Experimental time series data with and without anomalies from a continuous distillation mini-plant for development of machine learning anomaly detection methods

Aparna Muraleedharan,[†] Alvaro Ferre,[†] Justus Arweiler,[‡] Indra Jungjohann,[‡] Fabian Jirasek,[‡] Hans Hasse,[‡] and Jakob Burger*,[†]

†*Technical University of Munich, Campus Straubing for Biotechnology and Sustainability, Laboratory for Chemical Process Engineering, Uferstraße 53, 94315 Straubing, Germany*

‡*Laboratory of Engineering Thermodynamics (LTD), RPTU Kaiserslautern, Erwin-Schrödinger-Straße 44, 67663 Kaiserslautern, Germany*

E-mail: burger@tum.de

Abstract

Reliable detection of process anomalies remains a challenge in industrial chemical plants. The ability of machine learning (ML) to recognize patterns has triggered numerous research efforts to apply ML to anomaly detection (AD). Typically, simulation-based benchmarks like the Tennessee Eastman process are widely used in the development and training of AD methods. Real-world process data, which are crucial for meaningful research advancements, are lacking due to proprietary limitations in the industry. To overcome this issue, we present an openly accessible dataset of time series generated from an industry-like continuous distillation mini-plant under steady-state conditions. The data generated have different complexities: water runs, a heteroazeotropic

separation of n-butanol and water, and a reactive process to produce a fuel additive. Chemical systems, plant setup, and anomalies (encountered or induced manually) are described alongside the sample data. The complete dataset, including sensor and actuator data, annotations to mark anomalies, and other metadata, is available in open access in an online repository. It serves as training and testing data for ML-based AD and other data-driven applications.

1 Introduction

In the chemical industry, continuous processes are essential to optimize efficiency and produce high-quality products on large scales[1,2]. The stability of critical process parameters during plant operation is maintained with the help of control strategies[3], which counteract disturbances, such as fluctuations in input feeds, external conditions, or fluctuations inside the process. In spite of having robust control strategies, anomalies (deviation from normal behavior) can still occur[4]. Anomalies can indicate potential malfunctions, safety hazards, or inefficiencies that, if left unaddressed, may result in catastrophic failures or substantial economic losses. A survey and analysis by Kister[5] spanning five decades shows that the anomalies in distillation towers are repetitive and have continued to increase. Plugging and coking dominate the list of anomalies, followed by reboiler level and internal damages. With the industry's growing dependence on automated systems and advanced technologies, effective and early detection of anomalies through reliable techniques is crucial for maintaining safety, optimizing efficiency, and meeting regulatory requirements[6].

Even though anomalies are rare, machine learning (ML) methods have proven effective in detecting anomalies from large datasets and reached high accuracies in a wide set of application domains[7–10]. Also in the chemical process industry, which is the domain of the present work, ML-based AD has seen several research developments in the past decade, with new methods and techniques being proposed, such as deep learning or reinforcement learning[11–14]. More recent innovations in ML, such as Generative Adversarial Networks

(GANs) and Long Short-Term Memory (LSTM) networks, have shown promise in enhancing AD capabilities by effectively modeling complex data patterns, which are inherent in many chemical processes[15]. Although ML methods for AD are being developed at a high pace, the available data sources for training and testing have not kept up. As a rule, process data from the chemical industry is not openly available due to proprietary limitations. This led Downs and Vogel[16] in 1993 to create an imaginary industrial process, the Tennessee Eastman Process (TEP), and publish a dynamic process model for it.

For three decades, the TEP has continued to serve a broad audience as a robust educational and research tool, applicable to plant-wide control, multi-variable control, optimization, as well as AD[17]. The TEP is a chemical process with five unit operations: a reactor, a product condenser, a vapor-liquid separator, a recycle compressor, and a product stripper. From its simulation, time-series datasets can be produced and are available for download[17,18]. Although the TEP datasets provide a good benchmark for method development, it is crucial to gauge the applicability of the methods in a real-world setting, where the performance of methods trained to perform well with TEP data might be different. For example, we have performed a comprehensive evaluation of deep learning AD methods in the literature using the TEP dataset and concluded that generative model types perform the best[19]. However, when we applied the same methods to time-series data from an experimental scenario of the present work, hybrid model types perform the best[20]. Such results emphasize the importance of utilizing real-world data in the development of data-driven AD methods.

Publicly available datasets of real-world time series, such as the MIMII, IPAD, or VISION[21–23], focus on sound, video, or images from industrial machines such as valves and pumps. They contribute to industrial equipment maintenance, but are not typical time series data from chemical plants. Although several studies in the literature[24–27] actually use real-world data from chemical plants for AD, process optimization, and control applications, none of them have published the data. It is striking that there is still a lack of publicly available chemical process data[28]. Industrial companies often implement measures to anonymize the data

(anonymization involves removing or masking identifiable information to protect privacy and confidentiality) when sharing it with other parties. Ideal training datasets are, however, not anonymous[29] and contain additional metadata or context. Applied to chemical process data, this additional information can include, e.g., the process flow diagram, the piping and instrumentation diagram, property data of the chemicals used, or annotations to anomalies by the operators. Including this information ensures that models can detect realistic issues and that the results are interpretable and reproducible.

In the present work, we supply real-world process data from a continuous distillation plant in our lab. We put together historical data from an older project[30] with novel operating points that were run simply for the sake of data generation. Although operated in an academic environment, the plant is operated industry-like, including a process control system keeping it in a steady-state for hours or days (depending on the operating point). Distillation is one of the most common unit separations used in the chemical industry, making the data valuable for use in the development of data-driven machine learning methods. A particular focus of the present work is to provide data for training of AD methods; thus, we provide data with and without anomalies, which are labeled respectively. Together with a novel dataset for batch distillation, which we will report in a separate paper (currently in preparation phase), we created a solid database for testing and developing machine learning methods, see, e.g.,[20,31] for first applications in collaboration with project partners. By providing the dataset in open access through this publication, we aim to help the entire community to develop methods fit for use in later industrial operations.

2 Methodology

2.1 Continuous distillation mini-plant

The mini-plant used in the present work was originally erected for the production of poly (oxymethylene) dimethyl ethers ($OME_{3\text{-}5}$) by Ferre et al.[30], following a process developed

by Schmitz et al.[32]. It consisted of a tubular reactor, a membrane separator (DBI Gas and Umwelttechnik), and two distillation columns (Iludest GmbH). These two distillation columns are the object of the present work. Within the present work, we prepare and publish the data originally recorded by Ferre et al.[33]. Further, we ran additional experiments with other chemical systems. For this purpose, we augmented the setup with a decanter and additional control strategies for heteroazeotropic distillation, which will be described in detail in the following section. Figure 1 shows a P&I diagram of the two columns and the decanter. Detailed nomenclature is explained in the Supporting Information (Section S1). The first column, C1 (diameter: 70 mm), column on the left side of Figure 1, made of stainless steel, can operate up to a pressure of 2 bar and consists of six sections. The second column, C2 (diameter: 50 mm), made of glass, operates up to a pressure of 1 bar and consists of six sections. The cylindrical decanter (manufactured in-house) made of glass has a total height of 37 cm and a diameter of 13.36 cm. Each section in C1 and C2 is insulated with electric heating jackets (SAF) to minimize heat losses. Both columns are equipped with a partial reboiler and a nearly total condenser, respectively. Temperature sensors (e.g., T101) are positioned in key locations: reboiler, bottom of each section of the distillation column, and top of the column. Sampling valves (e.g., HV208) are located between the sections, facilitating the withdrawal of liquid samples during the experiment.

The head pressure sensors PIC101 and PIC201, record the top pressure in the columns. A small inert gas flow (nitrogen) is used to control the head pressure. The condensers E102 and E202 use water from the grid, and are located at the top of the columns. The distillate from C1 undergoes a second cooling process in exchanger E104, operated with a mixture of ethylene glycol and water, which is back-cooled in a cryostat (TH102). Uncondensed gases are collected in the cold traps filled with liquid nitrogen (E106, E206). The reboilers (E101, E202) are shell-and-tube heat exchangers with thermostats (TH101, TH201) heating the oil, that is fed into the heat exchanger's shell, and the liquid product evaporates in the tubes. Differential pressure measurement indicators (PDIC101, PDIC201) measure the liquid levels

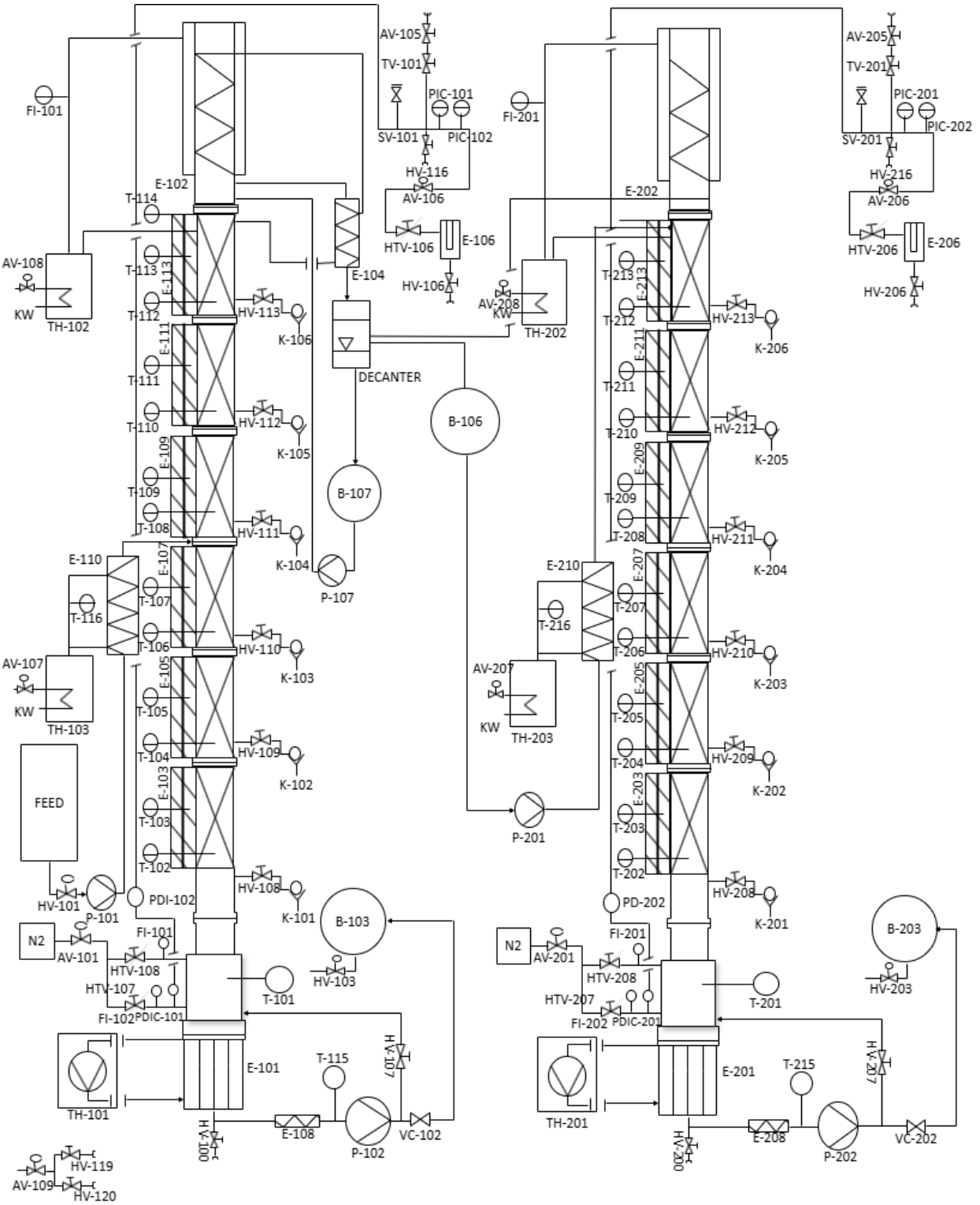


Figure 1: P&ID of the continuous distillation plant for the heteroazeotropic distillation of n-butanol and water.

in the reboiler. Both columns are equipped with 3.84 m of structured packing (Sulzer CY). The feed tank is connected to a membrane pump, which pumps the feed through a heat exchanger (E110), which is heated using oil from a thermostat (TH103). The bottom products of the columns are cooled in heat exchangers (E108, E208) using grid water to around room temperature and drawn using gear pumps (P102 and P202). If the feed mixture is heteroazeotropic, a decanter for liquid-liquid phase split is employed as a common reflux drum for both columns. The upper (organic) phase is fed as reflux into C2 with a gear pump (P101) and the lower (aqueous) phase as reflux to C1 with another gear pump (P107). Flow indicators (Coriflow) measure the flow rate of the feed (F104), the distillates from both columns (F102, F103), and the reflux stream to the first column (F101). Weighing scales (A101, A102, A103, A104, A106, A107, and A203) are used to measure the mass of buffer vessels, thereby calculating the mass flow rates of several other streams.

The software LabVIEW v.19 is used to control the operation and record the data of the mini-plant. Overall, the mini-plant has two pressure controllers, three flow controllers, four level controllers, and two temperature controllers. All controllers are PID controllers, except the level controllers in the reboilers, which are P-only controllers, which were sufficient for quick and reliable response. The control loops and strategies are described in later sections as they are scenario-specific. The mini-plant setup generates multivariate time series data from 38 different sensors. In addition to the sensor data, concentration analysis data of liquid samples from each experiment is generated via scenario-specific offline analytics, also described in later sections. Table S1 (Supporting Information) shows the range of operating parameters of the plant. Detailed information regarding the sensors, the associated P&ID labels, the related suppliers, and the precision of these instruments is provided in Table S2 (Supporting Information).

2.2 Scenarios for generating experimental data

Experiments in three scenarios (i.e., three different chemical systems) were conducted to generate steady-state time series data with varying complexity.

- A: Water runs with only the first column
- B: Separation of the heteroazeotropic mixture of n-butanol and water using both columns and a decanter
- C: Separation of OME_{3-5} from a reactive mixture of water, formaldehyde, methanol, and OME in only the first column

The chemical systems of the three scenarios and respective operating conditions were chosen considering factors like process safety, the utility of data, and their relevance to actual industrial applications. Experiments in Scenario A facilitated the remote operation of the plant and allowed for easier overnight functionality, consequently producing experimental data (62400 data points) over an extended period of 30 days. Scenario B is a typical separation process within the chemical and pharmaceutical sectors, e.g., in biofuel synthesis [34,35]. In Scenario C, chemical reactions occur inside the column, yielding a reactive distillation, which is another interesting application example.

Experiments in Scenario A involved feeding water into C1 at approximately 2.5 m of packing height; the distillate and the bottom products were collected and recycled back into the feed tank. Steady-state was achieved within one hour of plant startup. The bottom product flow rate, the feed flow rate, and the reflux ratio were controlled. The temperature in the oil thermostat (TH101) for the reboiler was regulated to maintain the reboiler level (PDIC101) using a P-only controller. The head pressure (PIC101) was controlled using a PID controller, and the liquid level in buffer tank B102 collecting the distillate was maintained by another PID controller(PDIC103). Further operational details for Scenario A are described in Supporting Information S2.

Scenario B involved a heteroazeotropic distillation of an n-butanol and water mixture. In several runs with feed compositions ranging from 0.06 g/g to 0.20 g/g of n-butanol, the mixture was fed into C1 at approximately 2.5 m of packing height. The bottom product from C1 was rich in water and the condensed distillate was sub-cooled and fed into the decanter, where it split into two liquid phases. The lower (aqueous) phase was fed back into C1 as the reflux, whereas the upper (organic) phase was fed into C2 at the top of the packing. The bottom product from C2 was rich in n-butanol and the condensed distillate was fed into the decanter. Both the columns operated at atmospheric pressure. Steady-state was achieved within 3 hours of plant startup.

The control strategy proposed by Luyben[36] was used here: During steady state, the oil of the reboiler thermostats (TH101, TH201) was set to a high temperature (413.15 K in C1 and 403.15 K in C2). The reboiler liquid levels (PDIC101, PDIC201) and the decanter overflow buffer tank levels (B106, B107) were maintained using level controllers by manipulating the flow through the respective pumps (P102, P202, P107, and P101). 17 liquid samples from C1, C2, the decanter, and the feed tank were collected and analyzed during each experiment. Since the mixture of n-butanol and water is heteroazeotropic, 1-propanol was added to the collected liquid samples to ensure a homogeneous mixture at room temperature. The mass fraction of n-butanol in the samples was measured using gas chromatography (GC) with Tetrahydrofuran (THF) as the internal standard. The gas chromatograph employed was a Thermofischer Trace 1610 (Flame Ionization Detector (FID), detector temperature 523.15 K, injection temperature 503.15 K, carrier gas Helium, split flow of 150 ml/min, Restek Rtx-Wax column). Pure n-butanol and 1-propanol were used for calibration. Each measurement was repeated twice, and the consecutive measurements repeated for each sample had an average standard deviation of less than 0.05%. The GC method was evaluated to have a relative error of 1.8%, calculated by measuring known samples of n-butanol. Karl Fischer titration was used to determine the water mass fractions in samples with less than 0.21 g/g water. Karl Fischer titration were also repeated twice for each sample, and the observed

standard deviation was less than 1%. The uncertainty of Karl Fischer titration was evaluated by measuring samples with a known amount of water, and the relative error was found to be 1.7%. Karl Fischer was preferred over GC when measurements from two different types of analysis were available. Examples of sample concentration profiles for scenario B are provided in the Supporting Information (Section S3).

In Scenario C, the experiments with the reactive OME system as described by Ferre et al. [33], involved feeding a methanolic formaldehyde solution into a tubular reactor filled with a catalyst (Amberlyst 46)[37]. The reactor product is then fed into C1 at approximately 2.5 m of the packing height. The main objective of C1, which is the focus of the present work, was to produce $OME_{3\text{-}5}$ with minimal formaldehyde impurities by separating formaldehyde and other components as the distillate from $OME_{3\text{-}5}$ as the bottom product. Steady-state was achieved within 4 to 5 hours of plant startup. In those experiments, the control strategy used is the same as that employed in the experiments with water in Scenario A: The reboiler oil temperature in thermostat TH101 was manipulated to maintain the reboiler level (PDIC101) using a P-only controller, the head pressure (PIC101) was controlled using a PID controller, and the distillate liquid level (PDIC103) was also controlled using a PID controller. In addition, the bottom product flow rate, feed flow rate and the reflux ratio were set. As analytical methods such as GC, Karl Fischer titration and sodium sulphite titration have been employed here to detect the concentration of different components; more details are provided in Ferre et al. [33].

2.3 Data preparation and quality

The following steps are taken to ensure data quality during the experimental campaigns.

Sensor calibration: All temperature, level, and flow sensors are calibrated using standard calibration protocols specified by the equipment manufacturers (Supporting Information S1). Sensor drift or failure is monitored and fixed to avoid erroneous data. GC methods for analysis are calibrated using pure components, and the accuracy of the measurements

is determined using test samples of known composition. Karl Fischer titration is calibrated using a standard coulometric solution, and accuracy is measured using known samples.

Identification of steady state: After startup (which is not part of the dataset), the process is in a steady state once the process variables, such as temperature, pressure, and concentrations, are constant over a period of time (30 minutes). Steady-state was confirmed from time-series quantitatively using statistical thresholds such as moving average plots.

Comprehensive variable logging: The data collected includes both manipulated variables and system responses. The raw process data includes column temperatures, mass flow rates, thermostat temperatures, column head pressure, pump power, pressure differential indicators, actuator commands, and analytical data derived from gas chromatography and Karl Fischer titration.

Data processing: Sensor data is reported every 30 seconds. For all scenarios, startup and shutdown phases were trimmed to retain only the steady-state data. For scenarios A & C, data is reported at the exact timestamp at which they were recorded. For scenario B with two distillation columns, technical reasons necessitated the use of two separate data acquisition systems that could not be synced during the recording. To provide a consistent dataset for training AD methods, we rounded the timestamps to the nearest 30s. In some early runs, one column produced readings every 20 seconds; in those cases, the values within the same 30-second window (e.g., between 10:20:00 and 10:20:30) were averaged and stored at the beginning of the window (e.g., 10:20:00).

Repeatability and metadata documentation: Multiple experimental runs are conducted at each operating condition. A detailed log is maintained documenting the column configuration, the time required for steady state, feed composition, column startup, shutdown, any deviations or anomalies, and the steps involved in fixing the anomaly. This metadata from each experiment is also published along with the data.

Identification of anomalies: Not all deviations from signals constant over time can be classified as an anomaly. For example, certain process variables, such as mass flow rates

obtained via weighing scales, may exhibit spikes and outliers that cannot be classified as anomalies, as this is the expected normal behavior. For manually induced anomalies, the operator is aware of the affected sensor and the time of occurrence. Naturally occurring anomalies, on the other hand, were identified through visual inspection of the data, and any unusual outliers or abnormal behavior patterns were labeled as anomalies based on the accepted threshold of the process variables, chosen based on the steady-state behavior expected from the thermodynamics of the system, operating parameters, and the operator's practical experience. The identified anomalies were labeled in the final dataset in separate columns. Detailed description of the anomaly labels is provided in the Supporting Information (Section S5). Anomaly details specific to each scenario are also provided in the metadata files, including the start and end times of the anomaly, the affected sensors, and the cause of the anomaly.

Accurate data labels: Steady-state windows and labels of normal process data vs. anomalies are labeled differently for naturally occurring anomalies vs manually induced anomalies. For naturally occurring anomalies, we have binary labels: 0 for normal data points and 1 for anomalous data points. For manually induced anomalies, there is a more detailed integer label scheme: 0 for normal data points, 1 for when an anomaly is induced but not yet visible in the data, 2 for when an anomaly is visible in the data, and 3 for the recovery phase when the cause of the anomaly is removed but the effects are still visible in the data. The labeled experimental data are then split into data from training runs (without any anomalies) and data from test runs (with anomalies). Further details of the headers in the data files are provided in Supporting Information (Section S5).

3 Resulting datasets

3.1 Data repository

Experiments involving water generated time-series data over a 30-day period. Each experiment with n-butanol and water typically lasted around 10 hours, with some extended overnight runs. Similarly, experiments using OME, as reported by Ferre et al.[33], also averaged 10 hours, including a few overnight sessions. The process dataset is publicly available[1]. An example of the folder layout is shown in Figure 2. Each scenario (e.g., Scenario A) has operating points that are grouped based on key operating parameters such as pressure, feed flow rate, and feed composition (e.g., operating_point_001). Within each operating point, data is provided for both anomalous (e.g., test_anormal_experiment_001.csv) and non-anomalous experiments (e.g., train_normal_experiment_001.csv). Every scenario includes a sensor description file (Features_Overview.csv), and for each operating point, metadata for individual experiments is provided with information on operating pressure, feed concentration, feed flow rate, and reflux ratio (e.g., test_anormal_experiment_001_-metadata.yaml). Table 1 shows the statistics regarding the number of experiments, duration of each experiment with the number of anomaly-free and anomaly runs. The steady-state time series data and concentration profiles can be visualized using a graphical and easy-to-use web interface[2] built on a Streamlit-based application[38]. Additional details on this web interface are provided in the Supporting Information S6.

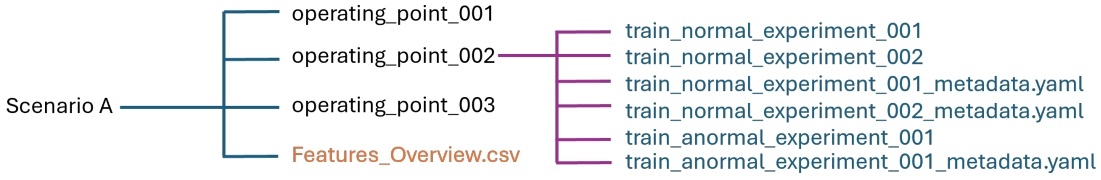


Figure 2: Data repository folder structure sample

[1]https://data.for5359.de/data/for5359b2/
[2]http://131.246.125.99:8501/

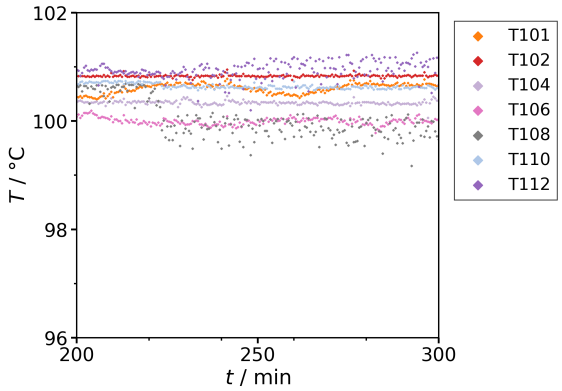
Table 1: Overview of experimental scenarios and extent of data collected

Scenario	Total runs	Normal runs	Anomalous runs	Total duration of steady state (hrs)	Time points
Scenario A	12	4	8	520	62400
Scenario B	9	3	6	80	9600
Scenario C	7	4	3	30	3600

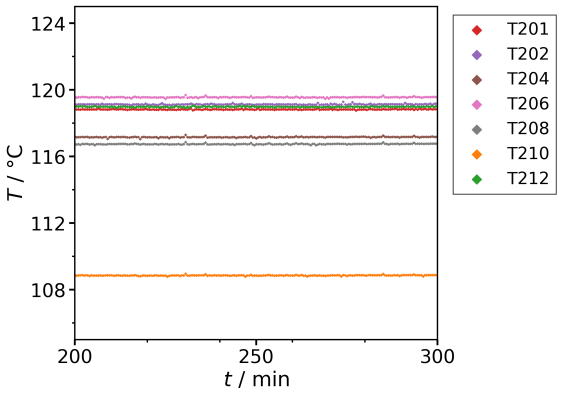
3.2 Sample data without anomalies

For the sake of brevity, we show sample data only from Scenario B in the main part of this publication (sample data from the other two scenarios are shown in the Supporting Information S2, S4). Figure 3 shows the temperature profiles in the columns, the differential pressures in the reboilers as a measure of their level, and the mass flow rates to and from the columns, respectively, during the fault-free steady state.

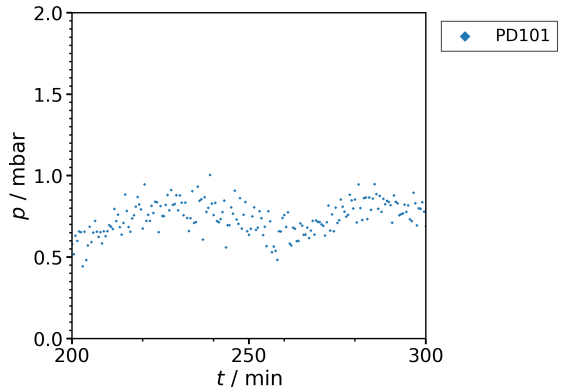
Figure 3a and 3b show that all temperatures are nearly constant with time, except for the temperature sensor at the feed stage (T108) and the reboiler temperature sensor (T101) in 3a. T108 shows small fluctuations due to disturbances of the incoming feed stream, and T101 shows steady oscillations due to the P-only level controller in the reboiler. Figures 3c and 3d represent the differential pressure in the reboilers (PDIC101, PDIC201), indicating also the effect of the P-only level controller, maintaining the reboiler level by manipulating the power of the corresponding bottom pump (P102, P202). In Figures 3e and 3f, the feed (F104), bottoms (M103, M203) and distillate (F102, F103) flow rates can be observed. Oscillations can be observed in bottom stream flow rates due to the behaviour of the P-only level controller. The feed in this setup is delivered using a membrane pump with a manual valve, without active flow control, so slight changes in feed rate over time are normal. This, in turn, influences the reboiler level and, through the P-only control of the bottoms pump, and leads to gradual drifts in the bottoms flow. Such variations arise naturally from the system's design rather than from any fault in operation. It is clear that normal operating points are not easily identifiable. There are relatively stable sensors, such as the temperature sensor; however, mass flow rates have spikes, drifts, and oscillations.



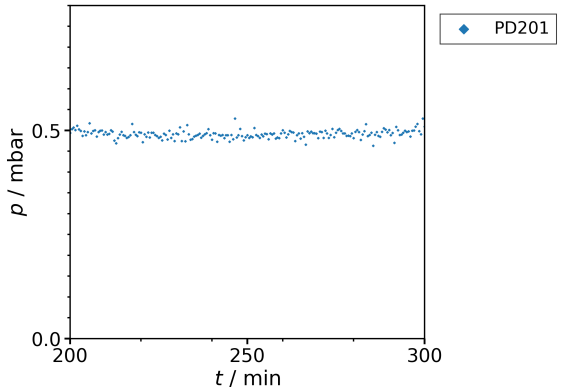
(a) Temperatures in column 1



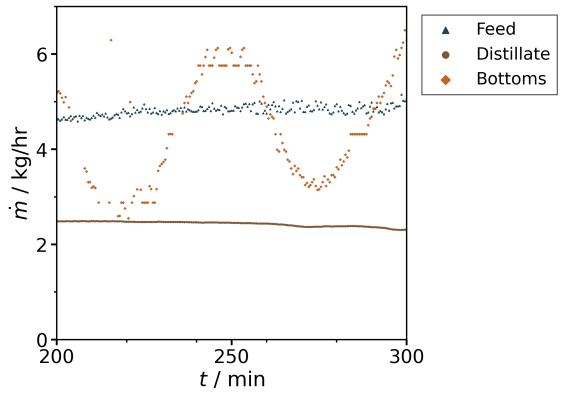
(b) Temperatures in column 2



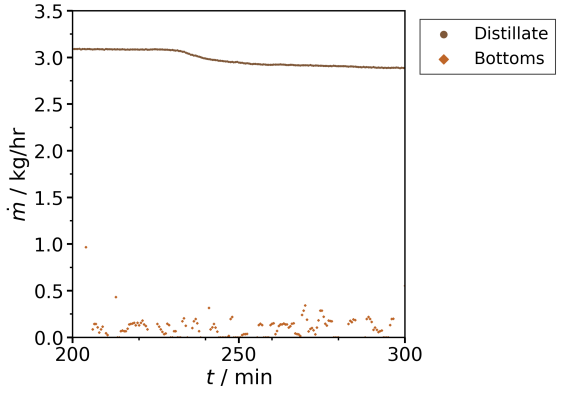
(c) Differential pressure reboiler in column 1



(d) Differential pressure reboiler in column 2



(e) Mass flow rates around column 1



(f) Mass flow rates around column 2

Figure 3: Steady state sample data with n-butanol and water (operating point 10) during normal operation (train_normal_experiment_001.csv)

3.3 Sample data with anomalies

Table 2 lists the observed anomalies in our continuous distillation plant and the affected sensors. Many anomalies were manually triggered. For example, a clogging in the feed pipeline could be manually triggered by closing a valve in the feed pipeline. The most common anomalies are clogging, reboiler level instability, overheating, faulty instruments, pressure control problems, and operating system crashes. Table 3 lists the methods used to induce the anomalies.

Table 2: Distillation column anomalies relevant to our mini-plant

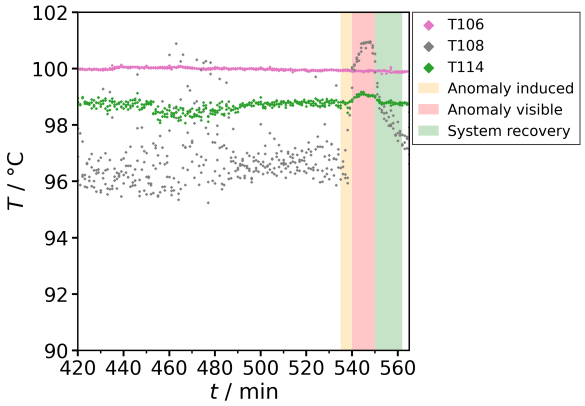
Anomaly category	**Affected sensors or actuators**	**Specific location**
Anomalous operation incidents	TH101, TH201, TH103, TH203	Reboiler overheating
	PIC101, PIC201	Pressure build-up
	PIC101, PIC201	Pressure loss
Clogging	A107, A106, PD101, PD201	Pipes, outlets, packing
	A103, A203	Pipes, outlets, packing
Condenser control problems	TH102, TH202, AV108, AV208	Cooling water line
Condenser malfunctions	TH102, TH202	Coolant liquid refill
Faulty instruments	PDIC101, PDIC201	Level sensor measurements
Reboiler level instability	PDIC101, PDIC201	Reboiler
Operating error	—	LabVIEW crash
Pressure control problems	PIC101, AV105, AV106	Inert gas flow

For all anomalies, we have provided information on the time frame of occurrence of the anomaly, along with the affected sensor in the metadata file. We also provided additional details on when the anomaly was induced, when it was visible in the data, and when the system recovery was performed, for cases where this information is available. Figures 4c and 4d show reboiler levels during an anomalous run with the occurence of three anomalies (a level sensor anomaly in C2, a temperature sensor anomaly in C1, and a feed line clogging anomaly in C1, respectively). A reboiler level anomaly could lead to flooding or drying up of the column if not identified in a timely manner. In 4f, effects of anomalies occurring in the plant, during the same experiment as the level anomalies. In comparison to normal operation, the C2 bottom mass flow remains close to zero for most of the anomalous experiment.

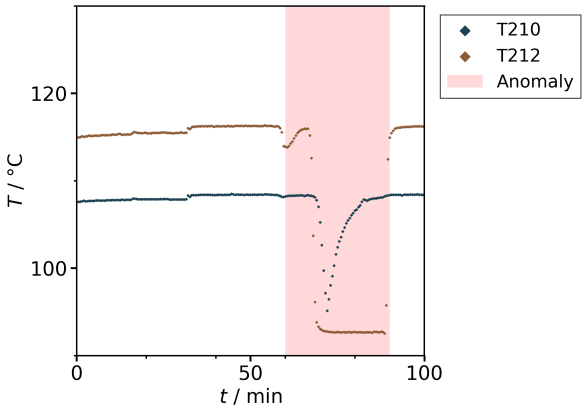
Table 3: Inducing anomalies in the mini-plant.

Anomaly category (Specific location)	**Strategy**
Anomalous operation incidents (Reboiler overheating)	Gradually increase the thermostat (TH101, TH201) temperature by 10 K and, when the reboiler level (PDIC101 at 1.1 mbar or PDIC201 at 0.35 mbar) falls below the minimum threshold, reset it to the steady-state value.
Condenser control (Cooling water line)	Gradually raise the thermostat (TH102 or TH202) by 10 K or switch off the automatic cooling water valve (AV108), then reset once the distillate flow rate falls below a safe threshold.
Condenser malfunction (Coolant liquid refill)	Not refilling thermostat coolant liquid timely, leading to insufficient cooling.
Clogging (Pipes)	Partially close the feed manual valve and reopen once the reboiler level (PDIC101 or PDIC201) drops below a safe threshold; similarly, partially close the bottoms manual valve and reopen when the level rises above the safe threshold (PDIC101: 2 mbar, PDIC201: 1 mbar).
Faulty instruments (Level sensor measurements)	Loosen the pipes to the level devices (PDIC101/PDIC102) or alter RS232 adapter settings to simulate faults.
Reboiler level instability (Reboiler)	Switch off TH101/TH201 until the reboiler level (PDIC101/PDIC201) increases above a safe threshold.
Pressure control problems (Inert gas flow)	Manually open/close the pressure release valve, then reset to the original position once pressure exceeds the threshold (PIC101 at 1.5 mbar).

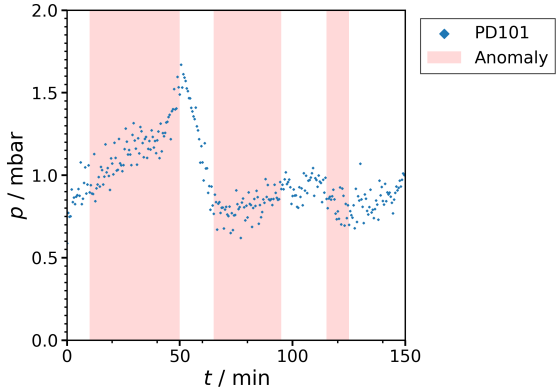
Figures 4a and 4b represent anomalous temperatures. Spikes can be observed in T108 in Figure 4a, due to manually triggered clogging in the feed line. Figures 4e shows the feed mass flow rate that was affected due to a manually induced pipeline clogging via a hand valve in the feed pipeline to C1. Similar plots for anomalies observed or created for the water system and the reactive OME system are provided in the Supporting Information (S2, S4). In the case of water system, the anomalies that were frequently encountered were unstable reboiler levels, reboiler overheating, feed line clogging, and distillate line clogging, temperature sensor, head pressure sensor, and faulty instruments. For the OME system, anomalies were encountered with temperature sensors, reboiler level, and faulty instruments.



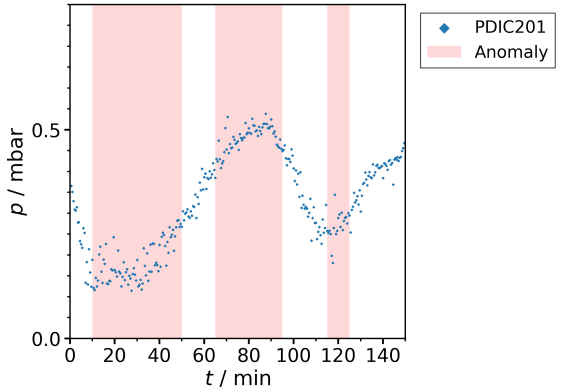
(a) Temperatures in column 1



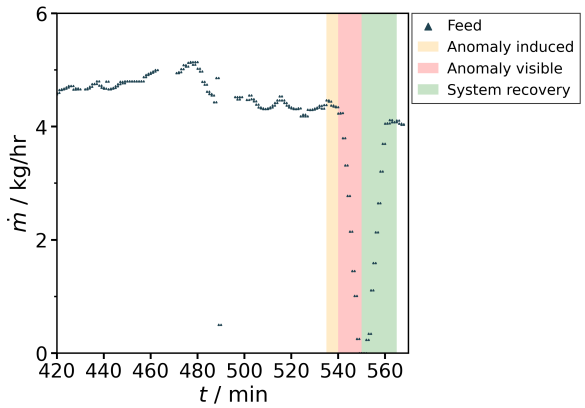
(b) Temperatures in column 2



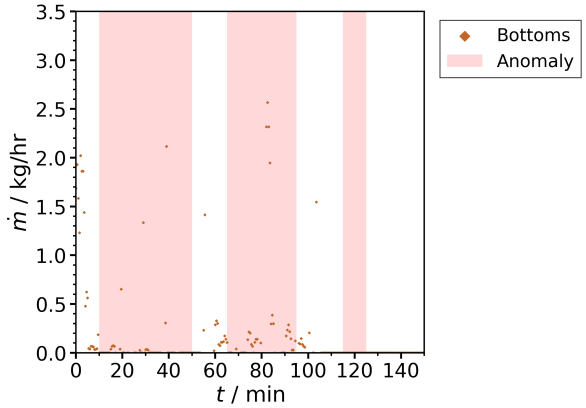
(c) Differential pressure reboiler in column 1



(d) Differential pressure reboiler in column 2



(e) Feed mass flow rate in column 1



(f) Bottoms mass flow rate in column 2

Figure 4: Steady state sample data from Scenario B (n-butanol and water) during anomalies
4a, 4e - operating point 006 - test_anormal_experiment_001
4b - operating point 001 - test_anormal_experiment_001
4c, 4d, 4f - operating point 001 - test_anormal_experiment_002

4 Conclusions and outlook

A real-word dataset from a continuous chemical distillation plant was created, annotated, and provided with open access in CC BY 4.0 license. The dataset contains process data in three scenarios with a total of more than 630 hours of steady-state operation. The dataset can be used in the following exemplary use cases (without the ambition for completeness):

Anomaly detection (AD). The dataset contains operating points with and without anomalies and the respective annotations. We have used parts of the data already to compare literature AD methods[20] and test newly developed methods[31].

Synthetic data generation. The dataset can be used to train generative models that produce synthetic data as examplified by[39–42]. Creating synthetic process data using solely data-driven methods is not very reliable, especially when extrapolation to novel operating points is the goal. Instead, one could create a mechanistic model[43] of the chemical process as a simulation environment and calibrate it to the dataset. For the chemical plant of the present work, we are planning to develop such a model in future work.

Visualization and exploration of data. With multiple variables recorded across several stable operating periods, the dataset is well-suited for testing and developing multivariate visualization techniques[44,45]. This includes dimensionality reduction methods and interactive plotting tools that help engineers and researchers in tasks like spotting and explaining anomalies and understanding sensor relationships and process behavior.

Thus, we believe the dataset is a valuable contribution for the development and testing of various machine learning methods.

5 Acknowledgements

This work was funded by the German Research Fundation (DFG) and conducted in collaboration with the FOR5359 Research Unit on Deep Learning on Sparse Chemical Process Data. The authors gratefully acknowledge Stefan Hartl and Kay Benjamin Wagner

for their assistance in conducting experiments with the mini-plant.

6 Supporting information

Overview of the mini-plant with images, specifications, and sensor descriptions. Scenario A - overview, sample data with and without anomalies. Scenario B - overview, sample concentration profiles, and actuator data. Scenario C - sample data with and without anomalies. Anomaly labeling scheme. Visualization website.

References

(1) Lapkin, A.; Plucinski, P. Engineering Factors for Efficient Flow Processes in Chemical Industries. *Chemical Reactions and Processes under Flow Conditions* **2010**,

(2) Wiles, C.; Watts, P. Continuous process technology: a tool for sustainable production. *Green Chem.* **2014**, *16*, 55–62.

(3) Ricker, N. L. *Chemometrics*; Springer Netherlands: Dordrecht, 1984; pp 205–223.

(4) Pearson, R. K. Outliers in process modeling and identification. *IEEE Trans. Control Syst. Technol.* **2002**, *10*, 55–63.

(5) Kister, H. What Caused Tower Malfunctions in the Last 50 Years? *Chemical Engineering Research and Design* **2003**, *81*, 5–26.

(6) Kumar, S.; Gupta, G. R.; Parihar, R. P. S. Fast anomaly detection for multivariate industrial time series data. 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT). 2023.

(7) Octiva, C. S.; Suryadi, D.; Judijanto, L.; Laia, M.; Irwan, D. The application of artificial intelligence for anomaly detection in big data systems for decision-making. *ijsecs* **2024**, *4*, 983–989.

(8) Garg, S.; Singh, A.; Batra, S.; Kumar, N.; Obaidat, M. S. EnClass: Ensemble-Based Classification Model for Network Anomaly Detection in Massive Datasets. GLOBECOM 2017 - IEEE Global Communications Conference. 2017; pp 1–7.

(9) Nassif, A. B.; Talib, M. A.; Nasir, Q.; Dakalbab, F. M. Machine Learning for Anomaly Detection: A Systematic Review. *IEEE Access* **2021**, *9*, 78658–78700.

(10) Bergmann, P.; Batzner, K.; Fauser, M.; Sattlegger, D.; Steger, C. The MVTec anomaly detection dataset: A comprehensive real-world dataset for unsupervised anomaly detection. *Int. J. Comput. Vis.* **2021**, *129*, 1038–1059.

(11) Mowbray, M.; Vallerio, M.; Perez-Galvan, C.; Zhang, D.; Del Rio Chanona, A.; Navarro-Brull, F. J. Industrial data science – a review of machine learning applications for chemical and process industries. *React. Chem. Eng.* **2022**, *7*, 1471–1509.

(12) Chadha, G. S.; Rabbani, A.; Schwung, A. Comparison of Semi-supervised Deep Neural Networks for Anomaly Detection in Industrial Processes. 2019 IEEE 17th International Conference on Industrial Informatics (INDIN). 2019; pp 214–219.

(13) Monroy, I.; Escudero, G.; Graells, M. *Computer Aided Chemical Engineering*; Elsevier, 2009; pp 255–260.

(14) Song, B.; Suh, Y. Narrative Texts-Based Anomaly Detection Using Accident Report Documents: The Case of Chemical Process Safety. *Journal of Loss Prevention in the Process Industries* **2019**, *57*, 47–54.

(15) Lu, D.; Li, S.; Zhao, Y.; Han, Q. *Lecture Notes in Computer Science*; Lecture notes in computer science; Springer Nature Singapore: Singapore, 2024; pp 88–100.

(16) Downs, J.; Vogel, E. A Plant-wide Industrial Process Control Problem. *Computers & Chemical Engineering* **1993**, *17*, 245–255.

(17) Rieth, C. A.; Amsel, B. D.; Tran, R.; Cook, M. B. Additional Tennessee Eastman process simulation data for Anomaly Detection Evaluation. 2017.

(18) Reinartz, C.; Kulahci, M.; Ravn, O. An extended Tennessee Eastman simulation dataset for fault-detection and decision support systems. *Comput. Chem. Eng.* **2021**, *149*, 107281.

(19) Hartung, F. et al. Deep Anomaly Detection on Tennessee Eastman Process Data. *Chemie Ingenieur Technik* **2023**, *95*, 1077–1082.

(20) Muraleedharan, A.; Hartung, F.; Wagner, D.; Kloft, M.; Burger, J. ESCAPE-34 / PSE 2024 Conference.

(21) Purohit, H.; Tanabe, R.; Ichige, T.; Endo, T.; Nikaido, Y.; Suefusa, K.; Kawaguchi, Y. MIMII dataset: Sound dataset for malfunctioning industrial machine investigation and inspection. Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019). 2019.

(22) Liu, J.; Yan, Y.; Li, J.; Zhao, W.; Chu, P.; Sheng, X.; Liu, Y.; Yang, X. IPAD: Industrial process anomaly detection dataset. *IEEE Trans. Circuits Syst. Video Technol.* **2025**, *35*, 380–393.

(23) Bai, H.; Mou, S.; Likhomanenko, T.; Cinbis, R. G.; Tuzel, O.; Huang, P.; Shan, J.; Shi, J.; Cao, M. VISION datasets: A benchmark for vision-based InduStrial InspectiON. **2023**,

(24) Martí, L.; Sanchez-Pi, N.; Molina, J. M.; Garcia, A. C. B. Anomaly detection based on sensor data in petroleum industry applications. *Sensors (Basel)* **2015**, *15*, 2774–2797.

(25) Ragab, A.; El-Koujok, M.; Poulin, B.; Amazouz, M.; Yacout, S. Fault diagnosis in industrial chemical processes using interpretable patterns based on Logical Analysis of Data. *Expert Syst. Appl.* **2018**, *95*, 368–383.

(26) Ralston, P.; DePuy, G.; Graham, J. H. Computer-based monitoring and fault diagnosis: a chemical process case study. *ISA Trans.* **2001**, *40*, 85–98.

(27) Sala, D. A.; Jalalvand, A.; Van Yperen-De Deyne, A.; Mannens, E. Multivariate time series for data-driven endpoint prediction in the basic oxygen furnace. 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA). 2018.

(28) Emmott, A.; Das, S.; Dietterich, T.; Fern, A.; Wong, W.-K. A meta-analysis of the anomaly detection problem. **2015**,

(29) Saebi, M.; Nan, B.; Herr, J. E.; Wahlers, J.; Guo, Z.; Zurański, A. M.; Kogej, T.; Norrby, P.-O.; Doyle, A. G.; Chawla, N. V.; Wiest, O. On the use of real-world datasets for reaction yield prediction. *Chem. Sci.* **2023**, *14*, 4997–5005.

(30) Ferre, A.; Voggenreiter, J.; Tönges, Y.; Burger, J. Demonstrationsanlage für die Synthese von OME-Kraftstoffen. *MTZ - Motortechnische Zeitschrift* **2021**, *82*, 28–33.

(31) Wagner, D.; Michels, T.; Schulz, F. C. F.; Nair, A.; Rudolph, M.; Kloft, M. *Transactions on Machine Learning Research*.

(32) Schmitz, N.; Ströfer, E.; Burger, J.; Hasse, H. Conceptual design of a novel process for the production of poly(oxymethylene) dimethyl ethers from formaldehyde and methanol. *Ind. Eng. Chem. Res.* **2017**, *56*, 11519–11530.

(33) Ferre, A.; Voggenreiter, J.; Breitkreuz, C. F.; Worch, D.; Lubenau, U.; Hasse, H.; Burger, J. Experimental Demonstration of the Production of Poly(Oxymethylene) Dimethyl Ethers from Methanolic Formaldehyde Solutions in a Closed-Loop Mini-Plant. *Chemical Engineering Research and Design* **2024**, *211*, 331–342.

(34) Thompson, A. B.; Cope, S. J.; Swift, T. D.; Notestein, J. M. Adsorption of n-butanol from dilute aqueous solution with grafted calixarenes. *Langmuir* **2011**, *27*, 11990–11998.

(35) Machida, H.; Watanabe, A.; Horizoe, H. Phase equilibrium measurement of n-butane/water/n-butanol system for development of an energy-saving biobutanol extraction process. *J. Adv. Chem. Eng.* **2018**, *08*.

(36) Luyben, W. L. Control of the Heterogeneous Azeotropic n-Butanol/Water Distillation System. *Energy & Fuels* **2008**, *22*, 4249–4258.

(37) Voggenreiter, J.; Ferre, A.; Burger, J. Scale-up of the Continuous Production of Poly(Oxymethylene) Dimethyl Ethers from Methanol and Formaldehyde in Tubular Reactors. *Industrial & Engineering Chemistry Research* **2022**, *61*, 10034–10046.

(38) Streamlit, Inc. Streamlit: A Faster Way to Build and Share Data Apps. https://streamlit.io/, 2025; Accessed: 2025-10-15.

(39) Sasiaowapak, T.; Boonsang, S.; Chuwongin, S.; Tongloy, T.; Lalitrojwong, P. Generative AI for industrial applications: Synthetic dataset. 2023 15th International Conference on Information Technology and Electrical Engineering (ICITEE). 2023.

(40) Baryshnikov, E.; Kanin, E.; Vainshtein, A.; Osiptsov, A.; Burnaev, E. ANNs trained on synthetic and lab data for modeling steady-state multiphase pipe flow. First EAGE Digitalization Conference and Exhibition. 2020.

(41) Albuquerque, G.; Löwe, T.; Magnor, M. Synthetic generation of high-dimensional datasets. *IEEE Trans. Vis. Comput. Graph.* **2011**, *17*, 2317–2324.

(42) Manduchi, L. et al. On the challenges and opportunities in generative AI. **2024**,

(43) Nagda, M.; Ostheimer, P.; Specht, T.; Rhein, F.; Jirasek, F.; Mandt, S.; Kloft, M.; Fellenz, S. Setpinns: Set-based physics-informed neural networks. *arXiv preprint arXiv:2409.20206* **2024**,

(44) Wang, R.; Baldea, M.; Edgar, T. Data visualization and visualization-based fault detection for chemical processes. *Processes (Basel)* **2017**, *5*, 45.

(45) Reinhardt, D.; Wagner, D.; Muraleedharan, A.; Arweiler, J.; Jungjohann, I.; Jirasek, F.; Burger, J.; Hasse, H.; Kloft, M.; Leitte, H. CPAX: Comparative visualization of known and novel anomalies for monitoring chemical plants. Accessed: 2025-4-17.

(46) Ahmed, C. M.; Gauthama Raman, M. R.; Mathur, A. P. Challenges in Machine Learning Based Approaches for Real-Time Anomaly Detection in Industrial Control Systems. Proceedings of the 6th ACM on Cyber-Physical System Security Workshop (ASIA CCS '20). 2020; pp 23–29.

(47) Kooijman, H. A.; Sorensen, E. Recent Advances and Future Perspectives on More Sustainable and Energy Efficient Distillation Processes. *Chemical Engineering Research and Design* **2022**, *188*, 473–482.

(48) Liaw, H.-J.; Chen, C.-T.; Gerbaud, V. Flash-point Prediction for Binary Partially Miscible Aqueous–Organic Mixtures. *Chemical Engineering Science* **2008**, *63*, 4543–4554.

(49) Madakyaru, M.; Kini, K. R. A Novel Anomaly Detection Scheme for High Dimensional Systems Using Kantorovich Distance Statistic. *International Journal of Information Technology* **2022**, *14*, 3001–3010.

(50) Md Nor, N.; Che Hassan, C. R.; Hussain, M. A. A Review of Data-Driven Fault Detection and Diagnosis Methods: Applications in Chemical Process Systems. *Reviews in Chemical Engineering* **2020**, *36*, 513–553.

(51) Rieth, C. A.; Amsel, B. D.; Tran, R.; Cook, M. B. Additional Tennessee Eastman Process Simulation Data for Anomaly Detection Evaluation. 2017; https://dataverse.harvard.edu/citation?persistentId=doi:10.7910/DVN/6C3JR1.

(52) Kini, K. R.; Madakyaru, M.; Harrou, F.; Vatti, A. K.; Sun, Y. Robust Fault Detection in Monitoring Chemical Processes Using Multi-Scale PCA with KD Approach. *ChemEngineering* **2024**, *8*, 45.

(53) Ardali, N. R.; Zarghami, R.; Gharebagh, R. S.; Mostoufi, N. *Computer Aided Chemical Engineering*; Elsevier, 2022; Vol. 49; pp 1447–1452.

(54) Schmitz, N.; Homberg, F.; Berje, J.; Burger, J.; Hasse, H. Chemical equilibrium of the synthesis of poly(oxymethylene) dimethyl ethers from formaldehyde and methanol in aqueous solutions. *Ind. Eng. Chem. Res.* **2015**, *54*, 6409–6417.

(55) Ochiai, S. Calculating process control parameters from steady state operating data. *ISA Trans.* **1997**, *36*, 313–320.

(56) Yi, C. K.; Luyben, W. L. Evaluation of plant-wide control structures by steady-state disturbance sensitivity analysis. *Ind. Eng. Chem. Res.* **1995**, *34*, 2393–2405.

(57) Tian, Z.; Hoo, K. A. Multiple model-based control of the Tennessee-Eastman process. *Ind. Eng. Chem. Res.* **2005**, *44*, 3187–3202.

(58) Zheng, A. Nonlinear model predictive control of the Tennessee Eastman process. Proceedings of the 1998 American Control Conference. ACC (IEEE Cat. No.98CH36207). 1998; pp 1700–1704 vol.3.

(59) He, R.; Chen, G.; Dong, C.; Sun, S.; Shen, X. Data-driven digital twin technology for optimized control in process systems. *ISA Trans.* **2019**, *95*, 221–234.

(60) Jha, A.; Okorafor, O. C. Optimal plantwide process control applied to the Tennessee Eastman problem. *Ind. Eng. Chem. Res.* **2014**, *53*, 738–751.

(61) Larsson, T.; Hestetun, K.; Hovland, E.; Skogestad, S. Self-optimizing control of a large-scale plant: The Tennessee Eastman process. *Ind. Eng. Chem. Res.* **2001**, *40*, 4889–4901.