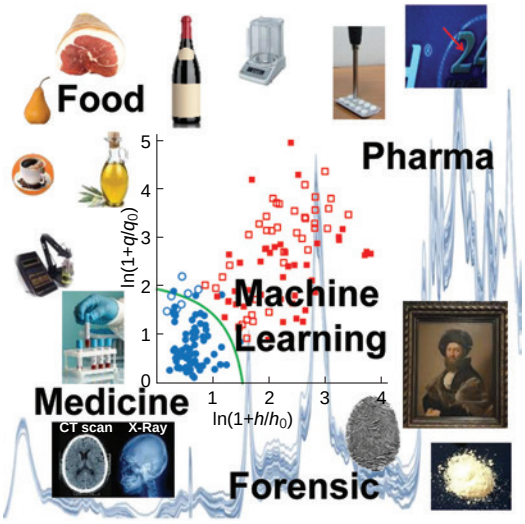
https://doi.org/10.59761/RCR5242

Chemometrics in qualitative analysis: identification, discrimination, and authentication

Oxana Ye. Rodionova,* Alexey L. Pomerantsev

N.N.Semenov Federal Research Centre for Chemical Physics, Russian Academy of Sciences, 119991 Moscow, Russian Federation

Non-targeted qualitative analysis has two aspects: instrumental, which involves collection of experimental data needed to solve a given problem, and mathematical, which involves analysis of these data to extract useful information and make reliable decisions. This review is devoted to the latter aspect, namely the general problems of chemometrics and machine learning, which are often incorrectly called artificial intelligence. Various problem formulations are considered, including discrimination and authentication; classification methods (binary, multi-class, or one-class); and types of made decisions (deterministic and probabilistic, soft and hard). Particular attention is given to analytical figures of merit such as sensitivity, specificity, and selectivity and how these characteristics are used for model optimization and validation. The concept of a cumulative analytical signal is also presented, and the prospects for its application in qualitative analysis are discussed.
The bibliography includes 162 references.
Keywords: classification; multi-class and one-class classifiers; optimization; validation; classification figures of merit; cumulative analytical signal.



Contents

1. Introduction

Among numerous analytical chemistry methods, qualitative analysis occupies a special, though not the most prestigious, place. It is responsible for the behind-the-scenes and often nasty work in forensic and customs expertise, medical pathology, food chemistry, and so on. However, unlike its successful 'big brother', quantitative analysis, this Cinderella lacks a rich array of luxurious accessories and adornments such as analysis of uncertainties, limits of detection and quantification, and so on. For example, it is well known what is the selectivity in quantitative analysis; there is a definition and methods for calculation. Meanwhile, until recently, there was nothing of the sort in qualitative analysis. Moreover, the International Union of

O.Ye.Rodionova. Doctor of Physical and Mathematical Sciences, Principal Researcher at FRC CP RAS.
E-mail: oksana@chph.ras.ru
Current research interests: chemometrics, numerical methods, vibrational spectroscopy.

A.L.Pomerantsev. Doctor of Physical and Mathematical Sciences, Principal Researcher at FRC CP RAS. E-mail: forecast@chph.ras.ru
Current research interests: chemometrics, mathematical statistics, chemical physics.

Translation: Z.P.Svitanko

Received 14 May 2026

Pure and Applied Chemistry (IUPAC) has not invented a proper definition for qualitative analysis. Officially, the old definition, adopted back in 1995, is still in effect: 'analysis in which substances are identified or classified on the basis of their chemical or physical properties such as chemical reactivity, solubility, molecular weight, melting point, radiative properties (emission, absorption), mass spectra, nuclear half-life, *etc*'.[1] According to the traditional understanding, qualitative analysis establishes the presence of particular elements/chemical compounds/functional groups in the analyte sample by means of qualitative test reactions.

However, the scope and methods of qualitative analysis have now markedly expanded beyond the identification of single chemical compounds.[2] Currently, the main trend is *non-targeted analysis*,[3–5] which focuses not on particular substances, but on analysis of those chemical and physical properties of a sample that make it similar to or distinct from the target sample. In addition, it was proposed to simplify the definition of qualitative analysis,[6] to make it 'classification according to specified criteria'; however, this minor revolution did not gain support within IUPAC.

Nevertheless, classification is an important and integral part of qualitative analysis. Machine learning and chemometrics offer a wide range of such methods.

Recent reviews address various aspects of the use of classification methods. For example a review by Strani *et al*.[7] is devoted to a strategy called one-class classification. Issues related to validation of screening methods in the context of qualitative analysis are considered in a review by Cuadros *et al*.[8] More specific classification issues related to application of experimental data gained using various analytical platforms are discussed in reviews on classification in vibrational spectroscopy[9] and in laser-induced breakdown spectroscopy (LIBS).[10]

The present review focuses not so much on specific methods as on their general characteristics, such as the way of presenting the results (binary or non-binary classification), the purpose of the method (single-class or multi-class one), and, of course, on the methods used to evaluate the quality of the results. In addition, the concept of a cumulative analytical signal is presented, and the prospects for its application are discussed.

To choose an appropriate method, it is necessary, first, to understand what is known and what question needs to be answered. Some typical scenarios in qualitative analysis are considered below.

Identification answers the question: what is it? The answer usually comes from comparison of a certain set of features, *e.g*., a spectrum or a chromatogram, with another similar set corresponding to a known instance. In practice, this is reduced to searching through known compounds stored in a library. Examples are identification using libraries of mass spectra[11–13] or Raman spectra.[14] This area is extensively addressed by Russian scientists;[15] hence, we will not examine it in detail.

Clusterization investigates a set of similar objects in order to understand to what extent they are uniform and whether they are subdivided into clusters or groups. This approach is called unsupervised classification, since information on the class to which the instances belong is either missing or not used in the modelling. An example of this approach is the analysis of Spanish wine samples using two-dimensional HPLC with fluorescence detection.[16] In essence, clusterization is an exploratory analysis technique used in various fields of analytical chemistry; therefore, it is also excluded from the scope of this review.

Discrimination determines to which of dedicated groups (classes) an instance belongs, with the exhaustive set of classes being known. This is a supervised classification in which the analytical chemist has either sets of samples for each class or ready analytical data describing each sample in the training set. To solve this problem, a model is developed and machine learning is used to form boundaries between the classes and to generate decision rules in order to determine to which of the specified classes the instance belongs. With this approach, it is critically important to define the exhaustive set of classes. For example, in the binary (two-class) classification, this involves discrimination between samples of sweet and bitter almonds using Fourier transform infrared spectroscopy.[17] An example of multi-class classification is subdivision of patients into three groups: healthy persons, type 1 diabetes patients, and type 2 diabetes patients. This is done by investigation of blood plasma using LIBS.[18] However, discrimination methods are often used even in those cases where it is impossible to define an exhaustive set of classes, for example, to classify various vegetable oils and their mixtures using NMR methods[19] or to recognize counterfeit medicines.[20] In these problems, there may be instances that do not belong to any of the classes used to build the model. In that case, only a preliminary or incomplete decision is possible. To solve problems of this type, it is necessary to proceed to the fourth type of methods.

Authentication answers the question of whether the presented sample is actually what it is declared to be. In other words, it is verification of authenticity of, for example, a medicine[21] or a specific variety of rice.[22] This problem is addressed using a one-class classifier (OCC), which builds a boundary encompassing the target class (often, this is the class of authentic specimens); this is used to determine whether or not a new sample can be assigned to the target class. In this case, training is also used to build the model. Here, the critical issue is to compose the set of samples that reliably and completely represent the target class.

Thus, we will focus on two basic types of non-targeted qualitative analysis: discrimination and authentication. In terms of methodology, qualitative analysis, that is, classification, is more complex than quantitative analysis, that is, calibration. When solving quantitative problems, the analytical chemist follows a well-known approach that has been tested and validated in numerous practical applications.[23] This procedure includes the following stages. The first stage is the use of a regression method, such as projection to latent structures (partial least squares, PLS, method).[24] The next one is performance evaluation using, for example, root mean squared error (RMSE) of calibration (RMSEC) and prediction (RMSEP). This is followed by optimization, for example, selection of the number of latent variables and other free parameters of the algorithm. The final stage is validation using, for example, a validation set. Of course, some details may vary; for example, principal component regression (PCR)[24] may be used instead of PLS, the correlation coefficient may be used instead of RMSE, cross-validation may be used instead of the validation set, and so on. However, the sequence of stages remains the same for any quantitative analysis method; that is, performance is estimated using calibration results, optimization is based on performance, and validation is carried out for the optimized model.

In qualitative analysis, the methodology is less well-developed, the sequence is less clear and has many gaps. The main difficulty is that the answer obtained upon classification is not a number defined with a certain degree of uncertainty.[25,26] Depending on the classification method used, very different answers are possible: YES/NO, or YES/NO/UNKNOWN, or even like

that: for a significance level of $\alpha = 0.01$, YES, but for $\alpha = 0.05$, NO.

Meanwhile, qualitative analysis, just like quantitative analysis, requires numerical assessment of the decision quality.[27] IUPAC and European Commission standards introduce two key quality parameters in qualitative analysis: *sensitivity* and *specificity*.[6] In the simplest case of binary classification, where a decision must be made as to whether a sample belongs to one of two classes, these metrics are easy to calculate. Therefore, their interpretation has been addressed in numerous publications dealing with qualitative analysis. However, as soon as the problem becomes more complicated, for example, for multi-class or one-class,[28,29] probabilistic[30] classification, evaluation of figures of merit is not an easy task.[31] Thus, this is a fairly typical case in which new classification methods are considerably ahead of quality assurance methodology.

The purpose of this review is to acquaint the reader with modern data analysis tools used in non-targeted qualitative analysis at each stage of development of a reliable solution, including: modelling, quality assessment, optimization, and validation.

2. Modelling

A classifier is a rule used to assign a sample to one of predefined classes relying on analysis of a set of data known as a fingerprint. A fingerprint of a sample is a multidimensional vector containing experimentally obtained characteristics. This could be a spectrum, a chromatogram, or a set of physicochemical parameters, such as component concentrations, colour, pH, *etc*. Although the main focus will be on vectors, the same approach can be extended to N-way data.[32,33]

2.1. Machine learning and artificial intelligence

The main methods of multivariate classification can be divided into several types depending on how the method uses the initial data. Methods of the first type directly handle the initial (pretreated) data. These are the so-called spectral matching methods. They include correlation coefficients and distances that express the similarity between the training set and new spectra.[34] Methods based on correlation coefficients are often used in software packages supplied with spectral instruments, for example, in the Micro Phazir analyzer, which is used for real-time monitoring of various food products.[35] The Euclidean distance is used in the software package supplied with Bruker instruments to treat vibrational spectroscopy data.[36] The Mahalanobis distance[37] is used in the UNEQ (unequal class models) method and is currently available in the Classification Toolbox for Matlab issued by the Milano Chemometrics and QSAR research group.[38] Recently, Rodionova and Pomerantsev[39] proposed a simple, but yet effective probabilistic classifier called NIMCA (Naïve DD-SIMCA), which is meant for multivariate classification, based on Euclidean distances, and implemented as a web-based application.[39] All the methods are simple, straightforward, and require adjusting only one parameter related to setting the significance level. Their main drawback is that they do not take account of the 'curse of dimensionality',[40] which can be stated as follows: the predictive power of a method decreases as the number of variables increases.

A distinctive feature of classification in chemical analysis is a specific structure of data that is called large p — small n. In these problems, the number of variables, elements of the fingerprint vector, markedly exceeds the number of samples. For example, the data may consist of tens or hundreds (n) of samples that are described by spectra including thousands (p) of variables.

Historically, the first classification methods were developed to analyze data in which $p \ll n$. First of all, these are classical methods such as linear discriminant analysis (LDA)[41] and quadratic discriminant analysis (QDA).[42] Among methods often used in various chemical applications, mention may be made of the k-nearest neighbours (kNN) method[43] and support vector machine (SVM).[44] These methods are based on the construction of boundaries between classes using distances. Another group of methods is based on *decision trees*,[45] *e.g*., random forest (RF) method[46] and extreme gradient boosting (XGBoost) method.[47] Thirty years ago, a boom around artificial neural networks (ANNs) started in chemistry, owing to efforts of Jure Zupan[48] and other authors.[49] By the start of the new century, this boom in chemistry had virtually come to an end, and interest shifted to artificial intelligence (AI).[50] However, it soon became clear that not everything could be called AI, and the focus began to shift toward *machine learning* (ML). This term encompasses all of the above methods. If they are used for high-dimensional data ($p \gg n$), this leads to major problems. Without going into details,[51] it can be simply stated that for using these methods, it is necessary first to reduce the problem dimensionality in order to make $p \ll n$. This can be done, for example, by principal component analysis (PCA).[52,53] The above-mentioned UNEQ method is also used after PCA projection. Recently, it was successfully used for geographical discrimination of saffron using ICP-MS data.[54]

One more problem faced while using machine learning methods is the difficulty of interpretation of the results. Despite their impressive success, deep learning (DL) models[55] are essentially black boxes that match input data to output data without revealing the contribution of predictors to the result.[56]

Chemometrics took a different approach: the external reduction of the number of variables is included into the data analysis techniques, thus becoming a part of the classification algorithm. This gave rise to classification methods that are most popular in chemistry. These methods are discussed in a separate section.

2.2. Various classifiers

In the case of supervised classification, the initial number of classes K is known. In addition, each class comprises a representative set of samples that are used to train the model. When $K = 1$, one deals with *one-class classification*. When $K = 2$, this is *binary* classification. The general case where $K > 2$, is called *multi-class* classification. The modelling methods for $K = 1$ and $K > 1$ are considerably different. A one-class classifier serves to build a boundary around the target class; therefore, this approach is also referred to as class modelling.[57] If there are several classes, a separate model is built for each class. If two or more classes are initially considered, it is necessary to construct boundaries separating these classes; that is, the discrimination problem is solved.[58] No matter how many classes are involved in a multi-class classification problem, a single model is built. The difference between one-class classification and discrimination is schematically illustrated in Fig. 1.

Each classifier has its own *decision rule* used to classify a new object: if the specified condition is met, then the object belongs to some class. These rules can be *deterministic* or

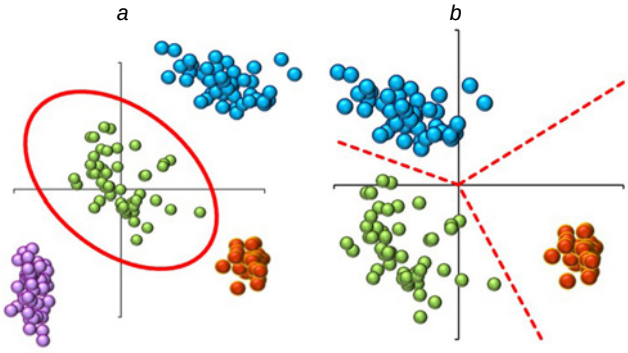


Figure 1. Schematic view of one-class (*a*) and multi-class (*b*) classification.

probabilistic. Deterministic classifiers simply give a result (YES/NO) for each object, whereas probabilistic classifiers additionally characterize the uncertainty of that decision. These approaches can be compared to point and interval estimates.

This difference can be illustrated by a simple numerical example. There are two small data sets, each containing 50 samples (Fig. 2). One class is called Blue (blue dots), and the other one is called Red (red squares). The deterministic approach is shown in the plot on the left (Fig. 2 *a*), while the probabilistic approach is shown on the right (Fig. 2 *b*). The green line represents the critical level used to make the decision.

In the left-hand plot, the deterministic classifier produces the following results: 49 samples in the Blue class are classified correctly (true positives), and only one sample, which lies above the threshold, is misclassified. In the Red class, 40 samples (true negatives) are correctly classified, while the ten samples that fall below the critical threshold are classified incorrectly.

The probabilistic classifier concept shown on the right (Fig. 2 *b*) is based on a different approach. First, each dataset is modelled with its own distribution, shown by the blue and red curves. For this purpose, the distribution parameters are estimated using the data. Next, the significance level (type I error) is selected, *e.g.*, $\alpha = 0.05$, and the critical level corresponding to this probability is calculated. Next, type II error is found as the area under the probability density function (p.d.f.) for the alternative class. For the case shown in the plot, $\beta = 0.09$. Using a probabilistic classifier, one can perform internal validation by comparing the empirical values for sensitivity and specificity with their theoretical analogues: $1-\alpha = 0.95$ and $1-\beta = 0.91$.

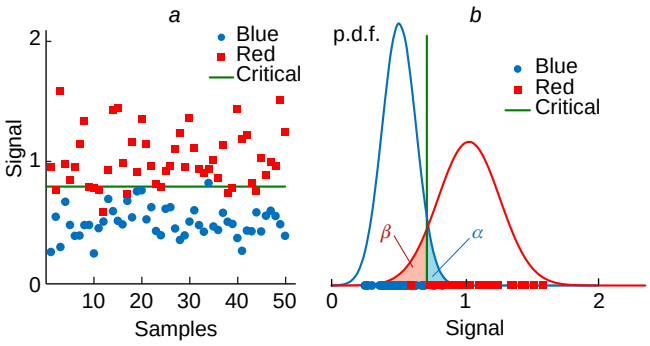


Figure 2. Concepts of the deterministic (*a*) and probabilistic (*b*) classifiers.

Classifiers also differ in the way they assign samples to one class or another. The most popular approach is *hard* classification,[31] in which each sample (training or new one) can be assigned only to one class. However, there is also another approach according to which a sample can be simultaneously assigned to a few classes or not classified at all. This approach is called *soft* classification.[58,59]

If we consider the current state of things using the Scopus database and find out how frequent particular discrimination methods have been used in chemical analysis in the last five years (2021–2026) (Fig. 3 *a*), then it will be seen that the partial least squares discriminant analysis (PLS-DA)[60] is obviously most popular, although this method was proposed rather long ago. The total number of mentions of PLS-DA in chemical applications has been more than 12 thousand since 1989.

The PLS-DA method is widely used in various fields, *e.g.*, to diagnose cancer using Raman spectroscopy data,[61] to identify respiratory diseases,[62] to analyze results of NMR spectroscopy in pharmaceutics,[63] to study foodstuffs,[64,65] and to analyze petroleum products using gas chromatography data[66] and IR spectroscopy data.[67]

One-class classifiers are directed toward modelling of the key properties of the target class rather than toward the search for differences between classes.[68] The method known as the one-class discriminator (OC-PLS) was an attempt to adapt PLS-DA for one-class classification,[69] but this did not gain much popularity. Machine learning methods that use a particular type of kernel or random variable selection for data conversion constitute a separate category. These methods include support vector domain description (SVDD)[70] and the related one class support vector machine (OC-SVM).[71] Despite their popularity in machine learning, these methods are not often used to analyze data in chemistry. Thus, out of a total of 744 references on the use of SVDD from 2021 to 2026, we were able to find only 57 studies in which this method was used to solve chemical problems (Fig. 3 *b*). An example of formal methods based on variable selection is the one-class random forest (OC-RF) or its variant that uses a genetic algorithm.[72] In essence, methods using support vectors, like those using random forest, are discrimination methods requiring the presence of an alternative class. If there is no alternative class, this class should be formed by some artificial method.

A conceptually different approach underlies a method that emerged in analytical chemistry more than 40 years ago thanks to S.Wold and B.Kowalski; this method is known as soft

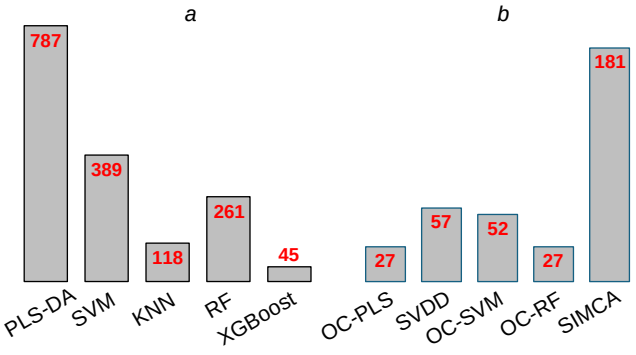


Figure 3. Results of the Scopus search for publications that use classification methods in chemical applications in 2021–2026, with the search being restricted to papers in chemistry: (*a*) discrimination ($K \geqslant 2$), (*b*) one-class classification ($K = 1$).

independent modelling of class analogy (SIMCA).[73,74] With this approach, an alternative class is not required, and each sample can be simultaneously assigned to several classes or not classified at all. Currently, the SIMCA approach has combined a whole series of techniques under the single umbrella term.[75,76] The possibility of this classification is attracting increased interest among both analytical chemists[77–79] and machine learning specialists.[80]

3. Chemometric approach

It has already been noted that chemometric modelling predominates in qualitative chemical analysis. This is due to the fact that this approach is better suited to data in which the number of variables (*e.g.*, wavelengths) is greater than the number of samples, which are always deficient in chemistry. In this Section, we provide a brief description of the two most popular chemometric methods used in qualitative analysis.

3.1. Discrimination using PLS-DA

PLS-DA is an exceptionally popular approach in chemometrics that is used, most often, for binary discrimination. However, PLS-DA in itself is not a classifier, but only a compression method that converts the initial data by transferring them from the multivariate space of variables to low-dimensional space of features. These features are then used in a classifier.

The modelling starts with PLS regression, which uses a categorical (dummy) matrix of responses **Y** consisting of zeros and ones that indicate the class assignment. This is a typical feature engineering stage, which gives a new set of features as score matrix $\hat{\mathbf{Y}}$, which depends on the number of latent variables (LV). The next step is to create a classifier (decision rule), which can use any of the numerous methods described above: soft, hard, deterministic, probabilistic, *etc*. Thus, the term PLS-DA is a general concept that encompasses a set of discriminators based on this idea.

Consider a general multi-class PLS-DA in which I samples are distributed over K classes. Feature extraction is based on PLS2 regression in which $(I\times J)$ data matrix **X** is used as predictors, while $(I\times K)$ dummy matrix **Y** serves as responses. The samples are split into K groups of sizes $I_1+I_2+\ldots+I_K=I$, with the corresponding index groups $\omega(1), \omega(2), \ldots, \omega(K)$, which indicate that sample i belongs to class k, that is, $i \in \omega(k)$.

The matrix of dummy variables **Y** contains the values {0,1}, which indicate assignment to classes. The matrix is built in the following way. Consider the identity matrix **E** of size K, which can be represented as a column of row vectors **e**.

$$\mathbf{E}=\begin{vmatrix}\mathbf{e}_1\\ \mathbf{e}_2\\ \cdots\\ \mathbf{e}_K\end{vmatrix}=\begin{vmatrix}1 & 0 & \cdots & 0\\ 0 & 1 & \cdots & 0\\ & & \cdots & \\ 0 & 0 & \cdots & 1\end{vmatrix} \qquad (1)$$

Each vector $\mathbf{e}_k$, $k = 1,\ldots, K$, is the pattern response for class k. Now matrix **Y** can be represented as a column containing the row vectors y_i, $i = 1, \ldots, I$.

$$\mathbf{Y}=\begin{vmatrix}\mathbf{y}_1\\ \cdots\\ \mathbf{y}_i\\ \cdots\end{vmatrix} \qquad (2)$$

We do not discuss PLS2 regression, as it is explained in many textbooks, *e.g.*, book by Martens and Naes.[24] The data pretreatment is standard: the matrix **X** is always centred (by columns) and may be scaled depending on the nature of the variables; the matrix **Y** is only centred. The regression results in the $(I\times K)$ matrix of predicted responses $\hat{\mathbf{Y}}$, which is used as input data for various discriminators, for example, LDA or QDA.

Figure 4 shows two examples of discrimination that used the same data: the set of near-IR (NIR) spectra recorded for samples of dried oregano (*Origanum vulgare*): purchased in a store (Store), collected in the wild (Nature), and counterfeit (Fake).[81]

The left plot (Fig. 4*a*) shows the result of deterministic discrimination in which the three classes are separated by boundaries.[58] It can be seen that each sample has been assigned to a certain class, in some cases, incorrectly. In the plot on the right (Fig. 4*b*), the same samples are classified using a probabilistic model, where the decision boundaries depend on the significance level α (in this example, α = 0.1). In this case, there are several samples assigned simultaneously to two classes and there are samples not assigned to any of the classes; therefore, this is a soft classification.

3.2. Authentication using DD-SIMCA

DD-SIMCA is a modern version of the one-class SIMCA method, which is a probabilistic classifier. The decision rule is formulated according to a predefined significance level α, and the distribution parameters are estimated using experimental data. If an alternative class is available, DD-SIMCA provides the possibility of calculating type II error β and construct the corresponding extended decision region, which guarantees that the risk of accepting a sample from the alternative class does not exceed β.

Quite a number of studies are devoted to various aspects of the DD-SIMCA method, and the final description of the method was reported by Kuchertavskiy *et al*.[82] In brief, the method can be described as a two-stage procedure. In the first stage, like in any sort of SIMCA method, the principal component analysis is applied to the training dataset collected from samples of the target class. The dataset matrix **X** of size $(I \times J)$, centred and/or scaled, is represented in the form:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{t}} + \mathbf{E} \qquad (3)$$

where $\mathbf{T} = \{t_{ia}\}$ is the $(I\times A)$ matrix of scores; $\mathbf{P} = \{p_{ja}\}$ is the loading matrix of size $(J\times A)$; $\mathbf{E} = \{e_{ij}\}$ is the residual matrix of size $(I\times J)$; A is the number of principal components (PCs).

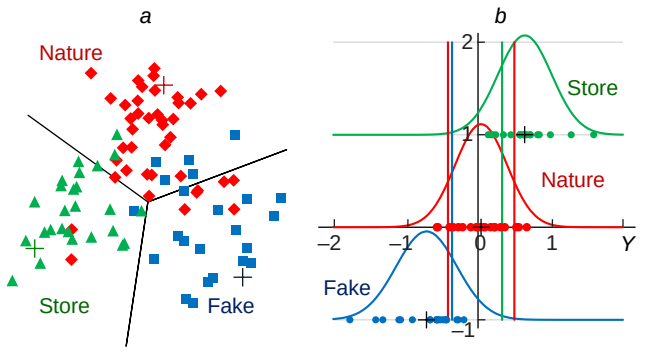


Figure 4. Examples of using PLS-DA for deterministic (*a*) and probabilistic (*b*) discrimination of oregano (*Origanum vulgare*) samples: purchased in a store (Store), collected in the wild (Nature), and counterfeit (Fake).

In the second stage, a decision rule is established. The way this rule is chosen determines which variant of the SIMCA method is used.[75] In the DD-SIMCA method, two distances are calculated for each sample of the training set $i = 1, \ldots, I$. The first one is orthogonal distance (OD), q_i, which is calculated as the squared Euclidean distance from the sample **x** to the subspace of principal components

$$q_i = \sum_{j=1}^{J} e_{ij}^2 \tag{4}$$

The second one is the score distance (SD), h_i, the squared Mahalanobis distance from the model centre to the sample projection in the subspace of principal components

$$h_i = \mathbf{t}_i^{\mathrm{t}} (\mathbf{T}^{\mathrm{t}}\mathbf{T})^{-1} \mathbf{t}_i \tag{5}$$

The normalized SD and OD values are described by the chi-square distribution

$$N_h \frac{h}{h_0} \sim \chi^2(N_h) \tag{6}$$

$$N_q \frac{q}{q_0} \sim \chi^2(N_q)$$

All distribution parameters such as the factors h_0, q_0 and the degrees of freedom N_h and N_q are estimated using the corresponding distances calculated for samples of the training set; therefore, the method is called data-driven (DD). The chi-square distribution parameters can be found using either the classical or the robust method. In the absence of outliers, the classical approach provides more accurate results. The robust approach is less sensitive to outliers and, as a result, helps to identify them.[82]

The full distance (FD), f, defined as the weighted sum of h and q, also obeys the chi-square distribution with $N_f = N_h + N_q$ *degrees of freedom* according to the definition of χ^2.

$$f = N_h \frac{h}{h_0} + N_q \frac{q}{q_0} \sim \chi^2(N_f) \tag{7}$$

where $N_f = N_h + N_q$. The full distance is used to establish the critical threshold depending on the significance level α in the following way:

$$f_{\mathrm{crit}} = \chi^{-2}(1 - \alpha, N_f) \tag{8}$$

The samples for which the full distance f_i satisfies the inequality

$$f_i \leqslant f_{\mathrm{crit}} \tag{9}$$

refer to the target class.

Graphically, the classification result can be represented using an acceptance plot in the q/q_0 axis against the h/h_0 axis (Fig. 5),[83] irrespective of the dimension of the principal component subspace. The green marks (Fig. 5 *a*) designate regular samples of the training set, while the green line indicates the acceptance threshold for the target class.

In addition, it is possible to specify the probability γ, which is an appropriate uncertainty in determination of outliers. Then the region of outliers is defined by the boundary

$$f_{\mathrm{out}} = \chi^{-2}\left((1 - \gamma)^{1/I}, N_f\right) \tag{10}$$

and all training set samples such that

$$f_i > f_{\mathrm{out}} \tag{11}$$

correspond to outliers. The red marks (Fig. 5 *a*) designate outliers among the training set samples; the red line is the outlier boundary. The training set samples located between the acceptance boundary and the outlier boundary (Fig. 5 *a*, yellow marks) are called extremes. If the boundary has been chosen appropriately, the proportion of extremes is close to the significance level α. If samples located within the acceptance region and samples located far beyond the acceptance boundary should be presented in the same plot, a logarithmic axis transformation should be used (Fig. 5 *b*).

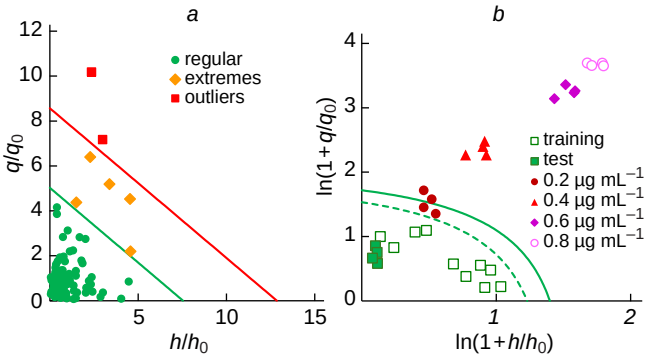


Figure 5. Examples of using DD-SIMCA. (*a*) Classification of metronidazole tablets.[83] (*b*) Determination of tetracycline in blood serum. The result for α = 0.05 (continuous line) and α = 0.15 (dashed line).

If alternative data are available, type II error can be estimated.[84] Let the vector **x** be a sample from an alternative dataset. If the vector **x** is projected onto the subspace of principal components based on the target set, the full distance FD for it can be calculated by equations (4) and (5), which use the parameters h_0, q_0, N_h, and N_q found for the target class. Thus, all calculations for the alternative set are similar to those for the target set. The main difference is that the distribution for the alternative set FD is not governed by the standard chi-square distribution, but rather by its generalization: the noncentral chi-square distribution $\chi'^2(N, s)$, where N is the number of the degrees of freedom, and s is the noncentrality parameter.

Thus, the type II error (β) can be estimated using the equation

$$\beta = \Pr\left\{\chi'^{-2}(N_f, s) < \frac{f_{\mathrm{crit}}}{f_0'}\right\} \tag{12}$$

where the f_{crit} value is defined in equation (8). The noncentrality parameter s and the scaling factor are estimated using the data of the alternative set.[84]

When β is specified, equation (12) can be inverted to find FD corresponding to this error. This method is useful for risk assessments[85] and for determining the limit of detection in one-class classification.

4. Figures of merit

The set of indicators that characterize the quality of a constructed model is collectively referred to as figures of merit and is commonly abbreviated as FoM. We will also use this abbreviation. The concept of FoM encompasses several characteristics: correct sample assignment to particular classes, evaluation of misclassification errors, assessment of model efficiency, and evaluation of the predictive capability of the model.

4.1. Confusion matrix

Irrespective of the classification method, the primary result of classification is the *confusion matrix*,[86] a table that shows the

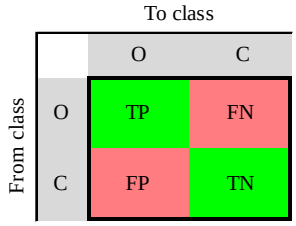


Figure 6. Confusion matrix in the binary classification, $K = 2$.

distribution of predicted samples among the actual classes (Fig. 6). For one-class classification, the confusion matrix is reduced to a single row.

4.2. Binary classification

In the case of binary classification, the size of the confusion matrix is (2 × 2) (see Fig. 6). If the target class is class O, then the elements of the matrix are defined as follows. The TP (true positive) cell shows the number of samples in class O that have been correctly assigned to this class. The FN (false negative) cell presents the number of class O samples that have been erroneously assigned to class C and, hence, they were falsely rejected. The FP (false positive) cell shows the number of samples of class C that have been erroneously assigned to O. Finally, in the TN (true negative) cell, there is the number of samples of class C assigned to class C and, hence, correctly rejected as not belonging to class O. In the binary classification, classes O and C are equivalent; that is, class C can be defined the target class, instead of O; in this case, the true positives and true negatives would exchange places.

The confusion matrix is used to calculate the basic FoMs.[6] They include sensitivity (SNS), which is the proportion or percentage of samples in the target class that are correctly assigned to this class, and specificity (SPC), which is the proportion or percentage of samples from another class that were correctly identified as not belonging to the target class. In the case of hard classification, where each sample necessarily belongs to one of the specified classes, SNS and SPC values are calculated by conventional equations:[86]

$$\mathrm{SNS} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN}) \qquad (13)$$

$$\mathrm{SPC} = 1 - \mathrm{FP}/(\mathrm{TN} + \mathrm{FP})$$

In addition, TP + FN = I_O (the number of samples in class O) and TN + FP = I_C (the number of samples in class C).

The main drawback of the hard classification is the absence of options such as 'the sample does not belong to any class at all', or 'the sample simultaneously belongs to several classes'. These options are possible only in the soft classification. Meanwhile, in practice, this is quite often the case. For example, a new sample may belong to neither of these two classes (O or C), but rather to a third class Z that was not taken into account in the modelling.[68] In addition, in biomedical problems, there is often no clear boundary between classes.[31,81]

Figure 7 shows the results of binary discrimination using hard and soft methods. Samples of Brazilian coffee grown by either organic or conventional method were studied. The model was built using 18 samples grown using organic methods (class O) and 26 samples grown using conventional methods (class C). A set of 22 parameters reflecting the composition, physicochemical properties, and antioxidant activity was used as variables.[87]

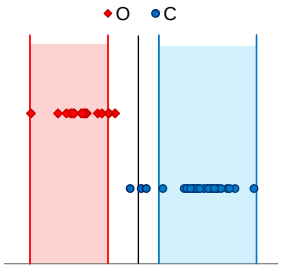


Figure 7. Hard and soft discrimination of Brazilian coffee in terms of the growth methods: organic (O) and conventional (C) techniques.

In Fig. 7, hard discrimination divides the plane into two parts along a vertical black line: O is on the left and C is on the right. It is clear that all samples are reliably discriminated, except for one sample of class C. Soft discrimination is specified by two areas: pink (O) and blue (C). These areas were generated using soft PLS-DA[58] for the significance level $\alpha = 0.1$. In this case, two samples from class C and three samples from class O remain unclassified, which is reflected in the corresponding confusion matrix (the upper part of Table 1). Here, the same principle as in Fig. 6 is used: the rows correspond to the known class and the columns correspond to the obtained class.

This example shows that in the case of soft classification, equations (13) are inapplicable for calculating sensitivity and specificity, since the total number of samples in each class is not equal to the sum of true positives and false negatives. Therefore, for the general case of discrimination between classes A and B, equations (13) must be written as

$$\mathrm{SNSA} = \mathrm{TP}/I_A \qquad (14)$$

$$\mathrm{SPCA} = 1 - \mathrm{FP}/I_B$$

$$\mathrm{SNSB} = \mathrm{TN}/I_B$$

$$\mathrm{SPCB} = 1 - \mathrm{FN}/I_A$$

It is useful to introduce another figure of merit that combines the sensitivity and specificity. This is efficiency, which is calculated as the geometric mean, or proportional mean, which allows using this value even when the number of samples in different classes is unbalanced.

$$\mathrm{EFF} = \sqrt{\mathrm{SNS} \cdot \mathrm{SPC}} \qquad (15)$$

All these values are given in the lower part of Table 1.

Apart from the efficiency, another frequently used parameter is accuracy (ACC), which is calculated as the proportion of correctly classified samples divided by the total number of samples

Table 1. Discrimination of Brazilian coffee.[87]

	Hard method		Soft method	
	O	C	O	C
O	18	0	15	0
C	1	25	0	23
	Figures of merit			
SNS	100%	96%	83%	88%
SPC	96%	100%	100%	100%
EFF	98%	98%	91%	94%

$$\mathrm{ACC} = (\mathrm{TP} + \mathrm{TN})/(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}) \quad (16)$$

The ACC index works best when the sample sizes for both classes are balanced. For unbalanced classes, in addition to equation (15), the F1-score is also used; it is particularly popular in biomedical research. In addition to the primary figures of merit, there are numerous secondary metrics and indices that characterize the results of classification. For example, Ferri *et al.*[88] analyzed 18 such metrics, while Sokolova and Lapalme [89] considered 24 metrics. All of these indices and characteristics are combinations of key values of the confusion matrix.

4.3. Multi-class classification

A general classification problem considers I samples representing K target classes with sizes $I_1+I_2+...+I_K = I$. The classification results can be presented as a confusion matrix of size $(K \times K)$. The matrix elements p_{kl} are the numbers of samples of class k assigned to class l. In the case of multi-class classification, the issues related to definition of FoMs have not yet been fully studied,[29,90] with the definitions being more complex. The TP(k) and FP(k) values are defined for each class on the basis of the confusion matrix [91]

$$\mathrm{TP}(k) = p_{kk} \quad (17)$$
$$\mathrm{FP}(k) = \sum_{l \neq k} p_{lk}$$

and the appropriate FoMs are calculated separately for each class k

$$\mathrm{SNS}(k) = \mathrm{TP}(k)/I_k \quad (18)$$
$$\mathrm{SPC}(k) = 1 - \mathrm{FP}(k)/(I - I_k)$$

Thus, $2K$ indices are collected, and in order to characterize the results of the multi-class model as a whole, Pomerantsev and Rodionova [91] proposed calculating the total figures using the following equations. The total sensitivity (TSNS) was defined as

$$\mathrm{TSNS} = \sum_{k=1}^{K} \frac{I_k}{I} \mathrm{SNS}(k) = \frac{1}{I} \sum_{k=1}^{K} \mathrm{TP}(k) \quad (19)$$

and the total specificity (TSPC) was defined in the following way:

$$\mathrm{TSPC} = \frac{1}{K-1} \sum_{k=1}^{K} \frac{I - I_k}{I} \mathrm{SPC}(k) = 1 - \frac{1}{I(K-1)} \sum_{k=1}^{K} \mathrm{FP}(k) \quad (20)$$

Formulas (18)–(20) are valid for both hard and soft classification methods. The total efficiency TEFF

$$\mathrm{TEFF} = \sqrt{\mathrm{TSNS} \cdot \mathrm{TSPC}} \quad (21)$$

characterizes the classification model as a whole and serves both for model optimization and for comparison of various models

4.4. One-class classification

In the case of one-class classification, there is only one target class, and all values are calculated relative to that class. As for the alternative (non-target) class, it may be absent; then it is impossible to calculate SPC, and EEF = SNS. Also, there may be several alternative classes. In this case, the specificity should be calculated separately for each of the alternative classes. When the constructed model is applied to new, unknown samples, it makes sense to evaluate only SPC, since there is no information on whether the samples belong to the target class, while the efficiency for a new set, in the absence of class assignment data, is calculated as EFF = SPC.

4.5. *A priori* characteristics of classification methods

In the description of classification models, it is necessary to distinguish two different sorts of indices. The first one includes, *e.g.*, type I error (α) and type II error (β), which have been borrowed from statistics. In analytical chemistry, they correspond to the sensitivity and specificity. The main difference between the statistical and analytical approaches is that the parameters α and β are specified *a priori* or calculated theoretically, whereas SNS and SPC are calculated *a posteriori*, after the model has been built. Type I error α is the significance level, that is, the probability of incorrect rejection of samples of the target class, while β is the probability of incorrect acceptance of samples of other classes as belonging to the target class.[92] For a specified α value, a classifier is developed, and the classification result is evaluated using SNS. If SNS is close to $1-\alpha$, the goal has been achieved, and the model is optimal. Otherwise, the classifier should be adjusted, *e.g.*, by increasing or decreasing the model complexity. The SPC is an *a posteriori* evaluation of the $1-\beta$ value; therefore, for a well-trained model, one can expect that

$$\mathrm{SNS} \approx 1 - \alpha \quad (22)$$
$$\mathrm{SPC} \approx 1 - \beta$$

Apart from above FoMs, there are other FoMs, *e.g.*, the *selectivity* and the *limit of detection*. They are actively used in quantitative analysis, but in qualitative analysis, their definitions and calculation methods were unknown until recently. These issues are discussed in separate parts of the review.

5. Optimization and validation

The optimization and validation of a model are, in essence, the same procedure carried out using the same tools but with different objectives. The procedure includes comparison of the figures of merit (FoMs) achieved upon calibration using the training set with those obtained in the prediction using the independent test set. Optimization is selection of the model parameters that ensure the best performance and robustness. Validation is verification of the fact that the model provides the expected results when used in practice. Each of these procedures requires its own test set.

5.1. Optimization

All classification models used in qualitative analysis have parameters that can be adjusted to improve the model performance. The number of these parameters varies greatly, ranging from one, as in the kNN method, to almost infinity, as in DL methods. In the case of projection methods, this parameter is the dimensionality of the projection space, for example, the number of principal components or latent variables. For algorithms based on the support vectors, these are the regularization parameter and the choice of kernel function; for methods based on search trees, these parameters include the number of trees, the maximum depth, minimum number of samples in a node, and the number of features for splitting. This choice is well formalized [93] in quantitative analysis for the

solution of calibration problems. The root-mean-square-errors RMSEC and RMSEP are plotted *vs*. a complexity parameter (*e.g.*, the number of principal components) and, after that, the point in which RMSEP is minimum is chosen.

Pomerantsev and Rodionova[91] proposed a similar approach for classification using the total efficiency of training (calibration), TEFFC, instead of the error of calibration, RMSEC, and the total efficiency of validation (prediction), TEFFP, instead of the error of prediction, RMSEP. Generally, as the model becomes more complex, the TEFFC first increases and then remains stable. Simultaneously, TEFFP first increases and then decreases. The point of convergence of these two characteristics is chosen as the optimal value for the complexity parameter. In the above example of binary classification of Brazilian coffee (see Fig. 7), there are two possible options. The first one is to choose the number of PLS latent variables, LV = 1; then TEEFC = 90% and TEFFP = 95%, which is a quite reasonable result considering the problem in question. The other option is LV = 3, then TEEFC = 97% and TEFFP = 95% (Fig. 8).

As the model complexity further increases, the efficiency on the test set declines, indicating that the classifier is overfitted. The selection of a smaller number of latent variables gives a more robust model, while the dimensionality equal to three reveals a greater difference between the classes. For multi-class classification, the optimization method remains the same, and the total efficiency is calculated using equations (19)–(21).

To optimize the deterministic PLS-DA model, it is necessary to find the unique free parameter: the number of latent variables. The probabilistic classifiers have one more parameter, significance level α, which can also be adjusted. The goal of this optimization is to balance the SNS and SPC values by making them approximately equal. This can be done by an approach based on analysis of receiver operating characteristic (ROC) curves.

In the traditional binary classification, the ROC curve characterizes the model performance in the SNS *vs*. 1–SPC coordinates and can be used to select the optimal threshold.[76] A convenient ROC approach for multi-class classification has not yet been developed.[94,95]

However, for the probabilistic discriminator PLS-DA presented in Section 3.1, the ROC curve can be constructed in the (1–TSPC; TSNS) coordinate system by varying the value of α. This curve is shown in Fig. 9. When α = 0, then TSPC = 0 and TSNS = 1, while in the case of α = 1, TSPC = 1 and TSNS = 0. The 'distance to (0, 1)' curve is calculated as the distance from the ROC curve to the point with the (0, 1) coordinates. The optimal α value, *i.e.*, the threshold between the classes, is chosen for the minimum value of this distance. The other way for selecting the α value is to plot TEFF as a function of α. The highest point in this curve corresponds to the optimal value of α. Both methods appear to be equivalent, but the former one provides an additional characteristic called 'area under the curve' (AUC), that is, the area confined by the ROC curve and the horizontal axis. The higher the AUC value, the better the classifier. In the example shown in Fig. 9, AUC = 0.94, which is a good result.

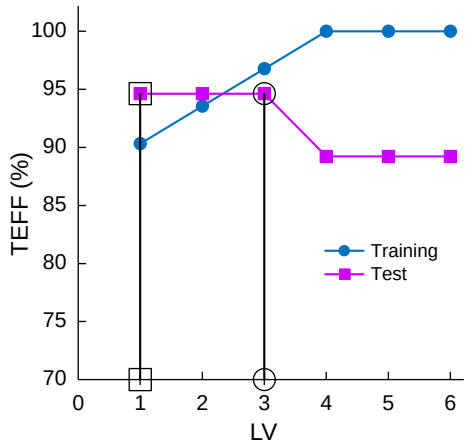


Figure 8. Optimization in the binary discrimination of Brazilian coffee: relationship between overall efficiency and model complexity.

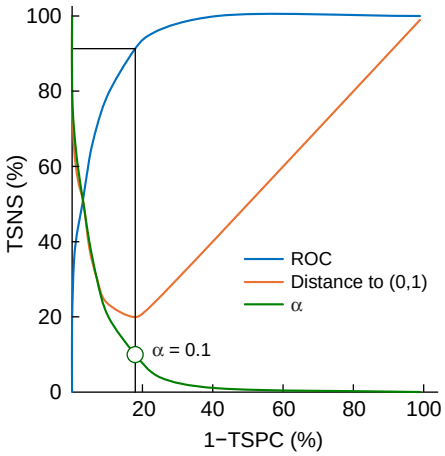


Figure 9. ROC curve for the optimization of the parameter α in the probabilistic model for discrimination of oregano samples.

5.2. Validation

The validation[96] is a necessary step in the construction of the final model based on formal modelling. The model should be tested using an independent test set, which (1) should not be involved in training of the model; (2) should not be a copy of the training set, *e.g.*, should not consist of repeated measurements of the training set; and (3) should be, as much as possible, a representative selection of samples that would be encountered in the practical use of the model. In an ideal study, it is preferable to use three data sets that represent the population of samples: a training set for model construction, test set No. 1 to optimize the model, and test set No 2 for validation (validation set). Often, a researcher does not have this large number of samples[97] due to various reasons. In this case, two sets are used: a training set and a test set. The test set is selected among the total initial data set either randomly or using various algorithms, such as the Kennard–Stone method, D-optimal design, and others.[98–100] A new approach[101] to the choice of an optimal subset is described in Section 6.6.

Another way to generate a test set is to simulate a new set using statistical modelling. The most popular method is the cross-validation procedure. The most frequently used versatile cross-validation procedures and evaluation of their efficiency in various scenarios were described in detail by Bro *et al.*[102] In the field of machine learning, the cross-validation procedure has become the gold standard for the model optimization stage.[103]

A double (or nested) cross-validation procedure uses two nested cross-validation loops, which simultaneously evaluate both the model complexity, *i.e.*, optimization step, and the

prediction efficiency.[104] However, in some situations, cross-validation is inapplicable. For example, this is the case where data are grouped or contain the results of repeated experiments; or data reflect a process that varies over time; or data have been collected using design of experiments; or a very small data set is used and, hence, every sample is critically important. Therefore, a new method called Procrustes cross-validation (PCV) was recently proposed for generating a simulated set.[105] The method is based on the k-fold cross-validation algorithm; however, unlike the conventional algorithm, the proposed method allows the generation of a new dataset that carries a correct variability evaluated by the cross-validation procedure. This dataset, known as a pseudo-validation (PV) set, can be used in the same way as an independent test set. It can serve to calculate residual distances, explained variance, scores, and other results that cannot be found by conventional cross-validation. This approach makes it possible to model and test small datasets.[106]

5.3. Rigorous and compliant approaches to one-class models

A one-class classifier (OCC) is always based on one (target) class, while other (alternative) classes are ignored in the model. Hence, we cannot calculate SPC, and EFF = SNS. Correspondingly, the number of principal components is optimized by comparing the sensitivity of calibration (SNSC) with the sensitivity of prediction (SNSP). This approach to modelling is called rigorous.[30] Its main advantage is that it is completely independent and can be applied to any alternative classes. However, an analytical chemist often focuses on only one alternative class and seeks to improve the classification quality specifically for this alternative. Then it is possible to calculate the specificity SPC and optimize the number of principal components by comparing EFFC with EFFP. Furthermore, the significance level α can also be optimized using the ROC method described above.[76] This approach is called compliant.[30]

Thus, the benefits of the rigorous approach include the possibility of authentication and versatility, while its drawback is low specificity. For the compliant approach, the opposite is true: the benefit is high specificity, while the drawback is the lack of possibility of authentication, since in this case, OCC simply becomes a discriminator that is not suited for classifying samples of new classes.[68] Note that compliant SIMCA is a poor discriminator compared to PLS-DA.

There is also an intermediate approach that can be called risk management.[85] In this case, several alternative classes are considered, and compliant OCC is formed to maximize the specificity values for all of these classes. This is done by choosing the α value that would minimize type II errors, β, that is, the risk of accepting an alien sample.

6. Cumulative analytical signal

An analytical signal is an important concept in chemistry, which currently lacks a universal definition. In quantitative analysis, an analytical signal is a part of the data that contains information about the substance of interest (analyte). There is no equivalent definition for qualitative analysis. This Section discusses how it can be defined and why it is needed.

6.1. Analytical signal in qualitative analysis

In the simple case of univariate calibration, an analytical signal is understood as a numerical value such as peak height or area under the peak that correlates with the analyte concentration.[107] Multivariate calibration introduces the analytical signal as a mathematical abstraction that represents a combination of regression coefficients. This quantity is called net analytical signal (NAS), that is, the signal that remains after the influence of other components has been eliminated.[108] This is a number or a vector calculated for each calibration sample as an implicit function of a dataset (**X**, **Y**) and the regression model used for calibration.

An analytical signal is a condensed expression of the relationship between the concentration of interest and the available experimental data. It is convenient for estimating the selectivity and uncertainty and for determining the limits of detection. Therefore, NAS has numerous applications, such as calculation of FoMs, selection of significant variables, detection of outliers, and model optimization.[109] In qualitative analysis, there is no simple objective such as concentration; instead, there are decisions that are difficult to formalize, *e.g.*, YES/NO/UNKNOWN, the relation of which to data properties (peak height and peak area) is not obvious. However, if we take a closer look at the essence of the classification methods used to make these decisions, we can see that a key role is played by distances. The distances determine the structure and interaction of classes; therefore, in qualitative analysis, we must switch from the 'more/less' to 'closer/further' indicators.

Three distances are used in analytical chemistry. The first one is the orthogonal distance (OD), q, a common Euclidean metric, that is, the sum of the squares of the coordinates, defined above in equation (4). It is used universally to assess the approximation error,[24] to verify the similarity of spectra,[34] and to address more challenging problems.[39]

The second one is the score distance (SD), h, the squared Mahalanobis distance, which is defined by equation (5). It is used to evaluate the positions of objects for low-dimensional data, so-called features, which appear upon the use of PCA or PLS. It is encountered in many classification methods such as SIMCA, UNEQ, QDA, and so on.

Finally, the third one is the Y-distance (YD) for calibration errors, z:

$$z_i = \sum_{k=1}^{K} (y_{ki} - \hat{y}_{ki})^2 \tag{23}$$

It is used to detect outliers during calibration.[110]

These three distances form the construction kit for the assembly of cumulative analytical signals (CAS). Each of them is a random variable that obeys the chi-square distribution up to parameters that can be estimated using standard statistical methods. In some cases, it is necessary to use the noncentral chi-square distribution, $\chi'^2(N, s)$, which includes the noncentrality parameter, s, and a scaling factor. Elementary CAS are non-negative random variables with the additivity property: $\chi^2(N_1)+\chi^2(N_2) = \chi^2(N_1+N_2)$. Therefore, they can be combined, and the result will also follow the chi-square distribution. In addition, they are often independent; therefore, the number of degrees of freedom can simply be added together. That is why this analytical signal is called 'cumulative'.

In the DD-SIMCA method, a single value, particularly the full distance FD defined in equation (7), can be assigned to each sample (multivariate 'fingerprint'). This value characterizes the proximity or remoteness of the target sample from the acceptance boundary through equations (8) and (9). Therefore, FD is an example of CAS.

We will consider a few examples of application of the CAS concept that demonstrate its practical significance.

6.2. Limit of detection

The CAS concept makes it possible to develop a new method for calculating the limit of detection for the DD-SIMCA classifier. The main advantage of the CAS approach over existing methods is that it is based on a distribution that is known up to parameters. This opens up the way to simplify the calculation of necessary statistical parameters, such as type I and type II errors, confidence intervals, critical levels, *etc*. Indeed, if the distribution is unknown, then, for example, to estimate the 0.95 quantile, the sample size must be at least 10 times larger than that in the case of a known parametric distribution. In authentication problems addressed using one-class methods, determination of the limits of detection is often irrelevant, since adulteration of many food products such as vegetable oils,[111–113] honey,[114,115] coffee,[116] and spices[117] by low-concentration adulterants is not economically reasonable. However, the adulteration of essential food products and medicines[21,118,119] requires determination of detection limits, since even low concentrations of impurities can pose a health risk, for example, the addition of ammonium chloride and melamine to milk[120] or the contamination of a propylene glycol-based children syrup with diethylene glycol.[121] Pomerantsev *et al*.[122] described in detail an algorithm for the calculation of the limit of detection within a one-class classification problem and gave examples of determination of the limits of detection for talc in wheat flour and tetracycline in blood serum.

Determination of tetracycline in blood serum using synchronous fluorescence spectra was first described in detail by Goicoechea and Olivieri[123] and was subsequently used as an example for calculating the detection limits in calibration[124] and qualitative analysis.[122] The classification problem is addressed by forming a target class consisting of 14 pure blood serum samples free of tetracycline contamination. Of these, ten samples were used to train the model and four samples were employed for testing. Four sets (four samples each) consisted of samples contaminated with tetracycline in increasing concentrations of 0.2, 0.4, 0.6, and 0.8 μg mL^{-1}. They were used as the alternative sets. All samples used in the study are depicted in Fig. 5*b*. The position of each sample indicates its proximity to the acceptance region. Training set samples (open squares) are used to build a model and estimate the boundary of the target class. Test set samples are used to verify that the class boundary was correctly established. All other samples belong to alternative sets. If the threshold is set at a significance level of $\alpha = 0.05$, then half of the samples with a low tetracycline concentration of 0.2 μg mL^{-1} are incorrectly identified as being pure. All other alternative classes are well separated, with samples for each particular concentration being grouped together and distinguished from samples of other classes. This example shows that in the presence of several alternative classes, it is advisable to calculate specificity separately for each alternative class. The total specificity for the whole set of tetracycline-containing samples SPC = 87.5%, for the set with concentration of 0.2 μg mL^{-1}, SPC = 50%, and for any other alternative sets containing ⩾ 0.4 μg mL^{-1} of tetracycline, SPC = 100%.

According to a definition,[125] the limit of detection (LOD) is the minimum amount of the analyte that can be distinguished from the blank with specified probabilities of type I (α) and type II (β) errors. The former error (α) is the probability that the analyte is detected in the sample, while in reality, this sample is blank. This error is characterized by the decision threshold (CC_α), that is, the CAS value corresponding to the critical level. Type II error (β) is the probability that a sample is reported as negative, while the analyte is actually present. This error is characterized by the limit of detection (CC_β), *i.e*., the lowest content of analyte that can be detected in the sample with the probability β. To avoid confusion, it is important to emphasize that CC_α is expressed in the same units as CAS (most often, dimensionless), whereas CC_β is expressed in analyte concentration units.

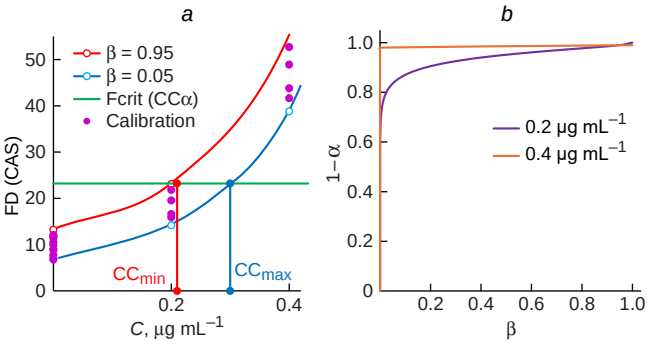


Figure 10. Determination of tetracycline in blood serum: (*a*) calculation of the limit of detection; (*b*) selectivity curves for two alternative classes.

Figure 10*a* illustrates the calculation of two limits of detection for $\alpha = 0.01$, which corresponds to $CC_\alpha = 23$. The minimum concentration at which tetracycline can be detected is $CC_{min} = 0.21$ μg mL^{-1}; however, the probability of error in this case is high and equals 0.95. The maximum concentration at which tetracycline can be detected in blood serum with a low probability of error of 0.05 is $CC_{max} = 0.30$ μg mL^{-1}. The curves (blue and red lines) were plotted using equation (12) inverted to find FD (β).

6.3. Selectivity

Qualitative analysis always includes three main elements: the object, the method, and the tool. The object is a part of reality that is being studied. For example, this can be melamine in powdered milk, which must be detected, or a medication that must be tested for authenticity. The method means the analytical platform used to gain data about the object: spectroscopy, chromatography, *etc*. Finally, the term 'tool' refers to the mathematical procedure used to analyze data and make decisions.

Each element of this triad can be compared to an analogue, provided that the other two elements are identical. For example, it can be stated that NIR spectroscopy is more effective than UV-visible spectroscopy for detecting talc in flour, or that a one-class classifier is more suitable for authentication than a discriminator. This brings us to the concept of selectivity (SLC), which is used to compare objects, methods, and tools and which is considered to be important by many analytical chemists.

It is not entirely clear how to define SLC in qualitative analysis. Only two facts are known for certain: (1) SLC is a limiting case of SPC[126] and (2) selectivity is the ability to minimize false positive results.[127]

In the DD-SIMCA method, the probability of false positive results is expressed in terms of type II error (β), which is calculated as described by Pomerantsev and Rodionova.[84] Thus, in the context of qualitative analysis, the following definition of

selectivity is proposed. Selectivity is the expected probability of true negative results:[128]

$$\mathrm{SLC} = 1 - \beta \tag{24}$$

It is important to find the right place for SLC among other figures of merit. SNS and SPC are empirical characteristics determined by counting the number of true positive and true negative decisions; therefore, they are subject to errors. Another pair consists of the probabilities α and β, one being set *a priori* (most often, α) and the other being calculated theoretically. SPC is an empirical estimate of the probability that the analytical process correctly classifies a non-target sample as not belonging to the target class. SLC is the theoretically expected limiting value of SPC that can be achieved in an ideal experiment in which the sampling variability is negligibly small or the sampling size is very large.

Consider how this works for the tetracycline example. If the value of α is changed to α = 0.15 (the dashed line in the plot, Fig. 5 *b*), then all samples containing tetracycline will be correctly classified as not belonging to the target class, *i.e.*, SPC = 100% for each alternative class. However, it is clear that the 0.2 μg mL^{-1} class is located much closer to the acceptance region than the 0.4 μg mL^{-1} class and, all the more, closer than other alternative samples. In this case, for the class of 0.2 μg mL^{-1}, SLC = 93%, since β = 0.07, while for the class of 0.4 μg mL^{-1}, SLC = 100% and $\beta = 7 \times 10^{-17}$.

Since SLC depends on the chosen value of α, it is useful to represent it as the curve shown in Fig. 10 *b*. This curve resembles the ROC curve. The area under this curve (AUC) can be used as an integral measure of selectivity. For example, for the 0.2 μg mL^{-1} class, AUC = 94%, while for the 0.4 μg mL^{-1} class, AUC = 100%; hence, the latter class is more selective.

The selectivity can be applied at various stages of qualitative analysis. First, to compare different datasets, for example, to determine which of the alternative sets is closer to the target class (Fig. 10 *b*). Second, the selectivity helps to understand which instrumental method has higher selectivity when applied to the same set of samples (Fig. 11 *a*). Third, the selectivity, especially in its integral form, allows for the comparison of data analysis tools[129] (Fig. 11 *b*).

6.4. Detection of outliers

A modern approach to outlier detection is based on the philosophy of machine learning. It includes the simultaneous use of various outlier detection techniques.[130,131] An example is a general-purpose tool that uses the sum of ranking differences,[132] which is sensitive to various types of outliers and requires significant computational resources.

Rodionova and Pomerantsev[110] proposed a simple approach to outlier detection that was called sequential focused trimming. For this purpose, the total distance TD (Total Distance), g, is used, which is a CAS composed of three basic elements:

$$g = N_h \frac{h}{h_0} + N_q \frac{q}{q_0} + N_z \frac{z}{z_0} \sim \chi^2(N_g) \tag{25}$$

where $N_g = N_h + N_q + N_z$. For the chosen significance level γ (an admissible false positive error), the decision rule is specified by the following condition:

$$g > \chi^{-2}\left((1-\gamma)^{1/I}, N_g\right) \tag{26}$$

where I is the number of samples in the dataset.

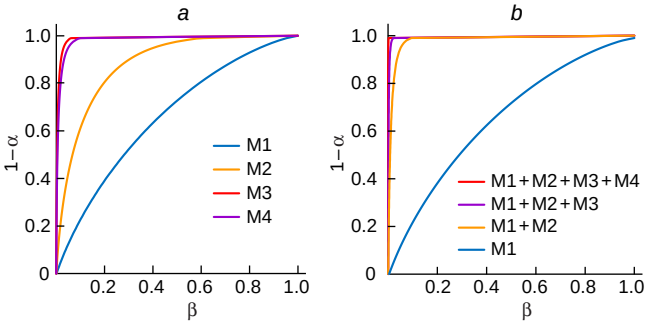


Figure 11. Authentication of olive oil. (*a*) Selectivity of various instrumental methods: M1, M2, ..., M4. (*b*) Changes in the selectivity upon successive fusion of the data blocks obtained using different tools: M1, M1+M2, ..., M1+M2+M3+M4.

The procedure starts with a model that uses the entire training set. This serves to estimate all three basic elements of CAS: q (OD), h (SD), and z (YD) [equations (4), (5), and (23)] and to calculate CAS g (TD). Unlike the conventional approach, which uses two separate cut-off thresholds for the X and Y residuals, the method based on CAS calculates a joint critical area. The advantages of this approach are as follows:[82] (1) all distribution parameters are calculated from the data; therefore, this is a data driven approach; (2) both classical and robust methods are used to estimate the parameters; (3) the probability γ used to calculate the boundary for the outlier region is specified *a priori* and has a statistical meaning; (4) the detected outliers are eliminated sequentially, one by one, which makes it possible to avoid the masking and swamping effects.

6.5. Data fusion

Data fusion methods are used for joint analysis of the data obtained using various analytical platforms for the same samples.[133] In this context, three levels of fusion are distinguished. The first one, also called low, involves merging of the results of measurements (blocks) obtained using various instruments into a single data block, which is then utilized to build a single common model. According to the mid-level fusion, each block is separately processed to extract information; for example, principal components that are identified using the PCA method, variables that are selected, and so on. After that, the features identified for each block are combined into a common predictor matrix for which the model is built.[134]

High-level fusion is performed by combining the modelling results for each data block. In practice, these methods are rarely encountered, as they are more complex than first- and second-level methods.[135] The use of the CAS-based approach makes high-level fusion simple and straightforward, since it is based on combining the single CAS values obtained by modelling separate data blocks. Any number of diverse blocks can be combined in this way. Although high-level fusion finally results in a fairly complex model, analysis of partial FD distances for each of the blocks makes it possible to establish a significant relationship between the classification results and the effect of particular samples and instruments. For example, the possibility of authentication of olive oil samples was investigated both using particular instruments (Fourier transform NIR spectroscopy, UV spectroscopy, electronic tongue, and electronic nose) and using joint classification incorporating all four analytical methods.[136] The result shown in Fig. 11 illustrates the benefits of data fusion.

Lozano *et al*.[129] used CAS-based approach for fusing a three-dimensional tensor of fluorescence data and a two-dimensional matrix of spectral data.

6.6. Selection of sample subset

Analysis of large datasets faces the challenge of selecting a subset of samples that meets particular criteria from the whole original dataset. This problem can be split into two tasks.[101] The first one is the choice of a reduced dataset that does not compromise the prediction accuracy, but is markedly smaller than the original dataset. This subset is typically used for calibration transfer[137] or for image analysis.[138] The second one is the selection of the test set. On the one hand, the test set should reflect the variability of the target class as fully as possible; on the other hand, it should differ from the training set. Typically, the test set comprises 25–30% of the total dataset and is used to optimize and validate the model.[97,139] The CAS concept can be used for effective selection of these subsets for both single-block (classification) and multi-block (calibration) data. As a result of simple calculations, sample importance (SI) index based on the full distance FD is assigned to each sample in the total dataset. The sample importance index takes values in the (0, 1) range, with the higher values corresponding to more important samples. This allows for a targeted, rather than random, selection of samples with specified properties. Samples with the maximum SI values are selected for the reduced set. In the formation of the test set, samples with the minimum and maximum SI values are excluded, and then a specified proportion is selected at random from the remaining samples. It was shown that the proposed method is more efficient than the Kennard–Stone algorithm[98] or D-optimal planning.[99] In addition, it was shown[140] that the importance index slightly depends on the model complexity, which breaks the vicious cycle in which the representative sampling depends on the model complexity, while the model complexity is tested against the representative sampling.

7. Prospects

It is always difficult to make predictions, especially about the future, but we will try. Previously, it was noted that every analytical procedure includes three components: object of the study, instrumental method, and data analysis tool. We will state our ideas on the prospects of qualitative analysis in terms of this structure.

7.1. Objects

Currently, qualitative analysis has been widely and successfully applied to study food products,[141,142] animal feed,[143] pharmaceuticals,[144] and the environment;[145] however, more important and complex objects related to medicine[146,147] and biology[148] are now on the horizon. First of all, this is diagnosis and personalized therapy.[149] Here, the focus is on the qualitative classification of disease stages and types and selection of therapy based on the patient's personal data.[150] In this field, qualitative analysis is difficult to perform for one reason, which can be summed up simply as human factor. First, the subjects themselves, which are patients, differ from one another much more appreciably than, *e.g*., samples of oil or tablets, which means higher variability with inevitable extremes and outliers. Second, there are no reliable reference methods, and all model training relies on the subjective opinions of experts, who are prone to mistakes.

7.2. Methods (instruments)

Here, two trends can be distinguished. The first one is the miniaturization of instruments. The fact that NIR spectrometers can be pocket-size devices is no longer surprising to anyone, but portable mass spectrometers[151] and even NMR spectrometers[152] are also appearing. Certainly, they are inferior to their desktop analogues, but they open up new opportunities for conducting qualitative analysis in the field.

The second trend is called point of care (POC):[153,154] this is medical diagnosis performed directly at the patient location in a clinic, in a pharmacy, or at home. This approach provides quick results without the need for specialized laboratories. This is done using compact devices and soft sensors,[155–158] which differ radically from the simple sensors used to detect target substances. This field is experiencing a real boom: suffice is to say that in just four months of 2026, 136 papers on POC were published in *Microchemical J*. alone, one of the publications being from Russia.[159]

7.3. Data analysis tools

Everything that is now happening in the world and will continue to happen for some time in the future is concentrated on artificial intelligence (AI). This term is misleading, but it has already become deeply ingrained in the consciousness of the general public, including scientists. This term is used, most often, to mean large language models (LLMs), that is, ANNs trained on huge datasets to understand and generate texts. They are good at tasks such as translation and abstracting. Examples are GPT-4, Claude, *etc*.

It is believed that AI can serve as a useful tool, especially in the field of image analysis. However, this does not mean that it should replace human decision makers, since recognition requires taking into account a multitude of additional factors that constitute domain knowledge.[160] Integration of the domain knowledge could bridge the gap between data analysis and decision making.[161,162]

In our opinion, this integration should be carried out at two levels. At the first basic level, the ML model is adapted to take account of the key characteristics of the domain, constraints, logical rules, *etc*. The second level is more complex, since it aims to take into account the human factor. This is achieved through collaboration between a group of subject matter experts and a group of data analysts, meant to address any shortcomings resulting from the decisions made by each group. The process begins with data collection, continues through the exploratory analysis phase, and concludes with the decision making stage.

8. List of abbreviations and symbols

This following abbreviations and symbols are used in the review:

α — significance level for acceptance,
γ — significance level of outlier,
χ'^2 — noncentral chi-square distribution,
χ^{-2} — quantile of the chi-square distribution,
AI — artificial intelligence,
ANN — artificial neural networks,
AUC — area under curve,
CAS — cumulative analytical signal,
DD-SIMCA — data driven SIMCA,
DL — deep learning,
EFF — efficiency,
FD — full distance,

FoM — figures of merit,
FN — false negative,
FP — false positive,
I — number of samples,
IUPAC — International Union of Pure and Applied Chemistry,
J — number of variables,
K — number of classes,
kNN — k-nearest neighbours,
LDA — linear discriminant analysis,
LOD — limit of detection,
LV — latent variables,
ML — machine learning,
N — number of degrees of freedom,
NAS — net analytical signal,
OCC — one class classifier,
OC-PLS — one class PLS,
OC-RF — one-class random forest,
OD — orthogonal distance,
P — loading matrix,
PCA — principal component analysis,
PCR — principal component regression,
PCV — Procrustes cross-validation,
PLS — partial least squares,
PLS-DA — PLS discriminant analysis,
POC — point of care,
PV set — pseudo validation set,
QDA — quadratic discriminant analysis,
RF — random forest,
RMSE — root mean squared error,
ROC — receiver operating characteristic,
SD — score distance,
SEL — selectivity,
SI — sample importance,
SIMCA — soft independent modelling of class analogy,
SNS — sensitivity,
SPC — specificity,
SVM — support vector machine,
T — score matrix,
TD — total distance,
TEFF — total efficiency,
TN — true negative,
TP — true positive,
TSNS — total sensitivity,
TSPC — total specificity,
UNEQ — unequal class models,
YD — Y-distance.

9. References

1. L.A.Currie. *Pure Appl. Chem.*, **67**, 1699 (1995); http://dx.doi.org/10.1351/pac199567101699
2. V.V.Apyari, Yu.A.Zolotov, S.G.Dmitrienko, A.A.Furletov, T.I.Tikhomirova. *J. Anal. Chem.*, **80**, 385 (2025); https://doi.org/10.1134/S1061934824701867
3. B.L.Milman, I.K.Zhurkovich. *J. Anal. Chem.*, **77**, 537 (2022); https://doi.org/10.1134/S1061934822050070; https://doi.org/10.31857/S0044450222050085]
4. X.Zhang, F.Zhan, C.Hao, Y.-D.Lei, F.Wania. *Environ. Sci. Technol.*, **59**, 3121 (2025); https://doi.org/10.1021/acs.est.4c10049
5. M.J.Strynar. *Environ. Int.*, **197**, 109332 (2025); https://doi.org/10.1016/j.envint.2025.109332
6. Commission Decision (EEC) No. 657, *Off. J. Eur. Commun*., L221, 8 (2002)
7. L.Strani, M.Cocchi, D.Tanzilli, A.Biancolillo, F.Marini, R.Vitale. *TrAC Trend. Anal. Chem.*, **183**, 118117 (2025); https://doi.org/10.1016/j.trac.2024.118117
8. L.Cuadros, L.Valverde-Som, A.M.Jimenez-Carvelo, M.Delgado-Aguilar. *TrAC Trend. Anal. Chem.*, **122**,1 15705, (2020); https://doi.org/10.1016/j.trac.2019.115705
9. B.Giussani, G.Gorla, J.Ezenarro, J.Riu, R.Boque. *TrAC Trend. Anal. Chem.*, **181**, 118051 (2024); https://doi.org/10.1016/j.trac.2024.118051
10. L.Brunnbauer, Z.Gajarska, H.Lohninger, A.Limbeck. *TrAC Trend. Anal. Chem.*, **159**, 116859 (2023); https://doi.org/10.1016/j.trac.2022.116859
11. D.D.Matyushin, A.Y.Sholokhova, A.K.Buryak. *Anal. Chem*., **92**, 11818 (2020); https://doi.org/10.1021/acs.analchem.0c02082
12. A.Samokhin, K.Sotnezova, I.Revelsky. *Eur. J. Mass Spectrom*., **25**, 439 (2019); https://doi.org/10.1177/1469066719855503
13. A.Samokhin, M.Khrisanfov. *J. Am. Soc. Mass Spectrom.*, **37**, 264 (2026); https://doi.org/10.1021/jasms.5c00322
14. M.Terán, J.J.Ruiz, P.Loza-Alvarez, D.Masip, D.Merino. *Chemom. Intell. Lab. Syst*., **264**, 105476 (2025); https://doi.org/10.1016/j.chemolab.2025.105476
15. B.L.Milman. *TrAC Trend. Anal. Chem.*, **69**, 24 (2015); https://doi.org/10.1016/j.trac.2014.12.009
16. M.R.Alcaraz, E.Martín-Tornero, H.C.Goicoechea, A.Muñoz de la Peña, I.Durán-Merás. *Talanta*, **301**, 129301 (2026); https://doi.org/10.1016/j.talanta.2025.129301
17. M.Afsharipour, M.Shamsi, F.Ghasemian, H.Khabazzadeh. *Microchem. J.*, **224**, 117583 (2026); https://doi.org/10.1016/j.microc.2026.117583
18. I.Rehan, K.Rehan, M.Aamir, S.Islam. *Microchem. J.*, **220**, 116416 (2026) https://doi.org/10.1016/j.microc.2025.116416
19. J.R.Belmonte-Sánchez, R.Romero-González, J.A.Tello-Jiménez, A.Garrido Frenich. *Food Control*, **179**, 111610 (2026); https://doi.org/10.1016/j.foodcont.2025.111610
20. E.Sufriadi, R.Idroes, H.Meilina, A.A.Munawar, G.Indrayanto. *ACS Omega*, **8**, 12348 (2023); https://doi.org/10.1021/acsomega.3c00080
21. O.Ye. Rodionova, A.V.Titova, K.S.Balyklova, A.L.Pomerantsev. *Talanta*, **205**, 120150 (2019); https://doi.org/10.1016/j.talanta.2019.120150
22. M.Arif, G.Chilvers, S.Day, S.A.Naveed, M.Woolfe, O.Ye.Rodionova, A.L.Pomerantsev, O.Kracht, C.Brodie, A.Mihailova, A.Abrahim, A.Cannavan, S.D.Kelly. *Food Control*, **123**, 107827 (2021); https://doi.org/10.1016/j.foodcont.2020.107827
23. R.G.Brereton, J.Jansen, J.Lopes, F.Marini, A.Pomerantsev, O.Rodionova, J.M.Roger, B.Walczak, R.Tauler. *Anal. Bioanal. Chem*., **410**, 6691 (2018); https://doi.org/10.1007/s00216-018-1283-4
24. H.Martens, T.Naes. *Multivariate Calibration*. (New York: Wiley, 1989)
25. E.Szymanska, J.Gerretzen, J.Engel, B.Geurts, L.Blanchet., L.M.C.Buydens. *TrAC Trend. Anal. Chem.*, **69**, 34 (2015); https://doi.org/10.1016/j.trac.2015.02.015
26. S.Cárdenas, M.Valcárcel. *TrAC Trend. Anal. Chem.*, **24**, 477 (2005); https://doi.org/10.1016/j.trac.2005.03.006
27. Yu.A.Zolotov. *J. Anal. Chem.*, **57**, 561 (2002); https://doi.org/10.1023/A:1016213114511
28. P.Oliveri, G.Downey. *TrAC Trend. Anal. Chem.*, **35**, 74 (2012); https://doi.org/10.1016/j.trac.2012.02.005
29. D.Ballabio, F.Grisoni, R.Todeschini. *Chemom. Intell. Lab. Syst*., **174**, 33 (2018); https://doi.org/10.1016/j.chemolab.2017.12.004
30. O.Ye.Rodionova, P.Oliveri, A.L.Pomerantsev. *Chemom. Intell. Lab. Syst.*, **159**, 89 (2016); http://dx.doi.org/10.1016/j.chemolab.2016.10.002
31. C.Beleites, R.Salzer, V.Sergo. *Chemom. Intell. Lab. Syst.*, **122**, 12 (2013); https://doi.org/10.1016/j.chemolab.2012.12.003

32. C.Durante, R.Bro, M.Cocchi. *Chemom. Intell. Lab. Syst.,* **106**, 73 (2011); https://doi.org/10.1016/j.chemolab.2010.09.004
33. P.Turova, I.Rodin, O.Shpigun, A.Stavrianidi. *Phytochem. Anal.*, **31**, 948 (2020); https://doi.org/10.1002/pca.2967
34. N.Rahmani, A.Varnosfaderani. *Microchem. J.*, **223**, 117392 (2026); https://doi.org/10.1016/j.microc.2026.117392
35. K.K.Ejeahalaka, S.L.W.On. *Food Chem.*, **295**, 198 (2019); https://doi.org/10.1016/j.foodchem.2019.05.120
36. S.Wartewig (Ed.). In *IR and Raman Spectroscopy.* (New York: Wiley, 2003); https://doi.org/10.1002/3527601635.ch3
37. R.De Maesschalck, D.Jouan-Rimbaud, D.L.Massart. *Chemom. Intell. Lab. Syst*., **50**, 1 (2000); https://doi.org/10.1016/S0169-7439(99)00047-7
38. Classification Toolbox for Matlab, https://michem.unimib.it/download/matlab-toolboxes/classification-toolbox-for-matlab/
39. O.Y.Rodionova, A.L.Pomerantsev. *Microchem. J*., **220**, 116662 (2026); https://doi.org/ 10.1016/j.microc.2025.116662
40. S.Kucheryavskiy. *J. Chemom.*, **32**, e2966 (2018); https://doi.org/10.1002/CEM.2966
41. R.A.Fisher. *Ann. Eugen.*, **7**, 179 (1936); https://doi.org/10.1111/j.1469-1809.1936.tb02137.x
42. C.R.Rao, J.R.Stat. *Proc. R. Soc. B: Biol. Sci.*, **10**, 159 (1948); https://doi.org/10.1111/j.2517-6161.1948.tb00008.x
43. T.Cover, P.Hart. *IEEE Trans. Inf. Theory*, **13**, 21 (1967); https://doi.org/10.1109/TIT.1967.1053964
44. C.Cortes, V.Vapnik. *Mach. Learn.*, **20**, 273 (1995); https://doi.org/10.1007/BF00994018
45. S.Kucheryavskiy. *J. Anal. Test.*, **2**, 274 (2018); https://doi.org/10.1007/s41664-018-0078-0
46. L.Breiman. *Mach. Learn.*, **45**, 5; (2001); https://doi.org/10.1023/A:1010933404324
47. T.Chen, C.Guestrin. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. (San Francisco, USA, 2016). P. 785; https://doi.org/10.1145/2939672.2939785
48. J.Zupan, J.Gasteiger. *Neural Networks for Chemists: An Introduction*. (New York: Wiley, 1993)
49. A.Bos, M.Bos, W.E.van der Linden. *Anal. Chim. Acta*, **256**, 133 (1992); https://doi.org/10.1016/0003-2670(92)85338-7
50. I.I.Baskin, T.I.Madzhidov, I.S.Antipin, A.A.Varnek. *Rus. Chem. Rev*., **86**, 1127 (2017); https://doi.org/10.1070/RCR4746
51. O.E.Rodionova, A.L Pomerantsev. *J. Anal. Chem*., **81**, 614 (2026); https://doi.org/10.7868/S3034512X26040122
52. A.L Pomerantsev, O.E.Rodionova. *J. Anal. Chem*., **81**, 618 (2026); https://doi.org/10.7868/S3034512X26040132
53. S.Wold, K.H.Esbensen, P.Geladi. *Chemom. Intell. Lab. Syst*., **2**, 37 (1987); https://doi.org/10.1016/0169-7439(87)80084-9
54. A.A.D'Archivio, M.L.Di Vacri, M.Ferrante, M.A.Maggi, S.Nisi, F.Ruggieri. *Food Anal. Meth*., **12**, 2572 (2019); https://doi.org/10.1007/s12161-019-01610-8
55. B.Debus, H.Parastar, P.Harrington, D.Kirsanov. *TrAC Trend. Anal. Chem.*, **145**, 116459 (2021); https://doi.org/10.1016/j.trac.2021.116459
56. S.Kucheryavskiy. *Anal. Chim. Acta*, **1407**, 345489 (2026); https://doi.org/10.1016/j.aca.2026.345489
57. M.Forina, P.Oliveri, S.Lanteri, M.Casale. *Chemom. Intell. Lab. Syst.*, **93**,132 (2008); https://doi.org/10.1016/j.chemolab.2008.05.003
58. A.L.Pomerantsev, O.Ye.Rodionova. *J. Chemom.*, **32**, e3030 (2018); https://doi.org/10.1002/cem.3030
59. S.D.Brown, L.Chen. *ACS Symp. Ser.*, **1199**, 159 (2015); https://doi.org/10.1021/bk-2015-1199.ch007
60. M.Barker, W.S.Rayens. *J. Chemom.*, **17**, 166 (2003); https://doi.org/10.1002/cem.785
61. D.N.Artemyev, L.A.Bratchenko. I.A.Matveeva, V.I.Kukushkin, D.V.Lystsev, A.I.Ischenko, A.A.Ischenko, V.M.Zuev, V.P.Zakharov. *J. Biomed. Photonic. Eng.*, **10**, 20307 (2024); https://doi.org/10.18287/JBPE24.10.020307
62. Y.Khristoforova, L.Bratchenko, V.Kupaev, D.Senyushkin, M.Skuratova, S.Wang, P.Lebedev, I.Bratchenko. *Diagnostics*, **15**, 660 (2025); https://doi.org/10.3390/diagnostics15060660
63. Y.B.Monakhova, B.W.K.Diehl, J.Fareed. *J. Pharm. Biomed. Anal.*, **149**,114 (2018); https://doi.org/10.1016/j.jpba.2017.10.020
64. D.A.Metlenkin, Y.T.Platov, R.A.Platova, E.V.Zhirkova, O.T.Teneva. *Agron. Res*., **20**, 326 (2022); https://doi.org/10.15159/AR.22.027
65. Y.V.Zubritskaya, A.V.Shik, I.A.Stepanova, S.A.Zolotov, P.Y.Borshchegovskaya, U.A.Bliznyuk, I.A Ananieva, A.P.Chernyaev, I.A Rodin, M.K.Beklemishev. *Foods*, **14**, 4285 (2025); https://doi.org/10.3390/foods14244285
66. A.S.Mukhametganeeva, A.R.Bayguzina. *Analitika i Kontrol*, **27**, 276 (2023); https://doi.org/10.15826/analitika.2023.27.4.009
67. P.N.Berezov, V.E.Goleva. *Analitika i Kontrol*, **27**, 292 (2023); https://doi.org/10.15826/analitika.2023.27.4.011
68. O.Ye. Rodionova, A.V.Titova, A.L.Pomerantsev. *TrAC Trend. Anal. Chem.*, **78**, 17 (2016); https://doi.org/10.1016/j.trac.2016.01.010
69. L.Xu, S.-M.Yan, C.-B.Cai, X.-P.Yu. *Chemom. Intell. Lab. Syst*., **126**, 1 (2013); https://doi.org/10.1016/j.chemolab.2013.04.008
70. D.M.J.Tax, R.P.W.Duin. *Patt. Recogn. Lett.*, **20**, 1191 (1991); https://doi.org/10.1016/s0167-8655(99)00087-2
71. B.Schölkopf, J.C.Platt, J.Shawe-Taylor, A.J.Smola, R.C.Williamson. *Neural Comput*., **13**, 1443 (2001); https://doi.org/10.1162/089976601750264965
72. Z.Huang, B.Li, S.Wang, R.Zhu, X.Cui, X.Yao. *Food Anal. Meth*., **16**, 933 (2023); https://doi.org/10.1007/s12161-023-02459-8
73. S.Wold, M.Sjostrom. In *Chemometrics Theory and Application, American Chemical Society Symposium Series.* Vol. 52. (Ed. B.R.Kowalski). (Washington, DC: American Chemical Society, 1977). P. 243
74. B.R.Kowalski, S.Wold. In *Handbook of Statistics, 2*. (Eds P.R.Krishnaiah, L.N.Kanal). (North-Holland Publishing Company, 1982). P. 673
75. A.L.Pomerantsev, O.Y.Rodionova. *J. Chemom*., **34**, e3250 (2020); https://doi.org/10.1002/cem.3250
76. R.Vitale, M.Cocchi, A.Biancolillo, C.Ruckebusch, F.Marini. *Anal. Chim. Acta*, **1270**, 341304 (2023); https://doi.org/10.1016/j.aca.2023.341304
77. P.Oliveri, M.I.Lopez, M.C.Casolino, I.Ruisanchez, M.P.Callao, L.Medini, S.Lanteri. *Anal. Chim. Acta*, **851**, 30 (2014); https://doi.org/10.1016/j.aca.2014.09.013
78. E.S.Ogorodnikova. *Analitika i Kontrol*, **27**, 267 (2023); https://doi.org/10.15826/analitika.2023.27.4.008
79. S.I.Azarenko, I.O.Kuznetsov. *Analitika i Kontrol*, **27**, 285 (2023); https://doi.org/10.15826/analitika.2023.27.4.01
80. Y.Liu, H.H.Zhang, Y.Wu. *J. Am. Stat. Assoc*., **106**, 166 (2011); https://doi.org/10.1198/jasa.2011.tm10319
81. A.L.Pomerantsev, O.Ye.Rodionova. *Microchem. J.*, **212**, 113329 (2025); https://doi.org/10.1016/j.microc.2025.113329
82. S.Kucheryavskiy, O.Rodionova, A.Pomerantsev. *J. Chemom*., **38**, e3556; (2024); https://doi.org/10.1002/cem.3556
83. O.Ye.Rodionova, A.L.Pomerantsev. *TrAC Trend. Anal. Chem.*, **29**, 79 (2010); https://doi.org/10.1016/j.trac.2010.05.004
84. A.L.Pomerantsev, O.Ye.Rodionova. *J. Chemom.*, **28**, 518 (2014); https://doi.org/10.1002/cem.2610
85. O.Ye.Rodionova, K.S.Balyklova, A.V.Titova, A.L.Pomerantsev. *J. Pharm. Biomed. Anal.*, **98**, 186 (2014); https://doi.org/10.1016/j.jpba.2014.05.033
86. S.L.R.Ellison, T.Fearn. *TrAC Trend. Anal. Chem.*, **24**, 468 (2005); https://doi.org/10.1016/j.trac.2005.03.007
87. P.I.Monteiro, J.Sousa Santos, O.Ye.Rodionova, A.Pomerantsev, E.Sidinei Chaves, N.Deliberali Rosso, D.Granato. *J. Food Sci.*, **84**, 3099 (2019); https://doi.org/10.1111/1750-3841.14815
88. C.Ferri, J.Hernandez-Orallo, R.Modroiu. *Patt. Recogn. Lett.*, **30** (1), 27 (2009); https://doi.org/10.1016/j.patrec.2008.08.010
89. M.Sokolova, G.Lapalme. *Inf. Process. Manag.*, **45**, 427 (2009) https://doi.org/10.1016/j.ipm.2009.03.002

90. G.Jurman, S.Riccadonna, C.Furlanello. *PLoS One*, **7**, e41882, (2012); https://doi.org/10.1371/journal.pone.0041882
91. A.L.Pomerantsev, O.Ye.Rodionova. *TrAC Trend. Anal. Chem.*, **143**, 116372 (2021); https://doi.org/10.1016/j.trac.2021.116372
92. E.Trullols, I.Ruisanchez, F.Xavier Rius. *TrAC Trend. Anal. Chem.*, **23**, 137 (2004); https://doi.org/10.1016/S0165-9936(04)00201-8
93. R.L.Green, J.H.Kalivas. *Chemom. Intell. Lab. Syst.*, **60**, 173 (2002); https://doi.org/10.1016/S0169-7439(01)00194-0
94. R.M.Everson, J.E.Fieldsend. *Patt. Recogn. Lett*., **27**, 918 (2006); https://doi.org/10.1016/j.patrec.2005.10.016
95. L.Wang; L.Carvalho. *arXiv*: 2404.13147 (2024); https://doi.org/10.48550/arXiv.2404.13147
96. J.Camacho. *J. Chemom*., **40**, e70110 (2026); https://doi.org/10.1002/cem.70110
97. K.H.Esbensen, P.Geladi. *J. Chemom*., **24**, 168 (2010); https://doi.org/10.1002/cem.1310
98. R.W.Kennard, L.A.Stone. *Technometrics*, **11**, 137 (1969); https://doi.org/10.1080/00401706.1969.10490666
99. J.Ferry, F.X.Rius. *TrAC Trend. Anal. Chem.*, **16**, 70 (1997); https://doi.org/10.1016/S0165-9936(96)00084-2
100. N.B.Gallagher, D.O'Sullivan. *Selection of Representative Learning and Test Sets using the Onion Method*. (Eigenvecror, 2019); https://eigenvector.com/wp-content/uploads/2022/10/Onion_SampleSelection.pdf
101. A.L.Pomerantsev, O.Ye. Rodionova. *Microchem. J*., **190**, 108654 (2023); https://doi.org/10.1016/j.microc.2023.108654
102. R.Bro, K.Kjeldahl, A.K.Smilde, H.Kiers. *Anal. Bioanal. Chem*., **390**, 1241 (2008); https://doi.org/10.1007/s00216-007-1790-1
103. C.A.Ramezan, T.A.Warner, A.E.Maxwell. *Remote Sens.*, **11**, 185 (2019; https://doi.org/10.3390/rs11020185
104. P.Filzmoser, B.Liebmann, K.Varmuza. *J. Chemom*., **23**, 160 (2009); https://doi.org/10.1002/cem.1225
105. S.Kucheryavskiy, S.Zhilin, O.Rodionova, A.Pomerantsev. *Anal. Chem*., **92**, 11842 (2020); https://doi.org/10.1021/acs.analchem.0c02175
106. A.L.Pomerantsev, O.Ye. Rodionova. *Talanta*, **226**, 122104 (2021); https://doi.org/10.1016/j.talanta.2021.122104
107. I.Lavagnini, F.Magno. *Mass Spectrom. Rev.*, **26**, 1 (2007); https://doi.org/10.1002/mas.20100
108. A.Lorber, K.Faber, B.R.Kowalski. *Anal. Chem.*, **69**, 1620 (1997); https://doi.org/10.1021/ac960862b
109. J.Ferry, N.M.Faber. *Chemom. Intell. Lab. Syst.*, **69**, 123 (2003); https://doi.org/10.1016/S0169-7439(03)00118-7
110. O.Ye.Rodionova, A.L.Pomerantsev. *Anal. Chem*., **92**, 2656 (2020); https://doi.org/10.1021/acs.analchem.9b04611
111. P.Cruz-Tirado, D.Muñoz-Pastor, I.A.de Moraes, A.F.Lima, H.T.Godoy, D.F.Barbin, R.Siche. *Chemom. Intell. Lab. Syst.*, **242**, 105004 (2023); https://doi.org/10.1016/j.chemolab.2023.105004
112. T.C.M.Pastore, L.R.Braga, D.C.G.da C.Kunze, L.F.Soares, F.Pastore, A.C.de O.Moreira, P.V.dos Anjos, C.S.Lara, V.T.R.Coradin, J.W.B.Braga. *Microchem. J.*, **182**, 107916 (2022); https://doi.org/10.1016/j.microc.2022.107916
113. V.G.Amelin, O.E.Emel'yanov. *J. Anal. Chem*., **80**, 354 (2025); https://doi.org/10.1134/S1061934824701843
114. D.Dimakopoulou-Papazoglou, S.Serrano, I.Rodríguez, N.Ploskas, K.Koutsoumanis, E.Katsanidis. *J. Food Comp. Anal.*, **144**, 107778 (2025); https://doi.org/10.1016/j.jfca.2025.107778
115. A.L.Panasyuk, E.I.Kuzmina, D.A.Sviridov, M.Yu.Ganin. *Food Systems*, **6**, 211 (2023); https://doi.org/10.21323/2618-9771-2023-6-2-211-223
116. L.Tessaro, Y.da Silva Mutz, M.M.de Bem, N.de Oliveira Souza, C.A.Nunes. *J. Sci. Food Agric*., **106**, 4038 (2026); https://doi.org/10.1002/jsfa.70493
117. V.Ferrari, R.Calvini, C.Menozzi, A.Ulrici, M.Bragolusi, R.Piro, A.Tata, M.Suman, G.Foca. *Chemom. Intell. Lab. Syst*., **249**, 105133 (2024); https://doi.org/10.1016/j.chemolab.2024.105133
118. Y.B.Monakhova, U.Holzgrabe, B.W.K.Diehl. *J. Pharm. Biomed. Anal*., **147**, 580 (2018); https://doi.org/10.1016/j.jpba.2017.05.034
119. N.A.Burmistrova, P.M.Soboleva, Y.B.Monakhova. *J. Pharm. Biomed. Anal.*, **194**, 113811 (2021); https://doi.org/10.1016/j.jpba.2020.113811
120. J.Müller-Maatsch, M.Alewijn, M.Wijtten, Y.Weesepoel. *Food Control,* **121**, 107744 (2021); https://doi.org/10.1016/j.foodcont.2020.107744
121. X.Li, S.Arzhantsev, J.F.Kauffman, J.A.Spencer. *J. Pharm. Biomed. Anal*., **54**, 1001 (2011); https://doi.org/10.1016/j.jpba.2010.11.042
122. A.L.Pomerantsev, D.N.Vtyurina, O.Ye.Rodionova. *Microchem. J.*, **195**, 109490 (2023); https://doi.org/10.1016/j.microc.2023.109490
123. H.C.Goicoechea, A.C.Olivieri. *Anal. Chem*., **71**, 4361 (1999); https://doi.org/10.1021/ac990374e
124. F.A.Allegrini, A.C.Olivieri. *Anal. Chem*., **86**, 7858 (2014); https://doi.org/10.1021/ac501786u
125. T.Wenzl, J.Haedrich, A.Schaechtele, P.Robouch, J.Stroka. *Guidance Document on the Estimation of LOD and LOQ for Measurements in the Field of Contaminants in Feed and Food, EUR 28099*. (Luxembourg: Publications Office of the European Union, 2016). P. 1; https://dx.doi.org/10.2787/8931
126. J.Vessman, R.I Stefan, J.F.Van Staden, K.Danzer, W.Lindner, D.T.Burns, A.Fajgelj, H.Müller. *Pure Appl. Chem.*, **73**, 1381(2001); https://doi.org/10.1351/pac200173081381
127. B.Ng, N.Quinete, P.R.Gardinali. *Sci. Total Environ.*, **713**, 136568 (2020); https://doi.org/10.1016/j.scitotenv.2020.136568
128. A.L.Pomerantsev, O.Y.Rodionova. *Anal. Chim. Acta*, **1332**, 343352 (2024); https://doi.org/10.1016/j.aca.2024.343352
129. V.A.Lozano, A.M.Jimenez Carvelo, A.C.Olivieri, S.V.Kucheryavskiy, O.Ye. Rodionova, A.L.Pomerantsev. *Food Chem.*, **463**, 141127 (2025); https://doi.org/10.1016/j.foodchem.2024.141127
130. R.Domingues, M.Filippone, P.Michiardi, J.Zouaoui. *Pattern. Recognit*., **74**, 406 (2018); https://doi.org/10.1016/j.patcog.2017.09.037
131. J.R.Pasillas-Diaz, S.Ratt. *Electron. Notes Theor. Comput. Sci.*, **329**, 61 (2016); https://doi.org/10.1016/j.entcs.2016.12.005
132. B.Brownfield, J.H.Kalivas. *Anal. Chem.*, **89**, 5087 (2017); https://doi.org/10.1021/acs.analchem.7b00637
133. P.Mishra, J.-M.Roger, D.J.-R.Bouveresse, A.Biancolillo, F.Marini, A.Nordon, D.N.Rutledge. *TrAC Trend. Anal. Chem.*, **137**, 116206 (2021); https://doi.org/10.1016/j.trac.2021.116206
134. S.M.Azcarate, R.Ríos-Reina, J.M.Amigo, H.C.Goicoechea. *TrAC Trend. Anal. Chem.*, **143**, 116355, (2021); https://doi.org/10.1016/j.trac.2021.116355
135. D.Ballabio, R.Todeschini, V.Consonni. In: *Data Fusion Methodology and Applications*. (Elsevier, 2019). P.129; https://doi.org/10.1016/B978-0-444-63984-4.00005-3
136. O.Rodionova, A.Pomerantsev. *Anal. Chim. Acta*, **1265**, 341328 (2023); https://doi.org/10.1016/j.aca.2023.341328
137. G.G.Siano, H.C.Goicoechea. *Chemom. Intell. Lab. Syst*., **88**, 204 (2007), https://doi.org/10.1016/j.chemolab.2007.05.002
138. M.Ghaffari, N.Omidikia, C.Ruckebusch. *Anal. Chem*., **91**, 10943 (2019); https://doi.org/10.1021/acs.analchem.9b02890
139. K.Rajer-Kanduc, J.Zupan, N.Majcen. *Chemom. Intell. Lab. Syst*., **65**, 221 (2003); https://doi.org/10.1016/S0169-7439(02)00110-7
140. A.L.Pomerantsev, O.Ye. Rodionova. *Microchem. J.*, **217**, 114955 (2025); https://doi.org/10.1016/j.microc.2025.114955
141. F.Boughattas, D.Vilkova, E.Kondratenko, R.Karoui. *Food Chem.*, **312**, 126000, (2020); https://doi.org/10.1016/j.foodchem.2019.126000
142. O.Ye.Rodionova, P.Oliveri, C.Malegori, A.L.Pomerantsev. *Trends Food Sci. Technol.*, **147**, 104429 (2024); https://doi.org/10.1016/j.tifs.2024.104429
143. V.S.Popov, T.V.Shelenga, E.V.Blinova, V.I.Khoreva. *Plant Biotech. Breed.,* **7**, 31 (2024); https://doi.org/10.30901/2658-6266-2024-2-o1

144. A.Dispas, P.-Y.Sacre, E.Ziemons, P.Hubert. *J. Pharm. Biomed. Anal.*, **221**, 115071 (2022); https://doi.org/10.1016/j.jpba.2022.115071
145. N.B.Zorov. A.M.Popov, S.M.Zaytsev, T.A.Labutin. *Russ. Chem. Rev*., **84**, 1021 (2015); https://doi.org/10.1070/RCR4538
146. V.Deev, S.Solovieva, E.Andreev, V.Protoshchak, E.Karpushchenko, A.Sleptsov, I.Kartsova, E.Bessonova, A.Legin, D.Kirsanov. *J. Chromatogr. B: Anal. Technol. Biomed. Life Sci.*, **1155**, 122298 (2020); https://doi.org/10.1016/j.jchromb.2020.122298
147. E.Yuskina, M.Mosoyan, I.Jahatspanian, A.Vasilev, V.Makeev, A.Gaponova, V.Protoshchak, E.Karpushchenko, A.Sleptsov, D.Kirsanov. *Microchem. J.*, **218**, 115589 (2025); https://doi.org/10.1016/j.microc.2025.115589
148. R.van Vorstenbosch, F.-J.van Schooten, Z.Mujagic, A.Smolinska. *Anal. Chim. Acta*, **1383**, 344889 (2026); https://doi.org/10.1016/j.aca.2025.344889
149. N.I.Kurysheva, O.Y.Rodionova, A.L.Pomerantsev, G.A.Sharova, O.Golubnitschaja. *EPMA J.*, **14**, 527 (2023); https://doi.org/10.1007/s13167-023-00337-1
150. O.Ye.Rodionova, N.I.Kurysheva, G.A.Sharova, A.L.Pomerantsev. *Anal. Chim. Acta*, **250**, 340958 (2023); https://doi.org/10.1016/j.aca.2023.340958
151. C.H.Chen, T.C.Chen, X.Zhou, R.Kline-Schoder, P.Sorensen, R.G.Cooks, Z.Ouyang. *J. Am. Soc. Mass Spectrom.*, **26**, 240 (2014); https://doi.org/10.1007/s13361-014-1026-5
152. K.-M.Lei, D.Ha, Y.-Q.Song, R.M.Westervelt, R.Martins, P.-I.Mak, D.Ham. *Anal. Chem.*, **92**, 2112 (2020); https://doi.org/10.1021/acs.analchem.9b04633
153. V.Gubala, L.F.Harris, A.J.Ricco, M.X.Tan, D.E.Williams. *Anal. Chem*., **84**, 487 (2012); https://doi.org/10.1021/ac2030199
154. F.Feng, Y.Ren, Y.Qi, Z.Li, H.Zhang, Y.Yuan, R.Fu, J.Hu, M.You. *TrAC Trend. Anal. Chem.*, **199**, 118811 (2026); https://doi.org/10.1016/j.trac.2026.118811
155. A.E.Urusov, A.V.Zherdev, B.B.Dzantiev. *Biosensors*, **9**, 89 (2019); https://doi.org/10.3390/bios9030089
156. I.Y.Goryacheva, P.Lenain, S.De Saeger. *TrAC Trend. Anal. Chem.*, **46**, 30 (2013); https://doi.org/10.1016/j.trac.2013.01.013
157. E.Martynko, D.Kirsanov. *Biosensors*, **10**, 100 (2020); https://doi.org/10.3390/bios10080100
158. S.Aftab, B.Aslam, S.Khan, Z.Baloch. *Russ. Chem. Rev*., **94** (6), RCR5158 (2025); https://doi.org/10.59761/RCR5168
159. S.E.Patlasova, E.I.Korotkova, V.Saqib, A.N.Solomonenko, E.V.Dorozhko. *Microchem. J.*, **225**, 118157 (2026); https://doi.org/10.1016/j.microc.2026.118157
160. D.Gondocs, V.Dorfler. *Artif. Intell. Med*., **149**, 102769 (2024); https://doi.org/10.1016/j.artmed.2024.102769
161. C.Deng, X.Ji, C.Rainey, J.Zhang, W.Lu. *iScience*, **23**, 101656 (2020); https://doi.org/10.1016/j.isci.2020.101656
162. H.Das Gupta, V.S.Sheng. *arXiv*: 2212.05712 (2022); https://doi.org/10.48550/arXiv.2212.05712