Journal of Pharmaceutical and Biomedical Analysis 193 (2021) 113744

Contents lists available at ScienceDirect

Journal of Pharmaceutical and Biomedical Analysis

journal homepage: www.elsevier.com/locate/jpba

Simultaneous automated image analysis and Raman spectroscopy of powders at an individual particle level

Andrea Sekulovic[a,b], Ruud Verrijk[b], Thomas Rades[a], Adam Grabarek[c,d], Wim Jiskoot[c,d], Andrea Hawe[c], Jukka Rantanen[a,*]

[a] University of Copenhagen, Department of Pharmacy, Denmark
[b] Dr Reddy's Research & Development B.V., Leiden, The Netherlands
[c] Coriolis Pharma, Martinsried, Germany
[d] Leiden University, Division of BioTherapeutics, The Netherlands

ARTICLE INFO

Article history:
Received 25 August 2020
Received in revised form 26 October 2020
Accepted 27 October 2020
Available online 4 November 2020

Keywords:
Polymorphism
Raman spectroscopy
Image analysis
Multivariate data analysis
Partial least squares - discriminant analysis
Modelling

ABSTRACT

Solid form diversity of raw materials can be critical for the performance of the final drug product. In this study, Raman spectroscopy, image analysis and combined Raman and image analysis were utilized to characterize the solid form composition of a particulate raw material. Raman spectroscopy provides chemical information and is complementary to the physical information provided by image analysis. To demonstrate this approach, binary mixtures of two solid forms of carbamazepine with a distinct shape, an anhydrate (prism shaped) and a dihydrate (needle shaped), were characterized at an individual particle level. Partial least squares discriminant analysis classification models were developed and tested with known, gravimetrically mixed test samples, followed by analysis of unknown, commercially supplied carbamazepine raw material samples. Classification of several thousands of particles was performed, and it was observed that with the known binary mixtures, the minimum number of particles needed for the combined Raman spectroscopy – image analysis classification model was approximately 100 particles per solid form. The carbamazepine anhydrate and dihydrate particles were detected and classified with a classification error of 1 % using the combined model. Further, this approach allowed the identification of raw material solid form impurity in unknown raw material samples. Simultaneous automated image analysis and Raman spectroscopy of powders at an individual particle level has its potential in accurate detection of low amounts of unwanted solid forms in particulate raw material samples.

1. Introduction

Polymorphism of drug compounds involves the occurrence of different types of packing of the same molecule in a crystal lattice [1]. The expression "solid form" can be used as a broader term to describe not only crystalline single component systems, but also amorphous matter and binary systems, such as salts, solvates, cocrystalline and coamorphous systems [2]. Different solid forms of a drug compound can be critical for the health outcome of a patient because they may affect product performance, especially when exhibiting a different particle size, shape, solubility, dissolution rate and bioavailability [3,4]. These critical material attributes can thus affect product quality, safety and efficacy [5]. Unexpected solid form changes, such as metastable polymorphs [6], elusive crystal forms [7] and unintentional seeding caused by very low amounts of an unwanted polymorph [8], have had negative effects on the availability of otherwise affordable drugs. An example of a drug product with a detrimental uncontrolled solid form change was Norvir® (ritonavir), leaving AIDS patients temporarily without a treatment [8]. To ensure that the polymorphism landscape is properly explored by the industry, the regulatory authorities have issued guidance documents on solid form characterization and control [9]. For reasons such as the ones mentioned above, the solid form diversity of particulate matter, both in solid, semi-solid and liquid products, is of particular interest to the pharmaceutical industry when aiming for a more detailed product and process understanding. Solid form screening has become an industrial practice to cope with these challenges and in this context, different methods for generating the maximal number of new forms as well as high throughput analytical methods for quantification and detection have been developed [10].

* Corresponding author.
E-mail address: jukka.rantanen@sund.ku.dk (J. Rantanen).

https://doi.org/10.1016/j.jpba.2020.113744
0731-7085/

Solid form diversity can be also a critical part of the patent portfolio of a given drug compound and its respective products. A number of litigation cases involving solid form issues have affected patent validity. For instance, the presence of forms I and II of ranitidine hydrochloride in anti-ulcer drug products led to litigation between Glaxo and Novopharm [8]. In many cases, such as in a case between Calgene and Dr Reddy's around the anti-cancer drug product Revlimid (lenalidomide), the key question is related to low amounts of a polymorphic impurity in the drug product [11]. It should be pointed out that many of these litigation cases have initiated intensive analytics in searching for very low amounts of a given solid form [8].

Conventional approaches to detect and quantify low levels of solid form impurities are based on the analysis of bulk materials [12]. Here we report an analytical technique at an individual particle level for the detection and quantification of solid form diversity. Crystals tend to grow in a specific crystallographic direction [13] and different polymorphs and crystallization/processing conditions typically result in different crystal morphologies [14]. Quantitative assessment of particle morphology has evolved with technological advancements in instrumentation and computing power. Methods based on automated particle tracking, bright/dark-field imaging and image analysis (IA) are capable of particle size and shape analysis of even hundreds of thousands of particles within a reasonable timeframe [15]. Particle morphology could therefore be used for solid form assessment, potentially resulting in a fast and sensitive analytical technique.

It has been estimated that about 90 % of studies related to polymorphism use at least two solid-state analytical techniques [16]. The use of a combination of characterization techniques enables scientific insight into the complexity of these phenomena with a higher accuracy [17]. Our study evaluates Raman spectroscopy, IA and the combination of Raman spectroscopy and IA at a single particle level in order to detect and quantify a low amount of a solid form impurity of a crystalline drug material. Partial least squares-discriminant analysis (PLS-DA) is a well-established data analytical method that combines dimensionality reduction and high prediction capability. PLS-DA is commonly used for variable selection as well as predictive and descriptive classification modeling. The PLS-DA algorithm is applicable for analyzing high dimensional data and does not assume the data to fit any distribution, making it suitable for imbalanced data with a high number of variables (n>1000) [18]. By using crystalline carbamazepine (CBZ) anhydrate (AH) and dihydrate (DH) as model solid forms, we aim to quantify solid forms in these binary mixtures, ultimately even at a single particle level. Quantification based on IA, Raman spectroscopy or a combination thereof is compared and a strategy is proposed for detection of low amounts of an unwanted solid form.

2. Experimental section

2.1. Materials

Carbamazepine (CAS 298−46 - 4) was purchased from three different commercial sources: Tokyo Chemical Industry, Co., LTD (Tokyo, Japan), Hawkins, Inc. (Minneapolis, MN, USA) and Carbosynth (Berkshire, UK). Methanol 99.8 % (67−65-1) was purchased from Sigma Aldrich Co. (St. Louis, Missouri, MO, USA). All chemicals used were of analytical reagent grade or higher. Highly purified water (Milli-Q, Millipore Inc., Denver, Massachusetts, USA) was used in all of the studies. Hydrophilic PTFE filters, Omnipore (with a 0.45 μm pore size and 47 mm diameter), were purchased from Merck (Darmstadt, Germany).

2.2. Methods

2.2.1. Preparation of carbamazepine physical mixtures (test samples)

Pure recrystallized CBZ AH and CBZ DH samples (Supporting Information, Section 1 A–D) were gravimetrically mixed by weighing a total mass of 20 mg material in an HPLC vial, mixing with a spatula and vortexing at 1500 rpm for 1 min. Powders were treated with a Zerostat 3 anti-static gun (Microtonano, Haarlem, The Netherlands) prior to mixing to reduce static charges and minimize adsorption to glass vials. Three different physical mixtures of CBZ DH and CBZ AH were prepared: test sample 1, composed of 20 % DH (w/w) (4 mg CBZ DH mixed with 16 mg CBZ AH), test sample 2, composed of 50 % DH (w/w) (10 mg CBZ DH mixed with 10 mg CBZ AH) and test sample 3, composed of 80 % DH (w/w) (16 mg CBZ DH mixed with 4 mg CBZ AH).

2.2.2. Characterization of the commercially supplied carbamazepine raw material samples

Commercially supplied CBZ raw material samples were characterized as indicated in Supporting Information, Section 1 C and 8.

2.2.3. Raman spectroscopy and IA

Pure recrystallized, test set, and commercially supplied CBZ samples were characterized by using an automatic optical microscope with an integrated Raman spectrometer (Malvern Panalytical, Worcestershire, UK). The setup allows for the measurement of Raman spectra and IA at a single particle level of up to several hundred thousand particles within a time frame of 1–10 hours. The instrument details are given in the Supporting Information Section 1 D.

Each sample was dispersed onto a glass plate by using an automated sample dispersion unit (SDU) and scanned for the particle size and shape IA of about 100,000 particles with a microscope objective that covers the particle size range of 1.8 μm – 100 μm. The particle size filter, equivalent circular diameter (ECD) 7 μm–300 μm, was used for Raman spectroscopy analysis of at least 1000 particles (Table S1, Supporting Information). A particle size filter was applied to ensure sufficient Raman intensity and adequate resolution for particle shape analysis and to avoid aggregates. A low filter cutoff of > 7 μm was used to ensure sufficient Raman intensity and adequate resolution for particle shape analysis. The spot diameter of the Raman laser beam was approximately 3 μm. The obtained spectra had a resolution of 6 cm^{-1} and comprised of a Raman wavenumber range of 150−1850 cm^{-1}.

2.2.4. Preparation of data for modeling

The raw data was exported as csv files, data matrices were prepared by using R (R statistical software environment version 3.4.3) and all data was imported into MATLAB (MATLAB Version 8.6.0.267246 R2015b) for subsequent multivariate modeling by using the PLS Toolbox (PLS Toolbox Version 8.1.1 20,131).

For each recrystallized pure solid form, at least 1000 particles were analyzed by Raman spectroscopy and IA (Table S1, Supporting Information). Subsequently, 1241 particles of each solid form were merged *in silico* into one data matrix with a total of 2482 particles representing 50 % (n/n) CBZ AH and 50 % (n/n) CBZ DH, and used for modeling (Fig. 1).

Raman spectroscopy raw data was composed of 1701 wavenumbers and IA raw data of 12 particle size and shape descriptors, per particle (Fig. 1). Both, Raman spectroscopy and IA data was randomly partitioned into 2/3 and 1/3 of the total number of particles for model training and validation. Raman spectroscopy data was

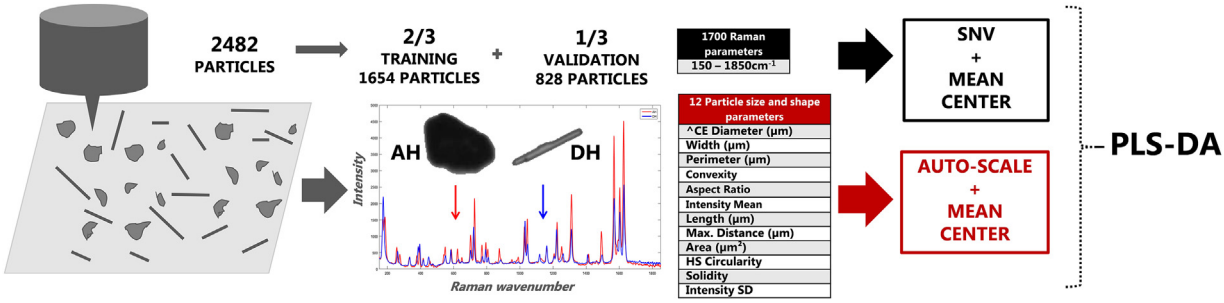

Fig. 1. Partial least squares-discriminant analysis (PLS-DA) training and validation datasets composed of IA and Raman spectroscopy data of recrystallized pure CBZ anhydrate (AH) and dihydrate (DH) samples.

smoothed by using a standard normal variate (SNV) algorithm and mean centered; IA data was auto-scaled and mean centered (Fig. 1).

2.2.5. Raman model

A supervised multivariate Raman classification model was created with the partial least squares-discriminant analysis (PLS-DA) algorithm. The number of latent variables (LVs) of the Raman PLS-DA model was determined by using internal validation, e.g. calibration/training and 10-fold venetian blinds cross validation. Raman PLS-DA model quality was expressed by the sensitivity, specificity, accuracy, precision and classification error. Performance of the model was evaluated with the test samples and commercially supplied CBZ raw material samples. Different Raman PLS-DA models were created with a stepwise increasing number of randomly sampled pure recrystallized particles of both solid state forms of the drug. To determine how many particles were needed for a classification model, the iterative process of random sampling continued with an increasing number of particles, until the predicted number of hydrate particles visually leveled off to a constant value.

2.2.6. IA model

The IA PLS-DA model was created the same way as the Raman model by using the same particles. Variable importance in projection (VIP) scores were calculated with the PLS toolbox software [19]. The minimum number of particles needed for the IA PLS-DA model was assessed in the same manner as for the Raman PLS-DA model.

2.2.7. Combined model

The combined PLS-DA model was created with concatenated, auto-scaled and mean centered Raman PLS-DA score values for the initial LVs and IA particle size and shape parmeters. Variable importance, modeling and assessment of the number of particles were evaluated the same way as for the IA model.

3. Results and discussion

3.1. Model development

3.1.1. Raman model

The Raman PLS-DA model was optimal with three LVs that cumulatively used 76.9 % of the variation in the data to classify CBZ AH and CBZ DH (Table S2, Supporting Information). Separation of CBZ AH and CBZ DH classes is visualized with the scores plot of the three LVs (Figure S4, Supporting Information). The loadings plot (Figure S5, Supporting Information) indicates that the Raman wavenumber range between 1400 cm^{-1} and 1600 cm^{-1} is the most significant for classification, in agreement with the literature [20].

The Raman model classified CBZ AH solid form with a true positive rate (TPR, sensitivity) of 98.7 % and a true negative rate (TNR, specificity) of 97.6 % (Table S3, Supporting Information). CBZ DH classification sensitivity and specificity were 97.6 % and 98.7 % respectively. The accuracy of the Raman model was on average 98.2 %, resulting in a classification error of 1.8 % (Table S3, Supporting Information). The CBZ DH classification precision was 98.7 %, compared to 97.7 % for the CBZ AH classification precision. In total, the predicted CBZ DH class was smaller and had less false positives, resulting in a higher percentage of true positives compared to CBZ AH.

The lower CBZ DH sensitivity compared to CBZ AH could have been caused by the spherical 3 µm diameter laser spot not optimally interacting with the needle-shaped CBZ DH particles, often recording the background scattering along with the particle, resulting in a low signal-to-noise ratio (Figure S3, Supporting Information). CBZ DH needle particles had a minimum width of 1.2 µm, which was smaller than the minimum width of 2.7 µm of the prism shaped CBZ AH particles. The difference in the minimum width could have been important for the detection of the particles by the camera and the laser, and particles could have been missed when the magnification was changed after particle tracking but before the Raman spectroscopy analysis, observed at 20x and 50x, respectively. A higher CBZ DH specificity compared to CBZ AH is caused by a low number of classified CBZ AH particles actually being CBZ DH (Table S4, Supporting Information). CBZ AH particles were not misclassified because their Raman spectra did not have a high background noise.

Many spectroscopic methods can be considered for fast and sensitive approaches to solid form detection. Terahertz pulsed spectroscopy has been reported to detect solid forms with a limit of detection of approximately 1 % [21]. Similarly, solid-state nuclear magnetic resonance (ss-NMR) has a detection limit of 1 % [22]. The approach presented here is based on detecting solid form impurities at a single particle level, which is a fundamentally different approach to these bulk methods. When analyzing a high number of particles, the level of detection can become as low as the measurement time and data storage allows, which ultimately may even lead to finding a single impurity particle amongst millions of particles [23].

3.1.2. Image analysis (IA) model

The IA based PLS-DA model had optimal classification with four LVs that cumulatively used 89.5 % of the variation in the data (Table S5, Supporting Information). Less clear visual separation of CBZ AH and CBZ DH classes was achieved based on IA model compared to Raman model (Figure S6, Supporting Information). The variable importance projection (VIP) plot indicated the intensity mean (IM), aspect ratio (AR) and HS as the most significant parameters in explaining the IA model (Figure S7, Supporting Information). CBZ AH particles appeared darker compared to CBZ DH particles (Figure S2, Supporting Information), resulting in a lower IM. IM is the mean greyscale value of the pixels (Figure S8, Supporting Information)

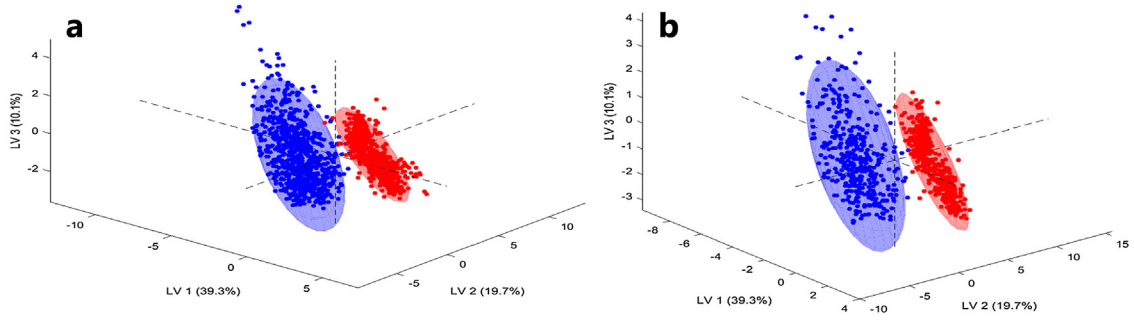

Fig. 2. CBZ AH (red) and CBZ DH (blue) classification achieved by the combined Raman-IA partial least square discriminant analysis (PLS-DA) model: Score values for the first, second and third latent variable (LV) of the auto-scaled and mean centered training dataset, n = 1654 (a) and the validation dataset n = 828 (b). The ellipsoids indicate the 95 % confidence intervals.

ranging from 0, representing black, to 255, representing white. This IM parameter is related to the optical properties of the crystals, ultimately determined by the crystal packing in these model systems. CBZ DH particles were needles, i.e., had a remarkably larger length than width compared to CBZ AH particles (Figure S2, Supporting Information), resulting in a higher AR for the CBZ AH particles (Figure S8, Supporting Information). AR is the ratio between the width and the length of the particles (Equation S6, Supporting Information). Because of the higher AR, CBZ DH particles had a lower HS circularity compared to CBZ AH particles (Figure S8, Supporting Information). HS circularity is defined as high sensitivity circularity, or squared circularity, where circularity is the measure of how close a particle is to a perfect circle based on the ratio between the area and the perimeter (Equation S8, Supporting Information).

The IA model classified CBZ AH with a sensitivity of 98 % and a specificity of 93 %. CBZ DH was classified with a sensitivity of 93 % and a specificity of 98 % (Table S3, Supporting Information). CBZ AH and CBZ DH classification was less sensitive when based on IA compared to Raman spectroscopy. Also specificity was lower and more CBZ AH and CBZ DH particles were misclassified based on IA compared to Raman spectroscopy. Moreover, the average IA model accuracy of 95.4 % was lower compared to the Raman model accuracy, resulting in a classification error of less than 5 %. The IA model precision of CBZ AH classification was 93.1 %, and 97.9 % for CBZ DH classification. For both, the IA and Raman model, precision of CBZ AH classification was lower compared to CBZ DH classification. The number of misclassified CBZ DH particles was higher compared to CBZ AH.

Even though the IM, AR and HS circularity are not significantly different between CBZ AH and CBZ DH particles (Figure S8, Supporting Information), the combination of them can be used by the multivariate PLS-DA model to discriminate between the CBZ AH and CBZ DH classes. PLS-DA is especially applicable to modeling of correlated variables such as particle shape, i.e. the AR and HS circularity. Both the IA and the Raman model classify CBZ DH with a lower sensitivity and higher specificity compared to CBZ AH. This may be because not all CBZ DH particles were needle shaped and not all CBZ DH particles that were needle shaped were observed as needles due to 2D imaging in not-longitudinal direction. Also if parts of CBZ DH particles were out of focus, they could have skewed the overall morphological parameters of these particles. IA resulted in a better detection of CBZ AH particles compared to CBZ DH due to the specific shape of CBZ AH particles, which makes it possible for the camera to focus/detect the prism shaped particles more easily. The IA model resulted in less misclassified CBZ AH particles compared to CBZ DH particles (Table S6, Supporting Information). This may be because not many CBZ AH particles were needle shaped or can be mistakenly identified as such by the IA.

Classification error of less than 5 % can be attributed to the clear visual difference between CBZ AH and CBZ DH particles and the information density offered by automatic IA. Generally used image based particle characterization methods related to the quality of the API are optical microscopy and scanning electron microscopy. Both methods have evolved with the current imaging systems towards automated timely analysis of thousands of particles leading to a larger amount of data on particulate systems [24]. A disadvantage of using the IA model was misclassification of CBZ DH particles that are not needle shaped, while a disadvantage of the Raman model was the difficulty of focusing the laser on the needle shaped CBZ DH particles. Therefore, a combination of Raman spectroscopy and IA is potentially able to classify CBZ DH with a lower classification error compared to Raman model or IA model alone, making these analytical techniques complementary.

3.1.3. Combined model

A combined model was created with four LVs that cumulatively use 77.0 % of the variation in the data for classification (Table S7, Supporting Information). The aim of using the combined model was to enable improved classification by increased information density compared to the Raman or IA model alone. The classification into CBZ AH and CBZ DH groups based on the combined model is shown in Fig. 2. CBZ AH and CBZ DH classes are visually narrower compared to the Raman model (Figure S4, Supporting Information) and more clearly separated compared to the IA model (Figure S6, Supporting information).

Based on the VIP analysis, Raman spectroscopy is the most significant variable of the combined model (Figure S9, Supporting Information). The combined model and Raman model are based on 77 % explained variance and the IA model is based on 90 % explained variance indicating the importance of the quality of the information, not the quantity. The amount of the explained variance is not related to the LV's ability to discriminate between the classes [25]. Raman spectroscopy provides chemical information compared to the physical information provided by IA. Consequently, the combined model has higher model quality parameters compared to the Raman model and the IA model alone (Fig. 3), demonstrating the complementary nature of these two analytical techniques.

3.2. Performance of the models using test samples

The performance of the IA model, the Raman model and the combined model was tested with the physically mixed test samples. The increase in the CBZ DH content between the test samples, i.e. 20 % to 50 % to 80 % (w/w) CBZ DH, was observed with the Raman model as 19.3 % to 48.3 % to 75.0 % (n/n) CBZ DH, as 17.1 % to 30.2 % to 63.3 % (n/n) CBZ DH with the IA model, and as 18.7 % to 46.9 % to 79.2 % (n/n) CBZ DH with the combined model (Table 1, Figures S10 - S11, Supporting Information). From the above it is clear that the IA model classified a lower CBZ DH content for all the test samples

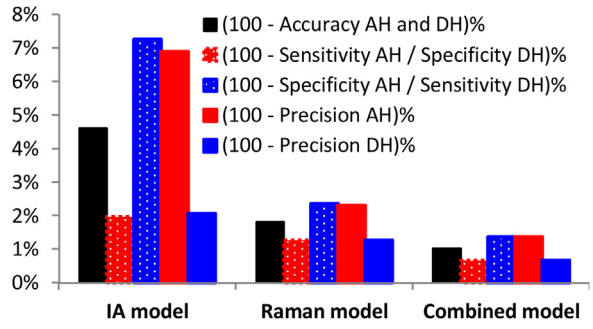

Fig. 3. Model quality attributes of the IA, Raman and combined partial least squares discriminant analysis (PLS-DA) models.

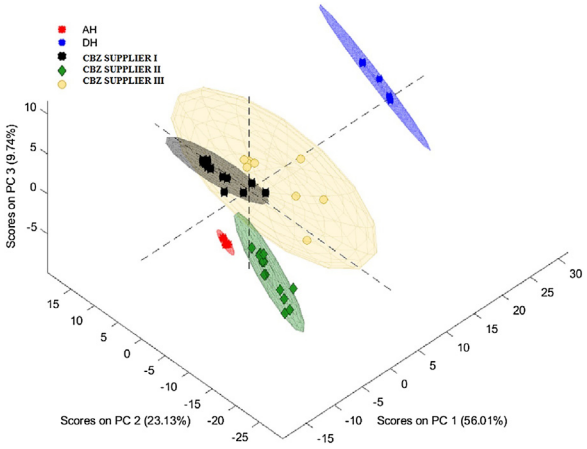

Fig. 4. Pure recrystallized CBZ AH (red), CBZ DH (blue), three different commercially supplied carbamazepine (CBZ) raw material samples (Supplier I, II and III) classification achieved by the principal component analysis (PCA) model: Score values for the first, second and third principal component (PC) of the SNV corrected and mean centered bulk Raman spectroscopy data. The ellipsoids indicate the 95 % confidence intervals.

compared to the Raman model, whereas the Raman model and the combined model determined a comparable CBZ DH content.

By combining chemical and physical information in the combined Raman - IA model, the PLS-DA approach gains improved predictive ability. On the one hand, as stated above, a challenge was observed with using Raman spectroscopy when the spot size of the laser beam in the current setup was not optimal for the needle shaped CBZ DH particles. On the other hand, IA had challenges with the detection of CBZ DH particles that were not needle shaped. The combined model was able to handle CBZ DH particles successfully because these methods are complementary. Needle shaped CBZ DH particles were successfully classified by IA and not-needle shaped CBZ DH particles by Raman spectroscopy. Having the classification error brought down to 1 % with the combined model, assessing thousands of particles will theoretically result in the detection of very low levels of CBZ AH or CBZ DH with an accuracy of 99 % (Table S3, Supporting Information).

Furthermore, a minor difference between the weight based CBZ DH content and the number based CBZ DH content was expected, as the weight of the test samples may contain a different number of particles depending on their respective density and volume. The density of CBZ DH and AH is 1.29 g/cm^3 [26] and 1.34 g/ cm^3 [27], respectively. The volume of the particles needs to be estimated based on assumptions of the third dimension of the particles as optical microscopy records only two dimensional images. When assuming the spherical equivalent volume, number based percentages 17.1 %, 19.3 % and 18.7 % DH (n/n) resulted in the respective weight based percentages 28.7 %, 32.4 % and 31.4 % DH (w/w). When assuming the third dimension was equal to the width of the particles, the weight based percentages were 24.3 %, 27.4 % and 26.6 % DH (w/w). These calculation examples highlight the challenge of transforming number to weight based percentages; careful consideration is needed when comparing numeric values of particle size results based on different analytical methods.

3.3. Performance of the models using commercially supplied (unknown) samples

The IA, Raman and combined models were tested for real life performance by assessing the DH content of commercially supplied CBZ raw material samples. The purpose was not to quantify the absolute DH content in these samples per se but to develop a multivariate based modeling strategy for such measurements in the future by comparing the performance and robustness of the models. In order to optimize the time used for these measurements, the particle size filter cutoff used for the analysis of the commercially supplied CBZ raw material samples was 40 μm. With this filter cutoff the maximum ECD of the analyzed particles was 39.99 μm, which means that the applied filter influenced the absolute composition of the samples determined by the models. For future research on the absolute composition of the samples, we recommend not to use a filter cut-off to avoid its possible influence on the absolute composition of the samples.

Commercially supplied CBZ samples had an inhomogeneous appearance when observed by scanning electron microscopy (Fig. 5 d, e, f; Supporting Information Section 1 C). This is typically an indication of variation in the solid form composition, and in this case, these samples contained a mixture of particle habits similar to both CBZ AH and DH (Fig. 5 g, h). Potential solid form diversity of commercially supplied CBZ material has been also reported earlier [3].

Solid form diversity of the commercially supplied CBZ samples was initially confirmed with unsupervised multivariate analysis of the bulk Raman spectra with principle component analysis (PCA) (Fig. 4).

In Fig. 5 a, b, c, this solid form variation is visualized by the combined Raman-IA partial least square discriminant analysis (PLS-DA) model as a score plot of the first three LVs. Inhomogeneity of these samples is clearly indicated by the mixture of the pure CBZ-AH or CBZ DH solid forms.

The Raman, IA and combined models were tested for performance by assessing the CBZ DH content of commercially supplied CBZ raw material samples. Based on the Raman model, the IA model and the combination model, CBZ material from supplier III had the highest CBZ DH content, while the material from supplier I had the lowest CBZ DH content amongst the three commercially supplied

Table 1
DH content in test samples as calculated by IA model, Raman model and combined Raman-IA model.

Test samples	DH [% (w/w)]	Particles [#]	DH [% (n/n)] IA model	DH [% (n/n)] Raman model	DH [% (n/n)] Combined model
1	20w%	2250	17.1 n%	19.3 n%	18.7 n%
2	50w%	3508	30.2 n%	48.3 n%	46.9 n%
3	80w%	1374	63.3 n%	75.0 n%	79.2 n%

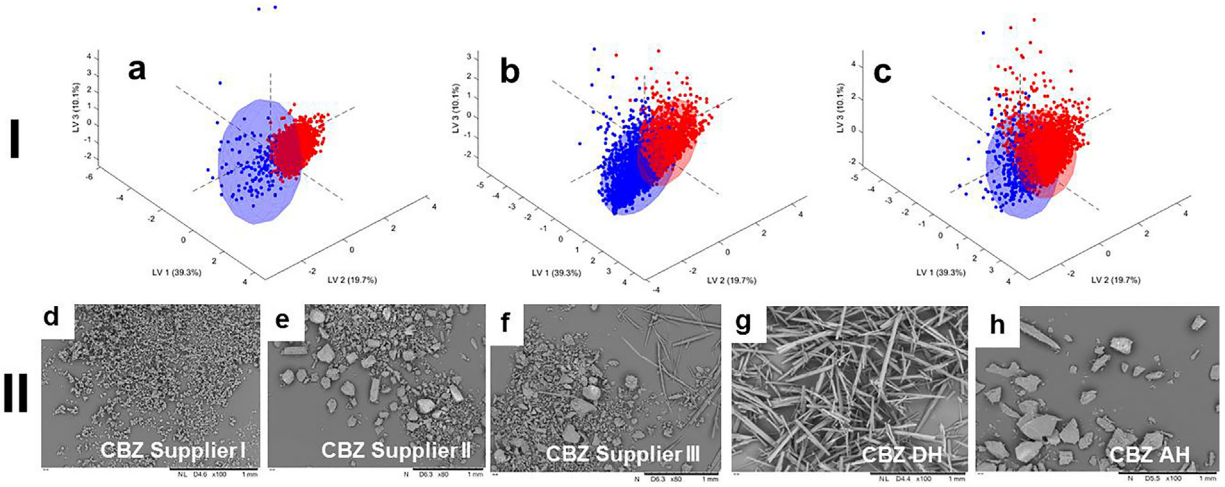

Fig. 5. (I) CBZ AH (red) and CBZ DH (blue) classification achieved by the combined Raman-IA partial least square discriminant analysis (PLS-DA) model: Score values for the first, second and third latent variable (LV) of the auto-scaled and mean centered commercially supplied CBZ samples. CBZ Supplier I, n = 3218 (a), CBZ Supplier II, n = 3428 (b) and CBZ Supplier III, n = 3712 (c). The ellipsoids indicate the 95 % confidence intervals. (II) Visual observation of the particle shape and homogeneity of commercially supplied CBZ materials (d, e, f) in comparison with pure recrystallized CBZ samples (g, h) was performed by using scanning electron microscopy.

Table 2
DH content in commercial samples as assessed by the IA, Raman and combined model.

Commercial samples	Particles [#]	DH [% (n/n)] IA model	DH [% (n/n)] Raman model	DH [% (n/n)] Combined model
CBZ Supplier I	3218	6.8%	6.1 %	5.1 %
CBZ Supplier II	3428	44.2%	64.9 %	61.8 %
CBZ Supplier III	3712	23.0%	15.2 %	14.6 %

materials (Table 2). As previously concluded based on accuracy, the CBZ DH content determined by the Raman and the combined model were similar to each other (Figure S13, Supporting Information), whereas the CBZ DH content determined by the IA model was lower for the commercial sample I and higher for commercial sample II and III compared to the Raman and the combined model results. It should be noted that the applied cutoff in particle size affected the absolute results.

3.4. Critical aspects of model development

The accuracy of prediction of the resulting models with an increasing number of particles is presented in Fig. 6. This aspect is an important part of model development, because it will directly affect the time needed for the analytical work, as well as the required computing power. The final number of particles needed for a classification model was estimated to be reached, when the predicted amount of hydrate was starting to visually level off to a constant value. It was observed that the number of particles leading to stable results is approximately 200 particles when only IA or Raman models were used, while for the combined model the results leveled off already at around 100 particles. Less particles were needed for a stable result with the combined model, because more complementary information is collected from each particle. This indicates that combining the information gained by various characterization techniques is valuable with respect to detection and quantification of solid form composition.

Particle size and shape analysis is an important analytical task related to many pharmaceutical challenges, however, this analysis is often performed manually, requiring a lot of time and effort [28]. Today, however, these methods have advanced to a level of automatic sample dispersion and particle tracking suitable for automated IA of hundreds of thousands of particles within a limited time frame. Imaging techniques such as flow imaging microscopy have been applied in analyzing the presence of small particles in parenterally administered pharmaceuticals [29]. An unsupervised model has been applied earlier for classification of particulates based on their morphology by using neural network analysis (ANN) [30]. These methods are often referred to as a 'black box' approach, because they do not allow means for controlling or deeper understanding of the root cause resulting in this classification. Taking this into the account, when analyzing unknown pharmaceutical materials, the following strategy is proposed. First, an unsupervised classification of the particulate matter could be performed based on IA as the fastest and the most accessible method for a large number of particles. The classes obtained by unsupervised learning can help in the process to identify impurities or inhomogeneity within the materials, i.e. the discriminative particle size and shape parameters can be extracted related to the inhomogeneity. Subsequently, more specific analytical chemistry method, such as Raman spectroscopy can be used for chemical and/or physical identification of the root cause for the diversity in the sample. If both the selected analytical chemistry approach and the IA of the unknown specimen are available, a combined model can be created for a fast screening of these materials with higher accuracy and precision. One can use this strategy to screen for inhomogeneity within pharmaceutical materials as well as other fine chemicals, to identify their origin based on identification of unique impurity profiles and finally, to create fast methods for quality control. By applying individual particle analysis approaches, it is theoretically possible to detect solid form impurities at very low levels, ultimately at a single particle level. IA allows for an analysis of a high number of individual particles and it has a lower theoretical sensitivity compared to Raman spectroscopy at the cost of approximately 5 % uncertainty, i.e. 5 times higher compared to the spectroscopic model.

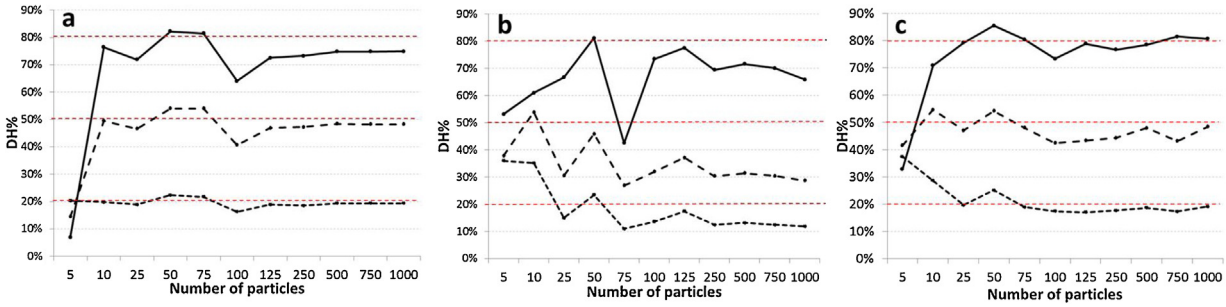

Fig. 6. CBZ DH content of test samples 1 (dotted line), 2 (broken line) and 3 (solid line), as a result of using the Raman (a), IA (b) and the combined Raman-IA PLS-DA models (c). Models are created by using data of an increasing number of randomly sampled particles.

4. Conclusion

In this study three different classification models were developed for solid form characterization at a single particle level. The IA model was the best for fast screening and comparison of different samples according to their hydrate content. The IA of at least 200 particles was sufficient for a classification error of less than 5 %. This can be relevant for a number of applications where time and cost of analysis are the highest priority. The Raman model could assess solid form composition with 2 % classification error and would therefore be the ideal choice where accuracy and precision above 98 % are needed. With a combined model based on IA and Raman spectroscopy only approximately 100 particles were needed to assess the solid form with 1 % classification error. The combination model had the highest precision, sensitivity and specificity compared to the IA model or the Raman model alone. IA and Raman spectroscopy are complementary in assessing the solid form diversity of materials and their combination resulted in an accuracy and precision of above 99 %. The reported approach is demonstrating the potential of integrated complementary measurement techniques in measuring very low amounts of solid form impurities.

CRediT authorship contribution statement

Andrea Sekulovic: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing - original draft, Writing - review & editing, Visualization, Supervision, Project administration, Funding acquisition. **Ruud Verrijk:** Conceptualization, Resources, Writing - review & editing, Project administration, Funding acquisition. **Thomas Rades:** Conceptualization, Writing - review & editing. **Adam Grabarek:** Investigation, Writing - review & editing. **Wim Jiskoot:** Resources, Writing - review & editing. **Andrea Hawe:** Resources, Writing - review & editing. **Jukka Rantanen:** Conceptualization, Methodology, Resources, Data curation, Writing - original draft, Writing - review & editing, Visualization, Supervision, Project administration, Funding acquisition.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.jpba.2020.113744.

Declaration of Competing Interest

Dr Reddy's IPDO Leiden (Leiden, Netherlands) has financed the PhD project of Andrea Sekulovic. Andrea Sekulovic and Ruud Verrijk are employed at Dr Reddy's. Jukka Rantanen and Thomas Rades have not received any consulting fees for this work.

References

[1] D.J.W. Grant, Theory and origin of polymorphism, in: Polymorphism in Pharmaceutical Solids, 1999, pp. 1–34.
[2] Q. Du, X. Xiong, Z. Suo, P. Tang, J. He, X. Zeng, Q. Hou, H. Li, Investigation of the solid forms of deferasirox: solvate, co-crystal, and amorphous form, RSC Adv. 7 (68) (2017) 43151–43160.
[3] F. Flicker, V.A. Eberle, G. Betz, Variability in commercial carbamazepine samples - impact on drug release, Int. J. Pharm. 410 (1-2) (2011) 99–106.
[4] R. Censi, P. Di Martino, Polymorph impact on the bioavailability and stability of poorly soluble drugs, Molecules 20 (10) (2015) 18759–18776.
[5] A. Banerjee, J. Qi, R. Gogoi, J. Wong, S. Mitragotri, Role of nanoparticle size, shape and surface chemistry in oral drug delivery, J. Control. Release 238 (2016) 176–185.
[6] J.D. Dunitz, J. Bernstein, Disappearing polymorphs, Acc. Chem. Res. 28 (4) (1995) 193–200.
[7] D.K. Bucar, G.M. Day, I. Halasz, G.G.Z. Zhang, J.R.G. Sander, D.G. Reid, L.R. Macgillivray, M.J. Duer, W. Jones, The curious case of (caffeine)(benzoic acid): how heteronuclear seeding allowed the formation of an elusive cocrystal, Chem. Sci. 4 (12) (2013) 4417–4425.
[8] D.K. Bucar, R.W. Lancaster, J. Bernstein, Disappearing polymorphs revisited, Angew. Chem. Int. Ed. Engl. 54 (24) (2015) 6972–6993.
[9] J.J. Maresca, Draft guidance for industry. ANDAs: pharmaceutical solid polymorphism. chemistry, manufacturing and controls information, Journal of Generic Medicines: The Business Journal for the Generic Medicines Sector 2 (3) (2005) 264–269.
[10] L.Y. Pfund, A.J. Matzger, Towards exhaustive and automated high-throughput screening for crystalline polymorphs, ACS Comb. Sci. 16 (7) (2014) 309.
[11] H.M.A. DeArment, Celgene vs. Dr. Reddy's Revlimid Patent Infringement Case Depends on Polymorph Stability; Lack Thereof May Favor Celgene Position – Experts (accessed 19 August 2019) https://www.drugpatentwatch.com/blog/celgene-vs-dr-reddys-revlimid-patent-infringement-case-depends-polymorph-stability-lack-thereof-may-favor-celgene-position-experts/.
[12] E.M. Paiva, V.H. da Silva, R.J. Poppi, C.F. Pereira, J.J.R. Rohwedder, Comparison of macro and micro Raman measurement for reliable quantitative analysis of pharmaceutical polymorphs, J. Pharm. Biomed. Anal. 157 (2018) 107–115.
[13] A. Tulcidas, N.M.T. Lourenço, R. Antunes, B. Santos, S. Pawlowski, F. Rocha, Crystal habit modification and polymorphic stability assessment of a long-acting β2-adrenergic agonist, CrystEngComm 21 (22) (2019) 3460–3470.
[14] M. Maghsoodi, Role of solvents in improvement of dissolution rate of drugs: crystal habit and crystal agglomeration, Adv. Pharm. Bull. 5 (1) (2015) 13–18.
[15] B.W. Kammrath, A. Koutrakos, J. Castillo, C. Langley, D. Huck-Jones, Morphologically-directed Raman spectroscopy for forensic soil analysis, Forensic Sci. Int. 285 (2018) 25–33.
[16] N. Chieng, T. Rades, J. Aaltonen, An overview of recent studies on the analysis of pharmaceutical polymorphs, J. Pharm. Biomed. Anal. 55 (4) (2011) 618–644.
[17] E.A. de Moura, M.V.C. Terto, E.A. de Moura Mendonca, J.V. Procopio, et al., Solid-state form characterization of riparin I, Molecules 22 (10) (2017).
[18] L.C. Lee, C.Y. Liong, A.A. Jemain, Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional (HD) data: a review of contemporary practice strategies and knowledge gaps, Analyst (Cambridge, U. K.) 143 (15) (2018) 3526–3539.
[19] S. Favilla, C. Durante, M.L. Vigni, M. Cocchi, Assessing feature relevance in NPLS models by VIP, Chemometr. Intell. Lab. Syst. 129 (2013) 76–86.
[20] K. Kogermann, J. Aaltonen, C.J. Strachan, K. Pollanen, P. Veski, J. Heinamaki, J. Yliruusi, J. Rantanen, Qualitative in situ analysis of multiple solid-state forms using spectroscopy and partial least squares discriminant modeling, J. Pharm. Sci. 96 (7) (2007) 1802–1820.
[21] C.J. Strachan, P.F. Taday, D.A. Newnham, K.C. Gordon, J.A. Zeitler, M. Pepper, T. Rades, Using terahertz pulsed spectroscopy to quantify pharmaceutical polymorphism and crystallinity, J. Pharm. Sci. 94 (4) (2005) 837–846.
[22] K. Maruyoshi, D. Iuga, A.E. Watts, C.E. Hughes, K.D.M. Harris, S.P. Brown, Assessing the detection limit of a minority solid-state form of a pharmaceutical by H double-quantum magic-angle spinning nuclear magnetic resonance spectroscopy, J. Pharm. Sci. 106 (11) (2017) 3372–3377.

[23] P.O. Okeyo, O. Ilchenko, R. Slipets, P.E. Larsen, A. Boisen, T. Rades, J. Rantanen, Imaging of dehydration in particulate matter using Raman line-focus microscopy, Sci. Rep. 9 (1) (2019), 7525-7525.
[24] J.F. Gamble, M. Tobyn, R. Hamey, Application of image-based particle size and shape characterization systems in the development of small molecule pharmaceuticals, J. Pharm. Sci. 104 (5) (2015) 1563–1574.
[25] D. Ballabio, V. Consonni, Classification tools in chemistry. Part 1: linear models. PLS-DA, Anal. Methods 5 (16) (2013) 3790–3798.
[26] A. Kogan, I. Popov, V. Uvarov, S. Cohen, A. Aserin, N. Garti, Crystallization of carbamazepine pseudopolymorphs from nonionic microemulsions, Langmuir 24 (3) (2008) 722–733.
[27] A.L. Grzesiak, M. Lang, K. Kim, A.J. Matzger, Comparison of the four anhydrous polymorphs of carbamazepine and the crystal structure of form I, J. Pharm. Sci. 92 (11) (2003) 2260–2271.
[28] L. Hellén, J. Yliruusi, P. Merkku, E. Kristoffersson, Process variables of instant granulator and spheroniser: I. Physical properties of granules, extrudate and pellets, Int. J. Pharm. (Amsterdam, Neth.) 96 (1-3) (1993) 197–204.
[29] M.M. Schack, E.H. Møller, J.F. Carpenter, T. Rades, M. Groenning, A platform for preparing homogeneous proteinaceous subvisible particles with distinct morphologies, J. Pharm. Sci. 107 (7) (2018) 1842–1851.
[30] C.P. Calderon, A.L. Daniels, T.W. Randolph, Deep convolutional neural network analysis of flow imaging microscopy data to classify subvisible particles in protein formulations, J. Pharm. Sci. 107 (4) (2018) 999–1008.