cPAX: Comparative Visualization Of Known And Novel Anomalies For Monitoring Chemical Plants

Daniel Reinhardt[1], Dennis Wagner[2], Aparna Muraleedharan[4], Justus Arweiler[3], Indra Jungjohann[3], Fabian Jirasek[3], Jakob Burger[4], Hans Hasse[3], Marius Kloft[2], and Heike Leitte[1]

[1] Visual Information Analysis, University of Kaiserslautern-Landau
{d.reinhardt,leitte}@rptu.de
[2] Machine Learning Group, University of Kaiserslautern-Landau
{dwagner,kloft}@cs.uni-kl.de
[3] Laboratory of Engineering Thermodynamics, University of Kaiserslautern-Landau
{justus.arweiler,indra.jungjohann,fabian.jirasek,hans.hasse}@rptu.de
[4] Chemical Process Engineering, Technical University of Munich
{aparna.muraleedharan,burger}@tum.de

Abstract. Anomalies in a chemical plant can have disastrous consequences from endangering personnel and the environment to significant costs caused by damages in the plant. Anomalies need to be identified and fixed as soon as possible to avoid the harshest consequences. Thus, online monitoring is essential for a safe and cost-efficient operation. However, the wealth of information available to operators at any given time can obscure signs of anomalous behavior. We propose a plant monitoring system called cPAX (the chemical Plant Anomaly eXplorer) that concisely and intuitively visualizes the data collected from a plant online, and incorporates anomaly detection methods to visualize the current state of the plant allowing for comparison with known anomalous scenarios. The system integrates a database of known anomaly cases to rank anomalous scenarios and provide suggestions for the operator. Additionally, the database is set up to incorporate additional information, such as instructions for fixing the anomaly, to provide more context to the operator. The system is developed and continuously tested with several domain experts to provide an effective user experience.

Keywords: Time Series · Anomalies · Visual Analysis · Explainable AI · Chemical Process Engineering

1 Introduction

Large chemical plants are expensive to build, operate, and maintain. Due to the complexity of the underlying chemical process, the inevitable degradation of the equipment, and other external factors, anomalies can occur with potentially devastating consequences if not addressed properly [21]. Anomalies in operation

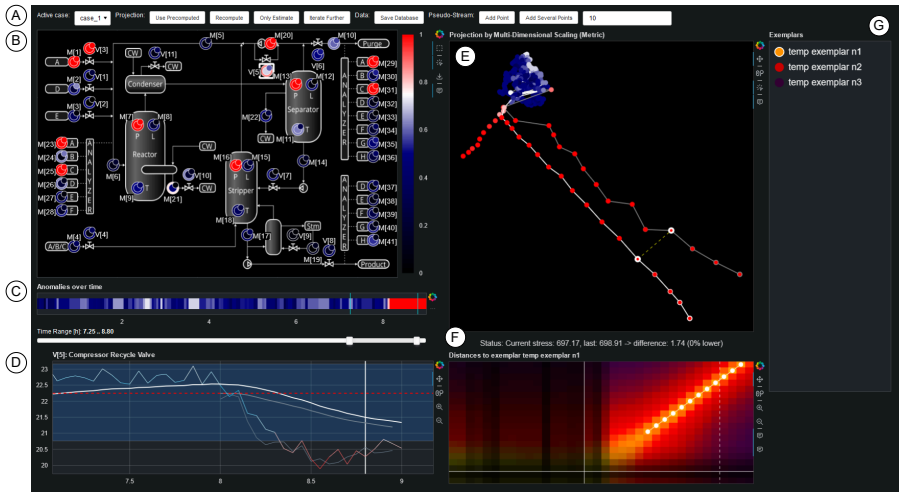

Fig. 1. Full view of the visualization tool at the end of a work flow. The data shown is the first case in the Tennessee Eastman Process benchmark.

can result in costly damage to the equipment, endanger the lives of the operating personnel, or deal irreversible damage to the environment, as most hazardous accidents in chemical plants start off as simple process anomalies [23, 31, 8]. Thus, it is of utmost importance to identify deviations from normal operating behavior and initiate counter measures as fast as possible. To this end, modern chemical plants use control systems that consist of physical sensors, strategically placed in the chemical plant, and software that processes the information from these sensors and presents the operator with information necessary to identify anomalies [6]. Due to the wealth of information collectible from any given chemical plant, presenting this information to an operator in a way that they can easily identify anomalous behavior is a non-trivial task.

In practice, some cases of anomalous behavior are known in advance. Anomalies that have occurred in the past are usually documented and certain anomalies can be characterized based on the physical constraints of the process itself. Modern operators could benefit greatly from this additional information, as a complex solution to a particular anomaly could have already been discovered in the past. However, existing tools do not integrate such information in the visualizations. To this end, we work together with domain and Machine Learning (ML) experts to develop a system called cPAX (the chemical Plant Anomaly eXplorer) as depicted in fig. 1 that uses additional information and carefully chosen visualizations to alert an operator early and suggest solutions to a running anomaly based on a database of known anomalies.

Additionally, we outline how state-of-the-art anomaly detection methods can be integrated into cPAX, displaying anomalies in more detail than existing sys-

tems, and highlighting their localities, in a system with rich interactivity to identify known anomalies, explain unknown ones, and extending the database with such new cases over time.

As a proof of concept, we use the publicly available Tennessee Eastman Process benchmark dataset[5] [19], a widely applied benchmark dataset with a large variety of anomaly cases including information on their root cause. cPAX is based on a modular design that allows for the easy integration of novel plants, as well as customized anomaly detectors and metrics.

The manuscript is structured as follows. In section 2, we discuss related work. To provide a base with which to work, in section 3 we detail how the used anomaly data is created. Methodological considerations underlying cPAX are detailed in section 4 including feature-wise visualization of anomalies (section 4.1), visualization of time series data including visualization in latent spaces (section 4.2), integration of a knowledge base to store previous anomalies as well as the integration of a database of known anomaly cases and it's application in matching (section 4.3). A case study demonstrating the abilities of cPAX is given in section 5.

2 Related Work

Visualization of Time Series Data Several papers present and compare basic methods to render univariate time-series, including line charts with or without augmentations, box plots, and color fields [2, 10], which help form a decision on how to plot the values of a single sensor. However, for high-dimensional data, it does not suffice to look at only single variables at once, as anomalous behavior can easily manifest in the interactions between multiple sensor measurements. In some settings we can expect multiple features to follow similar patterns during normal operation. In such a setting, [15] visualize multiple features as superimposed line charts, where they facilitate filtering the data. This can be seen as the opposite of our setting, but we remark that such a filter could potentially be used as an anomaly detector. [12] use stacked charts to visualize individual features which they rank and scale according to their importance to a given task. However, with the amount of sensors available for chemical plants this approach would be too overwhelming in our context. In order to detect correlations with latencies, lense-based systems like *ChronoLenses* [36] and *KronoMiner* [35] can be used.

Latent Space Based Visualization In order to use several variables at once in order to reveal relationships between time points in the bigger picture, a common approach is to make use of *latent spaces*, using low-dimensional projections of high-dimensional data. Examples include methods utilizing *Principal Component Analysis* (PCA) to find patterns in cyclic data regarding a variable [5]. This, however, would not admit integrating additional deviating data.

[5] See appendix A for a detailed description.

To analyze the development of a single multivariate time series, one can employ *Multi Dimensional Scaling* (MDS) [7] and then analyze the shape of the trace [3]. There are also approaches using a *Self-Organizing Map* (SOM) [22] as base for the projection [16, 27]. Such an approach could be tested in a future version in our program and evaluated against our current projection.

Several possibilities for augmenting projections exist. The renderings for each single point in a projection can display the data vector in radial glyph [33], and colored areas can represent clusters along with a representation of a typical member in form of a line chart [29]. While we do not use such a technique in the current version of our tool, this might become relevant in the future.

Anomaly Detection in Time Series Detecting anomalies in time-series data is an active research field with numerous contributions in recent years [28, 32]. Especially high-dimensional chemical process data presents an attractive opportunity [34, 19, 14, 30, 24, 25]. However, it is still not clear how different methods should be evaluated and compared [32], and, thus, it is still an open research question which methods are best suited for anomaly detection on time series and especially chemical process data [13]. For this reason, we keep the anomaly detector as an interchangeable component in our visualization tool.

3 Background – Anomaly detection in the Tennessee Eastman Process

3.1 The Tennessee Eastman Process

Over the past three decades, the Tennessee Eastman process (TEP) has become the standard benchmark for evaluating learning-based anomaly detection techniques on chemical process data [13, 19]. TEP is based on an existing plant and its operational processes, but the data itself is synthetic, derived from a simulation of the plant. It consists of five main modules that are monitored by 52 sensors. The version of the TEP data used here is available online and is referenced in [19]. This dataset includes error-free data for algorithm training, as well as 20 different types of erroneous data sets with their complete simulations. Each of these 21 data sets has 500 runs, initialized with different random values. Data points are sampled every three minutes over 25 hours for training data and 48 hours for test data, comprising 52 parameters.

Anomaly Detection Anomaly detection stands as a cornerstone challenge in machine learning, driving substantial efforts towards the creation of more resilient and effective models. The primary objective is to accurately characterize the normal operational patterns of a system, thereby identifying significant deviations as anomalies. Contemporary approaches often leverage deep neural networks, which excel in learning complex patterns from observational data.

This section investigates anomaly detection specifically within the context of time-series data, a pivotal component within cPAX. Here, an integrated anomaly

detector not only supports the control system but also generates insightful visualizations to aid operators in interactive exploration.

Anomaly Detection On Time Series Given example data from normal operating conditions $\mathbf{x}^1_{train}, \ldots, \mathbf{x}^n_{train}$, various anomaly detectors can be trained, as extensively reviewed in [28, 32]. These detectors yield a score function

$$s\colon (\mathbf{x}(0), \ldots, \mathbf{x}(t)) \mapsto [0, 1]$$

that assesses time series data up to time t, typically within a finite horizon $(\mathbf{x}(t-w), \ldots, \mathbf{x}(t))$ for some $w \in \mathbb{N}$. Importantly, this function assigns a score to each time step, where higher scores indicate a higher likelihood of an anomaly. Detailed evaluations of anomaly detectors for the TEP are provided in [13].

Sensor-based Anomaly Scores Our objective is to assist operators in pinpointing the cause of anomalies, necessitating detailed information on which sensors are implicated. Specifically, we require anomaly scores for each sensor variable, distinct from the system-wide scores used thus far. If an anomaly detector already provides variable-wise scores, all necessary information is inherently available, transforming them to system-wide scores is straightforward. Otherwise, methods are needed to allocate anomaly scores to individual features, quantifying their contributions to the overall anomaly score. Techniques such as *Layer-Wise Relevance Propagation* (LRP) [4] exemplify advancements in *Explainable Artificial Intelligence*, where scores generated by neural networks are back-propagated and normalized across network layers.

4 Methods

Existing anomaly detectors inform operators when plant data deviates from normal and identify sensors possibly responsible for anomalies. Managing this extensive information from numerous sensors is daunting. To address this, we introduce cPAX (cf. fig. 1), a system that provides concise and intuitive online visualization of plant data. Significantly, cPAX integrates anomaly detection methods to visually represent the plant's current state, enabling comparisons with known anomalous scenarios for enhanced monitoring and analysis.

System Overview cPAX consists of 7 major components (A–G) as annotated in fig. 1. (A) contains the user control paraments. (B) shows a schematic representation of the plant with all components (white shapes) and sensors (colored circles). (C) is a time slider including anomaly scores in color. (D) provides detailed information for a selected sensor. (E) is an annotated non-linear projection of the multivariate sensor data. (F) supports comparative analysis between new anomalies and the ones from the knowledge base. (G) gives an overview of relevant elements in the knowledge base.

4.1 Schematic Plant representation

To offer operators a familiar data perspective, we present a process flow diagram depicted in component (B) of fig. 1. This diagram provides a high-level abstraction of the plant, where each of the 52 sensors is symbolized by a spherical glyph conveying anomaly status information. Component (C) features a time slider illustrating the temporal evolution of the plant simulation and displaying system-wide anomaly scores over time using a color spectrum. It also allows operators to adjust the observed time period for detailed analysis.

Anomaly Scores As Features Operators can easily lose focus due to divided attention, distractions, or fatigue from extended work hours. Therefore, it is crucial to streamline the information presented and avoid overwhelming operators with excessive detail. Additionally, we advocate for an intuitive visualization method to pinpoint anomalies using the plant's schematic. The concept involves color-coding each sensor based on its anomaly score as provided by the anomaly detector.

We depict these scores on a schematic of the chemical plant, illustrated in fig. 2. On the left side, where no anomaly is detected, all sensors appear in dark blue, blending discreetly with the black background. Conversely, on the right side of the figure, during an anomaly occurrence, sensors indicating the anomaly are highlighted in bright red to emphasize critical areas. This design facilitates rapid and efficient monitoring with a straightforward visual inspection.

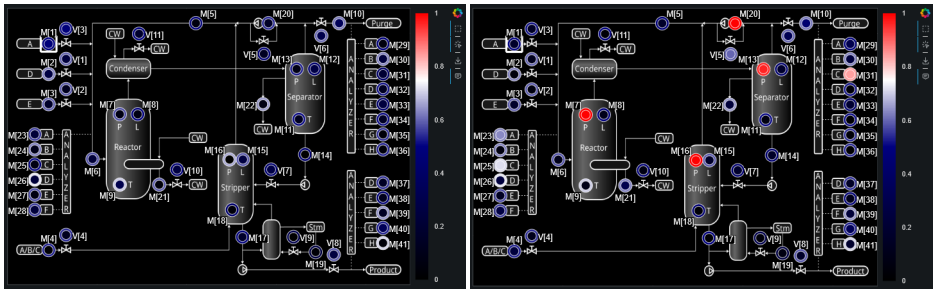

Fig. 2. The process flow diagram at two different points in time. The current variable-wise anomaly scores are shown (inner circles), as well as their maxima in the selected interval (outer rings). The left side shows a normal operating plant, the right side one that went anomalous in several parts of the system.

An anomaly might occur fleetingly, potentially unnoticed by the user if they are not actively monitoring the display. To ensure this critical information is captured, each sensor is depicted using a glyph featuring an inner circle and an outer ring. The inner circle's color represents the current anomaly score, while the outer ring displays the maximum score observed for that sensor over a user-

defined time interval. This ensures that potentially important information is retained.

4.2 Visualizing Sensor Data

A chemical plant can be equipped with a variety of sensors, including temperature, pressure, flow, mixture composition and level sensors. Each of these sensors can be sampled at a predefined fixed frequency during the operation of the plant allowing for continuous oversight. Let $x = (x_1, \ldots, x_k)$, $x_i \in \mathbb{R}$, $1 \leq i \leq k$ be a k-dimensional (kD) real-valued observation and let $T = \{\mathbf{x}(t)\}$, $1 \leq t \leq N$ be a time series with N kD observations. The raw sensor data is thus represented as a multivariate time-series $\mathbf{x}(t) \in \mathbb{R}^k$, where each feature $\mathbf{x}^i$ corresponds to a single sensor.

Augmented Time Series View (fig. 1(D)) Component (D) displays the raw data from an individual sensor, crucial for detailed and trustworthy analysis in a time series view. The sensor that is displayed can be interactively selected by the operator by clicking on a glyph in the process diagram (B), in the time range chosen in component (C).

A larger depiction of the view is shown in Figure 3, featuring examples of both normal operating conditions (top) and anomaly conditions (bottom).

Fig. 3. Two instances of the variable view, depicting raw sensor data in different variables. The top view shows a normal development, while the bottom one is abnormal. We see this by the line leaving the blue area, which marks the range of the normal operating conditions, as well as the line color, which encodes the anomaly score, changing to red.

The raw sensor data is depicted as a colored line where color encodes the anomaly score. The color legend is given in component (B) and is persistent throughout cPAX. The knowledge base of cPAX can store information on normal operation conditions for a sensor. These are depicted as a blue range in the time series plot. Additional statistical quantities are depicted by dashed red lines (mean and mean ± three standard deviations) as well as a white line for the trend (using exponential smoothing).

In cases where an anomaly is caused by failure of a single system, the augmented time series view already clearly depicts relevant information to identify the cause of an anomaly. In chemical plants this may occur in case of pipe clogging, pump or sensor failure.

Two-Dimensional Projections (fig. 1(E)) Often, anomalies stem from small deviations across multiple sensors, making them challenging to represent effectively with line charts alone. To aid in exploring the high-dimensional multivariate sensor space, we introduce component (E), which visualizes a two-dimensional projection of the k-dimensional signal.

A projection P is a function

$$P : \mathbb{R}^k \to \mathbb{R}^q \tag{1}$$

where $q \ll k$ and typically $q = 2$. If the target dimension q equals two the projected data can be represented as a scatter plot and a time series can be represented as a trajectory connecting subsequent points $(\mathbf{x}_t, \mathbf{x}_{t+1})$, $1 \leq t < N$.

Dimensionality reduction is a well-explored field with numerous techniques tailored to different data types and quality criteria [9]. In their comprehensive survey, Espadoto et al. [9] quantitatively compare 44 projection methods, highlighting their efficacy across various datasets. Among datasets similar to ours, such as *secom* and *seismic*, focusing respectively on manufacturing failure detection and seismic activity, a range of techniques including PCA, MDS, IDMAP, and LAMP demonstrate effective projection qualities. In their global analysis across diverse data sets, they found that best results are in general obtained by MDS, IDMAP, PBC, t-SNE and UMAP.

Espadoto et al. find that multidimensional scaling (MDS) [7] consistently performs well across diverse datasets, including those most akin to ours. Therefore, MDS has been chosen as the primary projection method for cPAX due to its robust performance. Notably, the projection technique is interchangeable and can be easily substituted with another method or switched interactively via the user control panel (cf. fig. 1).

MDS begins by constructing a distance matrix that quantifies pairwise dissimilarities among all items in the dataset. It then employs an iterative optimization algorithm to position each item in a lower-dimensional space, aiming to preserve the original dissimilarities as closely as possible.

For streaming data, we adapt the standard algorithm by integrating each new data point into the embedded space using a heuristic based on the trajectory defined by the last two points. This initialization method sets coordinates

for subsequent runs. Real-world data typically shows minimal deviation between consecutive points, ensuring rapid stabilization and efficient projection with minimal iterations thereafter.

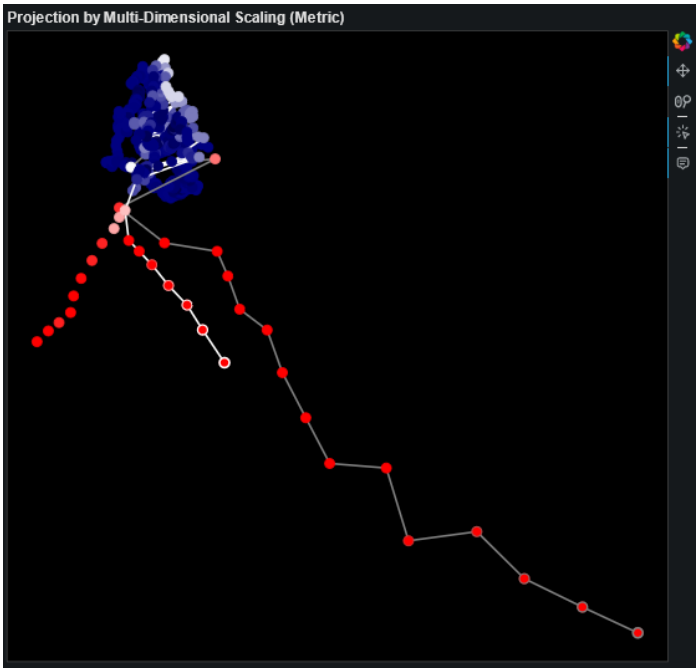

Fig. 4. The high-dimensional data points are projected into a two dimensional latent space using Multi-Dimensional Scaling. A white trace and outlines highlight the ongoing process, grey ones selected known fault cases. The colors of the nodes indicate the anomaly score, with red being high, blue being low.

In fig. 4, we observe the projection results of the ongoing process (white traced lines of nodes), a series under normal operating conditions (large blue bulb), and time series data from various known fault cases (grey traces) onto a two-dimensional plane.

The projection technique endeavors to preserve distances as accurately as possible in its two-dimensional embedding, facilitating intuitive interpretation for users based on normalized values. Traces are utilized as enhancements to emphasize specific time-series aspects.

4.3 Integrating A Database Of Known Anomalies

Investigating and resolving the cause of an anomaly is a challenging task that typically demands a profound understanding of the underlying processes and

substantial technical expertise. Previous instances of anomalies and their resolutions can offer invaluable insights to guide the mitigation of new anomalies.

In this section, we introduce a database dedicated to cataloging known anomalies and discuss methods for ranking these examples based on their similarity to a new anomaly.

A Database Of Examples During chemical plant operations, anomalies, while infrequent, are inevitable, and they often exhibit commonalities across different processes [20]. When anomalies occur, valuable metadata can be gathered, encompassing detailed insights into their causes, effects, such as the location of faults, affected components, mitigation strategies, and subsequent impacts on plant behavior. Additionally, data regarding the operators present during the anomaly and their responses provide crucial information for resolving similar situations.

All this information is systematically collected in a database alongside the raw sensor data for comprehensive analysis and reference.

Matching Multivariate Time-Series In fig. 4, we observed the projection results where data points representing normal operating conditions form a large blue bulb. This serves as a vital reference for what constitutes normal behavior. In contrast, fault cases are expected to exhibit distinct patterns from both normal conditions and other fault cases. Initially starting within the normal bulb, fault time-series gradually diverge outward, forming what we refer to as tendrils.

Armed with this understanding, users can visually compare the ongoing process to known faults. A normal process should remain within the bulb of normal operating conditions, whereas an abnormal run will begin to develop a tendril similar to known fault cases. When the ongoing process closely resembles a known fault, their traces are expected to exhibit proximity to each other.

Component (G) visualizes the database of known fault cases. Each known fault case is represented by a color-coded score indicating its similarity to the ongoing process, with the score derived from an inverted distance measure.

Currently, we utilize a modified version of Subsequence Dynamic Time Warping (SUBDTW) [26] as our distance measure. Our method involves extracting windowed subsequences from known fault time series and comparing them with the ongoing series, excluding points where the time series behave normally. The program offers several variations of this algorithm for flexibility.

One limitation of any projection method is the potential loss of information, which can lead to distortions. To directly assess the actual distances based on our point metric, in the matrix view (F), we visualize a submatrix of the point distance matrix D_{ij}. This submatrix is restricted to column indices representing the ongoing run and row indices corresponding to a specific known fault case, presented in the form of a color field. In fig. 5, we illustrate visualizations for two distinct known fault cases.

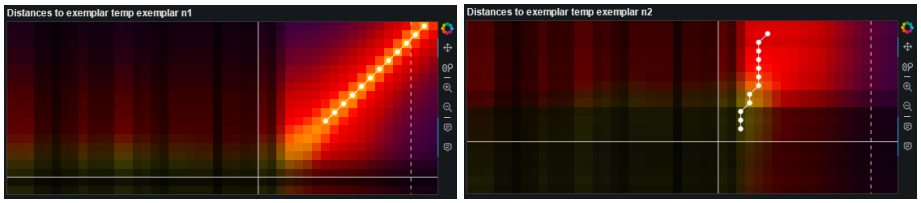

Fig. 5. The distance matrices between the ongoing process and two known fault cases. The x-axis denotes the point in time in the ongoing process, the y-axis that of the time series of the known fault. The color palette is the commonly used plasma palette, ranging from yellow (close) over red to dark violet (far). The opacity of the color is set to the product of the two system-wide anomaly scores at the points in time. Thus the uninteresting normal data points fade into black. The trace shows the optimal warping path found by the windowed SUBDTW algorithm. On the left side, we see a good match, on the right side a bad one.

5 Case Study

Applying cPAX to the TEP Benchmark We demonstrate the functionality of cPAX using the TEP benchmark, highlighting its interactive capabilities. Detailed images corresponding to this section can be found in appendix B.

We compiled a small database of known fault cases based on the first three instances in the TEP training data. For each case, we extracted the critical segment of the time-series around the introduction of the fault into the system. Subsequently, the first case from the testing data was chosen as the ongoing scenario.

We begin monitoring the system. Up to eight hours into the scenario, the case still exhibits normal behavior. The process flow diagram (B) and the time view (C) are predominantly displayed in inconspicuous dark blue, occasionally interrupted by lighter blues due to noise triggering the anomaly detector.

Fault Analysis Around 8:09 (fig. 8), we observe a slightly elevated anomaly score at the stripper pressure sensor. To analyze see the actual values, we click on the node. In the variable view (D) we see that the pressure exceeded the boundaries of the normal operating conditions.

By 8:12 (fig. 9), the variable view shows us that the stripper pressure has deviated more than three standard deviations from the mean of normal operating conditions. The corresponding sensor glyph has now turned into an alarming bright red. We also observe anomalies in the reactor and separator pressures, albeit to a lesser extent. Ideally, we can now match the run to a known fault case in the database. We take a look at the database view (G) to see which historical time series are closest to the ongoing run. In the projection view (E), we see that the ongoing scenario clearly departs from the normal operating condition

bulb, with the latest time point positioned between the first two fault cases. We do not have enough information to be sure about a particular match, though.

By 8:15, and more distinctly at 8:21 (fig. 10), we observe the trajectory of the ongoing case deviating from the second fault case and aligning closely with the first. To confirm this first impression, we engage in further analysis. We select the first two cases by clicking on them in the database view, resulting in the data point nodes of these in the projection view being highlighted by grey traces. This also creates plots in the matrix view (F), in which we see a ridge for the first case, indicating a good match, but a bad one for the second case.

We also want to compare the state of the plant in the current time point to that in the closest time points in the known cases. For that, we select the corresponding node in the projection view. Dashed lines now indicate the closest points in the selected cases to this point. We select one of these for comparison. This changes our visualizations in two ways:

Firstly, the glyphs in the process flow diagram (B) are extended by a small circle in the top right corner, as illustrated in fig. 6. This circle displays the anomaly score of the known fault case at the selected time point, facilitating comparison of anomaly distribution. Additionally, mismatches are highlighted by encircling the sensor in question in red. Secondly, the variable view (D) updates

Fig. 6. An updated glyph in the geometric plant view. The small circle in the upper right corner shows the anomaly score of the known fault case time point, and the sensor is encircled in red to indicate the mismatch.

to include the values of the known case displayed as a grey curve, along with it's trend. For straightforward comparison, the curves are aligned at the selected time points. Examples of both are illustrated in fig. 9.

Using these visualizations, we can clearly identify deviations in several sensors between the ongoing scenario and the second case. In the first case, only one sensor exhibits a significant deviation, and closer examination in the variable view suggests this is likely due to latency.

Additional Insights In a real operational scenario, immediate action is crucial once a fault is identified. In our case, we have identified the first known fault case as a match and can retrieve it's detailed description, informing us that this is a disturbance regarding reactant ratios in one specific inlet. This knowledge is pivotal as it directs attention to a specific area for investigation and intervention.

It's noteworthy that the sensor measuring the flow rate at the inlet in question shows normal values. Even with close monitoring of this variable, any anomalies in the reactant composition remain undetectable since that sensor does not directly measure the ratio, which is a hidden variable. This underscores the complexity of process monitoring, where domain expertise plays a decisive role.

A domain expert proficient in chemical processes gets a hint of the underlying issue from the observed surge in one reactant within the purge, which deviated from normal bounds around 8:21. However, the other relevant reactant shows normal values in the purge at this point in time, and the expert might require additional time to understand what is wrong.

Fault Documentation We now assume that the historical data would not be present, and thus no intervention did occur. Instead, we wait for more data that gives a better understanding of the situation.

At 8:30 (fig. 11), the process flow diagram (B) shows strong anomalies in the sensors regarding the two reactants both in the purge and in the reactor intake. This information equips a domain expert with a solid understanding of the problem, facilitating precise intervention strategies.

Furthermore, the ability to document the issue comprehensively and transform the ongoing case into a documented fault case for future reference underscores the value of integrated anomaly detection systems. This proactive approach not only aids in immediate resolution but also enriches the repository of known fault cases, enhancing the plant's operational resilience.

Such integrated systems exemplify the evolving landscape of anomaly detection in chemical processes, where advancements continue to refine our ability to detect, analyze, and mitigate operational disruptions effectively.

6 Discussion

In this paper, we introduce cPAX, a comprehensive tool designed for real-time monitoring of chemical plants. We critically evaluate various visualization methods in their effectiveness for depicting anomalous behavior and conclude that relying solely on univariate visualizations is insufficient. Consequently, we explore dimensionality reduction techniques to enhance anomaly detection capabilities.

Additionally, we incorporate a database of known anomalies into our tool and propose methods to integrate these examples into visualizations. This enhancement equips operators with crucial information about similar historical anomalies, thereby facilitating more informed resolution strategies.

Our development process involves close collaboration with operators of continuous distillation mini-plants and laboratory-sized batch-distillation plants, alongside machine learning experts. This collaborative effort ensures that our tool delivers an optimal user experience tailored to operational needs.

We leverage substantial datasets from these processing plants, including several induced anomalies, to establish a robust database of examples and rigorously test our tool in real-world applications.

Our tool features an interchangeable anomaly detector, enabling its adaptation to diverse applications and facilitating integration with advancements in anomaly detection techniques.

7 Future Work

The current data point metrics used by cPAX work on the data values alone and might not provide enough separation of seemingly similar data points that differ regarding non-trivial anomalies. Effectively integrating modern anomaly detection methods into the projection to maintain a better metric is an upcoming challenge. This could involve appending the variable-wise anomaly scores to the vector of the raw data values, in order to separate data points also by occurrence of non-trivial anomalies. Balancing the metric's abstraction is crucial to ensure operator interpretability, as overly abstract metrics may hinder effective tool utilization.

Testing both the metrics and anomaly detectors rigorously on real application data is imperative across all tool components, particularly in dimensionality reduction and root cause analysis.

Looking ahead, the experience collected over applying the tool on different plants promises insights that can then be reused to broaden overall efficiency.

Acknowledgements This work was conducted within the DFG research unit FOR 5359 on Deep Learning on Sparse Chemical Process Data.

References

1. Adeli, M., Mazinan, A.: High efficiency fault-detection and fault-tolerant control approach in tennessee eastman process via fuzzy-based neural network representation. Complex & Intelligent Systems **6** (02 2019). https://doi.org/10.1007/s40747-019-0094-3
2. Albers, D., Correll, M., Gleicher, M.: Task-driven evaluation of aggregation in time series visualization. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 551–560. Association for Computing Machinery (2014). https://doi.org/10.1145/2556288.2557200
3. Bach, B., Shi, C., Heulot, N., Madhyastha, T., Grabowski, T., Dragicevic, P.: Time curves: Folding time to visualize patterns of temporal evolution in data. IEEE Transactions on Visualization and Computer Graphics **22**, 559–568 (1 2016). https://doi.org/10.1109/TVCG.2015.2467851
4. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On Pixel-Wise explanations for Non-Linear classifier decisions by Layer-Wise relevance propagation. PLoS One **10**(7), e0130140 (Jul 2015)
5. Bernard, J., Wilhelm, N., Scherer, M., May, T., Schreck, T.: Timeseriespaths: Projection-based explorative analysis of multivariate time series data. In: International Conference in Central Europe on Computer Graphics and Visualization (2012)

6. Christofides, P.D., Davis, J.F., El-Farra, N.H., Clark, D., Harris, K.R.D., Gipson, J.N.: Smart plant operations: Vision, progress and challenges. AIChE Journal **53**(11), 2734–2741 (2007). https://doi.org/https://doi.org/10.1002/aic.11320
7. Cox, M.A.A., Cox, T.F.: Multidimensional Scaling, pp. 315–347. Springer Berlin Heidelberg, Berlin, Heidelberg (2008). https://doi.org/10.1007/978-3-540-33037-0_14
8. E., B.: The bhopal disaster and its aftermath: A review. Environmental Health **4**(6), 1–6 (2005). https://doi.org/https://doi.org/10.1186/1476-069X-4-6
9. Espadoto, M., Martins, R.M., Kerren, A., Hirata, N.S.T., Telea, A.C.: Toward a quantitative survey of dimension reduction techniques. IEEE Transactions on Visualization and Computer Graphics **27**(3), 2153–2173 (Mar 2021). https://doi.org/10.1109/tvcg.2019.2944182
10. Gogolou, A., Tsandilas, T., Palpanas, T., Bezerianos, A.: Comparing similarity perception in time series visualizations. IEEE Transactions on Visualization and Computer Graphics **25**, 523–533 (1 2019). https://doi.org/10.1109/TVCG.2018.2865077
11. Hajihosseini, P., Salahshoor, K., Moshiri, B.: Process fault isolation based on transfer entropy algorithm. ISA Transactions **53**, 230–240 (2014). https://doi.org/10.1016/j.isatra.2013.11.007
12. Hao, M.C., Dayal, U., Keim, D.A., Schreck, T.: Importance-driven visualization layouts for large time series data. Proceedings - IEEE Symposium on Information Visualization, INFO VIS pp. 203–210 (2005). https://doi.org/10.1109/INFVIS.2005.1532148
13. Hartung, F., Franks, B.J., Michels, T., Wagner, D., Liznerski, P., Reithermann, S., Fellenz, S., Jirasek, F., Rudolph, M., Neider, D., Leitte, H., Song, C., Kloepper, B., Mandt, S., Bortz, M., Burger, J., Hasse, H., Kloft, M.: Deep anomaly detection on tennessee eastman process data. Chemie Ingenieur Technik **95**(7), 1077–1082 (2023). https://doi.org/https://doi.org/10.1002/cite.202200238
14. Heo, S., Lee, J.H.: Statistical process monitoring of the tennessee eastman process using parallel autoassociative neural networks and a large dataset. Processes **7** (7 2019). https://doi.org/10.3390/pr7070411
15. Hochheiser, H., Shneiderman, B.: Dynamic query tools for time series data sets: Timebox widgets for interactive exploration. Information Visualization **3**, 1–18 (2004). https://doi.org/10.1057/palgrave.ivs.9500061
16. Hu, Y., Wu, S., Xia, S., Fu, J., Chen, W.: Motion track: Visualizing variations of human motion data. In: 2010 IEEE Pacific Visualization Symposium (PacificVis). pp. 153–160 (2010). https://doi.org/10.1109/PACIFICVIS.2010.5429596
17. Ji, C., Ma, F., Wang, J., Wang, J., Sun, W.: Real-time industrial process fault diagnosis based on time delayed mutual information analysis. Processes **9** (6 2021). https://doi.org/10.3390/pr9061027
18. Katser, I.: https://github.com/YKatser/CPDE, accessed May 22, 2024
19. Katser, I., Kozitsin, V., Lobachev, V., Maksimov, I.: Unsupervised offline change-point detection ensembles. Applied Sciences (Switzerland) **11** (5 2021). https://doi.org/10.3390/app11094280
20. Kister, H.: What caused tower malfunctions in the last 50 years? Chemical Engineering Research and Design **81**(1), 5–26 (2003). https://doi.org/https://doi.org/10.1205/026387603321158159, international Conference on Distillation and Absorption

21. Kleindorfer, P.R., Belke, J.C., Elliott, M.R., Lee, K., Lowe, R.A., Feldman, H.I.: Accident epidemiology and the u.s. chemical industry: Accident history and worst-case data from rmp*info. Risk Analysis **23**(5), 865–881 (2003). https://doi.org/https://doi.org/10.1111/1539-6924.00365
22. Kohonen, T.: Self-organized formation of topologically correct feature maps. Biological Cybernetics **43**(1), 59–69 (Jan 1982). https://doi.org/10.1007/BF00337288
23. Kourniotis, S., Kiranoudis, C., Markatos, N.: Statistical analysis of domino chemical accidents. Journal of Hazardous Materials **71**(1), 239–252 (2000). https://doi.org/https://doi.org/10.1016/S0304-3894(99)00081-3
24. Li, H., Xiao, D.Y.: Fault diagnosis of tennessee eastman process using signal geometry matching technique. EURASIP Journal on Advances in Signal Processing **2011**, 1–19 (2011)
25. Lohfink, A.P., Anton, S.D., Schotten, H.D., Leitte, H., Garth, C.: Security in process: Visually supported triage analysis in industrial process data. IEEE Transactions on Visualization and Computer Graphics **26**, 1638–1649 (4 2020). https://doi.org/10.1109/TVCG.2020.2969007
26. Müller, M.: Dynamic Time Warping, pp. 69–84. Springer Berlin Heidelberg (2007). https://doi.org/10.1007/978-3-540-74048-3
27. Sakamoto, Y., Kuriyama, S., Kaneko, T.: Motion map: Image-based retrieval and segmentation of motion data. In: Symposium on Computer Animation (2004)
28. Schmidl, S., Wenig, P., Papenbrock, T.: Anomaly detection in time series: a comprehensive evaluation. Proceedings of the VLDB Endowment **15**(9), 1779–1797 (2022)
29. Steiger, M., Bernard, J., Mittelstädt, S., Lücke-Tieke, H., Keim, D., May, T., Kohlhammer, J.: Visual analysis of time-series similarities for anomaly detection in sensor networks. Computer Graphics Forum **33**(3), 401–410 (2014). https://doi.org/https://doi.org/10.1111/cgf.12396
30. Sun, W., Paiva, A.R.C., Xu, P., Sundaram, A., Braatz, R.D.: Fault detection and identification using bayesian recurrent neural networks. Comput. Chem. Eng. **141** (11 2019). https://doi.org/10.1016/j.compchemeng.2020.106991
31. Vílchez, J.A., Sevilla, S., Montiel, H., Casal, J.: Historical analysis of accidents in chemical plants and in the transportation of hazardous materials. Journal of Loss Prevention in the Process Industries **8**(2), 87–96 (1995). https://doi.org/https://doi.org/10.1016/0950-4230(95)00006-M
32. Wagner, D., Michels, T., Schulz, F.C., Nair, A., Rudolph, M., Kloft, M.: Timesead: Benchmarking deep multivariate time-series anomaly detection. Transactions on Machine Learning Research (2023)
33. Ward, M.O., Guo, Z.: Visual Exploration of Time-Series Data with Shape Space Projections. Computer Graphics Forum (2011). https://doi.org/10.1111/j.1467-8659.2011.01919.x
34. Yin, S., Ding, S.X., Haghani, A., Hao, H., Zhang, P.: A comparison study of basic data-driven fault diagnosis and process monitoring methods on the benchmark tennessee eastman process. Journal of Process Control **22**, 1567–1581 (10 2012). https://doi.org/10.1016/j.jprocont.2012.06.009
35. Zhao, J., Chevalier, F., Balakrishnan, R.: Kronominer: Using multi-foci navigation for the visual exploration of time-series data. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (2011)
36. Zhao, J., Chevalier, F., Pietriga, E., Balakrishnan, R.: Exploratory analysis of time-series with chronolenses. IEEE Transactions on Visualization and Computer Graphics **17**, 2422–2431 (2011). https://doi.org/10.1109/TVCG.2011.195

Appendix

A The Tennessee Eastman Process Benchmark

In order to develop, test and present tools and visualizations for control systems in chemical plants, data of such a plant is required. Authentic data can be hard to come by, as companies are reluctant to share such sensitive information. This is where the *Tennessee Eastman Process* (TEP) benchmark data comes into play.

The TEP is a simulation of a chemical plant which imitates an actual plant owned by the Eastman Chemical Company. A sketch of the plant is seen in fig. 7.

Reactants named A, C, D and E in liquid and gaseous form as well as an inert E are piped in a reactor in order to produce products G and H, as well as the undesired byproduct F. The plant also features units which separate liquids and gasses, and a stripper that extracts the desired product from the stream. The remainder of this is recycled. A purge is used to maintain the balance.

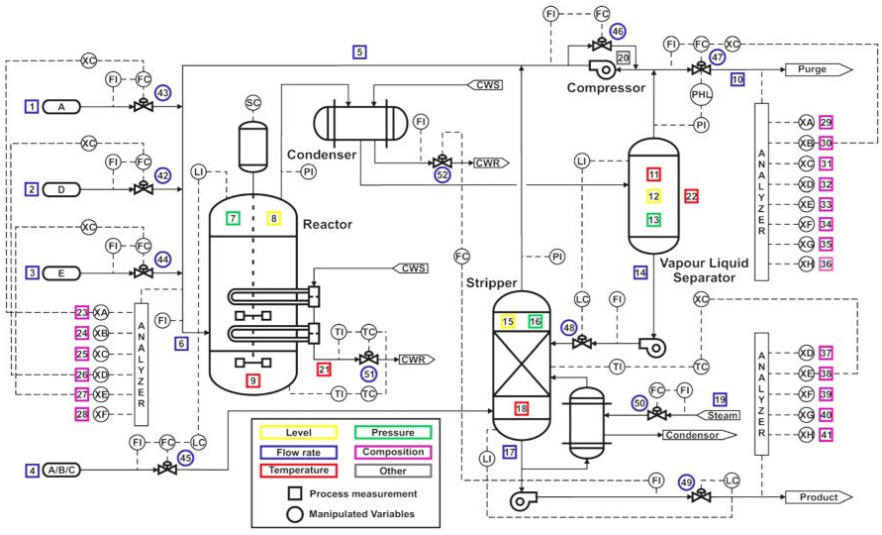

Fig. 7. Sketch of the TEP plant. The sensors are depicted by their number, with passive measured variables being squares and controlled variables being circles. Continuous lines are chemical flows, dashed lines control flows. Picture taken from Adeli et al [1].

The simulation contains data from 52 sensors, of which 41 represent what in a real process would be passive measurements, and the remaining 11 being controlled variables like valves. The system features automatic controls, the output

of several of the measured variables being used to steer some of the controlled variables. These simulated sensors represent what we would be able to perceive in a real control process, but they are not all variables that the actual simulation uses. For example, the ratio of reactants A and C in the fourth inlet is a *hidden variable*.

The TEP data comes in one set of training data and one set of testing data, which each consist of a case 0 with the *normal operating conditions* (NOC) and further 21 other cases, each representing a fault. In both training and testing data, the faults use the same numbers. As such, one can use the training data in order to perform supervised learning.

For most of the faults, an explanation is already provided, that is which variables of the simulation were made to behave out of order. For example, the first fault is an unusual ratio of reactants A and C in the fourth inlet, the hidden variable we previously talked about. While a supervised algorithm might simply associate each fault this with a pattern, it requires a human expert to create an explanation. Some fault cases, are without description, namely cases 16 and 21.

The TEP data as well as the Fortran routines can be found in [18], the repository tied to the paper by Katser et al [19].

More detailed descriptions of the TEP can be found in [11, 1, 17].

B Case Study Pictures

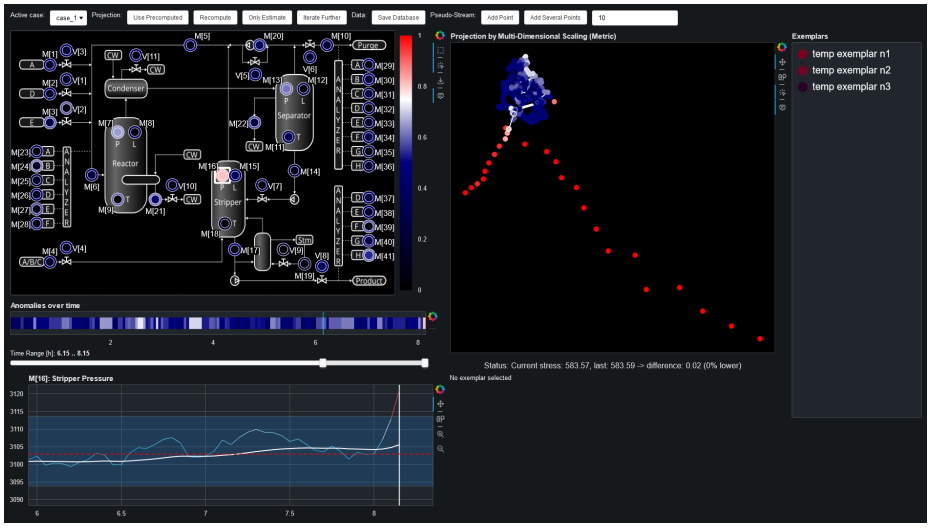

Fig. 8. The case study at 8:09. The light pink color of the stripper pressure node in the process flow diagram gives us an early hint of an anomaly.

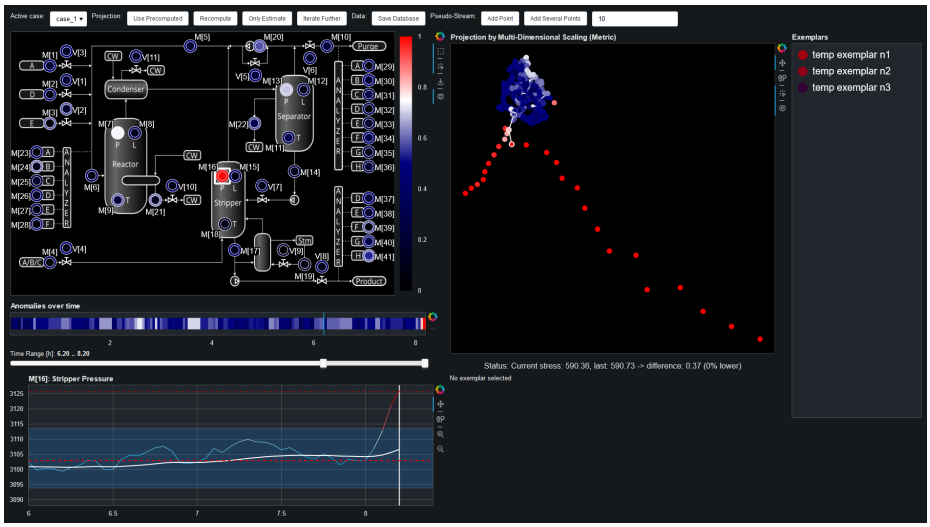

Fig. 9. The case study at 8:12. The views now clearly indicate an anomaly, affecting mostly the pressure of the stripper for now.

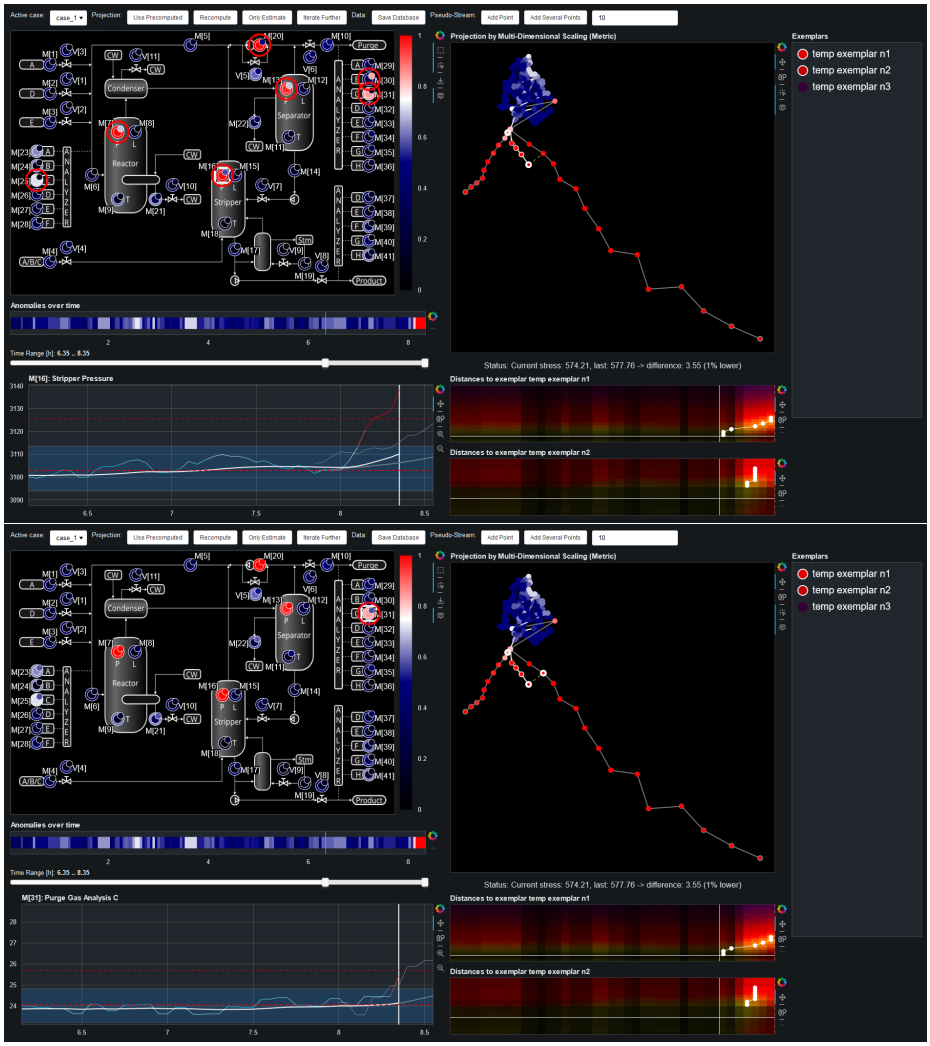

Fig. 10. The case study at 8:21, a time point in the second known fault case selected in the upper picture, one in the first case in the lower. In the process flow diagram, we see several mismatches for the former, but only one in the latter. For that, the variable view reveals that this is due to a latency. Both projection view and matrix view also indicate that the first case is the better match.

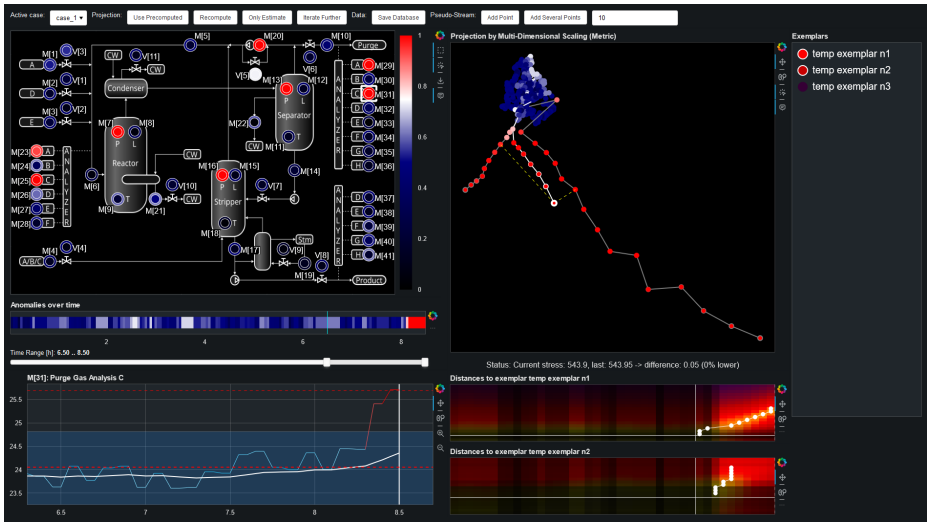

Fig. 11. The case study at 8:30. The flow diagram reveals anomalies of two reactants in the reactor inlet and purge, with which a domain expert could explain the fault even without the match.