

**The
Alan Turing
Institute**

Data Study Group Final Report: AstraZeneca

10-14 December 2018

Machine learning for enhanced
understanding of cell culture
bioprocess development



<https://doi.org/10.5281/zenodo.3367412>

This work was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1

Contents

1	Executive summary	2
1.1	Challenge overview	2
1.2	Data overview	2
1.3	Main objectives	3
1.4	Approach	3
1.5	Main conclusions	5
1.6	Limitations	5
1.7	Recommendations and future work	6
2	Quantitative problem formulation	7
3	Data overview	8
3.1	Dataset description	8
3.2	Variable description	9
3.3	Data quality issues	9
3.4	Exploratory data analysis	12
4	Experiments and results	23
4.1	Inference for identifying key controllable features	23
4.2	Prediction of drug quantity via feature extraction from partial time-series	24
4.3	Classification of low-yield processes via outlier detection from partial time-series	27
5	Future work and research avenues	28
5.1	Remedying data quality issues	28
5.2	Improving tried out approaches	29
5.3	Further research avenues	29
	References	31
	Team members	32

1 Executive summary

1.1 Challenge overview

A central part of the biopharmaceutical manufacturing process are bioreactors, i.e. sterile containers in which genetically modified cell cultures are grown in a controlled environment to produce drugs or intermediate compounds. One of the major – and mostly unsolved – challenges is to understand how the vast array of bioreactor settings influence amount and quality of the drugs produced.

Biopharmaceutical drug manufacturing is a very common but also expensive process, hence small improvements in efficiency have large impacts on manufacturers' operations and the availability of public healthcare. One major reason for currently high costs of biopharmaceutical drugs is that the manufacturing process based on bioreactors is highly complex and difficult to understand using the more classical engineering tool set.

AstraZeneca is driving efforts to make full use of the available bioreactor sensor data to enhance process control in a smart and automated way. This report presents the results from the Data Study Group, a week-long collaboration between AstraZeneca and The Alan Turing Institute, with the main goal to use modern data science and machine learning techniques in order to find out if one can accurately predict the amount of drugs produced and discover key controllable variables.

1.2 Data overview

AstraZeneca provided granular multivariate panel data from a typical series of bioreactor experiments, i.e. repeated measurements over time of several variables on multiple bioreactor runs. A bioreactor run covers the end-to-end process from initial inputs to final outputs.

In total, data was available for 168 bioreactor runs, each with sensor data of up to 30 different variables observed at roughly 450 time points throughout the cell culture process.

Drug quantity measurements, which is besides drug quality the most

important process output, were available at four time points, with the last one measuring the final amount of drugs produced.

Additional data was available on the bioreactor parameter settings for process control and the cell cultures used in each experiment.

For more details and a discussion of data issues, see section 3.

1.3 Main objectives

Given the overall goal to enhance drug production via modern data science and machine learning techniques and the available data, the main objectives were as follows:

1. Infer controllable features that can be leveraged to enhance current process control mechanisms in terms of the quantity produced,
2. Predict the final amount of drugs produced from partial time-series of sensor data, i.e. time-series sensor data observed up to a given cut-off,
3. Classify which bioreactor runs present the highest risk for unfavourable process outcomes from partial time-series of sensor data observed up to a cut-off point, where unfavourable process outcomes are defined in terms of low drug yields.

1.4 Approach

1.4.1 Inference with linear mixed effect models

For the first objective, we used statistical modelling and inference to discover key controllable variables of the manufacturing process from the available sample data.

For this purpose, we estimated a linear mixed effect model with varying slope and intercept for each bioreactor run using the complete time-series sensor data coupled with drug quantity measurements interpolated to all time points of the sensor data.

1.4.2 Supervised learning with time-series/panel data

Both the second and third objective represent prediction tasks, with the distinction that the second one is a classification task predicting a binary process outcome variable while the third one is a regression task predicting the exact quantity of drugs produced.

The standard framework for making such predictions is supervised learning. We used supervised learning algorithms to infer a function from the data that maps the time-series sensor data to the process output, so that the inferred function can then be used to make output predictions from new sensor data.

In particular, we used a combination of time-series feature engineering and common regression and classification algorithms. Feature engineering is necessary in this setting to extract useful information from the time-series sensor data in a format that can be used as inputs to the algorithms. We applied three feature engineering approaches:

- Time-series as features using each measurement point as a separate feature ignoring the temporal ordering of observations,
- Manual feature construction based on MedImmune's domain expertise and initial exploratory data analysis including lagged correlation analysis and outlier detection,
- Automatic feature extraction based on various time-series analysis and decomposition techniques using available off-the-shelf tools.

With the resultant sets of time-series features, we then trained the algorithms on a subset of the available bioreactor runs and evaluated their predictions on a held out test set.

In order to find out if it is possible to make accurate prediction early in the process, we evaluated the predictive performance at different cut-off points, limiting the amount of time-series sensor data used in feature engineering and fitting of the algorithms.

Tried out algorithms include random forest and gradient-boosting machines. In addition, we examined whether principal component analysis (PCA) as a common feature reduction technique helps improve predictive performance.

1.5 Main conclusions

The main conclusion is that the tried out data science and machine learning techniques helped to better understand and predict process output.

1. For the first objective, the estimated mixed effect model enabled us to identify injected oxygen as a key variable that shows statistically significant positive correlation with the amount of drugs produced.
2. For the regression task, we found that our tried out algorithms reduce the mean squared error between the predicted values and actual outcome in the test set by roughly 70% compared to a naive baseline. This represents a reduction of the mean relative error from roughly 30% to 20% around the actual values, indicating that this approach is promising.
3. For the classification task, we achieved an accuracy score of 94% on the test set compared to a naive baseline score of 89%.

1.6 Limitations

- Since the challenge demanded a specific combination of data science skills and domain knowledge, we focused on the fundamental tasks of inference and prediction rather than the more advanced topic of process control. Statistical process control not only requires to predict the process output from sensor data, but also to estimate the effect of process adjustments on output (see e.g. [1]).
- Due to the technical constraints, we concentrated our efforts on the data from only one type of bioreactor system, namely the micro-scale AMBR system mainly used in early process development, leaving open the question of how well our results generalise to other settings (i.e. other bioreactors, unseen bioreactor parameter configurations and production processes at larger scale).
- While the available data was rich in its temporal dimension, only relatively few bioreactor runs were available. Consequently, more independent samples are needed to statistically corroborate these

results.

- The estimated performance difference for the classification task is not statistically significant at the 95% level and can be plausibly explained by chance.
- Found correlations between sensor data and process output do not imply any causal relationship between them.

1.7 Recommendations and future work

Our work suggests several avenues for future research:

- As a general recommendation for future data scientific projects, it will be helpful to provide data in more consistent, easily accessible format and to provide more data on independent bioreactor runs,
- Results need to be statistically corroborated, including tests whether found performance differences between the tried out algorithms and a reasonable industry baseline are statistically significant and further investigating how well results generalise to processes at industrial scale,
- Improve and extend current approaches (e.g. by trying out other feature engineering techniques, specialised time-series prediction algorithms and model selection methods for tuning algorithms),
- Investigate other tasks of interest, including predictions of drug quality and estimation of forecast and transfer functions from the sensor data,
- Ultimately, deploy machine learning methods in the manufacturing process, with the goal to enhance decision making via early detection of unfavourable runs and re-design the bioreactor control system from static parameter configurations to a responsive control regime which is automatically adjusted by machine learning algorithms.

2 Quantitative problem formulation

In order to find opportunities to improve drug production, the pharmaceutical industry has implemented platform processes, i.e. standard bioreactor systems that allow to closely monitor and compare production runs.

Platform bioreactors collect sensor data of many variables throughout the process in almost real time. By contrast, process output measurements require more lengthy chemical analysis and only become available after the end of the process. This includes drug quantity, which is besides drug quality the most important process output variable.

In this context, reliable process output predictions are crucial for decision making, for example by signalling to adjust control parameters or terminate the process. Consequently, early detection is crucial and predictions will be more helpful the earlier one can reliably make them in the process.

One of our main objectives was therefore to find out if one can use partial time-series sensor data observed up to a cut-off point to predict the final amount of drugs produced. A closely related objective was to use partial time-series data to classify which processes are at high risk of unfavourable process outcomes, where an unfavourable outcome was defined as a process with final drug quantity below a given threshold.

The standard methodological framework for making such predictions is supervised learning [4]. However, the standard framework requires data to be formatted in a tabular form, which time-series and panel data does not naturally fit into. Consequently, it is necessary to either use more specialised methods or to transform the data first, so that one can then make use of the common algorithms. Either way, this entails a number of additional steps and choices in the prediction workflow. Here, we followed the more common second approach, reducing the data to the required tabular format using various feature engineering techniques.

Another central objective was to identify key controllable variables that influence the final amount of drugs produced. For this task, the standard approach is to estimate interpretable statistical models which can be

used to infer important relationships between sensor data and process output from the available sample data. For this purpose, we used panel-specific linear mixed effect regression models [3].

These objectives already posed considerable data scientific challenges and we therefore focused on solving them first, leaving the question of enhanced process control design to future research. Ultimately, the goal is to re-design the bioreactor control system from acting on static parameter configurations into a responsive regime which uses predictions from machine learning algorithms to adjust the cell culture process in a smart and automated way in real time.

3 Data overview

3.1 Dataset description

AstraZeneca provided multivariate panel data from a typical series of bioreactor experiments, i.e. repeated measurements over time of different variables on multiple bioreactor runs. A bioreactor run covers the end-to-end process from initial inputs to final outputs.

The available data came from two types of bioreactor systems, the micro-scale AMBR system and small-scale DASGIP system. Both are mainly employed in biopharmaceutical process development as opposed to industrial-scale production. For the AMBR system, there was data for 144 runs, for the DASGIP system for 24 runs.

For each bioreactor run, the time-series sensor data is collected through repeated measurements over the duration of the process of several variables. A list of key variables and their descriptions can be found in table 1.

In each run, cell cultures are grown for approximately 14 days and measurements were available for every hour, resulting in observations at roughly 450 time points. Note that this data had already been aggregated. Original measurements are taken at higher but not necessarily synchronised frequencies. Additional information was available on the cell cultures used in each run.

The output variable of interest was the total amount of drugs produced (known as titre), which is measured at 4 time points during the process, approximately at day 8, 10, 12 and 14, with the last one capturing the final process output. While these measurements are taken during the process, they only become available after the end of the process due to the required lengthy chemical analysis.

Additional data was available for the configuration of control parameters, called set-points. Based on these setting, bioreactors automatically adjust certain variables (e.g. stirring speed or dissolved oxygen) via preprogrammed proportional-integral-derivative controllers to maintain favourable conditions for the cell cultures.

3.2 Variable description

Table 1 lists key variables and their descriptions.

Airo and co are both flow variables measuring the injection of oxygen and carbon dioxide into the bioreactor. In order to capture the total inflow of oxygen and carbon dioxide up to any given time point in the process, we calculated their cumulative sums over time. Likewise, we computed the cumulative sum of the measured do-integral-error to capture the total deviation of the cell culture process.

3.3 Data quality issues

In our analysis, we noticed the following data quality issues:

1. **External validity.** There are a number of external validity issues:
 - The data comes from bioreactor systems used in process development, not industrial-scale production,
 - The drug produced and cell cultures used in each bioreactor run is not always the same,
 - There was data for only a subset of control parameter configurations of those used in real-world manufacturing processes,

Table 1: Description of key variables

Variable	Type	Frequency	Availability	Description
titre (amt)	output	day 8, 10, 12, 14	after the end of the run	Quantity of drugs produced (mg per litre)
airo	sensor	hourly	real time	Amount of oxygen injected to the bioreactor (ml per minute)
co	sensor	hourly	real time	Amount of carbon dioxide (CO2) injected to the bioreactor in (ml per minute)
do	sensor	hourly	real time	Dissolved oxygen (percentage)
do-integral-error	sensor	hourly	real time	Deviation of do from its set-point
ph	sensor	hourly	real time	pH value
glucose	sensor	hourly	real time	Glucose level (g per litre)
lactate	sensor	hourly	real time	Lactate level
temperature	sensor	hourly	real time	Temperature in block of 12 bioreactors for AMBR system (centigrade)
total cells	sensor	hourly	real time	Total number of viable cells
airo-cumsum	derived	hourly	real time	Cumulative sum of airo over time
co-cumsum	derived	hourly	real time	Cumulative sum of co over time
do-integral-error-cumsum	derived	hourly	real time	Cumulative sum of do-integral-error over time

- The available bioreactor runs were pre-selected based on particular temporal profiles of the sensor data and output (selection bias),
- Bioreactor runs may not be entirely independent, for example some dependencies may be introduced due to shared bioreactor control mechanism.

This raises important questions about the representativeness of our results for the real-world process of interest. Without further statistical investigation, our results do not generalise to the real-world manufacturing setting.

2. **Sample size of bioreactor runs.** There were relatively few bioreactor runs for AMBR bioreactors, and even less for DASGIP. We therefore focused on the AMBR data, even though DASGIP systems record more sensor measurements and results from DASGIP data may be easier to scale up to industrial production.
3. **Missing values.** For many sensor data variables, observations for time points were missing. For different batches of bioreactor runs, entire variables were missing. Consequently, these measurements could not be used when making predictions without additional preprocessing steps such as data interpolation.
4. **Data format and organisation.** There were a number of issues with the data format and organisation:
 - For each batch of bioreactor runs, the sensor data was stored in separate files.
 - For the DASGIP system, data were stored in excel files and spread across tabs and pages, some of which also had charts.
 - Column headers (i.e. variables names) were found to be inconsistent across files. For example, there is a “speed” measure in the “ambr414.csv” file that is missing from the other two.
 - Column names included special characters and empty spaces.

As a consequence, understanding and loading the data was difficult, and data preprocessing and cleaning took a significant amount of

effort.

3.4 Exploratory data analysis

Due to required data cleaning and preprocessing, we focused our exploratory data analysis on only a subsample of 48 AMBR bioreactor runs (AB414).

3.4.1 Data visualisation

To better understand the temporal dynamics of the cell culture process, we show plots for a selection of key variables over time.

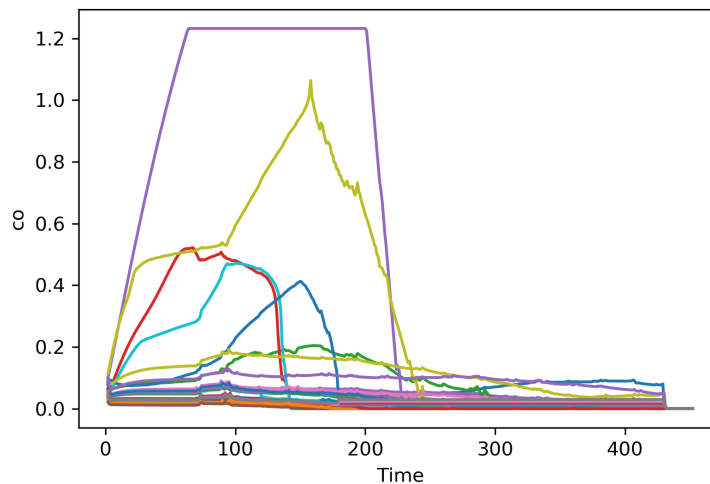


Figure 1: Measurements of carbon dioxide (co) (ml) over time (hourly intervals) for different bioreactor runs.

From the data visualisation, anomalies could be detected in the sensor data. These anomalies manifested as either high airo values, high co values or large pH integral errors.

In addition, the plots indicate that for certain variables certain time periods show higher variation in measurement values than others. For example, figure 1 shows that most variation of the injected carbon dioxide takes

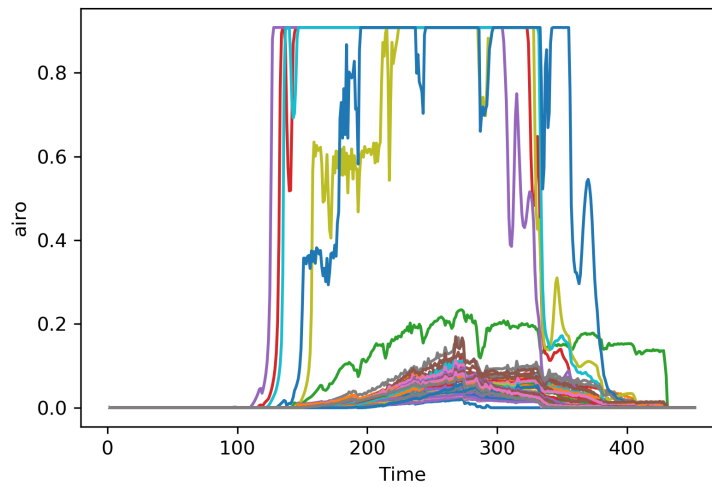


Figure 2: Measurements of injected oxygen (airo) (ml) over time for different bioreactor runs.

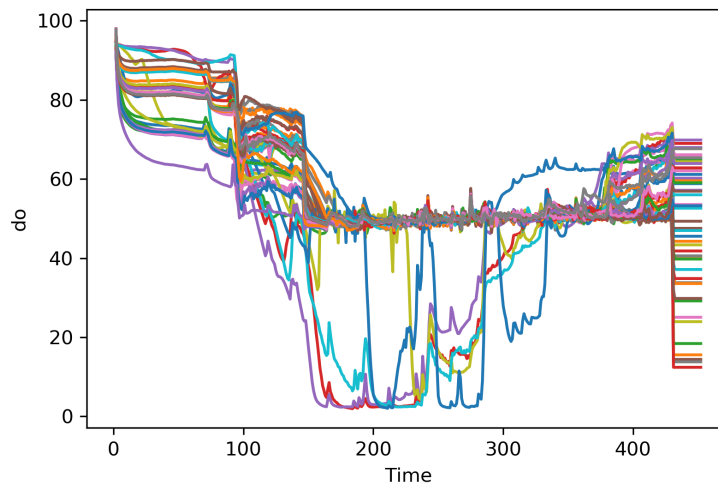


Figure 3: Measurements of dissolved oxygen (do) (percentage) over time for different bioreactor runs.

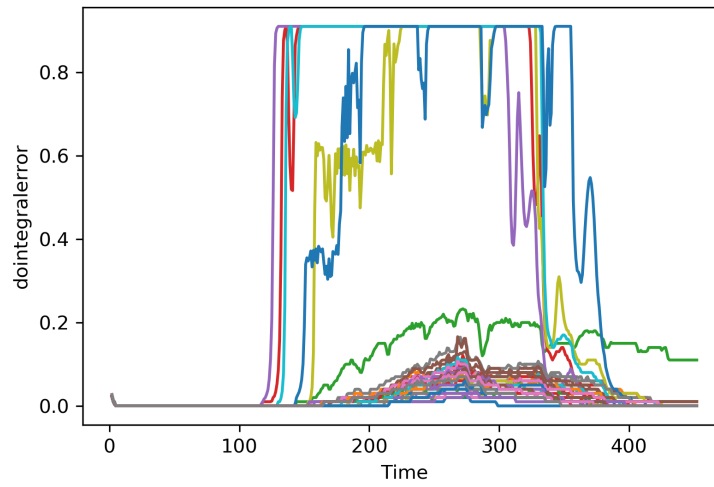


Figure 4: Measurements of the deviation of dissolved oxygen from the set-point (dointegralerror) over time for different bioreactor runs.

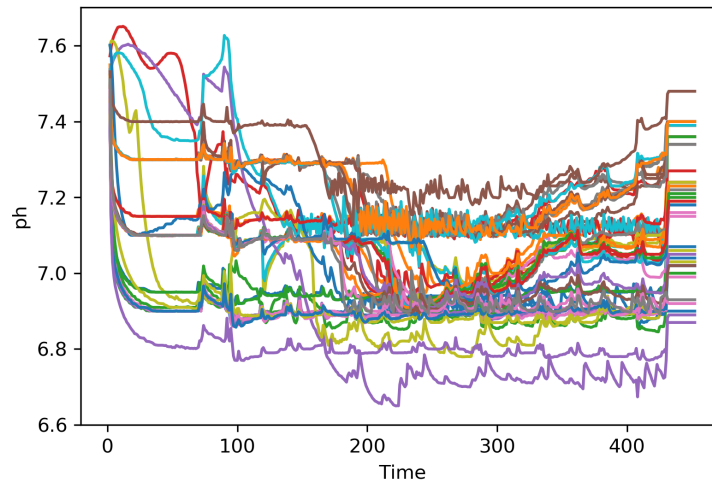


Figure 5: Measurements of the pH-value (ph) over time for different bioreactor runs.

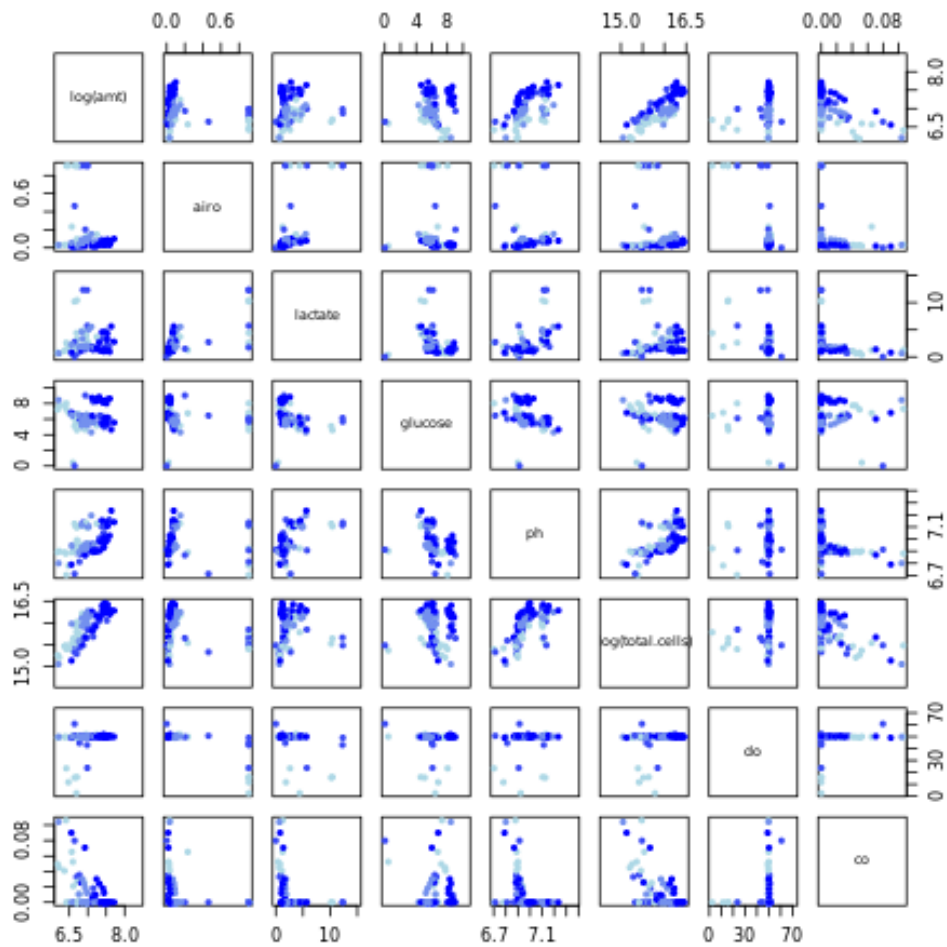


Figure 6: The paired scatter plot shows the amount of drugs (amt), airo, lactate, glucose, ph and total number of cells at the 4 measurement time points of titre (amt). To highlight the change of variables over the four time points, data points are coloured by the time point of the titre measurements, with darker colours representing later time points. To scale the data, the logarithm of the titre measurements and total number of cells is shown. From visual inspection, the plot suggests a positive linear association between the total number of cells and titre.

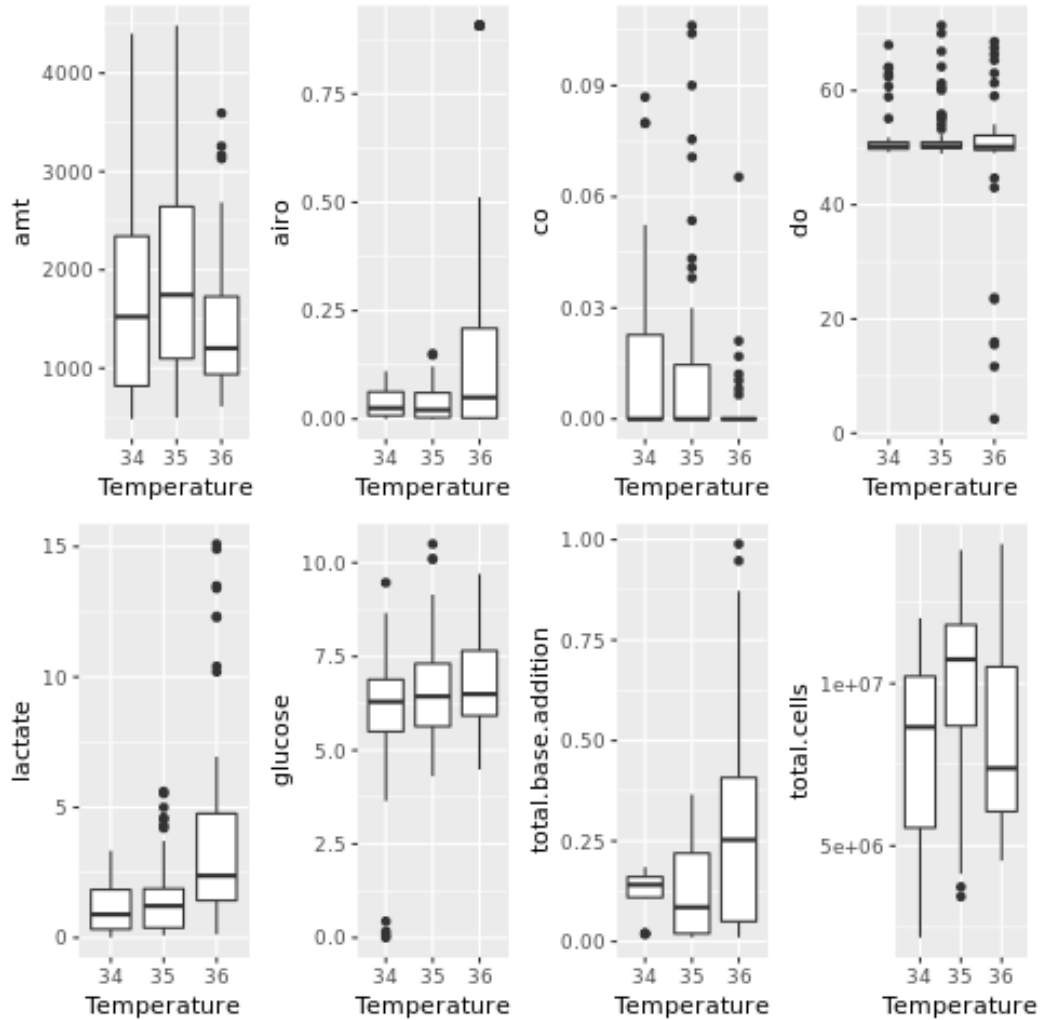


Figure 7: The boxplots show the distribution of amount of drugs (amt), airo, co, do, lactate, glucose, total base addition and total number of cells for different temperature values. Note that although temperature is not fixed, it only changes within a narrow interval around a certain fixed point and thus is here treated as fixed. The plots show how the variables are distributed from each centred temperature point. For example, the lactate and glucose tend to increase from the higher temperature. Also the range of lactate gets wider from the amount of drugs, the lactate and the total number of cells.

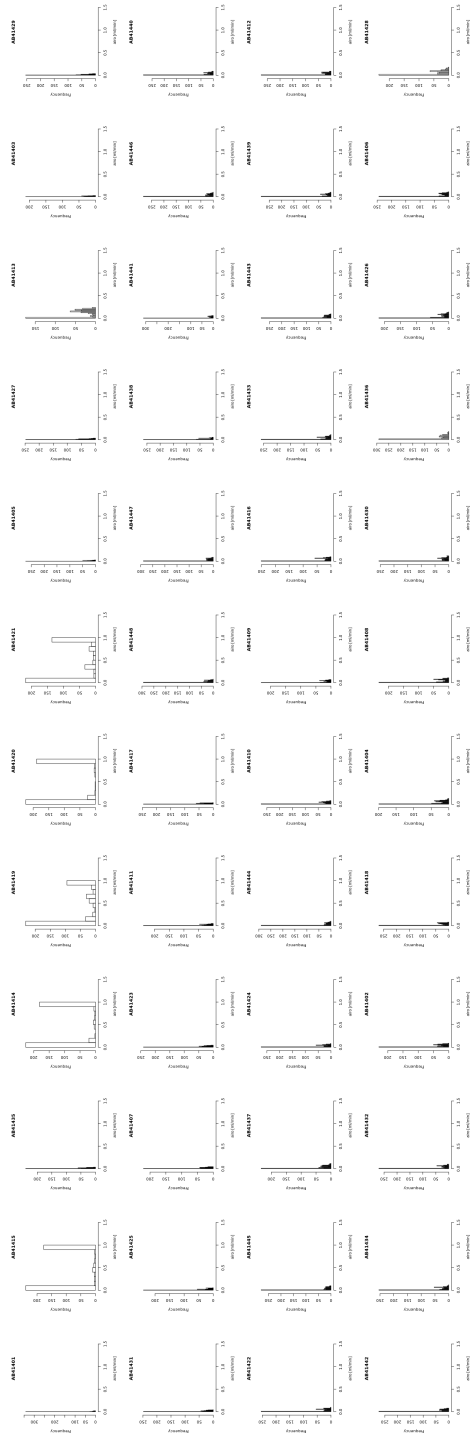


Figure 8: Histograms showing the frequency of observed measurements for airo (oxygen flow, ml/min) for the entire time series (ignoring their temporal ordering) for each bioreactor run. The histograms are sorted from runs producing the lowest amount of titre (top left) to runs producing the highest amount of titre at the end of the experiment (bottom right).

place from the start of the process until roughly 200 hours. After 200 hours, the processes appear to behave relatively similarly in terms of injected carbon dioxide. Similarly, figure 4 shows that not much variation occurs during the first 100 hours of the process with regard to the deviation of dissolved oxygen from its set-points.

While looking at univariate and bivariate distributions enabled us to gain first insights, many of the variables are inter-related and jointly determine the final amount of drugs.

3.4.2 Correlation analysis between titre intercept/slope and lagged sensor data

After the initial exploratory analysis, drug production was observed to grow linearly from approximately day 8 to day 14, but with varying initial value and slope, as can be seen in figure 9.

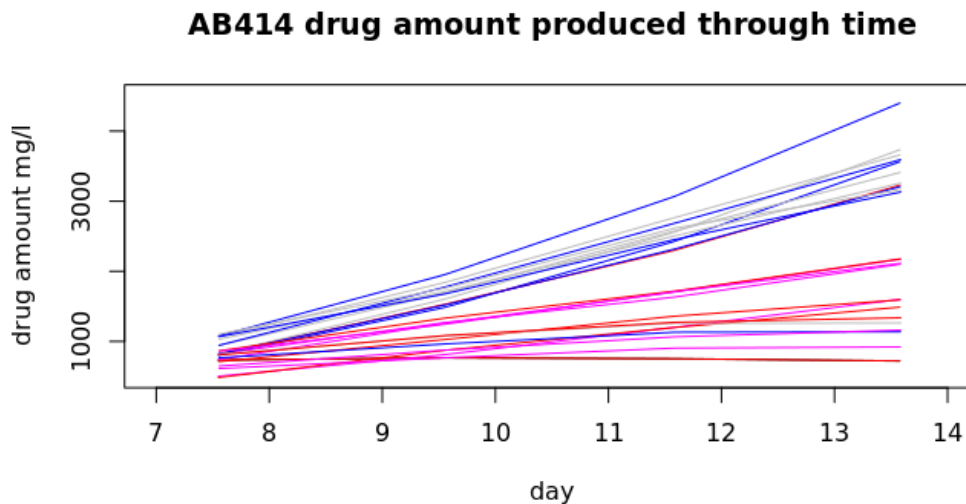


Figure 9: Drug quantity measurements (titre), linearly interpolated between the 4 measurements for a subsample of 48 AMBR experiments (AB414). As can be seen, titre measurements are not taken exactly at day 8, 10, 12 and 14, but rather a few hours earlier.

For this reason, we hypothesised that most information regarding drug production can be represented as an initial value and associated slope. We therefore computed the correlation between the initial value and slope of each bioreactor run with respect to time-series sensor data values at preceding time points [1]:

1. For each bioreactor run, we first performed smoothing with a 24-hour rolling mean over each time-series of the sensor data to remove some of the random noise,
2. For each run, we took the first titre measurement at day 4 as the

initial value. The slope was computed via a linear regression on the 4 titre production measurements,

3. We then computed a correlation coefficient across all runs between the initial values or slopes and the sensor data measurement at each of the time points before the initial titre value.

This gave us a correlation coefficient for both the slopes and initial values for each sensor data variable at each of the preceding time points. Figure 10 shows the results for the pH value, figure 11 for dissolved oxygen. In this way, we were able to examine temporal relationships between each biochemical sensor reading and drug-production across runs.

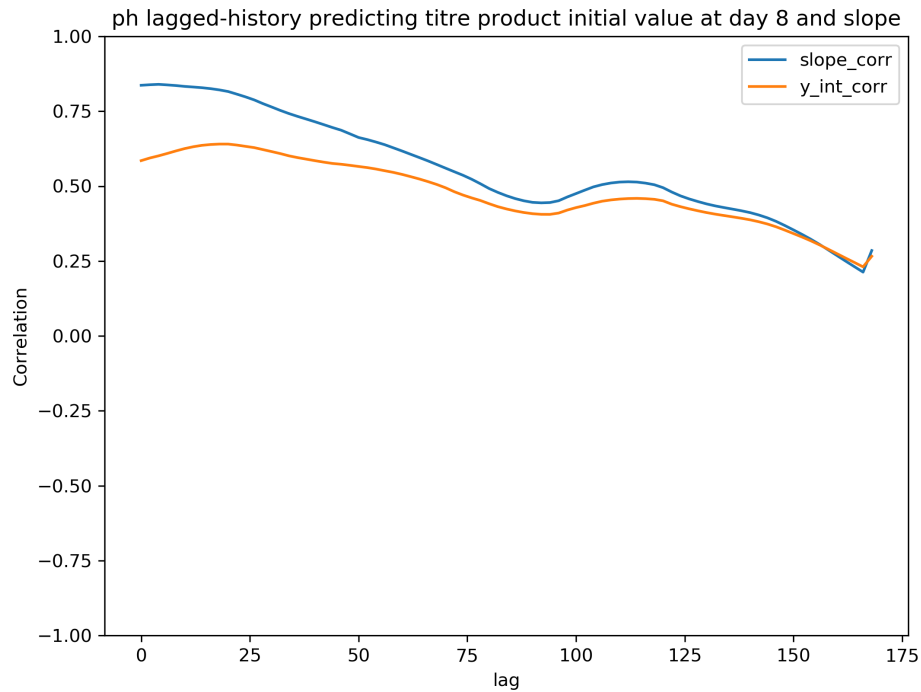


Figure 10: The graph shows the estimated correlation coefficient between pH values (ph) at each of the time points prior to the first titre measurement (lag) and the values of the first titre measurement and slopes. The lag is given in hours. The largest lag represents the start of the process, the smallest lag (zero) represents the time point of the first titre measurement. pH values was observed to have consistent positive correlation with the first measurement of drug amount and the slope of drug production. For smaller lags, the pH value becomes less correlated with the the initial value than the slope.

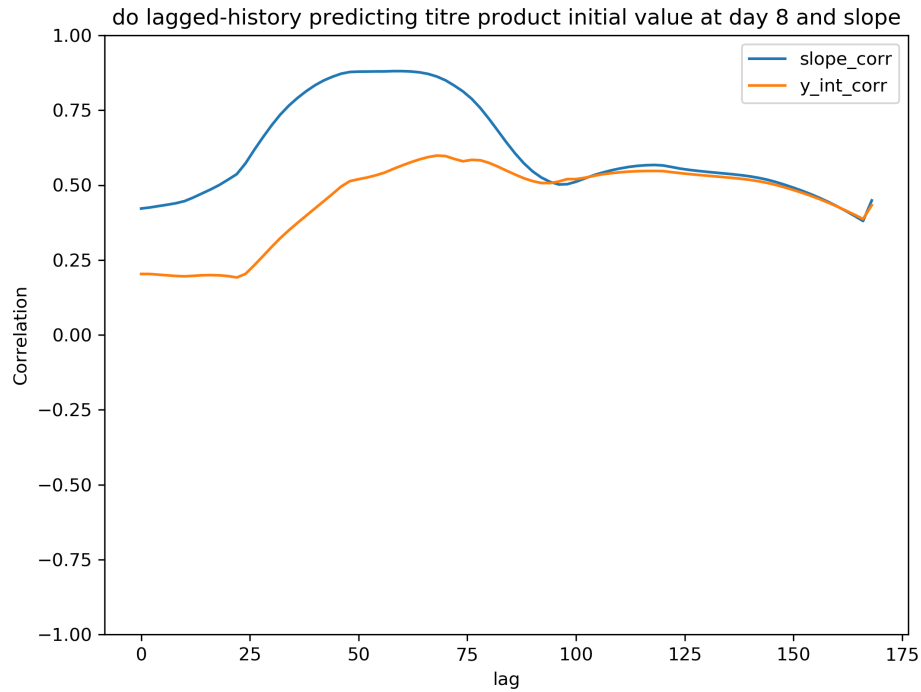


Figure 11: The graph shows the estimated correlation coefficient between dissolved oxygen (do) values at each of the time points prior to the first titre measurement (lag) and the values of the first titre measurement and slopes. The lag is given in hours. The largest lag represents the start of the process, the smallest lag (zero) represents the time point of the first titre measurement. In particular, do was observed to have consistent positive correlation with the first titre measurement value and slope of titre production. Between roughly lags 100 and 50, i.e. 100 and 150 hours after the start of the process, the level of dissolved oxygen shows a strong correlation with the slope of the production.

4 Experiments and results

4.1 Inference for identifying key controllable features

The first main objective is to infer controllable features that can be leveraged to enhance current process control mechanisms. Note that estimated associations between variables do not imply any causal relationships.

4.1.1 Model identification and estimation

For this task, a linear mixed model with random slope and intercept was fitted to the AMBR data [3]. The model is motivated by the observation that drug quantity grows roughly linearly over time with varying slope and intercept for each bioreactor run, as shown in figure 9.

For the mixed model, the response variable is titre, linearly interpolated between zero and each of the four titre measurements onto the hourly time points of the sensor data. The explanatory variables are time and airo-cumsum. For each bioreactor run, a random slope and intercept is estimated. For simplicity, the response variable was assumed to follow a normal distribution.

The model was fitted separately for each of the three batches of AMBR runs, each consisting of 48 bioreactor runs.

4.1.2 Results

For all three batches of runs, there was a statistically significant positive effect between airo-cumsum and titre. That is, higher values of cumulative oxygen have high correlation with an higher values of drug amount produced. Table 2 shows regression results for the AB414 set of experiments.

Note that while these results suggest that airo-cumsum is positively correlated with the process output, there are many potentially confounding variables that the model does not take into account. Thus,

the results do not imply any causal relationship between airo-cumsum and process output.

Table 2: Linear mixed model regression results for AB414 subsample

	Estimate	Std. error	t-value
Intercept	−325.06	38.33	−8.48
Airo-cumsum	6.58	0.29	22

4.1.3 Reproducing results

The script to fit the above model is in `prelim_modelling_v1.R`.

4.2 Prediction of drug quantity via feature extraction from partial time-series

The second main objectives is to predict the final quantity of drugs produced from partial time-series data observed up to a cut-off point. Being able to accurately predict the outcome early in the process can help identify low-yield runs and potentially signal interventions to steer runs into a more favourable direction or stop them early, and ultimately improve production efficiency.

4.2.1 Experimental Setup

In order to predict the final drug amount, the prediction experiment was set up as follows:

- After cleaning the data and removing missing values, the remaining 116 experiments from the AMBR data were randomly split into two disjoint subsets, 75% of the bioreactor runs were used for fitting the model, the remaining 25% were reserved for evaluating predictions on an held out test set.
- In order to investigate how early in the process one can reliably make predictions, cut-off points were set to 100, 150, 200, 300, 400 and 450 hours after the start of the process. For each cut-off point, only

the time-series data up to that point was used for fitting the model and making predictions. The last cut-off point served as a reference point for comparison, letting algorithms make use of the whole sensor data up to the end of the process.

- We used the following variables if they did not contain missing values: airo, co, do, do-integral-error, ph, lactate, airo-cumsum, co-cumsum and do-integral-error-cumsum.
- Using supervised learning algorithms with panel data requires to first transform the data into a tabular format via feature engineering. For the classification task, we consider outlier detection for feature engineering, as described below. Here, we considered the following three approaches:
 1. **Time-series as features.** We used each time point of each variable as a separate feature, ignoring any potential information carried by their temporal ordering.
 2. **Manual feature construction.** Based on our exploratory data analysis and MedImmune's domain expertise, we constructed the following features: the mean of pH and lactate value (if not missing) for subsequent 3-day intervals of the whole time-series from day 0 to day 12, i.e. for days 1-3, 3-6, 6-9 and 9-12.
 3. **Automatic feature extraction.** We used the tsfresh Python library [2] to automatically extract hundreds of features from each time-series using various time-series analysis and decomposition techniques.

Each approach was applied to each of the included sensor data variables. We then combined the resultant sets of features and used them as input features to the regression algorithm.

- In addition to feature engineering, we examined whether applying PCA as a feature reduction technique helps improve prediction performance. Due to time constraints, PCA was only used on the time-series as features approach.
- In order to have a point of reference, we compared our results to a naive baseline, simply predicting the mean titre value of the training set, ignoring any information carried by the input features.

- The tried out regression algorithms are random forest and gradient-boosted trees, both are sensible first choices as they work well as off-the-shelf methods without requiring much tuning and data preprocessing.
- To evaluate the prediction performance, we computed the mean squared error which measures the average squared difference between predicted and actual values.

4.2.2 Results

Figure 12 shows the results for all tried-out strategies over the chosen cut-off points. All strategies outperform the naive baseline. Out of the tried out strategies, automatic feature extraction combined with random forest shows the best performance across all cut-off points. This strategy reduces the mean squared error on the test set by roughly 70% compared to the naive baseline. This represents a reduction of the mean relative absolute error from roughly 30% to 20% around the actual values, highlighting the usefulness of this approach.

Using random forest with PCA as a feature reduction technique on the time-series as features approach does not seem to improve predictive performance.

Prediction performance also does not improve considerably for later cut-off points as more data becomes available. One reason is likely that many of the features are extracted from the earlier parts of the time-series or the time-series as a whole. Therefore, they may fail to pick out useful information from later segments. Another explanation may be that models start to over-fit on the training data when more data is used, so that predictive performance on the test set does not further improve.

Note that due to the time constraints of the challenge, we did not obtain confidence intervals, hence we could not test whether observed performance differences are statistically significant.

4.2.3 Reproducing results

The Python Jupyter notebook to replicate the results is `predict_all_strategies.ipynb`.

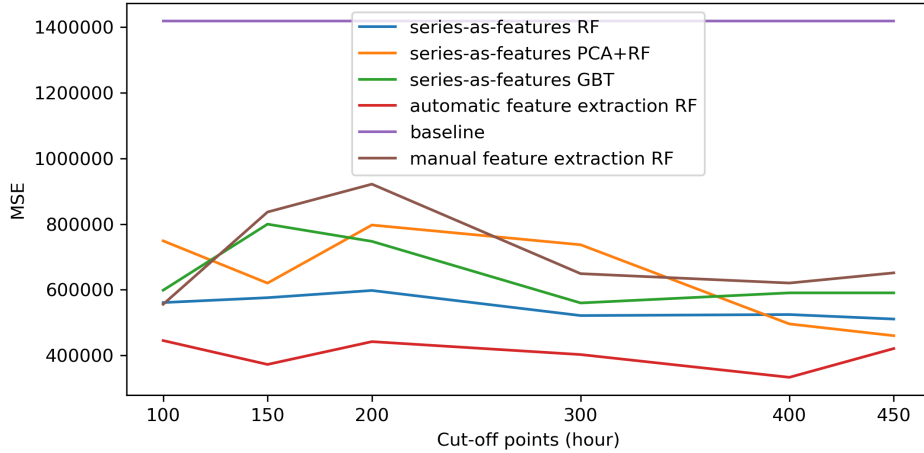


Figure 12: Mean squared error (MSE) for different prediction strategies over cut-off points.

4.3 Classification of low-yield processes via outlier detection from partial time-series

The third main objective is to classify which bioreactor runs present the highest risk for unfavourable process outcomes from partial time-series of sensor readings observed up to a cut-off point. Accurate classification early in the process can help identify unsuccessful experiments and thus help improve production efficiency.

4.3.1 Experimental setup

For this purpose, we set up the classification experiment as follows:

- Data for 48 AMBR experiments was randomly split into two disjoint subsets, 60% of the bioreactor runs were used for fitting, the remaining 40% were reserved for evaluating predictions of the fitted model on a held out test set.
- The target variable was created as a binary indicator equal to one if the process resulted in a titre concentration above a threshold of 1500 mg/l (high yield) and zero otherwise (low yield), where the

threshold was chosen based on MedImmune's domain experience.

- As input features, the variables do, glucose, lactate, temperature, ph and airo were considered. To use the time-series sensor data with standard supervised learning algorithms, we used outlier detection as a feature extraction technique. For each considered variable, we identified outliers via the interquartile range outlier detection method at each time point across bioreactor runs for the first 5 days of the process. The value of each time point was then encoded into a binary variable equal to one if it is an outlier and zero otherwise.
- The resultant matrix of binary features was used to predict the binary target variable using a random forest classification algorithm.
- 10-fold cross-validation was applied to the training set to optimise hyper-parameters of the algorithm.

4.3.2 Results

This approach achieved an accuracy score of 94.7%, a slight improvement over the naive baseline of simply predicting the most frequent class (high yield) of 89.5%.

Note that the performance difference is plausibly explained by chance for the given test set at the 95% confidence level using the binomial (or Wald) confidence interval.

4.3.3 Reproducing results

The R script for reproducing these results can be found in `codesPrediction.R`.

5 Future work and research avenues

5.1 Remedying data quality issues

- Collect or provide data on more bioreactor runs to statistically validate results,

- To speed up data analysis, implement systematic data curation, including standardised data cleaning (consistent, easily machine-readable encoding of missing values, measurement names) and data storage in readily accessible formats (e.g. single CSV-file for each type of bioreactor system). This will allow to understand data more quickly and start with predictive modelling sooner in future data scientific projects,
- Optimally, data would be stored in a relational database with standardised table schemata and automatic data quality checks. For each bioreactor systems, data would be recorded in the so-called long format, i.e. a matrix with columns for the identifier of the bioreactor run, the measurement time point and all observed variables and rows representing the measurement values at each time point.

5.2 Improving tried out approaches

- Investigate uncertainty associated with performance estimates, test statistical significance of results, investigate how well results generalise to real-world processes (e.g. processes at industrial scale, of other bioreactor systems with unseen parameter configurations, other cell cultures and drugs products),
- Include more extensive model selection techniques for algorithm tuning via hyper-parameter optimisation,
- Explore other algorithms, particularly more specialised time-series feature engineering approaches and algorithms,
- Regarding the mixed effect model, estimate non-linear mixed models with more appropriate distributional assumption on responses variable and additional explanatory variables.

5.3 Further research avenues

- Predict drug quality, which is besides drug quantity the most important process outcome,

- Develop forecasting and transfer function models for multivariate time-series data to better understand process dynamics between controlled and uncontrolled variables,
- Run intervention experiments to study effect of process adjustments with the aim to develop smarter process control systems,
- Create an interactive real-time visualisation of the sensor readings to monitor processes (e.g. using ShinyR apps).
- Ultimately, use predictions from machine learning algorithms to re-design bioreactor control system from following static parameter configurations into an responsive regime that makes process adjustments in a smart and automated way in real time, with the aim to optimise process stability and output.

References

- [1] George EP Box et al. *Time Series Analysis: Forecasting and Control*. John Wiley & Sons, 2015.
- [2] Maximilian Christ et al. “Time series feature extraction on the basis of scalable hypothesis tests (tsfresh - a Python package”. In: *Neurocomputing* 307 (2018), pp. 72–77.
- [3] Andrew Gelman and Jennifer Hill. *Data Analysis using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2007.
- [4] Gareth James et al. *An Introduction to Statistical Learning*. Springer, 2013.

Team members

Tracy Ballinger is a bioinformatician at the Institute of Genetics and Molecular Medicine at the University of Edinburgh. She is interested in cancer genomics and cancer evolution. She facilitated the group.

Magda Bucholt is a data scientist at the Intelligent Systems Research Centre at Ulster University. She is interested in developing healthcare solutions based on machine learning methods in order to transform delivery of care. She contributed to this report by exploring the data and implementing the classification of low-yield processes.

Jingjing Cui is a postdoctoral student at Queen Mary University of London. Her research focuses on optimization algorithms, game theory and machine learning in wireless networks. Currently, she is working on developing learning algorithms in UAVs. Her contribution to this project include the implementation of the time-series-as-feature approach and gradient-boosted regression algorithm.

Alex Gao is a second-year Statistics PhD student at the University of Toronto. He is interested in Bayesian methodology and spatio-temporal modelling. Most recently, he interned for the front-office of a professional basketball team in California, doing Bayesian modelling of basketball data. He contributed to this report by setting up and estimating the linear mixed effect model.

Tobias Hoejgaard Dovmark has a DPhil in cancer biology from Oxford University. His research focused on lactate, proton and bicarbonate transport in three dimensional pancreatic cancer tissue growths. He contributed to the report through exploratory data analysis and by bridging the chemical domain with the data scientific context.

Sangyu Lee is a PhD student in Statistics at the University of Leeds. Her research is on wavelet methods. She is particularly interested in analysing time changing spectra using locally stationary wavelets. She worked on exploratory data analysis and data visualisations.


Matthew Levine is a PhD student at Caltech in Computing and Mathematical Sciences. His research focuses on developing and applying novel methods in non-linear stochastic filtering, Bayesian

Inversion, and time-series analysis to biomedical problems. He contributed to this report by implementing the lagged correlation analysis and manual feature construction approach.

Markus Loning is a PhD student at UCL and an Enrichment Student at The Alan Turing Institute. His research focuses on supervised learning with time-series/panel data and toolbox development. He contributed to this report by mapping out the time-series/panel data prediction workflow and implementing the automatic feature extraction approach.

Daniela Perry is a Master's student in Data Science at the University of Munich. Her research interests lie in the fields of computational medicine and applied statistics. She contributed to data preprocessing and cleaning as well as exploratory data analysis.

Emma Vestesson is a data analyst at the Health Foundation. Her work focuses on using observational health care data in order to evaluate changes of the health care system. She worked on data preprocessing and cleaning and created an experimental ShinyR app for real-time monitoring of the sensor data.

The background of the image is split diagonally from the bottom-left to the top-right. The upper-left portion features a series of curved, overlapping lines in shades of blue and grey, creating a sense of depth and movement. The lower-right portion is a solid, light grey color.

turing.ac.uk
@turinginst