Optimization of classification and regression analysis of four monoclonal antibodies from Raman spectra using collaborative machine learning approach

Laetitia Minh Maï Le, Balázs Kégl, Alexandre Gramfort, Camille Marini, David Nguyen, Mehdi Cherti, Sana Tfaili, Ali Tfayli, Arlette Baillet-Guffroy, Patrice Prognon, et al.

www.elsevier.com/locate/talanta

Optimization of classification and regression analysis of four monoclonal antibodies from Raman spectra using collaborative machine learning approach

Laetitia Minh Maï Le, Balázs Kégl, Alexandre Gramfort, Camille Marini, David Nguyen, Mehdi Cherti, Sana Tfaili, Ali Tfayli, Arlette Baillet-Guffroy, Patrice Prognon, Pierre Chaminade, Eric Caudron

PII: S0039-9140(18)30227-3
DOI: https://doi.org/10.1016/j.talanta.2018.02.109
Reference: TAL18428

To appear in: *Talanta*

Received date: 25 October 2017
Revised date: 24 February 2018
Accepted date: 27 February 2018

Cite this article as: Laetitia Minh Maï Le, Balázs Kégl, Alexandre Gramfort, Camille Marini, David Nguyen, Mehdi Cherti, Sana Tfaili, Ali Tfayli, Arlette Baillet-Guffroy, Patrice Prognon, Pierre Chaminade and Eric Caudron, Optimization of classification and regression analysis of four monoclonal antibodies from Raman spectra using collaborative machine learning approach, *Talanta,* https://doi.org/10.1016/j.talanta.2018.02.109

Optimization of classification and regression analysis of four monoclonal antibodies from Raman spectra using collaborative machine learning approach

Laetitia Minh Maï Le[a,b 1] and Balázs Kégl[c,g1], Alexandre Gramfort[c,e,f], Camille Marini[c,d], David Nguyen[a], Mehdi Cherti[c,g], Sana Tfaili[b]*, Ali Tfayli[b], Arlette Baillet-Guffroy[b], Patrice Prognon[a,b], Pierre Chaminade[b], Eric Caudron[a,b]

[a] European Georges Pompidou Hospital (AP-HP), Pharmacy department, 75015 Paris, France

[b] Lip(Sys)[2]- EA7357 - Chimie Analytique Pharmaceutique (FKA EA4041 Groupe de Chimie Analytique de Paris-Sud), Univ. Paris-Sud, Université Paris-Saclay, F92290 Chatenay-Malabry, France.

[c] Center of Data Science, Université Paris-Saclay, 91440 Orsay, France.

[d] CMAP, Ecole Polytechnique, 91128 Palaiseau, France

[e] INRIA, Parietal team, Saclay, 91120 Palaiseau, France France

[f] LTCI, Télécom ParisTech, 75013 Paris, France

[g] LAL, CNRS, 91440 Orsay, France

***Corresponding author:** Sana Tfaili. 5, rue Jean Baptiste Clément, 92290 Châtenay-Malabry, France. sana.tfaili@u-psud.fr

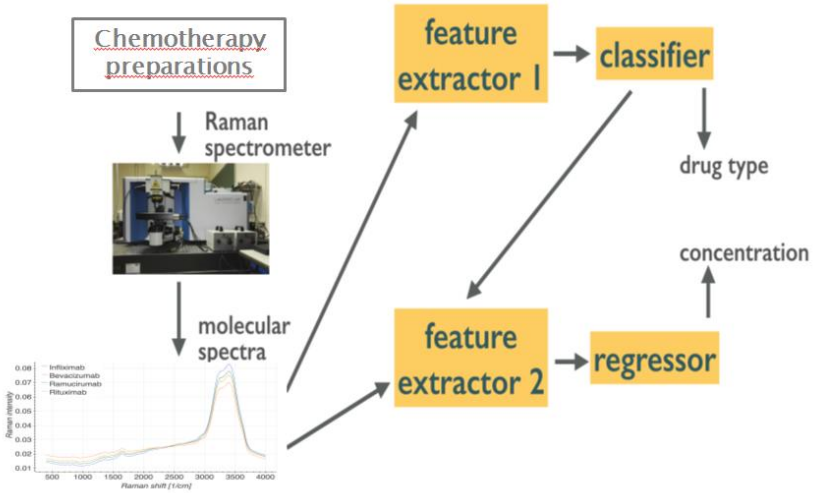
Abstract

The use of monoclonal antibodies (mAbs) constitutes one of the most important strategies to treat patients suffering from cancers such as hematological malignancies and solid tumors.

[1] **Equally contributed to this article**

These antibodies are prescribed by the physician and prepared by hospital pharmacists. An analytical control enables the quality of the preparations to be ensured. The aim of this study was to explore the development of a rapid analytical method for quality control. The method used four mAbs (Infliximab, Bevacizumab, Rituximab and Ramucirumab) at various concentrations and was based on recording Raman data and coupling them to a traditional chemometric and machine learning approach for data analysis. Compared to conventional linear approach, prediction errors are reduced with a data-driven approach using statistical machine learning methods. In the latter, preprocessing and predictive models are jointly optimized. An additional original aspect of the work involved on submitting the problem to a collaborative data challenge platform called Rapid Analytics and Model Prototyping (RAMP). This allowed using solutions from about 300 data scientists in collaborative work. Using machine learning, the prediction of the four mAbs samples was considerably improved. The best predictive model showed a combined error of 2.4% versus 14.6% using linear approach. The concentration and classification errors were 5.8% and 0.7%, only three spectra were misclassified over the 429 spectra of the test set. This large improvement obtained with machine learning techniques was uniform for all molecules but maximal for Bevacizumab with an 88.3% reduction on combined errors (2.1% *versus* 17.9%).

Graphical abstract



Keywords

Machine learning; Chemometrics; Raman spectroscopy; Monoclonal antibody; Classification analysis; Regression analysis

1. Introduction

Cancer can be treated using numerous strategies, such as surgery, radiation therapy, immunotherapy, hormone therapy, stem cell transplant, chemotherapy and more recently targeted therapy. Targeted therapy may use monoclonal antibodies (mAbs) and is the foundation of precision medicine. At the present time, it constitutes one of the most important strategies to treat patients suffering from cancers such as hematological malignancies and solid tumors. Monoclonal antibodies are proteins which bind to specific substances on cancer cells and act by immune-mediated cell-killing mechanisms. They may therefore help the immune system to destroy cancer cells, stop cancer cells from growing, stop angiogenesis signals, deliver cell-killing substances to cancer cells, and cause cancer cell death. As classic chemotherapy drugs, mAbs are commonly used alone or with other cytotoxic drugs or radioactive substances to kill cancer cells. Drugs designed for parenteral administration are aseptically prepared by pharmacists. Drugs are prepared just before use; they are reconstituted and/or diluted with 5% glucose or 0.9% sodium chloride to obtain a sterile final preparation at the dose prescribed by the physician. Even if pharmaceutical regulations do not require the final characterization of each compounding drug, many pharmacists have nevertheless implemented analytical control before release. These controls must be discriminant in order to ensure the nature of the drug in spite of similar physicochemical and spectral properties, sensitive enough to guarantee the dose even for low concentrations and also rapid to endure the medication process without delayed drug delivery. As a result, different analytical strategies using direct flow injection analysis, high performance liquid chromatography with UV detection **[1]**, or vibrational molecular spectroscopies such as Raman and infrared **[2,3]** have been investigated to control cytotoxic compounding drugs. Despite their importance in cancer therapy, only some studies have dealt with analytical methods to control mAbs preparations **[4–8]**. All of these techniques were invasive and required sampling the preparation for analysis. Because of the inherent toxicity of cytotoxic drugs resulting from their oncogenic, mutagenic and teratogenic properties, these drugs present a risk of occupational exposure for healthcare workers. This explains why we decided to explore the feasibility of a noninvasive, nondestructive, and rapid analytical method, i.e. Raman spectroscopy, to ensure the quality of the mAbs preparations produced in hospitals. Raman spectroscopy is a molecular vibrational spectroscopy based on the inelastic scattering of monochromatic light, usually from a laser in the visible, near infrared, or near ultraviolet

range, and in the past, Raman spectroscopy has been successfully investigated for chemotherapy drug control **[2,3,9,10]**.

Due to the complexity of Raman spectral data, multivariate analysis must be used to extract pertinent information. In this article, we report results using two different approaches of data analyses. Due to unsatisfactory results at the levels of both classification of the molecule and concentration regression using linear chemometrics techniques traditionally used in pharmaceutical field, we submitted the data to a collaborative data challenge platform called Rapid Analytics and Model Prototyping (RAMP), developed by the Paris-Saclay Center of Data Science to explore a different approach.

2. Material and methods

2.1. Preparation of mAbs samples

We evaluated four mAbs: Bevacizumab (Avastin® 25 mg/mL, Roche), Infliximab (Remicade® 100 mg, Schering-Plough), Ramucirumab (Cyramza® 10 mg/mL, Lilly) and Rituximab (Mabthera® 10 mg/mL, Roche). All drugs were prepared separately in aseptic conditions and analyzed them after dilution in 0.9% sodium chloride at various concentrations that cover the therapeutic range currently prepared to treat patients (10 concentrations for each drug: Bevacizumab from 0.5 to 25 mg/mL, Infliximab from 0.3 to 10 mg/mL, Ramucirumab from 1 to 10 mg/mL and Rituximab from 0.4 to 10 mg/mL). Twelve independent series were prepared (one batch of 0.9% chloride sodium per serie) for each drug. All compounded solutions were packaged in three glass vials (Interchim®, Montluçon, France), stored at +4 °C, and analyzed in accordance with laboratory requirements. The following abbreviations are used in this article: A = Infliximab, B = Bevacizumab; Q = Ramucirumab, R = Rituximab.

2.2. Spectral collection

Raman spectra were acquired with a Labram HR Evolution microspectrometer (Horiba Scientific, Lille, France). The excitation source was a 633 nm single-mode diode laser (Toptica Photonics, Germany) generating 35mW on the sample. The microspectrometer was equipped with an Olympus microscope and measurements were recorded using 10 X MPlan objective (Olympus, Japan). Light scattered by the sample was collected through the same

objective. Rayleigh elastic scattering was intercepted by an edge filter. A Peltier cooled (-70°C) multichannel CCD detector (charge-coupled device; 1024 x 256 pixels) detected the Raman Stokes signal dispersed with a 100 µm slit width and a 600 grooves/mm holographic grating. The spectral resolution from the full width at half maximum of the silica wafer band at 521 cm^{-1} was 2 cm^{-1} and the spectral region examined was 400-4000 cm^{-1}. The acquisition time of each spectrum was 2 x 15 s per collected spectrum to obtain better sensitivity. Spectral acquisition and data preprocessing were conducted with Labspec6 software (Horiba Scientific, Lille, France). The microspectrometer sample compartment was not suited for vials and so a vial adaptor was constructed to center the vial and to secure sample position on the base plate.

2.3. Preprocessing

The quality of predictive analysis increases with the number of input data, and so a large number of acquisitions were performed. Different sources of variabilities observed in real life (batch of the drug substance, batch of the diluent, container) were included to the dataset. A total of 360 spectra for each of the four mAbs were collected, except for Ramucirumab (348 spectra). All Raman spectra were pre-processed using LabSpec6 by normalization based on total area in order to correct spectral variation due to focusing variations. The result of this collection step was a spectrum containing 1866 values. During the preprocessing phase, a principal component analysis was performed for the spectra of each drug using Matlab® R2011a software (Mathworks, Natick, United States), to identify spectral errors and discard erroneous spectra. Among all acquisitions, no spectrum was eliminated from a total of 1428 spectra analyzed.

2.4. Flowchart of the spectral analysis

The methodology of spectral analysis is given in **Fig.1**. Data were split to develop a predictive model and evaluate its performance on a held-out set of data. The 1428 measurements were randomly divided into two datasets. A total of 999 spectra (calibration set) were used to developed predictive models whereas 429 other spectra constituting the test set were used to evaluate the predictive performances of the classification and the regression models

developed. All results reported here were from this test set, but they were unknown to the analysts throughout development.

In order to predict the drug and its concentration, two different methods were explored, a linear approach using chemometrics and a machine learning approach. For both approaches, the remaining 999 spectra were further split into random cross-validation folds. Chemometrics analyses were performed using 10-fold cross-validation whereas machine learning was based on cross-validation bootstrap aggregation (CV bagging). Using CV bagging, eight random cross-validation folds were considered. Each training and validation set contained 799 and 200 spectra, respectively. Each technique was trained eight times on the training sets. The eight trained models were then evaluated on each validation and test point. The final prediction is the average of these eight predictions on each validation and test point. These techniques take advantage of the variance reduction property of averaging and allowed to develop robust predictors despite the small data size. To further improve the machine learning results, the best models were blended using the technique of Caruana et al.**[11]**. Briefly, the pointwise mean prediction of the best models selected in a greedy loop was used until the validation result stops improving.

As shown in **Fig.1**, the workflow was slept into four modules. The first feature extractor g^{clf} converts the raw spectra **s** into a fixed-size feature vector $x^{clf} = g^{clf}(s)$. This is then followed by a classifier f^{clf} that outputs a vector $p = (p^A, p^B, p^Q, p^R) = f^{clf}(x^{clf})$**,** indexed by the four drug types, representing the estimated probability that the spectrum belongs to each of the four types. The predicted class is then:

$$\hat{l} = \arg\max p^j \text{ , avec } j \in \{\text{A,B,Q,R}\}.$$

The second feature extractor g^{reg} also receives the raw spectra **s** but also the probability vector $= f^{clf}(g^{clf}(s))$ and converts them into a fixed-size feature vector $x^{reg} = g^{reg}(s, c)$. The rationale is to allow the regression algorithm to know the (estimated) identity of the molecule whose concentration it has to predict. The final module is a regression model f^{reg} that uses this second feature vector x^{reg} and estimates the drug concentration:

$$\hat{y} = f^{reg}(x^{reg}) .$$

Given a validation or test set $D = \{(s_i, l_i, y_i)\}_{i=1}^{n}$ containing triplets of spectra, molecule types, and concentrations, the performance of the model $M = (g^{clf}, f^{clf}, g^{reg}, f^{reg})$ was evaluated as follows. We first computed the classification error (R^{clf}):

$$R^{clf}(M, D) = \frac{1}{n} \sum_{i=1}^{n} \text{II}\{ l_i \neq \hat{l}_i \},$$

where the indicator function II {X} is 1 if its argument X is true, and 0 if not. To assess the concentration predictor, the mean absolute relative error (R^{reg}) was used:

$$R^{reg}(M, D) = \frac{1}{n} \sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{y_i}.$$

In order to compare predictive models, combined errors (R^{comb}) were calculated:

$$R^{comb}(M, D) = \frac{2}{3} R^{clf}(M, D) + \frac{1}{3} R^{reg}(M, D)$$

The coefficients reflect the more stringent requirements posed on the classification task. The best model was considered for the lowest R^{comb}.

2.5. Conventional linear approach

Analyses were performed using Matlab® software R2011a. The first approach was based on linear discrimination and regression methods, common in traditional chemometry. Both preprocessing g^{clf}, g^{reg} and the predictor functions f^{clf}, , f^{reg} were linear. Classification was performed using partial least squares discriminant analysis (PLS-DA) and principal component analysis discriminant analysis (PCA-DA). PLS regression for was used to estimate concentrations. Various spectral processing techniques were tested for both classification and regression in order to limit the non-informative spectral background: baseline correction with a n^{th} degree polynomial curve, first and second Savitsky-Golay derivatives, standard normal variate (SNV), and combined preprocessing. For each model, the optimal number of latent variables was determined by 10-fold cross-validation for the lowest error of prediction. The predictive capacity of the calibration model was assessed by the root mean square error of cross-validation (RMSECV) and the root mean square error of prediction (RMSEP). The optimal predictive model was selected for the lowest RMSECV and RMSEP with the highest coefficient of determination (R^2).

2.6. Crowdsourced machine learning approach

In order to improve classification and regression prediction, a second approach based on machine learning was explored. Machine learning analyses were performed with the participation of the Paris-Saclay Center for Data Science (CDS) using the scikit-learn library **[12]**, which most of our analysts used.

Over the past several years, the CDS has developed a unique collaborative studio which allows a large number of data scientists to work together on various scientific and industrial data analytics workflows. Participants have access to a brief description, the public training dataset, and a first (non-optimized) solution in a starting kit. They study the problem, develop and refine solutions, and submit them through a web interface. The models are then trained and evaluated, and the validation performance score is fed back to the participants through a public leaderboard. The test score obtained on the held-out test set, reported in this paper, remains hidden from the participants. Unlike most of the data challenge sites that expect solutions to analytics problems, the CDS asks the data scientists to submit code. This provides more flexibility to the evaluation, generates a working prototype, and most importantly, enables participants to collaborate by examining each other's code and combining the different ideas. The code is publicly available on the website (after free signup), making the results fully reproducible by anyone. Moreover, any new method can be easily tested by submitting it to the site and so our results are fully transparent.

The classification and regression challenge proposed in this study was submitted to more than 300 students and data scientists in three different RAMP classrooms at Paris-Saclay University, Polytechnique School, and Mines ParisTech. Each RAMP involved about five days of work. In each class, the challenge started with a closed period when students did not see each other's code, only their scores. This allowed us to grade the students individually but also to make them explore the space of solutions independently. This was followed by an open collaborative period. A given class, however, could not see the solutions developed by the other classes. Interestingly, this led to different collective strategies to solve the problem. Models were blended in each RAMP separately which made it possible to compare the collaborative score of each group. During RAMPs, the analysts were instructed to minimize a combined error R^{comb}. All models and codes are available freely on the RAMP site (http://ramp.studio).

3. Results and discussion

Mean raw spectra for the four mAbs analyzed are shown in **Fig.2**. Raman spectra of mAbs, similarly to those of other proteins, are very difficult to interpret. Inter- and intramolecular effects including peptide bond angles and hydrogen-bonding patterns may influence Raman band positions. Structural information can nevertheless be deduced from Raman vibrational bands as amide I and amide III bands at 1650 and 1300 cm^{-1}.

The results of classification and regression obtained using both the linear approach and machine learning are listed in **Table 1**. The overall winner of the contest was the Mines RAMP with a combined error R^{comb} of 2.4%, a classification error R^{clf} of 0.7% and a concentration error R^{reg} of 5.8%. As shown in **Table 1**, the large improvement obtained with machine learning techniques is uniform for all molecules. The best improvement was obtained for Bevacizumab with an 88.3% reduction on combined errors (R^{comb} of 2.1% *versus* 17.9%).

A total of 96 models of classification were developed with the linear approach using PCA-DA and PLS-DA. For PLS-DA, the assignation criterion based on the Bayes theorem was applied in order to minimize the number of false positives and false negatives explaining that some samples were not assigned. Despite optimization, a maximum of 367 samples among the 429 samples of the test sets were correctly classified. The best prediction obtained by mean centering PLS-DA on first derivate spectra with five latent variables was characterized by a classification error of 14.5%. Using this model, 9.6% samples of the test set (n = 41) were not assigned and 4.9% were misclassified (n = 21). Using machine learning, all samples were assigned and only three spectra were misclassified, all from samples with low concentrations ($y \leq 0.6$mg/mL).

Concerning the regression challenge aiming to predict the concentration of the mAbs in solution, the results showed a better prediction for machine learning than with the linear approach, 14.7% *versus* 5.8% of concentration error. Details of the PLS regression models for the four mAbs are in **Appendix 1**. Using the linear approach, 63% of samples at a low concentration ($y \leq 1$ mg/mL) had a relative error more than 15% whereas an error of only 5.6% was reported in the case of machine learning. As shown in **Fig.3**, relative errors over all are higher with the linear approach, but as in the case of machine learning analysis, they depend on concentration. In addition, bias with the machine learning model is negligible at

concentrations higher than 0.1 mg/mL and relative errors decrease sharply with the concentration from 15% (below y < 1 mg/mL) to 5% (above y > 3 mg/mL).

Throughout RAMP sessions, the predictive performances of models progressed considerably. During the first RAMP performed at Paris-Saclay University, numerous preprocessing operations were evaluated. This session reached a classification error of 1.6% and a concentration error of 4.9% (R^{comb} of 2.7%). Spectra were smoothed with linear filters, such as the Savitzky–Golay filter **[13]**, that preserves peak amplitudes of the signal. Other models use simpler strategies based on convolution with Hanning windows. Following the smoothing step, baseline correction was conducted by subtracting a polynomial least-square fit of the data or simply by removing a constant. The polynomial order was 0 (constant) or 1 (linear). Order 0 corresponds to the subtraction of the mean of the spectra. Some of the solutions propose subtracting the median of the spectra. As the machine learning models employed in the second step are sensitive to the scale of the data, the spectra for some models were normalized by smoothing and baseline correction. The majority of winning predictions solutions used nonlinear kernel-based techniques: kernel PCA for non-linear dimensionality reduction **[14]** and support vector machines (SVM) for prediction **[15]**. The best approaches proposed used Gaussian kernels and polynomial kernels of order up to 4. The low concentration error was achieved by exploiting the fact that concentrations were discretized.

During the second RAMP performed at Polytechnique, the group explored only forest-based regression models (extra trees **[16]**, random forests **[17]**, gradient boosting **[18]**). This session reached a classification error of 2.1% and a concentration error of 12.2%. These otherwise popular and usually well-performing nonparametric classifiers seemed to be a suboptimal choice for the functional data of this problem.

As mentioned above, the Mines RAMP (the last one) was the dominant solution with a classification error of 0.7% and a concentration error of 5.8% corresponding to the lowest R^{comb} (2.4%). The optimal model that the group converged on included a log transformation of the spectra (but no other smoothing or preprocessing), followed by either factor analysis or principal component analysis to extract 10 features, followed by a small neural network. Interestingly, the same pipeline was successful for both classification and regression, but of course with different parameter settings of the neural net. In the regression step, all the best models used the same strategy: to learn a different regression model (although parameters set

the same way) for each molecule. According to this model, only three low-concentration test samples were missed.

The interest of statistical machine learning for biological applications has grown considerably even if this approach is not now applicable in routine in hospital pharmacies. As part of both computer science and statistics, machine learning as a scientific discipline aims to develop algorithms that can make sense of data. A typical outcome of machine learning is a model that can make predictions from new data after having learned from many training examples. Based on the hypothesis that spectra are influenced by the nature and concentration of mAbs in solution, we decided to explore how machine learning can help resolve the pharmaceutical challenge of classification and concentration estimation for pharmaceutical drugs particularly difficult to discriminate by classical linear methods. We found that combining Raman spectroscopy with machine learning methods presents an interesting potential to augment safety of the drug preparation process by the identification and quantification of chemotherapy preparations.

As opposed to traditional analysis using HPLC/UV or LC/MS/MS, Raman spectroscopy and near infrared spectroscopy enable direct measurement through glass and plastic packaging **[19,20]**. As a result of the rapidity of analysis and non-destructive and non-invasive measurements, these spectroscopies are widely used for Process Analytical Technology **[21]** with at-line and on-line measurements to control primary and secondary manufacturing processes in the pharmaceutical industry.

Despite promising results for ensuring the nature and dose of a drug in solution by classification and regression analysis, respectively, machine learning has some particular limitations and pitfalls that should be avoided. Machine learning algorithms usually require more input data to train than linear models. In addition, care must be taken to collect data with the same distribution as when the predictors are used in practice. First, if the training concentrations are over- or under-sampled in certain regions, the nonparametric concentration estimate may be biased towards the oversampled values. This is due to the fact that machine learning models attempt to minimize the average errors. Secondly, if only certain concentration levels are used for some molecules, the classifier can learn these values and use the concentration information for classification. Thirdly, if the concentration levels are discretized, the regression model can learn these discretization levels and quantify the

continuous estimate, creating a bias if the true concentrations fall between those levels. In fact, one of the student groups boosted prediction accuracy by forcing the model to predict only concentrations present in the training set. More concretely, they used regression model to round the predicted concentration to the closest discrete concentration. In our study, the predictors were so precise that these effects were negligible, but in more complicated measurements with larger uncertainties they must be taken into consideration. One practical way to consider this in future data collection protocols in order to prevent performance gains by rounding, is to randomize concentration values using a sampling distribution that matches realistic scenarios.

Extracting useful information from complex data is crucial for developing the best predictive models for both classification and regression. With the development of data science, the choice of feature extractors and filters has considerably increased. Selecting and tuning the best predictor, however, requires experience and many experiments, but scientists generally apply only several panels of algorithms to analyze their data. Despite promising results for classification and quantification of other antineoplastic drugs, the conventional approach using linear chemometrics methods was not sufficient in the case of mAbs. RAMP sessions organized by the CDS was the opportunity to explore a new approach to analyze our data. Throughout RAMPs, many scientists from different fields such as chemistry, mathematics, informatics, astronomy, computer games and the financial sphere participated. They tested new approaches, codes or combinations of known algorithms that we never thought to explore and contributed to considerably optimizing predictions and met the challenge.

4. Conclusions

As a result of similar structures and low concentration values, discrimination and quantification of mAbs preparations pose difficult challenges. Despite these models are not directly applicable in routine, this study highlighted the power of collaborative approach to solve a problem and in our case, the interest of statistical machine learning to interpret spectral Raman data to ensure the chemical quality of medications produced in hospitals before administration to patients.

References

[1] A. Delmas, J.B. Gordien, J.M. Bernadou, M. Roudaut, A. Gresser, L. Malki, M.C. Saux, D. Breilh, Quantitative and qualitative control of cytotoxic preparations by HPLC-UV in a centralized parenteral preparations unit, J. Pharm. Biomed. Anal. 49 (2009) 1213–1220. doi:10.1016/j.jpba.2009.03.007.

[2] L.M.M. Lê, A. Tfayli, J. Zhou, P. Prognon, A. Baillet-Guffroy, E. Caudron, Discrimination and quantification of two isomeric antineoplastic drugs by rapid and non-invasive analytical control using a handheld Raman spectrometer, Talanta. 161 (2016) 320–324. doi:10.1016/j.talanta.2016.07.025.

[3] P. Bourget, A. Amin, F. Vidal, C. Merlette, P. Troude, A. Baillet-Guffroy, The contribution of Raman spectroscopy to the analytical quality control of cytotoxic drugs in a hospital environment: Eliminating the exposure risks for staff members and their work environment, Int. J. Pharm. (2014). doi:10.1016/j.ijpharm.2014.04.064.

[4] C. Bazin, V. Vieillard, A. Astier, M. Paul, [Reliable real-time analytical control of monoclonal antibodies chemotherapies preparations on Multispec automaton], Ann. Pharm. Fr. 68 (2010) 163–177. doi:10.1016/j.pharma.2010.03.008.

[5] E. Jaccoulet, C. Smadja, M. Taverna, Quality Control of Therapeutic Monoclonal Antibodies at the Hospital After Their Compounding and Before Their Administration to Patients, Methods Mol. Biol. Clifton NJ. 1466 (2016) 179–184. doi:10.1007/978-1-4939-4014-1_14.

[6] E. Jaccoulet, J. Boccard, M. Taverna, A.S. Azevedos, S. Rudaz, C. Smadja, High-throughput identification of monoclonal antibodies after compounding by UV spectroscopy coupled to chemometrics analysis, Anal. Bioanal. Chem. 408 (2016) 5915–5924. doi:10.1007/s00216-016-9708-4.

[7] L. Ashton, Y. Xu, V.L. Brewster, D.P. Cowcher, C.A. Sellick, A.J. Dickson, G.M. Stephens, R. Goodacre, The challenge of applying Raman spectroscopy to monitor recombinant antibody production, The Analyst. 138 (2013) 6977–6985. doi:10.1039/c3an01341c.

[8] M. Paul, V. Vieillard, E. Jaccoulet, A. Astier, Long-term stability of diluted solutions of the monoclonal antibody rituximab, Int. J. Pharm. 436 (2012) 282–290. doi:10.1016/j.ijpharm.2012.06.063.

[9] L.M.M. Lê, E. Caudron, A. Baillet-Guffroy, L. Eveleigh, Non-invasive quantification of 5 fluorouracil and gemcitabine in aqueous matrix by direct measurement through glass vials using near-infrared spectroscopy, Talanta. 119 (2014) 361–366. doi:10.1016/j.talanta.2013.10.060.

[10] A. Amin, P. Bourget, F. Vidal, F. Ader, Routine application of Raman spectroscopy in the quality control of hospital compounded ganciclovir, Int. J. Pharm. 474 (2014) 193–201. doi:10.1016/j.ijpharm.2014.08.028.

[11] R. Caruana, A. Niculescu-Mizil, G. Crew, A. Ksikes, Ensemble Selection from Libraries of Models, in: Proc. Twenty-First Int. Conf. Mach. Learn., ACM, New York, NY, USA, 2004: p. 18–. doi:10.1145/1015330.1015432.

[12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: Machine Learning in Python, J Mach Learn Res. 12 (2011) 2825–2830.

[13] R. Schafer, What Is a Savitzky-Golay Filter? [Lecture Notes], IEEE Signal Process. Mag. 28 (2011) 111–117. doi:10.1109/MSP.2011.941097.

[14] B. Schölkopf, A. Smola, K. Müller, Nonlinear component analysis as a kernel eigenvalue problem, Neural Comput. 10 (1998) 1299–1319.

[15] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, Cambridge University Press, Cambridge ; New York, 2000.

[16] P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees, Mach. Learn. 63 (2006) 3–42. doi:10.1007/s10994-006-6226-1.

[17] L. Breiman, Random Forests, Mach. Learn. 45 (2001) 5–32. doi:10.1023/A:1010933404324.

[18] J.H. Friedman, Greedy Function Approximation: A Gradient Boosting Machine, Ann. Stat. 29 (2001) 1189–1232.

[19] N.W. Broad, R.D. Jee, A.C. Moffat, M.J. Eaves, W.C. Mann, W. Dziki, Non-invasive determination of ethanol, propylene glycol and water in a multi-component pharmaceutical oral liquid by direct measurement through amber plastic bottles using Fourier transform near-infrared spectroscopy, The Analyst. 125 (2000) 2054–2058.

[20] G.J. Vergote, C. Vervaet, J.P. Remon, T. Haemers, F. Verpoort, Near-infrared FT-Raman spectroscopy as a rapid analytical tool for the determination of diltiazem hydrochloride in tablets, Eur. J. Pharm. Sci. Off. J. Eur. Fed. Pharm. Sci. 16 (2002) 63–67.

[21] T. De Beer, A. Burggraeve, M. Fonteyne, L. Saerens, J.P. Remon, C. Vervaet, Near infrared and Raman spectroscopy for the in-process monitoring of pharmaceutical production processes, Int. J. Pharm. 417 (2011) 32–47. doi:10.1016/j.ijpharm.2010.12.012.

List of figures and tables

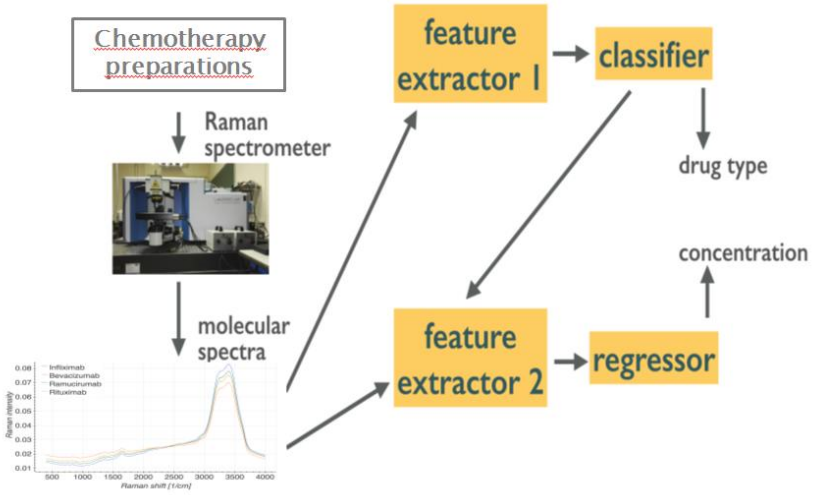


Fig.1. The data acquisition and data analytics pipeline.

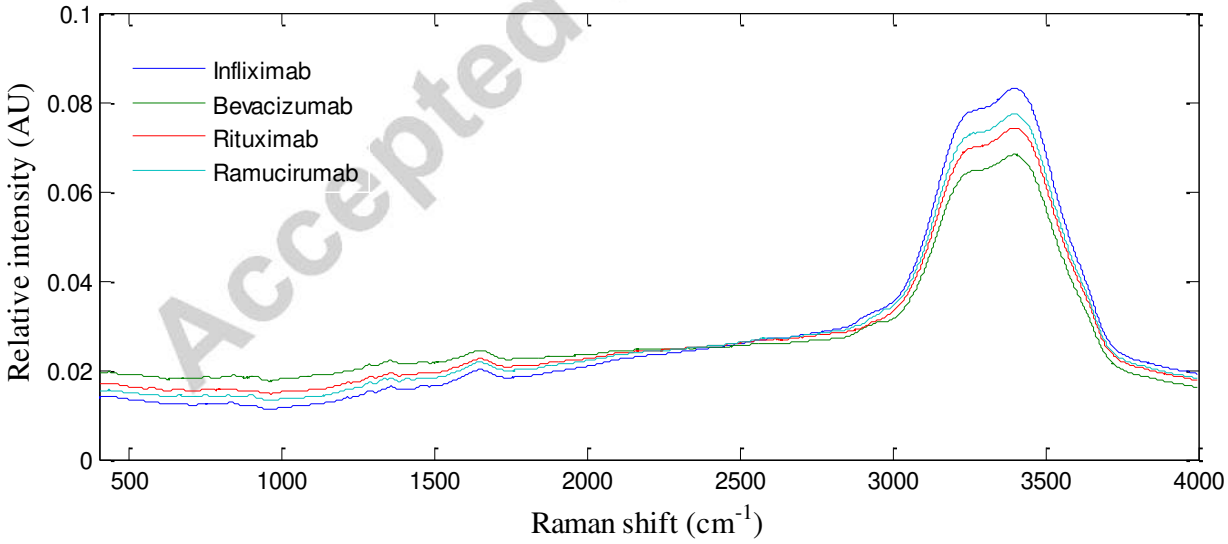
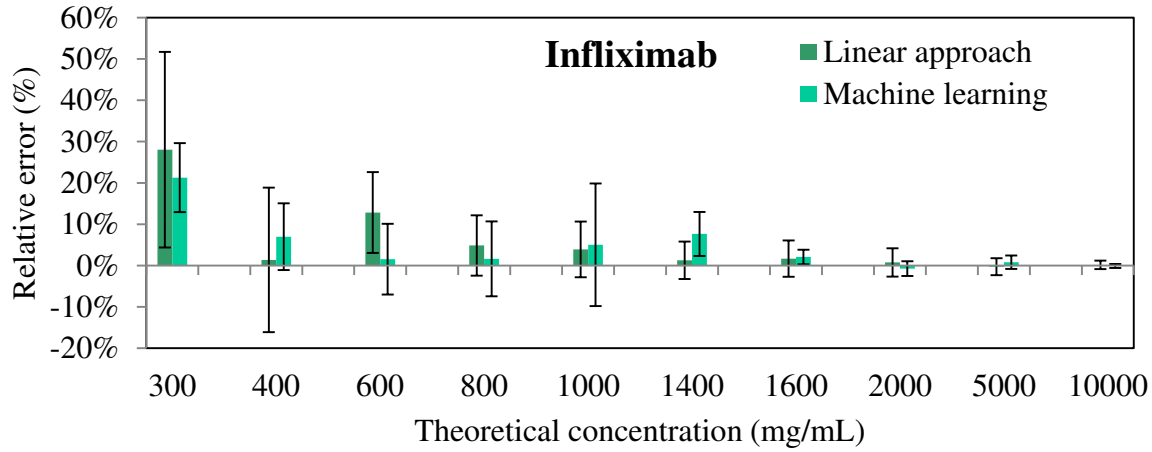
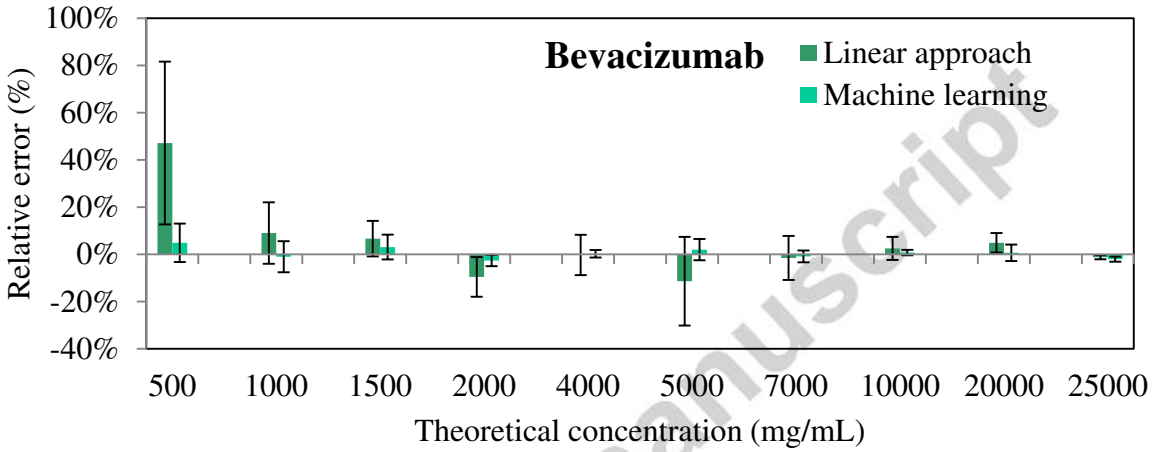
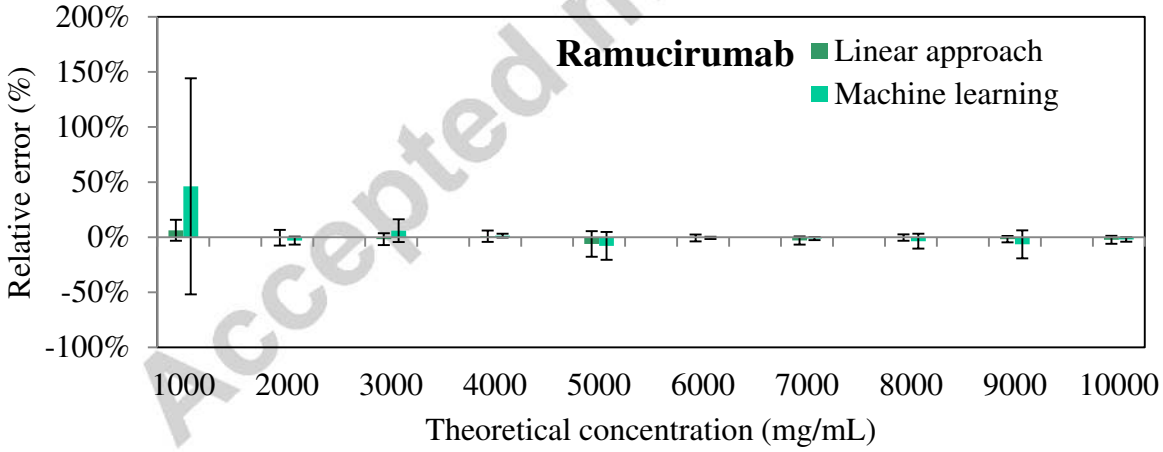


Fig.2. Mean raw Raman spectra for Infliximab, Bevacizumab, Ramucirumab and Rituximab from 400 to 4000 cm^{-1}







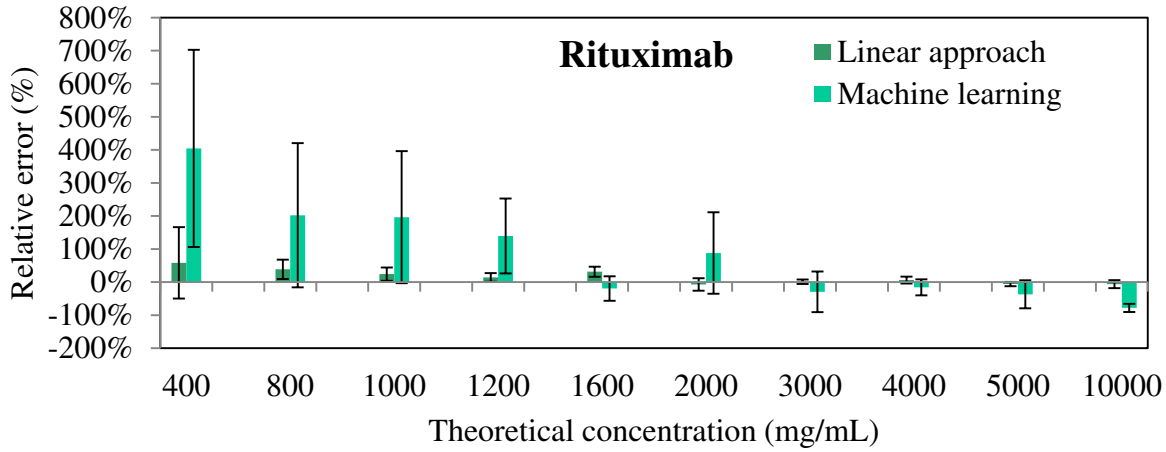


Fig.3. Relative error of the predicted concentrations for the four monoclonal antibodies obtained using the linear and machine learning approaches

Table 1 Misclassification errors ($\boldsymbol{R^{clf}}$), mean absolute relative errors of concentration ($\boldsymbol{R^{reg}}$) and combined errors ($\boldsymbol{R^{comb}}$) obtained for the prediction of the test set samples using the best predictive models developed by machine learning and chemometrics linear approaches. (Errors in percent of all predicted samples, %)

	Linear			Machine learning		
Molecule	R^{clf}	R^{reg}	$\boldsymbol{R^{comb}}$	R^{clf}	R^{reg}	$\boldsymbol{R^{comb}}$
Infliximab	13.7	12.3	**13.2**	0.9	8.4	**3.4**
Bevacizumab	19.8	14.0	**17.9**	1.0	4.3	**2.1**
Ramucirumab	9.0	7.3	**8.4**	0.0	3.5	**1.2**
Rituximab	16.0	26.7	**19.6**	1.0	6.9	**3.0**
Overall	14.5	14.7	**14.6**	0.7	5.8	**2.4**

Highlights

- A innovative and rapid analytical method is pertinent to discriminate and quantify monoclonal antibody drugs
-
- Interest of machine learning for classification and regression of anticancer drugs
- Interest of collaborative data analysis to optimize prediction