**ANN training with a generative *in silico* model for PAT in the biopharmaceutical DSP.**

Matthias Rüdt1, Jürgen Hubbuch1,\*

*1 Kar​lsruhe Institute of Technology (KIT), Institute of Engineering in Life Sciences, Section IV: Biomolecular Separation Engineering, Fritz-Haber-Weg 2, 76131 Karlsruhe, Germany*

*\*Corresponding author: juergen.hubbuch@kit.edu*

**Highlights**

* ANN applied to chromatography in the biopharmaceutical DSP.
* ANN training based on a generative *in silico* model.
* The calibration approach is versatile and scalable.
* RMSEPs compared to a standard PLS could be reduced by 14% to 40%.

**1. Introduction**

Biopharmaceutics are the fastest growing sector of the modern pharmaceutical industry. Nowadays, seven out of the ten most valuable pharmaceutical products are biopharmaceutics [1]. During the downstream processing (DSP) of biopharmaceuticals, chromatography is the main workhorse. While in production mainly univariate signals are used to monitor chromatographic steps, recently the benefits of multivariate spectroscopic methods for process monitoring have been demonstrated [2]. Data analysis was up to now mostly limited to linear methods such as partial least-squares (PLS) models. Machine learning provides much more powerful methods such as artificial neural networks (ANN). However, ANNs generally need a vast amount of data to be reliably calibrated. In wet lab experiments, this amount of data is difficult to obtain. To circumvent this problem, we developed an approach relying on a generative model to yield artificial process sensor data and the corresponding reference concentrations. Based on this data, the ANN could be successfully trained.

**2. Methods**

Experimental data from a publication by Brestrich et al. [3] were used in this study. Briefly, preparative chromatography experiments were performed on an Äkta Purifier 10 system (GE Healthcare, Chalfont St. Giles, UK) equipped with a diode array detector (optical pathlength 0.4mm) UltiMate 3000 DAD (Thermo Fisher Scientific, Waltham, US) with a HiTrap 7x25 mm column prepacked with SP Sepharose FF (both GE Healthcare). Experiments were performed for a ternary separation of ribonuclease A, cytochrome c and lysozyme (all Sigma Aldrich, St. Louis, US). Chromatographic separations were modeled with ChromX (GoSilico GmbH, Karlsruhe, Germany).

For generating the *in silico* dataset, first, the mechanistic chromatography model was calibrated based on multiple calibration runs. Simultaneously, a database for spectroscopic data including the relevant reference analytics was established. Next, the design space of the process was randomly sampled yielding 10,000 process chromatograms with simulated concentration traces resulting in 780,000 concentration vectors. For each concentration vector, the database was sampled for the nearest neighbors (Euclidean-distance). Scaled sensor responses were then obtained by multiple linear regression and subsequent matrix multiplication of concentration vectors and spectral matrices. On the obtained dataset, ANNs with different geometries were calibrated with the ANN framework Keras [4]. The regression models were tested on an independent chromatography run.

**3. Results and discussion**

A workflow was established for calibrating ANNs based on generative approach consisting of a mechanistic chromatography model and a spectral database. The generative approach allows to easily scale the calibration dataset. It thus circumvents some of the issues of using ANNs for process data. Most importantly, overfitting ANNs is reduced. A further benefit of the established workflow is the easy scalability of the spectral database. Thus, the database can easily be extended without significantly complicating calculations. Finally, the workflow captures non-linearities in the sensor response due to the local linear regression.

Based on the generated *in silico* data, two ANNs were calibrated. While a classical (dense) ANN geometry was on par with a reference PLS model, a convolutional ANN surpassed the PLS model in all calculated root mean squared error of calibration (RMSEC) and prediction (RMSEP, see Table 1). The results demonstrate the potential that lies in the application of more complex regression models compared to PLS.

**Table 1.** Results from the different calibrated models. Best results in each column are underlined.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Ribonuclease A | | Cytochrome c | | Lysozyme | |
|  | **RMSEC / (g/L)** | **RMSEP / (g/L)** | **RMSEC / (g/L)** | **RMSEP / (g/L)** | **RMSEC / (g/L)** | **RMSEP / (g/L)** |
| PLS | 0.0126 | 0.0149 | 0.0135 | 0.0098 | 0.0132 | 0.0116 |
| Classical ANN | 0.0176 | 0.0150 | 0.0122 | 0.0096 | 0.0140 | 0.0108 |
| Convolutional ANN | 0.0090 | 0.0130 | 0.0065 | 0.0059 | 0.0084 | 0.0073 |

**4. Conclusions**

We investigated on a generative *in silico* approach to calibrate ANNs for PAT applications. A workflow was established involving the calibration of a mechanistic model, spectroscopic database establishment, Monte-Carlo-based process sampling, spectra generation, and ANN model calibration. Based on the proposed approach, a convoluational ANN performed 15% to 40% better on an independent chromatography experiment compared to a standard PLS model. The results demonstrate significant potential of the generative approach to ANN calibration.

**References**

[1] Njardarson JT. Top 200 Pharmaceutical Products by Retail Sales 2018. 2018.

[2] Rüdt M, Briskot T, Hubbuch J. Advances in downstream processing of biologics – Spectroscopy: An emerging process analytical technology. J Chrom A 2017; 1490:2–9.

[3] Brestrich N, Hahn T, Hubbuch J. Application of spectral deconvolution and inverse mechanistic modelling as a tool for root cause investigation in protein chromatography. J Chrom A 2016; 1437:158–67.

[4] Chollet F and others. Keras. 2015. <https://keras.io>.