Data-based modeling of slug flow crystallization with uncertainty quantification

The application of artificial neural networks (ANN) for modeling of dynamic systems has seen widespread adoption in recent years. For model-based control algorithms, ANNs have become invaluable for replacing first-principle models that are based on physical differential equations, particularly when these models are numerically challenging to optimize or when reducing computation time is essential for real-time feasibility. However, a notable drawback of feedforward ANNs is their inability to provide uncertainty measures for their predictions. **In this project, students will analyze and process measured trajectories from a dynamic system, train artificial neural networks (ANNs) to model the system's temporal behavior, and implement an uncertainty quantification framework to assess the reliability of their data-driven predictions**.

Task Description

The first part of this project involves identifying distinct dynamic processes within the dataset, as the chemical reactor used to generate the data can produce different products. Students must use unsupervised machine learning techniques to cluster trajectories by product type, creating clean training datasets for each process. During this analysis, particular attention must be paid to data quality issues, including outliers from sensor failures and systematic measurement biases.
Following data preprocessing, students will develop machine learning models to predict the dynamic behavior of each identified process. These models should implement a NARX (Nonlinear AutoRegressive with eXogenous inputs) architecture, capable of predicting the next state y_{k+1} based on historical measurements $[y_k, y_{k-1}, \ldots, y_{k-n}]$ and control inputs $[u_k, u_{k-1}, \ldots, u_{k-n}]$, as illustrated in Figure 1 [1].

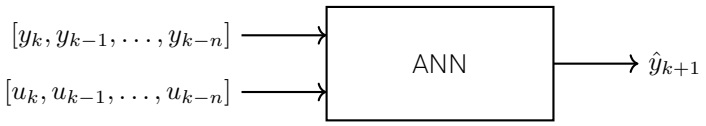

Figure 1: Schematic representation of the artificial neural network (ANN) used to predict the dynamic behavior of the SFC. The inputs to the ANN are the current and past measurements $[y_k, y_{k-1}, \ldots, y_{k-n}]$ and system inputs $[u_k, u_{k-1}, \ldots, u_{k-n}]$, and the output is the predicted measurement $\hat{y}_{k+1}$ for the next time step $k+1$. This kind of models are called Nonlinear AutoRegressive with eXogenous inputs (NARX) models in literature [1].

Once the data-based dynamic models are obtained, the next task will be the implementation of an uncertainty quantification framework to provide reliable prediction intervals for the trained models. For this purpose, Conformalized Quantile Regression (CQR) must be used [2]. The first step in this method is to generate new datasets that contain the approximation errors of the trained models, defined as $\varepsilon_{\text{App}} = \hat{y} - y$, where $\hat{y}$ denotes the ANN prediction and y the measurements. Next, the students must train quantile regressors to predict upper and lower quantiles of the approximation error datasets [2]. To train ANNs to predict quantiles of a distribution of datapoints, a special loss function called the pinball loss (Equation 1) can be used. The obtained quantile regressors should be able to predict the upper (0.95) and lower (0.05) quantiles of the prediction uncertainty of the data-based dynamic model based on the same inputs that are used to obtain y_{k+1}.

$$L_\tau(\hat{y}_k, y_k) = \begin{cases} \tau(y_k - \hat{y}_k) & \text{if } y_k \geq \hat{y}_k \\ (1-\tau)(\hat{y}_k - y_k) & \text{if } y_k < \hat{y}_k \end{cases} \tag{1}$$

where $\hat{y}_k$ is the prediction made by the ANN, y_k is the true label, and τ is the quantile, a value between 0 and 1. To account for the approximation error of the trained quantile regressors, the predicted inter-

vals must be conformalized using calibration data.
Since the quantile regressors themselves are subject to uncertainty, the final step of CQR uses a previously unused set of calibration data to compute correction factors for the predicted intervals. This conformalization step ensures the desired coverage probability of 90% (i.e., the true value falls within the prediction interval 90% of the time). Figure 2 demonstrates the effectiveness of Conformalized Quantile Regression (CQR) on a heteroscedastic regression task with $Y = \sin(X) + \varepsilon(X)$, where the noise variance increases with $|X|$. The comparison reveals how CQR's conformalization step corrects the undercoverage of standard quantile regression (89.33% vs. the nominal 90%) to achieve valid finite-sample coverage (92.67%), while maintaining adaptivity to the heteroscedastic noise structure. For a comprehensive explanation of the CQR methodology, we refer to [2].

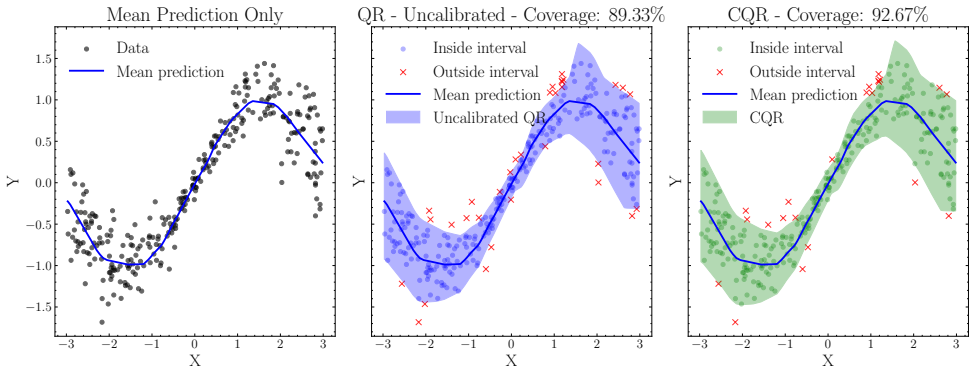

Figure 2: Demonstration of conformalized quantile regression (CQR) on heteroscedastic data. While standard quantile regression (middle) undercovers with 89.33% coverage for a 90% nominal rate, CQR (right) achieves the desired coverage of 92.67% through conformalization. Red crosses indicate points falling outside the prediction intervals. Note how both methods adapt interval widths to the heteroscedastic noise pattern, with wider intervals where uncertainty is higher.

The Data - Slug Flow Crystallization

An area where autonomous process operation can lead to considerable improvements is crystallization. Continuous crystallization promises more reliable and efficient processes, at the cost of more complex modeling and automation requirements. A promising apparatus in the field of continuous crystallization is the slug flow crystallizer (SFC), which is utilized for a broad range of products, especially in the pharmaceutical industry [3].
The general setup of such an SFC is shown in Figure 3. The crystallizer consists of an outer tube filled with water at temperature T_{TM} flowing with flow rate Q_{TM} to control the temperature of the inner tube. In the inner tube, liquid phase slugs that are fed to the reactor with a flow rate of Q_{PM} are separated by gas slugs. The parameter w_{cryst} represents the solid fraction within the liquid feed. The gas flow rate is denoted as Q_{air}. Crystal growth occurs within these liquid phase slugs, which are characterized by:

- the liquid temperature T_{PM}
- the product concentration within the liquid phase c_{PM}
- the particle size distribution of the formed crystals $n_{out}(L)$, where L is the diameter

These variables, together with the temperature in the outer tube T_{TM}, form the state vector that characterizes the system. Since the full particle size distribution $n_{out}(L)$ is impractical to use directly, it is represented by its 0.1, 0.5, and 0.9 quantiles (i.e., d_{10}, d_{50}, and d_{90}). Thus, the complete state vector becomes $\mathbf{y} = [T_{PM}, c_{PM}, d_{10}, d_{50}, d_{90}, T_{TM}]$. The system is controlled through the input vector

$\mathbf{u} = [Q_{PM}, Q_{TM}, Q_{air}, w_{cryst}, c_{in}, T_{PM,in}, T_{TM,in}]$, which includes the flow rates, the solid fraction in the feed, and the inlet conditions for concentration and temperatures.

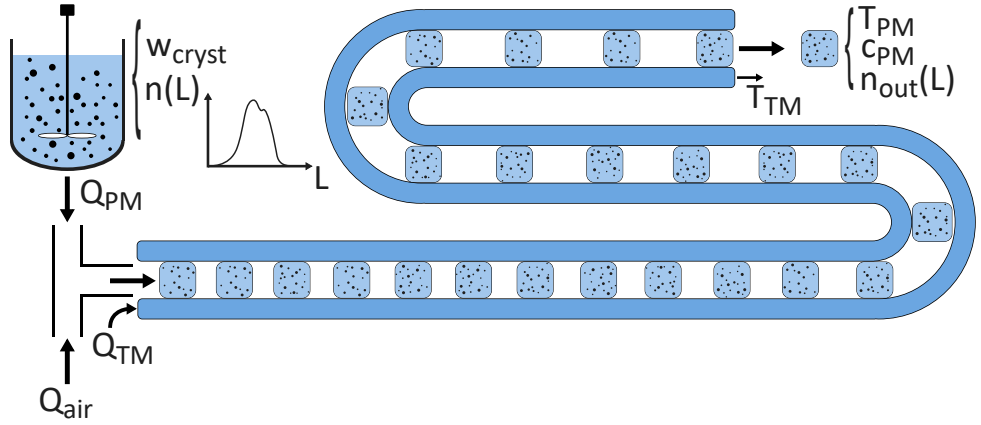

Figure 3: Schematic representation of a slug flow crystallizer (SFC)

Students are provided with measurements and process inputs from the continuous operation of such a slug flow crystallizer. This SFC has been used to produce several pharmaceutical agents, which differ mainly in the particle size distribution of the formed crystals $n_{out}(L)$. Each pharmaceutical agent corresponds to a distinct crystallization process, resulting in different dynamic trajectories within the dataset.
The particle size distribution measurements (d_{10}, d_{50}, and d_{90}) are inherently statistical in nature and therefore exhibit fluctuating behavior over time. This variability presents a significant challenge for process modeling and control. The developed in this project will enable robust predictions that account for both the inherent process variability and the model uncertainty, which is crucial for reliable process control in pharmaceutical manufacturing.
Disclaimer: In the collected data, measurements for the flow rates Q_{TM} and Q_{PM} are named MF_{TM} and MF_{PM}, respectively, and Q_{air} is named Q_g.

Mandatory tasks

The following tasks **have to be completed** in order to pass the project.

- Load, analyze and visualize the data
- Use unsupervised machine learning to identify different crystallization processes within the data
- Train ANNs to predict the dynamic behavior of the SFC
- Analyze and visualize the performance of the trained process models
- Generate the error datasets and use them to train quantile regressors for the 0.1 and 0.9 quantiles
- Use calibration data to conformalize the prediction intervals
- Analyze and visualize the performance of the CQR uncertainty quantification framework

Additional tasks

Below are **suggested** additional tasks to obtain good or excellent grades for the project. We want to emphasize that students are encouraged to come up with their own ideas for additional investigations and not all of the suggestions below must be included for an excellent grade.

- Test and tune the SFC models for open loop prediction: $ANN(\hat{y}_k, u_k) = \hat{y}_{k+1}$
- Investigate possible ways to propagate the predicted uncertainty information into the future
- Investigate alternative techniques for uncertainty quantification of time series predictions

Deliverables

The following materials have to be submitted **before the deadline** communicated via Moodle:

- **Recorded final presentation** (video screencast). The presentation must be **5-7 minutes** (for the entire group) and the file should not exceed **200 mb**. Highlight on the slides which group member(s) are responsible.
- **Written report** to present and discuss the obtained results. You must use the supplied template on Moodle and write no more than **3-4 pages** (for the entire group). Highlight which group member worked on which section.
- **Source code** of your project. Please ensure that the code is executable and optionally add a short explanation of the structure (readme).
- **AI usage disclosure** for your submitted files. If generative AI tools were used in your submission, please list: (a) the AI model(s) employed, (b) which portions of code/report were AI-assisted, (c) the nature of the assistance, and (d) two detailed examples showing your iterative process of using AI to develop and refine code for this project.
- **Beat-The-Felix competition**: To participate, your submission must include a directory named 'Beat-The-Felix' containing a Python script called main.py. This script must:
  - Process test data automatically when copied into the Beat-The-Felix directory
  - Perform all necessary preprocessing steps
  - Generate model predictions in open-loop mode
  - Output MSE and MAE performance metrics for each state variable

  Points will be awarded if your model performs better than or comparable to Felix's model for at least one error metric. The test data will be structured identically to the provided training data files. A dummy test file will be supplied to validate your main.py implementation. Benchmarks are provided below.

Please ensure that all formal conditions (e.g. page limits, highlight responsible author) are satisfied, as we will deduct points for significant violations. Please submit all deliverables via Moodle.

Benchmarks

Table 1: Mean performance metrics (R^2, MSE, MAE) for ANN-based slug flow crystallizer models across all identified clusters. Open-loop predictions were evaluated on test trajectories of 900 data points that were excluded from both training and validation sets. Open-loop predictions used the model's own outputs as inputs for subsequent predictions.

State	Open Loop		Closed Loop	
	MSE	MAE	MSE	MAE
c	2.472e-05	6.630e-04	4.808e-08	9.700e-05
T_PM	0.133	0.151	0.013	0.073
d50	1.895e-06	8.481e-05	1.438e-06	6.437e-05
d90	3.063e-07	5.572e-05	1.566e-06	8.005e-05
d10	4.578e-07	4.387e-05	1.823e-06	7.393e-05
T_TM	0.120	0.144	0.0139	0.0676

Responsible tutor

Please address questions to:

Name	Contact
Felix Brabender	felix.brabender@tu-dortmund.de

References

[1] Christopher M. Bishop. ***Pattern Recognition and Machine Learning***. Information Science and Statistics. New York: Springer, 2006. 738 pp. ISBN: 978-0-387-31073-2.

[2] Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. ***Conformalized Quantile Regression***. May 8, 2019. arXiv: 1905.03222 [stat]. URL: http://arxiv.org/abs/1905.03222 (visited on 11/02/2023). Pre-published.

[3] Maren Termühlen et al. "Continuous Slug Flow Crystallization: Impact of Design and Operating Parameters on Product Quality". In: ***Chemical Engineering Research and Design*** 170 (June 2021), pp. 290–303. ISSN: 0263-8762. DOI: 10.1016/j.cherd.2021.04.006. URL: https://linkinghub.elsevier.com/retrieve/pii/S026387622100160X (visited on 05/23/2025).