Classification of tablet formulations by desorption electrospray ionisation mass spectrometry and transmission Raman spectroscopy

Authors

- **Adam J Taylor**
  0000-0003-0501-8886
  Chemical and Biological Sciences Department, National Physical Laboratory, Teddington, UK
- **Dimitrios Tsikritsis**
  Chemical and Biological Sciences Department, National Physical Laboratory, Teddington, UK
- **Alex Dexter**
  Chemical and Biological Sciences Department, National Physical Laboratory, Teddington, UK
- **Amy Burton**
  0000-0002-6970-1492
  Chemical and Biological Sciences Department, National Physical Laboratory, Teddington, UK
- **Josephine Bunch**
  0000-0002-4257-1296
  Chemical and Biological Sciences Department, National Physical Laboratory, Teddington, UK; Department of Metabolism, Digestion and Reproduction, Imperial College London, London, UK; The Rosalind Frankilin Insitute, Harwell, UK
- **Natalie A Belsey**
  0000-0001-7773-0966 natalie.belsey@npl.co.uk
  Department of Chemical and Biological Sciences, National Physical Laboratory, Teddington, UK; Department of Chemical and Process Engineering, University of Surrey, Guildford, UK

Abstract

Discrepancies or defects in active ingredients, excipients and coatings that form solid oral dosage forms can both impact product quality and provide hallmarks of off-brand or counterfeit products. There is therefore a need for rapid and continuous analytical techniques that can assess and classify product differences of intact samples at- or near the production line, or in analytical labs, ideally without resorting to product dissolution.

Here we test the ability of two rapid ambient chemical characterisation methods to discriminate between solid dosage forms: desorption electrospray ionization mass spectrometry (DESI MS) and transmission Raman spectroscopy. These two techniques are highly complementary, offering greater sensitivity to the analysis of the surface and the tablet bulk, respectively. The data sets generated were then used to test a variety of classification algorithms including linear discriminate analysis (LDA), tree-based methods, a simple neural network, and support vector machines (SVM). The highest performing algorithms for DESI MS were the SVM, with an additional performance boost when used with a polynomial kernel. For transmission Raman data, an LDA model was found to be the most effective.

Introduction

Variability in active ingredients, excipients, the thickness and integrity of coatings and the presence of impurities in solid oral dosage forms all negatively affect product performance. Inferior quality attributes can be useful to identify off-brand or counterfeit products. There is a need for rapid and continuous analytical techniques that can assess and classify product differences of intact samples at- or near the production line, or in analytical labs, ideally without resorting to product dissolution[1]. Rapid measurement tools are particularly important to enable continuous monitoring, including necessary to support the change from batch to continuous manufacturing, and high-throughput batch release processes. Analytical methods are required to monitor both the actives, coatings and consistency of the product: For example, in addition to the total active pharmaceutical ingredient (API) content, insight is also needed on degradation products, impurities, (co-)crystallinity/presence of polymorphs, and content uniformity. The ability to monitor excipient deviations in tablet coatings is of great importance, particularly for functional coatings, such gastro-resistance[2].

Quantitative analysis of pharmaceutical tablets is routinely performed by high-pressure liquid chromatography (HPLC) which offers accurate and sensitive measurements of the active ingredient(s) and excipients, in addition to the presence of any contaminants. However, solution-based analytical methods are destructive and labor-intensive.

Mass spectrometric methods can provide unlabelled identification, both of expected ingredients in known samples and of contaminants or components of unknown formulations. Ambient ionisation mass spectrometry approaches including DESI (desorption electrospray ionisation) and DART (direct analysis in real time) facilitate the desorption and ionisation from the surface of samples at atmospheric conditions, without dissolution or additional sample preparation. They are therefore potentially useful tools for rapid assessment of solid oral dosage forms

Optical spectroscopy techniques offer rapid, non-destructive analysis, including polymorphic identification[3], and are also able to measure insoluble ingredients that may not be readily detected in dissolution testing. They have consequently been exploited for in-line process analytical testing and as quality control tools[1]. For example, infra-red-based techniques such as Fourier-transform infrared spectroscopy[4]; and infrared spectroscopy classification of 3,4-Methylenedioxymethamphetamine (MDMA) containing tablets[5]. Near-infrared spectroscopy (NIR) is one of the most commonly used process analytical tool[1], however Raman spectroscopy provides complementary information and has grown in popularity in recent years, since it provides more distinct spectral features, and is better-suited to analysis in aqueous environments owing to the relatively weak strength of the Raman O-H band. Technological advancements have facilitated miniaturization, increased speed and reduced cost, resulting in more widespread implementation[6].

Ambient ionisation mass spectrometry of tablets

Desorption electrospray ionisation uses a charged electrospray of organic solvent which, when directed at the sample surface in proximity to the mass spectrometry inlet, desorbs ions from the sample which may be taken up into a mass spectrometer[7]. As this process takes place at ambient pressure and with a flexible geometry, the technique is suited for the analysis of a wide range of samples including explosives on surfaces[8], fingerprints[9], plants[10] and tissues[11,12].

In one of the early applications of DESI MS, Chen *et al.* demonstrated the use of DESI MS to profile tablets containing loratadine, folic acid, acetaminophen (paracetamol), aspirin, melatonin or caffeine[13]. Optimization of DESI parameters including voltage, solvent delivery and capillary temperature facilitated analysis at up to three scans per second.
Subsequent studies using DESI MS of tablets have focused on targeted analysis for APIs. For example, the identification of MDMA and amphetamine derivatives in ecstasy tablets[14], counterfeit artesunate antimalarial tablets[4,15] and antiviral capsules[16].

For ambient mass spectrometry to be deployable in the field for counterfeiting applications, or in manufacturing environments for quality assurance and quality control (QA/QC), the mass spectrometer must be compact. Several designs for small field-deployable mass spectrometers have been demonstrated with DESI MS sources[17,18].

Each of these applications has targeted expected components of the tablet of interest, predominantly APIs or excipients. However, in manufacturing QA/QC and counterfeit-detection applications, additional information on unexpected changes in API or excipient source or quality, as well as the introduction of contaminants may be of importance. Untargeted multivariate and machine learning approaches are therefore of interest to determine differences between samples using all spectral information.

Classification approaches for mass spectrometry applications are proving powerful in a range of applications. The two most widespread applications of classification in mass spectrometry are in disease diagnosis and determination of bacterial type[19]. A range of classification algorithms have been applied to mass spectrometry and spectroscopy data. Partial least squares discriminant analysis (PLS-DA) is most commonly reported, although a range of algorithms including neural networks, and support vector machines[20,21] have been reported. Several publications have evaluated different classification algorithms for mass spectrometry in proteomics[22] and metabolomics applications[23], but unsurprisingly the optimal algorithm depends greatly on

the nature of the input data. Classification approaches are becoming more accessible through modelling tools with consistent grammar and data structure, and their integration into mass spectrometry software [ScilsLab,Waters software][23].

Notably, classification of rapid evaporative ionisation MS enables real-time classification of tissue types during surgery[24]. Classification of REIMS data has also found applications in food security[25] and bacterial speciation[26]. Classification approaches have also been widely employed in mass spectrometry imaging data, particularly in the classification of cancerous tissue[27].

Raman spectroscopy analysis of tablets

Raman spectroscopy exploits the inelastic scattering of light by the sample to reveal valuable chemical and structural information. Information can be obtained from the sample in a non-destructive manner, making it a popular process analytical technology tool. Raman spectroscopy can be performed in a variety of sampling configurations/geometries to suit different applications. Confocal Raman microscopy can provide detailed chemical mapping with high spatial resolution, however this is generally reserved for forensic investigation rather than continuous monitoring, since it requires lengthy acquisition times. Sub-sampling issues associated with conventional backscattered Raman can be overcome by strategies such as sample rotation in conjunction with spectral averaging, or simultaneous wide angle illumination[28].

Matousek et al demonstrated the ability of transmission Raman spectroscopy to probe deep into turbid materials such as pharmaceutical tablets[29] and provide information on their bulk properties[30]. In contrast to conventional backscattered Raman, in transmission Raman spectroscopy the beam passes through the full thickness of the tablet, sampling a much larger volume of the material, and consequently provides more representative sampling of the chemical composition of the sample[31]. Although Raman scattering intensity is linear with concentration within the same confocal plane, transmission Raman signal intensity is slightly biased towards the bulk of the tablet relative to the exterior due to internal scattering[32]. In contrast, DESI MS sampling is biased towards the surface/coating composition. Therefore, in combination, these two techniques provide a powerful toolkit with which to assess compositional differences between pharmaceutical tablet formulations.

Raman spectra of complex mixtures such as solid dosage forms often have complicated spectra with overlapping peaks. For this reason, multivariate techniques are often applied to help identify the components of interest and changes in chemistry. The selection and use of unsupervised and/or supervised techniques on Raman spectra rely on factors such as prior knowledge of the raw component spectra, and the quantity and complexity of the spectra[33].

As with mass spectrometry, classification of Raman spectroscopy data has been primarily focused on disease diagnostics[34–36] and bacterial analysis[37,38]. Other noteworthy examples of the use of classification in Raman spectroscopy include differentiation of narcotics[39], pharmaceuticals[40,41], and counterfeit tablets[42].

There have been relatively few comparisons of different classification methods for Raman spectroscopy data. Zheng et al. compared support vector machine (SVM), linear discriminant analysis (LDA) and k-nearest neighbours (KNN) methods to classify renin hypertension from Raman data from serum[43]. They found that SVM and LDA performed similarly, and both outperformed the KNN algorithm. Partial least squares (PLS) and PLS discriminant analysis are also commonly used methods in characterizing tablets. However, care is required

depending on the data quantity and the pre-processing performed[44]. Qun et al. tested the classification of expired drugs using PLS-DA, SVM and KNN, and reported that SVM gave the strongest performance[45]. Fransson *et al.* tested the performance of multivariate methods including PLS, classical least squares (CLS) and multivariate curve resolution (MCR) for classification of pharmaceutical tablets[46].

Objective

In this study we set out to explore the potential of DESI MS and transmission Raman spectroscopy to distinguish commercially available pharmaceutical tablets with similar or different formulations. Pairing DESI with transmission Raman spectroscopy was of particular interest due to their complementarity and relative abilities to sensitively probe the surface vs the bulk of the tablets. Classification of tablets based on both active ingredients and excipients has the potential to be used for in-line quality control measures during pharmaceutical manufacturing, and for rapid counterfeit testing. As such we have tested a range of classification algorithms on their capability to differentiate these tablets using a range of pre-processing methods to determine the best approaches to use in different applications.

Experimental

Samples

Samples were selected from commercially available off-the-shelf products and purchased from a local supplier. Their names, active ingredients, listed excipients and UK Medicines and Healthcare Products Regulatory Agency (MHRA) product license numbers are included in Table 1.

Table 1: Details of the tablet types analysed. Letter codes for each type are used throughout. Active ingredient list shows mass per tablet as stated on product information sheet. Total mass shows mean ± 1 SD for n = 8 tablets (n=7 for type B).

Type	Product name	Active ingredients	Listed excipients	Tablet mass	MHRA licence
A	Anadin Extra Tablets	300 mg Aspirin, 200 mg Paracetamol, 45 mg Caffeine	Maize starch, microcrystalline cellulose (E460), hydrogenated vegetable oil, hydroxypropyl methylcellulose (E464), polyethylene glycol, pregelatinised starch and povidone	662 ± 11 mg	PL 00165/5013R
B	Tesco Paracetamol Extra Tablets	500 mg Paracetamol, 65 mg Caffeine	tarch, povidone k-30, povidone k-90, croscarmellose sodium, talc, stearic acid and magnesium stearate	607 ± 6 mg	PL 08977/0025

Type	Product name	Active ingredients	Listed excipients	Tablet mass	MHRA licence
C	Tesco Paracetamol Tablets	500 mg Paracetamol	Potato Starch, pregelatinised starch, magnesium stearate, povidone, stearic acid and talc	550 ± 3 mg	PL 08977/0014
D	Tesco Extra Power Pain Control Tablets	300 mg Aspirin, 200 mg Paracetamol, 45 mg Caffeine	Povidone, hydroxypropylcellulose, stearic acid, microcrystalline cellulose, maize starch, pregelatinised starch, hydroxypropyl methylcellulose 5cPs, hydroxypropyl methylcellulose 15cPs, macrogol 4000	632 ± 6 mg	PL 29831/0164

DESI MS

DESI MS measurements were performed on a Synapt G2-Si Q-IM-ToF mass spectrometer (Waters Corp, Milford, MA, USA). The instrument was operated in 'resolution mode'. The ion mobility cell was not used. Positive ion mode spectra were collected with a scan time of 1 second across a mass range of *m/z* 50 to *m/z* 1200. The instrument was fitted with a prototype DESI source (Waters Corp, Milford, MA, USA), with the sprayer configured for electroflow focusing with a fused silica capillary sitting approximately 1 mm behind a 200 µm steel orifice. Methanol with 5 % water by volume was delivered at 2 µl/m by a pressure pump (Dolomite). Nitrogen gas was delivered at 0.2 MPa. The spray voltage was set at 5 kV. A heated inlet capillary was set to a calibrated temperature of 400 °C using a PID (Waters Research Centre, Budapest, Hungary). Tablets were sampled by holding the tablet 1-2 mm away from the DESI spray head using plastic tweezers. For training data, acquisition was started with the tablet already under the spray head, such that only data from the tablet surface was acquired, while validation data was collected continuously.

Transmission Raman spectroscopy

Transmission Raman spectra were acquired using a Renishaw InVia Qontor Raman microscope equipped with a 830 nm excitation source fibre-coupled to an InVia transmission Raman accessory (Renishaw plc, Wotton-under-Edge, Gloucestershire), in a temperature controlled environment. Light was collected in transmission with the 5x air objective lens (0.12 NA, N-PLAN, Leica, Wetzlar, Germany). Tablets were carefully placed onto a flat silicon sample support with a hole just smaller than the tablet dimensions, so that the excitation beam was able to pass through the tablet but not the sample support. Six tablets were analysed of each type. Three measurement replicates were acquired, each complete data set was collected on three separate days.

For all tablets, extended spectra were acquired using Renishaw Wire (version 5.3) software for the spectral range of 50 to 1800 cm^{-1}, with an acquisition time of 30 seconds, and 5 accumulations. Laser power was set to 100 % which has been measured at the sample to be approximately 117 mW. An internal silicon calibration reference spectrum was acquired each day to correct the Raman shift of the data.

Data analysis

All data were analyzed in R version 3.6.2 (2019-12-12) "Dark and Stormy Night" and RStudio Server version 1.2.5019. Analysis was conducted using the tidyverse[47] and tidymodels[48] metapackages. Raman data preprocessing was conducted in MATLAB 2020a. All analysis was performed on a Linux workstation (Intel Core i9-7900X CPU with 10 cores @ 3.30 GHz, 128G RAM, Ubuntu 16.04.6 LTS).

DESI preprocessing

For model development and comparison, data were converted from Waters *raw* format to *mzML* format using ProteoWizard MSConvert version 3.0.19239-0ae547798[49]. These were read into R using the *mzR* package[50,51]. All spectra were re-binned onto the same mass axis with a bin width of *m/z* 0.01. A mean spectrum of all training data was peak picked using the findPeaks function from the practical numerical math functions (pracma) package[52] with a peak intensity threshold of three times the median intensity of the spectrum. 1217 peaks were found. Each spectrum was then individually integrated across the found peak widths to form a datacube. Each scan of the validation dataset was similarly integrated across the peak widths from the training dataset. Reduced peak datacubes were generated by filtering for the top 1207 most intense peaks. Down-binning to simulate reduced mass resolving power was performed by rounding m/z values and summing intensities within each rounded *m/z* bin.

Transmission Raman spectroscopy preprocessing

Cosmic ray removal was performed automatically by Renishaw Wire (version 5.3) and spectra exported to .txt format. The Raman spectra were baseline corrected using the *msbackadj* Matlab function[53]. The baseline was estimated within multiple shifted windows of width 20 separation units, then a spline approximation was used to regress the varying baseline to the window points. While a spline fitting may not be appropriate for Raman datasets where broad peaks are present and should be used with caution; in this study peaks were relatively sharp and spline fitting was seen to provide a small improvement in qualitative fit over polynomial and Mexican-hat methods. The estimated baseline for each spectrum was then subtracted from the corresponding original. The background subtracted spectra were read in R for subsequent processing and analysis. The data were normalized to total spectrum intensity and the Raman shift recalibrated using the weighted-mean centroid to the 520.7 cm^{-1} peak from the daily Si wafer sample spectrum as a reference. Extended spectra were truncated to a wavenumber range between 250 cm^{-1} and 1700 cm^{-1}. Due to the limited number of wavenumber bins and the challenges of peak-picking Raman data, the continuous data were taken forward for classification.

Classification

Spectra were collated into a V-fold validation with 10 partitions and 10 repeats in a 9/1/10 (train/test/total) split. Highly correlating variables (Pearson correlation > 0.9) were removed from DESI MS data. Data were centered around the arithmetic mean and scaled to have a standard deviation of one. Underrepresented classes (for DESI MS, the background class) were up-sampled to have the same frequency as the most occurring level. Each training fold was applied to a range of classification algorithms using the tidymodels package. All models were implemented with their default parameters beyond setting to classification mode. The functions, engines and default parameters used for each model are provided in supplementary table 1. These models were then used

to predict each testing fold. For each fold the F1 score was calculated. For DESI MS the algorithm with the highest F1 score, a support vector machine with a polynomial kernel was selected for further model tuning on a single random sample (without replacement) of 10 % the training data. For transmission Raman data a LDA model was selected. A final model was fitted on all the training data. These models were used to predict the test independent test sets. Cosine similarity between spectra were calculated using the cosine function from the coop package[url?]"https://cran.r-project.org/package=coop". Considering the angle between vectors, rather than magnitude, cosine similarity provides a useful and robust measure of spectral similarity for highly multivariate datasets[54].

Results and discussion

Acquisition of DESI MS spectra from tablets

Rich mass spectra were rapidly obtained from each tablet type when held under the DESI sprayer and MS inlet capillary (Figure 1A). DESI MS spectra from tablet types A and D are dominated by repeating polymeric peaks above *m/z* ~700, indicating the presence of a polymer film coating while spectra from tablet types B and C are less complex, being dominated by peaks below *m/z* 700. The maximum intensity of the spectra collected in the absence of a tablet ("background") is lower than for spectra where a tablet is present.

The polymeric peak sequences observed in tablet type A and D above *m/z* 700 are distinct from one another with several different peak sequences observed (Figure 1B). Tablet types A and D both exhibit a clear sequence of peaks spaced by *m/z* 44 the a signally charged unit difference from $[C_2H_4O]^+$. For both tablet types peaks above *m/z* 1000 are most intense. In tablet type A this predominantly consists of isotope clusters separated by *m/z* 14.68, with peaks separated by *m/z* 0.33, indicating a 3+ charge state of the loss of [C2H4O]. Conversely, in tablet type D the isotope clusters are separated by *m/z* 11.01 with peak spacing of *m/z* 0.25, indicating a 4+ charge state and the loss of $[C_2H_4O]$. This indicates that while both are coated with a polyethylene glycol (PEG) polymer, the molecular weight or coating application may differ between the brand name (Type A) and generic product (Type D).

All three active ingredients were annotated from a mean spectrum within 15 ppm mass accuracy. Boxplots of single scan intensity show distinct differences in active intensity between types (Figure 1C, 1D, 1E). Caffeine is present in tablet types A, B and D and absent from type C. Median intensity is highest for the uncoated tablet type B. Intensity for tablet type C is similar to that in the background. While aspirin is present in tablet types A and D, the detected intensity in type B is high. This may represent an isobaric compound also present in type B or carry over from sampling tablet type A. Paracetamol is present in, and detected in all tablet types, although is most intense in tablets without a film-coating (types B and C) which may otherwise mask signal from the underlying bulk material. Although tablets A and D contained similar paracetamol content, tablet A exhibited much lower DESI signals for paracetamol, relative to tablet D. This could be explained by a difference in the integrity between the film coatings of tablets A and D, their relative solubility in the DESI solvent, or potentially drug migration into the film coating, in the case of tablet D. Variance for all actives and tablet types was high, indicating the need for consistent sampling procedures and robust classification approaches.

Acquisition of Transmission Raman spectra from tablets

Mean transmission Raman spectra for each tablet type (Figure 1F) show very similar profiles for tablet types A and D (the film coated aspirin, paracetamol, and caffeine tablets) and between types B (paracetamol and caffeine and C (paracetamol only). Peaks characteristic of each active ingredient are observed (Caffeine: 555 cm^{-1}, Paracetamol: 858 cm^{-1}, Aspirin: 1192 cm^{-1})[55–57]. However, no clear differences between tablets A and D were observed. As expected based on the tablet composition, the predominant spectral features relate to the actives rather than from the coating ingredients.

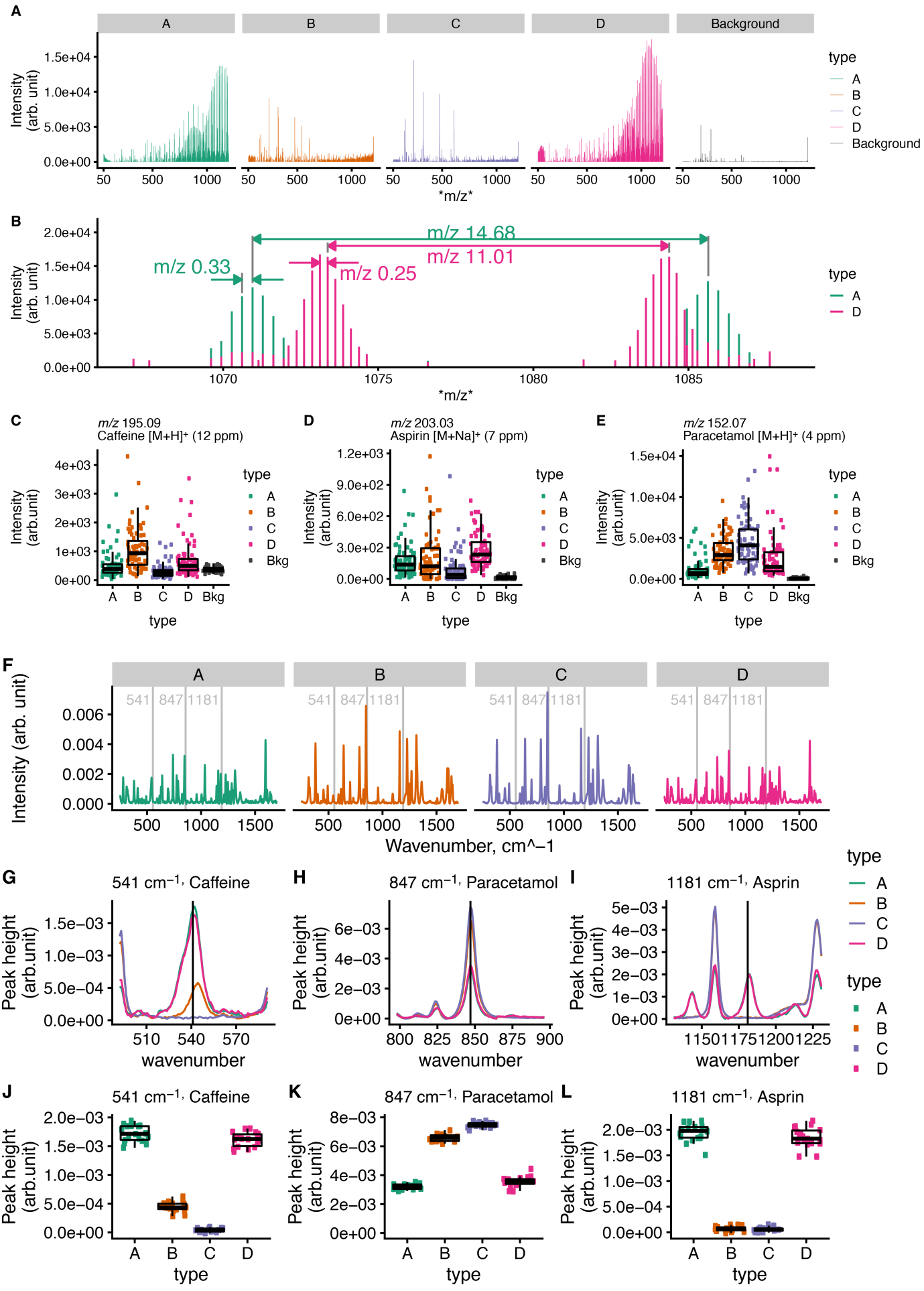
A
B
C
D
Background
Intensity (arb. unit)
m/z
type
m/z 14.68
m/z 11.01
m/z 0.33
m/z 0.25
m/z 195.09
Caffeine [M+H]+ (12 ppm)
m/z 203.03
Aspirin [M+Na]+ (7 ppm)
m/z 152.07
Paracetamol [M+H]+ (4 ppm)
Intensity (arb.unit)
Bkg
Intensity (arb. unit)
Wavenumber, cm^−1
541
847
1181
541 cm−1, Caffeine
847 cm−1, Paracetamol
1181 cm−1, Asprin
Peak height (arb.unit)
wavenumber

Figure 1: Representative spectra from DESI MS and transmission Raman spectroscopy analysis of solid-oral dosage forms (A-C) DESI MS results. (A) Mean spectra for tablet types A, B, C & D and background. (B) Mean spectra for tablet types A (teal) and D (pink) for mass range *m/z* 1065 to *m/z* 1090. (C-E) Boxplots for peaks assigned as caffeine (C), aspirin (D) and paracetamol (E) for each tablet type and background. (F-L) Transmission Raman spectroscopy results. (F) Mean spectra for tablet type A-D from training data. Peaks for the active ingredients shown in G-L are highlighted with a vertical line. (G-L) Mean spectra and (J-L) boxplots for peaks annotated as active ingredients (G & J) Caffeine, 541 cm^{-1}, (H & K) Paracetamol, 847 cm^{-1}, and (I & L) Aspirin, 1181 cm^{-1}). Boxplots show median (horizontal line), interquartile range (box) and range excluding outliers (whiskers). Points show single scan intensities.

Relative spectral similarity

It is notable that for both DESI MS and Transmission Raman data assessment of active ingredients or film coating, or excipients alone cannot robustly separate all tablet types. We can objectively assess the relative overall spectral similarity between and within tablet types by calculating the cosine distance of each spectrum of the same modality from one another. The cosine similarity matrix for DESI MS (Figure 2A) reveals that the highest cosine similarities are between spectra from tablet type D (D vs. D, 0.97 +/- 0.03) and between spectra from tablet type A (A vs. A, 0.96 +/- 0.02). The high value and low standard deviation of cosine similarity within these tablet types indicates the highly reproducible spectra achieved from these samples. Tablet types C and D have a lower mean cosine similarity and higher variance of similarity within their respective tablet types (B vs. B, 0.77 +/- 0.08. C vs. C 0.85 +/- 0.12). Tablet types B and C are relatively alike, with cosine similarity of 0.71. Conversely, tablet types A and D, are relatively dissimilar to one another (0.45 +/- 0.03).

For Transmission Raman spectroscopy, high cosine similarity is observed within tablet types (all greater than 0.97, Figure 2D), However, spectral similarity between tablets A & D is notably higher than for DESI MSI (0.97 +/- 0.01). This is also seen in the high similarity between tablet types B and C (0.99 +/- 0.01), where the only visible spectral difference is for Caffeine, (541, cm^{-1}, present in B, absent in C). As this peak is narrow it does not contribute noticeably to the cosine similarity. All other peaks are visibly identical contributing to the high similarity value.

The visible differences between mean spectra and differences in cosine similarities suggest that this DESI MS may be amenable for the training of classification algorithms to classify unseen data, but that transmission Raman spectroscopy may be more challenging.

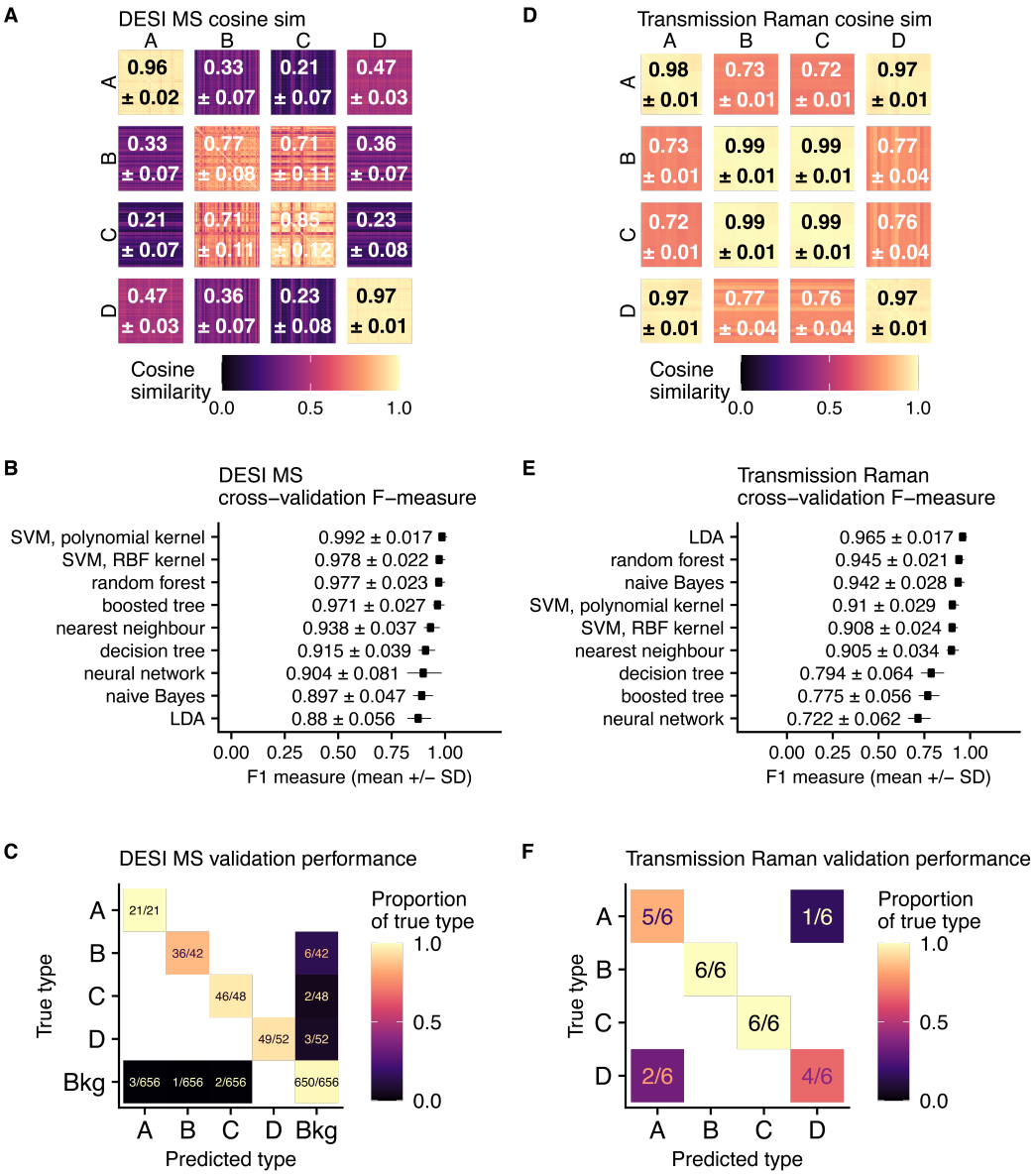

Figure 2: Relative spectral similarity and classification performance comparision (top row) Cosine similarity matrix for each scan of the training dataset from (A) DESI MS or (D) transmission Raman spectroscopy. (middle row) Cross validation F1 measures for (B) DESI MS and (E) transmission Raman spectroscopy. Bars and labels show mean +/- 1 SD for 10-folds with 10 repeats. (bottom row) Confusion matrix for the test set for (C) DESI MS classified by a SFM with polynomial kernel or (F) transmission Raman Spectroscopy classified LDA. Colour and labels show proportion of correct classifications.

Assessment of different classification algorithms

A wide range of classification algorithms are available and have been demonstrated for classification of spectrometric and spectroscopic data. Here we assessed a range of classification algorithms including linear discriminant analysis, tree-based methods, a simple neural network, and support vector machines. A 10-fold cross-validation enables the variance of each algorithm to be assessed. Figure 2B and 2E shows the F1 score of each algorithm for cross-validation for DESI MS and Transmission Raman. The F1 score summarises the precision and recall of the model with equal weights.

For DESI MS All tested algorithms performed well with mean F1 scores above 0.88. 4 algorithms provide a mean F1 score above 0.95, these include the two tree-based methods, random forest and boosted tree. However, the two highest performing algorithms are the support vector machines. The use of a SVM with a polynomial kernel provides a F1 score of 0.992 on the cross-validation training set.

For transmission Raman data a linear discriminant analysis (LDA) model is most robust yielding a F1 measure greater than 0.95. Random forest, naïve Bayes and nearest neighbor models also performed well with F1 measures greater than 0.9 a similar level of performance. Unlike for DESI MS, here a SVM with a polynomial kernel performed less well with a F1 measure of 0.876.

Comparing the best performing models for each analytical technique (DESI MS: SVM-poly, Transmission Raman: LDA), LDA finds a hyperplane that best separates all data points, while an SVM searches for a hyperplane that best separates only those data points in the frontier between classes. In DESI MS many peaks may be less informative, relating to solvent background or molecular fragmentation. A SVM is able to place less priority on these peaks to focus on those forming the class boundaries. Whereas in transmission Raman spectroscopy all peaks are informative of the sample, if not all assignable. A LDA model therefore takes advantage of the full spectrum in discriminating the classes.

These models were therefore taken forward for further exploration and testing. While not a limitation in the relatively small datasets demonstrated here it is worth noting the range of training times required for cross-validation, ranging from ~100 s for Transmission Raman with a boosted tree to ~500 s for DESI MS with naive Bayes (Figure S1). While a search grid for SVM hyperparameters (degree, cost and scale factor) was assessed, the default parameters proved optimal (Supplementary information B, Fig S3). There are no hyperparameters for the LDA.

Test set classification performance

DESI MS

As a support vector machine with a polynomial kernel provided the highest classification performance in the cross-validation of the training set, a model based on this algorithm was trained using the entire training set. This model preserved 410 variables. This model was used to predict tablet type from each scan of an independent test set. The test set was acquired 1 week after the training set was collected, with the instrument in active use and recalibrated in the intervening period. In the test set, tablets not seen in the training set, but from the same type and batch were held under the DESI source for approximately 5 seconds. Scans were individually annotated for known tablet type based on known acquisition order and in reference to the total ion chromatogram (Figure S4A).

The trained SVM model was then used to predict the classification of each scan (Figure S4B). The classification algorithm performed well on the test set with an F1 score of 0.956. A confusion matrix for the ground truth vs. the predicted class (Figure 2C) shows that all misclassifications are between the background and each tablet type. Selected correctly predicted spectra (Figure S4E) show notable characteristics, particularly in the polymer peak envelopes seen for tablet types A and D. Highlighting misclassifications on the total ion chromatogram (Figure S3C), shows that most misclassifications occur at the end of a tablet being presented to the mass spectrometer. Selected spectra from scans incorrectly predicted as background (Figure S4F, scans 184 and 380) show intense peaks at *m/z* 217.11 and 309.10 which are also seen in the mean spectra of the background (Figure S1A), highlighting the importance of accurate ground truth annotation in the assessment of classification models.

Transmission Raman

LDA trained on the whole Transmission Raman training set was used to classify an independently acquired test set of transmission Raman data from 24 tablets from the same batch. Classification performance was strong with an F1 score of 0.965 when using LDA (Figure 2F. Here classification is correct for tablet types B and C and A, which are classified correctly despite their high cosine similarity. However, in two cases tablet D misclassified as type A, and in one case type A misclassified as type D. While type A and D contain the same active ingredients and amounts, but differ in their film coating, as demonstrated by the notably different polymer profiles seen in the DESI MS data. This may contribute to altered peaks shape and or baseline in transmission Raman spectra potentially contributing to misclassification. We also note that the relatively low sample number in the test and train sets set used for cross-validation.

Variable importance

DESI MS

When assessing the classification performance of algorithms on multidimensional data it is useful to assess the variables that the model has placed importance on. Unlike some classification algorithms such as tree-based methods, support vector machines do not inherently provide a measure of variable importance. We therefore calculated a variance-based variable importance using a feature importance ranking measure (FIRM) approach[58,59], based on quantifying the relative flatness of each feature. Most variables are seen to have relatively low importance to the model (importance < 0.05), although several variables are prominent in a spectrum of variable importance (Figure 3A). The 30 variables with the highest importance were selected and boxplots of their single scan intensity per tablet type plotted (Figure S5B). Several of these import variables show high intensity for a single class over all others. These include peaks characteristic of the background (*m/z* 588.42 & *m/z* 309.20), tablet type B (*m/z* 693.18, *m/z* 164.13). Other important variables may be of greater intensity in two types (e/g *m/z* 821.8 in A and D, *m/z* 445.04 and B and C), or be present in all but one tablet type (e.g *m/z* 1013.59). Examination of variable importance, and the relative intensity of important peaks, enables an understandinbg of how the SVM charts a path in multivariate space to accurately classify the different tablet types.

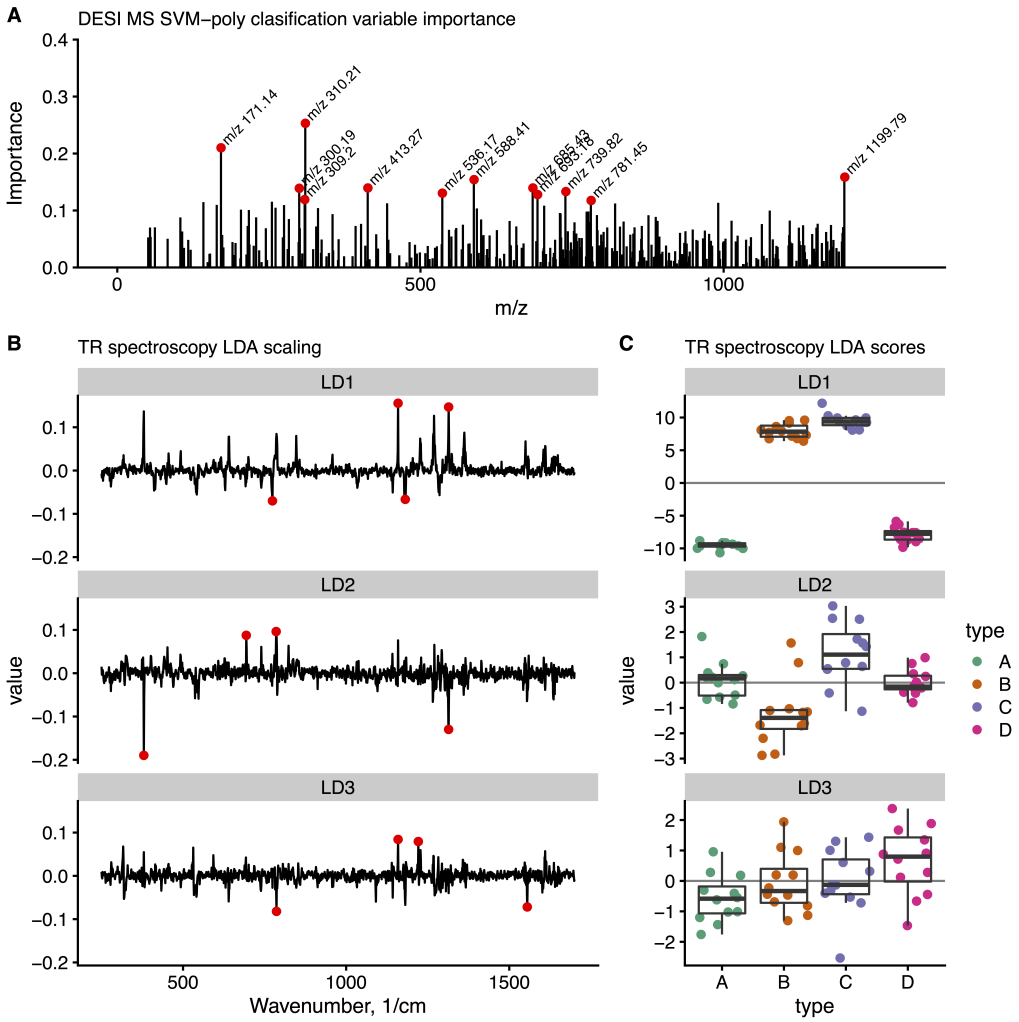

Figure 3: Variable importance for selected classification models (A) FIRM variable importance for SVM-poly classification of DESI MS data. 5 peaks with highest importance are highlighted. (B) Scaling values for each discriminant from LDA of transmission Raman data. Highest absolute loading variables are highlighted with red circles. (B) Boxplots showing LDA scores for the spectra from the transmission Raman training set.

Transmission Raman spectroscopy

The variable importance of the LDA model was assessed by inspection of the LDA scaling values per wavenumber and component (Figure 3B and S5) and the LDA scores of the training set (Figure 3C). LD1 (proportion of trace: 98.8 %) provides wide separation of coated tablet types A and D from uncoated tablets B and C. LD2 (proportion of trace: 0.9 %) broadly separates types C and D from one another. LD3 (proportion of trace: 0.2 %) provides some separation between types A and D, although this separation is less distinct.

LDA scaling spectra highlight key features in the transmission Raman spectra that contribute to the classification. Features with a broad range of Raman shifts contribute to each component. Most of the LDA scaling spectral features are consistent with key active ingredient Raman peaks. For example, the highest scoring peak in LD1 occurs at 1159 cm^{-1} which corresponds to aspirin C-H ring bending[57]. LD1 provides separation of tablet types A and D from B and C, and since A and D contain aspirin and B and C do not, strong weighting on wavenumbers consistent with aspirin Raman peaks would be expected. Several other LDA scaling spectral features are consistent with aspirin peak positions, such as 380, 786 and 1221 cm^{-1}. Other LDA scaling spectral features are consistent with Raman peaks of several ingredients, for example 1555 cm^{-1} (LD3) which

could correspond with spectral features present in all three active ingredients, paracetamol, aspirin, and caffeine, which have peaks centered at 1560, 1557, and 1554 cm^{-1} respectively[55–57].

Repeatability and reproducibility of DESI MS

In the same acquisition as the test set, tablets of type A (Anadin Extra) from a different manufacturing batch number as that analyzed in the training and test sets were profiled. The SVM model correctly classified 51 of 54 scans annotated as the tablet as being of type A (Supplementary figure S5). The three disagreements were between background and tablet type.

Spectral similarity between acquisitions date and tablet batch was assessed by cosine distance (Figure 5). Cosine distance between both date of acquisition and manufacturing batches were highly similar (Cosine distance >0.9) to the cosine distance of spectra within each data or batch. This suggests that sampling variance is higher than the variance between date of acquisition or manufacturing batch. In this work tablets were held manually, using tweezers. Sample variance may be reduced using a sample mounting system or guide to hold tablets in the same geometry respective to the DESI sprayer during acquisition.

Additional analysis (Supplementary information A) comparing DESI MSI cross-validation classification performance with reduced peak numbers (Figure S2A) and simulated lower mass resolving power (S2B) further demonstrate the robustness of classification performance for DESI MS, and its potential application on compact mass spectrometers at-line or in the field.

Conclusions

DESI mass spectrometry and transmission Raman spectroscopy are effective methods to acquire characteristic spectral information from solid oral dosage forms containing information about API and tablet coating components. These tools provide complementary information: transmission Raman is slightly biased in sensitivity towards the bulk of the tablet, whereas DESI MS provides sensitive analysis of the surface coating. A range of machine learning algorithms were found to be capable of classifying tablet type with a strong F-score performance. Of these, support vector machines showed the strongest performance for DESI MS, while LDA was the most effective for transmission Raman data.

DESI-based classification was primarily based on differences in the tablet coatings, whereas transmission Raman was more sensitive to differences in the active pharmaceutical ingredients due to their higher total content. Therefore, it may be advantageous to combine these two complementary analytical methods. Raman spectroscopy's non-destructive nature makes it potentially more suitable for in-line analysis than DESI MS, the destructive nature, even if minimally, of which makes it unsuitable for tablets remaining in the supply chain. Future classfication efforts could also seek to combine their orthogonal analytical advantages with data fusion. Classification performance was retained on datasets of reduced peak number and simulated reduced mass resolving power, indicating the robustness of this approach and its potential applicability to compact mass spectrometers suitable for deployment in counterfeiting, QA/QC, or production line environments.

Conflicts of Interest

There are no conflicts of interest to declare.

Acknowledgements

The authors thank Ariadna Gonzalez, Caterina Minelli and Spencer Thomas (NPL) for helpful discussion and guidance throughout the project. This work was funded by the UK Department of Business, Energy and Industrial Strategy through the projects NMS/IMM19 and NMS/IMM20 of the UK National Measurement System. This work was supported by the Community for Analytical Measurement Science through a 2020 CAMS Fellowship Award to NAB funded by the Analytical Chemistry Trust Fund.

Contributions

AJT, JB and NAB conceived and planned the experiments. AJT and AB acquired DESI MS data. NAB acquired transmission Raman spectroscopy data. AJT performed preprocessing of DESI MS data. DT performed preprocessing of transmission Raman data. AJT performed the classification and variable importance analysis. AJT, DT, NAB and JB interpreted the results. AJT, AD, DT, JB and NAB wrote the manuscript. All authors read, discussed and approved the final manuscript.

References

(1) Laske, S.; Paudel, A.; Scheibelhofer, O.; Sacher, S.; Hoermann, T.; Khinast, J.; Kelly, A.; Rantannen, J.; Korhonen, O.; Stauffer, F.; De Leersnyder, F.; De Beer, T.; Mantanus, J.; Chavez, P.-F.; Thoorens, B.; Ghiotti, P.; Schubert, M.; Tajarobi, P.; Haeffler, G.; Lakio, S.; Fransson, M.; Sparen, A.; Abrahmsen-Alami, S.; Folestad, S.; Funke, A.; Backx, I.; Kavsek, B.; Kjell, F.; Michaelis, M.; Page, T.; Palmer, J.; Schaepman, A.; Sekulic, S.; Hammond, S.; Braun, B.; Colegrove, B. A Review of PAT Strategies in Secondary Solid Oral Dosage Manufacturing of Small Molecules. *Journal of Pharmaceutical Sciences* **2017**, *106* (3), 667–712. https://doi.org/10.1016/j.xphs.2016.11.011.

(2) Knop, K.; Kleinebudde, P. PAT-Tools for Process Control in Pharmaceutical Film Coating Applications. *International Journal of Pharmaceutics* **2013**, *457* (2), 527–536. https://doi.org/10.1016/j.ijpharm.2013.01.062.

(3) Aina, A.; Hargreaves, M. D.; Matousek, P.; Burley, J. C. Transmission Raman Spectroscopy as a Tool for Quantifying Polymorphic Content of Pharmaceutical Formulations. *The Analyst* **2010**, *135* (9), 2328. https://doi.org/10.1039/c0an00352b.

(4) Ricci, C.; Nyadong, L.; Fernandez, F. M.; Newton, P. N.; Kazarian, S. G. Combined Fourier-Transform Infrared Imaging and Desorption Electrospray-Ionization Linear Ion-Trap Mass Spectrometry for Analysis of Counterfeit Antimalarial Tablets. *Analytical and Bioanalytical Chemistry* **2007**, *387* (2), 551–559. https://doi.org/10.1007/s00216-006-0950-z.

(5) Deconinck, E.; Van Campenhout, R.; Aouadi, C.; Canfyn, M.; Bothy, J. L.; Gremeaux, L.; Blanckaert, P.; Courselle, P. Combining Attenuated Total Reflectance- Infrared Spectroscopy and Chemometrics for the Identification and the Dosage Estimation of MDMA Tablets. *Talanta* **2019**, *195*, 142–151. https://doi.org/10.1016/j.talanta.2018.11.027.

(6) Paudel, A.; Raijada, D.; Rantanen, J. Raman Spectroscopy in Pharmaceutical Product Design. *Advanced Drug Delivery Reviews* **2015**, *89*, 3–20. https://doi.org/10.1016/j.addr.2015.04.003.

(7) Takáts, Z.; Wiseman, J. M.; Gologan, B.; Cooks, R. G. Mass Spectrometry Sampling Under Ambient Conditions with Desorption Electrospray Ionization. *Science* **2004**, *306* (5695), 471–473. https://doi.org/10.1126/science.1104404.

(8) Takáts, Z.; Cotte-Rodriguez, I.; Talaty, N.; Chen, H.; Cooks, R. G. Direct, Trace Level Detection of Explosives on Ambient Surfaces by Desorption Electrospray Ionization Mass Spectrometry. *Chem. Commun.* **2005**, No. 15, 1950–1952. https://doi.org/10.1039/b418697d.

(9) Bailey, M. J.; Bradshaw, R.; Francese, S.; Salter, T. L.; Costa, C.; Ismail, M.; P. Webb, R.; Bosman, I.; Wolff, K.; de Puit, M. Rapid Detection of Cocaine, Benzoylecgonine and Methylecgonine in Fingerprints Using Surface Mass Spectrometry. *The Analyst* **2015**, *140* (18), 6254–6259. https://doi.org/10.1039/c5an00112a.

(10) Li, B.; Bjarnholt, N.; Hansen, S. H.; Janfelt, C. Characterization of Barley Leaf Tissue Using Direct and Indirect Desorption Electrospray Ionization Imaging Mass Spectrometry: DESI Analysis and Imaging of Barley Leaves. *Journal of Mass Spectrometry* **2011**, *46* (12), 1241–1246. https://doi.org/10.1002/jms.2010.

(11) Dexter, A.; Steven, R. T.; Patel, A.; Dailey, L. A.; Taylor, A. J.; Ball, D.; Klapwijk, J.; Forbes, B.; Page, C. P.; Bunch, J. Imaging Drugs, Metabolites and Biomarkers in Rodent Lung: A DESI MS Strategy for the Evaluation of Drug-Induced Lipidosis. *Analytical and Bioanalytical Chemistry* **2019**, *411* (30), 8023–8032. https://doi.org/10.1007/s00216-019-02151-z.

(12) Eberlin, L. S.; Ferreira, C. R.; Dill, A. L.; Ifa, D. R.; Cooks, R. G. Desorption Electrospray Ionization Mass Spectrometry for Lipid Characterization and Biological Tissue Imaging. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* **2011**, *1811* (11), 946–960. https://doi.org/10.1016/j.bbalip.2011.05.006.

(13) Chen, H.; Talaty, N. N.; Takáts, Z.; Cooks, R. G. Desorption Electrospray Ionization Mass Spectrometry for High-Throughput Analysis of Pharmaceutical Samples in the Ambient Environment. *Analytical Chemistry* **2005**, *77* (21), 6915–6927. https://doi.org/10.1021/ac050989d.

(14) Leuthold, L. A.; Mandscheff, J.-F.; Fathi, M.; Giroud, C.; Augsburger, M.; Varesio, E.; Hopfgartner, G. Desorption Electrospray Ionization Mass Spectrometry: Direct Toxicological Screening and Analysis of Illicit Ecstasy Tablets. *Rapid Communications in Mass Spectrometry* **2006**, *20* (2), 103–110. https://doi.org/10.1002/rcm.2280.

(15) Nyadong, L.; Late, S.; Green, M. D.; Banga, A.; Fernández, F. M. Direct Quantitation of Active Ingredients in Solid Artesunate Antimalarials by Noncovalent Complex Forming Reactive Desorption Electrospray Ionization Mass Spectrometry. *Journal of the American Society for Mass Spectrometry* **2008**, *19* (3), 380–388. https://doi.org/10.1016/j.jasms.2007.11.016.

(16) Nyadong, L.; Hohenstein, E. G.; Johnson, K.; Sherrill, C. D.; Green, M. D.; Fernández, F. M. Desorption Electrospray Ionization Reactions Between Host Crown Ethers and the Influenza Neuraminidase Inhibitor Oseltamivir for the Rapid Screening of Tamiflu®. *The Analyst* **2008**, *133* (11), 1513. https://doi.org/10.1039/b809471c.

(17) Mulligan, C. C.; Talaty, N.; Cooks, R. G. Desorption Electrospray Ionization with a Portable Mass Spectrometer: In Situ Analysis of Ambient Surfaces. *Chemical Communications* **2006**, No. 16, 1709. https://doi.org/10.1039/b517357d.

(18) Liu, C.; Qi, K.; Yao, L.; Xiong, Y.; Zhang, X.; Zang, J.; Tian, C.; Xu, M.; Yang, J.; Lin, Z.; Lv, Y.; Xiong, W.; Pan, Y. Imaging of Polar and Nonpolar Species Using Compact Desorption Electrospray Ionization/Postphotoionization Mass Spectrometry. *Analytical Chemistry* **2019**, *91* (10), 6616–6623. https://doi.org/10.1021/acs.analchem.9b00520.

(19) Gromski, P. S.; Xu, Y.; Correa, E.; Ellis, D. I.; Turner, M. L.; Goodacre, R. A Comparative Investigation of Modern Feature Selection and Classification Approaches for the Analysis of Mass Spectrometry Data. *Analytica Chimica Acta* **2014**, *829*, 1–8. https://doi.org/10.1016/j.aca.2014.03.039.

(20) Maione, C.; Souza, V. C. de O.; Togni, L. R.; da Costa, J. L.; Campiglia, A. D.; Barbosa, F.; Barbosa, R. M. Establishing Chemical Profiling for Ecstasy Tablets Based on Trace Element Levels and Support Vector Machine. *Neural Computing and Applications* **2018**, *30* (3), 947–955. https://doi.org/10.1007/s00521-016-2736-3.

(21) Thomas, S. A.; Jin, Y.; Bunch, J.; Gilmore, I. S. Enhancing Classification of Mass Spectrometry Imaging Data with Deep Neural Networks. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*; IEEE: Honolulu, HI, 2017; pp 1–8. https://doi.org/10.1109/ssci.2017.8285223.

(22) Tang, J.; Wang, Y.; Luo, Y.; Fu, J.; Zhang, Y.; Li, Y.; Xiao, Z.; Lou, Y.; Qiu, Y.; Zhu, F. Computational Advances of Tumor Marker Selection and Sample Classification in Cancer Proteomics. *Computational and Structural Biotechnology Journal* **2020**, *18*, 2012–2025. https://doi.org/10.1016/j.csbj.2020.07.009.

(23) Spicer, R.; Salek, R. M.; Moreno, P.; Cañueto, D.; Steinbeck, C. Navigating Freely-Available Software Tools for Metabolomics Analysis. *Metabolomics* **2017**, *13* (9), 106. https://doi.org/10.1007/s11306-017-1242-7.

(24) Phelps, D. L.; Balog, J.; Gildea, L. F.; Bodai, Z.; Savage, A.; El-Bahrawy, M. A.; Speller, A. V.; Rosini, F.; Kudo, H.; McKenzie, J. S.; Brown, R.; Takáts, Z.; Ghaem-Maghami, S. The Surgical Intelligent Knife Distinguishes Normal, Borderline and Malignant Gynaecological Tissues Using Rapid Evaporative Ionisation Mass Spectrometry (REIMS). *British Journal of Cancer* **2018**, *118* (10), 1349–1358. https://doi.org/10.1038/s41416-018-0048-3.

(25) Balog, J.; Perenyi, D.; Guallar-Hoyas, C.; Egri, A.; Pringle, S. D.; Stead, S.; Chevallier, O. P.; Elliott, C. T.; Takats, Z. Identification of the Species of Origin for Meat Products by Rapid Evaporative Ionization Mass Spectrometry. *Journal of Agricultural and Food Chemistry* **2016**, *64* (23), 4793–4800. https://doi.org/10.1021/acs.jafc.6b01041.

(26) Golf, O.; Strittmatter, N.; Karancsi, T.; Pringle, S. D.; Speller, A. V. M.; Mroz, A.; Kinross, J. M.; Abbassi-Ghadi, N.; Jones, E. A.; Takats, Z. Rapid Evaporative Ionization Mass Spectrometry Imaging Platform for Direct Mapping from Bulk Tissue and Bacterial Growth Media. *Analytical Chemistry* **2015**, *87* (5), 2527–2534. https://doi.org/10.1021/ac5046752.

(27) St John, E. R.; Balog, J.; McKenzie, J. S.; Rossi, M.; Covington, A.; Muirhead, L.; Bodai, Z.; Rosini, F.; Speller, A. V. M.; Shousha, S.; Ramakrishnan, R.; Darzi, A.; Takats, Z.; Leff, D. R. Rapid Evaporative Ionisation Mass Spectrometry of Electrosurgical Vapours for the Identification of Breast Pathology: Towards an Intelligent Knife for Breast Cancer Surgery. *Breast Cancer Research* **2017**, *19* (1), 59. https://doi.org/10.1186/s13058-017-0845-2.

(28) Shin, K.; Chung, H. Wide Area Coverage Raman Spectroscopy for Reliable Quantitative Analysis and Its Applications. *The Analyst* **2013**, *138* (12), 3335. https://doi.org/10.1039/c3an36843b.

(29) Matousek, P.; Parker, A. W. Bulk Raman Analysis of Pharmaceutical Tablets. *Applied Spectroscopy* **2006**, *60* (12), 1353–1357. https://doi.org/10.1366/000370206779321463.

(30) Buckley, K.; Matousek, P. Recent Advances in the Application of Transmission Raman Spectroscopy to Pharmaceutical Analysis. *Journal of Pharmaceutical and Biomedical Analysis* **2011**, *55* (4), 645–652. https://doi.org/10.1016/j.jpba.2010.10.029.

(31) Johansson, J.; Sparén, A.; Svensson, O.; Folestad, S.; Claybourn, M. Quantitative Transmission Raman Spectroscopy of Pharmaceutical Tablets and Capsules. *Applied Spectroscopy* **2007**, *61* (11), 1211–1218. https://doi.org/10.1366/000370207782597085.

(32) Matousek, P.; Everall, N.; Littlejohn, D.; Nordon, A.; Bloomfield, M. Dependence of Signal on Depth in Transmission Raman Spectroscopy. *Applied Spectroscopy* **2011**, *65* (7), 724–733. https://doi.org/10.1366/11-06259.

(33) Peris-Díaz, M. D.; Krężel, A. A Guide to Good Practice in Chemometric Methods for Vibrational Spectroscopy, Electrochemistry, and Hyphenated Mass Spectrometry. *TrAC Trends in Analytical Chemistry* **2021**, *135*, 116157. https://doi.org/10.1016/j.trac.2020.116157.

(34) Stone, N.; Kendall, C.; Shepherd, N.; Crow, P.; Barr, H. Near-Infrared Raman Spectroscopy for the Classification of Epithelial Pre-Cancers and Cancers. *Journal of Raman Spectroscopy* **2002**, *33* (7), 564–573. https://doi.org/10.1002/jrs.882.

(35) Krafft, C.; Steiner, G.; Beleites, C.; Salzer, R. Disease Recognition by Infrared and Raman Spectroscopy. *Journal of Biophotonics* **2009**, *2* (1-2), 13–28. https://doi.org/10.1002/jbio.200810024.

(36) Liu, W.; Sun, Z.; Chen, J.; Jing, C. Raman Spectroscopy in Colorectal Cancer Diagnostics: Comparison of PCA-LDA and PLS-DA Models. *Journal of Spectroscopy* **2016**, *2016*, 1–6. https://doi.org/10.1155/2016/1603609.

(37) Gaus, K.; Rösch, P.; Petry, R.; Peschke, K.-D.; Ronneberger, O.; Burkhardt, H.; Baumann, K.; Popp, J. Classification of Lactic Acid Bacteria with UV-Resonance Raman Spectroscopy. *Biopolymers* **2006**, *82* (4), 286–290. https://doi.org/10.1002/bip.20448.

(38) Guo, S.; Heinke, R.; Stöckel, S.; Rösch, P.; Popp, J.; Bocklitz, T. Model Transfer for Raman-Spectroscopy-Based Bacterial Classification. *Journal of Raman Spectroscopy* **2018**, *49* (4), 627–637. https://doi.org/10.1002/jrs.5343.

(39) Ryder, A. G. Classification of Narcotics in Solid Mixtures Using Principal Component Analysis and Raman Spectroscopy. *Journal of Forensic Sciences* **2002**, *47* (2), 15244J. https://doi.org/10.1520/jfs15244j.

(40) Roggo, Y.; Degardin, K.; Margot, P. Identification of Pharmaceutical Tablets by Raman Spectroscopy and Chemometrics. *Talanta* **2010**, *81* (3), 988–995. https://doi.org/10.1016/j.talanta.2010.01.046.

(41) Romero-Torres, S.; Pérez-Ramos, J. D.; Morris, K. R.; Grant, E. R. Raman Spectroscopy for Tablet Coating Thickness Quantification and Coating Characterization in the Presence of Strong Fluorescent Interference. *Journal of Pharmaceutical and Biomedical Analysis* **2006**, *41* (3), 811–819. https://doi.org/10.1016/j.jpba.2006.01.033.

(42) de Veij, M.; Vandenabeele, P.; Hall, K. A.; Fernandez, F. M.; Green, M. D.; White, N. J.; Dondorp, A. M.; Newton, P. N.; Moens, L. Fast Detection and Identification of Counterfeit Antimalarial Tablets by Raman Spectroscopy. *Journal of Raman Spectroscopy* **2007**, *38* (2), 181–187. https://doi.org/10.1002/jrs.1621.

(43) Zheng, X.; Lv, G.; Zhang, Y.; Lv, X.; Gao, Z.; Tang, J.; Mo, J. Rapid and Non-Invasive Screening of High Renin Hypertension Using Raman Spectroscopy and Different Classification Algorithms. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **2019**, *215*, 244–248. https://doi.org/10.1016/j.saa.2019.02.063.

(44) Brereton, R. G.; Lloyd, G. R. Partial Least Squares Discriminant Analysis: Taking the Magic Away: PLS-DA: Taking the Magic Away. *Journal of Chemometrics* **2014**, *28* (4), 213–225. https://doi.org/10.1002/cem.2609.

(45) Gao, Q.; Liu, Y.; Li, H.; Chen, H.; Chai, Y.; Lu, F. Comparison of Several Chemometric Methods of Libraries and Classifiers for the Analysis of Expired Drugs Based on Raman Spectra. *Journal of Pharmaceutical and Biomedical Analysis* **2014**, *94*, 58–64. https://doi.org/10.1016/j.jpba.2014.01.027.

(46) Fransson, M.; Johansson, J.; Sparén, A.; Svensson, O. Comparison of Multivariate Methods for Quantitative Determination with Transmission Raman Spectroscopy in Pharmaceutical Formulations. *Journal of Chemometrics* **2010**, *24* (11-12), 674–680. https://doi.org/10.1002/cem.1330.

(47) Wickham, H.; Averick, M.; Bryan, J.; Chang, W.; McGowan, L.; François, R.; Grolemund, G.; Hayes, A.; Henry, L.; Hester, J.; Kuhn, M.; Pedersen, T.; Miller, E.; Bache, S.; Müller, K.; Ooms, J.; Robinson, D.; Seidel, D.; Spinu, V.; Takahashi, K.; Vaughan, D.; Wilke, C.; Woo, K.; Yutani, H. Welcome to the Tidyverse. *Journal of Open Source Software* **2019**, *4* (43), 1686. https://doi.org/10.21105/joss.01686.

(48) *Tidymodels*. https://www.tidymodels.org/ (accessed 2022-05-25).

(49) Holman, J. D.; Tabb, D. L.; Mallick, P. Employing ProteoWizard to Convert Raw Mass Spectrometry Data. *Current Protocols in Bioinformatics* **2014**, *46* (1). https://doi.org/10.1002/0471250953.bi1324s46.

(50) Chambers, M. C.; Maclean, B.; Burke, R.; Amodei, D.; Ruderman, D. L.; Neumann, S.; Gatto, L.; Fischer, B.; Pratt, B.; Egertson, J.; Hoff, K.; Kessner, D.; Tasman, N.; Shulman, N.; Frewen, B.; Baker, T. A.; Brusniak, M.-Y.; Paulse, C.; Creasy, D.; Flashner, L.; Kani, K.; Moulding, C.; Seymour, S. L.; Nuwaysir, L. M.; Lefebvre, B.; Kuhlmann, F.; Roark, J.; Rainer, P.; Detlev, S.; Hemenway, T.; Huhmer, A.; Langridge, J.; Connolly, B.; Chadick, T.; Holly, K.; Eckels, J.; Deutsch, E. W.; Moritz, R. L.; Katz, J. E.; Agus, D. B.; MacCoss, M.; Tabb, D. L.; Mallick, P. A Cross-Platform Toolkit for Mass Spectrometry and Proteomics. *Nature Biotechnology* **2012**, *30* (10), 918–920. https://doi.org/10.1038/nbt.2377.

(51) Fischer, B.; Neumann, S.; Gatto, L.; Kou, Q.; Rainer, J. *mzR: Parser for netCDF, mzXML, mzData and mzML and mzIdentML Files (Mass Spectrometry Data)*; Bioconductor version: Release (3.15), 2022.

(52) Borchers, H. W. *Pracma: Practical Numerical Math Functions*; 2022.

(53) *Correct baseline of signal with peaks - MATLAB msbackadj*. https://www.mathworks.com/help/bioinfo/ref/msbackadj.html (accessed 2022-05-25).

(54) Dexter, A.; Race, A. M.; Styles, I. B.; Bunch, J. Testing for Multivariate Normality in Mass Spectrometry Imaging Data: A Robust Statistical Approach for Clustering Evaluation and the Generation of Synthetic Mass Spectrometry Imaging Data Sets. *Analytical Chemistry* **2016**, *88* (22), 10893–10899. https://doi.org/10.1021/acs.analchem.6b02139.

(55) Baranska, M.; Proniewicz, L. M. Raman Mapping of Caffeine Alkaloid. *Vibrational Spectroscopy* **2008**, *48* (1), 153–157. https://doi.org/10.1016/j.vibspec.2007.12.016.

(56) Shende, C.; Smith, W.; Brouillette, C.; Farquharson, S. Drug Stability Analysis by Raman Spectroscopy. *Pharmaceutics* **2014**, *6* (4), 651–662. https://doi.org/10.3390/pharmaceutics6040651.

(57) Crowell, E. L.; Dreger, Z. A.; Gupta, Y. M. High-Pressure Polymorphism of Acetylsalicylic Acid (Aspirin): Raman Spectroscopy. *Journal of Molecular Structure* **2015**, *1082*, 29–37. https://doi.org/10.1016/j.molstruc.2014.10.079.

(58) Greenwell, B. M.; Boehmke, B. C.; McCarthy, A. J. *A Simple and Effective Model-Based Variable Importance Measure*; 1805.04755; arXiv, 2018.

(59) Scholbeck, C. A.; Molnar, C.; Heumann, C.; Bischl, B.; Casalicchio, G. Sampling, Intervention, Prediction, Aggregation: A Generalized Framework for Model-Agnostic Interpretations. In *Machine Learning and Knowledge Discovery in Databases*; Cellier, P., Driessens, K., Eds.; Springer International Publishing: Cham, 2020; Vol. 1167, pp 205–216. https://doi.org/10.1007/978-3-030-43823-4_18.

Supplementary information

Supplementary Notes

Note A. Reducing peaks and down binning data for DESI MS

Given the notable spectra differences observed for DESI MS, it is unsurprising that strong classification performance is achieved. In this model comparison we included all detected peaks in the algorithm (although some will be removed automatically as they correlate strongly with other variables. To assess the number of features required to maintain good classification performance we performed 10-fold cross validation using an SVM with a polynomial kernel and default parameters (degree: 1, cost: 0.1, scale factor: 1) on datasets in which fewer peaks were selected (from top 1200 to top 5 peaks). F1 metric was seen to be maintained above 0.97 (Figure S2B), but starts to drop off below 100 peaks, with 5 peaks providing a mean F1 score of 0.61 with a greatly increased variance.

Potential applications of tablet classification may demand the mass spectrometer to be located outside of an analytical chemistry laboratory, either at-line in a manufacturing environment, in an on-site QA/QC lab, on in the field. In these applications the use of a Q-ToF mass spectrometer may not be suitable due to size or cost. Compact, field-deployable mass spectrometers such as ion traps or single quadrupoles may offer reduced mass resolving power. To simulate reduced instrument performance, data were down binned at increasing bin width. No reduction in classification performance is observed when binning at 1 Dalton, simulating single quadrupole mass resolving power (Figure S2B). Classification performance is maintained even at 2 Dalton binning, indicating the applicability of this approach to low mass resolving power data from compact or portable mass spectrometers.

Note B. Model tuning

The hyperparameters of the SVM were tuned on the validation split of the training set over a regular grid with five levels of degree (1 to 5), cost (0.001 to 1) and scale factor (0.001 to 1). All variables and 0.01 Da binning. The F1 measure of each combination is shown in Figure S3. First-degree (linear) and third-degree polynomials with cost and scale factors above 0.001 show high F1 score (all = 1.0) for the validation split. As the default SVM parameters in the kernlab implementation are within this range (degree: 1, cost: 1, scale factor: 1) they were selected to fit a final model using the whole training set.

For Transmission Raman data the best performing classification approach, a linear discriminant analysis, as implemented by the MASS package has no parameters to optimize beyond the prior which was here set to the equal type probability. This would influence the suitability of the LDA for applications in cases such as defect or counterfeit detection where class probability is expected to be unbalanced with an unknown prior.

Supplementary Figures

Supplementary Figures

Table 2: Functions and engines for models evaluated

Model	Function	Engine
Logistic regression	logistic_reg	keras
naive Bayes	naive_Bayes	klaR
Nearest neighbour	nearest_neighbour	kknn
Support vector machinekernel	svm_poly	kernlab
Support vector machinebasis function kernel	svm_rbf	kernlab
Decision tree	decision_tree	rpart
Boosted tree	boost_tree	xgboost
Random forest	rand_forest	ranger
Neural network	mlp	keras

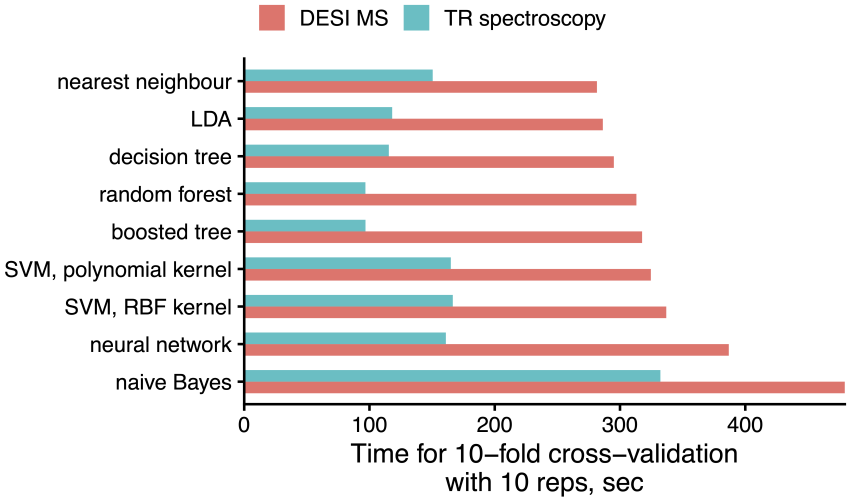

Figure S1: Timings for 10-fold cross validation with 10 replicates for selected classification algorithms for DESI MS (red) and transmission Raman spectroscopy (teal) training data.

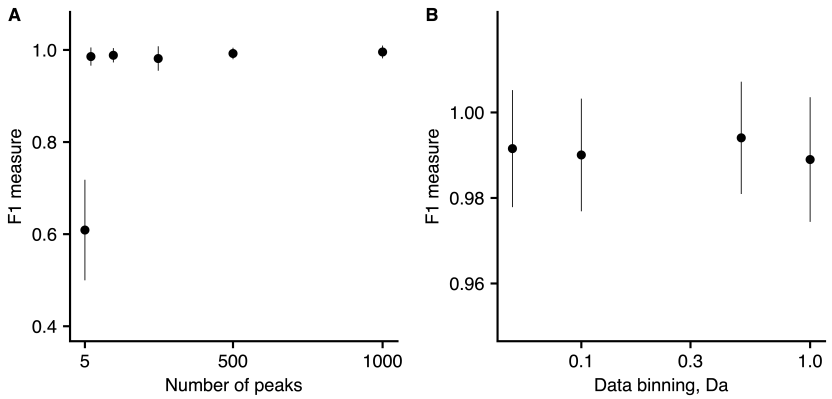

Figure S2: Exploration of reduced peak number (A) and down-binning (B) for DESI MS classification.

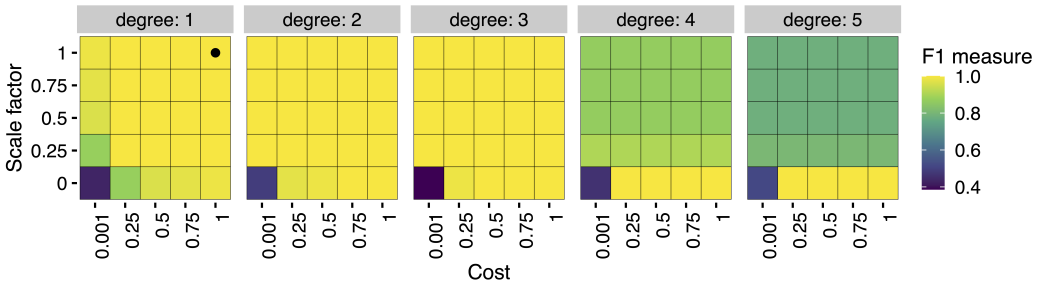

Figure S3: Model tuning grid for SVM-polynomial kernel for DESI MS data.

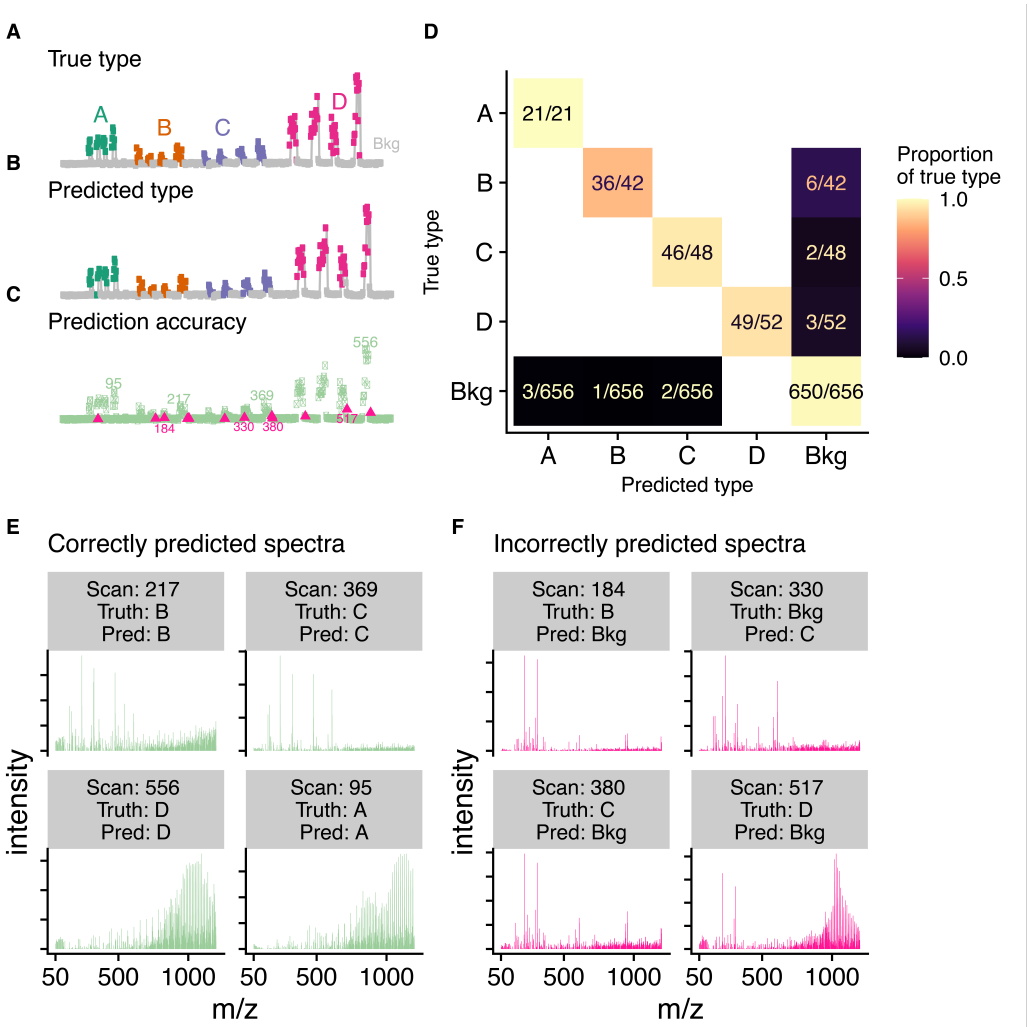

Figure S4: Classification of an independent test set of tablets by DESI MS and an SVM

(A) Total ion chromatogram (TIC) plot showing the ground truth labelling per scan for four sampling events per tablet type. Labelling was performed manually in reference to known sampling order and TIC. (B) TIC plot showing tablet type predicted by the SVM for each scan. (C) TIC plot highlighting prediction accuracy per scan (correct: green circles, incorrect: pink triangles). (D) Confusion matrix for each scan of the test set, classified by the SVM. Colour and labels show proportion of correct classifications. (E-F) Spectra for selected scans (labelled in C) that are correctly (E). or incorrectly classified (F).

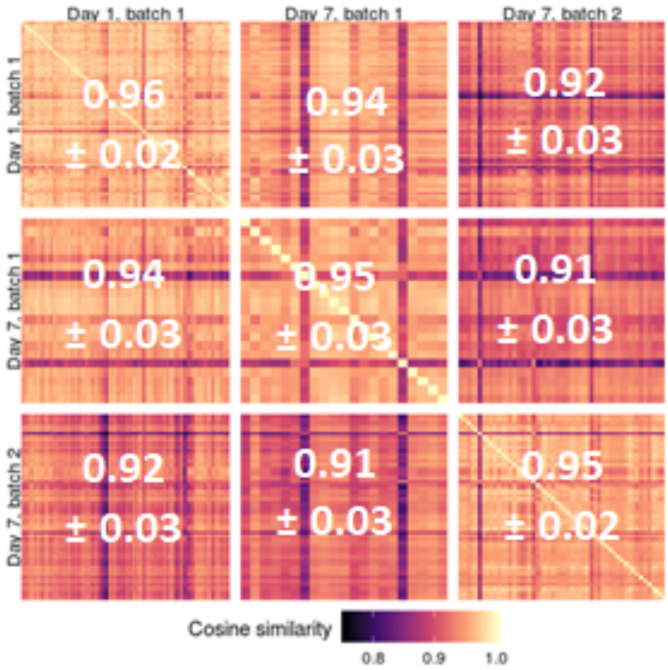

Figure S5: Cosine similarity matrix for DESI MS sampling of tablet type A from two batches on two days of analysis.

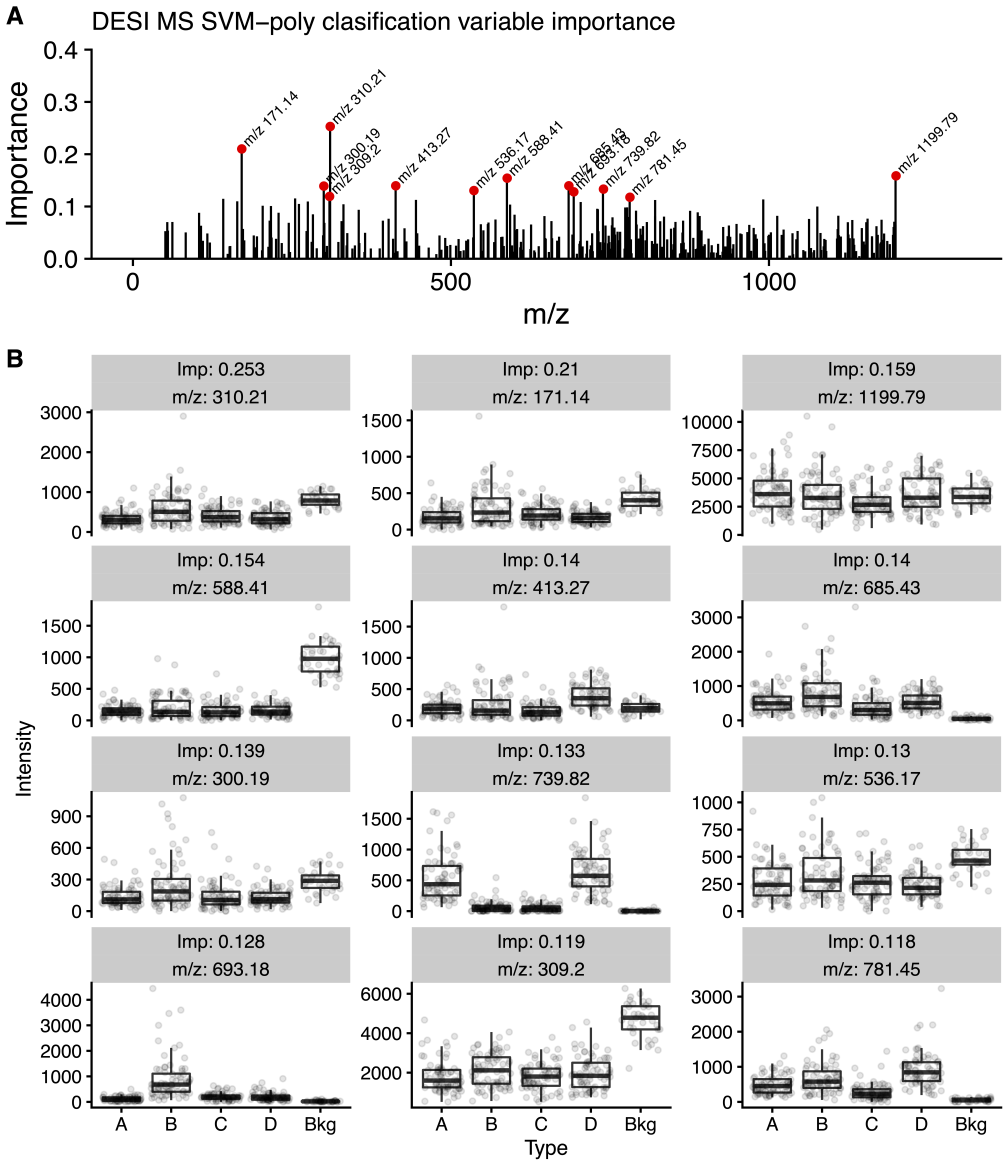

Figure S6: Variable importance plot for SVM with polynomial kernel trained on DESI MS data. Peaks with the 30 highest variable importance are highlighted by a black dot). (b) Boxplots for peaks with the 30 highest variable importance values (Line: median, box: Q2 & Q4, whiskers: range excluding outliers, points: single scan intensities).

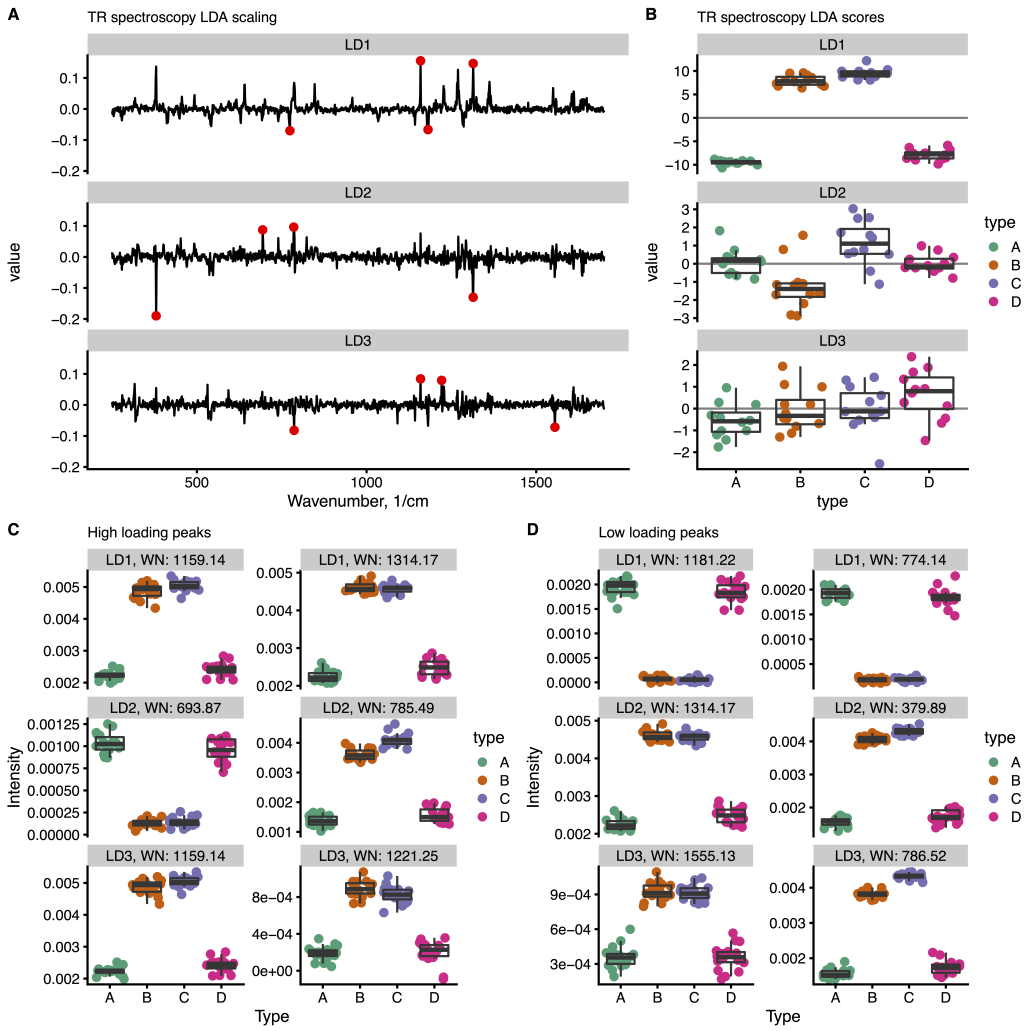

Figure S7: (a) Spectra of LDA scaling values for each discriminant. Selected variables are highlighted with red circles. Boxplots showing LDA scores for the spectra contained in the training set. (c & d) Boxplots of transmission Raman intensity for the wavenumber bins with the two highest (c) and lowest (d) scaling values per discriminant. Boxplots show mean (bar), first and third quartiles (bar) and range between the lowest and highest values no further than 1.5 times the IQR from the box. Points show individual observations.