The Role of Chemometrics in Establishing PAT Prediction Models for Analytical Concentration Determination

By Brad Swarbrick[1]

DOI: 10.62178/sst.004.008

ABSTRACT

The implementation of Process Analytical Technologies (PAT) has not been possible without a symbiotic relationship with chemometrics – multivariate data analysis and modelling. This article complements the context for this issue of SAMPLING SCIENCE and TECHNOLOGY (SST), 'The PAT issue', by introducing a compact brief of chemometrics: defined as the extraction of information from chemical data for the purposes of making informed decisions, especially when integrated with PAT, where quality decisions are made in real-time on the current state of the process. Using predictive chemometric models, changes can be made to the process so as to minimise the risk of deviations and out-of-specification (OOS) material.

1. Introduction

Implementation of process analytical technology (PAT) is not plug'n'play! As is evident from the articles presented elsewhere in this issue of Sampling Science and Technology, only in the rarest of cases concerning extremely low material heterogeneity will direct insertion of a PAT probe into a process stream result in the acquisition of representative spectral data. This article gives a perspective of what happens on the 'other side' of a successful PAT sensor implementation, where raw spectral data has been generated for the development of data models for prediction of analytical concentration (calibration of analytical instrumentation). For this ultimate PAT step chemometric multivariate calibration (and validation) come to the fore.

Chemometric data models have been instrumental for the successful development of PAT (see several other articles in this issue). The standard, proven chemometric data analytic approaches include methods such as Principal Component Analysis (PCA), Multiple Linear Regression (MLR) and Partial Least Squares regression (PLS) that are not only simple in application but are easily interpretable and can be validated to any contextual level.

This latter in deliberate opposition to current Artificial Intelligence (AI) and Machine Learning (ML) approaches that are wonderfully efficient data-wise, but which mainly work in the dark. By their nature these approaches are not amenable to proper validation.
A comprehensive introduction to chemometric MVDA is Esbensen and Swarbrick (2018), in which all relevant aspects regarding MVDA data models are presented in an authoritative context, including the critical sampling and model validation issues in full (chapters 3 and 6 respectively).

For all PAT implementations, there are a number of prerequisites that must be met with:

1. There must be a physico-chemical relationship between the PAT signal and the analyte(s) of interest.
2. There must be a 1:1 volume correspondence between the sample(s) characterised by the PAT sensor (either on an extracted sample or as a spectral signal [X] pertaining to a relevant stream segment) and the aliquot used for reference analysis [y]. [X,y] are the complementary matrix/vector data needed for the generation of chemometric models.

1 KAX Group, Penrith, Australia.

3. The samples used to generate a calibration MVDA model must span the greatest relevant range of analyte concentrations as defined by the PAT objective; for example, if a regulatory specification requires the range 75-125% of the target analyte concentration, the calibration sample set must span this range as well (analytical range and internal validation).
4. The calibration data set [X,y] must fit the relevant form of the MVDA model (for example, a straight-line fit for a linear model). If the model shows non-linearities, these must be addressed using an appropriate and interpretable set of pretreatment data transformations and/or by using an extra number of data analytical components.

The model must generate reliable prediction concentration results when applied to new samples which are independent of the calibration set (test set, or external validation).

It is of key importance for steps 3-5 to be reliable, that steps 1 and 2 are complied with without exception. There is an alarming recent trend where newcomers to PAT/ chemometrics are looking for automated ways of generating models - and ditto: automated way of validating model performance without putting in the effort to understand what a particular model is doing and why the particular preprocessing method applied is relevant.

There is a vast range of preprocessing methods available for this purpose, but their use requires considerable insight and experience. Even more alarming is today's indiscriminate use of non-linear, AI and Deep Learning approaches, which will fit any data set, but provide little or no interpretation insight, hence they have little or no scientific value.

Before providing examples of the way PAT has been implemented and the perils of poor sampling, the next section defines a compact roadmap for chemometric data modelling based on sound sampling practices. A first set of introductory literature references for newcomers to chemometrics can be found in (Wold, 1995; Esbensen and Swarbrick, 2018), augmented by four historical and general background references (Martens and Næs, 1991; Massart, 1997; Adams, 2004; Höskuldsson, 2024).

2. A Concise Introduction to Chemometrics Methods

Chemometrics is not application of mathematics – chemometrics is the prime data analytical tool for extraction of information from chemical data. While there is a mathematical component involved, chemometrics is very much also dependent upon knowledge and experience about spectroscopy, chemistry and pattern recognition - all bought together such that essential subject-matter interpretations can be made.

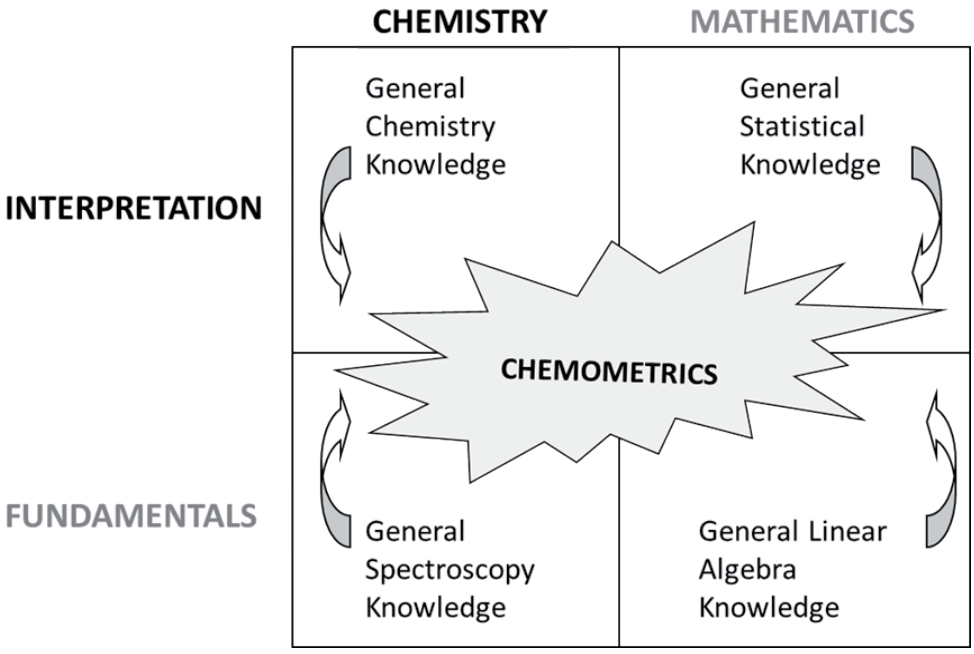

Figure 1: Chemometrics was born as a branch of chemistry concerned with the analysis of acquired data and ensuring that experimental and empirical data contain the maximum information [3]. Since its inception in 1972 chemometrics has developed into a general data analytical approach long since transgressing the boundaries of chemistry.

Figure 1 provides an infographic of how the disciplines of chemometrics come together to achieve the goal of understanding data structures embedded in data tables (data matrices), e.g., [X] and [X,y].

For the present compact perspective, chemometrics can be sub-divided into three main areas (Esbensen and Swarbrick, 2018; Martens and Næs, 1991; Massart, 1997; Höskuldsson, 2024; Adams, 2004):

1. Exploratory Data Analysis (EDA): This type of analysis is used to detect trends and patterns in spectroscopic and other types of multivariate data. Data may typically come from discrete samples generated at different locations or may be data generated over time from a process line.
2. Regression Analysis: Predictive models are generated from paired spectroscopic and 'reference' data [X,y] acquired on the same sample. In this way, the calibrated spectroscopic data [X] can generate many more predictions [Y] = [y1, y2, y3 ....] compared to physical sampling and sending samples to a reference laboratory.
3. Classification: The uniqueness of spectroscopic and other multivariate data for grouping into specific sample classes is used to develop discrimination models that can classify new, unknown samples w.r.t. known training classes. Training classes are bound by statistical limits, therefore providing a level of confidence to the predictions.

No matter which approach is used, chemometric data models follow the same foundational principle of partitioning data into a systematic structure part (the information part) and a noise part, Fig. 2.

In the general case of EDA, the data are partitioned into the structure part, i.e. groups, trends etc. and a noise part that the model cannot explain.

In the Regression/Classification case, a functional relationship between X- and Y-Data is established. The better fit of the data to the model, the less noise influences the model. However, if there is no subject-matter 1-to-1 information in the data, the noise part will be inflated and the resulting model performance will necessarily be poor.

Even if there is a lot of information in the data, or there is a strong relationship between X- and Y-Data, if proper sampling methods were not effectuated when generating the samples, the resulting analytical data will inherently be more influenced by noise than need be. In such cases, chemometric modelling will partition most of the data structure into the noise part – and the potential for generating a useful and acceptable model is often forfeited. It is never a good idea to ignore the good sampling imperative! A good primer for connecting the sampling, analysis and the data analysis realms can be found in (Esbensen, 2025a; Esbensen,2025b).

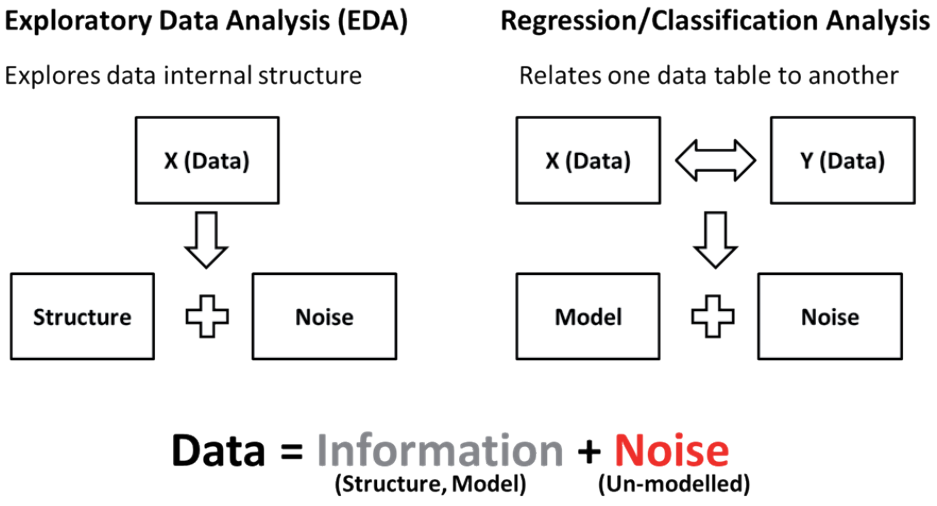

Figure 2: The general structure vs. noise decomposition assumption behind all chemometric data modelling

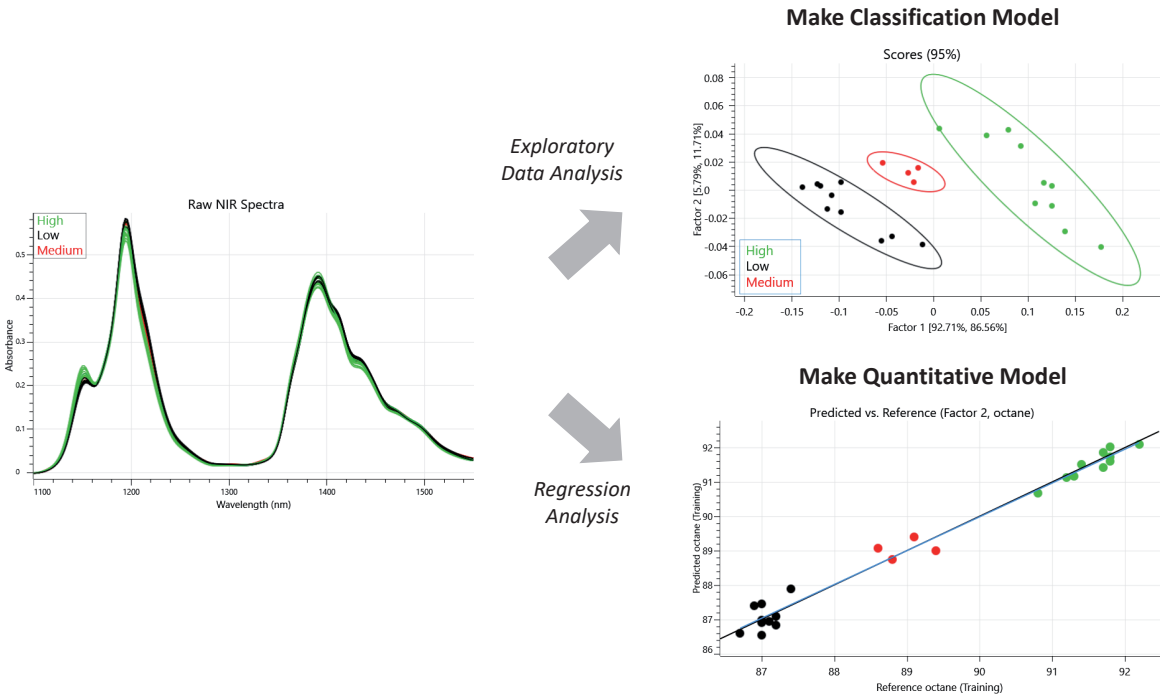

Figure 3: Generic example showing how spectroscopic data can be decomposed into structural patterns for classification or regression prediction. A multivariate chemometric model detects many feature the human eye does not, which is why these methods are called latent methods.

Fig. 3 shows illustrative examples of the regression and classification/EDA objectives in terms of chemometrics' key graphic performance facility, the so-called scores plot and the predicted vs. reference plot, which are common to methods such as PCA and PLSR. Full introduction in (Esbensen et al., 2018; Esbensen 2025a).

This brief chemometric overview has not considered the preprocessing, interpretation, optimisation and validation efforts which are also required to generate a reliable analytical prediction model.

In general, for most data analysis objectives, there is no need to look for more complex, non-linear, or more advanced methods to "sort through the rubbish" - if sampling is addressed and the right PAT technology is chosen(Esbensen and Swarbrick 2018; Esbensen, 2025a; Danish Standard, 2024).

The prime hallmark of chemometrics is the insistence of proper validation of all data models of whatever type (Esbensen and Geladi, 2010). Recently the place and role of multivariate data modelling/chemometrics was described in a broader philosophical perspective (Esbensen, 2025a; Esbensen 2025b).

Chemometric models can, in some sense, be viewed as a subset of Machine Learning (ML) models but by including the important aspects interpretability, problem-dependent outlier detection and proper validation they are in a class of their own. Chemometrics was founded in 1972; referral can be given to a bonanza of introductory literature developed over more than 50 years (Martens and Næs, 1991; Massart, 1997; Adams, 2004; Esbensen and Swarbrick, 2018; Höskuldsson, 2024; Danish Standard, 2024; Esbensen, 2025a; Esbensen 2025b) and further references herein.

3. Five "R's" of Chemometric Model Development

One of the most influential documents used as a guideline for the development of reliable analytical models was developed by the International Conference of Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) in the Q2(R2) (ICH, 2008) document.

This document, entitled "Validation of Analytical Procedures" and its latest counterpart ICH Q14 "Analytical Procedure Development" (ICH, 2022) have been referenced in many other regulatory documents, including the European Medicine Agency's guidance on the development of near infrared (NIR) spectroscopic methods (European Medicine Agency, 2014) and may be condensed into the four R's of analytical development.

1. Repeatability: Can the same aliquot be remeasured in a short time period and generate predictions that are not significantly different, typically at the 95% confidence level, or with a relative standard deviation < 2% (PAT instrument analytical precision).
2. Reproducibility: Typically relates to the transfer of a method from one PAT instrument to another. Reproducibility is an assessment of prediction accuracy and precision with respect to the aliquot mass.
3. Robustness: Does the PAT approach generate statistically similar results when small, but deliberate changes need to be accepted regarding the full measurement system (lot-to-aliquot). This is partly also an assessment of the reliability of the sub-sampling system implemented to generate the stable raw spectral data.
4. Reliability: Typically relates to the precision and accuracy statistics of the calibration model and the method ability to predict new samples (external validation).
5. Representativity. This the forgotten "R" which is the focus of this SST issue. Representativity relates to the support mass/volume for PAT sensor signals (can either be defined w.r.t. an individual stream segment (increment) or the full length of the streaming material batch, see (Esbensen 2025a, Esbensen 2025b) for a detailed discussion). Representativity of any analytical result must point back to the original target material or lot.

PAT representativity must be seen in relation to the positioning of the sampling sensor in the process stream and the ability to establish a representative process sampling interface, see Esbensen in this issue. This last point is of overwhelming importance for PAT, typically requiring ~75% of any PAT method development efforts. This also reinforces the old saying, "Garbage in, Garbage Out". If the data being generated are not representative, one cannot expect the chemometric model to perform miracles.

4. The Chemometric Model Development Roadmap

Armed with this information and a systematic approach to method development, the workflow in Figure 4 summarises the development effort required to generate reliable, robust and representative chemometric analytical prediction models.

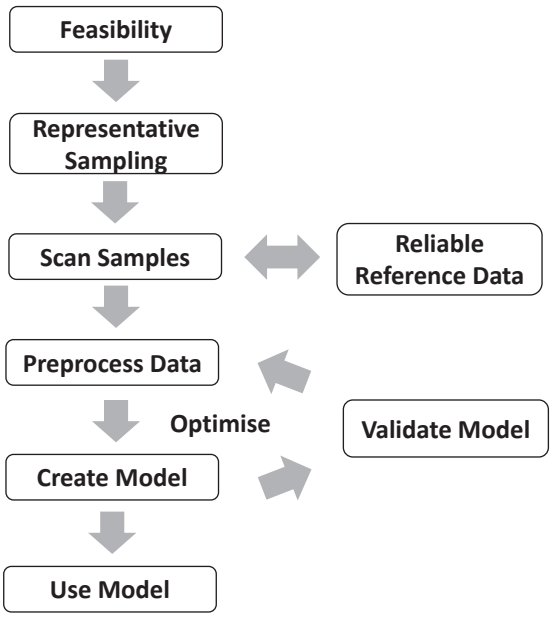

Figure 4: Generic workflow for developing reliable chemometric prediction models.

A brief substantiation of Figure 4 is as follows.

1. Feasibility Study: It is not necessarily an easy choice to find a candidate sensor, considering the very many sciences and applied fields which are opening up for PAT, e.g., biology, pharmaceutical, metallurgy, geoscience (minerals, commodities, ores), processing industry. A feasibility study is used to verify that the signal from the selected PAT sensor is indeed related in the physical, chemical or subject-matter sense to the analyte(s) of interest.
2. Representative Sampling: Is the PAT instrument located correctly in the process, and does it generate a response that is a true representation of the aliquot delineated by the process sampling interface (Esbensen in this issue). Be aware that it is not always possible to extract the exact same sample from the process that was characterised by the PAT sensor (volume mismatch error introduced in Esbensen this issue). Representativity is very much also related to the procurement of sample data sets (chemometrics: training and validation data sets) used to extend the range of constituent analytical min/max values.

3. Sample scanning: This is the critical element in analytical due diligence: proper optimisation of spectral acquisition parameters. This is the core of the analytical domain (Höskuldsson, 2024). Failure to optimise analytical instrument parameters will lead to the generation of imprecise raw spectral data, which cannot but influences negatively on the relevance and power of chemometric prediction models.
4. Reliable Reference Data: There must be a strict 1:1 reference sample-to-analytical-aliquot correspondence as any misalignment or disconnect between the two will result in imprecise and biased data. The reliability of the reference method [y] must also be established to generate the Standard Error of Laboratory (SEL) value. Without a critical evaluation of SEL, the precision and accuracy of the final prediction model cannot be properly assessed.
5. Preprocess Data: Preprocessing is *not* a supermarket of methods used in random combinations to generate a 'cleaned up' dataset. Proper application of preprocessing is intimately related to both spectroscopic type and wavelength region. Preprocessing is used to minimise the effects of residual spectral acquisition effects and can often be helpful in bringing out more clearly the subject-matter (e.g., chemical/biological) information in the acquired data. N.B. Preprocessing is no substitute for representative sampling!
6. Establish -, Validate -, Optimise Model: This is the iterative cycle of proper chemometric model development where fine tuning is applied to wavelength region selection, or to preprocessing methods such that the minimum number of components/factors are used in the final model and so the calibration and validation prediction statistics are as close (in the statistical sense) to the SEL as possible. Ideally one would prefer complete similarity, but this would be dependent upon successful elimination of all sampling, sub-sampling, all spectral acquisition errors, as well as all data modelling errors – a complete impossibility for the degree of complexity met with in technological and industrial data sets. The optimisation step results in a model that, when applied to a new, external sample set, will generate predictions that can meet the validation criteria defined for the model, often a minimum prediction error measure is used.
7. Model in use, Model Maintenance: The only way to improve a(ny) model in practice is by learning through its application. This is the Model Maintenance stage and must be implemented as a lifecycle model (Flåten, 2018). The model maintenance stage allows a user to identify whether process changes have occurred that result from identifiable causes, e.g., raw material changes, or observable changes that may indicate the need for PAT instrument maintenance, or changes pointing to the need for re-calibration and renewed qualification. This stage is obviously where informed competence and the largest possible experience with model use in practice will be of key importance.

5. Chemometrics for PAT - in practice

The following two examples illustrates the result of not following the above approach, i.e., of neglecting the principles of Good Data Modelling Practice (GDMP) outlined above.

5.1 Consequences of incorrect PAT Sampling for Chemometric Models

An important distinction: PAT is not about bringing the quality control laboratory to the process (an often-used euphemism) but is about PAT's ability to assess the current state of a process or a product. In this sense there are PAT implementations that work in an at-line capacity (Esbensen in this issue) where a specified number of representative samples are drawn from a process and measured at predefined intervals.

5.2 The Divided Sample Dilemma

This type of sampling error has been observed in so many PAT implementations that it is worthy of some consideration. This type of mistake typically occurs when newcomers to PAT are sold an instrument with the prospect of it generating more results so that 'faster and more informed quality decisions' can be made.
In this case, a specimen (grab sample) has been taken from the factory floor and sent to the QC laboratory for reference analysis (primary sample). Here a sub-sample is taken (alas also by grab sampling) from the primary sample which is used to generate reference values [y] (secondary aliquot). Then a second sub-sample is drawn from the primary sample for generating data on the at-line PAT instrument; this is done by grab-sampling.

This setup results in just about a maximum of errors committed:

Mistake 1: The primary sample is a grab sample; it is therefore not representative of the lot.

Mistake 2: The two sub-samples (one of which in this case is actually the analytical aliquot) are also grab samples (grab sub-samples), and thus neither are representative.

Mistake 3: The 1:1 reference-to-PAT [y-X] correspondence is lost as the sub-sample used for reference measurement (y] is different from the one used to acquire the sensor signal [X].

The result of this fatally misguided approach typically results in decreased precision in the fit of the data to the model due to the inflated uncertainty induced by the various sampling errors. Figure 5 shows the impact of this practice on the resulting chemometric model.

The regrettable consequence behind this situation is that both PAT and chemometrics get a bad name and the typical advice provided by non-experts is to acquire more samples and add more components/factors to the model to improve precision.

This is poor advice indeed as chemometric modelling and preprocessing methods cannot model the heterogeneity- induced sampling errors.

Figure 6 illustrates the performance of the same prediction model after making the following four changes to the methodology,

1. Acquire a proper primary sample from a representative sampling device. The Theory of Sampling (TOS) to the fore, see e.g., (Masser, 1997; Esbensen and Swarbrick, 2018; Hökuldsson, 2024; Esbensen in this issue).
2. Generate the analytical aliquot(s) using a representative sub-sampling procedure and equipment (Massert, 1997; Esbensen and Swarbrick, 2018).
3. Scan the analytical aliquots on the PAT instrument using optimised spectral data acquisition parameters.
4. Acquire the reference data on the same aliquot used to generate the PAT sensor data.

Figs. 5,6 illustrates how the only way to improve model precision is through due diligence in the pre-analysis sampling/sub-sampling domain (Esbensen, 2025a). It is fair to state that this understanding is not all-persuasive in today's PAT realm!

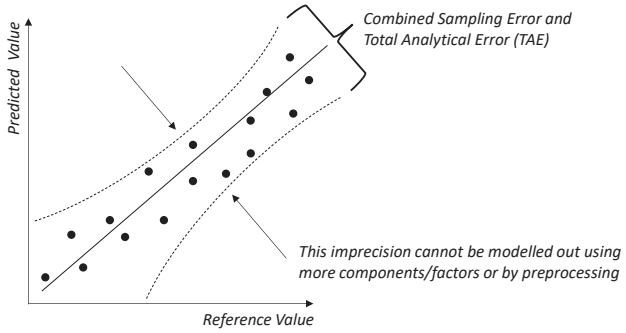

Figure 5: Impact of non-representative sampling/sub-sampling and the loss of the 1:1 reference-to-PAT correspondence.

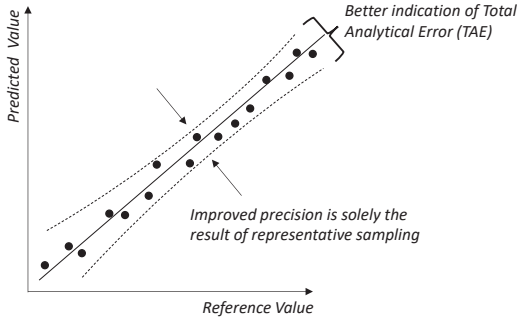

Figure 6: Prediction model generated using representative sampling and optimised spectral data acquisition parameters.

5.3 In-Line PAT Implementation - numerous problematic issues

This example represents a hybrid between at-line and in-line PAT sampling in the pharmaceutical industry. The manufacture of oral solid dose (OSD) formulations for pharmaceutical tablet production always requires a blending step of the powder components used for a specific formulation. In batch processes, this blending is typically performed using a rotating blending system ('V-blender' or IBC type) until a suitable endpoint has been reached. It is problematic that there are still organisations that attempt to sample such a 3-dimensional lot using non-representative insertion thief systems that not only generate biased results but also segregate the local blend through the introduction of a physical device into the powder bed. The use of this kind of device has resulted in endpoint times being set too high and this can result in degradation of the powder blend due to attrition, resulting in changes in particle size distributions and de-mixing phenomena (Muzzio, et al., 2003).

This problem has been somewhat reduced by the introduction of NIR spectrometers that can rotate with the blender and obtain a spectrum for the blend for each rotation. This single measurement represents a sensor grab sampling situation, but a new twist of aggregating a number of spectra, to generate a 'block', has led to implementation of the Moving Block Methods (MBM) approach for real-time detection of blend endpoints (Besseling, et al., 2015).

However, there are a number of optimisation steps that must be performed to generate reliable PAT data in this fashion of which the most important is block size, which must be validated with the aim to become representative of a unit dose prior to industrial implementation.

The major advantage of a PAT in-line NIR method is that there is now no disturbance of the powder bed during sampling. Provided the raw materials used in the blend are of similar particle size distribution, and the blender is not rotated too quickly, it is claimed from driven OEMs that repeatable estimates of blend endpoints can now be obtained. Perhaps ....

But there is still a remaining downside: There is still no way to extract a 1:1 correspondence sample for reference analysis. The European Medicines Agency (EMA) has addressed this situation in its NIR guidance document and terms such methods as "Dynamic PAT Methods"(European Medicines Agency, 2014).

In particular (excerpt from this guidance),

"Because PAT NIRS procedures are specific to the nature of the manufacturing processes (e.g. sampling frequency adapted to process dynamics), it is not appropriate to prescribe exact requirements for such procedures in this guideline.
An example of the use of NIRS in a PAT application is the monitoring of a powder blend for homogeneity. The blend may be monitored in terms of the measurement of the change of the NIR signal (e.g. its standard deviation) over time (also called moving block standard deviation (MBSD)), where this has been shown to be a valid measure of homogeneity."

Notice the improper use of the word 'homogeneity' in the context of the Theory of Sampling (TOS) in chapter six of Multivariate data Analysis: an introduction to multivariate analysis process analytical technology and quality by design (Esbensen and Swarbrick, 2018). But that aside, the guidance clearly states in section 5.1. that representative sampling of a 3-dimensional lot is difficult. But EMA states that this is impossible!

So, how can a quantitative measure of blend potency be made using PAT if proper sampling is impossible? This is where the risk-based approach of PAT must be considered. Figure 7 provides an example of a blend uniformity curve, measured by NIR spectroscopy combined with the Moving Block Standard Deviation (MBSD) measure.

The main assumption is that any point below the Endpoint Detection Limit represents a state of powder uniformity (N.B. not homogeneity!). This blend is now transferred to a compression room where the powder is either gravity fed, or vacuum transferred to the tablet press.

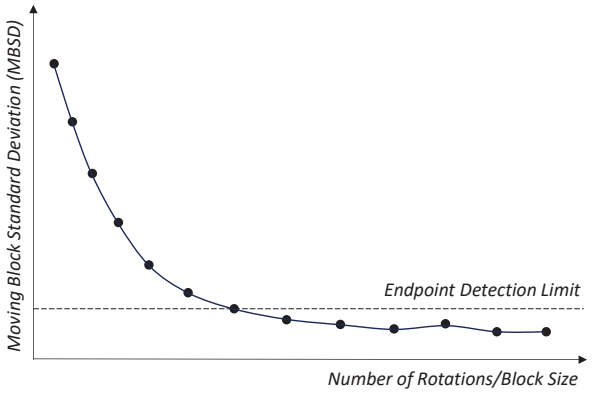

Figure 7: A blend uniformity endpoint curve using the Moving Block Standard Deviation (MSBD) measure.

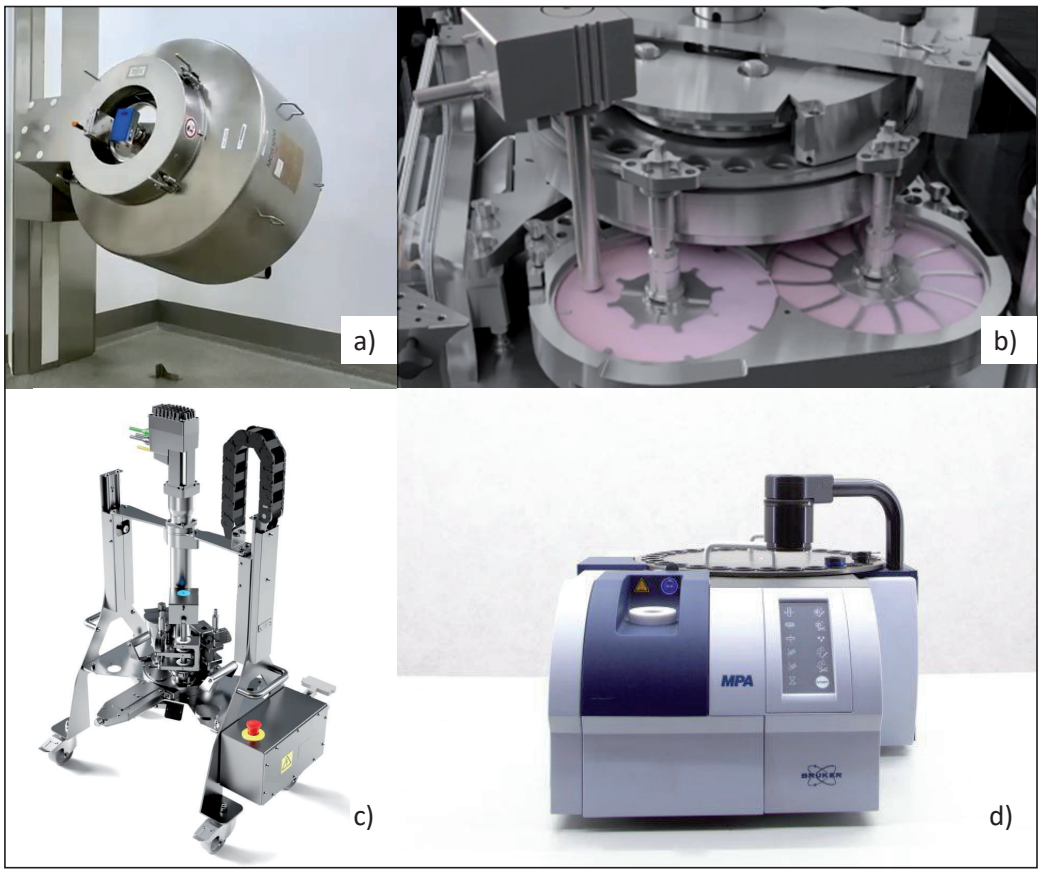

Figure 8: Sampling implementation strategy for assessing powder blend uniformity using PAT, a) NIR system attached to a rotating blender, b) NIR system implemented into a feed frame, c) a feed frame simulation device for calibration development and d) assessing pressed tablets using an at-line NIR spectrometer. *Caveat: examples are meant for illustration of principles only; no identification of company, personnel or equipment brand is intended).*

A tablet press can be considered to be a compacting spinning riffler (Romañach, 2015) and thus, a single tablet may be considered to be a representative sample. There is here a powerful opportunity in measuring a series of tablets and analyse the residual variability using a variogram (Esbensen and Swarbrick, 2018; Danish Standard, 2024). From a variogram can be estimated a Minimum Possible Error (MPE), which represents the minimum remaining total uncertainty due to sampling-and-analysis.

But there is even another approach than measuring discrete tablets - which is to place an in-line PAT NIR analyser into the feed frame of the tablet press. This is the spinning riffler section of the press and it is a bona fide representative sampling system (see e.g., the article by Romañach in this SST issue). To comply with the 1:1 correspondence between PAT measurement and reference analysis aliquot, there are two approaches,

1. A timing sequence can be implemented where a tablet(s) is extracted from the press which corresponds to the time the PAT measurement was made, or,
2. Use of a feed frame simulator (Expo Process Analytics, 2025) to make blends of various concentrations and measure them – 'as is' - in the feed frame using the full-scale on-line PAT device.

The impact of using a grab sampling approach for measuring powder characteristics in hoppers or chutes prior to tablet compression and analysing these samples using a reference method leads to the same adverse calibration model results as was shown in Figure 5 for the at-line case, whereas, using a feed frame simulator, or analysing 'timed' discrete tablets against the feed frame results in resolved calibration situations as depicted in Figure 6.

Figure 8 presents a potpourri of equipment and approaches for the entire process of calibration development using PAT for powder uniformity and quantitation. Caveat: examples are meant for illustration of principles only; no identification of company, personnel or equipment brand is intended).

The final implementation can be used as a PAT approach for two-fold assessment of blend and content uniformity of pharmaceutical powder blends, and can be summarised as follows,

1. Use the results of in-line NIR for endpoint detection of powder blending.
2. Develop a representative calibration on powder blends covering the specified analytical range using either a feed frame simulator or by implementing a PAT sensor directly into the full-scale feed frame and collecting compressed tablets for reference analysis.
3. Develop an at-line method based on tablets generated from production batches (and from batches made on the feed frame simulator as a back-up method).
4. Run the feed frame NIR during production to look for process trends and deviations.
5. Compare the batch process feed frame trends to the NIR blend uniformity data for joint assessment of blend uniformity and content uniformity.

6. Conclusions and Future Perspectives

It does not matter if a PAT approach is at-line, in-line or on-line, the principles of proper sampling and validation must be complied with in all cases. PAT analysers must be implemented and validated in the same way as any other analytical method, the main difference if that a PAT instrument is designed to measure continuously outside of a laboratory.

This requires that the PAT instrument is robust to its surroundings (see e.g., the article in this SST issue by Dusko Kadijevic) and must be positioned and implemented to be able to characterise a sample volume that is representative of the current state of the process – this can only be accomplished by a proper process sampling interface (Esbensen in this issue). Therefore, most of the development work behind chemometric PAT calibrations is often focused on acquiring a representative sample. There is absolutely nothing special about PAT in this regard.

The extraction of an analytical aliquot for sensor assessment has been the topic of endless debate and innumerable articles within the realm of PAT, but only a very few are based on full understanding of the three-domain complexities involved (sampling / analysis / data modelling), ibid.

The economically dominating pharmaceutical blend uniformity problem raises several issues that are also documented in foundational NIR method guidance documents. In particular, the extraction of aliquots using sample thieves results in the extraction of specimens with no certain provenance; and they disturb the uniformity of the entire 3-dimensional lot through induced segregation.

The blend uniformity problem requires some thinking outside the box and mandates correct usage of guidance documents to devise a risk-based strategy to achieve the critical objectives of assessing blend uniformity and content uniformity. A solution based on commercially available systems was indicated here, a solution that goes a long way to minimise sampling errors and focus the PAT measurements on the chemistry/biology of the samples being measured.

Future sampling systems must be designed where the exact aliquot measured can indeed be extracted from the sampling device without committing Incorrect Sampling Errors (ISE) (TOS) to improve the 1:1 sample to PAT correspondence. This will avoid situations where newcomers will be less tempted to use wrong preprocessing methods or revert to automated modelling scenarios based on AI that not only over-complicate the calibration development but are in reality black box in nature, are rarely interpretable, and cannot be properly validated. All these concerns will first be eliminated if/when a perfect process sampling interface is introduced – OEMs take notice!

Reliable PAT chemometric model development starts with selection of the optimal sensor technology for acquiring the most relevant multivariate sample spectra that are guaranteed representative of the contemporaneous process stream segments (see Esbensen in this issue) - and finishes with reliable prediction of analytical results, generated by properly calibrated and validated chemometric prediction models (Martens and Næs, 1991; Wold, 1995; Massert et al., 1997; Adams, 2004; Esbensen and Geladi, 2010; Esbensen and Swarbrick, 2018; Höskuldsson, 2024).

There is no short cut to the competences needed. The analytical domain notwithstanding, chemometricians have a lot to teach samplers about the power of multivariate spectra, but the opposite relationship: proper sampling before analysis, before data analysis is even more important. We are all in this together!

References

Adams, M.J. (2004). Chemometrics in Analytical Spectroscopy, 2nd ed. RSC Analytical Spectroscopy Monographs, Neil. W. Barnett (series ed.), Cambridge: Royal Society of Chemistry.

Besseling, R., Damen, M., Tran, T., Nguyan, T., can den Dries, K., Oostra, W., & Gerich, A. (2015). An efficient, maintenance free and approved method for spectroscopic control and monitoring of blend uniformity: The moving F-test. Journal of Pharmaceutical and Biomedical Analysis, 114, Oct.10, 471-481, DOI: 10.1016/j.jba.2015.06.019

Danish Standard (2024). Representive sampling—Horizontal standard (Repræsentativ prøvtagning—Horisontal standard) DS 3077:2024, Fonden Dansk Standard, Nordhavn, DK.

Esbensen, K.H. (2025a). Data Quality: Importance of the 'Before Analysis' Domain Theory of Sampling (TOS), Journal of Chemometrics 39 (4), DOI:10.1002/cem.70021

Esbensen, K.H. (2025b). Augmented Scope and Didactics for Initiation to the Theory of Sampling (TOS): Three domains behind valid data quality. Sampling Science & Technology, 3, 55-67. https://www.sst-magazine.info/issues/sst-003/article/augmented-scope-and-didactics-for-initiation-to-the-theory-of-sampling-tos-three-domains-behind-valid-data-quality/

Esbensen, K.H. & Geladi, P. (2010). Principles of Proper Validation: use and abuse of re-sampling for validation. Journal of Chemometrics, 24,168-187, DOI: 10.1002/cem.1310

Esbensen, K.H., Swarbrick, B. (2018). Multivariate Data Analysis: An introduction to multivariate analysis, process analytical technology and quality by design. 6th ed. Camo, Oslo, No. ISBN. 9788269110401

European Medicines Agency (2014). Guideline on the use of near infrared spectroscopy by the pharmaceutical industry and the data requirements for new submissions and variations, Committee for Human Medicinal Products (CHMP), London, UK.

Expo Process Analytics (2025). FFSIM Feed Frame Simulator; Cork, Ireland https://www.expoprocessanalytics.com/pat-systems/feed-frame-simulator/

Flåten, G. (2018). Model Maintenance, In Ana P. Ferreira, Jose C. Menezes, & Mike Tobyn (eds.) Multivariate Analysis for Pharmaceutical Industry, Academic Press, 313-321. ISBN13: 9780128110652.

Höskuldsson, A. (2024). PLS multi-step regressions in data paths. Chemometrics intelligent laboratory systems, 251, 105167

ICH (2008). International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use, Pharmaceutical Quality System Q10. European Union, Japan, United States. https://database.ich.org/sites/default/files/Q10%20Guideline.pdf

ICH (2022). International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. Analytical Procedure Development. Q14. ICH regions. https://database.ich.org/sites/default/files/ICH_Q14_Guideline_2023_1116_1.pdf

Martens, H. & Næs, T. (1991). Multivariate calibration, reprinted w/ corrections (April), Wiley, Chichester, UK ISBN. 0 047 90979 3

Massart, D.L. (ed.) (1997) Handbook of Chemometrics and Qualimetrics. Data handling in science and technology. Vol 20A, Amsterdam: Elsevier. ISBN 0.444.89724.0

Muzzio, F.J., Alexander, A., Goodridge, c., Shen, E., Shinbrot, T. (2003). Solid Mixing Part A, in Edward L. Paul, Victor A. Atieno-Obeng, & Suzanne M. Kresta (eds.) Handbook of Industrial Mixing: Science and Practice, Wiley, Hoboken, N.J. ISBN 0-471-26919-0

Romañach, R.J. (2015). Sampling and Determination of Adequacy of Mixing, in P.J. Cullen, Rodolfo Romañach, Nicolas Abatzoglou, Chris D. Riley (eds.) Pharmaceutical Blending and Mixing, Wiley & Sons, 57-58. DOI: 10.1002/9781118682692

Wold, S. (1995). Chemometrics; what do we mean with it, and what do we want from it? Chemometrics and Intelligent Laboratory Systems, 30(1), 109-115. https://www.sciencedirect.com/science/article/abs/pii/0169743995000429