Contents lists available at ScienceDirect

Analytica Chimica Acta

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

The inversion of multiresponse partial least squares models, a useful tool to improve analytical methods in the framework of analytical quality by design

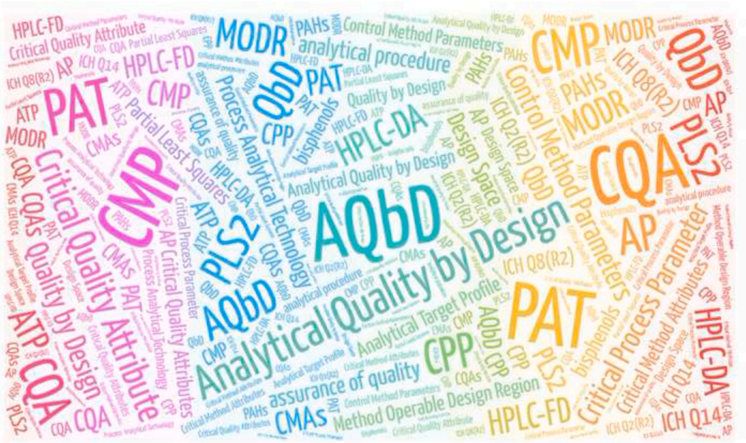
M.C. Ortiz [a,*], L.A. Sarabia [b], M.S. Sánchez [b]

[a] Dpt. Chemistry, Faculty of Sciences, Universidad de Burgos, Plaza Misael Bañuelos s/n, 09001, Burgos, Spain
[b] Dpt. Mathematics and Computation, Faculty of Sciences, Universidad de Burgos, Plaza Misael Bañuelos s/n, 09001, Burgos, Spain

HIGHLIGHTS

- Method Operable Design Region obtained with a Partial Least Squares model inversion.
- Finding, by PLS2, Control Method Parameters to fulfil Critical Quality Attributes.
- A computational algorithm to invert a PLS2 model to get Method Operable Design Region.
- D-optima criteria used to reduce experiments number in Analytical Quality by Design.
- AQbD useful tool in analytical procedures when liquid chromatography is carried out.

GRAPHICAL ABSTRACT


AQbD
Analytical Quality by Design
CMP
PAT
CQA
Critical Quality Attribute
Control Method Parameters
Critical Process Parameter
MODR
QbD
PLS2
HPLC-FD
HPLC-DA
ICH Q8(R2)
ICH Q14
Method Operable Design Region
assurance of quality
analytical procedure
Analytical Target Profile
Process Analytical Technology
Partial Least Squares
Design Space
CPP
CMAs
PAHs
bisphenols
ATP

ARTICLE INFO

Handling Editor: Prof. L. Buydens

Keywords:
Analytical quality by design
Partial least squares inversion
Liquid chromatography
Bisphenols
Polycyclic aromatic hydrocarbons
Method operable design region

ABSTRACT

Analytical Quality by Design (AQbD) is the adaptation of Quality by Design (QbD) when it is applied to the development of an analytical method. The main idea is to develop the analytical method in such a way that the desired quality of the Critical Quality Attributes (CQAs), stated via the analytical target profile (ATP), is maintained while allowing some variation in the Control Method Parameters (CMPs).

The paper presents a general procedure for selecting factor levels in the CMPs to achieve the desired responses, characterized by the CQAs, when liquid chromatographic methods are to be used for the simultaneous determination of several analytes. In such a case, the CMPs are usually the composition of the ternary mobile phase, its flow rate, column temperature, etc., while typical CQAs refer to the quality of the chromatograms in terms of the resolution between each pair of consecutive peaks, initial and final chromatographic time, etc. The analytical target profile in turn defines the desired characteristics for the CQAs, the reason for the whole approach.

[*] Corresponding author.
E-mail address: mcortiz@ubu.es (M.C. Ortiz).

https://doi.org/10.1016/j.aca.2023.341620
Received 24 March 2023; Received in revised form 11 July 2023; Accepted 12 July 2023
Available online 19 July 2023

The procedure consists of four steps. The first is to construct a D-optimal combined design (mixture-process design) to select the domain and levels of the CMPs. The second step is to fit a PLS2 model to predict the analytical responses expressed in the ATP (the good characteristics of the chromatogram) as a function of the CMPs. The third step is the inversion of the PLS2 model to obtain the conditions necessary to obtain the preset ATP in the corresponding CQAs. The inversion is performed computationally in order to estimate the Pareto front of these responses, namely, a set of experimental conditions to perform the chromatographic determination for which the desired critical quality attributes are met. The fourth final step is to obtain the Method Operable Design Region (MODR), that is, the region where the CMPs can vary while maintaining the quality of the CQAs.

The procedure has been applied to some cases involving different analytes, all of which are regulated by the European Union due to their toxicity to human health, namely five bisphenols and ten polycyclic aromatic hydrocarbons.

1. Introduction

The International Conference on Harmonization (ICH) in the ICH Q8 (R2) guideline, published in 2009, defines a systematic approach for the development of a product, including its manufacturing process [1], so that the resulting product meets preset quality requirements. The proposal starts with the definition of the objectives and uses scientific and risk management approaches to gain more knowledge about the product and the process and, ultimately, about process control.

The guide defines the conceptual framework of Process Analytical Technology (PAT) with the goal to enhance, understand, and control the manufacturing process, in coherence with the idea that it is not enough to verify the quality of the product, but that quality must be built into the process, that is, it must be designed. Hence, the term Quality by Design (QbD) to refer to it. Focusing more on building quality requires attention to multivariate relationships among material, manufacturing process, environmental variables, and their effects on quality. To establish these relationships and their operability, the guide defines three key concepts:

- Critical Quality Attribute (CQA): A physical, chemical, biological or microbiological property or characteristic of a product that should be within an appropriate limit, range, or distribution to ensure the desired product quality.
- Critical Process Parameter (CPP): A process parameter whose variability has an impact on a critical quality attribute and therefore should be monitored or controlled to ensure that the process produces the desired quality.
- Design Space: The multidimensional combination and interaction of input variables (e.g., material attributes) and process parameters that have demonstrated to provide assurance of quality.

It was quickly demonstrated that the QbD concept and methodology could be applied to the development of analytical methods that are also “processes” with quality objectives on the analytical measurement that have to be met [2,3]. This systematic way of developing an analytical method (AP) is known as Analytical Quality by Design (AQbD). In this context, there is still no consensus on the nomenclature, but it is an adaptation of QbD concepts to the field of chemical analysis. The most common terms, that can be consulted in the books in refs. [4–6], are the following:

- Critical Quality Attributes (CQAs), also known as Critical Method Attributes (CMAs) or Critical Analytical Attributes (CAAs). In AQbD, these are some “output” variables of an analytical method that should be maintained in a given range to ensure the desired method performance criteria.
- Control Method Parameters (CMPs). These are the “independent input” variables of the analytical method that influence the CQAs. CMPs are the analogous to CPP in QbD. They are related to the Analytical Target Profile (ATP), the statement that defines the purposes of the method, and that is then used to drive the selection, design, and development activities of the method.
- Method Operable Design Region (MODR). It is the region in the multidimensional space of the CMPs that contains their settings, i.e., the experimental conditions under which the analytical target profile defined by the CQAs is achieved. It is the AQbD version of the general Design Space (DS) in QbD.

In March 2022, the ICH Q14 [7] guideline on the development of analytical processes was released. Together with ICH Q2 (R2) [8], they specify the activities that should be performed during the lifespan of an analytical technique in relation to its development and validation when its intended use is to assess the quality of drug substances and medicinal products. However, while ICH Q14 describes the scientific principles for the development, management, and submission of analytical methods (for the minimal and enhanced approaches), ICH Q2 (R2) focuses on how to produce, submit, and maintain evidence that the analytical technique is fit for its purpose, which is the drug quality assurance. In both guidelines, particularly ICHQ14, the concepts of AQbD defined in the previous paragraphs are followed “to the letter”, although in ICH Q14 the CQA are referred to as “AP attributes” (based on ATP) and the CMP as “AP parameters with potential influence on AP attributes”. Despite the fact that both guidelines refer to the AP in the pharmaceutical field, they are a relevant sample of the consolidation of the AQbD methodology for the development of analytical methods in other fields (environment, food safety, health, etc.). Several papers that adapt the QbD paradigm to APs are reviewed by Orlandini et al. [9], which also includes an analysis of the two guidelines mentioned with their practical implementation. The review shows that AQbD is already established as a methodical approach to the development of an analytical method, following a well-defined framework to ensure that the AP remains fit for its intended use throughout its life cycle. Nevertheless, much remains to be done to ensure a clear understanding and efficient implementation of this new paradigm, AQbD, in practice, particularly regarding the specific lexicon. A contingency table is used in ref. [10] to analyse the degree of confusion in the terminology used in 37 papers, 16 on AQbD and 21 about QbD. The analysis shows that there is a need for better clarification and definition of the AQbD notation to achieve homogenization in the scientific community. This could be accomplished through the development of new regulatory documents.

In the present work, the terminology CMPs (control method parameters) and CQAs (critical quality attributes) related to the ATP (analytical target profile) would be followed, together with MODR (method operable design region) for the subspace where the CMPs can vary to assure the preset quality. In this context, the proposed procedure serves to construct the MODR with the aid of a Partial Least Squares model relating CMPs and CQAs, usually denoted as PLS2 because there is more than one response to be fitted. This PLS2 model should be inverted to get close to the intended ATP. The proposal is applied to the development of analytical methods using liquid chromatography, where the CMPs considered are, for example, the composition of the ternary mobile phase (e.g., water, methanol, acetonitrile), the flow rate of the mobile phase, or the column temperature. The Analytical Target Profile (ATP) specifications are related to the resolution between the identified peaks (that should be greater than a specific quantity) or the total chromatography time (that should be as short as possible). All these parameters would constitute the CQAs mentioned before.

A D-optimal design is used to efficiently explore the constrained experimental domain (input space) where the CMPs can vary. The results obtained after conducting the experiments, measured in terms of the CQAs, are used to build a PLS2 model that relates them. The inversion of the PLS2 is performed by means of an evolutionary algorithm that makes it possible to obtain the Pareto front of solutions that satisfy the desired multiple characteristics of the chromatograms, i.e., the CMPs for which the corresponding chromatograms are the closest to the required characteristics of the CQAs. The use of parallel coordinates plots to display CMPs and CQAs together on a single graph facilitates analysis of the Pareto front and selection of specific conditions.

Finally, the development of the chromatographic method with the Analytical Quality by Design (AQbD) methodology requires the construction of the Method Operable Design Region (MODR) to ensure analytical robustness, as the MODR is precisely the region where the CMPs can vary while maintaining the desired quality on the CQAs. This region is the 'core' of AQbD. After inversion of the PLS2 model, those Pareto optimal solutions that are expected to satisfy the CQAs constitute an initial estimate of this region.

The paper shows the application of the procedure to the development of two APs. The first, case I, focuses on the inversion of a PLS2 model looking for the analytical conditions to achieve a preset analytical target profile. The second, case II, extends the inversion of the PLS2 model to obtain the MODR.

The paper is organized as follows. After the present introduction, Section 2 summarizes the state of the art in AQbD, with details about the customary development stages of an AP. Software and instrumentation are described in Section 3, while Section 4 is devoted to the theoretical aspects of the proposed procedure. It includes the selection of the experimental design followed, which provides the training set for fitting the PLS2 model, its computational inversion, which gives the CMPs that guarantee the intended CQAs, and the method for obtaining the MODR. Results and discussion are presented in Section 5.

2. State of the art

The growing interest in the application of AQbD (or of QbD) has been evident in recent years. A search in the SCOPUS database, searching within the article title, abstract, or keywords with the query string (AQbD OR "analytical quality by design" OR QbD OR "quality by design") returned 5,242 references dating back to 1983. Among them, 2,231 were published between 2018 and 2022. Fig. 1 depicts the evolution of publications over these five years in the fields of Chemistry (CHEM) and Pharmacology, Toxicology and Pharmacy (PHARM), separately for each single field (ONLY CHEM or ONLY PHARM) and the number of papers in journals common to both fields. In the last five years, the number of references only in Chemistry (ONLY CHEM in Fig. 1) increased from 46 to 98 (that is, a 213% increase) whereas the total number of publications increased by 156% (from 342 to 534). In the pharmaceutical field (ONLY PHARM), the increase was a 149%. This means that, even though AQbD or QbD is used less in CHEM than in PHARM, the growth rate is greater in the former than in the latter. Several of the references refer to the implementation of on-line APs, but without a doubt the published guidelines are at the origin of these figures.

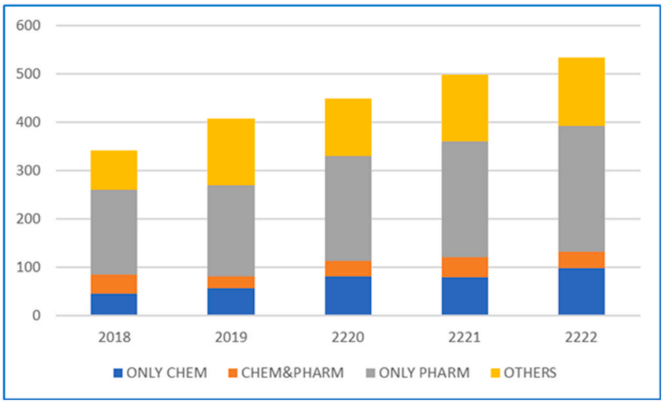


Fig. 1. Number of publications in the last five years (2018–2022) about QbD or AQbD in Chemistry (CHEM) and Pharmacology, Toxicology and Pharmaceutics (PHARM). Source: SCOPUS.

In general terms, the development stages of an AP are: 1) To define the ATP and the CMPs related to it. 2) To establish the CQAs and their individual ranges. 3) To build a function, $\mathbf{y} = f(\mathbf{x})$, that relates the values of the CMPs (predictor variables, $\mathbf{x}$) to the corresponding CQA values (response variables in $\mathbf{y}$). 4) To build the MODR using f. 5) To validate the method. 6) To monitor and manage the lifecycle, identifying trends and using the MODR for continuous improvement of the AP. From a methodological point of view, steps 3) and 4) are the core of the procedure.

Note that the function f in $\mathbf{y} = f(\mathbf{x})$ is a vector function defined in a multivariate domain, in other words, it applies a multivariate vector $\mathbf{x} = (x_1, x_2, \ldots, x_p)$ containing the values of the p CMPs into a multiresponse vector $\mathbf{y} = (y_1, y_2, \ldots, y_q)$ with the corresponding values of the q CQAs. Apart from exceptional cases, the function f must be estimated from experimental data. Therefore, n vectors of CMPs should be selected and arranged as the rows of an $\mathbf{X}$ matrix. In each of these n experimental conditions, the analytical method is applied to obtain the vector of the corresponding values of the q CQAs. These n response vectors are the rows of a matrix $\mathbf{Y}$ that correspond one to one with those of $\mathbf{X}$. In this way, the problem of obtaining f is a regression problem with training set $\{\mathbf{X}, \mathbf{Y}\}$.

This problem is divided into two parts. First, select the vectors that will form the $\mathbf{X}$ matrix, that is, design the experiment that, in practice, allows obtaining the $\mathbf{Y}$ matrix with the CQAs. Second, decide on the appropriate functional form of f.

Once the function f is built, the MODR must be obtained. In order to do it, it is necessary to define the range of the CQAs that is suitable to ensure the desired performance criteria of the method. This is equivalent to defining a region Q in the multivariate space of the CQAs. For example, if each CQA must lie between two values, that is, $y_i \in [a_i,b_i], i = 1,\ldots,q$, then Q would be the q-dimensional parallelepiped defined by the Cartesian product $\prod_{i=1}^{q}[a_i, b_i]$. Then, the "analytical knowledge space" should be determined, which is the set $f^{-1}(Q)$, in which the CMPs would theoretically produce a product that satisfies all CQAs. In practice, the uncertainty of the f model and of the variables used to compute it must also be considered. Therefore, the MODR is in fact a subset of $f^{-1}(Q)$ to ensure a high probability of meeting the CQAs. There are several methodologies to obtain a MODR in terms of the probability of meeting the specifications, ranging from Monte Carlo simulation (the most commonly used in practice) to bootstrap or Bayesian procedures [11]. Also, the MODR can be validated experimentally as described in option 2 of Annex B in Validation strategies for MODRs [7], with the advantage that there is no need to know (or impose) any probability distributions.

The multivariate-multiresponse nature of f (p predictors/q responses) is an aspect of AQbD that has not been much studied. However, it conditions the possibilities of its inversion and thus the determination of the MODR.

In the literature on the subject, attention has been paid to the multivariate nature of f, and therefore to the need to simultaneously consider the p CMPs in the matrix $\mathbf{X}$, so that the f model can capture the possible interactions among them that would influence the CQAs. The guidelines [1,7] insist on this point and recommend to abandon the "one factor at a time" (OFAT) method to develop an AP. In spite of this, its use is still maintained, as shown by a recent review of 36 papers (published from 2016 to 2021) dealing with supercritical fluid chromatography for pharmaceutical quality control [12], 17 of which use OFAT in AP development. This can lead to the blind selection of sub-optimal or even non-robust CMPs.

The design of experiments (DoE) methodology allows to adequately explore the p-dimensional space of the CMPs defined from the ATP [13, 14]. In particular, the usefulness and versatility of Response Surface Methodology (RSM [15]) designs that are generally used to build the $\mathbf{X}$ matrix. However, as the number of CMPs increases, so does the size n of the design, that is, the number of experiments to be carried out (rows of $\mathbf{X}$). For example, with $p = 7$ CMPs, the standard central composite design requires at least $n = 79$ experiments. In such situations, it is useful to resort to D-optimal designs [16] that significantly reduce n while maintaining sufficient quality in the estimated response, allowing further the introduction of experimental constraints, if any. In the previous example with $p = 7$ CMPs, instead of the 79 experiments, a D-optimal design with $n = 44$ experiments would suffice.

Another types of designs, especially interesting in HPLC, are those for mixtures, so as to handle the composition of ternary mobile phases, which, in turn, have to be combined with continuous CMP, such as the temperature or flow rate of the mobile phase [17]. In this case, the reduction of experimental effort with a D-optimal design would be remarkable. Nevertheless, there are only a few applications of this experimental procedure in AQbD [18–20].

In general, the methodology adopted in AQbD consists of least squares fittings of a polynomial (usually of second degree) using each individual CQA as a response. At the end, there are q functions $f_i(x_1, x_2, \ldots x_p)$, each with its own admissible variability range $[a_i, b_i]$ in the space of the CQAs, and, thus, with its own corresponding region $f_i^{-1}([a_i, b_i])$ in the space of the CMPs. Their intersection $\bigcap_{i=1}^{q} f_i^{-1}([a_i, b_i])$ would constitute the "analytical knowledge space". This way of working ignores the correlation among the CQAs, which is generally intrinsic to the problem, for example, the resolution between different chromatographic peaks and the total time of the chromatography are correlated CQAs. In addition, the collinearity between the CQAs must also be considered, meaning that the dimension of the subspace in which they vary may be less than q. If the preimage of each y_i under f_i is computed separately, it may happen that the inversion is performed for a vector $\mathbf{y} = (y_1, y_2, \ldots y_q)$ that in fact does not define any feasible characteristics.

There are different recent reviews about the selection of the type of function, f, as well as the procedure for obtaining the MODR. Ref. [21] reviewed 31 papers that develop AP according to AQbD and published between 2012 and 2018. All of them use a polynomial model and 24 of them determine the MODR using Monte Carlo simulation. Similarly, in ref. [22], the 17 articles reviewed use a polynomial model with a response surface design and, among them, 5 determine the MODR with Monte Carlo simulation, while another 2 do so with a Bayesian procedure. In Orlandini et al. [9], 21 articles from 2015 to 2022 are reviewed, only when Chiral Capillarity Electrophoresis methods are developed using AQbD principles. Polynomial models are used in all of them and those 12 among the 21 in which the MODR is determined, this was done with Monte Carlo simulation.

A review of 36 articles published in 2022 in which chromatographic methods are developed is shown in Table S1 in the supplementary material and confirms the trend already shown: quadratic models and response surface designs are used, except in 8 of them that use factorial designs, in which case the polynomial model does not contain pure quadratic terms. Of these articles, 18 do not determine the MODR and the rest use Monte Carlo simulation. On the other hand, a total of 162 analytes are treated in the reviewed analytical methods, although this number varies between 1 and 18 in the individual analyses (the most frequent being 1, 2 or 3 analytes in 8, 8 and 7 papers, respectively). Regarding the number of CMPs, it varies from 2 to 5, with 3 CMPs being the most frequent (22 articles). Regarding the CQAs, there are up to 13 different CQAs. When counting the frequencies, only those papers with 2–5 CQAs have a frequency greater than 1, with 3 CQAs being the most frequent case.

Overall, the highest frequency (10 papers) corresponds to 3 CMPs and 3 CQAs. The crosstabulations of CMP by CQA, CMP by analyte, and CQA by analyte are shown in Tables S2, S3 and S4 in the supplementary material, respectively. In all three cases, a χ^2 hypothesis test (run to determine whether or not to reject the idea that the row and column classifications are independent) gives P-values of $2.5\ 10^{-3}$, $2.3\ 10^{-3}$ and $3.0\ 10^{-4}$ respectively, much less than 0.05. Consequently, the hypothesis that the rows and columns are independent is rejected at the 95% confidence level.

This analysis supports the conclusion that the tendency to apply AQbD to the development of AP is to use RSM (Response Surface Methodology) with a polynomial model (linear or quadratic) and the determination of the MODR with certain "guarantees" is done with Monte Carlo methods. Therefore, it is of interest to extend the research on the use of PLS2 models, both to build the model and to determine the MODR. The approach in the present paper is based on multiresponse and multivariate PLS2 models to find the Pareto front of optimal conditions for the AP and to determine its MODR [23–25].

PLS2 is a regression method that handles the internal correlations among CMPs, those among CQAs, and also the correlation between CMPs and CQAs. Despite the widespread use of PLS2 as a multivariate regression technique in PAT, particularly in the pharmaceutical industry [26], its use in AQbD has been very limited [27–29]. In addition, in these applications, the inversion is performed separately for each CQA, analogous to the case of having used q regression models, one for each CQA. In other words, the PLS2 property of considering the projection of the predictor and response variables in a single subspace (the latent space) that captures the structure of each set of variables is neglected.

Fig. 2 shows the parallelism of an industrial process (QbD) and an analytical process in AQbD for the particular case of developing a method for liquid chromatography: the process variables are the method variables (Control Method Parameters, CMPs) and the quality characteristics are the Critical Quality Attributes (CQA). In the case of an AP as liquid chromatography, some specific characteristics are required for the chromatogram, for example, specificity, that is, good resolution between contiguous peaks, an adequate initial time in the chromatogram to avoid the dead volume, and also a short final time to save time in the analysis. After inversion of the fitted PLS2 model (with the available data), the appropriate CMPs should be selected to achieve the desired CQAs. An additional advantage in the case of a laboratory is that the levels of the factors (CMPs) and the domain of these parameters can be designed.

The inversion of the PLS2 model and the visualization of the "analytical knowledge space" and of the subsequent MODR raise specific methodological difficulties that are addressed in the two case studies detailed in the following sections.

3. Software and experimental

3.1. Software

OpenLab CDS ChemStation software for an Agilent 1260 Infinity HPLC chromatograph (Santa Clara, CA, USA) was used for data acquisition when measurements were recorded by means of HPLC-DAD and HPLC-FLD.

The PLS2 models were fitted with the PLS_Toolbox [30]. The inversion of the PLS2 model and the Pareto optimal front were calculated with in-house programs written in MATLAB [31]. The experimental design was selected using NEMRODW [32].

3.2. Instrumental details

Only some instrumental details are summarized in the present section; further details about sample preparation in cases I and II can be

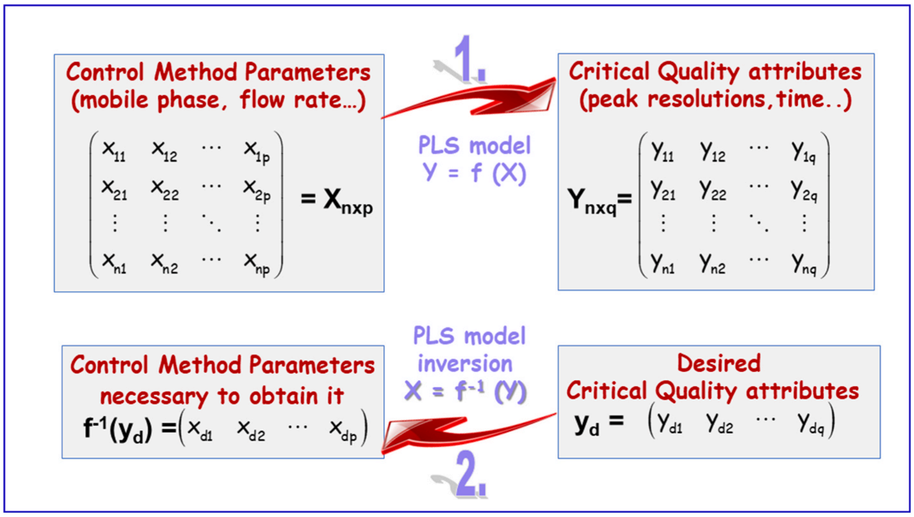

Control Method Parameters
(mobile phase, flow rate...)
X_{nxp}
1.
PLS model
Y = f (X)
Critical Quality attributes
(peak resolutions, time..)
Y_{nxq}
PLS model
inversion
X = f⁻¹ (Y)
Control Method Parameters
necessary to obtain it
$f^{-1}(y_d) = (x_{d1}\ x_{d2}\ \cdots\ x_{dp})$
2.
Desired
Critical Quality attributes
$y_d = (y_{d1}\ y_{d2}\ \cdots\ y_{dq})$

Fig. 2. Scheme of the fit of a PLS2 model and its inversion in the context of Analytical Quality by Design and Process Analytical Technology.

consulted in ref [23,24], respectively. In both cases, the chromatograms were recorded with an Agilent 1260 Infinity HPLC chromatograph (Santa Clara, CA, USA) consisting of a quaternary pump (G1311C), a sampler (G1329B), and a thermostatic column compartment (G1316 A). A Kinetex EVO-C18 column (150 mm × 4.6 mm, 5 μm, Phenomenex, Torrance, CA, USA) was used for the separation. Deionized water (solvent A), methanol (solvent B), and acetonitrile (solvent C) were used as mobile phases. The conditions for chromatographic analyses were programmed in isocratic mode. In all analyses, the injection volume was 10 μL.

3.2.1. *Case I. Inversion of a PLS2 model searching for a preset CQA*

The developed AP is for the determination of five bisphenols, BPA, BPS, BPF, BPZ and BPAF, by means of HPLC with a diode array detector (G7117C) programmed to measure the absorbance at a fixed wavelength of 225 nm. The CMPs considered were different percentages of a mixture of water/methanol/acetonitrile (X_1: X_2: X_3, v/v) and different flow rates (X_4, mL min^{-1}) of the mobile phase, depending on the conditions of the experimental design followed, which is explained in Section 4.2. For all analyses, the temperature of the column compartment was 20 °C.

3.2.2. *Case II. MODR obtained with a PLS model inversion*

In this case, the AP is for the determination of ten polycyclic aromatic hydrocarbons (PAHs), naphthalene (NAP), phenanthrene (PHE), anthracene (ANT), fluoranthene (FLN), pyrene (PYR), chrysene (CHR), Benzo [*a*]anthracene (BaA), perylene (PER), benzo [*b*]fluoranthene (BbF), and benzo [*a*]pyrene (BaP). The determinations were carried out using the described chromatograph with a fluorescence detector (G1321B) programmed to measure the fluorescence intensity at a fixed excitation wavelength of 274 nm. However, three emission wavelengths were selected to better identify the ten PAHs in the chromatograms: 345 nm was used for NAP and PHE, 405 nm for ANT, PYR, CHR, BaA, and BaP and 470 nm for FLN, PER, and BbF. The CMPs defined for this analysis were the percentages of a mixture of water/methanol/acetonitrile (X_1:X_2:X_3, v/v), the flow rate of the mobile phase (X_4, mL min^{-1}), and the column temperature (X_5, °C). As in the previous case, they were set to the conditions explained in Section 4.3 according to the different experiments performed.

4. PLS inversion, experimental design and MODR

4.1. *Steps of the general procedure*

With the preset ATP, the factors (CMPs) and the desired CQAs are stablished. Then, the procedure for obtaining the corresponding settings of the CMPs and the MODR has the following steps:

1. Selection of an appropriate experimental design (few experiments whenever possible) inside the experimental domain defined by the variation of the CMPs. Run the experiments with the particular settings of the CMPs in the design, that are part of the **X** matrix. Qualify the resulting chromatograms in terms of the responses of interest (CQAs) to form matrix **Y.** A training set {**X**, **Y**} is thus available to fit a PLS2 model.
2. Selection of some working experimental conditions (some settings for the CMPs). After the fitting and validation of the PLS2 model, its inversion towards approaching the multiresponse ATP gives the Pareto front for the desired CQAs. The exploration of this front, that represents the extension of the trade-off among responses, leads to the final selection of a set of CMPs (the experimental conditions) to perform the chromatographic determination.
3. Search of the Method Operable Design Region (MODR). Around the selected experimental conditions, the MODR is the subset of the experimental domain where the Control Method Parameters can vary without modifying the established values of the Critical Quality Attributes.
4. Validation of the analytical method with the selected CMPs. This is the experimental validation of the found conditions as well as the figures of merit (accuracy, precision, decision limit and capability of detection) of the proposed analytical method for the determination of the analytes.

The approach just described has already been applied to some analytical methods, such as the determination of five bisphenols by means of liquid chromatography (HPLC-DAD) [23], or eight triazines by means of SPE-HPLC-DAD [25]. Besides, the MODR is evaluated in ref. [24] for a procedure to determine ten PAHs measured by means of HPLC-FLD.

The specifics of the application of the general procedure outlined in steps 1 to 4 are explained in what follows for the development of two different analytical methods: the selection of settings via inversion in case I for the determination of five bisphenols by HPLC-DAD [23], and the additional computation of the MODR in case II for the determination of ten polycyclic aromatic hydrocarbons by liquid chromatography with fluorescence detection [24].

4.2. Case I. How to choose the settings of the CMPs

4.2.1. D-optimal design to select a proper training set

As it has been detailed in Section 2, a D-optimal design has to be used to explore broad domains of variability and, in particular, when continuous variables and mixtures should be handled together in combined designs. Otherwise, the experimental effort needed to complete the full designs becomes unfeasible in practice.

In the present case I, with liquid chromatography, some specific characteristics of the chromatogram are required (specificity peaks, good resolution between contiguous peaks, an adequate initial time and a short final time to save time in the analysis), characteristics that depend on the specific settings of the CMPs. To model this dependency via the PLS2 model, a training set is required. Therefore, the first step is to define the experimental domain of the CMPs, which in the present case are constraint ternary mixtures of water, methanol, and acetonitrile, and the mobile phase flow rate (a continuous process variable). The constraints imposed on the mixtures were that the percentage of water (X_1) should be between 20% and 50%, while methanol (X_2) and acetonitrile (X_3) should not exceed 70%. The constrained domain is shown in yellow in Fig. 3 in the form of the usual ternary diagram. Finally, the flow rate (X_4) is set at three levels: 0.6, 0.8, and 1 mL min^{-1}. The distribution of the experimental points in a standard mixture design has 19 experiments, which have to be repeated for each level of the continuous factor, meaning a total of 57 experiments that should be carried out. Applying the D-optimal criterion [16], this number of experiments is reduced because it suffices to take nine out of the 19 experiments in each level of the flow rate. This resulting design still maintains good reliability properties as measured by the maximum of the variance function in the experimental domain which was 0.8 (less than 1) [15]. Finally, some replicates and other ternary mixtures were added to improve the results, with a total of 43 experiments.

The next step is to propose a model for the mixtures-process design, one per response Y_i. The assumed model for each individual Y_i in the multiplicative mixture process design is quadratic in the continuous variable (flow rate, X_4), according to Eq. (1), and each coefficient γ_i of this model has a quadratic dependence on the mixture composition (X_1, X_2, X_3). For example, γ_4 in Eq. (1) is expressed as in Eq. (2).

$$Y = \gamma_0 + \gamma_4 X_4 + \gamma_{44} X_4^2 \tag{1}$$

$$\gamma_4 = \beta_{41}X_1 + \beta_{42}X_2 + \beta_{43}X_3 + \beta_{412}X_1X_2 + \beta_{413}X_1X_3 + \beta_{423}X_2X_3 \tag{2}$$

By substituting all the terms in Eq. (1), the final model has 18 coefficients. These models are only used to obtain the experiments according to the methodology of experimental design [15].

4.2.2. Fitting a PLS2 model

Once the training set is available, after carrying out the experiments planned in the previous section, the proposed model relating **X** and **Y** and fitted with PLS2 is the one in Eq. (3). Along with the four method variables (the CMPs), it includes several interactions among some of them, together with some quadratic effects.

$$\begin{aligned}\mathbf{Y} &= \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4 + \beta_{12}X_1X_2 + \beta_{13}X_1X_3 + \beta_{23}X_2X_3 + \\ &\beta_{41}X_4X_1 + \beta_{42}X_4X_2 + \beta_{43}X_4X_3 + \beta_{412}X_4X_1X_2 + \beta_{413}X_4X_1X_3 + \beta_{423}X_4X_2X_3 + \\ &\beta_{441}X_4^2X_1 + \beta_{442}X_4^2X_2 + \beta_{443}X_4^2X_3\end{aligned} \tag{3}$$

These interaction terms in the model (the different cross-terms) reflect the fact that the response does not depend exclusively on variables X_1, X_2, X_3 and X_4 (main factors, those that can be modified inside the experimental domain).

A latent variable regression model with more than one response (Partial Least Squares, PLS2) is fitted to predict the responses (**Y**) as a function of the predictors (**X**). PLS2 works by computing a new set of *c* orthogonal variables such that Eq. (4) and Eq. (5) hold.

$$\mathbf{X} = \mathbf{T}_c\mathbf{P}_c^T + \mathbf{R}_\mathbf{X} = \widehat{\mathbf{X}} + \mathbf{R}_\mathbf{X} \tag{4}$$

$$\mathbf{Y} = \mathbf{T}_c\mathbf{Q}_c^T + \mathbf{R}_\mathrm{Y} = \mathbf{XM} + \mathbf{R}_\mathrm{Y} \tag{5}$$

X (input variables) is representative of the control method parameters, CMPs, and **Y** contains the responses (critical quality attributes, CQAs) of the analytical method. The desired CQAs are known and can be defined as a vector $\mathbf{y}_\mathrm{des}$, not necessarily in the training set. Whether $\mathbf{y}_\mathrm{des}$ can be reached (i.e., whether the analysis with these characteristics is feasible) depends on the existence of values for the CMPs that would give this analytical target profile.

As stated in Eq. (3), **X** reflects the dependence on more than just the four CMPs. The size of the matrix **X** to fit the PLS2 model is 43 × 16, where 43 is the number of experiments (chromatograms in this case) and 16 is the number of predictor variables. The corresponding matrix **Y** is 43 × 6 because there were six CQAs in this case, namely the four resolutions between contiguous peaks, and the initial and final time of the chromatogram. The PLS2 model was fitted to the 6 CQAs simultaneously.

A cross-validation procedure was used to select 7 latent variables. In addition, a permutation test was performed to validate this selection. The *p*-value of this test was less than 0.05, which means that the model is significant at the 95% confidence level. The model explains more than 99% of the variance in the predictors (CMPs and their relations) and 97% of the variance in the responses (CQAs). The variance explained for each response and in both fitting and prediction are similar, which indicates a good predictive model.

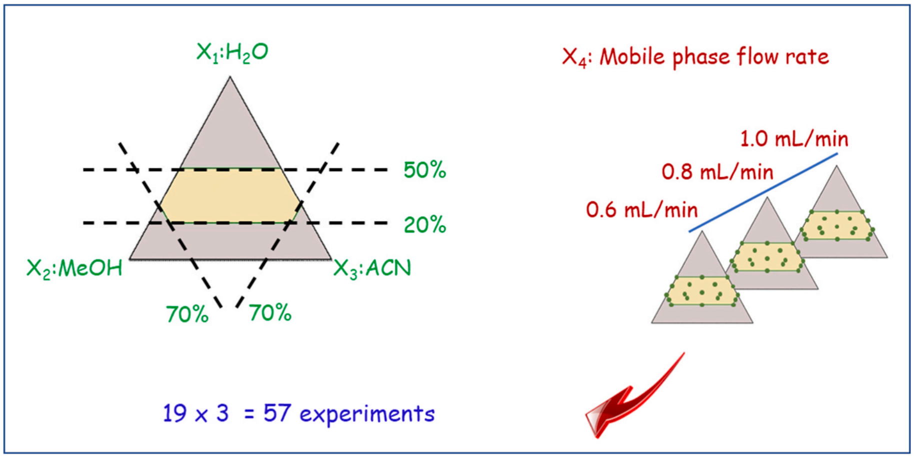


Fig. 3. Experimental domain in case I for a mixture-process design (combining ternary mixtures with constraints with process variables, in this case, the flow rate).

4.2.3. *PLS model inversion*

Finding the values of the input variables, $\mathbf{x}_{\text{des}}$, with which the preset quality $\mathbf{y}_{\text{des}}$ is obtained requires the inversion of the fitted PLS2 model. In general, the study of the viability of such inversion, with the necessary constraints, is known as Latent Variable Model Inversion (LVMI). In the reviewed literature, there are two alternatives to approach the LVMI. One of them is related to the inversion of the matrices in the decomposition in Eqs. (4) and (5). A summary of this approach, an algebraic or *direct* inversion, can be found in refs. [33,34]. The other alternative is a computational inversion by using an evolutionary algorithm to search for solutions in the input space while maintaining the constraints of the model [35].

The algebraic approach cannot be applied for the inversion of the fitted PLS2 model because the method variables are implicit in the PLS2 model [36]. In other words, the PLS2 model to be inverted includes not only the CMPs (i.e. the input variables related to the analytical method that can be modified in the laboratory), but also some of their cross-products. In general, it can include any other possible transformations of the CMPs that are necessary to model non-linearities. Consequently, the direct inversion of the matrices is not possible, as shown in ref. [36], but the computational approach can be applied to obtain the $\mathbf{x}_{\text{des}}$ solution, as it works only with the CMPs.

Another consideration is that the responses are conflicting, that is, the values of the CMPs that optimize one CQA do not necessarily optimize another. Therefore, a unique solution to the inversion is unlikely. In this multiobjective scenario with several conflicting responses, the Pareto optimal front represents the extent of the conflict. In other words, only Pareto front solutions are of interest, the others are worse in at least one of the quality characteristics.

The first step in applying this procedure is to define the desired target chromatogram characteristics. In this case the requirements are that the resolution between contiguous peak is at least 1.1, with an initial time greater than 2 min and a final time no longer than 4 min. Eq. (6) shows the coordinates of the desired vector, already taking into account the responses that had to be transformed to obtain the best PLS fit.

$$\mathbf{y}_{\text{des}} = \left(1.1, 1.1, 1.1, 2^{1.1}, 2.0, \log_{10}(4.0)\right)^T \qquad (6)$$

The problem is then to find, if any, the values of the CMPs, $\mathbf{x}$= (x_1, x_2, x_3, x_4), that provide these characteristics. To achieve this, the fitness function for each “candidate” $\mathbf{u}$ (to be minimized when looking for the Pareto front) is the absolute value of the differences between the response predicted by the PLS model $\widehat{\mathbf{y}}$ and the target value for each desired value, as expressed in Eq. (7).

$$\text{fitness}(\mathbf{u}) = \left(|\widehat{y}_1 - 1.1|, |\widehat{y}_2 - 1.1|, |\widehat{y}_3 - 1.1|, \left|\widehat{y}_4 - 2^{1.1}\right|, |\widehat{y}_5 - 2.0|, |\widehat{y}_6 - \log_{10}(4.0)|\right)^{\mathrm{T}} \qquad (7)$$

Fig. 4 shows a flowchart of the genetic algorithm used to calculate the Pareto front. In a pseudo-code it consists of the following steps:

1 Initial population. It consists of *N* different values of the composition of the mobile phase (x_1, x_2, x_3) and its flow rate (x_4), that is, four-dimensional vectors satisfying the constraints of being in the domain and in the projection space as defined by the limits imposed (here 95% confidence) on the *Q* and T^2 statistics [35]. These *N* values are chosen with a uniform distribution between the maximum and minimum of each CMP in the training set, considering also the constraints imposed in the composition of the mobile phase (domain shown in Fig. 3).
2 Evaluation of the population. Each member of the population is ‘fed’ into the PLS2 model of 16 predictor variables already built to obtain $\widehat{\mathbf{y}}$ (the 6 CQAs), which is then compared with $\mathbf{y}_{\text{des}}$ to compute its fitness function.
3 Apply selection, crossover, and mutation genetic operators to create a new population of valid experimental conditions (i.e. meeting the constraints on CMPs), which are also evaluated in terms of the fitness function.
4 Merge the old and newly generated populations.
5 Arrange the members of the expanded population according to the Pareto order for multidimensional vectors.

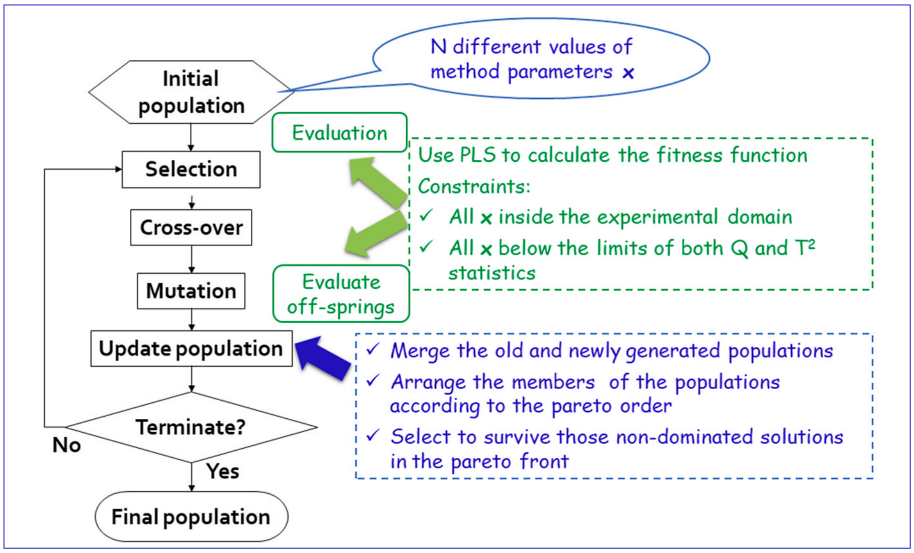

Initial population
N different values of method parameters x
Evaluation
Use PLS to calculate the fitness function
Constraints:
✓ All x inside the experimental domain
✓ All x below the limits of both Q and T²
statistics
Selection
Cross-over
Mutation
Evaluate off-springs
Update population
✓ Merge the old and newly generated populations
✓ Arrange the members of the populations according to the pareto order
✓ Select to survive those non-dominated solutions in the pareto front
Terminate?
No
Yes
Final population

Fig. 4. Flowchart of the genetic algorithm used.

6 Select those non-dominated solutions [37] in the Pareto front to survive for the next generation, and add the most dispersal according to the crowding distance [38].
7 Repeat steps 1) to 6) for a given number of generations.

The final population contains the settings of the CMPs that represent a compromise among all responses in the form of Pareto-optimal solutions, that is, those that cannot be improved in one response without worsening another.

Genetic or evolutionary algorithms are known to be computationally intensive, although the computational time in the cases explained here (2–3 min per run with populations of about a hundred elements) is negligible compared to the time required to perform the analytical experiments. On the other hand, like any other numerical algorithm, it is affected by the "course of dimensionality" as the number of responses increases. This is so because the population evolves in the experimental domain (or the knowledge space), but checks the non-dominance and convergence in the space of the objectives.

Fig. 5 shows some solutions of the estimated Pareto front in the form of a parallel coordinates plot. The method variables (CMPs) are in the first four coordinates and their predicted quality characteristics are in the last six. Since each variable is on a different scale, the values have been brought into a common range, with the original minimum and maximum values of each coordinate written at the bottom and top of the corresponding vertical line. In the graph, each broken line represents a single vector, whose coordinates contain together the values of the four CMPs (percentage of water, methanol and acetonitrile, and mobile phase flow rate) and of the six CQAs (pair-wise resolution between the first four peaks, and initial and final time of the chromatogram). The conflicting behavior already mentioned can also be seen in the graph: the first three resolutions R_{12}, R_{23}, R_{34} increase or decrease simultaneously although in different proportions. However, it is clear that their behavior is opposite to that of the resolution between peaks four and five (R_{45}). Note that while all the resolutions are greater than 1.1, their conflict with time is reflected, for example, in the cyan solution with dashed line: to improve all the resolutions approximately equal, the desired conditions on time are violated, exceeding the 4 min for the final time. In any case, there are some other adequate solutions, such as the one with the continuous blue line, which is the one chosen.

The filled circles in Fig. 5 for this selected solution mark the expected values of the CQAs: All resolutions are greater than 1.1, the initial time is greater than 2 min and the final time is less than 4 min (3.97 min), thus fulfilling all the conditions demanded for the CQAs. The corresponding $\mathbf{x}_{\text{des}}$ for obtaining $\mathbf{y}_{\text{des}}$ were: 22% water, 58% methanol, 20% acetonitrile and 0.66 mL min^{-1} for flow rate. Fig. 6 shows four chromatograms, the one on the lower left surrounded by a circle is the one obtained with the selected experimental conditions. Fig. 6 also shows some other chromatograms obtained in the process. Comparing the chromatographic time obtained with the proposed procedure and those found in other experimental conditions, the one on the top left was made with a binary mobile phase composition of water and methanol and took a long time (30 min). Another binary mixture, water and acetonitrile, gives the chromatogram shown at the top right, which takes less time (4 min) but a severe overlapping of the first three peaks is observed. The third chromatogram, made with a ternary mobile phase and shown at the bottom right, takes a short final time (3 min) but the first two peaks are completely overlapped. Following the proposed procedure, the time has been reduced without any overlap in any of the five chromatographic peaks. With the selected conditions, the validated procedure is also an advantage over similar work.

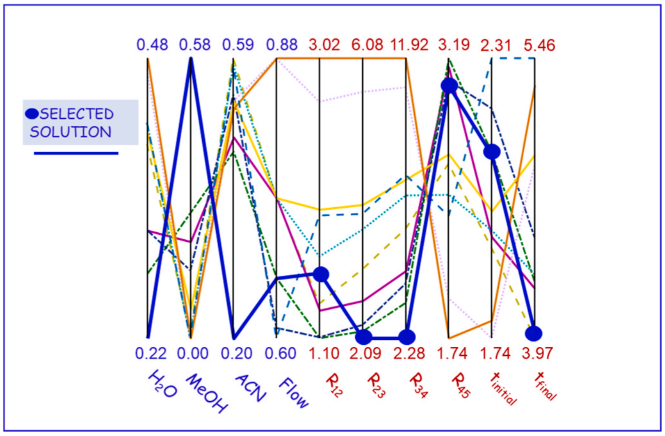


Fig. 5. Reduced Pareto front in the form of a parallel coordinates plot, case I (ref. [23]). R is peak resolution, t is time.

4.2.4. Experimental assessment of the CQAs

Finally, for experimental verification of the CMPs obtained by inversion (22% water, 58% methanol, 20% acetonitrile, and 0.66 mL min^{-1} for the mobile phase flow rate), 10 determinations were performed on a mixture of 4 mg L^{-1} of each bisphenol, and 95% confidence intervals were computed on the mean of the experimental values obtained in each of the six CQAs. Comparing them with the confidence intervals on the values predicted with the fitted PLS2 model, the actual chromatographic characteristics are included in the confidence intervals computed with the PLS2 model (the intervals and other details can be seen in ref. [23]).

4.3. Case II. How to build the method operable design region (MODR)

In this second case-study, the MODR will be obtained for the determination of 10 polycyclic aromatic hydrocarbons measured by liquid chromatography with fluorescence detection. The wide diffusion of polycyclic aromatic hydrocarbons (PAHs) in the environment and the toxicity of many of them explain the constant analytical interest, and thus the development of a multi-residue procedure for their determination.

The details about the selection of the desired CQAs to know how to handle the CMPs in an analytical method have been introduced with the previous case I. In the present case, the focus is on how to obtain the MODR from the inversion of a PLS2 model. As already mentioned, the MODR is a region where the CMPs can vary while maintaining the CQAs. That is, the region where the control method parameters allow us to ensure analytical robustness [24], and as such it is the 'core' of the AQbD. Its computation is essential to guarantee the required specifications in the quality characteristics of an analytical method.

4.3.1. D-optimal design to cover the domain of the CMPs

There are five CMPs for the analytical method. The first three CMPs specify the proportion of water (X_1), methanol (X_2), and acetonitrile (X_3) in the composition of the mobile phase. The composition of water in the mixture should be less than 40% with no restriction on the composition of methanol and acetonitrile. As in the previous case, the corresponding proportions are selected according to a mixture design in a restricted simplex, depicted in yellow in Fig. 7.

The flow rate of the mobile phase, factor X_4, and the column temperature, factor X_5, are continuous factors that vary from 0.5 to 1.5 mL min^{-1}, and from 20 to 44 °C, respectively. When the levels of these two continuous factors are added, the resulting experimental design is again a mixture-process design. The assumed model for each individual response in the multiplicative mixture-process design is linear in the continuous variables (flow rate and column temperature), with a quadratic dependence on the mixture composition (X_1, X_2, X_3), as in the case I already seen. The full experimental design has 405 experiments because there are 45 experiments in the mixture domain and 9 levels in the process variables (three levels per factor as illustrated in Fig. 7).

A D-optimal design [32] with 42 experiments was selected with the

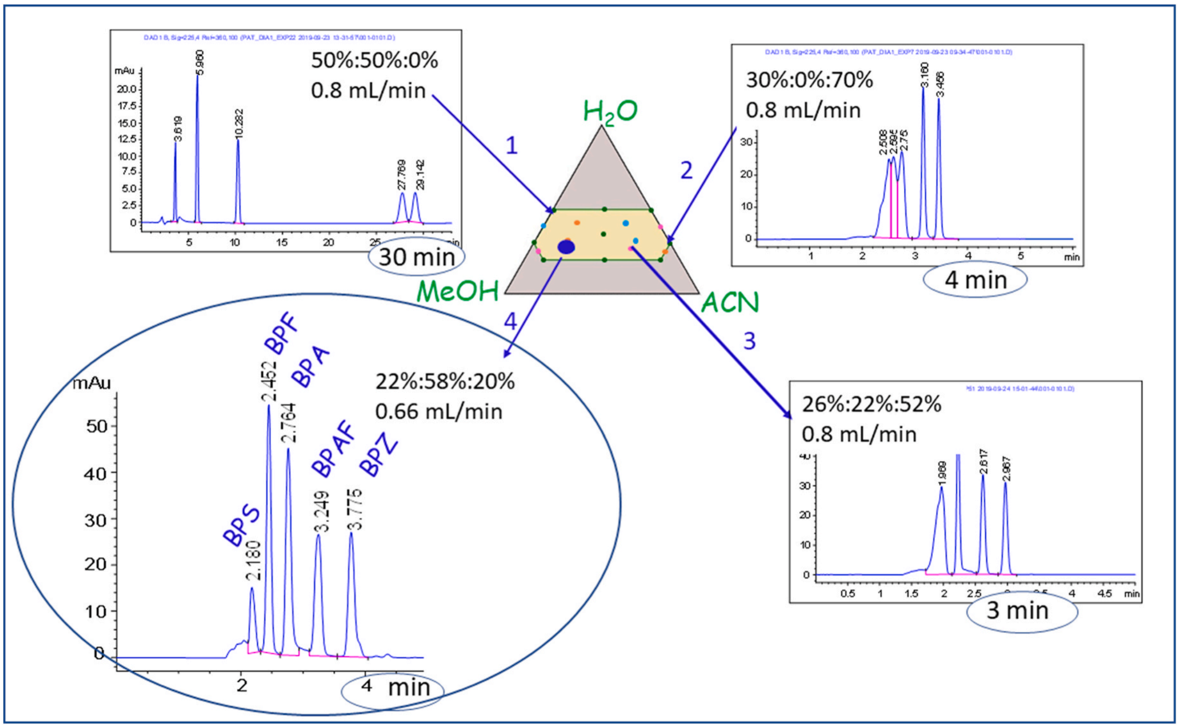


Fig. 6. Determination of five bisphenols: BPS (Bisphenol-S), BPF (Bisphenol-F), BPA (Bisphenol-A), BPAF (Bisphenol-AF) and BPZ (Bisphenol-Z) in ref. [23]. The chromatograms obtained with different CMPs are shown in each graph, with their position in a ternary diagram for the mixture indicated by the arrows.

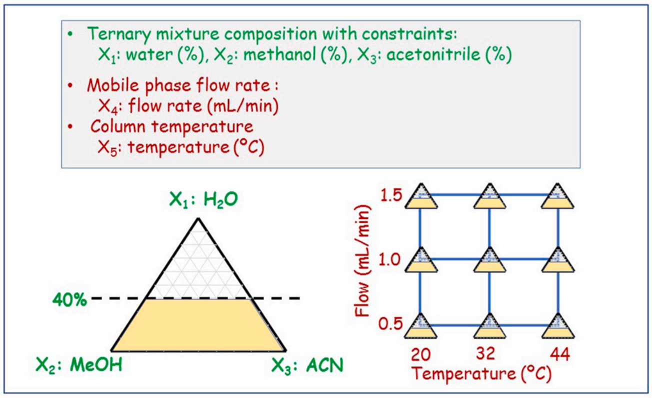


Fig. 7. Experimental domain for mixture-process design (ternary mixtures with constraints and the process variables flow rate and temperature) in case II.

maximum of the variance function equal to 0.91, including 16 protected experimental points, which were ternary mixtures at different flow rates and column temperatures from the 405 initial experiments.

4.3.2. *Fitting and inversion of a PLS2 model*

To fit the PLS2 model, there are five CMPs (the ternary mixture and flow rate of the mobile phase, and the column temperature) and eight CQAs, defined as the seven resolutions between contiguous chromatographic peaks (for the three emission wavelengths used to record the chromatograms) and the final time, which is the time needed to complete the chromatograms.

Interactions and/or strong nonlinear effects of the factors on the responses are expected, so the postulated model of Eq. (8) has 27 coefficients (β's) including the interactions between components of the mixture (X_1, X_2, X_3) and process variables (X_4 and X_5). Eq. (8) is similar to Eq. (3) but with two process variables.

$$\begin{aligned}\mathbf{Y} &= \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \\ &\beta_{12} X_1 X_2 + \beta_{13} X_1 X_3 + \beta_{23} X_2 X_3 + \sum_{j=1}^{3} \left(\beta_{4j} X_4 X_j + \beta_{5j} X_5 X_j\right) + \beta_{45} X_4 X_5 + \\ &\sum_{j=4}^{5} \left(\beta_{12j} X_1 X_2 X_j + \beta_{13j} X_1 X_3 X_j + \beta_{23j} X_2 X_3 X_j\right) + \sum_{j=1}^{3} \beta_{45j} X_4 X_5 X_j + \\ &\beta_{1245} X_1 X_2 X_4 X_5 + \beta_{1345} X_1 X_3 X_4 X_5 + \beta_{2345} X_2 X_3 X_4 X_5 \end{aligned} \tag{8}$$

With the 45 experiments (42 from the D-optimal design and 3 replicates), the **X** matrix of predictor variables is 45 × 27. The **Y** matrix, 45 × 8, contains the values of the CQAs obtained after running the experiments. With autoscaled **X** and **Y**, and crossvalidation with venetian blinds (five splits and blind thickness equal to one), a PLS2 model with 11 latent variables was built. The fitted model explains 97.93% of the total variance in **X** and 98.18% in **Y**.

The desired CQAs for the chromatograms for the joint determination of the ten PAHs were: resolution greater than 1.4 and the final time as short as possible, but not greater than 15 min. Eq. (9) shows $\mathbf{y}_{\text{des}}$ with the last coordinate in the form of the logarithmic transformation needed for the final time to fit the PLS2 model.

$$\mathrm{y}_{\text{des}} = (1.4, 1.4, 1.4, 1.4, 1.4, 1.4, 1.4, \log_{10}(15))^T \tag{9}$$

The inversion of the model should provide the experimental conditions (five-dimensional vector with the values of the CMPs) to obtain characteristics of the chromatograms close to the CQAs (eight-dimensional vector). Again, the PLS2 model is fitted with 27 predictor variables, not just with the five that can be modified in the laboratory, making the algebraic inversion infeasible. Therefore, in order to obtain experimental conditions that guarantee the required CQAs, the computational alternative explained before was used together with the Pareto front [35].

The parallel coordinates plot in Fig. 8 displays the 300 solutions in the Pareto front that meet the desired chromatogram characteristics. The first five coordinates correspond to the experimental factors (proportion of water, methanol and acetonitrile, mobile phase flow rate, and column temperature), the next seven are the peak resolutions, Y_i (i = 1,

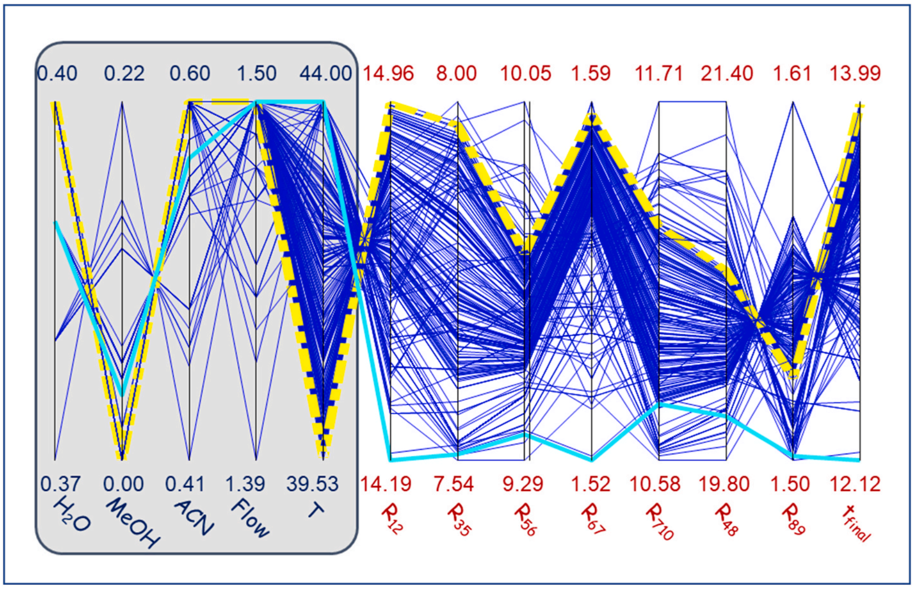


Fig. 8. Parallel coordinates plot of the Pareto front resulting from the inversion of the PLS2 model. The CMPs are those marked in grey, the remaining coordinates are the corresponding values of the CQAs. $Y_1 = R_{12}$, $Y_2 = R_{35}$, $Y_3 = R_{56}$, $Y_4 = R_{67}$, $Y_5 = R_{710}$, $Y_6 = R_{48}$, $Y_7 = R_{89}$, and $\log_{10}(t_f)$, which is Y_8. R is peak resolution, t is time.

…, 7), and the last vertical line is the coordinate of the final time, t_f, in the original minutes to facilitate exploration.

The bounds before range scaling, at the top and bottom of each coordinate, show that the found experimental conditions, in grey in Fig. 8, define a small region with only 37–40% water, mixed with up to 22% methanol, and less than 60% acetonitrile (in the corresponding proportions), linked to high values of flow rate (greater than 1.39 mL min^{-1}) and temperatures greater than 39.53 °C. The upper bounds of the last two factors are those already established for the experimental domain, 1.50 mL min^{-1} and 44 °C.

Although each setting of the CMPs shown in grey fulfils the specific AP attributes (also referred to as the “analytical knowledge space” [5]), they cannot be selected with any combination in the range shown. For example, to achieve the CQAs marked with the yellow dot dashed line, a mobile phase with a binary mixture of water (X_1, 0.40) and acetonitrile (X_3, 0.60), a flow rate of 1.5 mL min^{-1} and 44 °C must be used to obtain all resolutions greater than 1.53, and a final time of 13.99 min. Another extreme solution in terms of the CQAs is the continuous cyan line, which is the one with the shortest chromatogram (12.12 min) at the cost of having ‘limiting’ values of resolution in R_{89} (Y_7) and almost in R_{67} (Y_4). These characteristics are expected when using a ternary mixture of around 0.39 water (X_1), 0.06 methanol (X_2), and 0.55 acetonitrile (X_3) at the maximum flow rate and temperature, of 1.5 mL min^{-1} and 44 °C, respectively.

4.3.3. MODR computation and validation

All solutions in Fig. 8 meet the established requirements for the CQAs, the ATP, as predicted by the PLS2 model. This means that all the peaks are expected to be well resolved, with a final time of less than 14 min. However, these predicted CQAs are affected by uncertainty because they have been predicted with the fitted model. Consequently, it is necessary to determine the region of the CMPs with a guarantee that the required CQAs are met, which is the MODR.

The validation of this MODR in the proposal of this paper is done experimentally according to ref. [7]. To do that, the 300 solutions in the Pareto front are the starting point. The next step is to check that moving CMPs in the region defined by these solutions, in practice, give characteristics of the chromatograms that are similar enough to those predicted, so that they are equally valid to carry out the determination. It is clear that the experimental validation cannot include all 300 experiments in Fig. 8. In order to visualize them so as to choose some representative ones, a Principal Component Analysis (PCA) was performed with the CMPs in Fig. 8 (those marked in grey). With autoscaled data, two principal components were selected on the basis of crossvalidation with venetian blinds (ten splits and blind thickness equal to one). They explain 83.85% of the variance of the 300 experimental conditions.

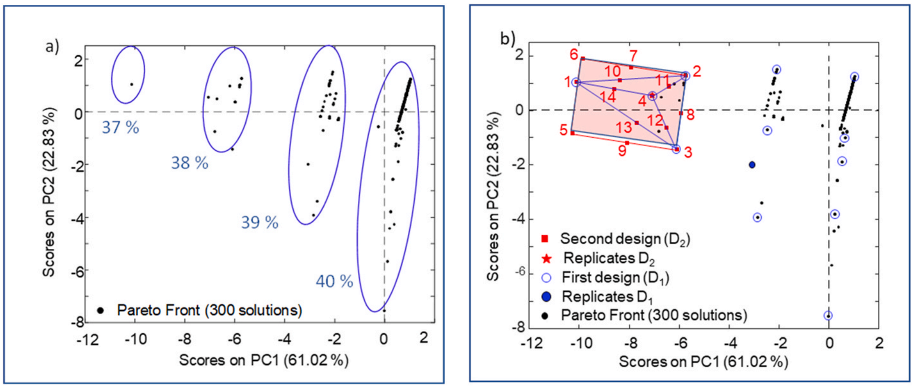


Fig. 9. Scores of the CMPs on the PCA plane of the 300 solutions of the Pareto front in case II, black dots in a) and b). In b) the sets of experiments in D_1 and D_2 are depicted in blue and red, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Table 1
CMPs of the 14 experiments that define the MODR. X_1, X_2 and X_3 are the proportions of H_2O, MeOH and ACN respectively. X_4 is the mobile phase flow rate (ml min^{-1}) and X_5 is the column temperature (°C).

Experiment	X_1	X_2	X_3	X_4	X_5
1	0.37	0.22	0.41	1.50	44.00
2	0.38	0.13	0.49	1.50	41.90
3	0.38	0.13	0.49	1.47	43.80
4	0.38	0.16	0.46	1.50	44.00
5	0.37	0.22	0.41	1.47	44.00
6	0.37	0.22	0.41	1.50	42.00
7	0.38	0.19	0.43	1.50	42.00
8	0.38	0.13	0.49	1.49	42.90
9	0.37	0.16	0.47	1.47	44.00
10	0.37	0.18	0.44	1.50	43.20
11	0.38	0.15	0.48	1.50	43.00
12	0.38	0.14	0.48	1.48	43.90
13	0.38	0.17	0.46	1.48	43.90
14	0.38	0.19	0.44	1.50	44.00

Fig. 9a shows the 300 scores. The circles surround the points with the same amount of water. It is seen that their percentage increases from 37 to 40% when moving from left to right on the first principal component, and that their dispersion along the second principal component also increases. Similar patterns would be seen if the scores were colored according to methanol or acetonitrile content (not shown here).

The small blue circles in Fig. 9b mark the projection of the CMPs selected for the experimental validation. Together with some replicates (filled circles), 16 experiments were carried out in a first set, D_1, to check the fulfilment of the CQAs on the obtained chromatograms, which was not the case except for four experiments. The settings of the CMPs of these four experiments are written in Table 1, numbers 1 to 4, and their scores are shown in red in Fig. 9b, identified with the same numbers.

These four experiments are part of a new set of planned experiments, D_2, whose scores are those in red in Fig. 9b (a total of 14 experiments). The superimposed red rectangle in Fig. 9b illustrates the reasoning behind the selection of these 14 conditions, which is explained bellow.

The scores 1 to 4 define a triangular shape that suggests consideration of the rectangle in which the triangle is inscribed by adding two other experiments, numbers 5 and 6 in both Fig. 9b and Table 1, that share the proportions of the mobile phase mixture with experiment number 1.

These 6 'points' (CMPs in Table 1, with their corresponding scores as dots in Fig. 9b) would form the vertices of the MODR in the five-dimensional space where the CMPs lie. Therefore, to better asses the experimental validity of the CQAs when moving in this restricted domain, intermediate experiments were chosen between any pair of the six already planned, up to the 14 points in Table 1, whose scores are marked in red in Fig. 9b.

The chromatograms obtained with the 14 CMPs in Table 1 all fulfilled the required CQAs, so the estimated MODR is made up of these 14 settings of the CMPs together with all their convex combinations, whose projection onto the PCA plane is inside the red polygon in Fig. 9b.

The discrete nature of the MODR in Table 1 is no obstacle to its use in other values of the different CMPs. For example, considering a mobile phase with 15% (0.15) MeOH, this 0.15 is a convex combination of the 13% and 22% MeOH in experiments 2 and 5 of Table 1, because for $\lambda = 0.78$, $0.15 = \lambda\ 0.13 + (1\text{-}\lambda)\ 0.22$ holds. Therefore, the corresponding settings of the CMPs to be used within the MODR would be $0.78 \times (0.38, 0.13, 0.49, 1.50, 41.90) + 0.22 \times (0.37, 0.22, 0.41, 1.47, 44.00) = (0.38, 0.15, 0.47, 1.49, 42.36)$. There are other solutions that can be considered as well.

The two cases just explained illustrate a way of working within the framework of AQbD that is an alternative to the usual approximations (as summarized in the introduction and described in the supplementary material). Among the limitations, the experimental validation requires additional steps in the laboratory with the consequent experimental effort, but avoids the need to know or to assume (in many cases arbitrarily) an a priori probability distribution to give a probabilistic confidence on the estimated design space/MODR. Another limitation comes from the little use of the parallel coordinates plot or the Pareto front in analytical chemistry, in spite of its increasing availability in the usual software.

Among the properties of interest, PLS2 is more versatile and the models are not reduced to (second-order) polynomials. There is a higher computational cost to estimate the knowledge space (and the design space/MODR inside it), but the procedure is still viable (minutes) compared to the laboratory experimental time (days).

Regarding the multiobjective approach, compared to the use of a desirability function, where the conditions are imposed before computing the solutions, the Pareto front allows exploration *a posteriori*, once the extent of the conflict is already observed. In fact, both cases studied the results are far superior to simply superimposing the level curves of the models for each pair of CMPs, which hinders the possible conflicts among CQAs.

Moreover, the time needed to estimate the Pareto front is also justified because the deep exploration of the Pareto front with the help of the parallel coordinates plot gives insight into the method and promotes the understanding of the process, which is also a requirement of AQbD. In other words, the joint exploration of CMP and CQA in the Pareto front provides a more detailed information about the analytical method than the simple numerical fulfillment of the requirements for a CQA.

Finally, the convex hull around the discrete estimation of the MODR does not need to impose that the MODR be a hypercube inside the "analytical knowledge space".

5. Conclusions

In the context of analytical quality by design, a systematic approach to the development of an analytical method is presented, illustrated with some liquid chromatographic procedures for the determination of several analytes.

The use of experimental designs, in particular combined designs, dealing with mixtures (composition of mobile phase) and process variables (flow, temperature), and the reduction of the number of experiments achieved with a D-optimal design provide a representative set of chromatograms to have a training set for modelling the relationship between predictor variables (CMPs) and responses (CQAs).

The modelling is done by building a predictive PLS2 model, whose computational inversion provides Pareto-optimal CMPs, which are the starting elements to estimate and experimentally validate the Method Operable Design Region.

The MODR established via PLS2 explicitly preserves the correlation between the CMPs and the CQAs, even when computing convex combinations of the chromatographic conditions that form the discrete MODR.

The parallel coordinates plot of the Pareto-optimal front, i.e., the settings of the CMPs together with the corresponding optimal values of the CQAs, allows the MODR to be displayed on a single graph, thus avoiding the need to hold some of the CMPs constant and/or to intersect "overlay maps" to ensure the specifications in the CQAs, with the consequent risk of misinterpretation.

The procedure developed in the context of AQbD represents an improvement over other approaches used to date.

Although the procedure has been illustrated with chromatographic procedures, it is general and can be applied to other instrumental applications (or industrial processes).

CRediT authorship contribution statement

M.C. Ortiz: Conceptualization, Data curation, Methodology, Funding acquisition. **L.A. Sarabia:** Conceptualization, Data curation, Methodology, Writing – original draft. **M.S. Sánchez:** Conceptualization, Methodology, Writing – review & editing, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

This work was supported by the Spanish MINECO (AEI/FEDER, UE), CTQ2017-88894-R and Consejería de Educación de la Junta de Castilla y León through project BU052P20. Both projects were co-financed with European Regional Development Funds.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.aca.2023.341620.

References

[1] ICH Harmonised Tripartite Guideline, Pharmaceutical Development Q8(R2), International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use, Geneva, Switzerland, 2009.

[2] F.G. Vogt, A.S. Kord, Development of quality-by-design analytical methods, J. Pharmaceut. Sci. 100 (3) (2011) 797–812.

[3] S. Orlandini, S. Pinzauti, S. Furlanetto, Application of quality by design to the development of analytical separation methods, Anal. Bioanal. Chem. 405 (2013) 443–450, https://doi.org/10.1007/s00216-012-6302-2.

[4] P. Ramalingam, B.Jahnavi, Chapter 5 QbD Considerations for Analytical Development in S. Beg, M.D.S. Hasnain (eds) Pharmaceutical Quality by Design, Academic Press London Wall, London EC2Y 5AS, United Kingdom. https://doi.org/10.1016/B978-0-12-815799-2.00005-8.

[5] S. Beg, J. Haneef, M. Rahman, R. Peraman, M. Taleuzzaman, W.H. Almalki, Chapter 1. Introduction to analytical quality by design, in: S. Beg, S. Hasnain, M. Rahman, W.H. Almalki (Eds.), Handbook of Analytical Quality by Design, Academic Press, 2021, https://doi.org/10.1016/B978-0-12-820332-3.00009-1.

[6] S. Beg, M. Rahman, Chapter 5. QbD Analytical quality by design for liquid chromatographic method development, in: S. Beg, M.S. Hasnain, M. Rahman, W. H. Almalki (Eds.), Handbook of Analytical Quality by Design, 2021, pp. 87–97, https://doi.org/10.1016/B978-0-12-820332-3.00010-8.

[7] ICH Harmonised Tripartite Guideline, Analytical Method Development Q14, International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use, Geneva, Switzerland, 2022.

[8] ICH Harmonised Tripartite Guideline, Validation of Analytical Methods Q2(R2), International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use, Geneva, Switzerland, 2022.

[9] S. Orlandini, G. Hancu, Z.-I. Szabó, A. Modroiu, L.-A. Papp, R. Gotti, S. Furlanetto, New trends in the quality control of enantiomeric drugs: quality by design compliant development of chiral capillary electrophoresis, Methods Mol. 27 (2022) 7058, https://doi.org/10.3390/molecules27207058.

[10] T. Bastogne, F. Caputo, A. Prina-Mello, S. Borgos, M. Barberi-Heyob, A state of the art in analytical quality-by-design and perspectives in characterization of nano-enabled medicinal products, J. Pharmaceut. Biomed. Anal. 219 (2022), 114911, https://doi.org/10.1016/j.jpba.2022.114911.

[11] E. Rozet, P. Lebrun, B. Debrus, B. Boulanger, P. Hubert, Design spaces for analytical methods, Trends Anal. Chem. 42 (2013) 157–167, https://doi.org/10.1016/j.trac.2012.09.007.

[12] H. Jambo, P. Hubert, A. Dispas, Supercritical fluid chromatography for pharmaceutical quality control: current challenges and perspectives, Trends Anal. Chem. 146 (2022), 116486, https://doi.org/10.1016/j.trac.2021.116486.

[13] D. Brynn Hibbert, Experimental design in chromatography: a tutorial review, J. Chromatogr. B 910 (2012) 2–13, https://doi.org/10.1016/j.jchromb.2012.01.020.

[14] I.M. Fukuda, C. Francin Fidelis Pinto, C. dos Santos Moreira, A. Morais Saviano, F. Rebello Lourenço, Design of experiments (DoE) applied to pharmaceutical and analytical quality by design (QbD), Braz. J. Pharmaceut. Sci. 54 (2018), e01006, https://doi.org/10.1590/s2175-97902018000001000.

[15] L.A. Sarabia, M.C. Ortiz, M.S. Sánchez, Response surface methodology, in: S. Brown, R. Tauler, B. Walczak (Eds.), Comprehensive Chemometrics. Chemical and Biochemical Data Analysis, second ed., Elsevier, 2020, pp. 287–326, https://doi.org/10.1016/B978-0-12-409547-2.14756-0.

[16] A. Beal, R. Phan-Tan-Luu, Nonclassical experimental designs, in: S. Brown, R. Tauler, B. Walczak (Eds.), Comprehensive Chemometrics. Chemical and Biochemical Data Analysis, second ed., Elsevier, 2020, pp. 385–410, https://doi.org/10.1016/B978-0-12-409547-2.14868-1.

[17] R. Phan-Tan-Luu, M. Sergent, Nonclassical experimental designs, in: S. Brown, R. Tauler, B. Walczak (Eds.), Comprehensive Chemometrics. Chemical and Biochemical Data Analysis, first ed., Elsevier, 2009, pp. 453–499, https://doi.org/10.1016/B978-044452701-1.00085-5.

[18] G. Piepel, B. Pasquini, S. Cooley, A. Heredia-Langner, V. Orlandini, S. Furlanetto, Mixture-process variable approach to optimize a microemulsion electrokinetic chromatography method for the quality control of a nutraceutical based on coenzyme Q10, Talanta 97 (2012) 73–82.

[19] S. Orlandini, B. Pasquini, M. Stocchero, S. Pinzauti, S. Furlanetto, An integrated quality by design and mixture-process variable approach in the development of a capillary electrophoresis method for the analysis of almotriptan and its impurities, J. Chromatogr. A 1339 (2014) 200–209.

[20] S. Orlandini, B. Pasquini, C. Caprini, M. Del Bubba, L. Squarcialupi, V. Colotta, S. Furlanetto, A comprehensive strategy in the development of a cyclodextrins-modified microemulsion electrokinetic chromatographic method for the assay of diclofenac and its impurities: mixture-process variable experiments and quality by design, J. Chromatogr. A 1466 (2016) 189–198.

[21] R. Deiddaa, S. Orlandini, P. Hubert, C. Hubert, Risk-based approach for method development in pharmaceutical quality control context: a critical review, J. Pharmaceut. Biomed. Anal. 161 (2018) 110–121, https://doi.org/10.1016/j.jpba.2018.07.050.

[22] T. Tome, N. Žigart, Z. Časar, A. Obreza, Development and optimization of liquid chromatography analytical methods by using AQbD principles: overview and recent advances, Org. Process Res. Dev. 23 (2019) 1784–1802, https://doi.org/10.1021/acs.oprd.9b00238.

[23] M.M. Arce, S. Ruiz, S. Sanllorente, M.C. Ortiz, L.A. Sarabia, M.S. Sánchez, A new approach based on inversion of a partial least squares model searching for a preset analytical target profile. Application to the determination of five bisphenols by liquid chromatography with diode array detector, Anal. Chim. Acta 1149 (2021), 338217, https://doi.org/10.1016/j.aca.2021.338217.

[24] M.M. Arce, S. Ruiz, M.S. Sánchez, L.A. Sarabia, M.C. Ortiz, Method operable design region obtained with a partial least squares model inversion in the determination of ten polycyclic aromatic hydrocarbons by liquid chromatography with fluorescence detection, J. Chromatogr. A 1657 (2021), 462577, https://doi.org/10.1016/j.chroma.2021.462577.

[25] M.C. Ortiz, L.A. Sarabia, A. Herrero, C. Reguera, S. Sanllorente, M.M. Arce, O. Valencia, S. Ruiz, M.S. Sánchez, Partial least squares model inversion in the chromatographic determination of triazines in water, Microchem. J. 164 (2021) 105971, https://doi.org/10.1016/j.microc.2021.105971.

[26] A.P. Ferreira, M. Tobyn, Multivariate analysis in the pharmaceutical industry: enabling process understanding and improvement in the PAT and QbD era. http://informahealthcare.com/phd. ISSN: 1083-7450 (print), 1097-9867 (electronic) Pharm Dev Technol, Early Online: 1–15.

[27] S. Orlandini, B. Pasquini, M. Stocchero, S. Pinzauti, S. Furlanetto, An integrated quality by design and mixture-process variable approach in the development of a capillary electrophoresis method for the analysis of almotriptan and its impurities, J. Chromatogr. A 1339 (2014) 200–209, https://doi.org/10.1016/j.chroma.2014.02.088.

[28] C. Ancillotti, S. Orlandini, L. Ciofi, B. Pasquini, V. Caprini, C. Droandi, S. Furlanetto, M. Del Bubba, Quality by design compliant strategy for the development of a liquid chromatography–tandem mass spectrometry method for the determination of selected polyphenols in *Diospyros kaki*, J. Chromatogr. A 1569 (2018) 79–90, https://doi.org/10.1016/j.chroma.2018.07.046.

[29] L. Nomparia, S. Orlandini, B. Pasquini, C. Campa, M. Rovini, M. Del Bubba, S. Furlanetto, Quality by design approach in the development of an ultra-high performance liquid chromatography method for Bexsero meningococcal group B vaccine, Talanta 178 (2018) 552–562, https://doi.org/10.1016/j.talanta.2017.09.077.

[30] M.B. Wise, N.B. Gallagher, R. Bro, J.M. Shaver, W. Winding, R.S. Koch, PLS Toolbox 8.8.1, Eigenvector Research Inc., Wenatchee, WA, USA, 2020.

[31] MATLAB, Version 9.7.0.1190202, The Mathworks, Inc., Natick, MA, USA, 2019. R2019b.

[32] D. Mathieu, J. Nony, R. Phan-Than-Lu, NEMRODW, L.P.R.A.I, Marseille, France, 2015. Version 2015.

[33] S. García-Muñoz, T. Kourti, J.F. MacGregor, F. Apruzzese, M. Champagne, Optimization of batch operating policies. Part I. Handling multiple solutions, Ind. Eng. Chem. Res. 45 (2006) 7856–7866, https://doi.org/10.1021/ie060314g.

[34] E. Tomba, M. Barolo, S. García-Muñoz, General framework for latent variable model inversion for the design and manufacturing of new products, Ind. Eng. Chem. Res. 51 (2012) 12886–12900, https://doi.org/10.1021/ie301214c.

[35] S. Ruiz, M.C. Ortiz, L.A. Sarabia, M.S. Sánchez, A computational approach to partial least squares model inversion in the framework of the process analytical technology and quality by design initiatives, Chemometr. Intell. Lab. Syst. 182 (2018) 70–78, https://doi.org/10.1016/j.chemolab.2018.08.014.

[36] S. Ruiz, L.A. Sarabia, M.C. Ortiz, M.S. Sánchez, Residual spaces in latent variable model inversion and their impact in the design space for given quality characteristics, Chemometr. Intell. Lab. Syst. 203 (2020), 104040, https://doi.org/10.1016/j.chemolab.2020.104040.

[37] L.A. Sarabia, M.S. Sánchez, M.C. Ortiz, Introduction to ranking methods, in: M. Pavan, R. Todeschini (Eds.), Scientific Data Ranking Methods: Theory and Applications, vol. 27, Data Handling Sci. Techn., 2008, pp. 1–50, https://doi.org/10.1016/S0922-3487(08)10001-6.

[38] K. Deb, Multi-Objective Optimization Using Evolutionary Algorithms, Wiley Interscience Series in Systems and Optimization, Wiley, Chichester, 2001.