Journal Pre-proofs

Simultaneous prediction of the API concentration and mass gain of film coated tablets using Near-Infrared and Raman spectroscopy and data fusion

Bence Szabó-Szőcs, Máté Ficzere, Orsolya Péterfi, Dorián László Galata

PII: S0378-5173(24)01191-8
DOI: https://doi.org/10.1016/j.ijpharm.2024.124957
Reference: IJP 124957

To appear in: *International Journal of Pharmaceutics*

Received Date: 13 June 2024
Revised Date: 28 October 2024
Accepted Date: 13 November 2024

Please cite this article as: B. Szabó-Szőcs, M. Ficzere, O. Péterfi, D.L. Galata, Simultaneous prediction of the API concentration and mass gain of film coated tablets using Near-Infrared and Raman spectroscopy and data fusion, *International Journal of Pharmaceutics* (2024), doi: https://doi.org/10.1016/j.ijpharm.2024.124957

Simultaneous Prediction of the API concentration and mass gain of film coated tablets using Near-Infrared and Raman Spectroscopy and Data Fusion

Bence Szabó-Szőcs[1], Máté Ficzere[1], Orsolya Péterfi[1], Dorián László Galata[1*]

1 Department of Organic Chemistry and Technology, Faculty of Chemical Technology and Biotechnology, Budapest University of Technology and Economics, Műegyetem rkp. 3., H-1111 Budapest, Hungary

*Correspondence galata.dorian.laszlo@vbk.bme.hu, Tel.: +36 1 463 5881

Abstract

This study investigates the simultaneous prediction of active pharmaceutical ingredient (API) concentration and mass gain in film-coated tablets using Partial Least Squares (PLS) regression combined with three data fusion (DF) techniques: Low-Level (LLDF), Mid-Level (MLDF), and High-Level (HLDF). Near-Infrared (NIR) and Raman spectroscopy were utilized in both reflection and transmission modes, providing four types of spectral data per tablet. Transmission models proved more effective for API prediction by capturing data from the entire tablet, while reflection models excelled in assessing mass gain by focusing on the surface layer. Among the DF strategies, MLDF with Principal Component Analysis (PCA) offered the most significant improvements in predictive accuracy by filtering out irrelevant information. Variable selection methods further enhanced model performance by reducing the number of latent variables required. Overall, the integration of multiple spectral datasets and DF techniques resulted in models that gave predictions for evaluation samples with lower errors, demonstrating their potential to optimize quality control in pharmaceutical manufacturing.

1. Introduction:

Orally administered solid dosage forms are the simplest and most commonly used formulations in the pharmaceutical industry, providing numerous advantages, including relatively easy manufacturing, cost-effectiveness, and high patient compliance. Tablets are the most important and widely used members of this group [1, 2]. In recent decades, it also became common to utilize film coatings to enhance certain properties of the tablets [3].

Film coating typically occurs in the final stages of pharmaceutical manufacturing. During this process, the surface of the products is covered with a continuous and uniform film layer [4]. The purpose of film coating can include modifying the kinetics of drug release (e.g., achieving extended drug release), protecting the drug and product from physical and environmental factors (mechanical damage, light, air, moisture, gastric acid, etc.), improving patient compliance through the enhancement of tablet appearance (e.g., colored coating), masking unpleasant tastes and odors, and making swallowing easier [5].

Film coating methods offer great flexibility, allowing for a wide range of products to be coated, such as tablets, granules, powders, capsules, and nonpareils. Film coating is typically applied gradually to a moving mass of product, usually using a spray atomization technique. The success of the process can be attributed to various factors [6]. Firstly, by applying a coating of only 2-3% of the tablet core weight, the attributes of the product can be significantly altered. Moreover, film coating opens up opportunities for branding and identification, further establishing its significance in the pharmaceutical industry [7, 8]. Tablet film coating is a commonly used but critical process that provides various functions to tablets, thereby meeting

different clinical needs. The evolution of dosage forms with film coatings is based on the development of coating technology, equipment, analytical techniques, and coating materials [9].

Two very important quality attributes of film-coated tablets are their active pharmaceutical ingredient (API) concentration and the mass gain that occurs during the film coating process. The cores must be adequately layered to ensure the product's safety and efficacy[10]. Acquiring real-time information during the process helps to enhance the production process and to detect possible problems [11, 12]. This can be accomplished using process analytical technology (PAT) tools, as recommended by the Food and Drug Administration in a guidance in 2004 [13].

In the past decade, near-infrared (NIR) [14-16] and Raman spectroscopy [10, 17-19] have become increasingly utilized for assessing critical attributes in pharmaceutical processing [20]. These techniques offer rapid, non-destructive analysis without requiring sample preparation, and their growing use is driven by their ability to provide multivariate qualitative and quantitative data [21, 22]. Full-spectrum calibration methods like partial least-squares (PLS) regression have been widely validated for developing fast spectral screening techniques [23, 24].

The collected data frequently encompasses extraneous variables that necessitate segregation from the primary variables. Algorithms for variable selection (VS) eradicate noisy spectral segments and redundant data to enhance predictive precision. VS improves the understanding and interpretability of these multivariate classification models [25]. The interval partial least squares (iPLS) is a local regression method, which can give prediction models with improved precision by selecting the optimum interval for the spectral data. Genetic algorithm (GA) is also a suitable method for selecting wavelengths in PLS, enabling the calibration of mixtures with nearly identical spectra without compromising prediction capacity, utilizing spectrophotometric data [26-28]. Extreme Gradient Boosting (XGBoost) is a scalable end-to-end tree boosting system, which is widely utilized by data scientists to achieve state-of-the-art results on numerous challenges in machine learning [29, 30].

These advanced analytical platforms that are readily accessible offer extensive and diverse datasets linked to manufacturing processes that can be used for monitoring and predictive purposes [31-34]. Data fusion (DF) refers to the process of combining multiple data sources, typically to increase the precision and accuracy of downstream predictive models. It has become a popular method in recent years due to the increased use of various spectroscopic analysis techniques [35]. Casian et al. gave an overview of opportunities of implementing DF in PAT [31]. For example Zomer et al. employed chemometric methods for multivariate statistical process modeling to oversee the ongoing wet granulation tableting process of a pharmaceutical product currently under development [36]. A study by Casian et al. represents an investigation of NIR spectroscopy, Raman spectroscopy, colorimetry and image analysis methods which were tested and compared considering the ability to quantify the API concentration and to detect production errors [33].

This study explores the use of NIR and Raman spectroscopy to simultaneously predict API concentration and mass gain in film-coated tablets. To accommodate different dosages with a single model, it is crucial that the model remains robust to variations in both coating thickness and API concentration. Both spectroscopic techniques were employed in reflection and transmission modes, resulting in four measurements per tablet. Data fusion techniques, which

have not been thoroughly tested before for film-coated tablets, were applied and compared to identify the most reliable modeling approach. Exploring various combinations of data fusion may lead to the development of a more robust model that is applicable across different dosages and unaffected by variations in API concentration. This is particularly important for products manufactured in multiple dosages, where the model must predict a quality attribute other than API content, despite changes in the API signal.

2. Materials and methods

2.1. Materials

In this study, anhydrous caffeine (BASF, Ludwigshafen, Germany) was used as model API. The formulated tablets contained three more excipients: microcrystalline cellulose (MCC) (Vivapur grade 200, JRS Pharma GmbH, Rosenberg, Germany) as filler, croscarmellose sodium (Ac-Di-Sol®, FMC BioPolymer, DuPont, Leiden, Netherlands) as disintegrant promoting dissolution and magnesium stearate (MgSt) (grade S, Faci S.P.A., Carasco, Italy) as lubricant to protect the tablet press from mechanical degradation due to repeated usage.

The material Opadry® QX (Colorcon, Budapest, Hungary) was used for the film coating, which is a commercially available preformulated yellow colored polyvinyl alcohol (PVA) based coating material. It is used to enhance the stability, appearance, and patient acceptability of tablets and capsules.

2.2. Methods

2.2.1. Sample preparation and tableting

In this research, 15% caffeine concentration was selected as the target, and six other formulations were prepared with smaller and higher concentrations. The target formulation of this study was defined as a 200 mg tablet containing 11 to 19% w/w API. The composition of various tablets is presented in *Table 1*.

Table 1 The composition of tablets used in the study.

Batch nr.	API (%w/w)	MCC (%w/w)	Ac-Di-Sol® (%w/w)	MgSt (%w/w)
1	11	83	5	1
2	13	81	5	1
3	14	80	5	1
4	15	79	5	1
5	17	77	5	1
6	18	76	5	1
7	19	75	5	1

The concentrations of Ac-Di-Sol® and MgSt were kept at the same amount to ensure that the final properties of the tablets remain consistent.

To prepare the tableting blends, a mixture of 130 g containing caffeine, MCC, and Ac-Di-Sol was manually blended in a bottle for 5 minutes. Following this step, MgSt was added, and the mixtures were further blended for an additional 1 minute.

Tableting was carried out in a Dott Bonapace CPR-6 (Dott Bonapace, Limbiate, Italy) eccentric tablet press. Biconvex tablets were produced by direct compression with a target compression force of 14 kN in automatic mode using a gravity feeder with a moving shoe and 9 mm round-shaped punches.

The hardness of the tablet cores was measured using a Dr. Schleuniger THP-4M crushing strength tester (Dr. Schleuniger Productronic, Switzerland). Five tablets were assessed from each formulation to verify and validate the adherence of tablet production.

2.2.2. Film coating

Film coating was carried out in a Glatt GB2 L50-10026 pan coating machine (Glatt GmbH, Binzen, Germany) equipped with a perforated drum and winglike baffles were installed on it to facilitate the thorough mixing of the tablet bed and ensure uniform coating. *Table 2* contains the values of parameters used during the coating process. The coating dispersion was prepared by mixing 90 g Opadry® QX powder with 360 mL distilled water for 30 min. A spray nozzle (diameter: 0.8 mm) was integrated into the set-up, which generated an elliptical spray pattern of the mixture.

Table2 Parameters of the film coating.

Process parameters	**Units**	**Value**
Coating liquid concentration	% w/w	20
Coating liquid mass flow rate	g/min	6.5
Drum rotation speed	rpm	20
Inlet air temperature	°C	60
Air out temperature	°C	40
Atomizing air pressure	bar	2
Inlet air flow rate	m^3/h	50
Air out flow rate	m^3/h	55

Throughout the film coating process, 100 g of the prepared tablets were mixed with placebo tablets of a different size to obtain a batch mass of 800 g, which is optimal for the coater. To obtain tablets with different amounts of coating, seven samples were collected at regular intervals during the procedure, with each sample containing five tablets. The process was stopped after 30 minutes. This way tablets were obtained with mass gains 0% and 6.14%, these values cover a wide range around 3%, similar mass gains are common in formulations with non-functional coatings.

The coating was performed with all seven formulations with different API concentrations. In addition, tablet cores were also retained for each formulation, therefore there were eight samples from each batch. The 5 tablets from each collected sample (total of 280 tablets) were analysed by offline NIR and Raman measurements.

2.2.3. NIR spectroscopy

In this study, near-infrared (NIR) spectra were acquired using a Bruker MPA Multi-Purpose FT-NIR Analyzer equipped with OPUS 7.5 software (Bruker OPTIK GmbH, Ettlingen, Germany), operating in both diffuse reflection and transmission measurement modes. The diffuse reflection spectra were collected using an integrating sphere, a high-intensity NIR source (Tungsten) and PbS detector. The obtained spectral range was 4000–12000 cm^{-1} with a resolution of 8 cm^{-1} and 32 scans were accumulated with 10 kHz scanner velocity. For measurement of transmission spectra an external transmission unit equipped with an InGaAs external detector was employed, focusing on the spectral range of 7000–15000 cm^{-1}. The operational parameters remained consistent with those employed for reflection measurements. For both measurement modes these parameters mean about 15 s spectral accumulation time.

2.2.4. Raman spectroscopy

Reflection and transmission Raman spectra were obtained utilizing a Kaiser Raman Rxn2® Hybrid in situ analyzer (Kaiser Optical Systems, Ann Arbor, MI, USA) equipped with a 400 mW, 785 nm diode laser (Invictus) and the PhAT (Pharmaceutical Area Testing) probe. For reflection measurements, tablets were illuminated from above through the PhAT probe, and Raman signals were collected using the same probe. The diameter of the laser spot size was adjusted to 6 mm and the nominal focal length was 250 mm. In the transmission mode, the Kaiser transmission accessory was positioned under the objective stage holding the tablets while Raman photons were collected by the PhAT probe positioned above the stage. Reflection and transmission spectra were acquired with 10 s and 45 s illumination time, respectively, with two repetitions. The studied spectral range spanned from 200 to 1890 cm^{-1} with a resolution of 4 cm^{-1}, which provided 1690 variables during data processing. Before each set of measurements, a dark spectrum was obtained and automatically subtracted from the acquired raw spectra.

2.2.5. Measurement of the API concentration by UV spectrometry

UV spectroscopy was used for validation of the caffeine concentration prediction, which was applied as a reference analytical method for verifying the predictions of the NIR and Raman spectroscopy measurements. For this measurement, 70 tablets were separated, ensuring that each batch was represented, and the API concentration of each tablet was measured. The mass of the tablets was weighed, then they were crushed in a mortar and pestle and washed into 1 L volumetric flasks. The flasks were filled up to 1 L with distilled water and placed on magnetic stirrers for 30 min. The solutions were filtered through a 0.45 µm filter before the UV analysis. An Agilent 8453 UV/VIS spectrometer (Hewlett-Packard, Palo Alto, CA, USA) was used to measure the API concentration at 272 nm in a 10 mm cuvette.

The calibration curve demonstrated linearity over the range of 0–50 mg/L caffeine concentration ($R^2 = 0.9996$). Validation tests of this method were also carried out in accordance with ICH Q2(R2) guideline [37]. Precision was demonstrated through repeatability, meaning the measurements were repeated over three consecutive days. The concentration of caffeine was measured in a solution containing only caffeine and in another solution that also included excipients in the appropriate amounts. Since these excipients do not affect the measurement, specificity was confirmed. The results of these tests are shown in the supplementary: *Table s1*

2.2.6. Coating thickness measurement with machine vision

Therefore, a machine vision-based method that was presented in one of our earlier works by Ficzere et al. was utilized [38]. This method involves acquiring images of the tablet from above with illumination from below. This way, the outline of the tablet can be seen in good detail. This enables the quantification of the diameter of the tablet, and consequently, the thickness of the coating. Using MATLAB software, tablet diameters were measured along with a 10 mm length on a scale to obtain the pixel-to-distance ratio. From the obtained pixel values, the tablet diameters were calculated. The diameter of the uncoated tablet was subtracted from that of the coated tablets, and by dividing this value by two, the thickness of the coating was determined.

2.3. Data analysis

All spectroscopic data analysis and prediction modeling were implemented in MATLAB R2022a (MathWorks, USA) using the PLS Toolbox 8.8.1. (Eigenvector Research, USA).

2.3.1. Preprocessing methods

The NIR and Raman spectra were preprocessed and different spectral regions were examined before building the predictive models [39, 40]. In the case of NIR spectra, standard normal variate (SNV), Savitzky-Golay derivative (DV, 15 points/window, second order smoothing), multiplicative scatter correction (MSC) and mean centering (MC) were applied, while for Raman spectra Automatic Whittaker Filter baseline correction (BLC) with asymmetry parameter $p = 0.001$ and smoothing parameter $\lambda = 10^5$, Savitzky-Golay smoothing, normalization methods (MSC, SNV) and MC were used. Due to their poor signal-to-noise ratio, some spectral regions were omitted from the analysis. The *Table 3* contains the combination of preprocessing methods and the examined regions yielding the best results.

Table 3 Preprocessing methods for each spectrum.

	Preprocessing	Range (cm^{-1})
Reflection NIR	SNV, MC + DV in case of models for weight gain	7700-3800
Transmission NIR	SNV, MC	15000-7500
Reflection Raman	BLC, SNV, MC	350-1890
Transmission Raman	BLC, SNV, MC	350-1890

BLC: baseline correction; DV: derivative; MC: mean centering; SNV: standard normal variate.

2.3.2. Partial Least Squares regression

A total of 280 samples (with various API concentrations) were divided into a training set (CAL - 210 tablets) for calibrating the PLS models, with cross-validation samples selected by the PLS algorithm from this set, and an independent external test set (TEST - 70 tablets) to evaluate the predictive ability of the models. This division is demonstrated in *Table 4*:

Table 4 Division of samples into calibration and test groups.

Sample / Batch	0 min	4 min	8 min	12 min	16 min	20 min	24 min	28 min
11%	CAL	CAL	TEST	CAL	CAL	TEST	CAL	CAL
13%	TEST	CAL	CAL	CAL	CAL	CAL	CAL	TEST
14%	CAL	CAL	CAL	TEST	CAL	TEST	CAL	CAL

15%	CAL	CAL	CAL	CAL	TEST	CAL	TEST	CAL
17%	CAL	CAL	TEST	CAL	CAL	CAL	CAL	TEST
18%	TEST	CAL	CAL	TEST	CAL	CAL	CAL	CAL
19%	CAL	CAL	CAL	CAL	TEST	CAL	TEST	CAL

PLS regression models were created to predict the API concentration and the weight gain of tablets during the film coating process. The PLS algorithm transforms the original data space to maximize the covariance between the spectral dataset (X) and the dependent data (Y variable) through the creation of new variables known as latent variables (LVs). The models were contrasted by the root mean square error of calibration, cross-validation, and prediction (RMSEC, RMSECV, RMSEP) and the coefficient of determination for calibration, cross-validation, and prediction (R^2 C, R^2 CV, R^2 P).

The coefficient of determination quantifies how much of the variability in the dataset is explained by the model (Eq.1-2) [41]:

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad (1)$$

$$\bar{y}_i = \sum_{i=1}^{n} \frac{y_i}{n} \quad (2)$$

where RSS is the residual sum of squares, TSS is the total sum of squares, y_i is the actual value of the variable, $\hat{y}_i$ is the value predicted by the model, $\bar{y}_i$is the average value of the variable and *n* is the number of samples. The value of this attribute falls within the range of 0 to 1, with higher values indicating a better fit.

The Root Mean Squared Error can be computed using the following equation (Eq.3)[42]:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \quad (3)$$

In Equations (3) the symbols have the same meaning as in Equation (1). Low RMSE values imply a small amount of unexplained variance and a good fit.

2.3.3. Variable selection

In this research, the effectiveness of variable selection was tested to evaluated whether it can improve the predictive ability of PLS regression models. The uninformative and noise-affected variables have been excluded using separately two VS methods: interval-PLS (iPLS) and genetic algorithms (GA). iPLS was utilized in forward mode with an automatic step size, and the interval size was set to 30. As for GA, the window width was adjusted to 30, and the maximum number of LVs was set to 6.

2.3.4. Data fusion

Data fusion methodologies are typically categorized into three overarching strategies, contingent upon the manner in which fusion occurs: low-level (LLDF), wherein raw data serves as the input for the fusion process; mid-level (MLDF), characterized by the utilization of extracted features from the data; and high-level (HLDF), which involves the combination of data at the classification or prediction decision level [34]. Considering the complementary attributes of the examined PAT instruments, all three data fusion strategies were investigated to enhance the predictive efficacy of the models. During model evaluation and construction,

all four types of datasets and their combinations were analyzed. This involved fusing the datasets in pairs, triplets, and ultimately combining all four, as illustrated in *Figure 1*. Hence, every possible combination was investigated.

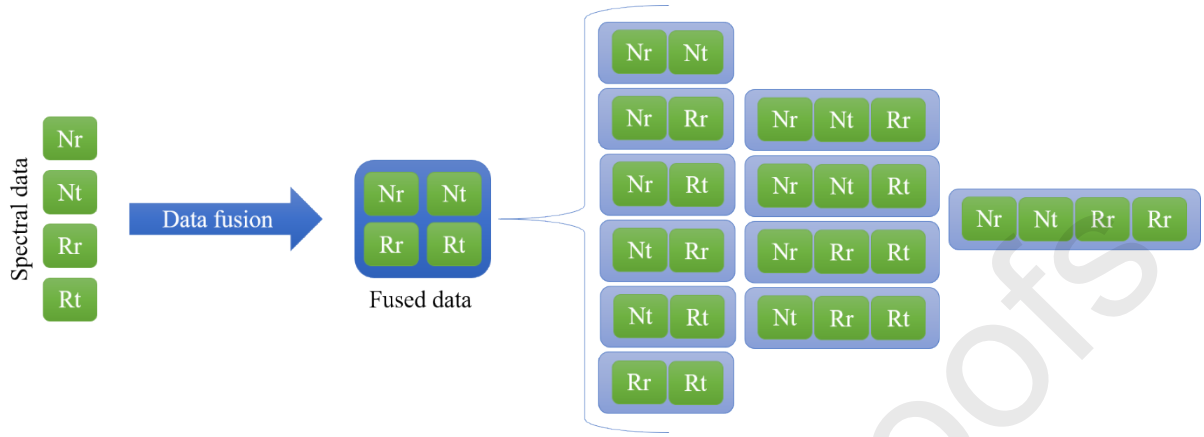

Figure 1: Combination of all four spectral data: reflection NIR (Nr), transmission NIR (Nt), reflection Raman (Rr) and transmission Raman (Rt).

2.3.4.1. Low-level data fusion

In LLDF, the pretreated spectra were concatenated into a single vector, where the variables coming from different sources are placed one after the other. Mean centering was applied for the fused data and then this was used as input of the PLS models. This merger method was applied to both the calibration and test datasets. VS methods were employed on these combined inputs, resulting in 3 models for each combination: one without VS, one with iPLS and one with GA.

2.3.4.2. Mid-level data fusion

The difference between LLDF and MLDF is the process of feature selecting in MLDF. However, the key issue on this level of DF is the extraction and screening of the most significant features, given the limited number of scores that can be obtained. Principal component analysis (PCA) was utilized for feature extraction. During PCA, the original $n \times \lambda$ size dataset (where n is the number of samples and λ is the number of variables, such as wavenumbers) undergoes a coordinate transformation such that the new variables are orthogonal to each other, with the first few variables representing the highest variance in the dataset. As a result, each spectrum was described with 10 principal components (PCs), and these PC values were then combined into a vector to serve as the input for the PLS models. At this level of DF, no VS was applied.

2.3.4.3. High-level data fusion

In HLDF, a supervised learning model, specifically a regression model, is initially applied to the spectra to predict the desired output variable. Subsequently, these predictions are utilized as input in another regression model, which has the potential to correct any inaccuracies made by the models and produce a more accurate final prediction. For this purpose, the XGBoost (XGB) algorithm, based on an ensemble of gradient boosted regression trees, was selected. In this phase of data fusion, the input assessment involves predictions derived from fundamental PLS models and their various combinations. The eta parameter (learning rate) for XGB was initially set to 0.15 and then gradually decreased to 0.05. In this scenario, different aspects of variable selection were employed, similar to those observed in LLDF.

Figure 2: Schematic representation of data analysis workflow.

3. Results and discussion

3.1. Physical characterization of tablets

The weight and crushing strength of five tablets from each formulation was measured. The measurement results and the mean weight of the tablet cores, as well as the relative standard deviations, are displayed in *Table 5*.

Table 5 Weight and crush strength measurements. RSD: relative standard deviation.

Batch nr.	Mean weight of tablet cores (mg)	RSD of tablet weight (%)	Mean crushing strength (N)	RSD of crushing strength (%)
1	219.9	0.21	50.4	3.60
2	202.3	0.62	51.2	6.08
3	198.7	1.49	50.2	11.73
4	202.7	1.06	54.6	4.01
5	199.5	0.63	49.0	5.95
6	201.7	0.65	47.6	3.82
7	202.3	0.53	50.2	6.52

For the prediction of mass gain during film coating, tablets from the samples were weighed using an analytical balance. The experimental mass of coating materials was determined by subtracting the average weight of tablet cores from the average weight of coated tablets. The results are provided in *Table 6*.

Table 6 Final mass gain after film coating

Batch nr.	Average weight of uncoated tablets (mg)	Average weight of film-coated tablets (mg)	Mass gain (%)
1	219.9	229.9	4.58
2	202.3	211.3	4.47
3	198.7	210.7	6.06

4	202.7	212.2	4.70
5	199.5	211.7	6.14
6	201.7	210.4	4.33
7	202.3	209.7	3.64

To verify that there is a continuous mass increase during the film coating process, the coating thickness was determined using a machine vision measurement method for the batch containing 15% caffeine. Five tablets were examined from each sample, and their average coating thickness is presented in the following table:

Table 7 The average diameter and the coating thickness of the tablets from the batch of 15% caffeine

Sample	Average diameter (mm)	Coating thickness (µm)
0	9.612	-
1	9.636	11.8±3.1
2	9.647	17.3±1.6
3	9.657	22.7±1.5
4	9.663	25.4±2.6
5	9.682	35.0±3.4
6	9.704	46.0±1.2
7	9.714	50.9±4.3

In the case of building the models developed for the determination of API concentration, the API concentration of the samples from the validation sets was measured as discussed in Section *2.2.5. Measurement of API Concentration by UV Spectrometry*.

3.2. Data handling and model optimisation

In this study, reflection and transmission NIR and Raman spectra of caffeine tablets were collected; thus, four measurements were conducted for each tablet. A total of 280 samples were divided into a training set (210 tablets) to calibrate the PLS models and a validation set (70 tablets) to test the predictive ability of the models. A similar division is also used in data analysis in comparable studies [19, 20].

PLS regression models were constructed with the calibration models developed by correlating the preprocessed spectra. The selection of LVs and the spectral range impacting the model aimed to minimize the RMSEP. The optimal number of LVs was determined individually for each model.

3.3. Single spectra PLS models

During the initial phase of result evaluation, individual PLS models were developed separately for predicting API concentration (PLS-API) [43, 44] and mass gain of the tablets (PLS-MG) [45, 46]. In both cases, models were constructed using four spectroscopic methods: reflection NIR (Nr), transmission NIR (Nt), reflection Raman (Rr), and transmission Raman (Rt). Subsequently, variable selection procedures were tested to assess whether they improve predictive ability.

The parameters of the models are compared using table heatmap, as shown in *Figure 3*. The values fluctuate between red and green hues in every column. The intensity of these hues serves as an indicator of model quality: green hues refer to higher quality, while red indicates lower quality:

PLS models	Model	Variable selecion	LV	RMSEC (%)	RMSECV (%)	RMSEP (%)	R^2 Cal	R^2 CV	R^2 Pred
PLS-API	Nr	-	6	1.099	1.160	1.280	0.831	0.812	0.812
		IPLS	6	1.131	1.175	1.402	0.821	0.807	0.768
		GA	6	0.884	0.953	1.084	0.891	0.873	0.861
	Nt	-	6	0.718	0.806	0.852	0.928	0.909	0.917
		IPLS	6	0.700	0.776	0.849	0.931	0.916	0.917
		GA	4	0.756	0.784	0.880	0.920	0.914	0.916
	Rr	-	3	1.106	1.189	1.082	0.829	0.802	0.878
		IPLS	3	1.028	1.138	1.047	0.852	0.819	0.890
		GA	3	0.996	1.034	1.092	0.861	0.850	0.865
	Rt	-	2	0.856	1.063	0.868	0.897	0.843	0.923
		IPLS	2	1.034	1.067	0.938	0.850	0.841	0.911
		GA	2	0.784	0.848	0.949	0.914	0.899	0.900
PLS-MG	Nr	-	6	0.233	0.283	0.357	0.973	0.960	0.947
		IPLS	4	0.240	0.268	0.371	0.971	0.964	0.939
		GA	3	0.259	0.267	0.365	0.967	0.965	0.944
	Nt	-	6	0.402	0.570	0.644	0.920	0.841	0.813
		IPLS	5	0.377	0.411	0.442	0.929	0.916	0.908
		GA	5	0.392	0.450	0.543	0.924	0.900	0.864
	Rr	-	6	0.147	0.340	0.416	0.989	0.943	0.921
		IPLS	6	0.272	0.324	0.433	0.963	0.948	0.913
		GA	4	0.324	0.360	0.485	0.948	0.935	0.889
	Rt	-	3	0.638	0.745	0.788	0.797	0.724	0.714
		IPLS	3	0.546	0.576	0.708	0.852	0.835	0.767
		GA	3	0.516	0.550	0.836	0.868	0.849	0.686

Figure 3: Table heatmap of PLS-API and PLS-MG models for comparison of various spectroscopic methods and variable selection approaches.

The results of PLS-API models indicate that VS improves the performance metrics of the model in certain instances, yet deteriorates them in numerous cases However, VS consistently reduces the number of required LVs. The best outcomes are seen in the Nt and Rt models. Considering the number of latent variables, the Rt model has proven to be much more suitable. This is likely attributed to transmission-based measurements gauging the radiation intensity traversing the tablet, thus examining the entire tablet. In these cases, VS decreased the accuracy of the models. The reflection models are presumably noticeably worse because the coating strongly influences the signal obtained with it, and since the quantity of this coating fluctuates significantly, it may disrupt the estimation of the active ingredient.

In the case of PLS-MG models, similar to PLS-API models, it can be said that VS methods do not always improve the performance metrics; in fact, they can deteriorate them. It is observable that the best values were obtained for the reflection models, especially the Nr models. This is presumably because, in reflection spectroscopy measurements, the rays reflected from the sample are assessed, yielding information about regions near the surface of the tablet, which includes the film layer.

3.4. Data fusion

Informed by the studies available in the literature [31-35], this work systematically explored all three DF strategies to enhance predictive performance. While single spectra PLS models provided satisfactory results, integrating information from various spectroscopic techniques allowed for regression models with lower prediction errors. By combining all four spectra based on selected spectral regions and preprocessing methods, DF strategies were employed to establish robust models with improved accuracy.

The individual PLS models and the DF models were compared based on RMSEP values, a parameter selected for its ability to best demonstrate the predictive capability of the regression model. All other parameters can be viewed in the supplementary material: *Table s2-s3*.

		PLS-API			PLS-MG		
Model	**VS**	**LLDF**	**MLDF**	**HLDF**	**LLDF**	**MLDF**	**HLDF**
Nr (single spectra)	-	1.280	1.279	1.400	0.357	0.477	0.374
	IPLS	1.402	-	-	0.371	-	-
	GA	1.084	-	-	0.365	-	-
Nt (single spectra)	-	0.852	0.839	1.171	0.644	1.877	0.648
	IPLS	0.849	-	-	0.442	-	-
	GA	0.880	-	-	0.543	-	-
Rr (single spectra)	-	1.082	1.118	1.303	0.416	0.480	0.486
	IPLS	1.047	-	-	0.433	-	-
	GA	1.092	-	-	0.485	-	-
Rt (single spectra)	-	0.868	0.885	0.939	0.788	2.001	0.740
	IPLS	0.938	-	-	0.708	-	-
	GA	0.949	-	-	0.836	-	-
Nr+Nt	-	0.915	0.836	1.020	0.500	0.915	0.344
	IPLS	0.994	-	1.185	0.549	-	0.501
	GA	0.827	-	0.947	0.424	-	0.485
Nr+Rr	-	1.019	0.809	0.860	0.397	0.342	0.359
	IPLS	1.098	-	1.220	0.555	-	0.439
	GA	1.112	-	1.155	0.491	-	0.342
Nr+Rt	-	0.839	0.820	0.919	0.839	0.504	0.369
	IPLS	0.802	-	0.959	0.478	-	0.691
	GA	0.794	-	0.879	0.521	-	0.435
Nt+Rr	-	0.893	0.917	0.969	0.399	0.577	0.482
	IPLS	1.137	-	1.335	0.447	-	0.491
	GA	0.933	-	1.133	0.437	-	0.463
Nt+Rt	-	0.905	0.777	1.008	1.608	1.887	0.630
	IPLS	0.820	-	0.998	0.941	-	3.353
	GA	0.871	-	0.945	1.385	-	1.359
Rr+Rt	-	0.810	0.823	0.899	0.577	0.472	0.490
	IPLS	0.809	-	0.833	0.463	-	0.386
	GA	0.812	-	0.872	0.535	-	0.543
Nr+Nt+Rr	-	0.845	0.770	0.931	0.369	0.503	0.359
	IPLS	1.049	-	1.249	0.421	-	0.479
	GA	0.836	-	0.973	0.511	-	0.380
Nr+Nt+Rt	-	0.835	0.729	0.872	0.792	0.920	0.360
	IPLS	0.795	-	0.899	1.722	-	1.940
	GA	0.781	-	0.864	0.392	-	0.674
Nr+Rr+Rt	-	0.775	0.786	0.871	0.544	0.420	0.353
	IPLS	0.806	-	1.003	0.555	-	0.443
	GA	0.839	-	0.920	0.322	-	0.396
Nt+Rr+Rt	-	0.799	0.745	0.866	0.571	0.565	0.491
	IPLS	0.810	-	0.958	0.490	-	0.485
	GA	0.847	-	0.924	0.492	-	0.465
Nr+Nt+Rr+Rt	-	0.766	0.691	0.815	0.548	0.522	0.357
	IPLS	0.792	-	0.897	0.421	-	0.479
	GA	0.832	-	0.891	0.362	-	0.353

Figure 4: Heatmap of RMSEP values for PLS-API and PLS-MG models.

In both cases of PLS-API and PLS-MG models, the single spectra models serve as a baseline, and they are shown in the LLDF column. These were also tested at the other two DF levels.

For MLDF, the results of PCA conducted with a single spectrum were used to create the regression models. Regarding HLDF, attempts were made to improve the basic PLS models using the XGB algorithm. These attempts degraded the values of the model parameters in most cases.

3.4.1. Data fusion PLS-API models

In this segment of the study, it is evident that DF enhances the precision of prediction. Similar to the single spectra models, it can be noted that VS may improve performance metrics in some instances, yet it does not induce large alterations. No systematic regularities are discernible in the results obtained from the application of iPLS and GA.

3.4.1.1. LLDF PLS-API

Observing the models created using LLDF, it is observed that the RMSEP values decrease from top to bottom (non-systematically), indicating that the more types of spectra are used for evaluation, the better the results obtained, especially when utilizing transmission spectra. On this level of DF, the best result (RMSEP=0.766) was obtained when combining the four types of spectra with no VS.

3.4.1.2. MLDF PLS-API

The best results were obtained at this level. The models derived from factors acquired through PCA of the Nt and Rt spectra exhibited a substantial improvement in RMSEP values. Additionally, the most optimal outcomes were consistently attained through the fusion of all four spectra (RMSEP=0.691).

3.4.1.3. HLDF PLS-API

The results clearly indicate that this data analysis technique is not suitable for predicting the API concentration in this study. In all cases, inferior results were obtained compared to those achieved with LLDF and MLDF.

After examining all PLS-API models, both LLDF and MLDF proved to be suitable. The results also indicate that MLDF yields better results than LLDF, likely because PCA filters out unnecessary information. It is also observed that combining more spectra yields better values, particularly when using transmission spectra. Transmission measurements provide information about the entire tablet, enabling more accurate measurement of API concentration. However, reflection measurements may also contain relevant information, which likely explains why the best results were obtained with the regression model constructed using PCs extracted from all spectra.

3.4.2. Data fusion PLS-MG models

During the evaluation of PLS-MG models, the Nr spectrum-based regression model yielded strong results (RMSEP=0.357) even without DF, though several DF models exhibited lower prediction errors. Enhancements are observable at every level of DF. It is also noteworthy that the application of VS frequently enhances the parameter values. However, no systematic regularities are discernible in the results.

3.4.2.1. LLDF PLS-MG

As previously noted, the individual PLS models exhibited satisfactory performance. Analysis of the LLDF results reveals that the composite spectra derived from reflection spectra offer optimal outcomes. VS frequently contributes to substantial improvements in RMSEP values.

Post-application of genetic algorithms, the Nr+Rr+Rt model emerged as the top performer among all PLS-MG models.

3.4.2.2. MLDF PLS-MG

Within the MLDF regression models, the Nr+Rr model demonstrated exceptional performance (RMSEP=0.342). This is likely attributed to their formulation utilizing two reflection spectra. Conversely, all other models exhibited inferior results compared to the LLDF models.

3.4.2.3. HLDF PLS-MG

Similar to LLDF, several models were created here that produced similarly good results. The models that integrate the Nr-based PLS model with XGB, all yielded good results. In some cases, variable selection improved the prediction performance here as well, but not by large amounts.

The comparison of PLS-MG models shows that good results were obtained in several cases. Among the basic models, the reflection models provided optimal results. The Nr and Rr pairs showed excellent results across all strategies. Combining multiple spectra (especially in the case of transmission spectra) increases the amount of irrelevant information, thereby increasing the RMSEP value. This extraneous information cannot be optimized by PCA as the results also demonstrate (except for the Nr+Rr model). However, at the LLDF level, VS has proven suitable for this purpose in several instances, and in the case of HLDF, XGBoost has also demonstrated suitability.

3.5. Comparison with similar works

The presented study highlights the importance of suitable data sources in monitoring the manufacturing of film-coated tablets. The spectroscopic tools were utilized off-line; however, these techniques can also be applied in-line. Radtke et al. presented a study where Raman spectroscopy was use inside a pan coater to estimate the mass of three different coating materials on tablet cores containing caffeine. In their approach, the total mass of applied coating was predicted, but they did not measure the individual weight gain of tablets [47]. In a similar research, Nishii et al. used NIR hyperspectral imaging on tablets moving on a conveyor and investigated how both the API concentration and the coating mass increase could be measured simultaneously. The RMSECV value of their best model was 3.48 in case of 32-48% caffeine content. Comparing this result to our work, it is clear that the shorter measurement time and reflection mode measurement leads to less accurate model, and including a transmission measurement via data fusion would likely increase predictive ability. Their models for mass gain were more accurate, with RMSECV of 0.46 mg in the 0-7 mg mass gain range. This in in line with our result, which showed that the reflection measurements are better for measuring the amount of coating on the tablets [48]. There is also a study published by Casian at al. where they use NIR transmission and Raman reflection spectroscopy for measuring the concentration of two APIs in a combination product. They obtained more accurate results for API concentration measurement, which shows that without the interference of changing coating thickness, API concentration can be measured more reliably [49].

4. Conclusion

This work investigated the application of PLS models for predicting API concentration and mass gain of film coated tablets. The analysis provided insights into the effectiveness of different spectroscopic techniques and data fusion strategies. It is important to note that these

models were evaluated solely based on RMSEP values. Further investigations are necessary to assess the models' quality and robustness comprehensively.

Transmission models exhibited higher accuracy in predicting individual API concentration. This is likely because transmission spectroscopy captures the rays passing through the entire tablet, providing extensive information about the tablet's whole volume. In contrast, reflection models were less effective as they measure reflected rays, which only capture the top layer of the tablets, including the film layer, resulting in less comprehensive information. Data fusion strategies significantly enhanced predictive accuracy for API concentration models. Both LLDF and MLDF techniques showed considerable improvements, with MLDF models excelling due to PCA filtering that eliminates irrelevant information. Combining multiple spectra, particularly integrating transmission spectra, yielded the best results by providing thorough information about the entire tablet.

For mass gain prediction, the models that processed the reflection spectra provided the best results. Reflection spectroscopy's ability to assess the top layer of the tablets, where the film coating resides, contributed to higher accuracy. Similar to the API concentration models, data fusion approaches, including LLDF, MLDF, and HLDF, consistently enhanced model performance. LLDF models were particularly effective, and MLDF models showed exceptional results when integrating the two reflection spectra. Although HLDF models generally performed well, they did not outperform LLDF models by a large margin.

Overall, the study highlights that while variable selection can reduce model complexity by lowering the number of latent variables, it does not always lead to better model performance. In contrast, data fusion strategies, particularly those integrating multiple spectra types, enhance predictive accuracy for both API concentration and mass gain during film coating processes.

In conclusion, the combination of reflection NIR and transmission Raman spectroscopy techniques could be highly effective for simultaneously predicting API concentration and mass gain during the film coating process. For API-PLS models, the LLDF strategy with GA application proves to be appropriate, whereas for MG-PLS models, the HLDF technique is optimal without the necessity for variable selection.

These research findings demonstrate the potential for applying advanced spectroscopic techniques and PLS modeling in pharmaceutical manufacturing to enhance quality control processes. By using an at-line method for rapid, non-destructive analysis, the industry can more efficiently determine whether formulations meet required standards for active ingredient concentration and film coating quality. The method's ability to independently assess active ingredient concentration and coating thickness without interference eliminates the need for separate models for different dosage forms of the same product. This simplification of the quality control process not only improves efficiency and reliability but also aligns with the pharmaceutical industry’s ongoing transition toward more advanced, data-driven manufacturing practices, optimizing production processes and ensuring regulatory compliance.

Declaration of competing interest

The authors declare that they have no conflicts of interest.

Acknowledgement

Project no. RRF-2.3.1-21-2022-00015 has been implemented with the support provided by the European Union. It was also supported by the EKÖP-24-3-BME-41, EKÖP-24-3-BME-103 and EKÖP-24-4-II-BME-44 New National Excellence Program of the Ministry for Culture and Innovation from the source of the National Research, Development and Innovation Fund. The project supported by the Doctoral Excellence Fellowship Programme (DCEP) is funded by the National Research Development and Innovation Fund of the Ministry of Culture and Innovation and the Budapest University of Technology and Economics, under a grant agreement with the National Research, Development and Innovation Office.

The research was supported by the Agency for Credits and Study Grants coordinated by the Romanian Ministry of National Education from the source of the research grant established through the Government Decision no. 118/2023.

References

1. Augsburger, L.L. and S.W. Hoag, *Pharmaceutical dosage forms-tablets*. 2016: CRC press.
2. Zaid, A.N., *A comprehensive review on pharmaceutical film coating: past, present, and future.* Drug Design, Development and Therapy, 2020: p. 4613-4623.
3. Sastry, S.V., J.R. Nyshadham, and J.A. Fix, *Recent technological advances in oral drug delivery–a review.* Pharmaceutical science & technology today, 2000. **3**(4): p. 138-145.
4. Kapoor, D., et al., *Coating technologies in pharmaceutical product development*, in *Drug delivery systems*. 2020, Elsevier. p. 665-719.
5. Felton, L.A., *Film coating of oral solid dosage forms.* Encyclopedia of pharmaceutical technology, 2007. **3**: p. 1729-747.
6. Wang, J., et al., *An evaluation of process parameters to improve coating efficiency of an active tablet film-coating process.* International journal of pharmaceutics, 2012. **427**(2): p. 163-169.
7. Porter, S.C., *Coating of pharmaceutical dosage forms*, in *Remington*. 2021, Elsevier. p. 551-564.
8. Felton, L.A. and S.C. Porter, *An update on pharmaceutical film coating for drug delivery.* Expert opinion on drug delivery, 2013. **10**(4): p. 421-435.
9. Seo, K.-S., et al., *Pharmaceutical application of tablet film coating.* Pharmaceutics, 2020. **12**(9): p. 853.
10. Barimani, S. and P. Kleinebudde, *Monitoring of tablet coating processes with colored coatings.* Talanta, 2018. **178**: p. 686-697.
11. Peng, T., et al., *Study progression in application of process analytical technologies on film coating.* asian journal of pharmaceutical sciences, 2015. **10**(3): p. 176-185.
12. Bakeev, K.A., *Process analytical technology: spectroscopic tools and implementation strategies for the chemical and pharmaceutical industries*. 2010: John Wiley & Sons.
13. Food and D. Administration, *Guidance for industry, PAT-A framework for innovative pharmaceutical development, manufacturing and quality assurance.* http://www.fda.gov/cder/guidance/published.html, 2004.
14. Togashi, D., et al., *Evaluation of diffuse reflectance near infrared fibre optical sensors in measurements for chemical identification and quantification for binary granule blends.* Journal of Near Infrared Spectroscopy, 2015. **23**(3): p. 133-144.
15. Tabasi, S.H., et al., *Quality by design, part I: application of NIR spectroscopy to monitor tablet manufacturing process.* Journal of pharmaceutical sciences, 2008. **97**(9): p. 4040-4051.
16. Römer, M., et al., *Prediction of tablet film-coating thickness using a rotating plate coating system and NIR spectroscopy.* Aaps Pharmscitech, 2008. **9**: p. 1047-1053.
17. Nagy, B., et al., *Raman spectroscopy for process analytical technologies of pharmaceutical secondary manufacturing.* Aaps Pharmscitech, 2019. **20**: p. 1-16.
18. Wabuyele, B.W., et al., *Dispersive Raman spectroscopy for quantifying amorphous drug content in intact tablets.* Journal of pharmaceutical sciences, 2017. **106**(2): p. 579-588.
19. Müller, J., et al., *Prediction of dissolution time and coating thickness of sustained release formulations using Raman spectroscopy and terahertz pulsed imaging.* European journal of pharmaceutics and biopharmaceutics, 2012. **80**(3): p. 690-697.
20. Kandpal, L.M., et al., *Quality assessment of pharmaceutical tablet samples using Fourier transform near infrared spectroscopy and multivariate analysis.* Infrared Physics & Technology, 2017. **85**: p. 300-306.

21. De Beer, T., et al., *Near infrared and Raman spectroscopy for the in-process monitoring of pharmaceutical production processes.* International journal of pharmaceutics, 2011. **417**(1-2): p. 32-47.
22. Ciza, P., et al., *Comparing the qualitative performances of handheld NIR and Raman spectrophotometers for the detection of falsified pharmaceutical products.* Talanta, 2019. **202**: p. 469-478.
23. Esposito Vinzi, V. and G. Russolillo, *Partial least squares algorithms and methods.* Wiley Interdisciplinary Reviews: Computational Statistics, 2013. **5**(1): p. 1-19.
24. Pirouz, D.M., *An overview of partial least squares.* Available at SSRN 1631359, 2006.
25. Alsberg, B.K., D.B. Kell, and R. Goodacre, *Variable selection in discriminant partial least-squares analysis.* Analytical Chemistry, 1998. **70**(19): p. 4126-4133.
26. Xiaobo, Z., et al., *Genetic algorithm interval partial least squares regression combined successive projections algorithm for variable selection in near-infrared quantitative analysis of pigment in cucumber leaves.* Applied spectroscopy, 2010. **64**(7): p. 786-794.
27. Nørgaard, L., et al., *Interval partial least-squares regression (i PLS): A comparative chemometric study with an example from near-infrared spectroscopy.* Applied spectroscopy, 2000. **54**(3): p. 413-419.
28. Ji, G., et al., *Using consensus interval partial least square in near infrared spectra analysis.* Chemometrics and Intelligent Laboratory Systems, 2015. **144**: p. 56-62.
29. Chen, T. and C. Guestrin. *Xgboost: A scalable tree boosting system*. in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016.
30. Hayashi, Y., et al., *Development of concentration prediction models for personalized tablet manufacturing using near-infrared spectroscopy.* Chemical Engineering Research and Design, 2023. **199**: p. 507-514.
31. Casian, T., et al., *Challenges and opportunities of implementing data fusion in process analytical technology—a review.* Molecules, 2022. **27**(15): p. 4846.
32. Casian, T., et al., *Development of a PAT platform for the prediction of granule tableting properties.* International Journal of Pharmaceutics, 2023. **648**: p. 123610.
33. Casian, T., et al., *Data fusion strategies for performance improvement of a Process Analytical Technology platform consisting of four instruments: An electrospinning case study.* International Journal of Pharmaceutics, 2019. **567**: p. 118473.
34. Azcarate, S.M., et al., *Data handling in data fusion: Methodologies and applications.* TrAC Trends in Analytical Chemistry, 2021. **143**: p. 116355.
35. Hayes, E., et al., *Spectroscopic technologies and data fusion: Applications for the dairy industry.* Frontiers in Nutrition, 2023. **9**: p. 1074688.
36. Zomer, S., et al., *Multivariate monitoring for the industrialisation of a continuous wet granulation tableting process.* International Journal of Pharmaceutics, 2018. **547**(1-2): p. 506-519.
37. Guideline, I., *Validation of analytical procedures Q2 (R1).* ICH: Geneva, Switzerland, 2022.
38. Ficzere, M., et al., *Real-time coating thickness measurement and defect recognition of film coated tablets with machine vision and deep learning.* International Journal of Pharmaceutics, 2022. **623**: p. 121957.
39. Bocklitz, T., et al., *How to pre-process Raman spectra for reliable and stable models?* Analytica chimica acta, 2011. **704**(1-2): p. 47-56.
40. Gautam, R., et al., *Review of multidimensional data processing approaches for Raman and infrared spectroscopy.* EPJ Techniques and Instrumentation, 2015. **2**: p. 1-38.

41. Renaud, O. and M.-P. Victoria-Feser, *A robust coefficient of determination for regression.* Journal of Statistical Planning and Inference, 2010. **140**(7): p. 1852-1862.
42. Porep, J.U., D.R. Kammerer, and R. Carle, *On-line application of near infrared (NIR) spectroscopy in food production.* Trends in Food Science & Technology, 2015. **46**(2): p. 211-230.
43. Blanco, M. and A. Peguero, *A new and simple PLS calibration method for NIR spectroscopy. API determination in intact solid formulations.* Analytical Methods, 2011. **4**(6): p. 1507-1512.
44. Chavez, P.-F., et al., *Active content determination of pharmaceutical tablets using near infrared spectroscopy as Process Analytical Technology tool.* Talanta, 2015. **144**: p. 1352-1359.
45. Radtke, J., H. Rehbaum, and P. Kleinebudde, *Raman spectroscopy as a PAT-Tool for film-coating processes: In-Line Predictions Using one PLS Model for Different Cores.* Pharmaceutics, 2020. **12**(9): p. 796.
46. Andersson, M., et al., *Monitoring of a film coating process for tablets using near infrared reflectance spectrometry.* Journal of pharmaceutical and biomedical analysis, 1999. **20**(1-2): p. 27-37.
47. Radtke, J. and P. Kleinebudde, *Real-time monitoring of multi-layered film coating processes using Raman spectroscopy.* European Journal of Pharmaceutics and Biopharmaceutics, 2020. **153**: p. 43-51.
48. Nishii, T., K. Matsuzaki, and S. Morita, *Real-time determination and visualization of two independent quantities during a manufacturing process of pharmaceutical tablets by near-infrared hyperspectral imaging combined with multivariate analysis.* International Journal of Pharmaceutics, 2020. **590**: p. 119871.
49. Casian, T., et al., *Development, validation and comparison of near infrared and Raman spectroscopic methods for fast characterization of tablets with amlodipine and valsartan.* Talanta, 2017. **167**: p. 333-343.bn

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Highlights:

- The study aims to evaluate the effectiveness of Partial Least Squares (PLS) models combined with different spectroscopic techniques (NIR and Raman spectroscopy) for predicting Active Pharmaceutical Ingredient (API) concentration and mass gain during film coating process. It explores low-, mid-, and high-level data fusion (DF) techniques to enhance prediction accuracy.
- Transmission spectroscopy models showed superior performance in predicting API content, capturing comprehensive information about the tablet's volume. Data fusion strategies, especially mid-level data fusion (MLDF) using Principal Component Analysis (PCA), significantly enhanced prediction accuracy. The best results were achieved by combining all four spectra types.
- Reflection spectroscopy models were most effective for predicting mass gain, focusing on the tablet's top layer where the film coating resides. Data fusion approaches consistently improved model performance, with MLDF models integrating two reflection spectra showing exceptional results. High-level data fusion models (HLDF), using Extreme Gradient Boosting (XGBoost), generally performed well but did not significantly outperform low-level data fusion (LLDF) models.