TRAIL'26 Workshop -TReC (Lausanne)

- Full Name of the team leader (Researcher)

Noémie Vlaminck – Cenaero – Senior ML Researcher

- Contact Email

noemie.vlaminck@cenaero.be

- Project Title

Understanding industrial process dynamics through AI: the case of a recycle reactor process

- Profile of the team leader(s) & Expected Team Composition

* Provide a brief description highlighting your expertise, your experience relevant to this project, and the profiles you are looking to recruit during the camp. (max 2000 signs)

Noémie Vlaminck: after completing her mechatronics engineering studies at the University of Mons, she worked as a machine learning engineer for more than eight years across various fields: computer vision, speech recognition, manufacturing data analysis and time-series forecasting.

We are looking for people interested in going beyond black-box models by incorporating explainability, interpretability, causality, or physics-based knowledge into their AI methods.

- **Have you already identified potential team members for your project?**
  Yes

- **List the team members you have identified and briefly describe their profiles/roles (e.g., expertise, affiliation, expected contribution). (max 2000 signs)**

Aysu Özden is a Research Engineer at Cenaero in the Machine Learning and Optimisation Group. She holds a PhD in multi-fidelity reduced-order modelling for the development of digital twins for combustion systems from Université Libre de Bruxelles and Université Catholique de Louvain.

Cyriac Delie is a PhD student at the University of Liège (Montefiore Institute of Electrical Engineering and Computer Science) as part of the LIMPID project. He previously worked on Dynamical Systems

Reconstruction at Cenaero in collaboration with Carmeuse and has now shifted his focus towards physics-based Machine Learning methods.

Zahra Zamanian is a PhD student at the University of Mons in the SECO (Systems, Estimation, Control and Optimisation) group, within the ARIAC project. She focuses on integrating models and artificial intelligence (AI) into Battery Management Systems (BMS) and their applications, including health-aware control in autonomous vehicles.

This project is supported by Professors Pierre Geurts and Vân Anh Huynh-Thu at the University of Liège and builds upon Cyriac's thesis.

- **Domain of Application**

  Industry 4.0
- **Scientific Theme**

  Physics-based models and digital twins

- Abstract (3000 signs max)
  *Concise summary of the project

Even today, many industrial processes remain sub-optimised. In practice, classical controllers such as PID are typically used to regulate an output based on a single manipulated variable, often neglecting system couplings and the process's inherent nonlinearities. As a result, there is strong potential for improvement through advanced control techniques, which could lead to significant economic gains.

However, modern industrial systems are becoming increasingly complex. They integrate sophisticated equipment, real-time data streams, and highly interconnected processes. Beyond this intrinsic complexity, such systems are influenced by numerous interdependent parameters that evolve over time and are subject to uncertainties, including variations in operating conditions or component failures. This complexity makes the design of advanced control strategies particularly challenging.

With the rise of machine learning, new approaches have emerged to model process dynamics, either through purely data-driven methods or by leveraging physical knowledge of the system. However, in industrial contexts, the adoption of such methods critically depends on trust: operators and engineers are unlikely to rely on models they cannot understand. Providing clear explanations of model behaviour is therefore essential for acceptance and deployment.

This project could investigate four complementary directions depending on the number of participants:

- Purely data-driven "black-box" models (typically deep learning). Given their limited interpretability, post-hoc explainability techniques will be used and adapted to better understand their predictions.
- The development of intrinsically interpretable models, with a focus on approaches that automatically identify mathematical expressions directly from data.

- The development of methods to infer causal models enabling the identification of cause-and-effect relationships between system variables.
- Hybrid approaches combine machine learning with physical knowledge, embedding physical laws such as differential equations or conservation principles directly into learning algorithms.

These approaches aim to improve the understanding of system dynamics, leading to better control, more accurate predictions, and early detection of drifts or anomalies. Ultimately, they enable the construction of digital twins capable of simulating the impact of previously unobserved actions, while integrating causal structures to account for multivariate interventions. By enhancing both predictive performance and interpretability, the project seeks to build trustworthy models that can be effectively adopted in real industrial environments.

As a case study, the well-known Tennessee Eastman process (TEP) will be considered. This benchmark represents a reactor-separator-recycle chemical system that is unstable in an open loop. The goal is to analyse its dynamics and systematically compare the different modelling approaches described above.

- Background information & Problem Statement -
  *Explain the problem you aim to address, the rationale behind your project, and how it fits into the broader research landscape. (3000 signs max)

Modern chemical plants are complex, nonlinear, and tightly interconnected systems where safe and efficient operation depends on reliable monitoring, diagnosis, and control. The Tennessee Eastman Process (TEP), consisting of five main units (reactor, product condenser, separator, recycle compressor, and stripper), is a widely used benchmark for developing and testing such methods [1-2].

Over the past three decades, extensive research has focused on control and monitoring of the TEP, including decentralized PID control [3], plant-wide Model Predictive Control (MPC) [4], and model-based input-output approaches [5,6]. More recently, open-source implementations and datasets have enabled the widespread application of machine learning and optimisation methods for fault detection and process monitoring [7-10]. While these approaches improve detection and classification performance, they often fail to provide insight into underlying process behaviour, limiting their industrial adoption where interpretability is essential.

This lack of interpretability is increasingly recognised as a barrier to adoption, leading to growing interest in explainable and trustworthy AI for process monitoring and fault diagnosis [11-12]. For example, interpretable approaches such as Sparse Identification of Nonlinear Dynamics (SINDy) learn simpler governing equations from data, making them attractive for industrial use. However, they struggle with noise, high dimensionality, and limited observability in large-scale systems [13].

In addition to interpretability, understanding causes is crucial for decision-making in industry. Causal modelling aims to identify cause-and-effect relationships to improve diagnosis and decision-making, but is difficult in systems like the TEP due to feedback loops, delays, and hidden variables [14-15].

Hybrid and physics-informed modelling have also emerged as an important research direction. Such models incorporate physical laws into data-driven learning, improving consistency and efficiency, yet their integration with interpretability and causal reasoning remains an open challenge for industrial systems [16].

Overall, black-box models provide high accuracy but low interpretability, while interpretable, causal, and physics-based models offer better insight but are harder to apply in practice. This project addresses the lack of systematic comparison and integration of these complementary approaches. Using the TEP as a benchmark, it aims to develop modelling frameworks that support trustworthy monitoring, anomaly detection, control, and digital-twin development. The goal is to move beyond purely predictive performance toward models that are accurate, interpretable, causally meaningful, and suitable for real industrial deployment.

- Project Objectives & Concrete Implementation
  * Describe exactly what you intend to implement during the 2-week Camp. Clearly define the tangible outputs (e.g., prototype, algorithm, proof of concept) and advancements you expect to achieve by the end of the event. **(6000 signs max)**

The increasing complexity of modern industrial systems makes their monitoring, prediction, and control particularly challenging. These systems involve nonlinear dynamics, strong variable couplings, and time-varying behaviours, often operating under uncertain conditions. This leads to sub-optimal control due to a limited understanding of the underlying process. In addition, although industrial processes generate large amounts of data, these datasets are often not readily usable for machine learning due to noise, missing values, and limited accessibility.

This project aims to develop machine learning approaches to model, predict, and better understand the dynamics of industrial processes. The focus will be on building models that are not only accurate but also interpretable and trustworthy, enabling their adoption in real industrial environments.

During the 2-week Camp, the project will focus on a benchmark industrial case study: the Tennessee Eastman process, a widely used simulated reactor-separator-recycle system known for its value in dynamical and anomaly detection studies. The objective is to learn the system dynamics from simulated data and evaluate different modelling strategies.

To achieve this, several complementary approaches could be implemented and compared:

- Data-driven models with explainability: Development of predictive models (e.g., DNN, RNN, Transformer, etc.) to forecast system behaviour based solely on historical data. Given their black-box nature, post-hoc interpretability techniques will be applied to identify key variables and better understand model predictions, improving transparency and trust.
- Intrinsically interpretable models: Exploration of approaches capable of identifying explicit mathematical relationships directly from data (e.g., symbolic regression, KANs, Temporal Fusion Transformer), providing direct insight into the system structure.
- Causal modelling: Development of methods to infer causal relationships between process variables, enabling identification of cause-and-effect mechanisms rather than simple correlations.
- Hybrid physics-informed models: Integration of physical knowledge into machine learning approaches, for instance through Physics-Informed Neural Networks (PINNs), embedding physical laws such as differential equations or conservation principles into the learning process. It could also be models based on the system control theory like state-space models.

The implementation will consist of benchmarking these approaches on the Tennessee Eastman dataset, comparing their predictive performance, interpretability, and robustness. Particular attention will be given to their ability to generalise and to provide actionable insights for process monitoring and control.

The expected outcome is a proof of concept demonstrating how different machine learning paradigms can be used to model complex industrial dynamics. The research targeted within this project will contribute, on its own scale, to the development of digital twins, enabling simulation of system behaviour under new operating conditions. Moreover, the identified methods will be useful in the framework of the LIMPID Win2Wal project led by ULiège with Cenaero as research partner and Carmeuse (lime industry) as supporting industrial partner. The objective of this project is to develop a set of methods based mainly on machine learning approaches for predicting the dynamics of complex systems representative of industrial processes.

- **Do you plan to deliver, as an outcome of your project, a reusable "brick" for the TRAIL Factory (https://factory.trail.ac/en/home_page) that could later be transferred and converted into a company process?**

  Yes

When mature enough, the codes developed during the Camp should be released on the TRAIL Factory. A link referencing the dataset could also be stored on the TRAIL Factory.

- Project Dataset

*Provide a comprehensive description of the dataset you plan to use. Include details such as its source, format, and any relevant metadata. If possible, supply a direct link for access or proof of availability (max 3000 signs)

The Tennessee Eastman Process (TEP) [1] is a open-source simulated industrial process meant to replicate the core attributes of real-life chemical engineering processes. It describes the refinement of five unnamed chemical inputs into two products and a by-product. The overall process is constructed using standard chemical engineering subsystems, namely a condenser, a compressor, a separator, a reactor and a stripper. It also allows simulation of different faults.

The TEP was introduced over 30 years ago and is frequently used in process control and fault detection benchmarks. A dataset, including faulty and fault-free runs of the simulation, is available on Kaggle [17] and provided as RData files with a pre-existing training and testing split. Both training and testing sets each contain 10500 (500 fault-free, 10000 faulty) time series of the 41 process variables (XMEAS) and 12 control variables (XMV). Training time series are 500 points long and test time series are 960 points long, corresponding to a sample every 3 minutes for 25 and 48 hours, respectively. This dataset has already been used in fault detection benchmarks.

Since we intend to use the dataset for a different purpose (physic-based modelling), the usage of faulty runs will need to be discussed. A new implementation of the TEP using MATLAB and its Simulink toolbox was recently developed [18] which allows a full and easy access to every simulation parameter and variable, as well as a comprehensible view of the process. This enables the generation of a custom-made dataset if deemed necessary.

While purely synthetic, the TEP remains a complex use case and delivers realistic results. For the purposes of the LIMPID project, it offers a needed transparency over the behaviours a physics-based or interpretable model will have to manifest. Indeed, when evaluating the modelling capacities of such architectures, it will be required to not only validate it quantitatively (using metrics) but also qualitatively, by extracting some of its properties and matching them to those of the original process.

- Detailed work plan

*Provide a clear timeline for the 2 weeks of the Camp. Outline key milestones, task distribution, and deliverables **(5000 signs max)**

A brainstorming session will be organized online before the beginning of the workshop to introduce the project in detail, to exchange ideas, and to organize the team (depending on the number of participants, a split off into groups of 2-3 people could be done).

During the workshop, the project will be structured over two weeks, combining data preparation, model development, and comparative evaluation. The work will be carried out collaboratively using Python and Jupyter Notebooks, with version control through a shared GitHub repository. Daily short meetings with the whole team will ensure coordination and progress tracking.

The first 1-2 days will be dedicated to data exploration providing that the project was already setup before the workshop through the brainstorming session. The dataset will be explored to understand variable distributions, correlations, and time dependencies. It will lead to Milestone 1 – a clean and structured dataset ready for modelling.

Thereafter, the model development will be done in parallel:

- Baseline and purely data-driven models: initial predictive models will be developed to establish performance baselines (e.g. linear regression, Random Forest, Gradient Boosting, LSTM, ...).
  Milestone 2: baseline predictive performance established.
- Explainability for black-box models: post-hoc explainability techniques will be applied to the trained models
  - Feature importance analysis
  - Sensitivity analysis
  - Local explanation methods (e.g. SHAP-like approaches if feasible)

  Milestone 3: First insight into model interpretability
- Advanced modeling and comparison – interpretable and symbolic models: development of intrinsically interpretable approaches
  - Symbolic regression to extract explicit equations
  - Exploration of interpretable architectures like KANs and Temporal Fusion Transformer for example.

  Milestone 4: Add models with deeper interpretability capabilities into the models' benchmark
- Causal modeling: implementation of methods to infer causal relationships between variables
  - Analysis of dependencies and potential causal graphs
  - Identification of key drivers in system dynamics

  This step aims to move beyond correlation toward understanding cause-effect mechanisms.

  Milestone 5: Preliminary causal structure of the system
- Hybrid physics-informed approaches: implementation of simplified physics-informed ML approaches (e.g. energy conservation constraints or PINN-inspired methods if feasible)

Milestone 6: first hybrid modeling results

Depending on the number of participants and their expertise, the focus may be placed on a subset of these development paths to ensure the project remains achievable within the timeframe of the

workshop. All approaches will be systematically compared according to predictive performance, interpretability and usability as well as robustness and generalization.

By the end of the 2-week Camp, the project will deliver:

- Python code (shared via GitHub) implementing the different modelling approaches and benchmarking pipeline.
- A comparative analysis of the methods, including performance metrics and interpretability results.
- Visualization plots illustrating system dynamics, model predictions, and key influencing variables.
- Final slides and a report summarizing methodology, results, and insights.

The project will result in a proof of concept demonstrating how different machine learning paradigms can model complex industrial systems while balancing accuracy and interpretability. It will also provide insights into the most suitable approaches for building trustworthy digital twins in industrial contexts. Depending on the results and their maturity, the work may lead to the preparation of a scientific publication and to a reusable brick on the TRAIL Factory.

Finally, the resources needed for the project are the following:

- A laptop equipped with a Python environment (>= 3.12) with at least the following packages is needed: SciPy, NumPy, Matplotlib, Pandas, Scikit-Learn, PyTorch and Jupyter Notebook.
- A remote access to a supercomputing infrastructure could be necessary. Researchers from Walloon universities can ask access to Lucia through the CECI. Instructions will be provided to participants.
- Good Internet connection.

- Bibliographic references

[1] Downs, J. J., & Vogel, E. F. (1993). A plant-wide industrial process control problem. *Computers & chemical engineering*, *17*(3), 245-255.

[2] Juricek, B. C., Seborg, D. E., & Larimore, W. E. (2001). Identification of the Tennessee Eastman challenge process with subspace methods. *Control Engineering Practice*, *9*(12), 1337-1351.

[3] McAvoy, T. J., & Ye, N. (1994). Base control for the Tennessee Eastman problem. *Computers & Chemical Engineering*, *18*(5), 383-413.

[4] Ricker, N. L., & Lee, J. H. (1995). Nonlinear modelling and state estimation for the Tennessee Eastman challenge process. Computers & chemical engineering, *19*(9), 983-1005.

[5] Sriniwas, G. R., & Arkun, Y. (1997). Control of the Tennessee Eastman process using input-output models. Journal of Process Control, *7*(5), 387-400.

[6] Tian, Z., & Hoo, K. A. (2005). Multiple Model-Based Control of the Tennessee– Eastman Process. Industrial & engineering chemistry research, *44*(9), 3187-3202.

[7] Lyu, N., Botcha, S., Kulkarni, E., Pagaria, S., Alves, V., Sunshine, E. M., & Kitchin, J. R. (2026). Benchmarking Machine Learning Fault Detection Methods on the Tennessee Eastman Process Dataset.

[8] Salahshoor, K., & Kiasi, F. (2008). Online statistical monitoring and fault classification of the tennessee eastman challenge process based on dynamic independent component analysis and support vector machine. IFAC Proceedings Volumes, *41*(2), 7405-7412.

[9] Yin, S., Ding, S. X., Haghani, A., Hao, H., & Zhang, P. (2012). A comparison study of basic data-driven fault diagnosis and process monitoring methods on the benchmark Tennessee Eastman process. Journal of process control, *22*(9), 1567-1581.

[10] Hartung, F., Franks, B. J., Michels, T., Wagner, D., Liznerski, P., Reithermann, S., ... & Kloft, M. (2023). Deep anomaly detection on tennessee eastman process data. Chemie Ingenieur Technik, *95*(7), 1077-1082.

[11] Bhakte, A., Pakkiriswamy, V., & Srinivasan, R. (2022). An explainable artificial intelligence based approach for interpretation of fault classification results from deep neural networks. Chemical Engineering Science, *250*, 117373.

[12] Jang, K., Pilario, K. E. S., Lee, N., Moon, I., & Na, J. (2023). Explainable artificial intelligence for fault diagnosis of industrial processes. *IEEE Transactions on Industrial Informatics*, *21*(1), 4-11.

[13] Brunton, S. L., Proctor, J. L., & Kutz, J. N. (2016). Discovering governing equations from data by sparse identification of nonlinear dynamical systems. Proceedings of the national academy of sciences, *113*(15), 3932-3937.

[14] Wang, S., Zhao, Q., Han, Y., & Wang, J. (2023). Root cause diagnosis for complex industrial process faults via spatiotemporal coalescent based time series prediction and optimized Granger causality. *Chemometrics and Intelligent Laboratory Systems*, *233*, 104728.

[15] Dewantoro, H., Smith, A., & Daoutidis, P. (2024). Causal discovery for topology reconstruction in industrial chemical processes. Industrial & Engineering Chemistry Research, *63*(26), 11530-11543.

[16] Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. Journal of Computational physics, *378*, 686-707.

[17] Averkiev S., Tennessee Eastman Process Simulation Dataset (2017). URL: https://www.kaggle.com/datasets/averkij/tennessee-eastman-process-simulation-dataset/.

[18] Vosloo, J., Uren, K. & Van Schoor, G. Complete and open Simulink model of the Tennessee Eastman process. (2024). URL: https://github.com/kennyuren/COSTEP.

- Does the project include multidisciplinarity between STEM & SSH?

No

- We confirm that the Team Leader will be present **for the full duration of TReC'26 if the project is selected (August 24th - September 4th, 2026, Lausanne, Switzerland)**

I agree and confirm

- Additional comments