High accuracy discriminant analysis of preparation from 22 commercial antibodies based on UV spectroscopy combined with chemometrics

Antoine Dowek, Maxime Annereau, Lucas Denis, Thomas Fleury, Emmanuel Daguet, Hail Aboudagga, André Rieutord, Eric Caudron, Laetitia Lê

Highlights

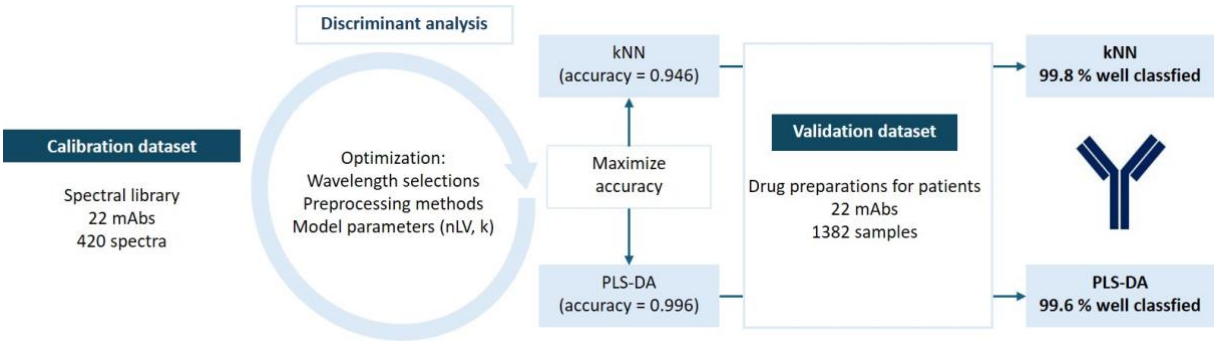
- Comprehensive analysis of 22 therapeutic monoclonal antibodies, covering the main commercial antibodies commonly used in therapeutic applications.
- Utilization of both linear and non-linear discriminant analyzed models.
- Validation dataset from real-life production, ensuring the robustness of the findings.
- Rapid analysis methodology demonstrating excellent accuracy performance.

High accuracy discriminant analysis of preparation from 22 commercial antibodies based on UV spectroscopy combined with chemometrics

Antoine Dowek[1,2*], Maxime Annereau[1,3], Lucas Denis [1,4], Thomas Fleury[1], Emmanuel Daguet[1], Hail Aboudagga[1], André Rieutord[1], Eric Caudron[2,5], Laetitia Lê[2,5]

1- Clinical Pharmacy Department, Gustave Roussy Cancer Campus, 94800 Villejuif, France
2- Lipides, Systèmes Analytiques et Biologiques, Université Paris-Saclay, 91400, Orsay, France
3- Université Paris-Saclay, Matériaux et santé, 91400, Université Paris-Saclay, Orsay, France
4- Université Paris-Saclay, CNRS, Institut Galien Paris-Saclay, 91400, Orsay, France.
5- Service de Pharmacie, Hôpital Européen Georges Pompidou, APHP. Centre -Université Paris Cité, 20 Rue Leblanc, 75015, Paris, France
*Corresponding author

Graphical abstract



Abstract

Keywords: Chemometrics, discriminant analysis, monoclonal antibodies, ultraviolet spectroscopy

The demand for monoclonal antibodies (mAbs) in hospitals is continually rising due to the advancement of targeted therapy or immunotherapy in cancer treatment. Typically, mAbs are produced in centralized chemotherapy production units and require rigorous control to ensure the right drug at the right dose before administration to patients. Discriminant analysis of mAbs is particularly challenging due to their closely related chemical and spectral properties. Various techniques, including flow injection analysis, electrophoresis and spectroscopies such as ultraviolet (UV), infrared or Raman have been attempted with limited success, particularly on a small subset of mAbs, underscoring the difficulty in discriminating them.
This study aimed to develop a robust discriminant analysis method using a large dataset comprising 22 mAbs used in therapeutic. The discriminant analysis was based on UV spectroscopy coupled with chemometrics. A calibration dataset consisting of 404 samples was prepared and analyzed using various preprocessing methods, wavelength selection and discriminant analysis algorithms to maximize overall accuracy. Validation of the newly developed discriminant analysis was conducted using samples collected from real manufactured preparations for patient (over 1,380 samples). The study achieved an overall accuracy of 0.996 and 0.998 on the 22 mAbs using partial least squares-discriminant analysis (PLS-DA) and k-nearest neighbors (kNN), respectively. The utilization of chemometrics for discriminating mAbs based on their UV spectra, following total area normalization, has been successfully demonstrated. This method has undergone validation, affirming its capability to swiftly and dependably distinguish different mAbs. Moreover, it seamlessly integrates with the workflow involved in the preparation of mAbs drugs within a cancer hospital setting. This advancement holds promise for streamlining processes and enhancing efficiency in the production and utilization of mAbs, thereby potentially benefiting patient care and treatment outcomes.

1. Introduction

Monoclonal antibodies (mAbs) play a major role in cancer therapy, they are immune system proteins that constitute a main component of cancer therapy [1]. These advancements have significantly enhanced overall survival rates across various cancers [1–5] leading to a surge of the number of commercial mAbs drugs. The preparation of each mAb involves one or two major steps, depending on whether it is in ready-to-used form or in powdered form, and requires

reconstitution in a solvent before the dilution (commonly 5% glucose or 0.9% sodium chloride). They are produced in specific production unit by pharmacy technicians under controlled conditions[6–8]. Various analytical strategies have been explored for analysis of mAbs, including electrophoresis [9] and flow injection analysis [10], both coupled with UV-detection. Direct analysis using Raman spectroscopy [11,12] have also been conducted but is not currently applicable and involves machine learning approach [12]. UV spectroscopy approach, up to now, has limited success and has only been effective in discriminating a few mAbs [10]. At Gustave Roussy cancer center, over 20 mAbs are produced. To perform analytical quality control (AQC) for all anticancer drug production, a system called QCRx (Icone), based on UV and Raman spectroscopies is currently employed. This system seamlessly integrates with the drug preparation workflow, and the analysis takes no longer than two minutes. However, given that discriminant analysis using Raman spectroscopy primarily recognizes the excipient rather than the mAbs directly [11], we have opted to use UV spectra for this study. Despite limitations in specificity, UV spectroscopy may still be suitable for mAbs discrimination due to inherent properties of mAbs which are large proteins rich in aromatic amino acids, which strongly absorb UV light in the range of 250 – 300 nm. In contrast, common excipients in mAb formulation, such as surfactants (polysorbates, poloxamers), amino acids for buffering capacity (glycine, arginine, histidine) and carbohydrates (sucrose, mannitol) typically absorb below 230 nm [13]. The aim of this work was to develop an accurate and robust discriminant analysis based on the UV spectra of 22 mAbs using chememotric tools.

2. Materials and methods

2.1.Materials

The manufactured mAbs included atezolumab (Tecentriq, Roche), avelumab (Bavencio, Merck Serono), bevacizumab (Mvasi, Amgen), rituximab (Ruxience, Pfizer), brentuximab-vedotin (Adcetris, Takeda), cemiplimab (Libtayo, Sanofi Winthrop), cetuximab (Erbitux, Merck Serono), daratumumab (Darzalex, Janssen Cilag), durvalumab (Imfinzi, Astrazeneca), enfortumab-vednotin (Padcev, Astellas Pharma), ipilimumab (Yervoy, Bristol Myers Squibb), isatuximab (Sarclisa, Sanofi Winthrop), nivolumab (Opdivo, Bristol Myers Squibb), obinutuzumab (Gazyvaro, Roche), panitumumab (Vectibix, Amgen), pembrolizumab (Keytruda, MSD), pertuzumab (Perjeta, Roche), sacituzumab-govotecan (Trodelvy, Gilead sciences), tafasitamab (Minjuvi, Incyte biosciences), trastuzumab (Trazimera, Pfizer), trastuzumab-emtansine (Kadcyla, Roche), trastuzumab-deruxtecan (Enherthu, Daiichi Sankyo). These mAbs were prepared according to the manufacturer recommendations (summary of product characteristics, SPC) and were mostly diluted in sodium chloride 0.9% (NaCl 0.9%, Fresenius Kaby). Adcteris, Padcev, Minjuvi, Kadcyla, Enhertu must be reconstituted with water for injection while Trodelvy must be reconstituted with 0.9 % NaCl. Adcteris, Padcev, Trodelvy, Kadcyla, Enhertu are conjugated mAbs, so they are linked to another molecule which is cytotoxic. These mAbs are expected to be more readily distinguishable from others non-conjugated mAbs.

2.2.Instrumentation

The UV analyses were performed using an i-QCRx system (B&W Tek Europe GmbH, Germany, Icones service, France). This system comprised an autosampler, fluidic control and main detection module. The detection module was equipped with a 0.5 mm quartz-cuvette. For

UV detection, a deuterium lamp was used with a diode array detector. The acquisition with QCRx system was controlled by a computer with QCRx-iC analyses software version 9.0.

2.3.Data analysis

An identification library was constructed with QL_Analyst software (version 1.0). Chemometric analyses were performed with R software, version 4.3.2 (2023-10-31). Preprocessing methods evaluated included normalization, asymmetric least square (ALS) baseline correction, Savitzky-Golay (SG) smoothing, first derivative (d1), and second derivative (d2) implemented with the baseline package (version 1.3-5). Discriminant analysis, including PLS-DA and kNN, was conducted using the Caret package, version 6.0-94 [14]. Partial least square regression for discriminant analysis (PLS-DA) were performed using "softmax" algorithm for probability calculation. Additionally, k-Nearest Neighbor (kNN) analysis was performed.
Principal component analysis (PCA) [15] was firstly applied, to quickly visualize all dataset by projecting data on principal components (PC) which are linear combinations of original variables that maximize the variance. Therefore, it is easier to visualize large dataset onto a lower dimensional space and to identify patterns.
PLS-DA is primarily used to enhance the discrimination between different sample groups by maximizing the covariance between the independent variables X (UV spectra) and the corresponding dependent variable Y (mAbs classes). This is achieved by identifying a linear subspace of the explanatory variables [16]. This new subspace enables the prediction of the Y variable based on a reduced number of factors, known as latent variables (LV). In this approach, it is important to define the number of LV to prevent overfitting, a process typically carried out using the training set to maximize accuracy. PLS-DA primary advantage lies in its ability to manage highly collinear and noisy data, characteristics often encountered in spectral measurements [17].
kNN is a simple supervised modelling technique that utilizes the nearest neighbor to classify objects. This algorithm operates by computing the Euclidean distances between the unknown spectra and the spectra of the training set. Subsequently, the k spectra with the shortest Euclidean distances are selected [18], forming a group of similar spectra. A centroid, calculated as a geometric mean, is then calculated from this group. When presented with a new unknown sample, its class attribution is determined by the distance from the centroid. In this approach, the number of neighbors (k) must be selected at the start of the algorithm. This parameter was optimized using the training set to maximize accuracy.

The reference spectral library was developed by measuring the 22 mAbs at various concentrations, following the recommendation outlined in international conferences of harmonization (ICH) for quantitative validation method [19]. The concentrations of molecules used to construct the calibration dataset are summarized in table S1 of the supplementary information (SI). Specifically, measurements were taken at a minimum of five concentrations with three concentrations in triplicate (*i.e.* about 15 to 20 samples by mAb, depending on analytical range). These spectra formed the basis of the calibration dataset, which was randomly divided into 70 % for the training set and 30 % for the testing set. During this step, wavelength selection, preprocessing optimization, discriminant analysis and selection of the number of latent variables (LV) or the number of neighbors (k) were performed to maximize accuracy. The number of LV and k were determined by repeated k-fold cross-validation on the test set of the calibration dataset with ten repetitions (r = 10). The validation dataset of independent samples, used to access the prediction performance of the model, comprised real preparations of mAbs intended for patient administration.

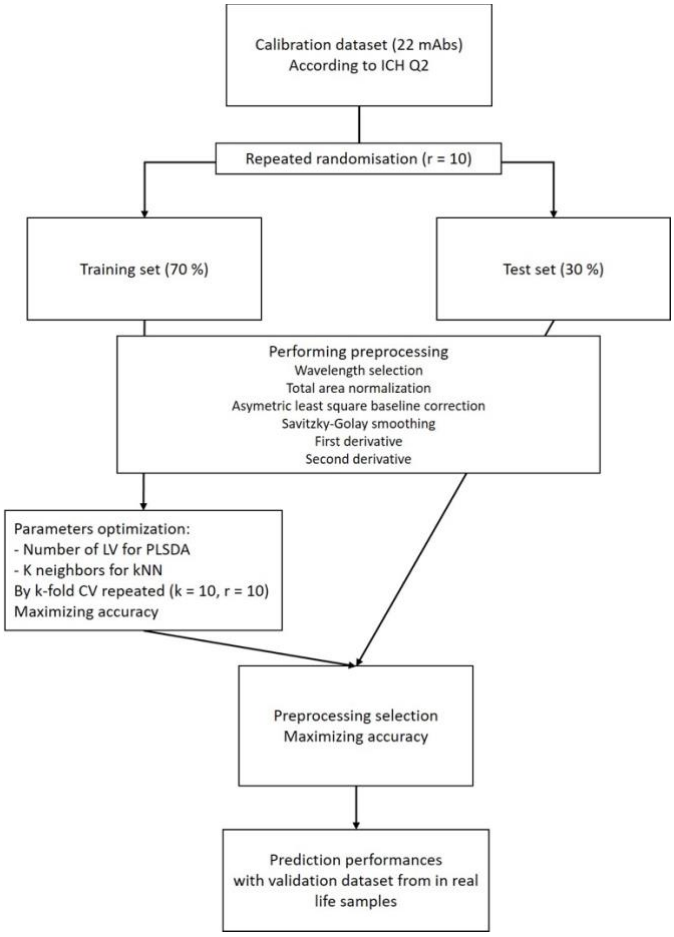


Figure 1: analytical strategy for discriminant analysis on monoclonal antibodies spectral data.

3. Result and discussion

The first step involved constructing a library containing spectra of all mAbs, covering the entire concentration range used in therapeutic. Chemometric analysis was then applied to highlight spectral differences. First, various preprocessing methods were tested to reduce noise background. Then, two widely used supervised discriminant analysis [20] were performed: partial least square-discriminant analysis (PLS-DA) [21] and k-nearest neighbor (kNN) [22]. PLS-DA is favored in spectral analysis due to its ability to handle collinearity structure, whereas kNN is capable to handling non-linear models. The main objective was to maximize accuracy representing the proportion of well-classified samples.
Once the best model selected with calibration dataset, the final discriminant analysis was performed on new data obtained from mAbs preparations intended for patient injection. Each sample was treated as unknown, and drug determination was carried out blindly.

A total of 404 spectra were collected to build the calibration dataset, including UV spectra of 22 mAbs at concentrations ranging from 0.1 $mg.mL^{-1}$ to 25 $mg.mL^{-1}$. The concentrations varied for each mAbs based on their therapeutic use (table S1).

The data were initially examined by PCA to access data variability. As shown in Figure *2* (A), significant signal saturation was observed in the region below 240 nm. Consequently,

wavelengths above 250 nm were selected for further analysis. Upon examination of the first two principal components (PCs) as illustrated in Figure *2*(B), it was evident that the concentration contributed substantially to the observed variability. To mitigate the influence of concentration, total area normalization was implemented. Subsequently, various preprocessing commonly used in spectral analysis were evaluated (ALS, SG, d1, d2). PC1 and PC2 visibly illustrated an initial distinction between mAbs and conjugated mAbs, which denote mAbs linked to a cytotoxic drug: tastuzumab deruxtecan and trastuzumab emtansine.

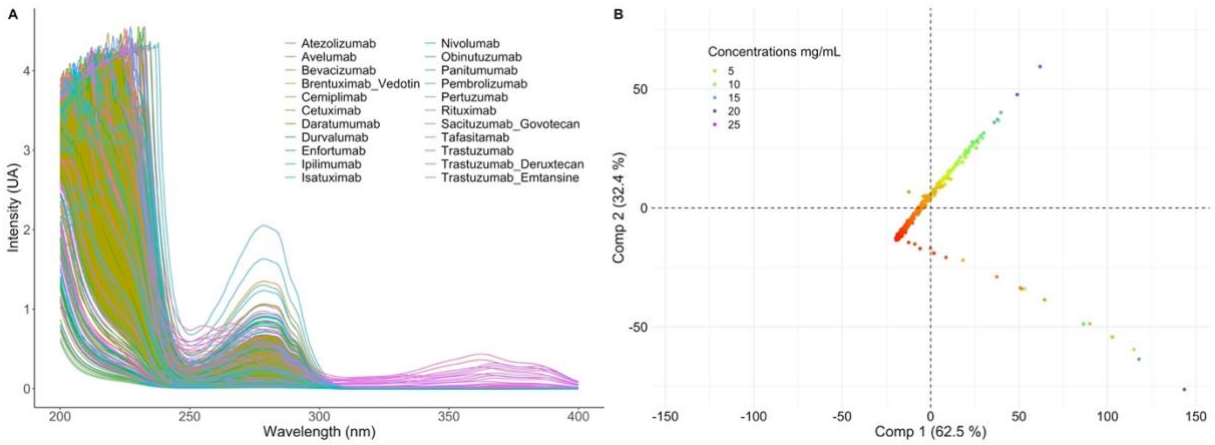


Figure 2: raw spectra of 404 measurements (A), two first principal components (Comp 1, Comp 2) of principal component analysis on raw spectra, colored by concentrations (B).

In Figure *3*A, the spectra selection is depicted following wavelength selection (250-400 nm), total area normalization and removal outliers based only on spectra structure and not on PCA calculation. A total of 17 spectra were removed due to errors encountered during the acquisition process.. A subsequent PCA was performed post-preprocessing (Figure *3* B) and shown a first separation between mAbs and all 5 conjugated mAbs. Applying PCA on preprocess data suggested that conjugated mAbs should be easy to discriminate while discriminate non-conjugated mAbs remind difficult.

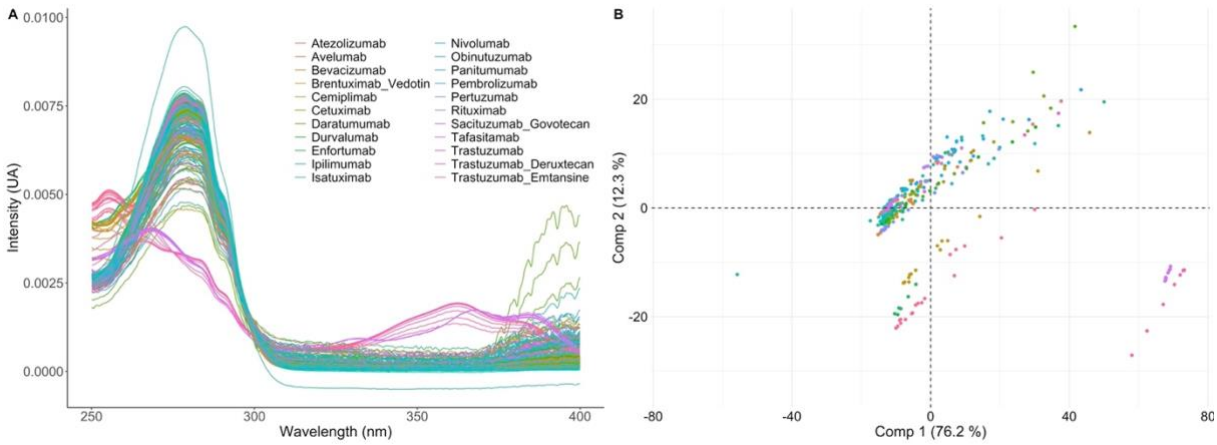


Figure 3: preprocess spectra (387) by wavelength selection (250-400 nm) and total area normalization (A), principal component analysis on preprocess data showing class separation between mAbs and conjugated mAbs (B).

Based on this groundwork, we performed data preprocessing which included the removal of outliers based on their spectral characteristics. Subsequently, 387 spectra were selected, and wavelength selection was set at the range of 250 nm to 400 nm. The next step involved optimizing model parameters, such as the number of LV for PLS-DA and the number of k for kNN. An illustration of parameter selection, conducted on the training set of the calibration dataset, is presented in figure 4.

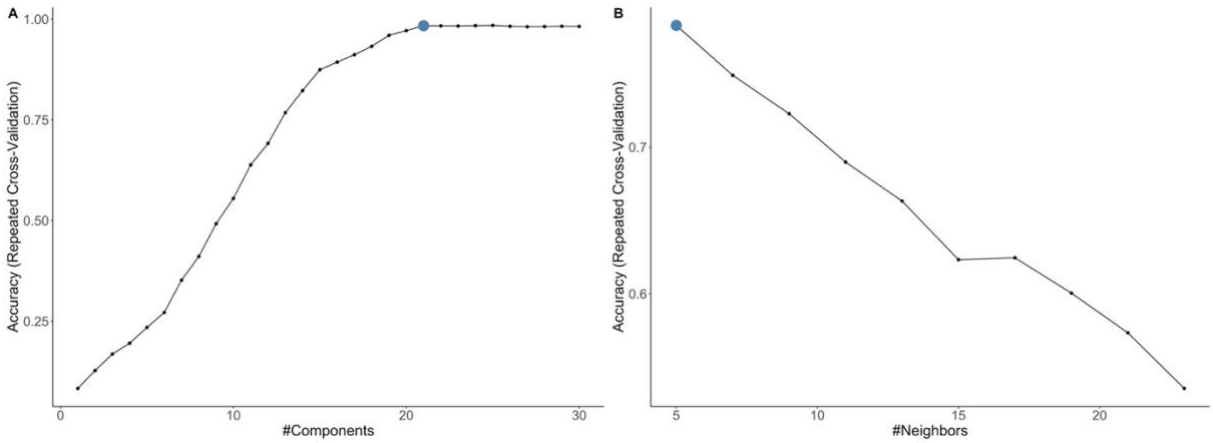


Figure 4: parameters selection for the number of latent variables (LV) for PLS-DA model (A), number of neighbors (k) for kNN model (B). The blue dot corresponds to the selection, LV = 21 for PLS-DA, k = 5 for kNN.

The main results for various preprocessing methods applied to PLS-DA or kNN algorithms are summarized in table S2. These results were derived from accuracy calculations on the test set of the calibration dataset, averaging across ten repetitions. The findings underscore the significance of wavelength selection and total area normalization in maximizing accuracy. While others preprocessing methods such as ALS, d1, and d2 had minimal impact on accuracy with PLS-DA model, they exhibited influence on the performance of the kNN model.

The results demonstrated superior performance with PLS-DA compared to kNN across all preprocessing methods. Upon comparing the performance of PLS-DA and kNN, we concluded that linear classification was highly effective in discriminating mAbs and likely better suited to handle collinearity issues compared to the kNN approach. Nonetheless, it is noteworthy that kNN also achieved high levels of accuracy.

In conclusion, the highest accuracy obtained with the calibration dataset was achieved using Narea, ALS, SG, and d1 preprocessing for PSL-DA, resulting in a mean accuracy of 0.973 CI95 [0.960 – 0.986]. For kNN, the Narea, ALS, SG, and d2 preprocessing methods yielding the best performance, with a mean accuracy of 0.907 CI95 [0.881 – 0.933]. Although there were no significant differences between the best result and Narea preprocessing for PLS-DA, it was ultimately chosen to retain Narea preprocessing for PLS-DA to conserve the structure of the spectra.

To ensure the proper select of wavelengths, we assessed the importance of each variable (*i.e.* wavelength) in predictive performances (Figure 5). The analysis revealed that the most significant area lay between 260 and 310 nm, with wavelengths above 310 nm contributing less to algorithm construction. This result is consistent with mAbs composition and focus on aromatic amino acids absorbances: phenylalanine (245 – 260 nm), tyrosin (265 – 284 nm) and tryptophan (265 – 300 nm) [10].

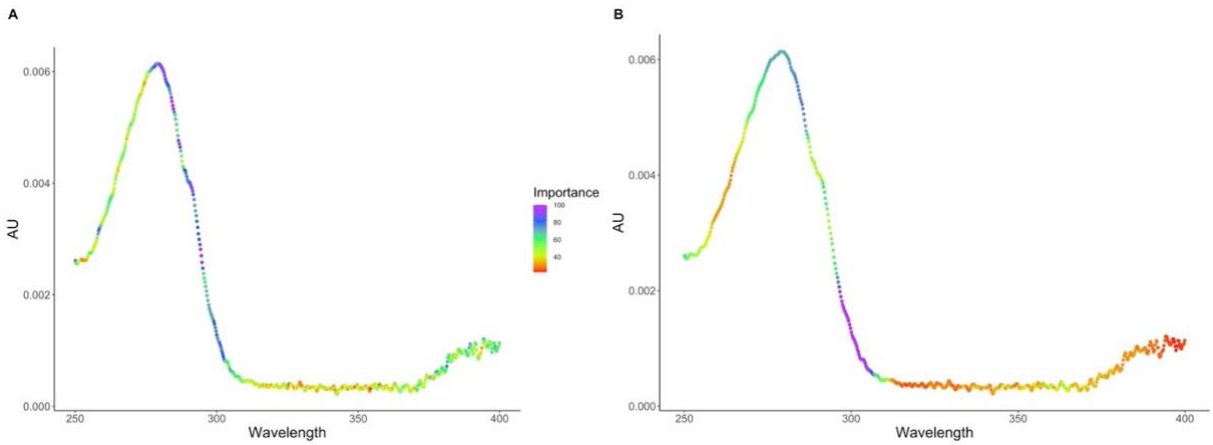


Figure 5: cumulative importance of variables, for PLSDA model (A), kNN model (B).

Therefore, the data analysis on the calibration dataset were reprocessed following a new wavelength selection, resulting in the best performance achieved with a spectral range from 240 to 335 nm for both discriminant analysis algorithms (PLS-DA and kNN). Main results with this wavelength selection, for several preprocessing methods, are represented in Table *1*.

Table 1: main results obtain with test set from calibration dataset, accuracy CI95[] represents mean accuracy with confidence interval of 95 % for ten repetitions (randomization of calibration dataset). Preprocessing methods: wavelength selection (240 – 335 nm) total area normalization (Narea), asymmetric least square (ALS) for baseline correction, smoothing by Savitzky-Golay (SG), first derivative (d1) and second derivative (d2).

Preprocessing	**Accuracy CI95[]** (PLS-DA)	**Accuracy CI95[]** (kNN)
Raw spectra	0.849 [0.830 – 0.868]	0.322 [0.292 – 0.352]
Narea	**0.996 [0.992 – 1.000]**	0.833 [0.810 – 0.856]
Narea + ALS	0.990 [0.984 – 0.996]	0.823 [0.796 – 0.850]
Narea + ALS + SG	0.993 [0.985 – 1.000]	0.821 [0.803 – 0.839]
Narea + ALS + SG + d1	0.991 [0.986 – 0.996]	0.904 [0.882 – 0.926]
Narea + ALS + SG + d2	0.946 [0.931 – 0.961]	**0.946 [0.930 – 0.962]**

Wavelength selection greatly improve accuracy for both discriminant models whatever preprocessing methods. The final models selected with calibration dataset were those with the highest prediction performances, *i.e.* the highest accuracy: for PLS-DA (nLV = 21) after Narea preprocessing, and for kNN (k = 5) after Narea, ALS, SG, d2 preprocessing.

Upon determination of the final models, prediction performances were calculated on a new validation dataset comprising of sampling from real mAbs preparations intended for patient administration. A total of 1,382 samples were collected and analyzed using UV spectroscopy.

Predictive performances were exceptional: for PLS-DA, an accuracy of 0.996 (or 99,6 % of well classified, equivalent to 5 misclassified out of 1,382), and for kNN, an accuracy of 0.998

(or 99,8 % correctly classified, corresponding to 2 misclassified out of 1,382). Detailed confusion matrices for the validation dataset for PLS-DA and kNN classification are available in supplementary information (SI, table S3 and table S4). Upon analyzing the spectra of misclassified samples, no specific reason could be identified, and thus, there were considered genuine prediction errors in all cases.

Finally, both approaches were highly accurate. Based on the stability and superior accuracy achieved with PLS-DA across all preprocessing methods, particularly with only Narea preprocessing, we believe selecting the PLS-DA algorithm is the optimal choice. PLS-DA not only provides consistent results but is also effective in addressing collinearity issues and can fit UV spectra, which are assumed to be linear while for kNN its seems that more samples are needed than PLS-DA to perform high accurate analysis. In practice, both discriminant analysis could be used with equivalent results.

4. Conclusion

In this study, we analyzed twenty-two mAbs using UV spectroscopy to perform discriminant analysis. Employing wavelength selection along with total area normalization significatively enhanced accuracy performance. Our predictive models achieved remarkable accuracy, correctly classifying 99.6% and 99.8% of a large dataset comprising 1,382 real-life prepared mAbs samples using PLS-DA or kNN, respectively. These methods are rapid and can be efficiently applied in various applications for mAbs discrimination.

References

[1] D. Zahavi, L. Weiner, Monoclonal Antibodies in Cancer Therapy, Antibodies (Basel) 9 (2020) 34. https://doi.org/10.3390/antib9030034.

[2] M. Vanneman, G. Dranoff, Combining immunotherapy and targeted therapies in cancer treatment, Nat Rev Cancer 12 (2012) 237–251. https://doi.org/10.1038/nrc3237.

[3] P. Gotwals, S. Cameron, D. Cipolletta, V. Cremasco, A. Crystal, B. Hewes, B. Mueller, S. Quaratino, C. Sabatos-Peyton, L. Petruzzelli, J.A. Engelman, G. Dranoff, Prospects for combining targeted and conventional cancer therapy with immunotherapy, Nat Rev Cancer 17 (2017) 286–301. https://doi.org/10.1038/nrc.2017.17.

[4] F.J. Esteva, V.M. Hubbard-Lucey, J. Tang, L. Pusztai, Immunotherapy and targeted therapy combinations in metastatic breast cancer, The Lancet Oncology 20 (2019) e175–e186. https://doi.org/10.1016/S1470-2045(19)30026-9.

[5] O. Hemminki, N. Perlis, J. Bjorklund, A. Finelli, A.R. Zlotta, A. Hemminki, Treatment of Advanced Renal Cell Carcinoma: Immunotherapies Have Demonstrated Overall Survival Benefits While Targeted Therapies Have Not, European Urology Open Science 22 (2020) 61–73. https://doi.org/10.1016/j.euros.2020.11.003.

[6] J. Cazin, P. Gosselin, Implementing a multiple-isolator unit for centralized preparation of cytotoxic drugs in a cancer center pharmacy, Pharm World Sci 21 (1999) 177–183. https://doi.org/10.1023/A:1008731406552.

[7] ASHP guidelines on quality assurance for pharmacy-prepared sterile products. American Society of Health System Pharmacists, American Journal of Health-System Pharmacy 57 (2000) 1150–1169. https://doi.org/10.1093/ajhp/57.12.1150.

[8] F. Drapeau, G. Claustre, S. Gaimard, C. Rossard, Preparation of injectable anticancer drugs within chemotherapy preparation unit: What levers are available to optimize and secure production?, Bull Cancer 110 (2023) 301–307. https://doi.org/10.1016/j.bulcan.2022.12.012.

[9] E. Jaccoulet, C. Smadja, P. Prognon, M. Taverna, Capillary electrophoresis for rapid identification of monoclonal antibodies for routine application in hospital, ELECTROPHORESIS 36 (2015) 2050–2056. https://doi.org/10.1002/elps.201400603.

[10] E. Jaccoulet, A. Schweitzer-Chaput, B. Toussaint, P. Prognon, E. Caudron, Simple and ultra-fast recognition and quantitation of compounded monoclonal antibodies: Application to flow injection analysis combined to UV spectroscopy and matching method, Talanta 187 (2018) 279–286. https://doi.org/10.1016/j.talanta.2018.05.042.

[11] A.A. Makki, V. Massot, H.J. Byrne, R. Respaud, D. Bertrand, E. Mohammed, I. Chourpa, F. Bonnier, Understanding the discrimination and quantification of monoclonal antibodies preparations using Raman spectroscopy, Journal of Pharmaceutical and Biomedical Analysis 194 (2021) 113734. https://doi.org/10.1016/j.jpba.2020.113734.

[12] L.M.M. Le, B. Kégl, A. Gramfort, C. Marini, D. Nguyen, M. Cherti, S. Tfaili, A. Tfayli, A. Baillet-Guffroy, P. Prognon, P. Chaminade, E. Caudron, Optimization of classification and regression analysis of four monoclonal antibodies from Raman spectra using collaborative machine learning approach, Talanta 184 (2018) 260–265. https://doi.org/10.1016/j.talanta.2018.02.109.

[13] W. Wang, S. Singh, D.L. Zeng, K. King, S. Nema, Antibody Structure, Instability, and Formulation, JPharmSci 96 (2007) 1–26. https://doi.org/10.1002/jps.20727.

[14] M. Kuhn, Building Predictive Models in R Using the caret Package, Journal of Statistical Software 28 (2008) 1–26. https://doi.org/10.18637/jss.v028.i05.

[15] H. Hotelling, Analysis of a complex of statistical variables into principal components, Journal of Educational Psychology 24 (1933) 417–441. https://doi.org/10.1037/h0071325.

[16] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemometrics and Intelligent Laboratory Systems 58 (2001) 109–130. https://doi.org/10.1016/S0169-7439(01)00155-1.

[17] A tutorial review: Metabolomics and partial least squares-discriminant analysis – a marriage of convenience or a shotgun wedding, Analytica Chimica Acta 879 (2015) 10–23. https://doi.org/10.1016/j.aca.2015.02.012.

[18] Y. Tie, C. Duchateau, S. Van de Steene, C. Mees, K. De Braekeleer, T. De Beer, E. Adams, E. Deconinck, Spectroscopic techniques combined with chemometrics for fast on-site characterization of suspected illegal antimicrobials, Talanta 217 (2020) 121026. https://doi.org/10.1016/j.talanta.2020.121026.

[19] P. Borman, D. Elder, Q2(R1) Validation of Analytical Procedures, in: ICH Quality Guidelines, John Wiley & Sons, Ltd, 2017: pp. 127–166. https://doi.org/10.1002/9781118971147.ch5.

[20] R.G. Brereton, Pattern recognition in chemometrics, Chemometrics and Intelligent Laboratory Systems 149 (2015) 90–96. https://doi.org/10.1016/j.chemolab.2015.06.012.

[21] M. Barker, W. Rayens, Partial least squares for discrimination, Journal of Chemometrics 17 (2003) 166–173. https://doi.org/10.1002/cem.785.

[22] T. Cover, P. Hart, Nearest neighbor pattern classification, IEEE Transactions on Information Theory 13 (1967) 21–27. https://doi.org/10.1109/TIT.1967.1053964.