



Machine Learning-Enabled NIR Spectroscopy in Assessing Powder Blend Uniformity: Clear-Up Disparities and Biases Induced by Physical Artefacts

Prakash Muthudoss^{1,2} · Ishan Tewari^{3,4} · Rayce Lim Rui Chi¹ · Kwok Jia Young¹ · Eddy Yii Chung Ann¹ · Doreen Ng Sean Hui¹ · Ooi Yee Khai⁵ · Ravikiran Allada⁶ · Manohar Rao⁷ · Saurabh Shahane³ · Samir Das¹ · Irfan Babla¹ · Sandeep Mhetre¹ · Amrit Paudel^{8,9}

Received: 3 May 2022 / Accepted: 22 August 2022
© The Author(s) 2022

Abstract

NIR spectroscopy is a non-destructive characterization tool for the blend uniformity (BU) assessment. However, NIR spectra of powder blends often contain overlapping physical and chemical information of the samples. Deconvoluting the information related to chemical properties from that associated with the physical effects is one of the major objectives of this work. We achieve this aim in two ways. Firstly, we identified various sources of variability that might affect the BU results. Secondly, we leverage the machine learning-based sophisticated data analytics processes. To accomplish the aforementioned objectives, calibration samples of amlodipine as an active pharmaceutical ingredient (API) with the concentrations ranging between 67 and 133% w/w (dose ~3.6% w/w), in powder blends containing excipients, were prepared using a gravimetric approach and assessed using NIR spectroscopic analysis, followed by HPLC measurements. The bias in NIR results was investigated by employing data quality metrics (DQM) and bias-variance decomposition (BVD). To overcome the bias, the clustered regression (non-parametric and linear) was applied. We assessed the model's performance by employing the hold-out and k-fold internal cross-validation (CV). NIR-based blend homogeneity with low mean absolute error and an interval estimates of $0.674 (\text{mean}) \pm 0.218 (\text{standard deviation})$ w/w was established. Additionally, bootstrapping-based CV was leveraged as part of the NIR method lifecycle management that demonstrated the mean absolute error (MAE) of $\text{BU} \pm 3.5\%$ w/w and $\text{BU} \pm 1.5\%$ w/w for model generalizability and model transferability, respectively. A workflow integrating machine learning to NIR spectral analysis was established and implemented.

Keywords bias-variance decomposition · blend uniformity · clustered linear regression · cross-validation · data pre-processing · data quality metrics · machine learning · near infrared spectroscopy · performance metrics · statistics

Abbreviations

AR	AdaBoost regression
BR	Bagging regression
BU	Blend uniformity
BVD	Bias-variance decomposition
CR	CatBoost regression
CU	Content uniformity
CV	Cross-validation
DQM	Data quality metrics
DT	Decision tree
DoE	Design of experiment
EMSC	Extended multiplicative scatter correction
ETR	Extreme tree regression

Synopsis NIR spectra of samples often contain overlapping physical and chemical information. This study proposes a machine learning framework (clustering regression) in deconvoluting the sample chemical information from those related to physical properties. A versatile NIR method to estimate amlodipine API as part of blend uniformity was successfully developed using different calibration datasets. The results demonstrate that K-fold cross-validation produced MAE of $\text{BU} \pm 0.9\%$ w/w whereas bootstrap at 95% CI yielded MAE of $\text{BU} \pm 1.5\%$ w/w. Additionally, our research overcomes the barrier, bringing us a step closer to the routine implementation of NIR-based BU assessment.

Extended author information available on the last page of the article

FT	Fourier transform
GBM	Gradient boosting machine
HPLC-UV	High-performance liquid chromatography ultraviolet
KNN	K-nearest neighbours
LightGBM	Light gradient boosting machine
LOOCV	Leave one out cross-validation
LoR	Logistic regression
LR	Linear regression
MAE	Mean absolute error
MSC	Multiplicative scatter correction
MSE	Mean square error
NIR/NIRS	Near-infrared spectroscopy
OFAT	One factor at a time
PLS	Partial least squares
R ²	Coefficient of determination
CI	Confidence interval
RF	Random forest
RMSE	Root mean square error
RTRT	Real time release testing
SA	Sensitivity analysis
SG	Savitzky-Golay
SNV	Standard normal variate
UV-Vis	Ultraviolet or visible
VIF	Variance inflation factor
XGB	Extreme gradient boost

Introduction

Powder blending is one of the critical steps in pharmaceutical solid dosage form manufacturing that needs to be monitored to ensure content uniformity of the final drug product [1–4]. To ensure patient safety and therapeutic efficacy of the drug products, powder blend uniformity (BU) or content uniformity (CU) is highly relevant for highly potent as well as low-dose drug products [2]. BU method development currently entails dry powder blending for a predetermined time period, manually stopping the blender and removing representative unit dose samples, which are then analysed using time-consuming ultraviolet or visible (UV/Vis) spectroscopy or high-performance liquid chromatography (HPLC) approaches [3, 5]. Near-infrared spectroscopy (also known as NIR or NIRS) is non-intrusive and non-destructive, thus requiring relatively none to minimal sample preparation procedures, and the total analysis time could become much shorter than the current approaches making it a real-time analysis tool [6–9]. NIR spectroscopy is a vibrational spectroscopic method that utilizes the absorbance or transmittance by the sample within the near-infrared region (wavenumbers ranging between 4,000 and ~15,000 cm⁻¹) [6–9]. Fine chemicals, agricultural, food and dairy sector, pharmaceutical, cosmetics, pulp and paper, 3D printing and

precision medications, petrochemicals, polymer synthesis, and oil industry are just a few of the industries that use NIRS to appropriately assess both chemical (related to functional groups) and physical characteristics of solid samples [4, 10–20].

Because NIR data is rich in chemical and physical information, it can detect several metrics relevant to drug product performance and quality in a single measurement. To name some, this can be API/excipient characteristics, assay, content uniformity, dissolution, process analysis, water content, viscosity, etc. [3–5, 12] Although NIRS provides a plethora of information on a sample's physical and chemical characteristics, separating the two can often be difficult, particularly for BU/CU analysis where the data is hierarchical and regression analysis is one of the first steps. Similarly, physical features of the samples (e.g. particle size, shape, and surface areas), as well as experimental variables, can alter NIR spectra, masking chemical information relating to the analyte(s) of interest [1, 2, 21–26]. Various data pre-processing approaches are implemented to overcome the effect of artefacts on NIR data.

Savitzky-Golay (SG)-based derivatization (first or second derivative), normalization (min–max normalization), standard normal variate (SNV), multiplicative scatter correction (MSC), and extended MSC (EMSC) are commonly used parametric data pre-processing approaches [3, 27–29]. On the contrary, testing every possible combination of pre-processing methods to address the spectral artefacts would be unsurmountable. Hence, Mishra and co-workers proposed parallel pre-processing through orthogonalization (PORTO), pre-processing ensembles with response oriented sequential alternation calibration (PROSAC), sequential pre-processing through orthogonalization (SPORT), and various other approaches [12, 24–26]. With respect to chemical properties, NIR peaks are often broad; therefore, occasionally overlapping peaks can limit their precision. As a result, NIR cannot provide precise spectroscopic fingerprints of different chemical functional groups. In other words, NIR spectroscopy has two significant drawbacks: (i) overlapping peaks and the (ii) inability to detect or quantify trace or multicomponent analytes. One of the most significant and beneficial approaches available to spectroscopists, chemometricians, and data scientists in general is deconvolution of spectral peaks with respect to reduced chi-square value (which, in principle, signifies good spectral resolution). To deconvolute the overlapping spectral patterns, the other spectral resolution enhancement approaches are derivatization (1st, 2nd, or 4th), difference spectroscopy, curve fitting, two-dimensional (2D) correlation spectroscopy, and chemometrics like self-modelling curve resolution methods (SMCR) [30, 31]. As mentioned, spectral analysis in NIR is quite a daunting task. The recent add-on to the resolution enhancement approaches is machine learning (ML) and deep learning (DL) [32]. Similarly, ML approaches have been

extensively employed in formulation development to learn and predict various complex systems [33–40]. However, the integration of machine learning in pharmaceutical analytical development has so far been limited [41, 42].

For NIR-based BU or CU, current pre-processing approaches involving normalization, second derivative, MSC, EMSC, and SNV might not be pertinent if any of the following conditions exist: (1) data demonstrating heteroscedasticity, non-normality, multicollinearity which are the prerequisites for the parametric models [43–45]; (2) also, if physical properties are of simultaneous importance; (3) the types of physical fluctuations are uncontrollable or may not be representative of future samples; (4) the artefacts are non-linear/non-parametric. To the best of our knowledge, this is the first work to provide an intuitive understanding of the gaps in existing NIR-based BU analysis. We employ the following multidisciplinary strategy to accomplish this, which incorporates the section on AQbD “development of multivariate analytical procedures”.

1. Data transformation: Several data quality evaluation metrics to assess the applicability of existing pre-processing procedures is carried out, and an artefact-agnostic framework is investigated.
2. Variable selection and model development: Cause-and-effect nature of various data learning approaches is recognised and ranked.
3. Robustness of model: Fit, reliability, and validity of algorithms, as well as their generalizability, are investigated utilizing multiple model evaluation and cross-validation methodologies.
4. Recalibration and model maintenance: Lifecycle management of the developed ML models are explored.

Materials and Methods

Materials

Amlodipine besylate (Glochem Industries Pvt. Ltd., India) was used as the model active pharmaceutical ingredient (API), mannitol (Pearlitol® 100SD, Roquette Frères, France), microcrystalline celluloses (Avicel® PH112, DuPont Pharma, USA), croscarmellose sodium (Ac-Di-Sol®, DuPont Pharma, USA), and magnesium stearate (Ligamed MF 2-V®, Peter Greven, Germany) were used as excipients.

Sample Preparation for BU Analysis-Physical Mixture

Sample preparation for the off-line calibration involved accurately weighing API and excipients equivalent to unit dose (gravimetric approach) which was filled in vial as

physical mixture (~1.5 g). To keep measuring system variability under control, only single batches of API and excipients were used. To establish repeatability, five measurements per unit dose vial were collected. A $1 \times 1 \times 5$ design was adopted to depict batch, reproducibility, and repeatability, respectively. A total of 13 calibration samples were prepared and analysed ranging between 67 and 133% w/w of API dose (~3.6% w/w). As part of the database building, individual components were contained in vials and their NIR spectra were collected.

Sample Preparation for BU Analysis-Processed Mixture

A GEA bin blender with a stainless-steel shell was used for preparation of powder mixture/blend. Blender was operated for 15 min at a rotation speed of 25 rpm to blend API and excipients. A total of 3 blend mixtures each equivalent to ~3,000 tablets were prepared comprising API content of 70% w/w, 100% w/w, and 130% w/w. After the mixing was completed, the entire contents of the blender were transferred into an HDPE bottle. Samples were filled in triplicate in NIR vials that were approximately the same size as the unit dose, such as 1.5 g per vial. Three NIR scans were performed on each vial. A $1 \times 3 \times 3$ design was used to emphasize batch, reproducibility, and repeatability, respectively.

NIR Spectroscopy

Frontier transform-NIR (FT-NIR) spectrometer (PerkinElmer, Waltham, MA, USA) equipped with a reflectance accessory and with a liquid nitrogen cooled mercury cadmium telluride (MCT) detector was utilized for raw material analysis. Each spectrum was collected over the range of 15,000 to 4,000 cm^{-1} wavenumbers at an average of 64 scans with a spectral resolution of 8 cm^{-1} . The spectra of the investigated samples were collected using the spectrum 10 software (version 10.7). Off-line calibration samples were prepared and data acquisition was carried out by placing the sample holders to allow the incident NIR source to probe the contents within the sample. Five NIR measurements were acquired on the individual samples to ensure repeatability. Between each measurement, random shaking and tapping was employed to determine any significant variations. This procedure was carried out in a similar fashion for all the samples.

HPLC Analysis

Amlodipine blend uniformity in powder physical mixture blends was performed using a validated high-performance liquid chromatography-UV (HPLC–UV) method as the reference. The chromatographic parameter employed in this

investigation are as follows: column is Waters symmetry C18 150×3.9 mm, 5 µm; mobile phase triethylamine buffer (1% v/v, pH 2.8): mixture of methanol & ACN (70:30 v/v) in gradient elution program of (0–3 min, 50:50 v/v; 15–20 min 30:70 v/v; 20.1–25 min 50:50 v/v); flow rate of 1 mL/min; column temperature 30 °C; and sample temperature of 10 °C. The detection was performed at 237 nm for amlodipine. Under the given chromatographic conditions, the total run time is set to 25 min and retention time was ~6–7 min for amlodipine.

Statistical Machine Learning

Understanding NIR Blend Uniformity Data

The NIR data for the blend uniformity was organized into a table with rows and columns. Within the table, the rows represent observations, and the columns represent attributes/features for those observations. The NIR BU dataset employed in this study consist of 65 observations (rows). The data consist of ~ 11,001 columns (termed features/attributes) of which 11,000 represents independent features termed wavenumber while 1 column represents dependent or target variable which is blend uniformity. Each row represents spectral log 1/R values (arbitrary units/a.u) for each repeat run of sample. The target variable represents 13 different concentrations of content uniformity in % w/w with 5 repeat measurements per concentration (13 concentration×5 NIR repeat measurements yielding a total of 65 observations). Typical representation to understand the nir blend uniformity data is depicted in Table S1.

Data Quality Metrics (DQM)

To choose among the various modelling approaches, the following quality metrics were performed: (i) whether the data follows a normal distribution (Shapiro–Wilk test); (ii) whether the variances across sample points are homogeneous (Levene’s test); (iii) presence of multi-collinearity (variance inflation factor/VIF test); (iv) test for linearity (Pearson R and scatter plot matrix); (v) test for outliers (Mahalanobis, T2 Hotelling, Jack-knife test); and (vi) Hopkin’s test for dimensionality reduction of instances or rows.

Model Selection

In supervised machine learning, we assume a definitive link between a feature(s) and the target. To estimate this unknown relationship, a model function called “f” is used, which can accurately depict the relationship between features and target (Eq. 1). In general, the following supervised approaches were executed: univariate linear regression (LR), multivariate linear regression (MLR), and machine learning.

Univariate regression required the selection of a single peak for either API or mannitol, whereas MLR required the selection of multiple peaks of API, mannitol, or both. On the other hand, various machine learning algorithms from the class of (1) conventional linear machine learning (partial least squares (PLS), linear regression (LR), multiple linear regression (MLR), logistic regression (LoR), ridge regression (RR), lasso regression (LaR)); (2) conventional non-linear machine learning (support vector machine (SVM), k-nearest neighbours (kNN), decision tree (DT)); (3) bagging-based ensemble learning (bagging regression (BR), random forest (RF), extreme tree regression (ETR)); (4) boosting-based ensemble learning (AdaBoost regression (AR), gradient boosting machine (GBM), extreme gradient boost (XGB), CatBoost regression (CR), light gradient boosting machine (LightGBM)); and (5). hybrid approach (clustered regression). The mathematical representation of the estimation of function (f), which maps the dependent variable (y) against the independent variable (x) with a certain error (residual error, E), is as follows:

$$y = f(x) + \text{Model Error} \quad (1)$$

Clustered Regression

It is a hybrid strategy that involves binary encoding of each blend of uniformity samples. Although the total number of samples is 65, the BU data can be considered as cohorts or grouped based on either amlodipine or mannitol content. In such scenarios, one-hot encoding of features can be accomplished [46]. For example, samples consisting of BU 67% w/w will be encoded as 1, and the rest of the samples like 72% w/w or 133% w/w or others will be encoded as 0. Using the same approach, the remaining BU composition samples were encoded. These samples were then subjected to a supervised ML regression model (as stated in the “Model Selection” section).

Feature or Variable Selection

Different feature/variable selection involving (i) full dataset: 15,000 to 4,000 cm^{-1} ; (ii) truncated data 1: 10,000 to 4000 cm^{-1} ; (iii) truncated data 2: 6,099 to 4,000 cm^{-1} ; and (iv) truncated data 3: only encoded features were prepared to finalize the model based on performance metrics and perform sensitivity analysis and bias-variance decomposition.

Performance Metrics

To measure the model’s prediction abilities, the coefficient of determination (R^2), mean absolute error (MAE), mean square error (MSE), and root mean square error (RMSE),

as defined in Eqs. 2–5, respectively, were implemented [47, 48]. MAE is a metric that quantifies the accuracy of a prediction. It is defined as the average difference between the actual and predicted value, MSE calculates the average of the square of the difference between the actual and predicted values. Decomposition of MSE provides irreducible and reducible error (for more details refer to section on bias-variance decomposition). RMSE is the square root of MSE. The extent of variance in the dependent (y -variable) component that can be described by independent (x -variable) characteristics is measured by R -squared, a goodness-of-fit metric. It quantifies the strength of the relationship between the model and the dependent variable on a scale of 0 to 1. The R^2 should be higher (> 0.95) while MAE, MSE, and RMSE results should be as low as possible.

$$R^2 = 1 - \left(\frac{\sum_{i=1}^n (y_i^{\text{actual}} - y_i^{\text{pred}})^2}{\sum_{i=1}^n (y_i^{\text{actual}} - y_i^{\text{actual,mean}})^2} \right) \quad (2)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i^{\text{actual}} - y_i^{\text{pred}}| \quad (3)$$

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i^{\text{actual}} - y_i^{\text{pred}})^2 \quad (4)$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i^{\text{actual}} - y_i^{\text{pred}})^2} \quad (5)$$

where y_i^{actual} and y_i^{pred} were the reference and ML predicted content uniformity, respectively. And $y^{\text{actual,mean}}$ was the mean of experimental value and n was the number of datapoints.

Training-Test Split

In spectroscopy and analytical method development, we frequently employ the strategy of first generating a calibration, and if it fulfils the requisite standards, then the validation step is executed. In this study, the calibration data is divided in an 80 (train):20 (test) proportions using a random selection approach, hereinafter referred as “hold-out data”. This approach is followed to ensure that the train and test datasets represent the original calibration data. Separating test data from training data enables unbiased evaluation of the machine learning algorithms’ performance. The 80:20 split was employed throughout this investigation, unless otherwise mentioned. To gauge the robustness of the generated model, external validation samples to evaluate the real-world prediction abilities were also included. The abovementioned performance metrics were computed on following

approaches: (i) hold-out data (80:20 split) and (ii) cross-validation (internal and external). Hold-out and cross-validation generates point estimates; additionally, bootstrapping was carried out to generate interval estimates.

Internal Validation-Hold-Out The NIR spectra datasets on powder blend uniformity data were used to train the machine learning models. In the field of machine learning, generally, the approach is to split the available dataset into two parts: (i) training (80%) and (ii) test (20%). The training subsets were used to build the models and perform feature engineering like determine NIR spectral regions of interest, using only encoded features, while the test subsets were used to assess model generalizability. The dataset has 13 subgroupings with 5 repetition data, regardless of the fact that the target data is continuous. As a result, on training-test splitting to obtain balanced data and achieve better performance, the stratified sampling approach was employed. In case of clustered regression, the data was first encoded (as mentioned in the “[Clustered Regression](#)” section), and then the data were split. This approach was deemed acceptable because as part of life cycle management of the model, we can keep adding new data and improvising the model. Point estimates generated from hold-out data are not always reliable; hence, sensitivity analysis needs to be carried out.

Internal Validation-Sensitivity Analysis (Cross-Validation and Bootstrapping) The examination of divergence in a model’s output linked to sources of variance/bias in the model input is known as sensitivity analysis (SA). It is employed for improving the accuracy and quantifying the uncertainty of a performance measure. In general, there are four approaches [49–52]: (i) comparing predictions and coefficients to physical theory, (ii) outcome comparison between theory and simulation, (iii) validating the prediction model on new data (external validation), and (iv) resampling methodologies such as k -fold cross-validation (internal validation) or bootstrap. Resampling is an economic approach for sensitivity analysis of the developed model, and it performs such analysis by leveraging available data. In k -fold cross-validation, a dataset is partitioned into k groups, with each group having the option of being utilized as a training set while the remaining groups serve as the test set [37]. When $k = 1$, it is leave one out cross-validation (LOOCV). Bootstrapping is a statistical process that creates several simulated samples by resampling a single dataset with replacement [43, 53]. Those instances that were not resampled are used in the test set. Also known as uncertainty estimates, bootstrapping creates resampled datasets of size similar to original dataset enabling generation of standard errors, confidence intervals, or interval estimates. In summary, cross-validation as well as bootstrapping aids in understanding model generalizability

(predicting future unseen data) and transferability (extrapolation to similar problems).

External Validation Samples On the other hand, new samples as in-process blends could be employed as the external validation sample. That said, to combine the power of external validation with the benefits of a prediction model based on available internal validation data, this research utilized an internal–external validation design/architecture. That is, physical mixture samples at 72% w/w, 100% w/w, and 128% w/w were substituted with processed samples at the lowest (70% w/w), mid (100% w/w), and highest (130% w/w) concentrations to generate an internal–external cross-validation dataset (resembling 13 clusters).

Bias-Variance Decomposition (BVD)

The bias-variance decomposition (BVD) of error is another valuable approach employed to unravel both data as well as learning algorithm's performance characteristics [54]. Based on these results, the need for hyperparameter tuning was approached. The BVD demonstrates that mean squared error of a model generated by a certain algorithm is indeed made up of two components: (1) irreducible error (as noise) and (2) reducible error (as bias, and variance), as shown in Eqs. 6 and 7. Lowering bias and/or variance would allow in developing more accurate models. Bias measures how closely the learning algorithm's average prediction matches the optimal prediction, while variance describes how much the prediction varies over different training sets of a particular size. Irreducible error includes instrument, sample, or sampling-related causes. A model with minimal bias and variance is better often difficult to achieve. Hence, the bias-variance trade-off principle is used.

$$\text{Mean Squared Error (Model)} = \text{Reducible Error} + \text{Irreducible Error} \quad (6)$$

$$\text{Mean Squared Error (Model)} = \text{Bias}^2(\text{model}) + \text{Variance}(\text{model}) + \text{Irreducible Error} \quad (7)$$

Data Analysis and Statistics

NIR data from the instrument were converted to.csv file on Microsoft® Excel® for Microsoft 365. DQM was carried out using JMP standard package (JMP®, Version 16, SAS Institute Inc. Cary, NC, 1989–2022) and R 3.6.2 (R Development Core Team, The R Foundation for Statistical Computing, 2020) using the RStudio Desktop 1.2.5019 (RStudio, PBC., Boston, MA, USA) which is an integrated development environment (IDE) for R <http://www.rstudio.com/>. Data analysis based on peak, ML were performed using Python (version 3.9.0). Machine learning involving the LightGBM, XGB, and CatBoost models were built by

using the LightGBM package (version 3.2.1) [55], Xgboost package (version 1.5.2) [56], and CatBoost package (version 1.0.4) [57, 58] in Python. Univariate linear regression, multivariate linear regression, and other models were built by using the sci-kit sklearn package (version '0.24.0') in Python [59]. The bias variance decomposition was performed using the mlxtend package (version '0.18.0') [54] in Python (Raschka, 2018). Matplotlib package (version '3.4.1') [60] and JMP were employed in generating plots.

Results

Risk-Based Approach in Assessing NIR-Based Blend Uniformity

Potential sources of variability that could influence the blend uniformity results measured using NIRS are indicated in Fig. 1. As much as six primary categories (albeit not exhaustive) for the Ishikawa diagram with all components stated under each category are captured. Several major areas were kept constant, including sampling related, instrument related, measurement related, and environment related. Errors contributing to bias can exist, even when the equipment and standards are routinely calibrated and are under the control. The source of bias with respect to NIR measurements arising due to instrument-based was controlled using routine calibrations and employing FT-NIR; the errors due to instrument were mostly reduced. The influence of temperature ($25^\circ \pm 5^\circ\text{C}$) and humidity variations ($60\% \pm 5\% \text{ RH}$) on the chemical characteristics and powder properties (such as agglomeration, caking) were mitigated by incorporating the API that is stable to such extraneous effects. Because the API is in a low dose ($\sim < 5\% \text{ w/w}$), the sample preparation technique was considered critical. To avoid issues about segregation or blend uniformity caused by powder mixing procedures, the samples were externally prepared (“offline”)

as a “single dose” in the NIR glass vial. Similarly, glass vials were also NIR inspected to rule out any inconsistency. A total 13 calibration samples were prepared and analysed ranging between 67 and 133% w/w. Sampling for HPLC was carried out including the entire powder blend contained in the vial once the NIR measurements were completed. That is, the “offline unit dose vial” utilized for NIR data acquisition is considered “a population”, while for HPLC, it is subsampling of NIR vial or statistically “a sample”. The entire idea is if HPLC results would be linear and yields correlation coefficient $R^2 = 0.99$ or so, the NIR results should be also at par. To avoid people-related errors, one analyst trained in both HPLC and NIR technique was employed.

Using such an approach, the authors believe that NIR results are primarily dependent on scattering or physical variations, as well as accompanying data analysis methodologies like pre-processing and processing which is *prima focus* of this paper.

Interpretation of NIR Spectra of API, Excipients, and Blend-NIR Feasibility Studies

The NIR spectra (raw data or zero derivative) of pure components (amlodipine, mannitol, microcrystalline cellulose, croscarmellose, magnesium stearate) are displayed in Fig. 2. Figure 3 depicts the second derivative NIR spectra for pure components from 6,100 to 4,000 cm^{-1} and expanded region from 5,000 to 4,500 cm^{-1} , respectively. The other expanded regions for clarity are shown in Figure S1 (from 6,100 to 5,700 cm^{-1}) and Figure S2 (from 4,500 to 4,000 cm^{-1}). The characteristic peak centres of API assessed were 4,336 cm^{-1} (CH stretch + CH deformation combination), 4,450 cm^{-1} (OH stretch + C-O stretch combination), 4,578 cm^{-1} (CH stretch + Carbonyl stretch), 4,661 cm^{-1} (aromatic stretch + aromatic bend combination), and 6,004 cm^{-1} (aromatic first overtone), while the unique peaks for mannitol were monitored at 4,310 cm^{-1} (polysaccharide CH stretch + CH deformation combination), 4,619 cm^{-1} (asymmetric C-H stretch/C-H deformation), 4,805 cm^{-1} (aliphatic OH combination), 5,748 cm^{-1} (pyranose or furanose CH stretch first overtone), and 5,931 cm^{-1} (CH stretch first overtone). The precise NIR band assignments are challenging since a single band may be attributed to multiple different possible combinations of fundamental and overtone vibrations, all of which are severely overlapping. Various resolution enhancement approaches are available. In this study, the calibration curve for blend homogeneity, however, is based on changes in the concentrations of

amlodipine and mannitol which ranged between 67 and 133% w/w. The concentration of other ingredients, however, remains unchanged. Additionally, for qualitative reasons, the intensity of regions between $\sim 5,000 \text{ cm}^{-1}$ and $\sim 4,500 \text{ cm}^{-1}$ were tracked as a function of concentration variations; see Fig. 4. The spectral patterns for 133% w/w and 67% w/w can be interpreted as being qualitatively similar to amlodipine and mannitol, respectively, whereas, 100% is between 133 and 100% w/w. The prior information allowed it to be determined that the selected peaks were unique. Note that the second derivative spectra utilized for diagnostic and/or visualisation purposes only. The ability of machine learning and deep learning to extract and decode the NIR signal from noise has been demonstrated to be effective, particularly in terms of quantitative analysis [30]. Hence approaches like machine learning and/or deep learning are employed [32, 61–63] wherein the raw NIR spectral data was utilized directly.

Samples for Calibration and Validation

To explore the precision and accuracy of the NIRS measurements as well as to optimize the machine learning parameters, 65 blend uniformity samples (a set of 13 class with 5 repeatable measurements) ranging from about 67 to 133% w/w were generated. NIR spectra of these samples along with the regions of interest (ROI) are shown in Fig. 5. Initial analysis suggests that there is no apparent linear demarcation between the samples, as hypothesized.

Model and Feature Selection: Pre-Screening

A framework for the selection of a suitable baseline model, including both linear and non-linear models, is

Fig. 1 Cause-and-effect chart depicting the factors influencing the NIR method performance in assessing blend uniformity. (Instrument-related, measurement-related, environment-related, sampling-related factors were kept constant, while the factors of scattering and data analysis as indicated in “red fonts” were intensively studied) (Note: OFAT is one factor at a time)

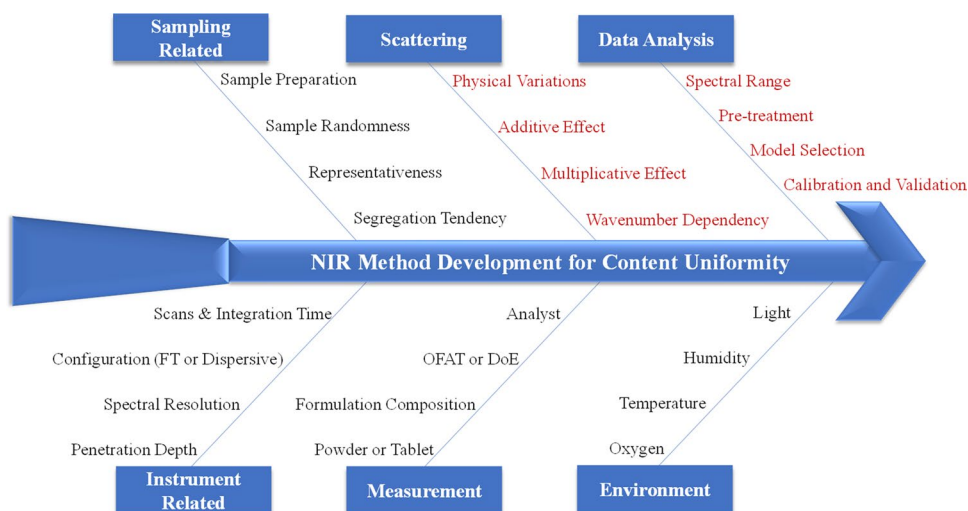


Fig. 2 Comparison of the zero derivative NIR spectra of API and excipients

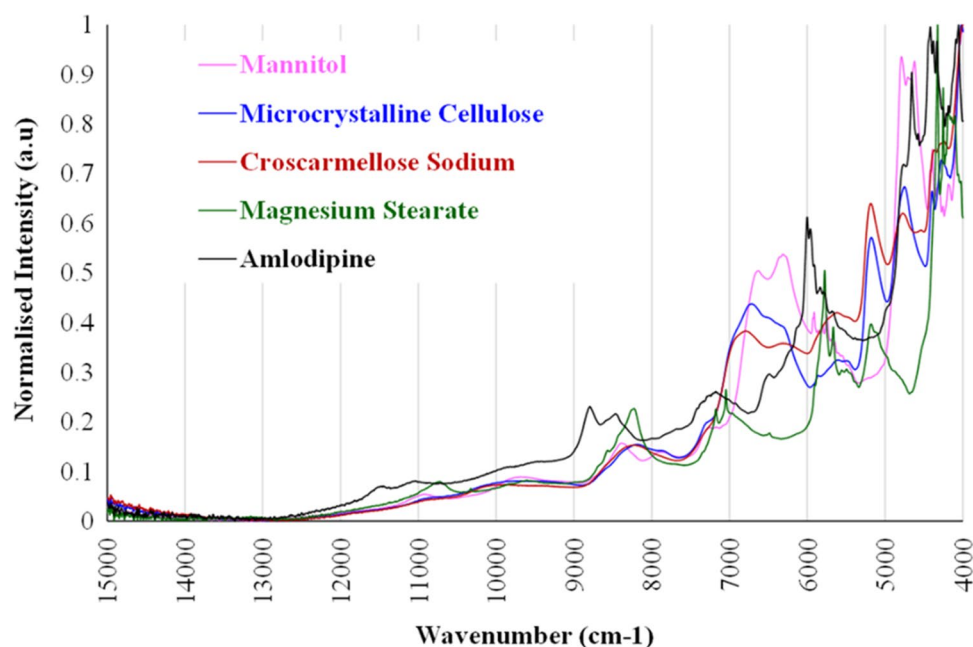
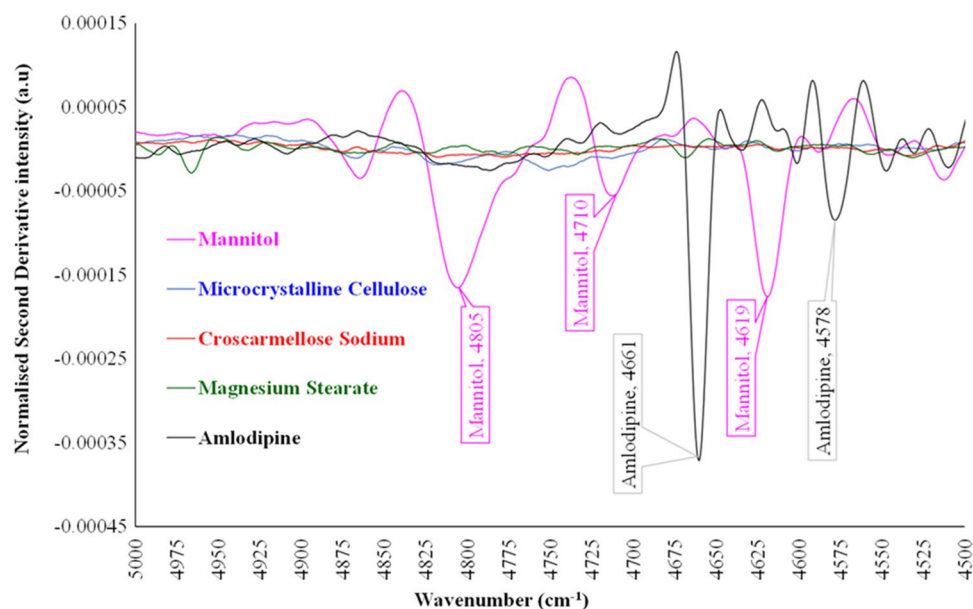


Fig. 3 Expanded second derivative spectra (region between 5000 and 4500 cm^{-1}) of the pure components (amlodipine/API and excipients)

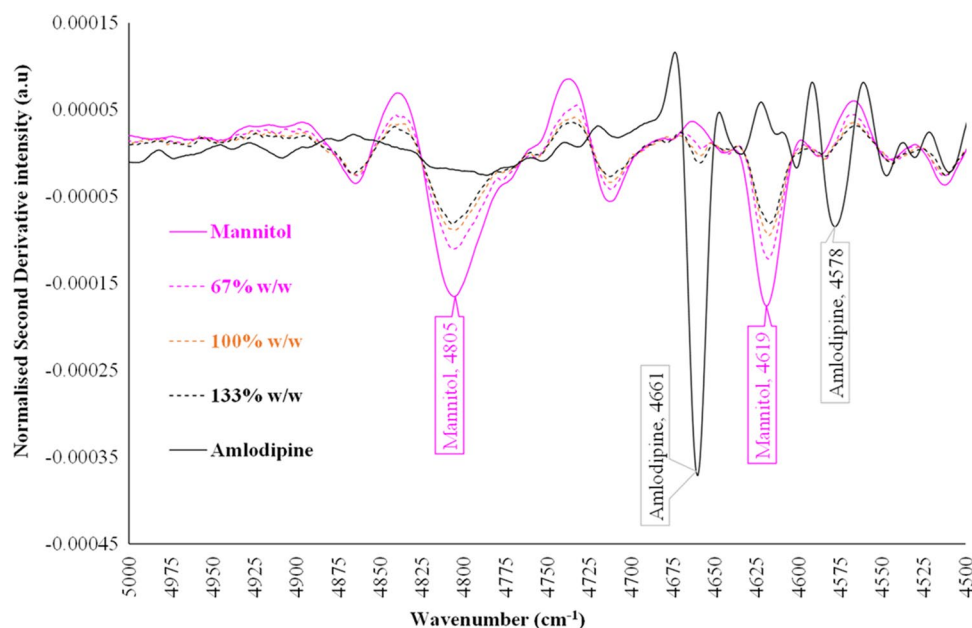


used as a preliminary method to analyse blend content uniformity of amlodipine. The three separate spectral regions of interest that were considered as features will also be presented in this section. The performance assessment metrics R^2 , MAE, MSE, and RMSE were employed to rank order the best-performing models. In summary, Table 1 shows the outcomes of several models (top 10) as well as various feature selection approaches and performance indicators. Linear models perform better, and the ROI 3 is the most efficient, as evidenced by its spectral appearance. For clarity, please refer the second derivative spectral region Fig. 4.

Comparison of Performance of NIR with Primary Tools

Figure 6 displays the content uniformity results of amlodipine blends as a measure of gravimetry, HPLC, and NIR. The peak purity determination and the profiles for amlodipine employing HPLC are shown in Figures S3 and S4. No interferences and co-elution with amlodipine peak were apparent. The linear regression-predicted results were used to compare NIR and HPLC performance with gravimetry as reference measurements. The results are surprising, and further analysis to decode this non-performance is carried out.

Fig. 4 Expanded second derivative spectra (region between 5,000 and 4,500 cm^{-1}) of the pure components (amlodipine and mannitol) and blend samples (physical mixtures)



Data Quality Metrics

Understanding the reliability of data through the application of diverse data quality metrics is a prerequisite; otherwise, poor quality might have adverse effects on model prediction [43, 44]. In this regard, the BU data was subjected to rigorous tests to examine for normality, linearity, multicollinearity, homogeneity in variance/homoscedasticity, presence of outliers, sample adequacy, and clustering tendency. The results indicated that the data is non-normal, non-linear, inhomogeneity in variance/heteroscedasticity. That is NIR BU data were found to

have violations of parametric assumptions [43, 44]. On the other hand, no outliers were identified. Moreover, data demonstrated clusterable tendency. The results of DQM are shown in Table II.

Clustering Regression

Clustering is a method for splitting a dataset into a collection of groups or clusters which is based on unsupervised machine learning. The Hopkins test demonstrated that the data in this study can be clustered; nevertheless, the number and quality of clusters were determined using cluster

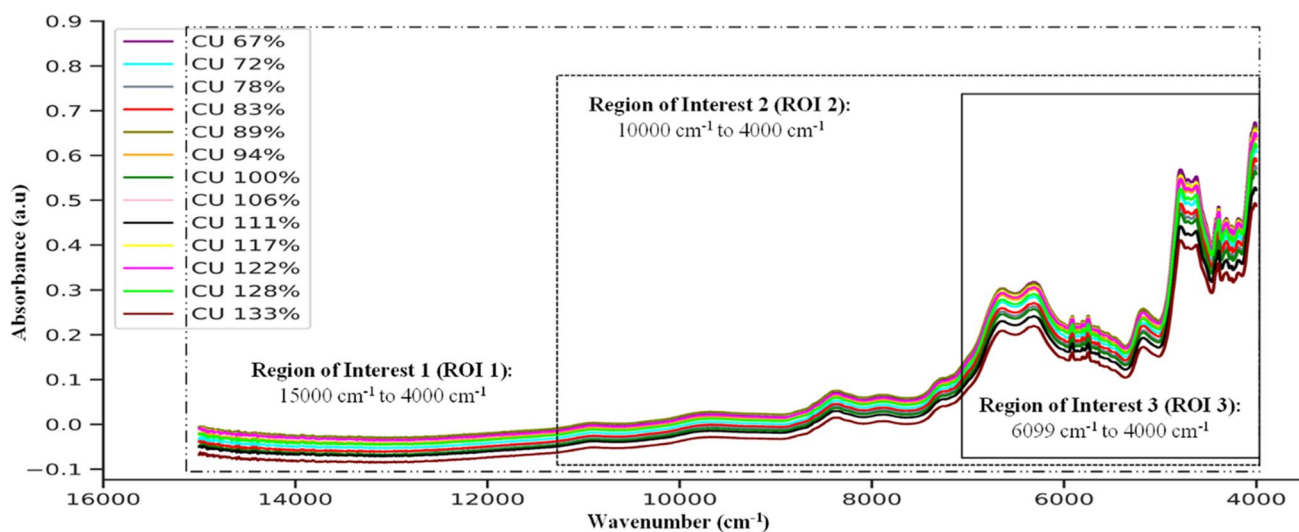


Fig. 5 Overlay NIR spectra of averaged amlodipine BU calibration samples with different regions of interest

validation indices (CVIs), which are divided into four categories: (I) relative clustering validation, (II) internal cluster validation, (III) external cluster validation, and (IV) cluster stability validation. The intrinsic approaches are preferable over the other CVIs since they do not require prior knowledge of subgroups. In general, intrinsic validation methods examine how effectively clusters are separated and compacted. This study employed ICVs (distortion score, silhouette index, Calinski-Harabasz/CH index) as well as domain knowledge. The distortion score metric calculates the sum of squared distances between each point and its corresponding centre. Silhouette score measures the similarity of datapoints within cluster (cohesion) as well as between clusters (separation) thereby determining the mean silhouette coefficient across all samples. On the contrary, the Calinski-Harabasz index is computed as the ratio of inter-cluster and intra-cluster dispersion for all clusters (where dispersion is the sum of squared distances). Elbow method based on distortion metric and elbow method based on silhouette score are shown in Fig. 7A and B, respectively. The scores for distortion metric as well as silhouette metric should be lower. The computed Calinski-Harabasz index or score is shown in Fig. 7C, and higher the CH scores, the more defined the clusters. As shown in Fig. 7D, 13 BU calibration samples

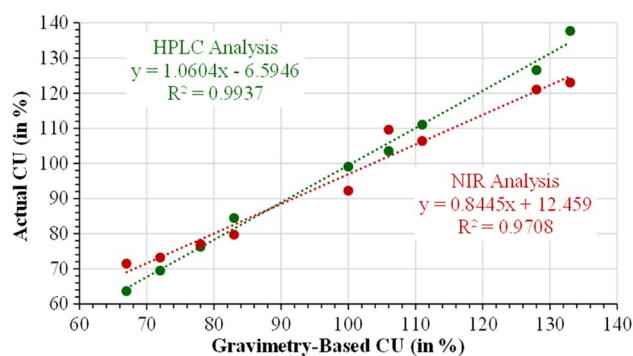


Fig. 6 Comparison of HPLC and NIR (conventional machine learning) performance against gravimetry method

were produced, ranging from 67 to 133% w/w, and ideally, there should be 13 clusters based on domain knowledge.

Following the determination of the number of clusters, clustered regression a hybrid method involving the following steps was carried out: (1) cluster analysis of data with independent variable (three different spectral regions of interest) using the K-means algorithm; (2) the samples are then categorized according to how similar they are in clusters (number of clusters determined using CVIs and

Table 1 Performance Metrics (R^2 , MAE, MSE) Results for Various Models (Top 10) as a Function of Feature Selection (Region of Interest based on Wavenumber)

Feature selection 1: 15,000 to 4,000 cm^{-1}											
Performance metrics		Linear	PLS	Ridge	SVM	Lasso	RF	kNN	LightGBM	AdaBoost	XGB
R^2	Train	1.00	0.99	0.21	0.00	0.00	0.87	0.39	0.27	0.85	1.00
	Test	0.77	0.83	0.00	-0.03	-0.05	-0.81	-0.70	-0.36	-0.86	-0.78
MAE	Train	0.00	1.86	16.66	18.49	18.55	6.24	13.76	15.26	7.14	0.00
	Test	6.69	6.73	14.23	15.92	16.10	17.81	17.83	17.86	18.21	18.29
MSE	Train	0.00	5.93	362.86	457.57	459.80	57.85	282.70	334.87	67.66	0.00
	Test	71.15	52.02	304.73	315.40	319.75	554.48	519.05	416.11	567.12	544.30
Feature selection 2: 10,000 to 4,000 cm^{-1}											
Performance metrics		PLS	Linear	AdaBoost	Bagging	RF	LightGBM	Ridge	ETR	CatBoost	GB
R^2	Train	0.93	1.00	0.75	0.85	0.89	0.60	0.17	1.00	1.00	0.99
	Test	0.95	0.94	0.15	0.15	0.14	0.09	0.09	0.02	0.02	-0.01
MAE	Train	4.37	0.00	9.12	6.18	5.72	9.72	16.60	0.00	0.85	1.10
	Test	4.44	4.67	20.06	18.74	19.71	21.38	21.49	19.79	19.46	18.51
MSE	Train	29.10	0.00	108.22	63.39	49.94	175.95	360.76	0.00	1.28	2.37
	Test	27.31	35.34	512.49	514.06	516.99	546.48	551.75	590.23	592.03	611.15
Feature selection 3: 6,099 to 4,000 cm^{-1}											
Performance metrics		PLS	Linear	XGB	Bagging	CatBoost	RF	Ridge	GB	LightGBM	AdaBoost
R^2	Train	0.94	1.00	1.00	0.84	1.00	0.87	0.14	0.99	0.46	0.73
	Test	0.94	0.94	0.17	0.13	0.09	0.08	0.04	0.02	0.01	0.01
MAE	Train	4.27	0.00	0.00	6.79	1.08	6.17	16.82	1.64	12.30	9.52
	Test	5.00	4.96	18.45	19.47	17.86	20.70	21.99	19.17	22.14	20.80
MSE	Train	25.30	0.00	0.00	70.15	1.96	57.31	376.75	4.49	237.16	117.02
	Test	34.49	36.52	501.89	527.84	552.10	556.02	581.62	591.13	594.80	597.01

domain knowledge; (3) the new sorted data is reassembled, and the regression-based machine learning is performed; and (4) the optimal model (linear/non-linear) is chosen based on the lowest MSE/MAE/RMSE in relation to the test data, whereas the R^2 metric was employed as a support tool to understand the linearity.

Comparison of Performance Metrics to Finalize the Feature Selection and Linear Model

Comparison for various univariate, multivariate, conventional machine learning, and hybrid machine learning approaches and their respective performance is sorted in descending order using the “mean squared error of test” (Fig. 8A) and “mean absolute error of test” (Fig. 8B), while the “coefficient of determination (R^2)” is depicted in ascending order in Fig. 8C. The hybrid machine learning models were found to outperform the univariate, multivariate, and conventional machine learning models as indicated through the very low MAE, RMSE scores and very high or equal to unity R^2 values.

Bias-Variance Decomposition

In order to understand the overfitting/variance and underfitting/bias, these models were subjected to bias-variance decomposition. Traditional univariate and multivariate regression, as well as regression based on machine learning, did not work effectively. To this end, the component of the mean-squared error was separated into bias and variance

error. Bias represents the error from the erroneous assumptions about the training data (quality). On the other hand, variance arises due to error from sensitivity of small changes in the training data (quantity). Because MSE and bias both have similar deviations, under-fitting is assumed as the main reason (refer to Fig. 9). The error may be reduced because of diverse methodologies, but the error is considerably larger than that of primary analytical methods like HPLC and gravimetry. This is surprising because samples for HPLC procedures were withdrawn from NIR vial. That is, NIR outcomes should be on par with or better. When nonparametric pre-processing (clustering) was combined with linear regression (termed clustered linear regression), the findings showed a ~350% reduction in bias when compared to conventional machine learning approaches.

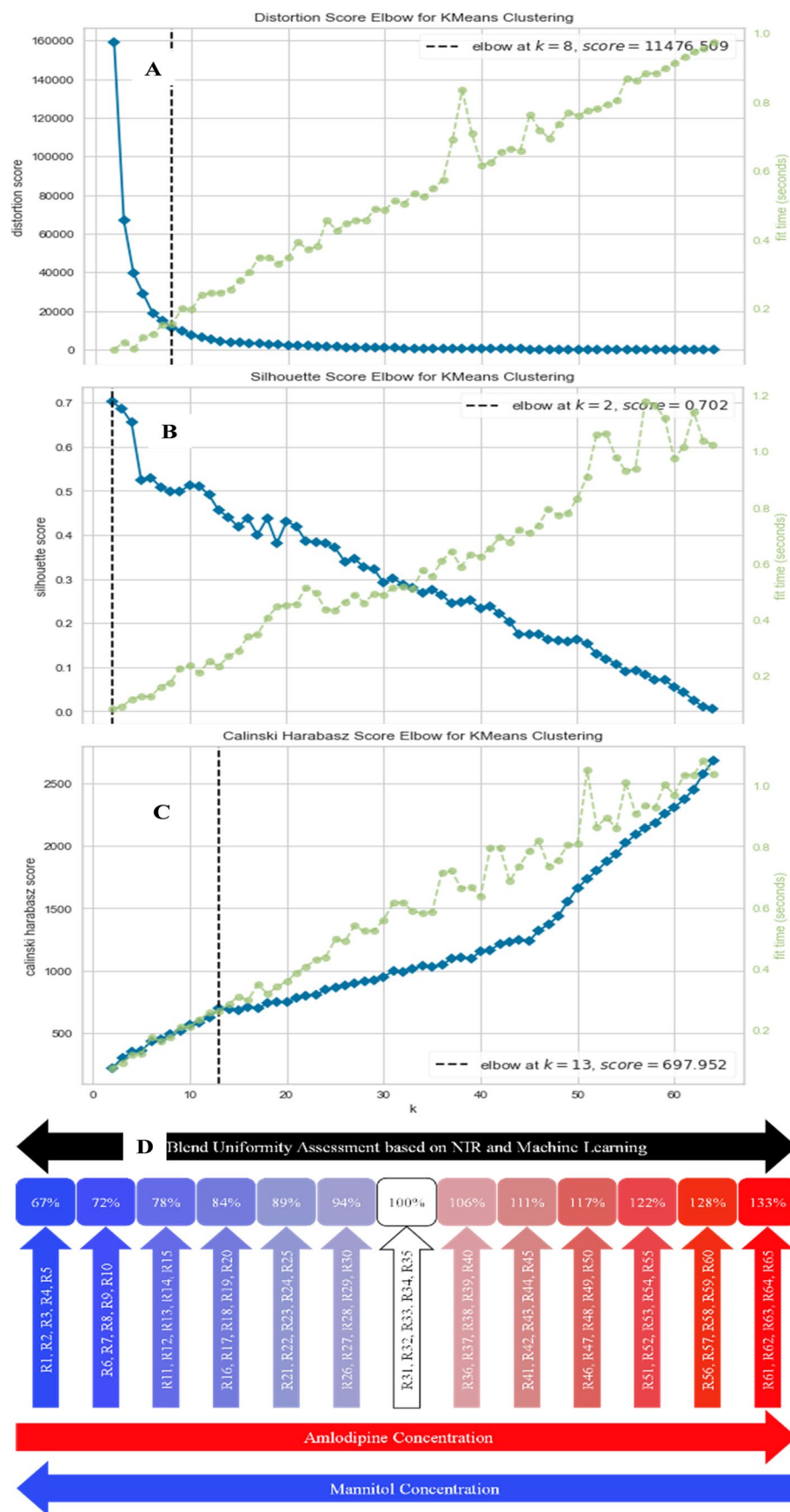
Sensitivity Analysis-Cross-Validation

The findings (point-estimates) of the hold-out sample strategy are well-known to be inefficient. Hence, cross-validation methods (like internal, external, or combination) to extract interval-estimates are required. Internal CV techniques such as fivefold cross-validation, LOOCV, and bootstrapping are used, with the results shown in Table III. The mean absolute error based on tenfold cross-validation was $BU \pm 0.9\%$ w/w while bootstrapping interval estimate yielded about $BU \pm 3.5\%$ w/w at 95% confidence interval. To evaluate these disparities, leave one out cross-validation was performed. There were few

Table II Data quality metrics

Test	What for?	Decision Rule	Results
Shapiro–Wilk test	Test for normality	<ul style="list-style-type: none"> Null hypothesis: $p > 0.05$, data follow normal distribution Alternate hypothesis: p value < 0.05 implies data do not follow normality 	<ul style="list-style-type: none"> Non-normal (Figure S5)
Levene’s test	Test for homogeneity of variance (homoscedasticity)	<ul style="list-style-type: none"> Null hypothesis: $p > 0.05$, data shows homoscedasticity Alternate hypothesis: p value < 0.05 implies data is not show homoscedastic 	<ul style="list-style-type: none"> Heteroscedasticity (Figures S6 and S7)
Scatterplot matrix (SPLOM)	Test for linearity	<ul style="list-style-type: none"> Pairwise combinations of continuous variables 	<ul style="list-style-type: none"> Independent features are linear non-linear with target cell (Figure S8)
Variance inflation factor (VIF)	Test for multicollinearity	<ul style="list-style-type: none"> Values < 5 indicate less or no, Values 5–10 imply moderate Values > 10 indicate severe 	<ul style="list-style-type: none"> Multicollinear (Figure S9)
Mahalanobis T2 hotelling Jackknife test	Test for outliers	<ul style="list-style-type: none"> A datapoint that has extreme values (either large or small) than nearest value 	<ul style="list-style-type: none"> No outliers observed (Figure S10)
Hopkin’s test	Test for clustering as dimensionality reduction in row-wise or instance-wise or x -direction	<ul style="list-style-type: none"> To identify the presence of clusters in data 	<ul style="list-style-type: none"> H-statistic value 0.05 (should be < 0.25 a.u) hence the data clusterable

Fig. 7 Expected number of clusters based on “distortion metric (A)”, “silhouette metric (B)”, and “Calinski-Harabasz metric (C)”, and domain understanding representation (D)



datapoints that yielded MAE from 1.5 to 5.1% w/w which further evidences the disparities. That is, these outlier datapoints may cause bootstrap interval estimations to be skewed. In this study, an internal–external validation design involving new samples (like processed amlodipine blend) were added to the existing internal validation data. Importantly, physical mixture samples at 72% w/w, 100% w/w, and 128% w/w were substituted with processed samples at the lowest (70% w/w), mid (100% w/w), and highest (130% w/w) concentrations to generate an internal–external cross-validation dataset (resembling 13 clusters). As previously indicated, the clustered linear regression model is re-run on this data, and the results are presented in Table III. However, the target labels were obtained from reference techniques like gravimetry. Entire procedure of clustered linear regression was run on the internal–external validation samples. The tenfold cross-validation results were $0.332\% \text{ w/w} \pm 0.285\% \text{ w/w}$, and bootstrapping CI at 95% were $\text{BU} \pm 1.5\% \text{ w/w}$.

Discussion

The work carried out is in accordance with the analytical quality by design (AQbD) for multivariate statistical approach. The various stages of AQbD can be classified into broad section involving development, verification, and lifecycle management.

Analytical Target Profile and Risk Assessment

The robustness of an analytical methodology during its application can be addressed by using a systematic and

scientific approach to method development. The requirements associated to a measurement on a quality attribute that must be met by an analytical technique are prospectively described by employing the analytical target profile (ATP). This frequently necessitates selecting the optimal tool followed by screening the potentially fit-for-purpose analytical procedure. In this study, ATP selected is in determining the blend uniformity of amlodipine in the presence of excipients from 67 to 133% w/w. Since NIR results are prone to non-linearities, in one of the study, Rantanen *et al.* [64] identified various physical factors influencing the performance of the model. The author employed more than one quality profile in increasing the robustness of the NIR analytical procedure. That said, the major goal of this study is to ascertain whether the selection of pre-processing and processing techniques influences the results of the analytical data. To this end, risk assessment-based approach involving Ishikawa/fishbone analysis wherein variability due to material, process, environment, analyst, and product were accounted, enabling the focus to concentrate on data analysis.

Variable Selection and Model Selection

For NIR-based BU analysis, the amlodipine and mannitol were blended at 13 concentrations (between 67 and 133% w/w), while other excipients remained unchanged. HPLC and gravimetric analysis were utilized as primary tool to estimate the true API content. Different feature/variable selection involving (i) full dataset: 15,000 to $4,000 \text{ cm}^{-1}$; (ii) truncated data 1: 10,000 to $4,000 \text{ cm}^{-1}$; (iii) truncated data 2: 6,099 to 4000 cm^{-1} ; and (iv) truncated data 3: only encoded features were prepared. The top ML models on

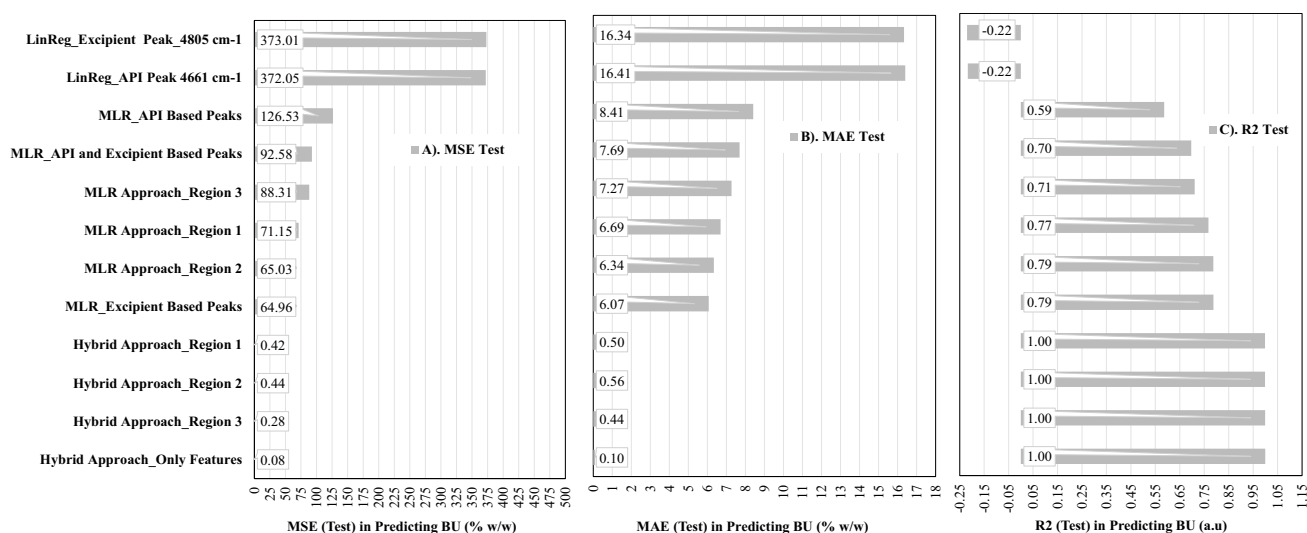
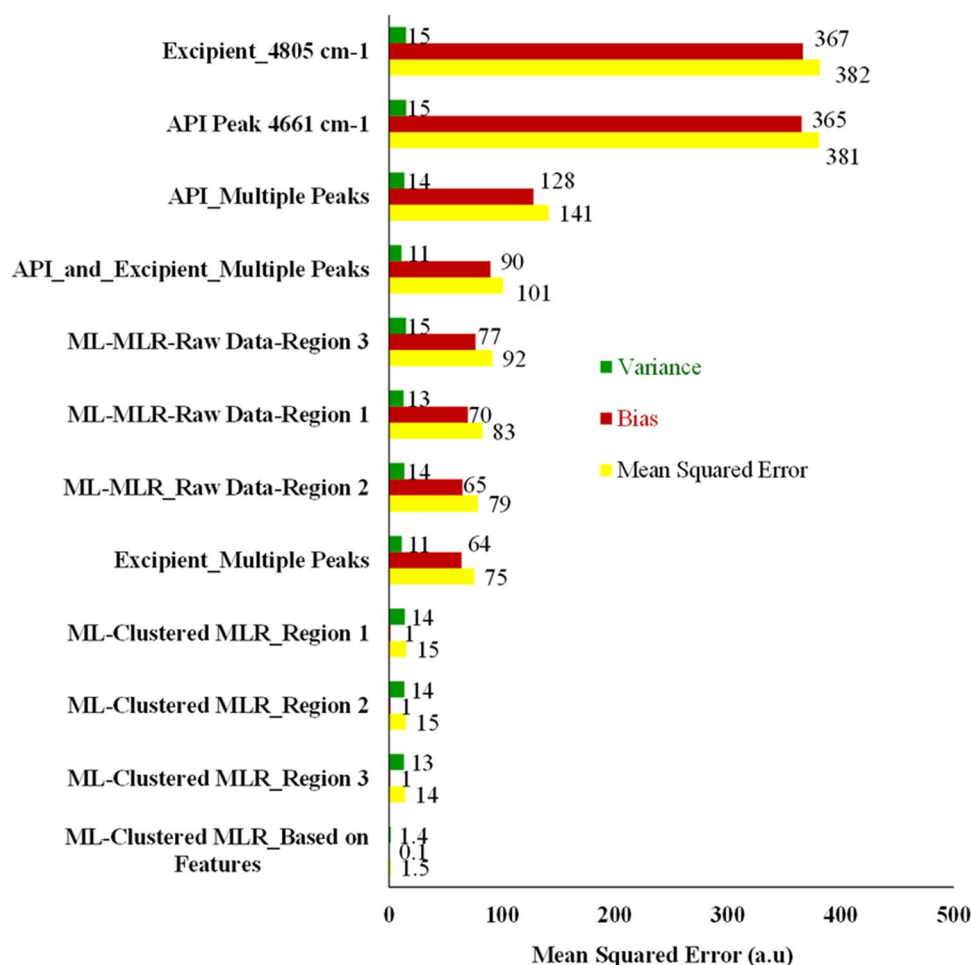


Fig. 8 Performance metrics (MSE, MAE, R^2) of test comparison for various univariate, multivariate, conventional machine learning, and hybrid machine learning approaches

Fig. 9 Bias-variance decomposition for various univariate machine learning conventional and machine learning hybrid approaches



these samples were ranked using performance metrics, sensitivity analysis, and bias-variance decomposition results; the findings are discussed below. With an MAE of 5–9% w/w, linear baseline models outperformed non-linear baseline models, which means HPLC (MAE~2.5% w/w) is more precise than NIR. Because HPLC sampling occurs after NIR analysis, its results should have been equivalent. This non-performance was attributed to violation of assumptions with respect to linear regression like normality, inhomogeneity in variance, and multicollinearity leading to bias in the results. Clustered regression models, in which the BU data is first grouped over which regression is executed, are used to overcome these restrictions. As illustrated in Fig. 10, this technique produces good linearity while reducing bias, resulting in improved BU outcomes comparable to primary analytical procedures. In low-risk scenarios, interpretability, explainability, generalizability, and/or transferability of the machine learning models are not considered to pose any concerns [65]. On the other hand, it is crucial and needs to be high risk with respect to critical quality attributes like BU [66]. Because the purpose of this project is to develop approaches to address disparities and biases caused by physical artefacts,

the ML models have been validated. Similarly, the sample size and population selection are discussed. On the other hand, the validation of the NIR method, on the other hand, is beyond the scope of this study.

Sample Size, Population, and Data Transformation

In this investigation, a sample design with 13 concentrations ranging from 67 to 133% w/w is employed. NIR data were collected five times at each concentration. In total, there are 65 samples. BVD of mean squared error (cf. section on bias-variance decomposition) was adopted to comprehend the cause-and-effect (interpretability) [67, 68]. According to BVD, it is interpreted that bias/underfitting of algorithms is the cause-and-effect of NIR non-performance. The variance factor of for all the models remained a constant. As a result, integrating and leveraging data clustering reduced bias, which enhanced the regression model's performance. DQM was employed to demonstrate how significant is the selection of data pre-processing approaches that contribute to the success of the final model (explainability) [67, 68]. Data quality metrics

Table III MAE Results of Hold-Out Data and Resampling Approaches on Hybrid Cluster Linear Regression

Internal validation (model generalizability)				
Contents	6,099–4,000 cm ^{−1} (MAE in % w/w)		Cluster features (MAE in % w/w)	
Hold-out test results	0.4389		0.077	
fivefold cross validation	0.674 ± 0.218		0.211 ± 0.174	
LOOCV (only abnormal or high error values are presented with instance/sample number)	Observation (55)	0.9638	Observation (55)	1.00000
	Observation (56)	0.7254	Observation (56)	1.00000
	Observation (57)	0.6673	Observation (57)	1.00000
	Observation (58)	1.4390	Observation (58)	1.00000
	Observation (59)	0.1505	Observation (59)	1.00000
	Observation (62)	5.0001	Observation (62)	5.00000
Bootstrapping at 95% CI	(0.078, 2.265)		(0.077, 3.866)	
Internal–external validation data (model transferability)				
Contents	6,099–4,000 cm ^{−1} (MAE in % w/w)		Cluster features (MAE in % w/w)	
Hold-out test results	0.001		0.000	
fivefold cross validation	0.327 ± 0.650		0.325 ± 0.975	
Bootstrapping at 95% CI	(0.000, 1.547)		(0.000, 1.580)	

The discrepancies in the performance error metrics MAE against internal validation/cross-validation/model generalizability (hold-out, fivefold CV, LOOCV, Bootstrapping) are due to the following observations: 55 (122% w/w), 56–59 (128% w/w), 62 (133% w/w)

The results of univariate outlier test using Mahalanobis, Jackknife, and Hotelling T2 indicate that the rows or observations describing 9 (72% w/w), 12 (78% w/w), 13 (78% w/w), 34 (100% w/w), 64 (133% w/w) were found to be closer to the upper control limit (UCL) and probable outliers (refer to Figure S10).

On the other hand, these observations in the scatter plot (refer to Figure S8) exist as usual count; hence, these were not considered as outliers (therefore were not be removed/transformed)

In summary, the observations causing performance metric errors are not anomalous or outlier datapoints. The investigation is beyond the scope of this work, however, might be a subject of future research.

assessment indicated that the multitude of linear parametric assumptions like normality, linearity, multicollinearity, and homoscedasticity was violated which might mislead prediction or inferences due to biased and/or imprecise coefficient estimates. It is well-known that NIR data is prone to physical artefacts, which requires spectral pre-processing, that utilizes derivatization (1st/2nd), SNV, and MSC/EMSC. These are, however, parametric approaches; hence, non-parametric feature transformation/categorization like data clustering/segmentation was adopted. On the segmented data, there was ~350% reduction in bias, and the clustered-linear regression model performed better than HPLC with a MAE of ~0.5% w/w. Interestingly, when the bias was addressed, there was profound decrease in variance. These results also reflect that the data or sample or population is of appropriate quality and quantity for ML analysis.

Robustness of Model

To examine the model's robustness/generalization capability, sensitivity analysis like cross-validation (such as tenfold, bootstrapping, and leave one-out) was performed [40, 67, 69]. As with hold-out validation, tenfold

cross-validation had lower error rates. Conversely, the implemented bootstrapping (at 95% confidence interval) on the internal validation dataset produced marginally wider cross-validated estimates, emphasizing a degree of variability. LOOCV was also performed, and few observations ascribing to this variability were also identified. These observations were not anomalies (refer Table III for details), according to orthogonal evaluation for outlier analysis (employing Hotelling T2 and Mahalanobis distance). In general, a range of performance metric approaches have been utilized to identify the models' robustness or generalization capabilities of the developed ML model. Reiterating, near-infrared spectra are known to be sensitive to physical artefacts, and the observed deviations in NIR-based BU results can be attributed to strong scattering effects of pharmaceutical powders [70]. Therefore, additionally, we adopted leave one out cross-validation-based performance metrics in assessing the error. Results indicated that the error was limited to few datapoints (<5); hence, no additional investigations were performed. In summary, the model generalization error of BU ± 3.5% w/w (95% confidence interval) was tolerable and suggest that the CLR model is robust [70].

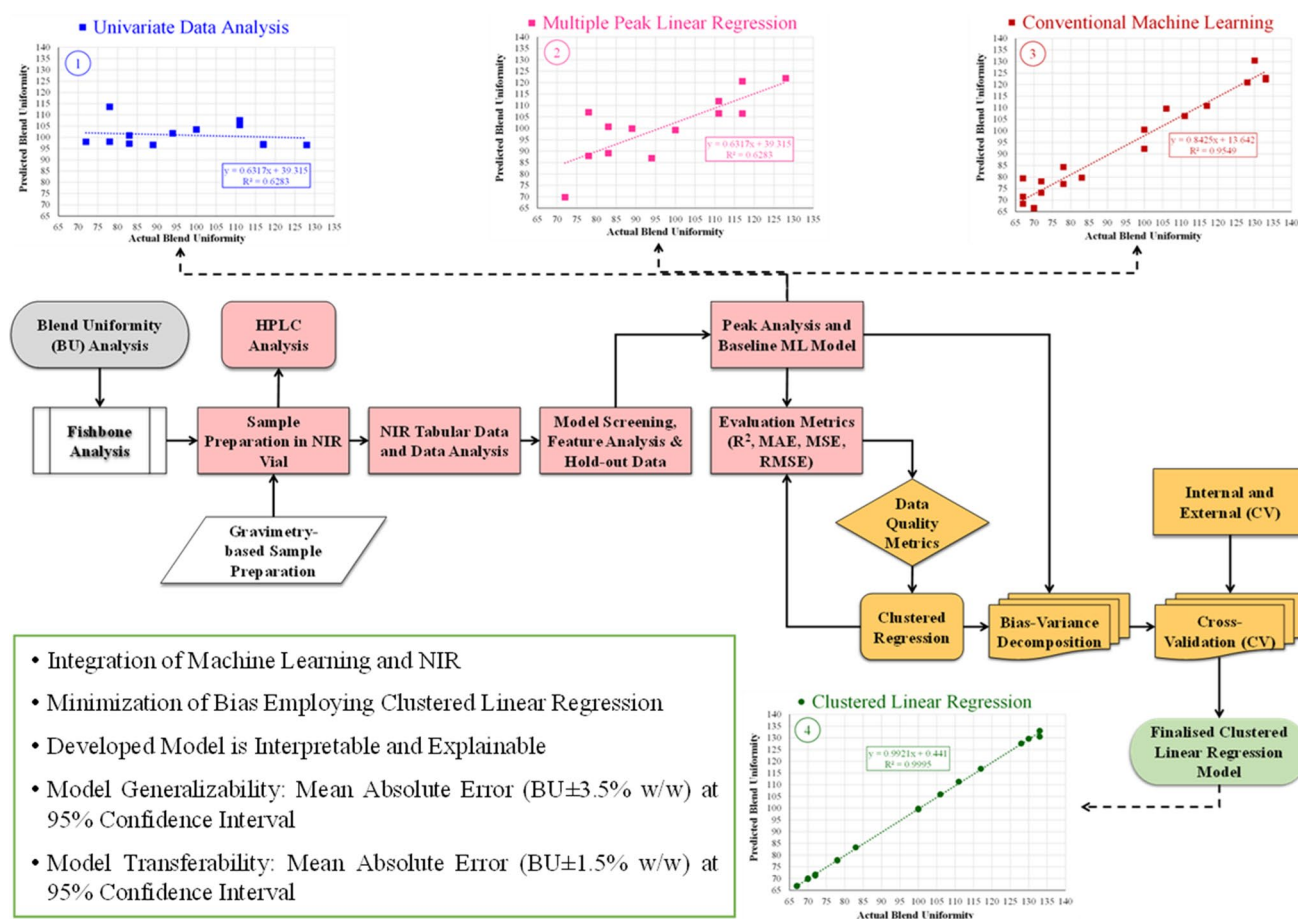


Fig. 10 Outline of the NIR-based content uniformity assessment using methodical statistical machine learning approach

Recalibration and Model Maintenance

It is well-known that when sample size is small (as in this study), cross-validated estimates are susceptible to variability or heterogeneity in the dataset. Determining the model's transferability, recalibration, and/or model maintenance cannot be a binary concept because there are no universally accepted procedures [67, 69, 71, 72]. In this study, an internal–external validation architecture was employed where in-process blends were incorporated to an already generalized internally cross-validated model. These samples were the most realistic approximations and could resemble data from alternate suppliers or new batches of raw materials/processed samples. The described procedure should be considered as a catalyst for the timely and effective implementation of post-approval changes to the NIRS procedure in conformity with regulatory bodies. Economically efficient methodology for a given dataset is chosen using the “quadro” approach (like DQM, BVD, clustering, and cross-validation) from a pool of machine learning algorithms and feature transformation modalities.

Summary and Limitations

Since the preliminary study presented herein was focussed on model's interpretability, explainability, generalizability, and transferability, the procedures employed were carried out manually like determining the number of clusters (requiring domain knowledge as well as statistical approaches) and determining the cluster label, which requires the results of primary techniques such as HPLC or gravimetry. Without automating these two steps, the regression coefficients are hard to determine on new datasets. As part of future work, we intend to address the limitations of clustered linear regression in an unsupervised manner or automate the present methodology.

Conclusion

Interestingly, machine learning on spectroscopy data is a field which is not researched much, at least within the pharmaceutical domain. In this study, a versatile NIR method

to estimate quantify amlodipine API as part of blend uniformity was successfully developed using different calibration dataset. The research presented here is based on the formalism of data analytics, which includes understanding the data, leveraging statistics, selecting appropriate machine learning approaches, and incorporating domain knowledge or first principles approach. The results show that at 95% CI, the MAE is $BU \pm 3.5\%$ w/w and $BU \pm 1.5\%$ w/w for generalizability and transferability, respectively. As a result, this approach can be regarded as a unique toolbox for solving various technical and business-related challenges. In summary, for the first time, a domain-agnostic, artefact-agnostic, interpretable, explainable, generalizable, and transferable machine learning model to measure blend uniformity of API/excipient under investigation has been demonstrated. The approaches employed were part of the development of multivariate statistical approaches (as per ICH Q14 AQB) which are extensible to machine and deep learning methodologies. According to the authors, this will enable implementation of NIR-based BU evaluation in the pharmaceutical environment for applications such as batch and continuous manufacturing, PAT-related routines, and to the best, real-time release testing.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1208/s12249-022-02403-9>.

Acknowledgements PM acknowledges interaction on AI and Machine Learning with Vetrivel PS, Accenture, Chennai, India, and blog at TheHackWeekly.com. The Open Access Publication Charge is supported by the Funding from the Graz University of Technology.

Author Contribution 1. Prakash Muthudoss: Conceptualization; software; methodology; validation; formal analysis; writing, original draft; writing, review and editing; supervision.

2. Ishan Tewari: Software, validation
3. Rayce Lim Rui Chi: Writing—investigation.
4. Kwok Jia Young: Investigation
5. Eddy Yii Chung Ann: Investigation
6. Doreen Ng Sean Hui: Resources
7. Ooi Yee Khai: Investigation
8. Ravikiran Allada: Validation, supervision, writing—review and editing.
9. Manohar Rao: Resources, review
10. Saurabh Shahane: Software, validation, supervision
11. Samir Das: Resources, writing—review and editing
12. Irfan B. Babla: Resources
13. Sandeep Mhetre: Funding acquisition, supervision, project administration
14. Amrit Paudel: Supervision, project administration, writing—review and editing

Funding Open access funding provided by Graz University of Technology.

Declarations

Conflict of Interest The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Li W, Bashai-Woldu A, Ballard J, Johnson M, Agresta M, Rasmussen H, *et al.* Applications of NIR in early stage formulation development: part I. Semi-quantitative blend uniformity and content uniformity analyses by reflectance NIR without calibration models. *Int J Pharm.* Elsevier; 2007;340:97–103.
2. Li W, Bagnol L, Berman M, Chiarella RA, Gerber M. Applications of NIR in early stage formulation development. Part II. Content uniformity evaluation of low dose tablets by principal component analysis. *Int J Pharm.* Elsevier; 2009;380:49–54.
3. Sulub Y, Konigsberger M, Cheney J. Blend uniformity end-point determination using near-infrared spectroscopy and multivariate calibration. *J Pharm Biomed Anal.* Elsevier. 2011;55:429–34.
4. Sulub Y, Wabuyele B, Gargiulo P, Pazdan J, Cheney J, Berry J, *et al.* Real-time on-line blend uniformity monitoring using near-infrared reflectance spectrometry: a noninvasive off-line calibration approach. *J Pharm Biomed Anal.* 2009;49:48–54.
5. Bakri B, Weimer M, Hauck G, Reich G. Assessment of powder blend uniformity: comparison of real-time NIR blend monitoring with stratified sampling in combination with HPLC and at-line NIR Chemical Imaging. *Eur J Pharm Biopharm Elsevier.* 2015;97:78–89.
6. Blanco M, Coello J, Iturriaga H, MasPOCH S, De La Pezuela C. Near-infrared spectroscopy in the pharmaceutical industry. Critical review. *Analyst.* Royal Society of Chemistry; 1998;123:135R–150R.
7. Luybaert J, Massart DL, Vander HY. Near-infrared spectroscopy applications in pharmaceutical analysis. *Talanta Elsevier.* 2007;72:865–83.
8. Pasquini C. Near infrared spectroscopy: a mature analytical technique with new perspectives—A review. *Anal Chim Acta Elsevier.* 2018;1026:8–36.
9. Razuc M, Grafia A, Gallo L, Ramírez-Rigo MV, Romañach RJ. Near-infrared spectroscopic applications in pharmaceutical particle technology. *Drug Dev Ind Pharm.* Taylor & Francis; 2019;45:1565–89.
10. Okubo N, Kurata Y. Nondestructive classification analysis of green coffee beans by using near-infrared spectroscopy. *Foods. Multidisciplinary Digital Publishing Institute;* 2019;8:82.
11. Cayuela-Sánchez, José A., Javier Palarea-Albaladejo, Juan Francisco García-Martín and M del CP-C. Olive oil nutritional labeling by using Vis/NIR spectroscopy and compositional statistical methods. *Innov Food Sci & Emerg Technol.* Elsevier; 2019;51:139–47.
12. Mishra P, Nordon A, Roger J-M. Improved prediction of tablet properties with near-infrared spectroscopy by a fusion of scatter correction techniques. *J Pharm Biomed Anal.* Elsevier; 2021;192:113684.

13. Mishra P, Herrmann I, Angileri M. Improved prediction of potassium and nitrogen in dried bell pepper leaves with visible and near-infrared spectroscopy utilising wavelength selection techniques. *Talanta*. Elsevier; 2021;225:121971.
14. Mishra P, Verkleij T, Klont R. Improved prediction of minced pork meat chemical properties with near-infrared spectroscopy by a fusion of scatter-correction techniques. *Infrared Phys & Technol*. Elsevier; 2021;113:103643.
15. Domokos A, Nagy B, Gyürkés M, Farkas A, Tacsí K, Pataki H, *et al.* End-to-end continuous manufacturing of conventional compressed tablets: from flow synthesis to tableting through integrated crystallization and filtration. *Int J Pharm*. Elsevier; 2020;581:119297.
16. de Oliveira Moreira AC, Braga JWB. Authenticity identification of copaiba oil using a handheld NIR spectrometer and DD-SIMCA. *Food Anal Methods Springer*. 2021;14:865–72.
17. Zhu L, Lu SH, Zhang YH, Zhai HL, Yin B, Mi JY. An effective and rapid approach to predict molecular composition of naphtha based on raw NIR spectra. *Vib Spectrosc*. Elsevier; 2020;109:103071.
18. Liu Y, Fearn T, Strlič M. Quantitative NIR spectroscopy for determination of degree of polymerisation of historical paper. *Chemom Intell Lab Syst*. Elsevier; 2021;214:104337.
19. Trenfield SJ, Tan HX, Goyanes A, Wilsdon D, Rowland M, Gaisford S, *et al.* Non-destructive dose verification of two drugs within 3D printed polyprintlets. *Int J Pharm*. Elsevier; 2020;577:119066.
20. Beć KB, Grabska J, Badzoka J, Huck CW. Spectra-structure correlations in NIR region of polymers from quantum chemical calculations. The cases of aromatic ring, C=O, C≡N and C-Cl functionalities. *Spectrochim Acta Part A Mol Biomol Spectrosc*. Elsevier; 2021;262:120085.
21. Pawar P, Talwar S, Reddy D, Bandi CK, Wu H, Sowrirajan K, *et al.* A “Large-N” content uniformity process analytical technology (PAT) method for phenytoin sodium tablets. *J Pharm Sci Elsevier*. 2019;108:494–505.
22. Xu X, Khan MA, Burgess DJ. A quality by design (QbD) case study on liposomes containing hydrophilic API: I. Formulation, processing design and risk assessment. *Int J Pharm*. Elsevier; 2011;419:52–9.
23. Xu X, Khan MA, Burgess DJ. A quality by design (QbD) case study on liposomes containing hydrophilic API: II. Screening of critical variables, and establishment of design space at laboratory scale. *Int J Pharm*. Elsevier; 2012;423:543–53.
24. Mishra P, Roger JM, Marini F, Biancolillo A, Rutledge DN. Parallel pre-processing through orthogonalization (PORTO) and its application to near-infrared spectroscopy. *Chemom Intell Lab Syst*. Elsevier; 2021;212:104190.
25. Mishra P, Roger JM, Rutledge DN, Woltering E. SPORT pre-processing can improve near-infrared quality prediction models for fresh fruits and agro-materials. *Postharvest Biol Technol*. Elsevier; 2020;168:111271.
26. Mishra P, Roger JM, Marini F, Biancolillo A, Rutledge DN. Pre-processing ensembles with response oriented sequential alternation calibration (PROSAC): a step towards ending the pre-processing search and optimization quest for near-infrared spectral modelling. *Chemom Intell Lab Syst*. Elsevier; 2022;104497.
27. Xiao-Li L, Hua L. Quantitative analysis of amlodipine besylate powder using near infrared spectroscopy combined with partial least-squares. *ICAE 2011 Proc 2011 Int Conf New Technol Agric Eng*. 2011;874–7.
28. Jiao Y, Li Z, Chen X, Fei S. Preprocessing methods for near-infrared spectrum calibration. *J Chemom*. Wiley Online Library; 2020;34:e3306.
29. Stordrange L, Libnau FO, Malthe-Sørensen D, Kvalheim OM. Feasibility study of NIR for surveillance of a pharmaceutical process, including a study of different preprocessing techniques. *J Chemom A J Chemom Soc*. Wiley Online Library; 2002;16:529–41.
30. Ozaki Y, Šašić S, Jiang JH. How can we unravel complicated near infrared spectra?—Recent progress in spectral analysis methods for resolution enhancement and band assignments in the near infrared region. *J Near Infrared Spectrosc*. SAGE Publications Sage UK: London, England; 2001;9:63–95.
31. Sadat A, Joye JJ. Peak fitting applied to fourier transform infrared and raman spectroscopic analysis of proteins. *Appl Sci*. MDPI; 2020;10:5918.
32. Roggo Y, Jelsch M, Heger P, Ensslin S, Krumme M. Deep learning for continuous manufacturing of pharmaceutical solid dosage form. *Eur J Pharm Biopharm Elsevier*. 2020;153:95–105.
33. Zhao Q, Ye Z, Su Y, Ouyang D. Predicting complexation performance between cyclodextrins and guest molecules by integrated machine learning and molecular modeling techniques. *Acta Pharm Sin B*. Chinese Academy of Medical Sciences; 2019;9:1241–52.
34. Dong J, Gao H, Ouyang D. PharmSD: A novel AI-based computational platform for solid dispersion formulation design. *Int J Pharm [Internet]*. 2021;604:120705. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S037851732100510X>
35. Gao H, Ye Z, Dong J, Gao H, Yu H, Li H, *et al.* Predicting drug/phospholipid complexation by the lightGBM method. *Chem Phys Lett [Internet]*. 2020;747:137354. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0009261420302694>
36. Ye Z, Yang W, Yang Y, Ouyang D. Interpretable machine learning methods for in vitro pharmaceutical formulation development. *Food Front*. 2021;2.
37. Yang Y, Ye Z, Su Y, Zhao Q, Li X, Ouyang D. Deep learning for in vitro prediction of pharmaceutical formulations. *Acta Pharm Sin B Elsevier*. 2019;9:177–85.
38. Gao H, Jia H, Dong J, Yang X, Li H, Ouyang D. Integrated in silico formulation design of self-emulsifying drug delivery systems. *Acta Pharm Sin B [Internet]*. 2021; Available from: <https://linkinghub.elsevier.com/retrieve/pii/S2211383521001568>
39. Han R, Xiong H, Ye Z, Yang Y, Huang T, Jing Q, *et al.* Predicting physical stability of solid dispersions by machine learning techniques. *J Control Release*. 2019;311–312.
40. Mendyk A, Paclawski A, Szafraniec-Szczyński J, Antosik A, Jamróz W, Paluch M, *et al.* Data-Driven Modeling of the Bicalutamide Dissolution from Powder Systems. *AAPS PharmSciTech*. 2020;21.
41. Miyamoto K, Mizuno H, Sugiyama E, Toyooka T, Todoroki K. Machine learning guided prediction of liquid chromatography-mass spectrometry ionization efficiency for genotoxic impurities in pharmaceutical products. *J Pharm Biomed Anal*. Elsevier; 2021;194:113781.
42. Zhao Y, Li J, Xie H, Li H, Chen X. Covalent organic nanospheres as a fiber coating for solid-phase microextraction of genotoxic impurities followed by analysis using gas chromatography-mass spectrometry. *J Pharm Anal*. Elsevier; 2021.
43. Saravanan D, Muthudoss P, Khullar P, Rose VA. Quantitative Microscopy: Particle Size/Shape Characterization, Addressing Common Errors Using ‘Analytics Continuum’ Approach. *J Pharm Sci*. 2021;110:833–49.
44. Muthudoss P, Kumar S, Ann EYC, Young KJ, Chi RLR, Allada R, *et al.* Topologically directed confocal raman imaging (TD-CRI): advanced raman imaging towards compositional and micromeritic profiling of a commercial tablet components. *J Pharm Biomed Anal*. Elsevier; 2022;114581.
45. Mishra P, Rutledge DN, Roger J-M, Wali K, Khan HA. Chemometric pre-processing can negatively affect the performance of

- near-infrared spectroscopy models for fruit quality prediction. *Talanta*. Elsevier; 2021;229:122303.
46. Alaya MZ, Bussy S, Gaiffas S, Guilloux A. Binsarity: a penalization for one-hot encoded features in linear supervised learning. *J Mach Learn Res*. 2019;20:1–34.
 47. Andersen CM, Bro R. Variable selection in regression—a tutorial. *J Chemom Wiley Online Library*. 2010;24:728–37.
 48. Rajalahti T, Kvalheim OM. Multivariate data analysis in pharmaceuticals: a tutorial review. *Int J Pharm Elsevier*. 2011;417:280–90.
 49. Sileoni V, van den Berg F, Marconi O, Perretti G, Fantozzi P. Internal and external validation strategies for the evaluation of long-term effects in NIR calibration models. *J Agric Food Chem ACS Publications*. 2011;59:1541–7.
 50. Sileoni V, Marconi O, Perretti G, Fantozzi P. Evaluation of different validation strategies and long term effects in NIR calibration models. *Food Chem Elsevier*. 2013;141:2639–48.
 51. Westad F, Marini F. Validation of chemometric models—a tutorial. *Anal Chim Acta Elsevier*. 2015;893:14–24.
 52. Snee RD. Validation of regression models: methods and examples. *Technometrics*. Taylor & Francis; 1977;19:415–28.
 53. Muthudoss P, Kumar S, Ann EYC, Young KJ, Chi RLR, Allada R, *et al*. Topologically directed confocal Raman imaging (TD-CRI): advanced Raman imaging towards compositional and micromeritic profiling of a commercial tablet components. *J Pharm Biomed Anal [Internet]*. 2022;210:114581. Available from: <https://www.sciencedirect.com/science/article/pii/S0731708522000024>
 54. Raschka S. MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack. *J open source Softw. The Open Journal*; 2018;3:638.
 55. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, *et al*. Lightgbm: a highly efficient gradient boosting decision tree. *Adv Neural Inf Process Syst*. 2017;30.
 56. Chen T, Guestrin C. Xgboost: a scalable tree boosting system. *Proc 22nd acm sigkdd Int Conf Knowl Discov data Min*. 2016. p. 785–94.
 57. Dorogush AV, Ershov V, Gulin A. CatBoost: gradient boosting with categorical features support. *arXiv Prepr arXiv181011363*. 2018;
 58. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. *Adv Neural Inf Process Syst*. 2018;31.
 59. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, *et al*. Scikit-learn: machine learning in Python. *J Mach Learn Res JMLR org*. 2011;12:2825–30.
 60. Hunter JD. Matplotlib: A 2D graphics environment. *Comput Sci & Eng. IEEE Computer Society*; 2007;9:90–5.
 61. Amigo JM. Data mining, machine learning, deep learning, chemometrics: definitions, common points and trends (Spoiler Alert: VALIDATE your models!). *Brazilian J Anal Chem*. 2021;8:45–61.
 62. Houhou R, Bocklitz T. Trends in artificial intelligence, machine learning, and chemometrics applied to chemical data. *Anal Sci Adv Wiley Online Library*. 2021;2:128–41.
 63. Montavon G, Samek W, Müller K-R. Methods for interpreting and understanding deep neural networks. *Digit Signal Process Elsevier*. 2018;73:1–15.
 64. Rantanen J, Räsänen E, Antikainen O, Mannermaa JP, Yliruusi J. In-line moisture measurement during granulation with a four-wavelength near-infrared sensor: an evaluation of process-related variables and a development of non-linear calibration model. *Chemom Intell Lab Syst*. 2001;56:51–8.
 65. Arrieta AB, Diaz-Rodriguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, *et al*. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf fusion Elsevier*. 2020;58:82–115.
 66. Szlek J, Khalid MH, Paclawski A, Czub N, Mendyk A. Puzzle out machine learning model-explaining disintegration process in ODTs. *Pharmaceutics. Multidisciplinary Digital Publishing Institute*; 2022;14:859.
 67. Mowbray M, Vallerio M, Perez-galvan C, Zhang D, Del A, Chanona ADR, *et al*. Reaction chemistry & engineering industries †. *React Chem Eng [Internet]*. Royal Society of Chemistry; 2022; Available from: <https://pubs.rsc.org/en/content/articlepdf/2022/re/d1re00541c>
 68. Oviedo F, Ferres JL, Buonassisi T, Butler K. Interpretable and explainable machine learning for materials science and chemistry. *arXiv Prepr arXiv211101037*. 2021;
 69. Salehinejad H, Kitamura J, Ditzko N, Lin A, Bharatha A, Suthip-hosuwat S, *et al*. A real-world demonstration of machine learning generalizability in the detection of intracranial hemorrhage on head computerized tomography. *Sci Rep Nature Publishing Group*. 2021;11:1–11.
 70. Rish AJ, Henson SR, Alam A, Liu Y, Drennen JK, Anderson CA. Comparison between pure component modeling approaches for monitoring pharmaceutical powder blends with near - infrared spectroscopy in continuous manufacturing schemes. *AAPS J [Internet]*. Springer International Publishing; 2022;24:1–10. Available from: <https://doi.org/10.1208/s12248-022-00725-x>
 71. Liu S, Zibetti C, Wan J, Wang G, Blackshaw S, Qian J. Assessing the model transferability for prediction of transcription factor binding sites based on chromatin accessibility. *BMC Bioinformatics BioMed Central*. 2017;18:1–11.
 72. Korolev VV, Mitrofanov A, Marchenko EI, Eremin NN, Tkachenko V, Kalmykov SN. Transferable and extensible machine learning-derived atomic charges for modeling hybrid nanoporous materials. *Chem Mater ACS Publications*. 2020;32:7822–31.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Prakash Muthudoss^{1,2} · Ishan Tewari^{3,4} · Rayce Lim Rui Chi¹ · Kwok Jia Young¹ · Eddy Yii Chung Ann¹ · Doreen Ng Sean Hui¹ · Ooi Yee Khai⁵ · Ravikiran Allada⁶ · Manohar Rao⁷ · Saurabh Shahane³ · Samir Das¹ · Irfan Babla¹ · Sandeep Mhetre¹ · Amrit Paudel^{8,9}

✉ Amrit Paudel
amrit.paudel@tugraz.at

¹ Oncogen Pharma (Malaysia), Sdn Bhd, 3, Jalan Jururancang U1/21, Hicom-glenmarie Industrial Park, 40150 Shah Alam, Selangor, Malaysia

² A2Z4.0 Research and Analytics Private Limited, Old No:810, New No:62, CTH Road, Behind Lenskart, Thirumullaivoil, Chennai, Tamilnadu, India

³ The Machine Learning Company, Beed, Maharashtra, India

⁴ Institute of Technology, Nirma University, Ahmedabad, Gujarat, India

⁵ Perkin Elmer Sdn Bhd, L2, 2-01, Wisma Academy, Jalan 19/1, Seksyen 19, 46300 Petaling Jaya, Selangor, Malaysia

⁶ Novugen Pharma (Malaysia), Sdn Bhd, 3, Jalan Jururancang U1/21, Hicom-glenmarie Industrial Park, 40150 Shah Alam, Selangor, Malaysia

⁷ PerkinElmer (India) Private Limited, Vayudooth Chambers, 12th floor, Trinity Circle, Mahatma Gandhi Rd, Bengaluru, Karnataka 560001, India

⁸ Research Center Pharmaceutical Engineering GmbH (RCPE), Inffeldgasse 13, 8010 Graz, Austria

⁹ Institute of Process and Particle Engineering, Graz University of Technology, Inffeldgasse 13/3, 8010 Graz, Austria