Without knowledge	**3**	**11**	**3**	**3**	**6**	**3**
With partial knowledge	3	9	3	3	3	2
With complete knowledge	**5**	**4**	**1**	**5**	**2**	**1**
	Precision	Recall	g-score	Precision	Recall	g-score
Without knowledge	**0.21**	**0.5**	**0**	**0.33**	**0.5**	**0**
With partial knowledge	0.25	0.5	0	0.5	0.6	0
With complete knowledge	**0.56**	**0.83**	**0.17**	**0.71**	**0.83**	**0.5**

application in industrial processes.

The first crucial challenge is that assumptions should but often are not satisfied. For example, the distribution of disturbance is often not independent of industrial process variables, which violates the assumptions of DirectLiNGAM. In GFCI, the faithfulness assumption may be violated if the positive effect and negative effect happen to exactly balance and there will be no correlation, which is very common in the industrial process with controllers. The violation of assumptions will produce undesirable results.

The second crucial challenge is incomplete data. The incomplete data includes two types, missing samples, and unobserved causal variables. Both types will render spurious causal relations. In the industrial process, missing samples can be caused by imbalanced data, time sub-sampling, etc. For unobserved causal variables, it is difficult to ensure that all potential causal variables are considered since there may be hundreds of potential causal variables in only one industrial process unit.

The third crucial challenge is large and time-varying time delays. As modern industrial processes are often accompanied by the transfer of complex material flow, information flow, and energy flow, these special characteristics will lead to large and time-varying time delays. The existence of a time delay will affect the reliability of the independence test and then lead to incorrect causal conclusions.

We have noticed that the application of causal discovery algorithms has some major impediments and needs to be carefully addressed. We believe that there are two opportunities that may substantially advance state-of-the-art algorithms. One is to develop novel algorithms that require fewer assumptions. For example, reference 29 can partially handle the problem of time delay, and reference 30 can deal with the problem of nonlinearity. The other one is to integrate process knowledge. Process knowledge, like process topological information and control loop diagrams, can provide guidance on the existence of partial causal structures, which greatly improve the performance of causal discovery algorithms.

5. Conclusions

Discovering causal graphs from massive industrial process data opens up new ways of improving the interpretability, reliability, and robustness of the model. Although state-of-art causal discovery algorithms are able to mine valuable causal information from data, it also detects lots of spurious causal links. In this work, we use process knowledge as a regularizer to guide the discovery of causal graphs. The commercial-scale FCC unit has shown that the performance of causal discovery is significantly improved with the integration of process knowledge. As a way forward, we further discuss the challenges and opportunities for methodological research when applying causal discovery algorithms to extract causal relationships in the context of complex industrial processes.

Acknowledgement

The authors thank Mitacs and Parkland Corporation for their financial support. Yankai Cao acknowledges funding from the discovery program of the Natural Science and Engineering Research Council of Canada under grant RGPIN-2019-05499.

References

(1) Yin, S.; Ding, S. X.; Xie, X.; Luo, H. A Review on Basic Data-Driven Approaches for Industrial Process Monitoring. *IEEE Transactions on Industrial Electronics* **2014**, *61*, 6418–6428.

(2) Mehta, B.; Reddy, Y. *Industrial Process Automation Systems Design and Implementation*; Butterworth-Heinemann: Oxford, 2015.

(3) He, Y.-L.; Zhao, Y.; Zhu, Q.-X.; Xu, Y. Online Distributed Process Monitoring and Alarm Analysis Using Novel Canonical Variate Analysis with Multicorrelation Blocks

and Enhanced Contribution Plot. *Industrial & Engineering Chemistry Research* **2020**, *59*, 20045–20057.

(4) Gao, H.; Struble, T. J.; Coley, C. W.; Wang, Y.; Green, W. H.; Jensen, K. F. Using Machine Learning To Predict Suitable Conditions for Organic Reactions. *ACS Central Science* **2018**, *4*, 1465–1476.

(5) Fan, J.; Qin, S. J.; Wang, Y. Online monitoring of nonlinear multivariate industrial processes using filtering KICA–PCA. *Control Engineering Practice* **2014**, *22*, 205–216.

(6) Zhu, Q.; Joe Qin, S.; Dong, Y. Dynamic latent variable regression for inferential sensor modeling and monitoring. *Computers & Chemical Engineering* **2020**, *137*, 106809.

(7) Zhang, X.; Cui, P.; Xu, R.; Zhou, L.; He, Y.; Shen, Z. Deep Stable Learning for Out-of-Distribution Generalization. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021; pp 5372–5382.

(8) Pearl, J. *Causality*, 2nd ed.; Cambridge University Press:Cambridge, 2009.

(9) Hu, H.; Li, Z.; Vetta, A. R. Randomized Experimental Design for Causal Graph Discovery. *Advances in Neural Information Processing Systems* **2014**, *27*, 1–10.

(10) Matiasz, N. J.; Wood, J.; Wang, W.; Silva, A. J.; Hsu, W. Computer-Aided Experiment Planning toward Causal Discovery in Neuroscience. *Frontiers in Neuroinformatics* **2017**, *11*, 12.

(11) Spirtes, P.; Glymour, C. An Algorithm for Fast Recovery of Sparse Causal Graphs. *Social Science Computer Review* **1991**, *9*, 62–72.

(12) Verma, T.; Pearl, J. Equivalence and Synthesis of Causal Models. Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence. USA, 1990; p 255–270.

(13) Shimizu, S.; Hoyer, P. O.; Hyvarinen, A.; Kerminen, A. A Linear Non-Gaussian Acyclic Model for Causal Discovery. *Journal of Machine Learning Research* **2006**, *7*, 2003–2030.

(14) Shimizu, S.; Inazumi, T.; Sogawa, Y.; Hyvärinen, A.; Kawahara, Y.; Washio, T.; Hoyer, P. O.; Bollen, K. DirectLiNGAM: A Direct Method for Learning a Linear Non-Gaussian Structural Equation Model. *J. Mach. Learn. Res.* **2011**, *12*, 1225–1248.

(15) Spirtes, P.; Meek, C.; Richardson, T. Causal Inference in the Presence of Latent Variables and Selection Bias. Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence. San Francisco, CA, USA, 1995; p 499–506.

(16) Glymour, C.; Zhang, K.; Spirtes, P. Review of Causal Discovery Methods Based on Graphical Models. *Frontiers in Genetics* **2019**, *10*, 524.

(17) Ogarrio, J. M.; Spirtes, P.; Ramsey, J. A Hybrid Causal Search Algorithm for Latent Variable Models. Proceedings of the Eighth International Conference on Probabilistic Graphical Models. Lugano, Switzerland, 2016; pp 368–379.

(18) Cai, R.; Zhang, Z.; Hao, Z. SADA: A General Framework to Support Robust Causation Discovery. Proceedings of the 30th International Conference on Machine Learning. Atlanta, Georgia, USA, 2013; pp 208–216.

(19) Cooper, G. F.; Glymour, C. *Computation, Causation, and Discovery*; AAAI Press: California, 1999.

(20) Ramsey, J.; Glymour, M.; Sanchez-Romero, R.; Glymour, C. A million variables and more: the Fast Greedy Equivalence Search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *International Journal of Data Science and Analytics* **2017**, *3*, 121–129.

(21) Hoyer, P. O.; Janzing, D.; Mooij, J. M.; Peters, J.; Schölkopf, B. Nonlinear causal dis-

covery with additive noise models. *Advances in Neural Information Processing Systems* **2008**, *689*, 689–696.

(22) Zhang, K.; Hyvärinen, A. On the Identifiability of the Post-Nonlinear Causal Model. Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence. Arlington, Virginia, USA, 2009; p 647–655.

(23) Runge, J.; Bathiany, S.; Bollt, E.; Camps-Valls, G.; Zscheischler, J. Inferring causation from time series in Earth system sciences. *Nature Communications* **2019**, *10*.

(24) Runge, J. Causal network reconstruction from time series: From theoretical assumptions to practical estimation. *Chaos: An Interdisciplinary Journal of Nonlinear Science* **2018**, *28*, 075310.

(25) Schreiber, T. Measuring Information Transfer. *Phys. Rev. Lett.* **2000**, *85*, 461–464.

(26) Granger, C. W. J. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica* **1969**, *37*, 424–438.

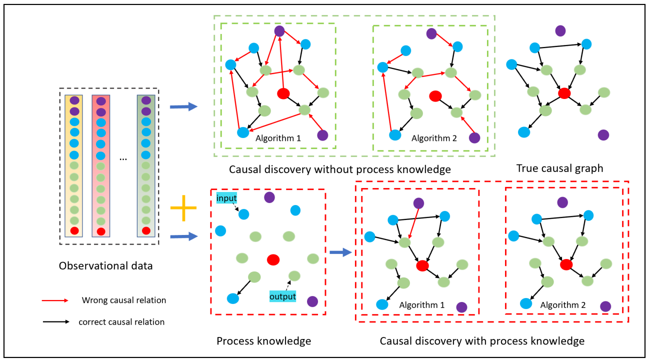
(27) Cao, L.; Yu, F.; Yang, F.; Cao, Y.; Gopaluni, R. B. Data-driven dynamic inferential sensors based on causality analysis. *Control Engineering Practice* **2020**, *104*, 104626.

(28) Su, J.; Cao, L.; Lee, G.; Tyler, J.; Ringsred, A.; Rensing, M.; van Dyk, S.; O'Connor, D.; Pinchuk, R.; Saddler, J. 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.

(29) Luo, Y.; Gopaluni, B.; Xu, Y.; Cao, L.; Zhu, Q.-X. A Novel Approach to Alarm Causality Analysis Using Active Dynamic Transfer Entropy. *Industrial & Engineering Chemistry Research* **2020**, *59*, 8661–8673.

(30) Sanchez-Romero, R.; Ramsey, J. D.; Zhang, K.; Glymour, M. R. K.; Huang, B.; Glymour, C. Estimating feedforward and feedback effective connections from fMRI time series: Assessments of statistical methods. *Network Neuroscience* **2019**, *3*, 274–306.

TOC Graphic


Observational data
...
Causal discovery without process knowledge
Algorithm 1
Algorithm 2
True causal graph
input
output
Process knowledge
Causal discovery with process knowledge
Algorithm 1
Algorithm 2
Wrong causal relation
correct causal relation