Machine learning of Raman spectra predicts drug release from polysaccharide coatings for targeted colonic delivery

Youssef Abdalla, Laura Mccoubrey, Fabiana Ferraro, Lisa Sonnleitner, Yannick Guinet, Florence Siepmann, Alain Hedoux, Juergen Siepmann, Abdul Basit, Mine Orlu, et al.

Journal of Controlled Release 374 (2024) 103–111

Contents lists available at ScienceDirect

Journal of Controlled Release

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

Machine learning of Raman spectra predicts drug release from polysaccharide coatings for targeted colonic delivery

Youssef Abdalla [a], Laura E. McCoubrey [a], Fabiana Ferraro [b], Lisa Maria Sonnleitner [b], Yannick Guinet [c], Florence Siepmann [b], Alain Hédoux [c], Juergen Siepmann [b], Abdul W. Basit [a,*], Mine Orlu [a,*], David Shorthouse [a,*]

[a] UCL School of Pharmacy, University College London, 29-39 Brunswick Square, London WC1N 1AX, UK
[b] Univ. Lille, Inserm, CHU Lille, U1008, F-59000 Lille, France
[c] Univ. Lille, CNRS, INRAE, Centrale Lille, UMR 8207 - UMET - Unité Matériaux et Transformations, F-59000 Lille, France

ARTICLE INFO

Keywords:
Machine learning
Oral colonic drug delivery film coatings
Mesalazine
Mesalamine
Gut microbiome
Raman spectroscopy

ABSTRACT

Colonic drug delivery offers numerous pharmaceutical opportunities, including direct access to local therapeutic targets and drug bioavailability benefits arising from the colonic epithelium's reduced abundance of cytochrome P450 enzymes and particular efflux transporters. Current workflows for developing colonic drug delivery systems involve time-consuming, low throughput *in vitro* and *in vivo* screening methods, which hinder the identification of suitable enabling materials. Polysaccharides are useful materials for colonic targeting, as they can be utilised as dosage form coatings that are selectively digested by the colonic microbiota. However, polysaccharides are a heterogeneous family of molecules with varying suitability for this purpose. To address the need for high-throughput material selection tools for colonic drug delivery, we leveraged machine learning (ML) and publicly accessible experimental data to predict the release of the drug 5-aminosalicylic acid from polysaccharide-based coatings in simulated human, rat, and dog colonic environments. For the first time, Raman spectra alone were used to characterise polysaccharides for input as ML features. Models were validated on 8 unseen drug release profiles from new polysaccharide coatings, demonstrating the generalisability and reliability of the method. Further, model analysis facilitated an understanding of the chemical features that influence a polysaccharide's suitability for colonic drug delivery. This work represents a major step in employing spectral data for forecasting drug release from pharmaceutical formulations and marks a significant advancement in the field of colonic drug delivery. It offers a powerful tool for the efficient, sustainable, and successful development and pre-ranking of colon-targeted formulation coatings, paving the way for future more effective and targeted drug delivery strategies.

1. Introduction

Currently, most medicines on the market are formulated as oral solid dosage forms that release drugs within the stomach and small intestine. Patients and healthcare providers widely prefer oral administration due to its ease, lack of invasiveness, and lower cost than other administration routes [1]. Though immediate-release dosage forms are typically the default choice during formulation, some drugs may benefit from targeted delivery to the colon. Although the colon has a lower surface area for absorption than the small intestine, and a thicker epithelial mucus layer, it may still provide pharmacokinetic, safety, and therapeutic benefits [2–4]. This is partly due to its reduced expression of the drug-metabolising enzyme cytochrome P450 3A4 (CYP3A4) [5] and the efflux transporter permeability-glycoprotein (P-gp) compared to the small intestine [6]. Colonic drug delivery is also attractive for treating local diseases, such as inflammatory bowel disease (IBD) and colorectal cancer [7–9]. Similarly, targeting the colon can provide access to local therapeutic targets related to systemic diseases, such as the colonic microbiome and nutrient-sensing receptors [10,11]. Delivering drugs to their target site can improve therapeutic efficacy and reduce response variability, dose requirements, and adverse effects [12].

The design and development of the targeted dosage form is a critical aspect of colonic drug delivery. Solid oral dosage forms can be tuned for colonic release by coating them with materials that are sensitive to

* Corresponding authors.
E-mail addresses: a.basit@ucl.ac.uk (A.W. Basit), m.orlu@ucl.ac.uk (M. Orlu), d.shorthouse@ucl.ac.uk (D. Shorthouse).

https://doi.org/10.1016/j.jconrel.2024.08.010
Received 19 February 2024; Received in revised form 4 August 2024; Accepted 7 August 2024
Available online 12 August 2024

gastrointestinal (GI) physiology, namely by utilising pH-dependent, time-dependent, and/or microbiota-dependent trigger mechanisms [2,13]. These mechanisms may also be combined to enhance the reliability of formulations' colonic drug release [14]. Identifying materials that reliably enable colonic drug release can be challenging and time-consuming, frequently relying on low throughput in vitro and in vivo screening programs [15,16]. As such, there is a need for tools that can increase the throughput of material selection for colonic drug delivery.

Polysaccharides hold great promise as materials for enabling drug delivery to the colon. With the aid of a structuring agent to prevent swelling, such as ethyl cellulose or an enteric polymer, polysaccharide-based coatings should be stable during transit through the proximal GI tract and only allow drug release when they are digested by the microbiota in the colon [17,18]. Currently, only one commercial formulation, OPTICORE®, utilises polysaccharides for colonic drug delivery [14]. As part of its dual trigger mechanism, OPTICORE® employs high amylose starch as the polysaccharide that is digested by patients' colonic microbiota, facilitating 5-ASA delivery for the treatment of ulcerative colitis. This technology demonstrates the potential of polysaccharides and encourages the development of new polysaccharide-based formulations for colonic drug targeting. However, a major challenge with using polysaccharides for this purpose is that they are a highly chemically diverse family of molecules with varying suitability for inclusion into dosage forms. This means the digestibility and, therefore, drug release is extremely hard to predict. Hence, a major bottleneck in designing these materials is the time and effort required to study and validate dosage forms.

Machine learning (ML) methods hold great promise for many areas of pharmaceutical research. ML models can be trained to predict an output based on relationships between input features and have been shown to accurately predict material properties in pharmaceutics before [19–22], including controlled release drug delivery systems, such as long acting injectables [23,24]. An advantage of ML methods over traditional statistical tools is that they do not assume linearity and, therefore, typically have high predictive power for complex and multi-dimensional data at the risk of a loss of interpretability compared to traditional statistical methods.

Here, we show how Raman spectra can be used to predict drug release from polysaccharide-based coatings. We trained an ML model on publicly available data composed of novel polysaccharide-based coatings that were investigated for colonic drug delivery for the first time [25]. Drug release for untested polysaccharides coatings was then predicted and validated by generating new experimental data composed of 8 unseen release profiles. Interpreting these models demonstrated that Raman peaks identifying glycosidic linkages were essential for predicting drug release. All models and the code for generating them have been made freely available in the GitHub repository for this article (https://github.com/y-babdalla/coating_release).

2. Materials and methods

2.1. Materials

The data used to train the ML models in this study have been published by Ferraro et al. [25]. In this work, pellets were loaded with 5-aminosalicylic acid (5-ASA) and coated with polysaccharide-based materials. The full manufacturing method used for coating production is presented in the original publication by Ferraro et al. [25] (Section 2.5). Briefly, aqueous ethylcellulose dispersion was plasticized with dibutyl sebacate (DBS). A solution/dispersion of a second polysaccharide aiming at colon targeting was added. The blend was stirred for 1 h and subsequently used to coat 5-ASA loaded pellets in a fluidised bed coater.

It should be appreciated that the polysaccharide-based coatings developed in the study by Ferraro et al. [25] contained DBS and ethylcellulose, ethylcellulose avoided the dissolution and/or substantial swelling of the "colon targeting" polysaccharide in the upper gastro intestinal tract. DBS is a plasticizer for ethylcellulose, assuring appropriate mechanical properties of the film coatings. As the ratio of DBS and ethylcellulose was maintained as a constant in every coating ("colon targeting" polysaccharide: ethylcellulose = 2:3 *w*/w), the identity of the "colon–targeting" polysaccharide can be viewed as the independent variable for ML purposes. The "colon–targeting" polysaccharides used were Abelmoscus esculentus extract (Specialty Natural Products Co. Ltd., Chon Buri, Thailand), *Coix lacryma* esculentus extract (Specialty Natural Products Co. Ltd., Chon Buri, Thailand), maize maltodextrin (Roquette Freres, Lestrem, France), raffinose (Alfa Aesar, Kendel, Germany), pregelatinized starch (Starch 1500; Colorcon, Kent, UK), cook-up maize starch (Roquette Freres, Lestrem, France), xylan (Tokyo Chemical industry, Zwijndrecht, Belgium), inulin (Orafti Synergy 1; Beneo-Orafti, Oreye, Belgium), resistant maize starch (Novelose 240; Ingredion, Hamburg, Germany), inulin (Orafti HIS; Beneo-Orafti, Oreye, Belgium), rice starch (Specialty Natural Products Co. Ltd., Chon Buri, Thailand), goji berry extract (Specialty Natural Products Co. Ltd., Chon Buri, Thailand), isomaltulose (Beneo-Orafti, Oreye, Belgium), maltitol (Roquette Freres, Lestrem, France), and inulin (Orafti HP; Beneo-Orafti, Oreye, Belgium). These coatings, composed of different polysaccharides, were tested for their colon-targeting potential by incubating them in faecal material sourced from humans diagnosed with inflammatory bowel disease (IBD); rats with 2,4,6-trinitrobenzene sulfonic acid (TNBS)-stimulated colitis; and healthy dogs. A faeces-free medium was also utilised as a microbiota-free control. The release of 5-ASA from the polysaccharide-based coatings in the four incubation media was measured using high-performance liquid chromatography. The polysaccharide-based coatings that were efficiently digested by the faecal microbiota, enabling 5-ASA release, were deemed potential candidates for new formulations designed to deliver drugs to the colon following oral administration.

The raw data from the work by Ferraro et al. [25], subsequently utilised in this study as an ML dataset, are presented in the GitHub repository (https://github.com/y-babdalla/coating_release). Here, the names of the polysaccharides utilised in the coatings are arranged alongside their incubation medium and 5-ASA release percentage at either 2, 8, or 24 h. 5-ASA release at these three incubation time points was included in the dataset to enable the ML models to predict drug release at a range of times rather than providing a single snapshot, as this could better predict the drug release profiles from different polysaccharides. We also captured the Raman spectra for all the polysaccharide coatings and each polysaccharide was paired with its Raman spectrum to capture its unique chemical structure, which was used as an input for ML.

2.2. Raman spectroscopy

Raman spectra were captured using a Renishaw InVia Raman spectrometer (Renishaw plc, Wottonunder-Edge, Gloucestershire, UK) composed of a single-grating spectrograph coupled with an optical Leica microscope. The 785 nm line of a Cobolt diode laser was used for excitation. Focusing the laser beam via a × 50 long-working distance objective, a volume of about 400 mm^3 of each raw polysaccharide powder was analysed. The scattered light was collected in backscattering geometry, with an acquisition time of 120 s and a resolution of around 2 cm^{-1} in the 150–1500 cm^{-1} spectral range to cover the molecular fingerprint region.

2.3. Data analysis and statistics

2.3.1. Machine learning models

Five distinct ML models were employed in this study. These included 3 tree-based ensembles: Extreme Gradient Boosting (XGBoost) [26], Light Gradient-Boosting Machine (LightGBM) [27], and Random Forest (RF) [28], along with a kernel-based method, Support Vector Machine (SVM) [29] and a memory-based learning algorithm, K-nearest

neighbours (KNN) [30]. Additionally, K-means clustering was used to cluster the Raman spectra. The models were run on a Server (Operating System: Ubuntu 20.04 LTS; Processor: AMD EPYC 7282 16-core 2.8GHz; RAM Memory: 512GB, GPU: RTX 3090 24GB). Python (Version 3.10.4) was used to run the ML models (Python Software Foundation). All ML models were trained using the Scikit-learn (Version 1.1.3) Python package [31], except for XGBoost (Xgboost Version 1.6.2) and LightGBM (lightgbm Version 4.1.0), which were trained through their respective libraries. The complete code is available in the GitHub repository (https://github.com/y-babdalla/coating_release).

2.3.2. Unsupervised learning

Uniform Manifold Approximation and Projection (UMAP) [32] is a dimensionality reduction technique, it was used to visualise the Raman spectra into a single point in 2-dimensional space. UMAP maps a high-dimensional graph as a low-dimensional graph whilst trying to maintain structural similarity between the two graphs. A fuzzy simplicial complex is generated, which is a connected graph representing the topological representation of the high-dimensional data; the edge weights show the likelihood that points are connected. UMAP optimizes a low-dimensional embedding of the data while preserving the local and global structures of the data.

2.3.3. Feature processing

Before being input into the machine learning models, the incubation media (IBD human (patient), IBD rat (rat), healthy dog (dog), or control) were label encoded. The Raman spectra were uniformly processed using the Savitzky-Golay filter for noise elimination from the spectra [33]. This was followed by a baseline adjustment to diminish the background signal and improve peak discernibility [9]. Downsampling was then performed on the spectra to reduce their dimensionality while preserving their structure; downsampling between 1 and 40 was trialled, and the effect on the produced spectrum and model performance was evaluated. Finally, all data was normalised within a range of 0–1.

2.3.4. Hyperparameter optimisation

The nested cross-validation method utilised by Bannigan et al. [23] was followed to optimise the models' hyperparameters. This approach included an inner loop that employed 5-fold cross-validation within a random search of 100 different hyperparameter combinations. The outer loop was structured into 5 cycles, within which the data was randomly split into a training set for hyperparameter tuning and a testing set for validation of the model selected through this tuning process.

2.3.5. Uncertainty quantification

Conformal predictions [34] were used to identify confidence intervals for the predictions made by the models. The training data was divided into train (70 %) and calibration (30 %) sets. Models were trained using the train set, and predictions were subsequently made on the calibration set. The discrepancy between actual and predicted values on this set was utilised to identify the models' confidence. This confidence was then used to compute the uncertainty interval for the predictions on the test data.

2.3.6. Explainability analysis

SHAPley additive explanations (SHAP), an algorithm based on Game theory [35], was employed to investigate the features influencing the best model's decision-making. This analysis was conducted using the Python SHAP package (Version 0.42.1). The most influential features and their SHAP values were selected, representing their contribution to the model's final decision.

2.3.7. Determining model performance

The models' performances were evaluated using 5-fold cross-validation. In this method, the data was divided into five equal segments. Each segment was used as the test set once, while the remaining data was the training set. The models' overall performances were determined by calculating the mean of the results across the five iterations. The regression metrics used were the coefficient of determination (R^2), mean square error (MSE) and mean absolute error (MAE) (Table 1).

2.4. Experimental validation

The Raman spectra of two polysaccharides – rice starch and Abelmoscus esculentus extract - were acquired. Their spectra were then inputted into models to predict their release in the incubation media. Subsequently, 5-ASA-loaded pellets coated with these polysaccharides were manufactured [25] and incubated in all four media types to determine their release profiles. The predicted release profile was compared to the actual release observed to validate the models.

3. Results and discussion

3.1. Exploratory data analysis

Since Raman spectra describe the chemistry and bonding of the molecules involved in a drug coating, and as this chemistry should be intrinsically linked to the ability of the microbiota to digest it and release the drug, we surmised that these spectra may be suitable for predicting drug release profiles. We turned to a recently published dataset of 60 drug release profiles, with associated Raman spectra and aimed to generate predictive models that could take Raman spectra of coatings with unknown release profiles as input.

The drug release characteristics of the 13 polysaccharide coatings in the training dataset are presented in **Fig. 1**A. Variability of 5-ASA release increased at the later timepoints, with the widest spread of drug release measured at 24 h. The extent of drug release also increased between all three timepoints (**Fig. 1**B). No statistical difference was observed between any of the media at 2 h. At 8 h, drug release in the control and dog media was significantly lower than in patients and rats (ANOVA, $p < 0.05$). At 24 h, release in the control medium was significantly lower than in all other media (ANOVA, $p < 0.05$). This similarity means that a single model can be employed to predict all release profiles by label encoding the medium, rather than developing separate models for each medium. Consequently, this approach provides the model with more data to learn from, enhancing the accuracy of its predictions. However, predicting drug release in dog media at 8 h remains particularly challenging due to its distinct release trend compared to other media.

To further understand the nature of the data used to train the model, the Raman spectra of the polysaccharide coatings were analysed. Analysis of the Raman spectra of the training dataset revealed that all coatings had between 12 and 20 peaks in their spectra (Fig. 2A). Most of the peaks were identified between 500 and 1200 cm^{-1}, corresponding to aliphatic chains, ethers and aromatic rings, and 1200–1700 cm^{-1} (Fig. 2B), corresponding to aromatic and hetero rings, methyl and ethyl groups [36]. Furthermore, UMAP dimensionality reduction and K-means clustering [37] using an n of 2 was carried out to cluster the

Table 1
Regression ML metrics. Abbreviations- N: population size, y_i: actual value, $\bar{y}$: mean value, $\hat{y}_i$: predicted value.

Metric	Focus	Calculation
R^2	Measures the goodness of fit of a regression model, higher values indicate a better fit.	$1-\frac{\sum_i (y_i-\hat{y}_i)^2}{\sum_i (y_i-\bar{y})^2}$
MSE	Mean of the square of the difference between the actual and predicted values	$\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2$
MAE	Mean of the absolute difference between the actual and predicted values	$\frac{1}{N}\sum_{i=1}^{N}\lvert y_i-\hat{y}_i\rvert$

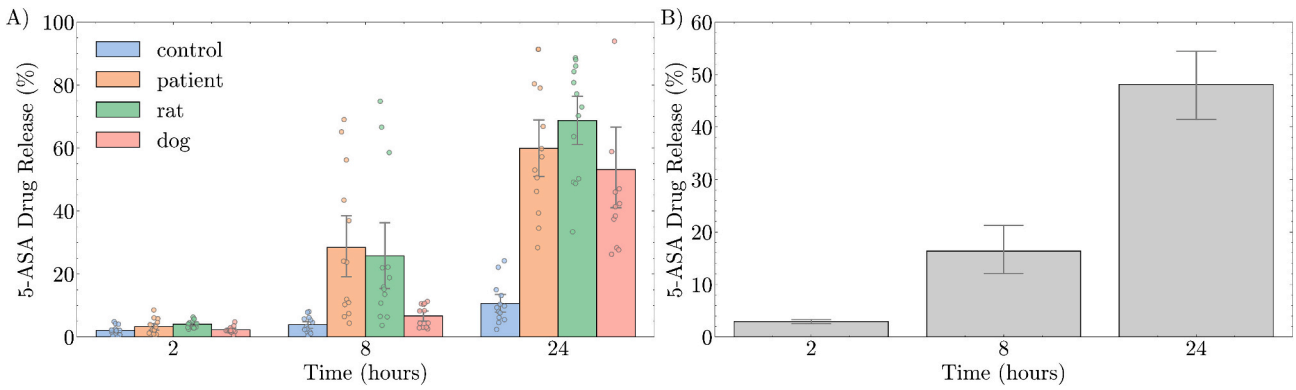

Fig. 1. The 5-aminosalicylic acid (5-ASA) release from polysaccharide-based coatings in the original training dataset A) separated according to the timepoint and incubation media (individual points overlayed) and B) separated by incubation time.

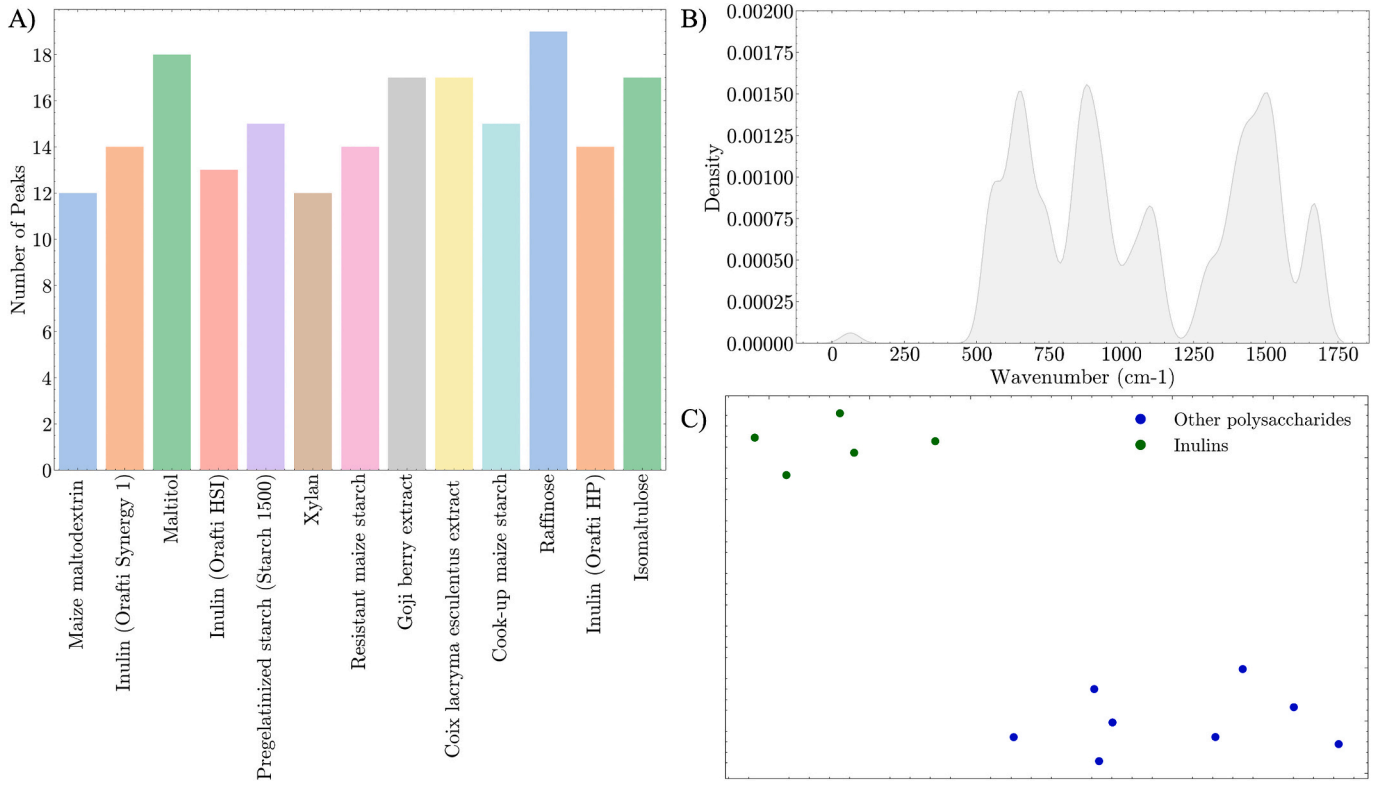

Fig. 2. Analysis of the polysaccharides coatings' Raman spectra. A) The number of peaks in the Raman spectra of the coatings used in the training dataset. B) The kernel density estimate (KDE) distribution of peaks in the Raman spectra of the coatings used in the training dataset. C) UMAP of the Raman spectra coloured by K-means clusters.

wavelengths (Fig. 2C). The two identified clusters represent inulins (and similar structures) and other polysaccharides.

3.2. *Raman Spectra successfully train machine learning models*

We evaluated the ability of five different ML models - LightGBM, XGBoost, RF, KNN and SVM - to predict 5-ASA release profiles given the polysaccharides Raman Spectra as inputs. These models encompass a range of algorithmic approaches, including tree-based, kernel-based, and memory-based methodologies. Models were trained to predict the percentage of 5-ASA release from polysaccharide-coated pellets, with data describing polysaccharide-coating, incubation medium, and incubation time used as input features. A 5-fold CV approach was utilised on the training data, and the R^2 was computed to assess model performance. An initial examination was conducted to compare the impact of spectrum processing on model accuracy: the raw Raman spectra, the normalised, denoised spectrum and the normalised, denoised spectrum reduced to 20 dimensions using partial least squares regression (PLSR) (Fig. 3A). We found that all spectra performed similarly for tree-based models. However, SVM and KNN performed better for the PLSR processed spectrum, likely attributed to the greater simplicity of the data. Furthermore, we observed that the spectrum processing reduced variability in XGBoost and LightGBM, consistent with the literature showing the effect of noise reduction and normalisation on model performance [38]. While PLSR is commonly used for spectrum pre-processing, and it exhibited strong predictive performance, its performance was similar to that of the processed spectrum, and it cannot be easily interpreted to understand which components of the Raman spectrum are important for release. Consequently, use of the processed Raman spectrum was used in further work.

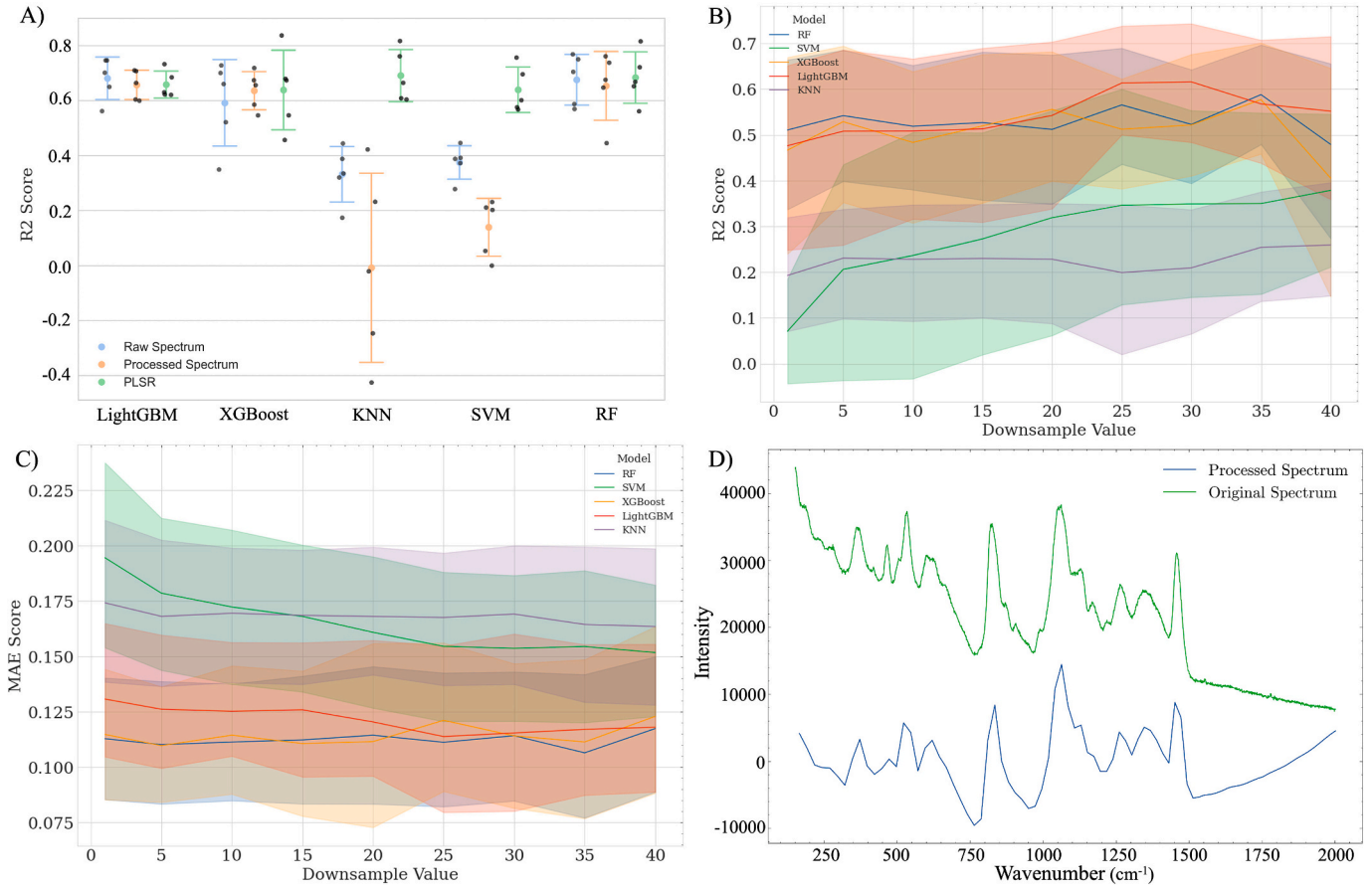

Fig. 3. A) R2 scores for predictions made following different Raman spectrum pre-processing of the polysaccharides coating. The individual coloured points represent the mean of the data along with the standard deviation, while the black points denote the individual data points. B) The effect of downsampling the Raman spectra on R^2 and C) MAE. D) Comparison of an example raw spectrum and its denoised, baseline shifted, downsampled spectrum.

Having established that Raman spectra have predictive power for drug release, the impact of downsampling the processed spectra was assessed. Specifically, the effects of downsampling by selecting up to every 40th wavenumber intensity reading were evaluated. Both the R^2 (Fig. 3B) and MAE (Fig. 3C) scores were measured to determine the impact of this process on model performance. We found that downsampling did not significantly affect the overall performance of the models. However, this approach reduces computational requirements while preserving the data's original structure and therefore, downsampling is an effective dimensionality reduction method that maintains data integrity and explainability. This method is notably simple and requires minimal computational resources. In contrast, other dimensionality reduction methods, such as principal component analysis (PCA), which are frequently employed for similar purposes lead to a loss in the data's explainability, as interpretation of principal components can be difficult [39]. Based on the findings, downsampling at a factor of 20 was deemed optimal and thus included in future models. This level of downsampling effectively reduced the dimensions of the data while conserving its original structure and characteristics (Fig. 3D).

3.3. *Raman spectra accurately predict drug release profiles for polysaccharide coatings*

We next sought to understand and improve the generalisability of our model through cross-validation. The nested cross-validation method employed by Bannigan et al. [23] was used to tune the hyperparameters of the five ML models and evaluate their performance. In this process, for each iteration, a validation set was separated from the dataset, while the remaining training data underwent a random search cross-validation ($n = 100$) to identify the optimal hyperparameters for each model. Subsequently, these tuned models were tested against the validation set.

As shown in Fig. 4, tree-based models outperformed their counterparts. This outcome aligns with the existing literature, suggesting that tree-based models are particularly well-suited for handling tabular data [40]. This was evident in the higher R^2 scores and lower MAE and MSE, indicating more accurate and reliable predictions. Among the models tested, XGBoost and RF emerged as the top performers, achieving the highest R^2 of 0.81 and 0.80, respectively and the same lowest MAE score of 0.08, underscoring their efficacy in this application. This is consistent with literature highlighting the compatibility of these models with small datasets [41]. Due to their superior performance over other models, we progressed them both for further evaluation.

Subsequent analysis of predictions revealed that the models had higher performance when predicting 5-ASA release at lower values, as illustrated in (Fig. 5A and B). This enhanced accuracy at lower release levels can be attributed to the training data distribution, predominantly composed of instances with less than 40 % drug release (Fig. 1A). Consequently, the model was inherently more reliable in estimating release percentages within this more frequently represented lower range. However, the model could still predict higher release values and thus could capture the range of values present in a complete drug release profile, as shown in Fig. 5C.

3.4. *Machine Learning of Raman spectra is generalisable*

To fully validate the predictive power of our model, we chose two additional polysaccharide coatings to experimentally characterise and test against model predictions — Rice starch and Abelmoscus esculentus

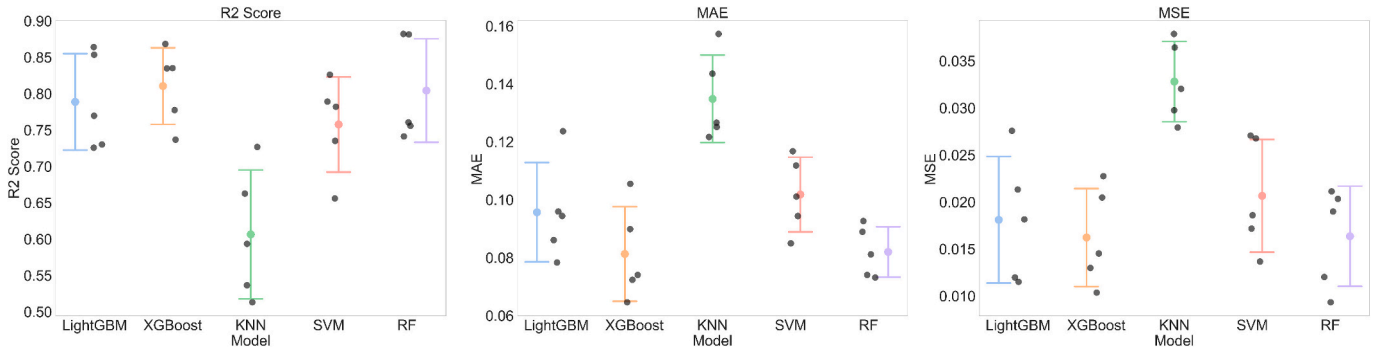

Fig. 4. Performance of the outer loop in the nested cross-validation. The individual coloured points represent the mean of the data along with the standard deviation, while the black points denote the individual data points.

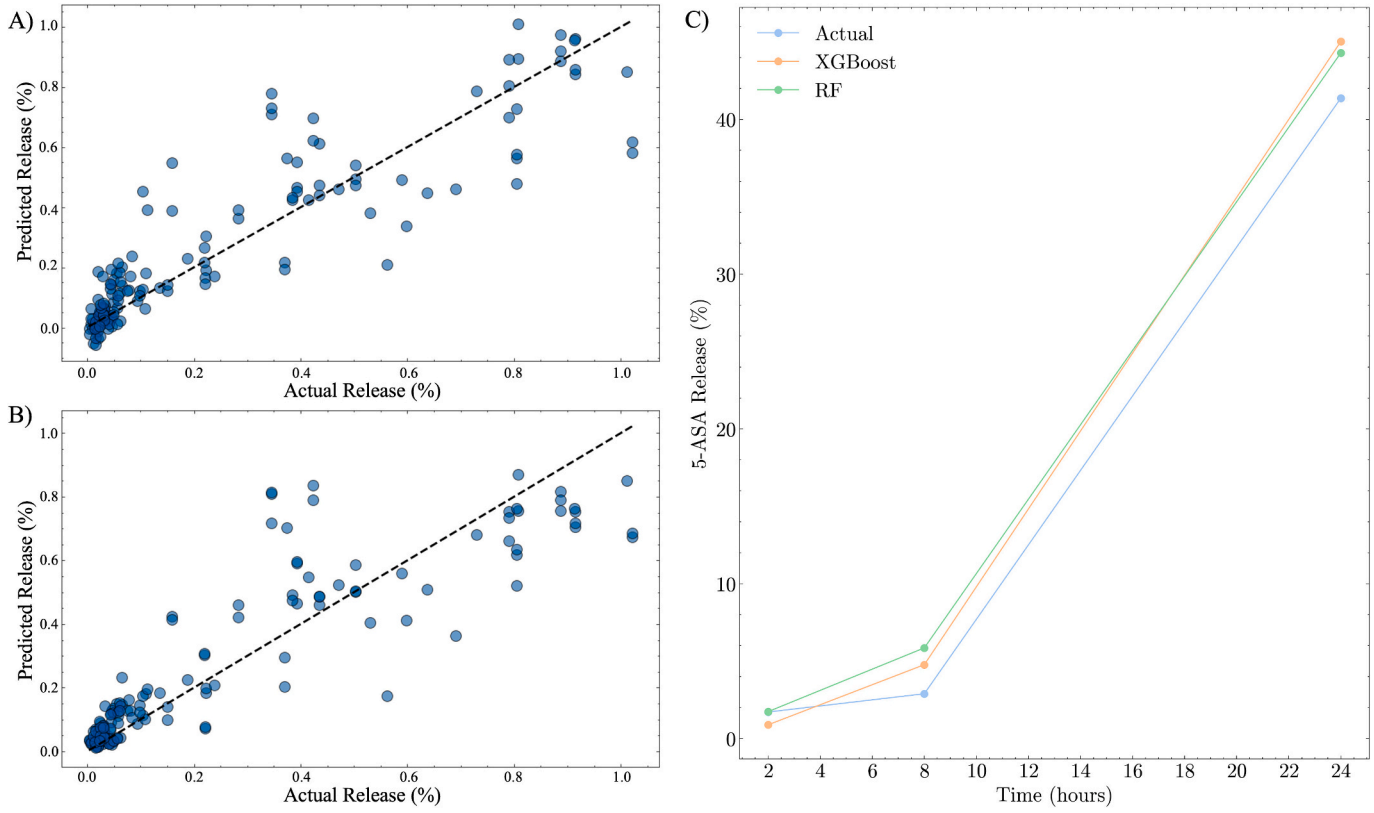

Fig. 5. A) Actual vs predicted release for XGBoost in the outer loop of the nested cv. B) Actual vs predicted release for RF in the outer loop of the nested cv. C) Example release profile for Goji berry extract incubated in IBD dog medium.

extract. Raman spectra from these coatings were fed into the final ML model and predictions of release profiles in media compared to experimentally measured equivalents. When comparing the Raman spectra of polysaccharides in the new validation set to those in the original training set, it was observed that they had a comparable number of peaks (Fig. 6A) located at similar positions (Fig. 6B). Therefore, the model was expected to perform well in predicting their release.

Both RF and XGBoost, the top-performing models, were evaluated. There was a slight reduction in XGBoost performance when applied to this new data, with the R^2 dropping to 0.72 and no change to the MAE and MSE. However, RF proved to be more robust, with only a minute reduction in R^2 to 0.79, and no change in MAE and MSE. This suggests that RF was less overfit and more generalisable, therefore this model was chosen for further evaluation. Fig. 7 shows the RF predicted and actual drug release profile, including a 90 % confidence interval determined via conformal predictions. Overall, the model showed great predictive performance, with very similar predicted and release profiles. Additionally, it was observed that the majority of the actual data points fell within the model's predicted confidence interval, suggesting a high reliability in the model's predictions. The model could predict drug release in dog media at 8 h, although this had been projected to be the most difficult to predict. Consistent with prior observations, the model exhibited reduced prediction accuracy for higher drug release percentages, as identified for prior observations, and as seen with the release profile of both coatings in rat media. Incorporating more data points representing these higher ranges can likely further enhance the model's performance. However, it is important to note that since the available data was limited to 2, 8, and 24 h, model performance could only be assessed at these specific intervals. Therefore, future work should incorporate additional time points in model training and evaluation to enable a more comprehensive assessment of model performance. Furthermore, while the model demonstrated generalisability to the two

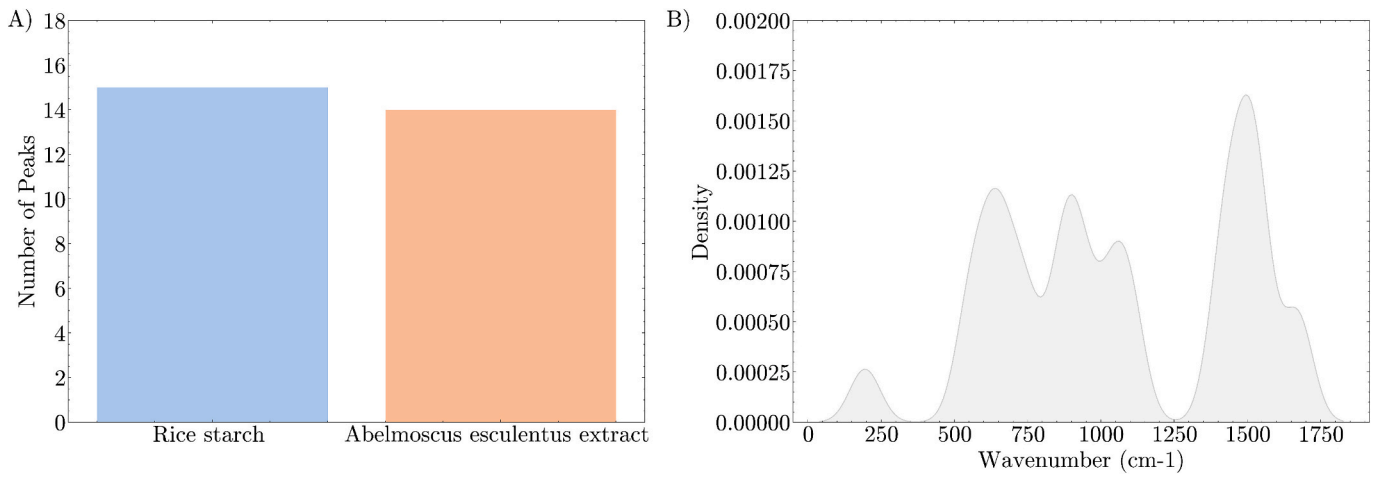

Fig. 6. Analysis of the polysaccharide coatings' Raman spectra. A) The number of peaks in the Raman spectra of the coatings used in the validation dataset. B) The KDE distribution of peaks in the Raman spectra of the coatings used in the training dataset.

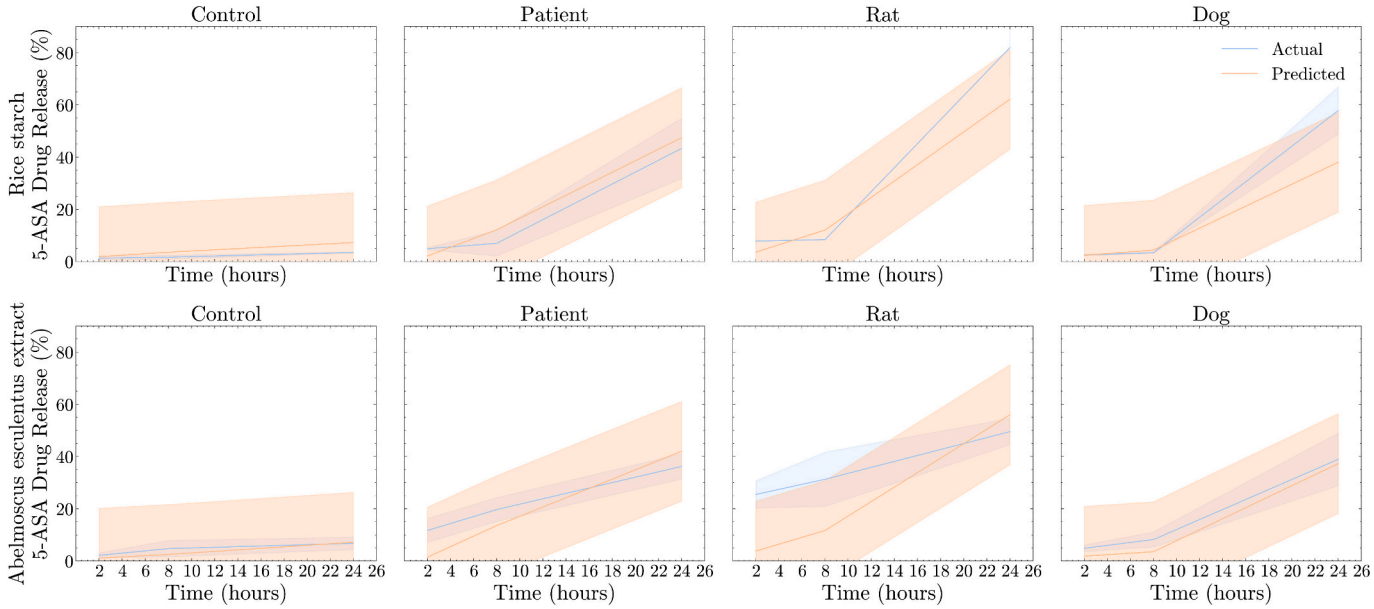

Fig. 7. Predicted and actual release profile of 5-ASA from the coatings.

polysaccharides studied, additional data is required to fully evaluate its generalisability and to test its limits. Nonetheless, the model is fit for purpose specifically for predicting 5-ASA release from polysaccharide coatings in the four media. More data would be necessary to expand the remit of this model beyond its current scope.

3.5. *Explainability analysis reveals glycosidic linkages as key predictors of polysaccharide degradation*

The model's decision-making process was analysed using SHAP to gain deeper insights into the factors influencing drug release predictions (Fig. 8). Consistent with expectations, time and the incubation medium were the most influential factors. Additionally, the analysis highlighted several key wavenumbers, including 522 cm^{-1}, 620 cm^{-1}, 858 cm^{-1}, 927 cm^{-1}, and 1085 cm^{-1}. These wavenumbers align with the predominant peaks observed in the exploratory data analysis. They are associated with molecular vibrations such as C—H bending, C—C stretching, C—O stretching, and ethyl bending, respectively [36]. Most notably, the 522 cm^{-1}, 858 cm^{-1}, 927 cm^{-1} and 1085 cm^{-1} peaks correspond to regions where different glycosidic linkages are found [42,43]. These bonds, which interlink sugar moieties, are metabolised by bacteria during the degradation of polysaccharides [44,45]. Shifts in the Raman peak can indicate factors such as the bond strength, molecular interactions, polymorphism and the overall structure of the polysaccharide [46–49]. As well as influencing the solubility of the polysaccharides in the GI tract, these factors impact the stability of the glycosidic bond to hydrolysis and enzymatic degradation by intestinal bacteria [50,51] and hence influence the release of the drug from the coating. Consequently, the model accurately identified these crucial areas in the spectrum as key indicators for predicting polysaccharide degradation.

The ML models presented in this study were developed to assist in selecting suitable polysaccharides for colonic drug delivery coatings. This predictive capability significantly reduces the time that would otherwise be required for in vitro screening. The developed models were robust and notably demonstrated a strong alignment with the drug release data used for external validation, as evidenced by most data points falling within the model's 90 % confidence interval. This indicates a reasonable degree of reliability and applicability in practical scenarios. Though none of the polysaccharides in the training/testing

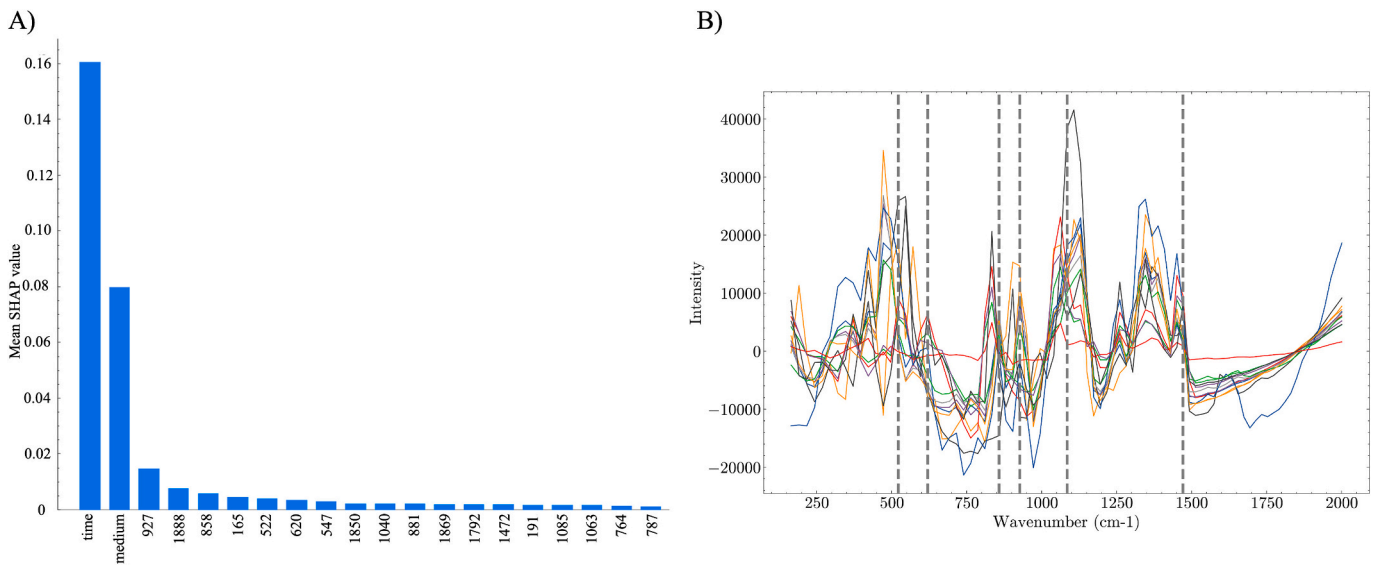

Fig. 8. A) SHAP analysis for the RF model for predicting drug release. B) Overlay of the polysaccharides Raman spectra, dotted lines represent wavenumbers that SHAP analysis identified as important.

dataset underwent pre-treatment prior to incorporation into the colon targeting coatings (e.g., with heat), it should be noted that polysaccharide pre-treatment should be considered when generating predictions for new polysaccharides. For example, if a polysaccharide has undergone heat pre-treatment then it is advised that the Raman spectrum of the treated polysaccharide sample is used for model predictions rather than the raw starting material.

The RF model developed in this research can predict the 5-ASA release profile from polysaccharide-based coatings designed for colonic drug delivery. Notably, its robust performance on an external validation set confirmed the model's generalisability. Previous studies have predicted drug release from various pharmaceutical formulations via other methods. For instance, Barmpalexis et al. [52] successfully employed an artificial neural network (ANN) to predict the release of nimodipine from matrix tablet formulations, achieving an R^2 of 0.90. Similarly, Petrović et al. [53] modelled the dissolution profiles of different matrix tablet types, achieving a Pearson's correlation of 0.9991 between predicted and observed to optimise controlled drug release. Salem et al. [54] used an ANN to predict drug release from polyethene oxide-based promethazine tablets and obtained an R^2 of 0.9381. The features used to characterise the formulations in these models included polymer composition, processing parameters and physical properties. While these models exhibit superior performance metrics compared to the one presented herein, it is essential to consider the differences in the test environments. The previous models operated in more controlled settings, contrasting with the complexities and variability inherent in the gut microbiome. The microbiome presents a diverse and less predictable environment, which poses additional challenges for accurate modelling. Furthermore, the model herein encompasses multiple media types, adding to its complexity, whereas previous studies typically focused on a single medium. Another distinction lies in using Raman spectroscopy in our study, enhancing the model's generalisability. The Raman-based approach allows for a broader application across different drug formulations and conditions, surpassing the specificity of previous models. Consequently, despite the seemingly lower performance metrics, the model developed in this study demonstrates greater robustness and a wider range of potential applications. This advancement holds significant promise in accelerating the production of targeted drug coatings, offering an alternative to the lengthy, expensive, and labour-intensive in vitro methods currently in use.

4. Conclusion

This study marks a significant advance as the first to utilise spectral data to predict drug release from pharmaceutical formulations. It introduces an innovative approach to preparing Raman spectra for ML via dimensionality reduction, successfully preserving the original data's explainability, structure, and integrity. The RF model developed in this study provides a tool for the streamlined development of colon-targeted formulation coatings, which can be employed to increase the efficiency, sustainability, and success of future colonic drug delivery programmes. A notable use case for this methodology is pre-ranking polysaccharide coatings before experimental validation of top hits, significantly reducing experimental requirements for developing new formulations. Whilst this work has produced a seemingly robust model, we note that the dataset is still small by ML standards and additionally performed with only a single drug. Future work will involve expanding the range of polysaccharide coatings studied and exploring the role of drug properties on release, particularly in light of previous work that has demonstrated interactions between drugs and microbiota.

CRediT authorship contribution statement

Youssef Abdalla: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Laura E. McCoubrey:** Writing – original draft, Methodology, Formal analysis, Data curation, Conceptualization. **Fabiana Ferraro:** Writing – review & editing, Data curation. **Lisa Maria Sonnleitner:** Writing – review & editing, Data curation. **Yannick Guinet:** Data curation. **Florence Siepmann:** Writing – review & editing, Supervision. **Alain Hédoux:** Writing – review & editing, Supervision. **Juergen Siepmann:** Writing – review & editing, Supervision. **Abdul W. Basit:** Writing – review & editing, Supervision, Conceptualization. **Mine Orlu:** Writing – review & editing, Supervision. **David Shorthouse:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

This work was funded by University College London Research Excellence Scholarship and the University College London Centre for Digital Innovation Scholarship (Youssef Abdalla).

Data availability

Data available in the supplementary

References

[1] M.S. Alqahtani, M. Kazi, M.A. Alsenaidy, M.Z. Ahmad, Advances in Oral drug delivery, Front. Pharmacol. 12 (2021) 618411.
[2] L.E. McCoubrey, A. Favaron, A. Awad, M. Orlu, S. Gaisford, A.W. Basit, Colonic drug delivery: formulating the next generation of colon-targeted therapeutics, J. Control. Release 353 (2022) 1107–1126.
[3] H. Jadhav, P. Augustijns, C. Tannergren, Approaches to account for colon absorption in physiologically based biopharmaceutics modeling of extended-release drug products, Mol. Pharm. 20 (12) (2023) 6272–6288.
[4] F. Taherali, F. Varum, A.W. Basit, A slippery slope: on the origin, role and physiology of mucus, Adv. Drug Deliv. Rev. 124 (2018) 16–33.
[5] M. Thörn, N. Finnström, S. Lundgren, A. Rane, L. Lööf, Cytochromes P450 and MDR1 mRNA expression along the human gastrointestinal tract, Br. J. Clin. Pharmacol. 60 (2005) 54–60.
[6] M. Drozdzik, D. Busch, J. Lapczuk, J. Müller, M. Ostrowski, M. Kurzawski, S. Oswald, Protein abundance of clinically relevant drug transporters in the human liver and intestine: a comparative analysis in paired tissue specimens, Clin. Pharmacol. Ther. (St. Louis, MO, U. S.) 105 (2019) 1204–1212.
[7] V. Yadav, A. House, S. Matiz, L.E. McCoubrey, K.A. Bettano, L. Bhave, M. Wang, P. Fan, S. Zhou, J.D. Woodhouse, E. Poimenidou, L. Dou, A.W. Basit, L.Y. Moy, R. Saklatvala, L.G. Hegde, H. Yu, Ileocolonic-targeted JAK inhibitor: a safer and more effective treatment for inflammatory bowel disease, Pharmaceutics 14 (2022).
[8] P. Singh, P. Waghambare, T.A. Khan, A. Omri, Colorectal cancer management: strategies in drug delivery, Expert. Opin. Drug Deliv. 19 (2022) 653–670.
[9] L.E. McCoubrey, N. Seegobin, N. Sangfuang, F. Moens, H. Duyvejonck, E. Declerck, A. Dierick, M. Marzorati, A.W. Basit, The colon targeting efficacies of mesalazine medications and their impacts on the gut microbiome, J. Control. Release 369 (2024) 630–641.
[10] M. Peiris, R. Aktar, D. Reed, V. Cibert-Goton, A. Zdanaviciene, W. Halder, A. Robinow, S. Corke, H. Dogra, C.H. Knowles, A. Blackshaw, Decoy bypass for appetite suppression in obese adults: role of synergistic nutrient sensing receptors GPR84 and FFAR4 on colonic endocrine cells, Gut 71 (2022) 928–937.
[11] L.E. McCoubrey, M. Elbadawi, A.W. Basit, Current clinical translation of microbiome medicines, Trends Pharmacol. Sci. 43 (2022) 281–292.
[12] B. Verstockt, D. Alsoud, J. van Oostrom, J. Smith, J. Stylli, S. Singh, S. van Gennep, P. Rahimian, J. Sabino, M. Ferrante, S. Singh, G. D'Haens, S. Vermeire, Tofacitinib tissue exposure correlates with endoscopic outcome, J. Crohn's Colitis 16 (2022) i394–i395.
[13] S. Moutaharrik, A. Maroni, C. Neut, C. Dubuquoy, L. Dubuquoy, A. Foppoli, M. Cerea, L. Palugan, F. Siepmann, J. Siepmann, A. Gazzaniga, In vitro and in vivo evaluation of a pH-, microbiota- and time-based oral delivery platform for colonic release, Eur. J. Pharm. Biopharm. 183 (2023) 13–23.
[14] F. Varum, A. Cristina Freire, R. Bravo, A.W. Basit, OPTICORE, an innovative and accurate colonic targeting technology, Int. J. Pharm. 583 (2020) 119372.
[15] V. Doggwiler, C. Puorger, V. Paredes, M. Lanz, K.M. Nuss, G. Lipps, G. Imanidis, Efficient colonic drug delivery in domestic pigs employing a tablet formulation with dual control concept, J. Control. Release 358 (2023) 420–438.
[16] V. Doggwiler, M. Lanz, V. Paredes, G. Lipps, G. Imanidis, Tablet formulation with dual control concept for efficient colonic drug delivery, Int. J. Pharm. 631 (2023) 122499.
[17] A. Awad, C.M. Madla, L.E. McCoubrey, F. Ferraro, F.K.H. Gavins, A. Buanz, S. Gaisford, M. Orlu, F. Siepmann, J. Siepmann, A.W. Basit, Clinical translation of advanced colonic drug delivery technologies, Adv. Drug Deliv. Rev. 181 (2022) 114076.
[18] S. Moutaharrik, G. Meroni, A. Soggiu, A. Foppoli, M. Cerea, L. Palugan, F. Caloni, P.A. Martino, A. Gazzaniga, A. Maroni, Guar gum as a microbially degradable component for an oral colon delivery system based on a combination strategy: formulation and in vitro evaluation, Drug Deliv. Transl. Res. 14 (3) (2024) 826–838.
[19] H.M. Zawbaa, J. Szlęk, C. Grosan, R. Jachowicz, A. Mendyk, Computational intelligence modeling of the macromolecules release from PLGA microspheres—focus on feature selection, PLoS One 11 (2016) e0157610.
[20] P. Carou-Senra, J.J. Ong, B.M. Castro, I. Seoane-Viaño, L. Rodríguez-Pombo, P. Cabalar, C. Alvarez-Lorenzo, A.W. Basit, G. Pérez, A. Goyanes, Predicting pharmaceutical inkjet printing outcomes using machine learning, Int. J. Pharmaceut.: X 5 (2023) 100181.
[21] Y. Li, M.R. Abbaspour, P.V. Grootendorst, A.M. Rauth, X.Y. Wu, Optimization of controlled release nanoparticle formulation of verapamil hydrochloride using artificial neural networks with genetic algorithm and response surface methodology, Eur. J. Pharm. Biopharm. 94 (2015) 170–179.
[22] Y. Abdalla, M. Elbadawi, M. Ji, M. Alkahtani, A. Awad, M. Orlu, S. Gaisford, A. W. Basit, Machine learning using multi-modal data predicts the production of selective laser sintered 3D printed drug products, Int. J. Pharm. 633 (2023) 122628.
[23] P. Bannigan, Z. Bao, R.J. Hickman, M. Aldeghi, F. Häse, A. Aspuru-Guzik, C. Allen, Machine learning models to accelerate the design of polymeric long-acting injectables, Nat. Commun. 14 (2023) 35.
[24] M. Li, R. Wang, Q. Bao, Hyper-spectra imaging analysis of PLGA microspheres via machine learning enhanced Raman spectroscopy, J. Control. Release 367 (2024) 676–686.
[25] F. Ferraro, L.M. Sonnleitner, C. Neut, S. Mahieux, J. Verin, J. Siepmann, F. Siepmann, Colon targeting in rats, dogs and IBD patients with species-independent film coatings, Int. J. Pharm. X (2024) 100233.
[26] T. Chen, C. Guestrin, XGBoost: A Scalable Tree Boosting System, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, San Francisco, California, USA, 2016, pp. 785–794.
[27] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, LightGBM: a highly efficient gradient boosting decision tree, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, Curran Associates Inc., Long Beach, California, USA, 2017, pp. 3149–3157.
[28] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[29] M.A. Hearst, S.T. Dumais, E. Osuna, J. Platt, B. Scholkopf, Support vector machines, IEEE Intellig. Syst. Applicat. 13 (1998) 18–28.
[30] Z. Zhang, Introduction to machine learning: k-nearest neighbors, Ann. Transl. Med. 4 (2016) 218.
[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, Scikit-learn: machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[32] L. McInnes, J. Healy, J. Melville, Umap: Uniform manifold approximation and projection for dimension reduction, Journal of Open Source Software 3 (29) (2018) 861.
[33] A. Savitzky, M.J.E. Golay, Smoothing and differentiation of data by simplified least squares procedures, Anal. Chem. 36 (1964) 1627–1639.
[34] A.N. Angelopoulos, S. Bates, Conformal prediction: A gentle introduction, Foundations and Trends in Machine Learning 16 (4) (2023) 494–591.
[35] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Proces. Syst. 30 (2017).
[36] A. Gałat, Study of the Raman scattering and infrared absorption spectra of branched polysaccharides, Acta Biochim. Pol. 27 (1980) 135–142.
[37] X. Jin, J. Han, K-means clustering, in: C. Sammut, G.I. Webb (Eds.), Encyclopedia of Machine Learning, Springer, US, Boston, MA, 2010, pp. 563–564.
[38] S. Gupta, A. Gupta, Dealing with noise problem in machine learning data-sets: a systematic review, Proced. Comput. Sci. 161 (2019) 466–474.
[39] I.T. Jolliffe, Principal Component Analysis for Special Types of Data, Springer, 2002.
[40] L. Grinsztajn, E. Oyallon, G. Varoquaux, Why do tree-based models still outperform deep learning on typical tabular data? Adv. Neural Inf. Proces. Syst. 35 (2022) 507–520.
[41] P. Xu, X. Ji, M. Li, W. Lu, Small data machine learning in materials science, npj Comput. Mater. 9 (2023) 42.
[42] E. Wiercigroch, E. Szafraniec, K. Czamara, M.Z. Pacia, K. Majzner, K. Kochan, A. Kaczor, M. Baranska, K. Malek, Raman and infrared spectroscopy of carbohydrates: a review, Spectrochim. Acta A Mol. Biomol. Spectrosc. 185 (2017) 317–335.
[43] P. Talik, P. Moskal, M. Kucharska, L.M. Proniewicz, A. Wesełucha-Birczyńska, Raman spectroscopy investigations of hydrated hydroxypropyl cellulose mixtures with low-soluble salicylic acid: molecular interactions and the water-binding structure, J. Mol. Struct. 1294 (2023) 136452.
[44] M. Inman, How bacteria turn fiber into food, PLoS Biol. 9 (2011) e1001227.
[45] Y.A. Knirel, O.I. Naumenko, S.Y.N. Senchenkova, A.V. Perepelov, Chemical methods for selective cleavage of glycosidic bonds in the structural analysis of bacterial polysaccharides, Russ. Chem. Rev. 88 (2019) 406–424.
[46] F. Zhu, N.W. Isaacs, L. Hecht, L.D. Barron, Polypeptide and carbohydrate structure of an intact glycoprotein from Raman optical activity, J. Am. Chem. Soc. 127 (2005) 6142–6143.
[47] A. Hédoux, Y. Guinet, M. Descamps, The contribution of Raman spectroscopy to the analysis of phase transformations in pharmaceutical compounds, Int. J. Pharm. 417 (2011) 17–31.
[48] M. Kačuráková, M. Mathlouthi, FTIR and laser-Raman spectra of oligosaccharides in water: characterization of the glycosidic bond, Carbohydr. Res. 284 (1996) 145–157.
[49] L. Ashton, P.D. Pudney, E.W. Blanch, G.E. Yakubov, Understanding glycoprotein behaviours using Raman and Raman optical activity spectroscopies: characterising the entanglement induced conformational changes in oligosaccharide chains of mucin, Adv. Colloid Interf. Sci. 199-200 (2013) 66–77.
[50] H. Leemhuis, T. Pijning, J.M. Dobruchowska, B.W. Dijkstra, L. Dijkhuizen, Glycosidic bond specificity of glucansucrases: on the role of acceptor substrate binding residues, Biocatal. Biotransformat. 30 (2012) 366–376.
[51] B. Wei, Y.-K. Wang, W.-H. Qiu, S. Wang, Y.-H. Wu, X.-W. Xu, H. Wang, Discovery and mechanism of intestinal bacteria in enzymatic cleavage of C–C glycosidic bonds, Appl. Microbiol. Biotechnol. 104 (2020) 1883–1890.
[52] P. Barmpalexis, F.I. Kanaze, K. Kachrimanis, E. Georgarakis, Artificial neural networks in the optimization of a nimodipine controlled release tablet formulation, Eur. J. Pharm. Biopharm. 74 (2010) 316–323.
[53] J. Petrović, S. Ibrić, G. Betz, Z. Đurić, Optimization of matrix tablets controlled drug release using Elman dynamic neural networks and decision trees, Int. J. Pharm. 428 (2012) 57–67.
[54] S. Salem, S.R. Byrn, D.T. Smith, V.J. Gurvich, S.W. Hoag, F. Zhang, R.O. Williams, K.L. Clase, Impact assessment of the variables affecting the drug release and extraction of polyethylene oxide based tablets, J. Drug Deliv. Sci. Technol. 71 (2022) 103337.