# Capturing prediction uncertainty in upstream cell culture models using conformal prediction and Gaussian processes

**Tien Dung Pham** TIENDUNG@UNIMELB.EDU.AU
**Uwe Aickelin** UAICKELIN@UNIMELB.EDU.AU
*School of Computing and Information Systems, The University of Melbourne, Victoria, Australia*

**Robert Bassett** ROBERT.BASSETT@CSL.COM.AU
*CSL Innovation Pty Ltd, Victoria, Australia*

**Editor:** Harris Papadopoulos, Khuong An Nguyen, Henrik Boström and Lars Carlsson

## Abstract

This extended abstract compares the efficacy of Gaussian process and conformal XGBoost regressions in capturing prediction uncertainty in simulated and industrial cell culture data.
**Keywords:** conformal regression, Gaussian process, cell culture modelling

## 1. Motivation

Capturing uncertainty in cell culture modeling is a non-trivial task for the biopharmaceutical industry due to the complex dynamics of the cellular bio-systems. However, most modeling studies have not applied a proper uncertainty quantification (UQ) framework to their designs, relying on bootstrapped prediction intervals to generate uncertainty estimates. A major challenge for UQ in cell culture modeling is that the sources of uncertainty are often indeterminate. Variations in data trends could stem from either input process parameters, measurement errors, or random noises (Lazic and Williams, 2021). To evaluate the efficacy of any UQ framework on real-life bioprocess data, a benchmark should first be obtained from simulated data with a controlled level of uncertainty.

UQ studies in cell culture modelling often use Gaussian processes (GP) as the standard model (Pham et al., 2023). Despite their high applicability, GPs are expensive to train. In this work, we propose using a XGBoost-based cross-conformal regressor (CP) as an alternative UQ model to GP. As the CP prediction intervals are conformalised to the "difficulty" of the input, they can potentially capture how much the input process parameters contribute to the bioprocess variations. To the best of our knowledge, our study is one of the few applications of conformal prediction in the cell culture domain.

## 2. Experimental Designs and Results

We designed two experiments to investigate the performance of XGBoost CP (XGB+CP) versus GP in predicting viable cell density (VCD) - a key measure in cell culture. Each observation (sample) consists of a 13-point time-series, with the first data point representing the initial input condition and the other 12 data points representing daily measurements. In Experiment 1, we used the initial experimental conditions to predict the 12-point trajectory of each measure. In Experiment 2, we gradually added daily measurements to the models

as new predicting features to examine whether obtaining more data throughout the process affects the UQ frameworks.

For each Experiment, we obtained results using 3 different datasets: 2 simulated datasets with 500 and 1500 time series; and a real-life bioreactor dataset from a biopharmaceutical manufacturing partner consisting of 91 extremely noisy time series.

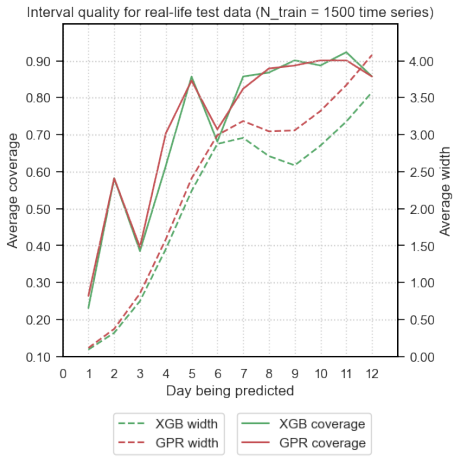


Figure 1: Experiment 1 results on real data

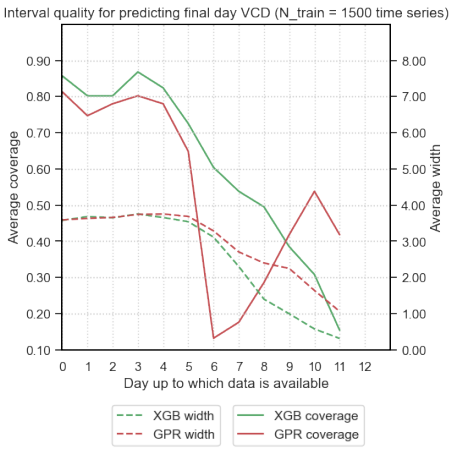


Figure 2: Experiment 2 results on real data

For Experiment 1, on simulated training datasets, XGB+CP resulted in 21% larger intervals compared to GP, but achieved the nominal coverage of 95% on each day, whereas GP only achieved 95% coverage for 2 of 12 predictions. When applied to the real dataset (Fig. 2), the XGB+CP outperformed GP with lower MAPE and widths from day 6 to day 12, while maintaining similar coverages. The underperformance of both models in predicting earlier days might be due to the scale of the noise in the data, as the models, trained on simulated data, were not adjusted to the heteroscedasticity of real-life data. Results for Experiment 2 showed that if we retrain the models as more experimental data comes through, the interval widths will contract at the cost of significant drops in coverage (Fig. 2). Lastly, it was observed that the training time of XGB+CP scales much better compared to GP. As the dataset increased 3-fold in size, the average XGB+CP training time merely doubled from 1.71 to 3.22 seconds, whereas GP increased 8-fold from 0.48 to 3.9 seconds. In summary, our Experiments showed that with proper tuning, XGB+CP is a potential candidate for capturing uncertainty in cell culture modelling.

## Acknowledgement

This research was supported under the Australian Research Council's Industrial Transformation Research Program (ITRP) funding scheme (project number IH210100051). The ARC Digital Bioprocess Development Hub is a collaboration between The University of Melbourne, University of Technology Sydney, RMIT University, CSL Innovation Pty Ltd, Cytiva (Global Life Science Solutions Australia Pty Ltd) and Patheon Biologics Australia Pty Ltd.

## References

Stanley E Lazic and Dominic P Williams. Quantifying sources of uncertainty in drug discovery predictions with probabilistic models. *Artificial Intelligence in the Life Sciences*, 1:100004, 2021. ISSN 2667-3185. doi: https://doi.org/10.1016/j.ailsci.2021.100004.

Tien Dung Pham, Chaitanya Manapragada, Yuan Sun, Robert Bassett, and Uwe Aickelin. A scoping review of supervised learning modelling and data-driven optimisation in monoclonal antibody process development. *Digital Chemical Engineering*, 7:100080, 2023. ISSN 2772-5081. doi: https://doi.org/10.1016/j.dche.2022.100080.