Applied Analytics Practicum - Fall 2022

Justin Huang

GT ID: 903651369

Georgia Institute of Technology

Sponsored by Sutro Biopharma

Background: Sutro Biopharma is a clinical stage biopharmaceutical company which discovers, develops, and manufactures therapeutics. Sutro focuses on next-generation cancer therapeutics, including antibody drug conjugates, bispecific antibodies and cytokine derivatives. Sutro's pipeline of products are based off of the platform Xpress CF (XCF) technology which allows customized manufacture, design and development of therapeutic proteins using cell-free protein synthesis. A core component of the XCF technology is the extract. From Sutro's website:

Sutro's technology platform is made possible by the separation, into an extract, of the cellular components required to produce proteins from the process of protein generation itself. The extract includes all the necessary biochemical components for energy production, transcription and translation and can be used to support cell-free biochemical protein synthesis by the addition of the specific DNA sequence for the desired protein. The process produces single proteins at g/L yields in 8-10 hours at any scale.

Although a basic understanding exists of how process affects product quality from literature and historical data, there is a need to improve process (increasing titers) as well as improve manufacturing consistency (robustness), along with understanding the "why" behind such improvements. There is tremendous business value in improved process understanding as it would impact all the therapeutics in the pipeline that depend on the platform XCF technology.

One approach is to use analytical tools and approaches such as Raman Spectroscopy to provide spectra, or a "molecular fingerprint", of liquid samples taken from the XCF reaction. Chemometric models are employed to match these spectra to offline measurement values, which are taken separately. Such methods rely on analytical and computational capability as well as understanding behind the model generation process.

Background on Raman Spectroscopy:

Raman spectroscopy has been used in the biopharma, chemical, and biomedical industries to improve process understanding. Essentially Raman spectroscopy can detect certain interactions of light with matter, and can pick up on rotational and vibrational modes of a system. A laser in conjunction with a probe can detect Rayleigh, Stokes, and Anti-Stoke's scattering as wave shifts across a spectra. These spectra contain chemical information about the sample itself. One advantage Raman spectroscopy provides is a nondestructive form of measurement, meaning the sample is kept intact. Another advantage is there is very low water interference, which makes it preferable for such applications (measuring solutions).

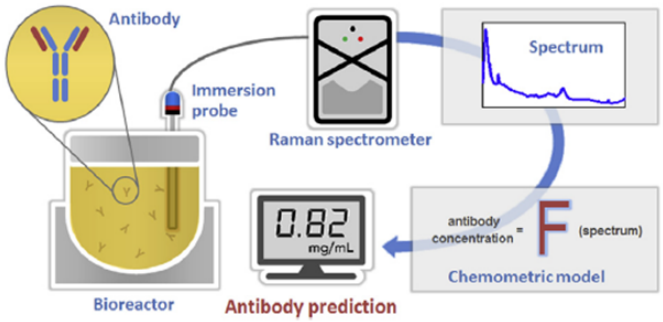

Figure 1 - Workflow of using Raman spectroscopy to predict offline measurements with chemometric modeling.

Recent advances in laser sources, high-speed optical fibers, and detectors has enabled Raman spectroscopy to find robust commercial applications. Raman has been established in many industries and has found applications in manufacturing processes to provide inline monitoring and real-time process corrections, leading to improved product quality. Furthermore, from the regulatory perspective, there has been increased expectation in process and product quality control from FDA, with an emphasis on continuous real time monitoring for quality.

Project scope/Deliverables:

The project focuses on characterizing performance across 11 runs based off 4 different extract lots. The 4 lots of extract differ in process conditions as well as point of manufacture.

Raman spectroscopy was set up for online continuous monitoring across the course of the 11 runs. For the 11 reactions, spectra were taken continuously and a total of 1086 spectra were collected. Samples were taken manually at intervals across each run for reference offline measurements. Chemometric models were built to correlate the spectra to the reference offline measurements (101 total). The models and results were then evaluated and tested.

The goal is to demonstrate applicability of Raman spectroscopy and analysis and gain insights on the XCF process. With this in mind, the scope and project deliverables are:

- Develop model(s) that can predict titer and metabolites from Raman Spectra
- Assess model performance and limitations.
- Analyze results and interpret findings.
- Build a scalable data analysis pipeline for building future models.

Methods:

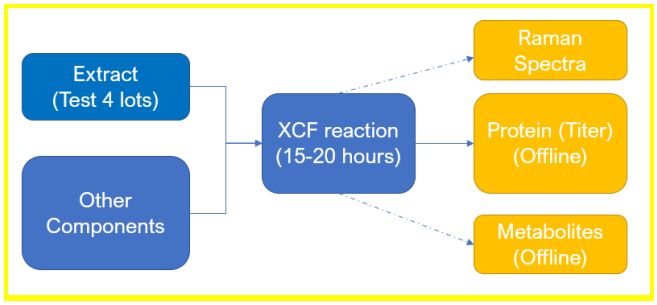

Figure 2 - Schematic for XCF reaction and data collection.

Data collection via Raman

A Raman spectrometer equipped with a 785 nm laser source was used for spectral acquisition. Reaction mixture was recirculated from the reaction vessel through a measurement chamber where an immersion probe was in contact with the flowing mixture. Care was taken to shield the measurement chamber from natural or artificial light. A suitable acquisition time consisting of a number of accumulations of scans gave a good balance of good signal-to-noise without oversaturating the detector. Acquisition ranges were across the 100-3425 cm-1 range. Examples of spectra can be seen in Figure 4.

Data collection titer and metabolites

Samples were collected and aliquoted manually at several timepoints for the 11 runs. Offline reference measurements for titer were taken through analytical methods developed at Sutro. Offline metabolite measurements (Glucose, Phosphate, Acetate, Ammonia, Magnesium were taken with a bioprocess analyzer. A significant amount of resources were dedicated to collecting these offline measurements for a total of 101 samples.

The spectral data and offline measurements (titer and metabolites) were joined through a sample ID using python script. All computation, algorithm implementation, analysis, and visualization was performed using Python and Jupyter notebook. Modules used include sklearn, plotly, pandas, numpy, and Pyspectra.

	Titer (mg/mL)	File Name	Solution A	Phase	100.0	101.0	102.0	103.0	104.0	105.0	...	3416.0	3417.0	3418.0	3419.(
0	0.003471	210923 run 1_20210923-093305_CONFIDENTIAL.spc	1	0.0	9.618279	9.026614	8.434949	7.843284	7.251619	6.659954	...	1.730245	1.706250	1.682256	1.65826
1	0.105501	210923 run 1_20210923-110303_CONFIDENTIAL.spc	1	0.0	9.713284	9.115023	8.516761	7.918500	7.320238	6.721976	...	1.727594	1.700147	1.672699	1.64525
2	0.216834	210923 run 1_20210923-131803_CONFIDENTIAL.spc	1	0.0	9.602399	9.016977	8.431556	7.846134	7.260713	6.675291	...	1.939261	1.926983	1.914706	1.90242
3	0.301073	210923 run 1_20210923-150303_CONFIDENTIAL.spc	1	0.0	9.466411	8.909972	8.353532	7.797093	7.240653	6.684214	...	1.783151	1.750451	1.717750	1.68505
4	0.330496	210923 run 1_20210923-161803_CONFIDENTIAL.spc	1	0.0	9.548907	8.988496	8.428085	7.867674	7.307264	6.746853	...	1.825770	1.792076	1.758382	1.72468

Figure 3 - Dataframe format in Python using Pandas module.

Preprocessing spectras

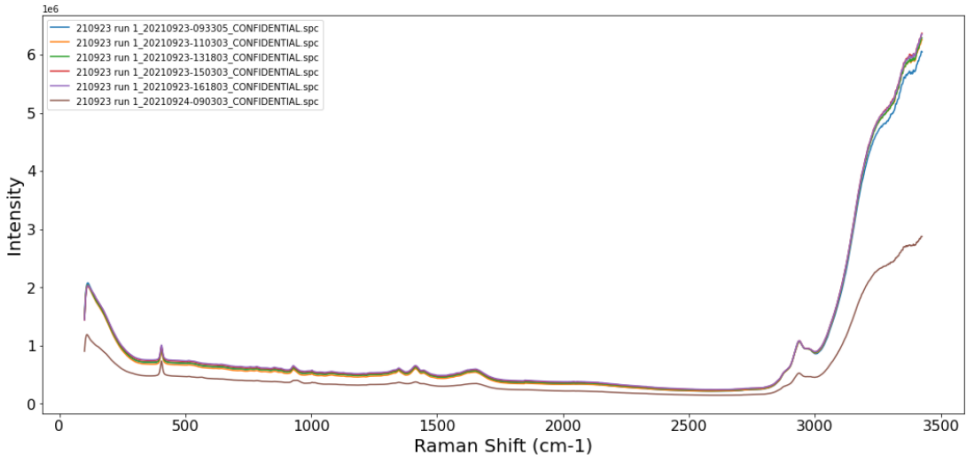

Figure 4 - Selected Raw Raman Spectra

The Raman spectra were passed through a Savitsky-Golay 1st derivative filter with a second order polynomial and window length of 29 to remove fluorescence background as well as smooth the data. Standard Normal Variate (SNV) was used to standardize the spectra collected at different times and across runs. This correction is also useful to standardize for global intensity variations due to potential probe misalignment when taking the measurements. The resulting curve of the derivative and standardization shows the effectiveness of this preprocessing step (Figure 5).

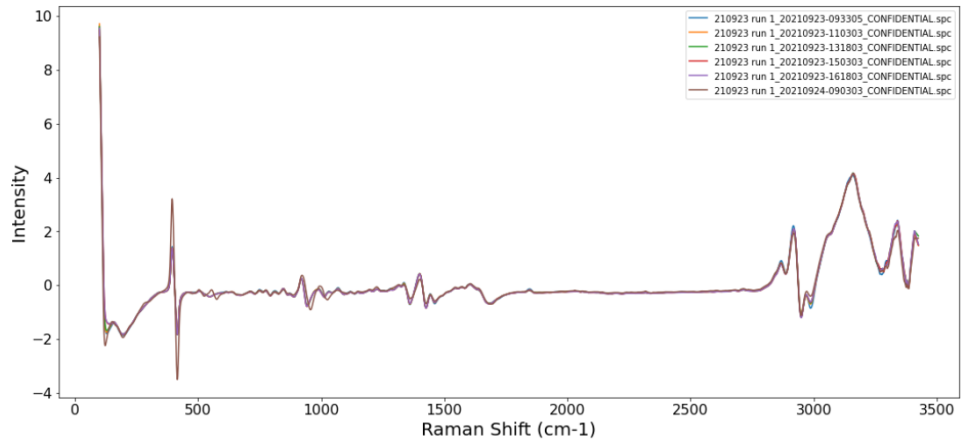

Figure 5 - Preprocessed Raman Spectra (First derivative, Smoothing, Normalization)

Model generation

Partial least squares (PLS) regression is used to establish a correlation between the preprocessed Raman Spectra and the reference measurements. PLS combines elements of principal component analysis and multiple regression, to find the principal components, or latent variables, that best explain the variance in X and y, as well as explain as much of the covariance between X and Y. PLS regression is particularly useful when there the number of features is high relative to the number of observations, and when there is multicollinearity among the features. Both of these cases are true in this application (101 observations, with 3500 wave features), and it is safe to say that spectra include multicollinear features.

Model Evaluation

For the model, optimal tuning parameters are found by training and cross validating on the training set. The tuning parameter here is the number of latent variables to use in the regression. A bias-variance tradeoff exists in the number of latent variables to include - too many and there is a risk of overfitting the model (Variance). Too little and the model may be underfit (Bias). We use repeated leave-one-out cross validation (LOOCV) to determine the number of latent variables to use in the model. Repeat LOOCV is preferable to k-fold CV in this case because the size of the training set is relatively small which may lead to high variance due to the small amount of data in each fold. Mean Squared Error is the metric used to calculate optimum amount of PLS components.

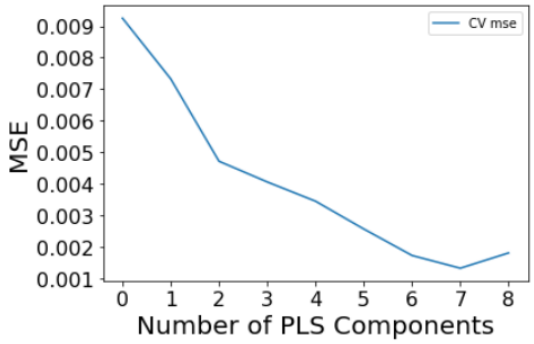

Figure 6 - Elbow Diagram of Cross Validated Mean Squared Error versus # PLS components.

The number of latent variables is selected manually by examining the elbow plot. PLS model predicts values for the test set (21 observations). 6 latent variables were selected for the model.

After fitting the model, titer was predicted for 21 observations in the test set. There was a positive linear relation between the predicted titer and the offline reference values for titer with R-squared of 0.90 (Figure 7). A possible outlier is observed and will be discussed in more depth.

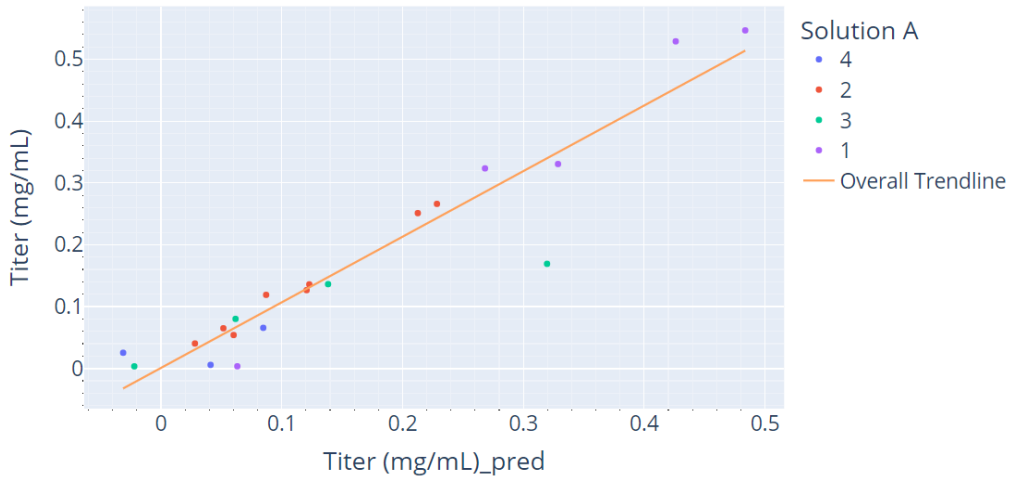

Figure 7 - Predicted titers vs actual offline titer measurements on test set.

Due to inherent differences in the process and source of the 4 extract lots, it made sense to examine different partitions of data used to train the model. A full partition means 1 model was trained on the entire training set and used to predict and compare against the test set. An individual partition means 4 separate models trained on their own respective lots to make predictions. Overall training 1 model on the entire set gave the best performance in terms of MSE and R-squared on the predicted values versus the actual values on the test set. Too many partitions simply increased variance on cross validation since the sample size after splitting was low. The full model was less prone to variance in cross validation. The individual and custom partitions may perform better as sample size is increased

Partition Type	MSE	# Latent Variables	R-squared
Full [1,2,3,4]	0.0027	6	0.900
Individual [1],[2],[3],[4]	0.0047	2	0.878
Custom [1],[2,3,4]	0.0019	3	0.743

Figure 8 - Summary of model evaluation with selective partitioning of training set.

Regression residuals were examined and plotted to check regression model assumptions (Figure 9). Again, one potential outlier exists and will be discussed later. The residuals appear to be centered and constant around zero without the presence of clusters. The histogram shows a normal distribution. The regression model is an appropriate choice for fitting, given that linearity, constant variance, independence, normal assumptions hold.

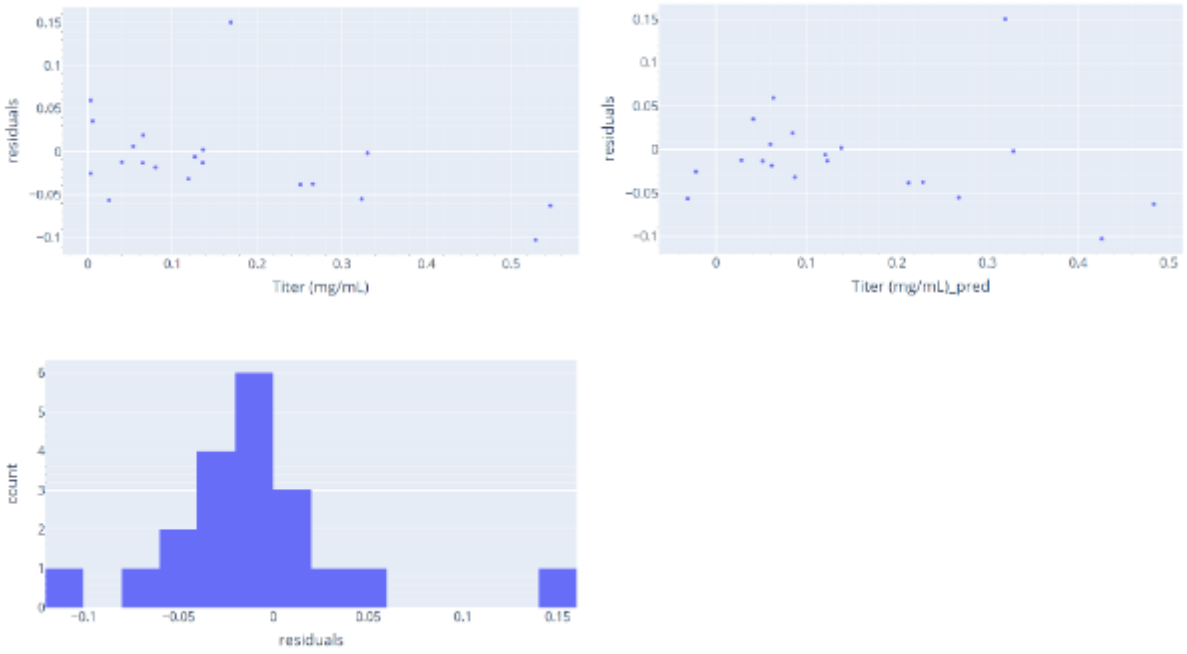

Figure 9 - Residual analysis for test set (21 observations). Top left - residuals vs titer. Top right - residuals vs predicted titer. Bottom left - Histogram of residuals.

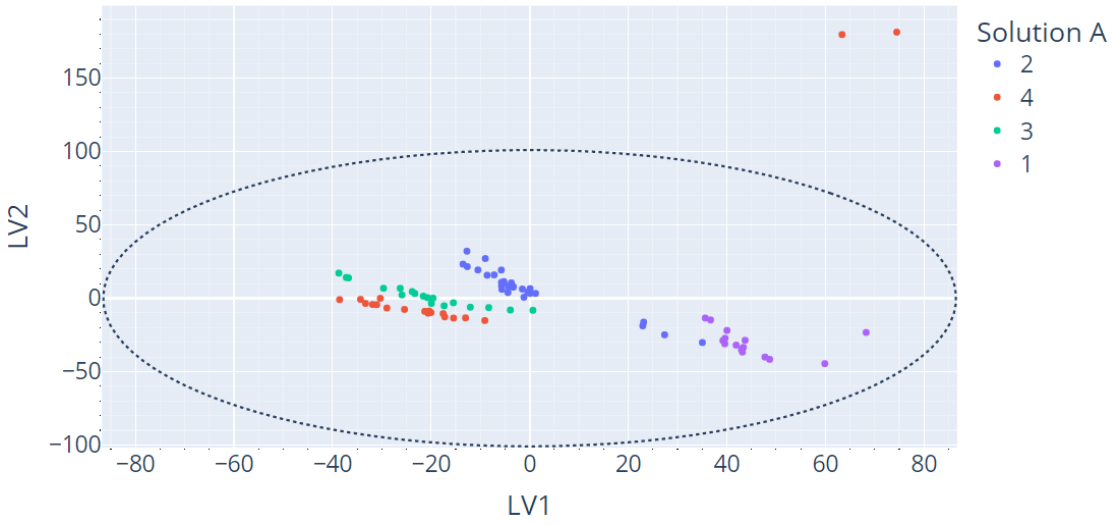

Figure 10 - Plot of training set (101 spectra) of 1st and 2nd PLS components projected space. Majority of the points fall within the 99% confidence ellipse, or 3 standard deviations. 2 outliers are present.

Interpretation

After PLS is able to correlate the Raman Spectra (X) and the offline measurements (y), Variable importance in projection (ViP) scores were calculated to evaluate the contributions each variable (wave shifts) had towards the variance explained by the PLS model.

$$\mathrm{VIP_j} = \sqrt{d \sum_{i=1}^{h} \left(\boldsymbol{b}_i^2 \boldsymbol{t}_i^T \boldsymbol{t}_i \right) \left(\boldsymbol{w}_{ji} / \|\boldsymbol{w}_i\|^2 \right) \Big/ \sum_{i=1}^{h} \boldsymbol{b}_i^2 \boldsymbol{t}_i^T \boldsymbol{t}_i}$$

With this framework each variable's (wave shift) importance could be evaluated. D is the number of variables from spectra matrix X. $b_i^2 t_i^T t_i$ is the variance explained by ith component. w_{ji}/w_i represents the importance of jth variable.

A user defined threshold can be used for variable selection. It is generally accepted that a variable should be selected if vj > 1, but this can vary. The VIP scores are interpreted by tying peaks to certain molecules in solution according to their molecular 'fingerprints'. For example, there are spectral patterns within the 1000 cm-1 to 1200 cm-1 that are characteristic of glucose.

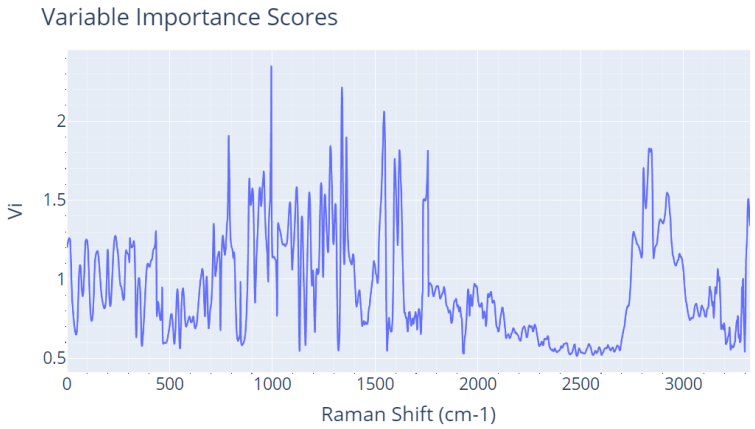

Figure 11 - Variable importance scores of spectral features.

The most important variable is in the 980 cm-1 region with a score greater than 2. The following variables of importance are 1300 cm-1, 1550 cm-1, and 800 cm-1 respectively.

There are other ways to explore if the Raman model is 'really' measuring the value of interest. Direct models can be employed where glucose or other analytes are directly added into the solution, looking for selective peaks in the spectral analysis. Inferential models can be used as well to see if for multiple titers and products, similar regions of peaks are consistently selected as the most important variables.

Analysis and Experiments:

Predictions and Titer Profiles

The full partition fitted model was used to predict on 1000+ collected spectra. The resulting titer predictions were plotted in time course for all 11 runs. The model was able to predict titer values (Figure 12). The profiles are consistent with expected kinetics as well as the offline reference values. This demonstrates the chemometric model's ability to make accurate predictions once a robust model is built. The analyzed spectra were treated as time independent, yet we are able to build time course profiles of the titer and gain insight into the kinetics. Potential applications of Raman modeling could reveal timing of certain events such as when the titer stops increasing during a run.

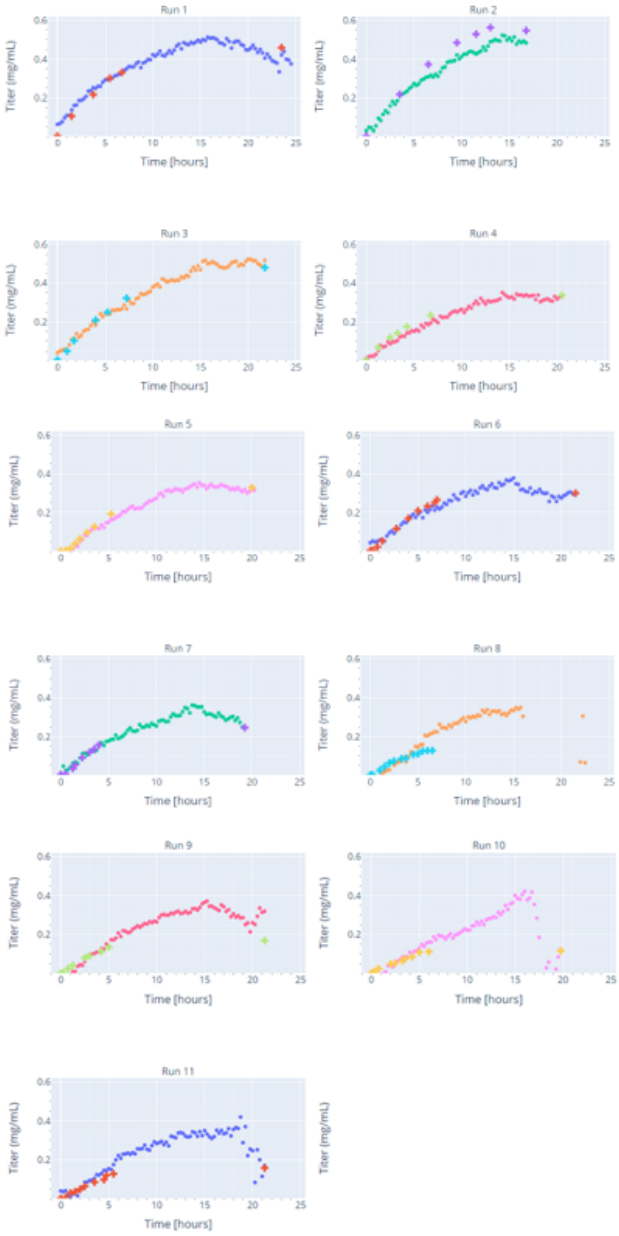

Figure 12 – Titer predictions (•) overlaid with reference Measurements (+) for 11 XCF runs.

The predicted values reveal some unexpected profiles. See the profiles for runs 9,10,11 in particular, where the predicted titers 'fall' to the offline reference measurements for titer at reaction endpoint. Given the context that there is a process shift on a critical parameter in the later stage of reaction, this makes more sense. When examining the potential outliers in the previous residuals plots (Figure 9), as well as projected plots on 1st two latent variables (Figure 10) it becomes clear that the potential outliers all came from values taken after the process shift. We conclude that separate models should be built before and after this process shift. Given the majority of samples were taken early on, the current model is most accurate for predictions before the process shift, and would likely further improve accuracy if more offline reference measurements are taken over the course of the reaction.

As concluded in the model comparison by partitioning, it may be possible to improve model accuracy by constructing individual models for each category of extract (individual partition). There is a need however to collect more offline reference values to increase the training set size.

Metabolite profiles and distinguishing characteristics

Models were constructed for metabolite reference measurements (Glucose, Phosphate, Acetate, Ammonia, Magnesium). For phosphate, distinguishing trends could be identified in the runs depending on the source of extract. From extract source 2,3,4, the starting phosphate was lower and a characteristic dip in the near the beginning of reaction could be seen. Color-coded groupings of the runs with different extract source can be seen in plots below.

Raman spectra were collected for two 'unlabeled' runs (Runs 12 and 13) for which no offline measurements were taken and a model was built correlating spectra to phosphate concentration. The model predicted phosphate profiles (Figure 13) of runs for 12 and 13 similar to runs with their respective extract source without offline reference values. This demonstrates the model was able to detect differences in features which related back to the extract source.

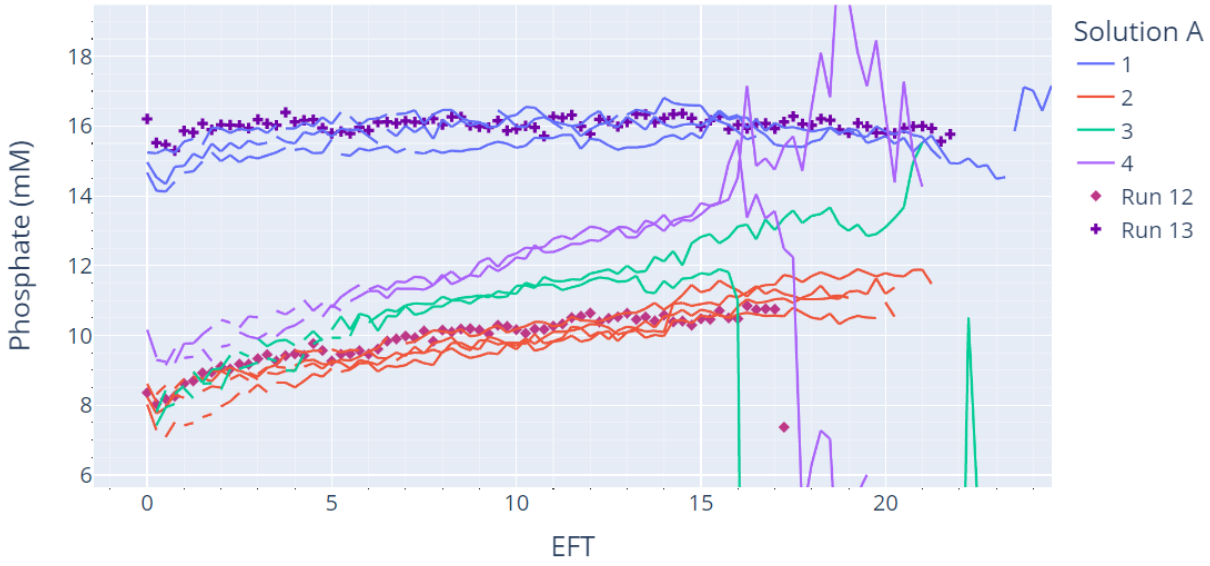

Figure 13 - Phosphate predictions for 11 runs + 2 unlabeled runs (◆). Characteristic groupings of run profiles are identifiable per extract source (1,2,3,4)

The same approach was taken for glucose measurements as well as there are distinguishing trends between runs dependent on extract source. Groupings of the run profiles can be seen in the plot below. In extract source 2,3,4 the starting glucose concentration starts around 1 g/L and there is some further accumulation throughout the run.The model predicted glucose profiles (Figure 14) of runs 12 and 13 similar to runs with their respective extract source without offline reference values. Once again this demonstrates that the model was able to detect differences in features which related back to the extract source.

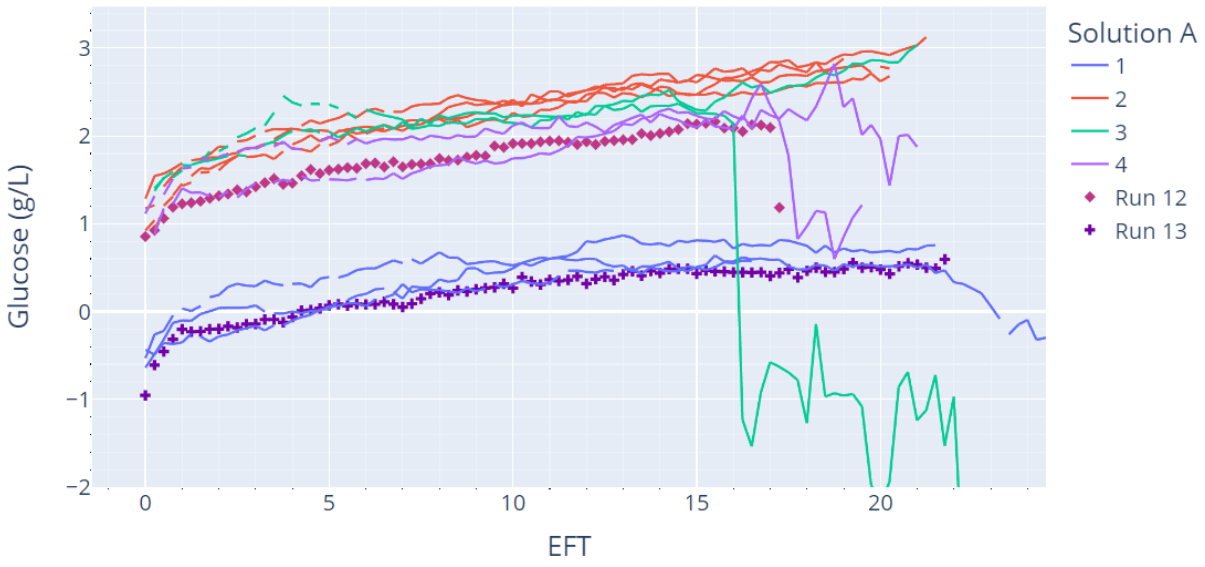

Figure 14 - Glucose predictions for 11 runs + 2 unlabeled runs (◆).

Model limitations

Another test of model performance involved prediction of titer on two ‘unlabeled’ runs. In runs 12 and 13, a key component of XCF was omitted on purpose where no protein, or titer, can theoretically be assembled. As such the expected titer is 0 throughout. The model predicted nonzero titers for these runs (Figure 15). This demonstrates the model is not measuring the spectra of the actual protein itself but of other biological or chemical components that are proximal for the protein of interest. This is not unexpected as a model’s predictions are only as good as the training data, and the model was untrained on such runs with missing component.

There are also some oscillatory periods on the plots of predicted titer and metabolites. These unexpected readings are explained by the process shift in the later stage of the reactions. The use of separate models before and after the shift may give more sensible predictions. It may be necessary to employ several models for processes with several shifts.

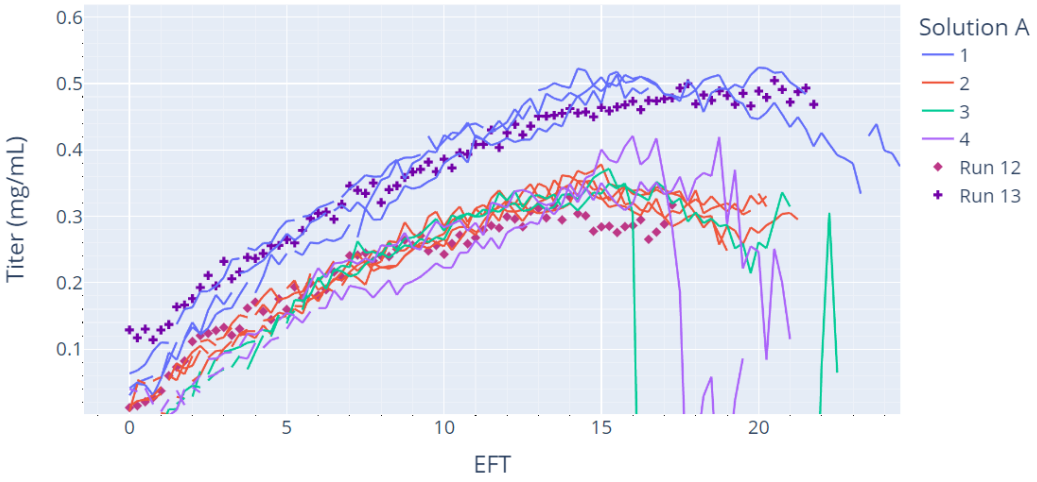

Figure 15 - Grouped titer predictions for 11 runs + 2 unlabeled runs (◆).

Conclusions:

Chemometrics and analytics were applied on a dataset provided by Sutro to derive insights on Sutro's platform technology. A dataset consisting of 11 XCF runs from 4 extract lots was analyzed. A robust set of Raman spectra was taken continuously throughout the reaction course. Offline reference values for titer and metabolites were acquired through various analytical methods. PLS regression was used to correlate preprocessed spectras and offline reference values. A full partition model was chosen after cross validation and was able to predict values such as titer and metabolite concentrations with quantifiable accuracy with respect to mean squared error and R-squared of 0.900. Using Variable Importance Scores, the model parameters were interpretable in terms of the original spectra features.

The Raman model was used to predict titer and metabolites based on several unlabeled spectra. Using this approach we were able to determine reaction profiles over time and gain insight into kinetics. The Raman model was able to pick up 'characteristic' phosphate and glucose profiles on two unlabeled runs. After the process shift, the models behaved unexpectedly and overall had reduced accuracy, demonstrating a need for separate models for different reaction phases. When subject to a test of 2 runs missing an essential reaction component, the model predicted titer when the expected titer was 0. Thus the prediction accuracy and overall usefulness was recognized to be limited to the training data available.

The applications of this work include reduced need for offline measurements, saving time and consumables costs. Combining Raman spectroscopy's real-time monitoring with the know-how on robust model construction, limitations of resources and operator sampling can be bypassed. Furthermore we analyzed and interpreted the models and gained an improved process understanding of the XCF platform process, of interest to both Research and Process Development groups. We expect to uncover many more insights with this methodology. Finally,

a scalable data analysis pipeline integrating Raman spectroscopy was developed in this project and is applicable for all future models and work to be done at Sutro.

References:

Tahir Mehmood, Kristian Hovde Liland, Lars Snipen, Solve Sæbø. A review of variable selection methods in Partial Least Squares Regression.

Silvere Andre, Lydia Saint Cristau, Sabine Gaillard, Olivier Devos a, Eric Calvosa, Ludovic Duponchel. In-line and real-time prediction of recombinant antibody titer by in situ Raman spectroscopy.

Karen A. Esmonde-White, Maryann Cuellar, Carsten Uerpmann, Bruno Lenain, Ian R. Lewis. Raman spectroscopy as a process analytical technology for pharmaceutical manufacturing and bioprocessing.

Jens A. Iversen, Rolf W. Berg & Birgitte K. Ahring. Quantitative monitoring of yeast fermentation using Raman spectroscopy.

Bjørn-Helge Mevik, Ron Wehrens. Introduction to the pls Package. July 14, 2022
Kevin Dunn. Process Improvement Using Data. 04 May 2022.

Acknowledgements:

David Burgenson, PhD, Scientist

Dawei Yuan, PhD, Principal Scientist

Marcella Yu, PhD, Senior Director