MACHINE LEARNING AND ADVANCED DATA ANALYTICS AUTOMATING THE EXPLOITATION OF RAMAN SPECTROSCOPY: FROM MICRO-SCALE TO LARGE-SCALE OPERATION

Stephen Goldrick, Advanced Centre for Biochemical Engineering, University College London
s.goldrick@ucl.ac.uk
Matthew Banner, Advanced Centre for Biochemical Engineering, University College London & AstraZeneca, Cambridge, UK
Michael Thomas, Advanced Centre for Biochemical Engineering, University College London
Jonathon Welsh, AstraZeneca, Cambridge, UK
Suzanne S. Farid, Advanced Centre for Biochemical Engineering, University College London, UK

Key Words: Raman spectroscopy, machine learning, on-line monitoring, high-throughput, control, critical quality attributes, critical process parameters, scale-up/scale-down.

Raman spectroscopy measures differences in vibrational states of chemical bonds enabling quantification and qualification of different molecular species. This technology has the potential to revolutionise many aspects of biopharmaceutical process development enabling on-line monitoring and control of critical process parameters (CPPs) and critical quality attributes (CQAs). However, the widespread adoption of this promising process analytical technology (PAT) is rather limited across the biopharmaceutical sector. This can be attributed to the challenges involved in the development, maintenance and transfer of Raman spectroscopy models across different cell lines, scales and facilities.

This work applies machine learning and advanced data analytics to automate the analysis of Raman spectra recorded on multiple cell lines producing a variety of therapeutic antibody-related modalities. The work involved collecting spectra using a 795 nm Raman spectroscopy for multiple laboratory scale (7 L) glass bioreactors and a large-scale (50 L) single-use bioreactor (SUB). Additionally, a high-throughput Raman device was investigated to measure at-line daily samples of multiple cell lines recorded on a micro-bioreactor system (ambr-15). An advanced machine learning approach was undertaken to automate all necessary pre-processing and model-building steps enabling accurate mathematical models to be easily generated. This included the application of an advanced genetic algorithm (GA) to automatically determine the optimum wavelengths and tweaking multiple model parameters ensuring the generation of highly robust mathematical models. Several model approaches were explored including principal component analysis (PCA), partial least squares (PLS) and neural networks in the generation of these mathematical models. The automated models were generated with no human intervention and were found to significantly outperform standard MVDA models enabling accurate predictions of the glucose, lactate, viable cell density, total cell density, protein concentrations and product quality attributes. The developed models were highly robust and validated across multiple projects. An additional challenge this research overcomes is the ability of the developed mathematical models to handle spectra that contain high fluorescence. Finally, this research explores the ability to integrate these on-line and at-line predictions for advanced process control leading to improved performance and optimization of biopharmaceutical process development.