Available online at www.sciencedirect.com

SCIENCE DIRECT®

Chemometrics and Intelligent Laboratory Systems 73 (2004) 73–84

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemolab

Time-resolved QSAR: an approach to PLS modelling of three-way biological data

Lennart Eriksson[a,*], Johan Gottfries[b,c], Erik Johansson[a], Svante Wold[c]

[a] *Umetrics AB, POB 7960, SE-907 19 Umeå, Sweden*
[b] *AstraZeneca R&D, Mölndal, Sweden*
[c] *Res. Group of Chemometrics, Institute of Chemistry, Umeå University, Umeå, Sweden*

Available online 19 July 2004

Abstract

This paper outlines a novel approach to the analysis of three-way *Y*-data in quantitative structure–activity relationship (QSAR) modelling. The new method represents a modification of an existing approach for multivariate modelling of batch process data. It is based on unfolding the three-way **Y**-matrix into a two-way matrix according to a sequential order of an external variable. In QSAR, time, pH, or temperature at which the biological data were gathered, are conceivably such external variables. Thus, unfolding can be done differently depending on the objective of the investigation, thereby shifting the focus of the QSAR analysis. The ensuing multivariate data analysis uses two levels of modelling. (1) On the lower (observation) level a projections to latent structures (PLS) model is developed between the unfolded biological data and the external variable. This model will identify compounds with biological data being sensitive to changes in the external variable (like time, pH, or temperature). (2) The scores of the lower level model are then re-arranged to enable the upper (QSAR) level model. In this model, a battery of structure descriptors (*X*) is related to the **Y**-matrix of scores of the lower level model. As an example, a series of 35 compounds and their anti-microbial activity towards the bacterial strain *Escherichia coli* CCM2260 is used. This biological activity has been determined at different times (2 to 10 h) and pH-values (pH 5.6 to 8.0).

Keywords: QSAR; Time-resolved QSAR; Three-way *Y*-data; PCA; PLS

1. Introduction

The data analytical part of quantitative structure–activity relationship (QSAR) modelling has undergone a dramatic development at the end of the 20th century. A short historical recollection is found in Ref. [1]. The traditional approach to QSAR relies heavily on multiple linear regression (MLR), where a few and fairly uncorrelated structure descriptors (*X*) are related to a single biological activity variable, *y* [2]. With the advent of partial least squares projections to latent structures (PLS), QSAR analysis underwent a developmental leap, and QSARs involving multiple X's and Y's were suddenly a realistic option (see, e.g., Ref. [3]). And recent methodological development allows, e.g., hierarchical PLS, whereby very many (in the order of hundreds or thousands) variables, both in the **X**- and **Y**-matrices, are modelled in terms of block-relations connected across two or more model layers (see, e.g., Ref. [4]).

We here outline another methodological extension, addressing QSARs where the **Y**-matrix is three-dimensional. Fig. 1 shows the data arrangement for the present example. Toxicity data for 35 compounds have been gathered across variations in time and pH. Conventional QSAR analysis with two-way data matrices does not directly apply. Some kind of unfolding of the **Y**-matrix is needed to accomplish the QSAR analysis.

The new approach for handling three-way *Y*-data is based on a method of Wold et al. [5], originally developed to analyze three-way batch process data. A small modification of the original approach is proposed in order to handle three-way *Y*-data. This modification will focus the QSAR modelling, so that not only will it express the change in toxicity between compounds, but also be able to reveal time- and pH-dependent trends in the prevailing structure–activity relationships. This method is further presented in Section 4.

Thus, our objective is to outline this novel approach of analyzing three-way *Y*-data in QSAR. In so doing, we will

* Corresponding author. Tel.: +46-90-184582, +46-73-6824852 (mobile); fax: +46-90-184899.
E-mail address: lennart.eriksson@umetrics.com (L. Eriksson).

0169-7439/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2004.04.004

be using two different ways of unfolding the example *Y*-data, namely unfolding to focus on (i) time-dependent and (ii) on pH-dependent trends in the toxicity data.

2. Example data set

The example data set comprises a series of 4-pyranones (Fig. 2), for which Pirselova et al. [6,7] have determined antimicrobial activity towards the bacterial strain *Escherichia coli* CCM2260. The data set consists of $N=35$ compounds, of which 14 were ionizable and 21 non-ionizable in the testing protocol used [6,7]. To describe the physico-chemical properties a set of $K=26$ chemical descriptors (matrix **X**) was compiled. The *X*-descriptors were predominantly 2D-based structure descriptors [8,7,9–11].

The following *X*-descriptors were compiled: $x_1=\log P$ ("log P", [6,7]); $x_2=$ number of atoms ("N_At"); $x_3=$ number of bonds ("N_Bo"); $x_4=$ ratio number of bonds/number of atoms ("Nbo/Nat"); $x_5=$ number of rings ("N_Ring"); $x_6=$ number of rigid bonds ("N_RigBo"); $x_7=$ number of carbon atoms ("N_C"); $x_8=$ number of nonpolar atoms ("N_NonPolAt"); $x_9=$ highest molecular orbital Π-energy ("Pi-energy"); $x_{10}=$ highest molecular resonance energy ("Reson. Ene"); $x_{11}=$ highest occupied molecular orbital energy ("HOMO"); $x_{12}=$ lowest unoccupied molecular orbital energy ("LUMO"); x_{13} molecular weight ("Mw"); $x_{14}=$ molecular volume ("MolVol"); $x_{15}=$ polarizability ("Polariz"); $x_{16}=$ Ratio [Sum hydrogen bond donors and acceptors)/number of atoms] ("Sum (HBD + HBA)/N_At"); $x_{17}=$ calculated log *P* ("clogP"); $x_{18}=$ calculated molecular refractivity ("CMR"); $x_{19}=$ hydrogen bond donors ("HB-Donors"); $x_{20}=$ sum donors ("Sum_donor"); $x_{21}=$ sum total ("Sum_total"); $x_{22}=$ polar surface area ("PSA"); $x_{23}=$ nonpolar surface area ("NPSA"); $x_{24}=$ percentage polar surface area ("%PSA"); $x_{25}=$ percentage nonpolar surface area ("%NPSA"); $x_{26}=$ total surface area ("TSA").

The biological (*Y*-) data, expressed as the concentration decreasing bacterial growth by 50% (log 1/IC50), were acquired as a function of changing time (2 to 10 h after exposure) and pH (between 5.6 and 8.0). Hence, a low number means low toxicity and vice versa. Fig. 1 shows an overview of the **X**- and **Y**-matrix structures. As seen, the **Y**-matrix is built up of "directions" relating to $N=35$ compounds, $L=5$ time points (2, 4, 6, 8 and 10 h) and $M=6$ pH-values (5.6, 6.0, 6.6, 7.0, 7.6 and 8.0).

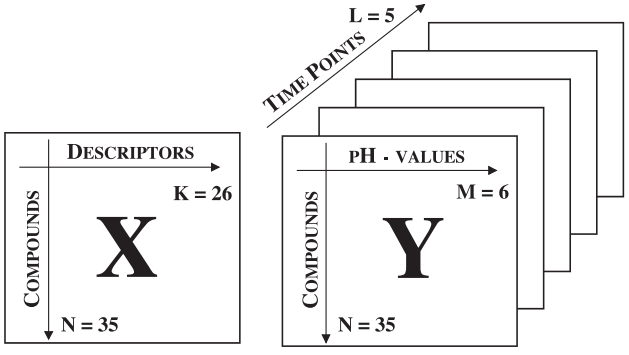

Fig. 1. Schematic overview of data arrangement of **X**- and **Y**-matrices.

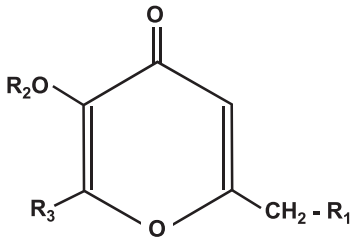

Fig. 2. Basic molecular structure of the investigated compounds. R_1, R_2, and R_3 are the three varied substituent positions.

All raw data are available from the corresponding author.

3. Data analytical method

In this paper we have used the software SIMCA-P, version 10 [12], and its implementations of standard and batch-based PLS [5,13–16].

3.1. Partial least squares projections to latent structures, PLS

PLS is a regression method that works with two matrices, **X** (e.g., chemical descriptors) and **Y** (e.g., biological responses), and has two objectives, namely to well approximate **X** and **Y**, and to model the relationship between them. The chemical variation in the predictor block (**X**) is summarized by the ($N*A$) *X*-scores, **T**, and the corresponding variation in the response block (**Y**) is described by the ($N*A$) *Y*-scores, *U*. Basically, PLS maximizes the covariance between **T** and **Y**. For each model dimension, a weight vector $\mathbf{w}'$, is computed, which reflects the contribution of each *X*-variable to the modelling of **Y**, in that particular model dimension. The resulting ($A*K$) **X**-weight matrix, $\mathbf{W}'$, is important since it reflects the structure in **X** that relates to **Y**. The corresponding matrix of *Y*-weights is designated $\mathbf{C}'$. Additionally, a matrix of *X*-loadings, $\mathbf{P}'$, is calculated in order to deflate **X** appropriately.

The decomposition in PLS of **X** and **Y** can be described as:

$$\boldsymbol{X}=\boldsymbol{TP}'+\boldsymbol{E};\ \boldsymbol{Y}=\boldsymbol{TC}'+\boldsymbol{F} \tag{2}$$

The set of PLS regression coefficients can then be computed according to:

$$\boldsymbol{B}=\boldsymbol{W}(\boldsymbol{P}'\boldsymbol{W})^{-1}\boldsymbol{C}' \tag{3}$$

Subsequently, an estimate of **Y**, $\hat{\mathbf{Y}}$, is obtained as:

$$\hat{\boldsymbol{Y}}=\boldsymbol{XW}(\boldsymbol{P}'\boldsymbol{W})^{-1}\boldsymbol{C}'=\boldsymbol{XB} \tag{4}$$

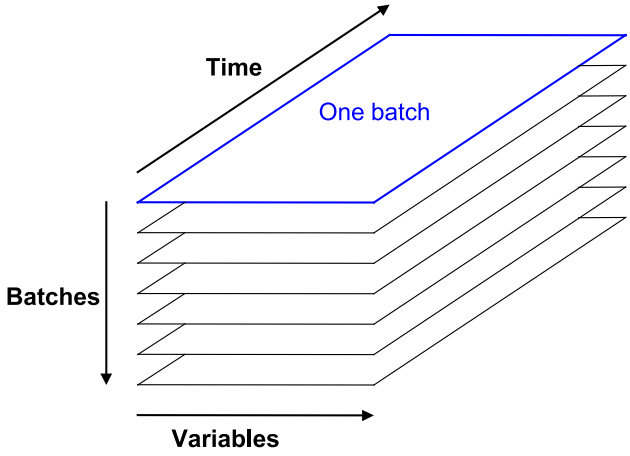

Fig. 3. A three-way table of historical batch process data. The data table comprises N batches, J time points, and K variables.

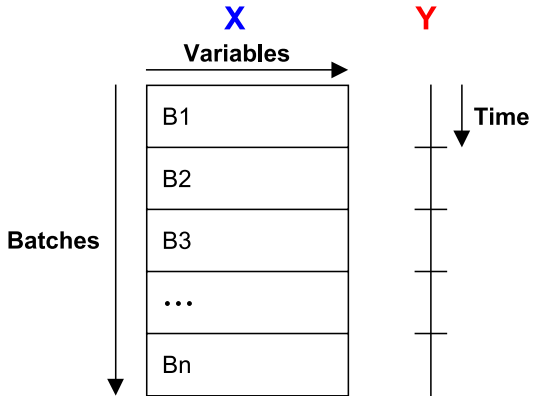

Fig. 5. The three-way data table of Fig. 3 is unfolded by preserving the direction of the variables. This gives a two-way matrix with $N*J$ rows and K columns. Each row contains data points x_{ijk} from a single batch observation (batch i, time j, variable k). If regression is made against local batch time, the resulting PLS scores usually reflect linear (t_1), quadratic (t_2), and cubic (t_3) relationships to local batch time.

Prior to the analysis, data were column-wise mean-centered and scaled to unit variance.

4. Time-resolved QSAR: a way to deal with three-way *Y*-data

4.1. Review of original approach for batch process data

The basic idea of the approach presented in Ref. [5] is to analyze three-way batch data in two model layers. Typical configurations of batch-data are seen in Figs. 3 and 4. On the lower (observation) level the three-way batch data are unfolded preserving the variable direction (Fig. 5), and a PLS model is computed between the unfolded X-data and time or a suitable maturity variable. The X-score vectors of this PLS model usually capture linear, quadratic, cubic, etc., dependencies between the measured process data and time or maturity. Subsequently, these score vectors are re-arranged (Fig. 6) and used on the upper (batch) level where relationships among whole batches are investigated (Fig. 7). The re-arrangement (cf. Fig. 6) of the scores and other model statistics (DModX, predicted time) enables batch control charts [5] to be produced. The resulting control charts can be used to follow the trace of a developing batch, and extract warnings when it tends to depart from the typical trace of a normal, good batch.

4.2. Implementation in QSAR when Y-matrix is three-way

We will now outline how the batch-approach can be tailored to suit the needs of three-way **Y**-matrices in QSAR studies. Thus, henceforth, any mention of unfolding refers to Y-data, not X-data.

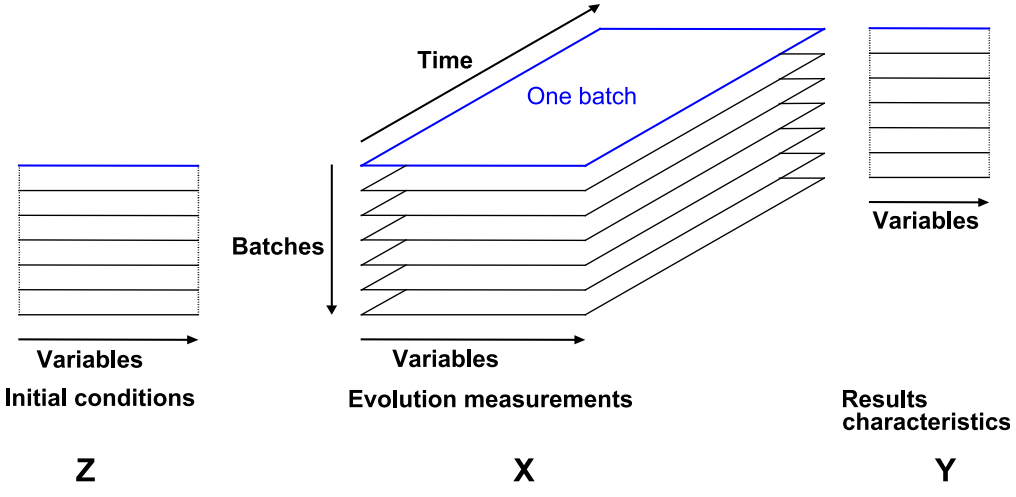

Fig. 4. Batch-data often involve three distinct blocks of data, i.e., initial conditions (the **Z**-matrix), evolution measurements (the **X**-matrix), and results and quality characteristics (the **Y**-matrix). These data tables can be analyzed independently with PCA or related to each other by PLS.

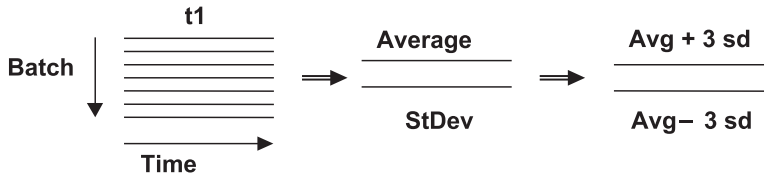

Fig. 6. The score values for each batch are arranged as row vectors underneath each other, giving a new **X**-matrix that has the number of rows equal to the number of batches in the reference data set. Observe that the drawing only outlines the re-arranging principle for the first score t_1. From this new matrix one calculates the averages and standard deviations (SDs) of the matrix columns, and subsequently control limits as averages ± 3 SD.

4.2.1. Lower level model: extracting time- or pH-dependent trends in toxicity data

In this context, there are two relevant ways of unfolding the **Y**-matrix. The first preserves the pH-direction (Fig. 8a) and the second the Time-direction (Fig. 8b). The former way captures time-dependent features in the data, and we call this "Time-unfolding" (Fig. 8a). The latter approach indicates whether there are features in the toxicity data critically influenced by pH-variation. We call this "pH-unfolding" (Fig. 8b).

The resulting two-way matrices then have either $N*L$ rows ($35*5=175$ observations; Time-unfolding) and $M=6$ columns (toxicities measured at different pH values), *or*, $N*M$ rows ($35*6=210$ observations; pH-unfolding) and $L=5$ columns (toxicities registered at different times). Hence, in both these cases, the unit in the unfolded matrices are the individual observations, not the entire compounds. The scores of the resulting PLS models may later be used on the upper model level.

4.2.2. Upper level model: time- or pH-resolved QSAR modelling

In order to accomplish the upper level model, i.e., the actual QSAR analysis, a re-arrangement of the PLS scores of the observation level models is necessary. The principle is the same as the one sketched in Figs. 6 and 7, the difference being the lower level score vectors are used to construct a **Y**-matrix, not an **X**-matrix.

Firstly, the score vectors of the observation level PLS model are divided in segments corresponding to each compound in the training set data. Secondly, these segmented score vectors are re-ordered such that all the score-values of one compound form one row vector of a new score matrix $\mathbf{Y}_T$. This is a new matrix where, in each row, all t_1-values of one compound are followed by all t_2-values of the same compound, which are followed by all t_3-values of the same compound, and so on (Fig. 9). This matrix $\mathbf{Y}_T$ has one row per training set compound. Observe that the drawing in Fig. 9 outlines the re-arranging principle for several sets of score vectors t_1, t_2, t_3, etc., whereas in reality any number of score vectors might be employed. As will be apparent below, only one segmented score vector will be used here to formulate the QSAR models.

Subsequently, the QSAR model can be calculated where the *X*-block of original chemical descriptors is related to the $\mathbf{Y}_T$-matrix. Depending on the choice of unfolding principle on the lower lever, the upper level QSAR modelling then interrogates the chemical descriptors from different viewpoints, i.e., to possibly uncover time-influences or pH-influences on the structure–activity relationship.

5. Results

5.1. Time-unfolding

5.1.1. Lower level reference PLS model based on 35 compounds

The first PLS model (Model 1) was calculated using the data arrangement shown in Fig. 8a, i.e., where the two-way matrix of toxicity data has $N*L$ rows ($35*5=175$) and $M=6$

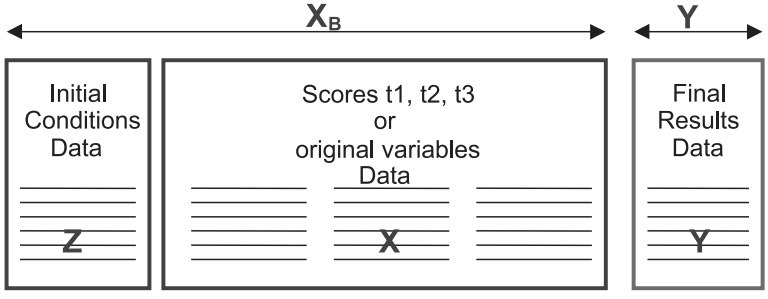

Fig. 7. In the batch level modelling, all available data are used to obtain a model of whole batches. Note that each row corresponds to one batch. Initial conditions data are often pooled with process evolution data to form a new **X**-matrix, $\mathbf{X}_B$. This $\mathbf{X}_B$-matrix is regressed against the final results contained in the **Y**-matrix. When used for batch monitoring, the resulting PLS-model may be used to categorize evolving batches as good or bad. It is also possible to interpret which initial condition data and process evolution data exert the highest influence on the type and quality of the resulting product.

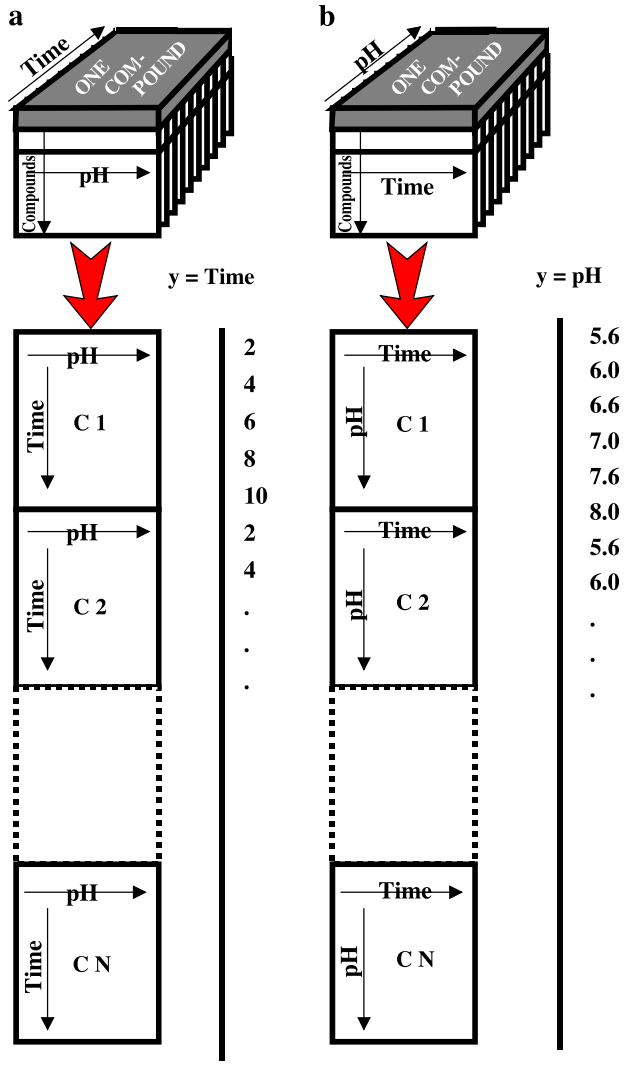

Fig. 8. Unfolding of *Y*-data preserving the pH-direction (a) or the Time-direction (b). The resulting two-way matrices then have either $N*L$ rows ($35*5=175$ observations; time-unfolding) and $M=6$ columns (pH-variables), *or*, $N*M$ rows ($35*6=210$ observations; pH-unfolding) and $L=5$ columns (Time-variables).

columns (toxicity measurements at six pH-values). This model had only one component explaining 93% of the variation in the unfolded toxicity data and 14% of the variation in the time-variable (see performance statistics in Table 1). This indicates there are no strong time-dependent features in the toxicity data.

However, the score line plot of this model (Fig. 10) strongly contradicts this initial statement, because 14 compounds (numbers 14–19, 24, 26, 27, 29, and 31–34) have very "steep" lines, i.e., toxicity data for these change significantly with time. Hence, we understand the reason for the initial poor time relation is data inhomogeneity.

5.1.2. Refined lower level PLS model based on 14 compounds

A second PLS model (Model 2, Table 1) was fitted to the 14 compounds exhibiting time-sensitive toxicity data. Now, the result was a strongly significant two-component model, which used 95% (86 + 9) of the variation in the toxicity data to explain 84% (62 + 22) of the time variable. Fig. 11a and b shows the scores and loadings of Model 2.

Time points corresponding to the same compound have been connected by a line (Fig. 11a). For all 14 compounds, the early time points are found in the left-hand part of the score plot, and the late ones in the right-hand area. Increasing t_1 score values correlate with decreasing toxicity. Hence, we may interpret the first score vector as a general toxicity scale, capturing a linear and monotonically decreasing trend in toxicity over time.

The second component uncovers a difference in toxicity profiles between two sub-groups of compounds. The "top" group encompasses 10 compounds (numbers 14–19, 24, 26, 27, and 29) and the "bottom" group four chemicals (numbers 31–34). Compounds 31–34 are much smaller (have lower molecular weight and molar volume) and less hydrophobic than the other 10, and also have non-hydrogen substituents in all three varied positions (cf. Fig. 2). Furthermore, the 10 compounds in the top cluster are all ionizable whereas the four compounds in the bottom group are not.

5.1.3. Upper level (QSAR) model

On the upper (QSAR) level attention will be given to the first score vector of the lower level model (Model 2). This is reasonable since this component accounts for 86% of the sum of squares of the toxicity data and was found to mirror a time trend of decreasing toxicity with increasing time.

The t_1 score vector of Model 2 was re-arranged as depicted in Fig. 9. Hence, a $\mathbf{Y}_T$-matrix consisting of 14 rows

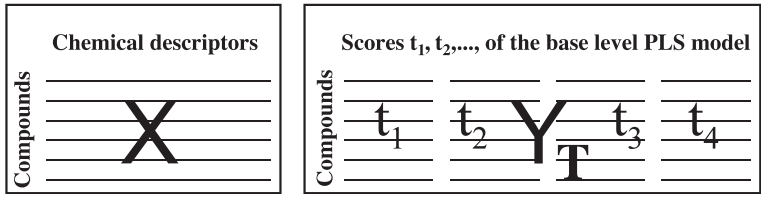

Fig. 9. Arrangement of data on the upper (QSAR) level. The new **Y**-matrix, $\mathbf{Y}_T$, consists of segmented and rotated score vectors drawn from the lower (observation) level model.

Table 1
Summary statistics of lower and upper level PLS models

Model no.	No. of compounds	R^2X[a]	R^2Y[b]	Q^2Y[c]	Remark
M1	35	0.93	0.14	0.13	Reference lower level model (time-unfolding)
M2	14	0.95	0.85	0.59	Revised lower level model (time-unfolding)
M3	14	0.46	0.75	0.66	Upper level QSAR model (time-unfolding)
M4	35	0.96	0.02	0.02	Reference lower level model (pH-unfolding)
M5	10	0.97	0.44	0.43	Revised lower level model (pH-unfolding)
M6	10	0.52	0.72	0.53	Upper level QSAR model (pH-unfolding)

[a] Explained sum of squares of *X*. In models M1, M2, M4 and M5, the *X*-block contains appropriately unfolded *Y*-data. In models M3 and M6, the *X*-block contains the chemical descriptor variables and their squared counterparts.

[b] Explained sum of squares of *Y*. In models M1 and M2, the single *Y*-variable is the time order of the measurements. In models M4 and M5, the single *Y*-variable is the pH-value at which the experiments took place. In models M3 and M6, the **Y**-matrix is composed of re-arranged score vectors derived from the relevant lower level PLS model. Score vectors from M2 build up the **Y**-matrix of M3, and score vectors from M5 build up the **Y**-matrix of M6.

[c] Predicted sum of squares of *Y*, estimated with cross-validation [27].

(compounds) and 5 columns (segmented and transposed t_1 score vector with five time points) was constructed. The $\mathbf{Y}_T$-matrix thus created contains a stable summary (i.e., score values) of the original toxicity measurements zooming-in their time-dependent properties. Below, we will develop a QSAR model between these $\mathbf{Y}_T$-data and the **X**-matrix, the latter of which consisting of the 26 chemical descriptors and their squared terms. The squared terms were introduced in order to search for possible non-linearities in the structure–activity relationship.

The resulting PLS model (Model 3, Table 1) had only one significant component with the following statistics: $R^2X=0.46$, $R^2Y=0.75$, and $Q^2Y=0.66$, which are satisfactory considering the model treating five *Y*-variables. Its PLS weights are plotted in Fig. 12. We can see that most of the *X*-variables are correlated within this subset of 14 compounds. An increased toxicity (lowered values of the five score responses) is coupled to an increased hydrophobicity, molecular weight and molar volume. Thus, although the raw toxicity data first have been summarized by a latent variable, and then rotated towards time, classical physico-chemical parameters are still meaningful in the interpretation of the QSAR model.

5.2. PH-unfolding

5.2.1. Lower level reference model based on 35 compounds

To investigate the relation between toxicity data and pH, the **Y**-matrix was unfolded as shown in Fig. 8b. The resulting two-way matrix of toxicity data has $N*M$ rows ($35*6=210$) and $L=5$ columns (toxicity readings at five time points). The PLS model fitted to this data set (Model 4, Table 1) comprised one component explaining 96% of the variation in the unfolded toxicity data and 2% of the variation in the pH-variable. This result implies there is no systematic change in toxicity data associated with the change in pH-values, at least not for the entire set of 35 compounds.

Interestingly, and analogously to Fig. 10, however, the score line plot given in Fig. 13 *does* indeed point to a subset of 10 compounds for which toxicity readings have an intrinsic pH-trend. These compounds are numbers 2–4, 6–8, and 31–

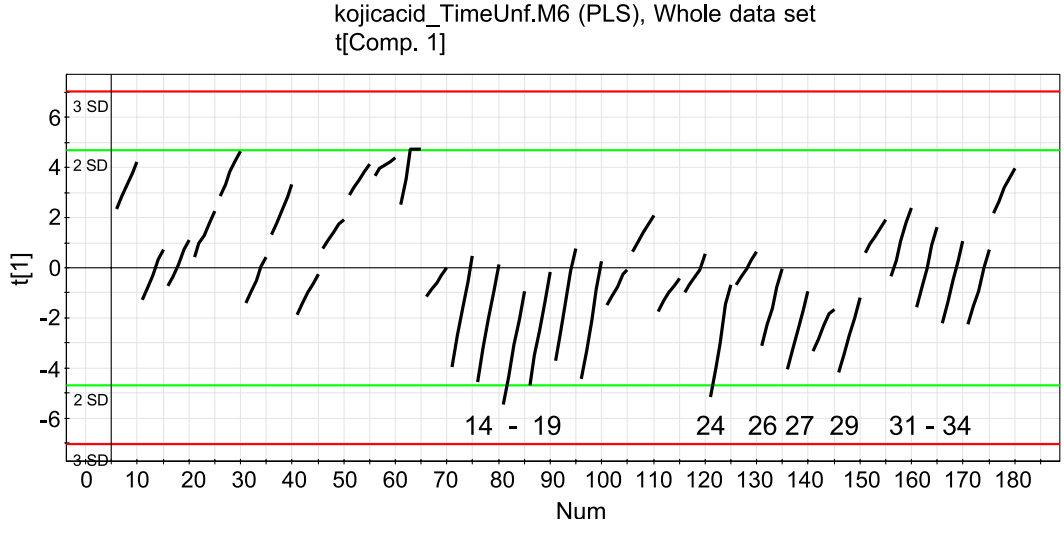

Fig. 10. Score line plot of the reference model (M1) based on time-unfolding. Score points originating from the same compound are shown as jointed lines. A vertical line would indicate very strong time dependencies and a horizontal line no connection with time whatsoever. Fourteen compounds, i.e., numbers 14–19, 24, 26, 27, and 31–34, have rather "steep" lines. This suggests toxicity data for these compounds change significantly with time.

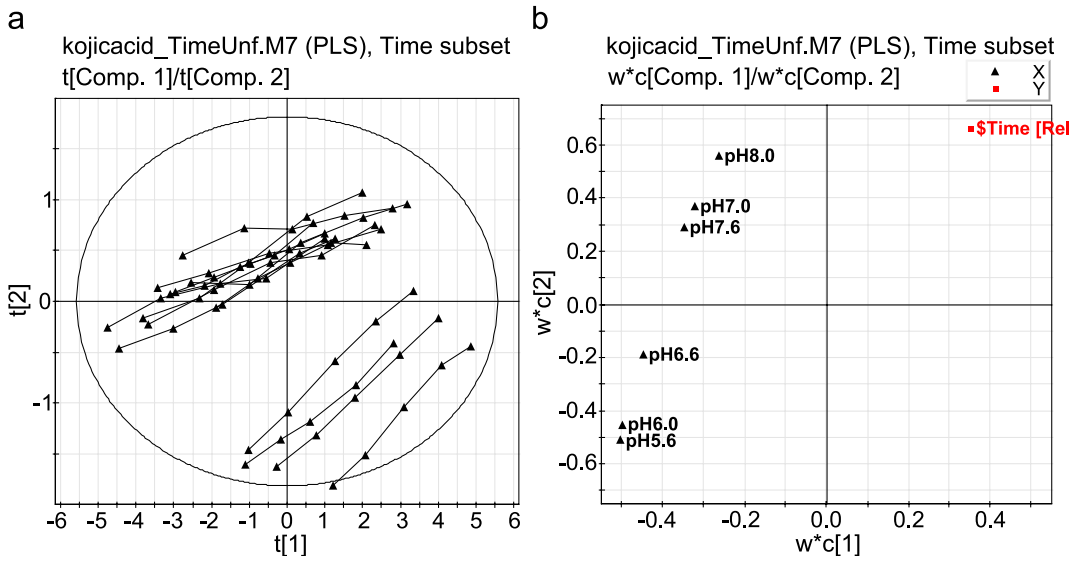

Fig. 11. (a, left) Score plot of the lower level PLS model (M2) based on the time-unfolding subset. Each line designates one compounds. (b, right) PLS loading plot corresponding to (a). The six pH-value-variables are seen to be correlated and are all positioned in the left-hand area of the loading plot. This suggests that toxicity decreases with time.

34, and they are all of the ionizable type. Hence, the reason for the initial poor pH relation is data inhomogeneity.

5.2.2. Refined lower level PLS model based on 10 compounds

The PLS model (Model 5, Table 1) fitted to the 10 compounds with pH-sensitive toxicity data consisted of one component utilizing 97% of the variation in the toxicity data to explain 43% of the pH-variable. Fig. 14a and b shows the scores and loadings of this model. Measurements corresponding to the same compound have been connected by a line.

When comparing (Figs. 10, 11a, and 14a), we realize that—for the two chosen subsets of compounds—the toxicity data are more sensitive to the time point than the pH-value. This is inferred from the fact that the "score-lines" are generally more straight and monotonically increasing when time rather than pH is involved as the external

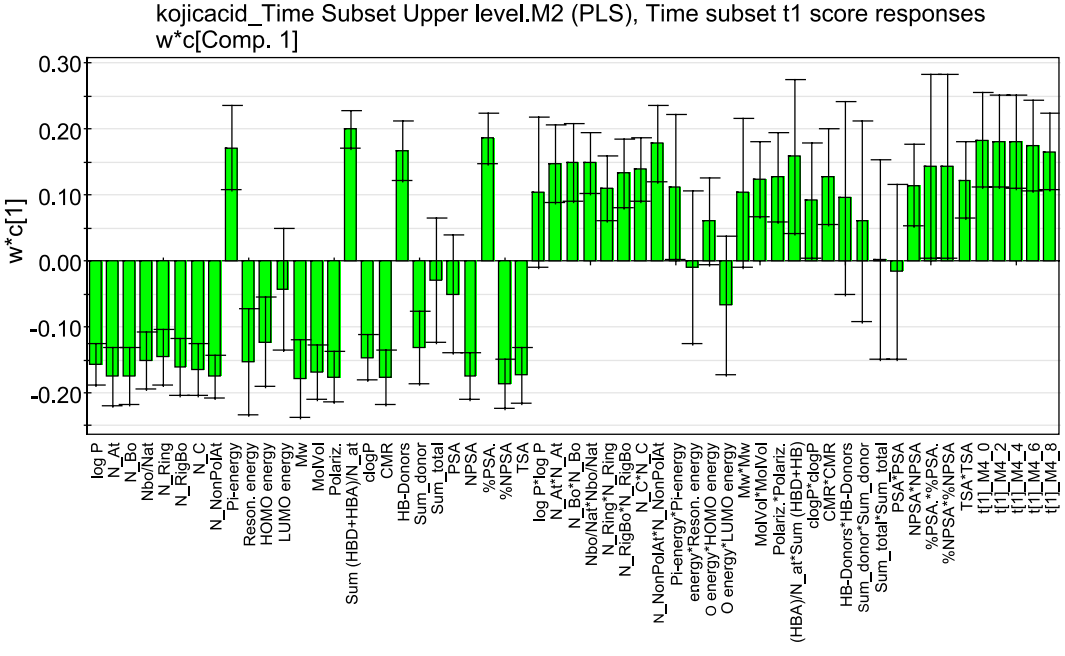

Fig. 12. PLS loadings of the upper level QSAR model (M3) based on the time subset. For a description of the *X*-variables, please see text in Section 2.

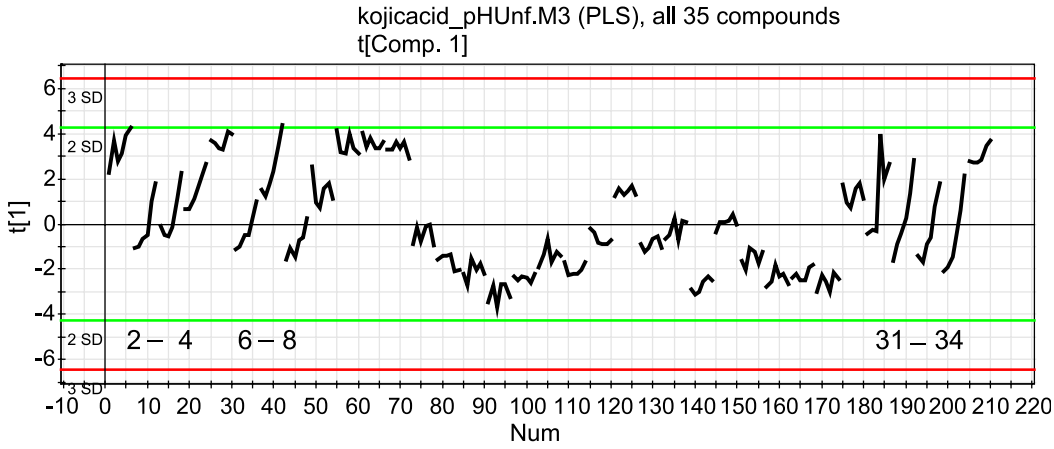

Fig. 13. Score line plot of the reference model (M4) based on pH-unfolding. This plot is interpreted similarly to Fig. 10. Ten compounds (2–4, 6–8, and 31–34) are seen to have toxicity data exhibiting pH-dependence.

variable. We recall that toxicity data were able to model 84% of the variation in time (Model 2) and here merely 43% of pH (Model 5).

For all 10 compounds plotted in Fig. 14a, the measurement conducted at the lowest pH (5.6) marks the left-hand "start" of a score-line, and at the other end we find the corresponding value measured at the highest pH (8.0). The corresponding loadings plotted in Fig. 14b demonstrate an inverse relationship between toxicity data and pH. Thus, remembering that numerically low toxicity values imply low toxicity, we can deduce that toxicity generally diminishes with increasing pH. This phenomenon is most pronounced for compounds 31–34.

5.2.3. Upper level (QSAR) model

To bring about the upper (QSAR) level model, the t_1 score vector of Model 5 was re-arranged as depicted in Fig. 9. The resulting $\mathbf{Y}_T$-matrix constituted 10 rows (compounds) and 6 columns (segmented and transposed t_1 score vector with six pH-values).

An initial QSAR model was computed between the $\mathbf{Y}_T$-matrix and the $\mathbf{X}$-matrix, the latter of which contained the 26 linear X-descriptors plus their squared counterparts. From this provisional model it was obvious that a majority of the chemical descriptors were not relevant in this application (no plot shown). Cautious variable selection according to the VIP-parameter in SIMCA-P [12,14] remedied this problem. Only X-variables with a VIP larger than 1.0 were kept [12,14], and hence the final model had 13 X-variables.

The resulting PLS model (Model 6, Table 1) contained only one component with the following performance statistics: $R^2X=0.52$, $R^2Y=0.72$, and $Q^2Y=0.53$. According to Fig. 15, which provides the PLS weights, a joint

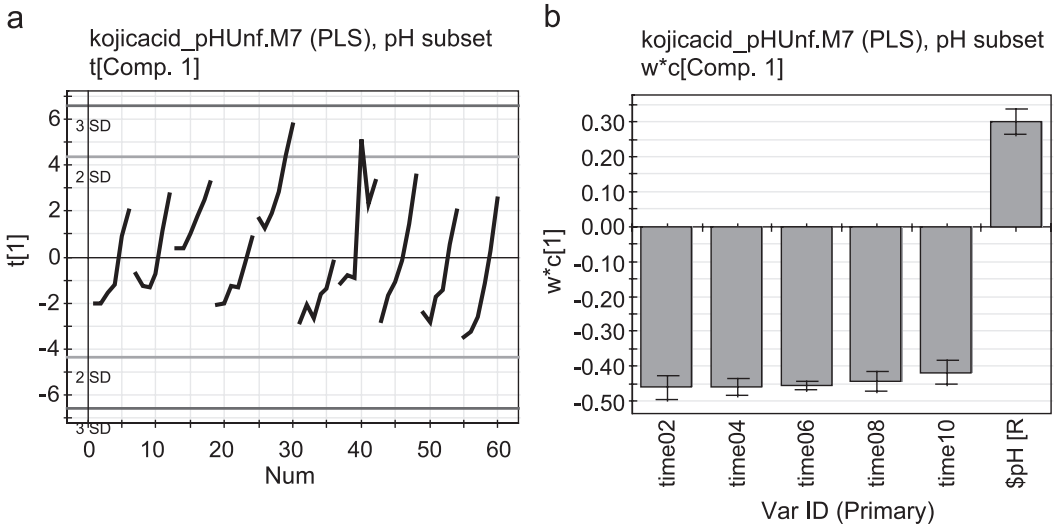

Fig. 14. (a, left) Score plot of the lower level PLS model (M5) based on the time-unfolding subset. Each line designates one compound. (b, right) PLS loading plot corresponding to Fig. 11a. The five time-point-variables are seen to be correlated and the model interpretation indicates that toxicity decreases with pH.

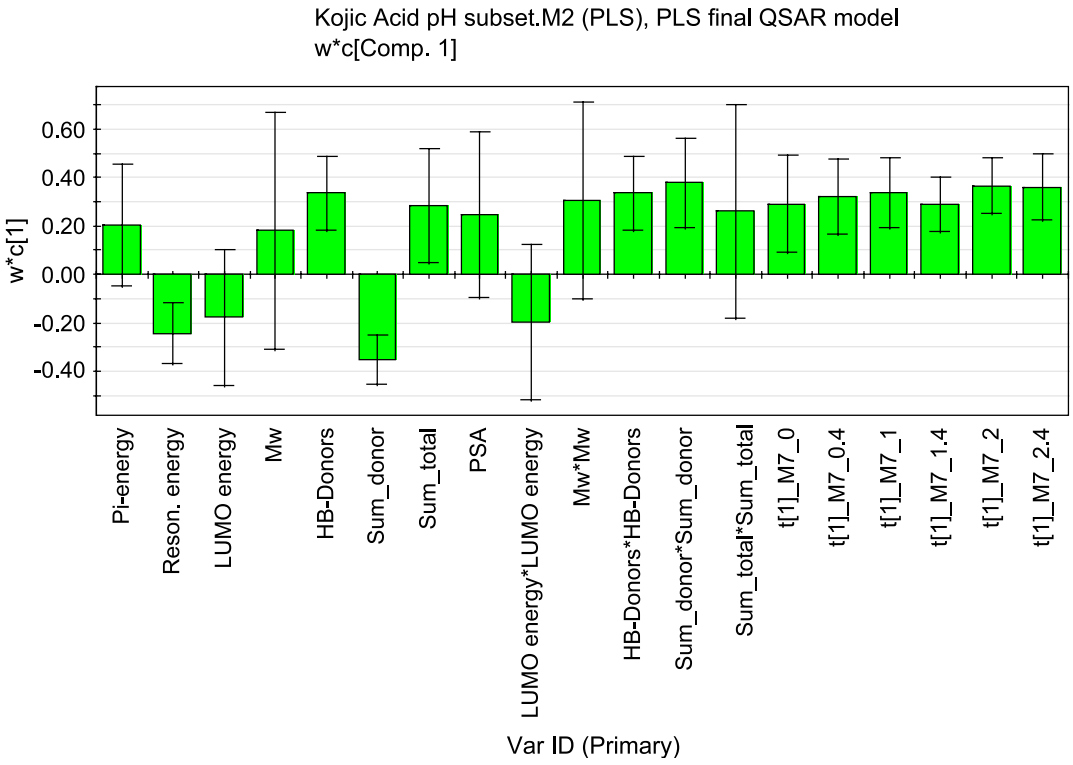

Fig. 15. PLS loadings of the upper level QSAR model (M6) based on the pH subset. For a description of the *X*-variables, please see text in Section 2.

assessment of linear and quadratic terms highlights that HB-Donors, Sum_donor, Sum_total, LUMO energy, and molecular weight are most appropriate to capture the pH-influenced change in toxicity levels. For instance, the property profile of many hydrogen bond donors, large polar surface area, and high molecular weight is associated with high toxicity, a relationship which is valid across all six pH-values.

6. Discussion

6.1. Three-way data structures

Three-way data structures are common and methods for their analysis are constantly evolving. Remote sensing applications, where, e.g., satellite images measured using multi-channel instruments are stacked together, quickly produce three-way data which can be explored using multivariate image analysis [17]. Similarly, multiple signals assembled over time on many batches in a manufacturing process can also be understood as a three-way data structure. Such data can be conveniently interrogated using the approach to BSPC of Wold et al. [5], or by related methods established by MacGregor et al. [18,19], or PARAFAC [20]. Yet another field in which three-way (and higher order) data structures are often encountered is analytical chemistry, a prime example being the use of various hyphenated techniques, like LC-MS, GC-MS, and 2D-NMR, on a series of analytical samples [20,21]. Interesting results have also been reported for three-way data within the "-omics" disciplines [22,23] using the approach presented in Ref. [5].

In the QSAR field, three-way data structures are not as abundant as in other areas of research, development and production. Also note that when mentioning "three-way" data in the present context, allusion is *not* made to 3D-QSAR applications, like COMFA [24] or GRID [25], which involve placement of molecules in a three-dimensional grid with the purpose of computing grid point interaction energies between the target molecules and various probes allowed to migrate around them.

Interestingly, however, one can easily envision QSAR situations where three-way data would be the norm rather than the exception. This would for instance relate to sets of compounds or chemical mixtures being chemically mapped by a multitude of chemical model systems, and where the parameters measured are also charted over time. Other applications of the QSAR-type might involve molecular characterization efforts using, e.g., the combination of liquid chromatography and mass spectrometry (LC-MS), or perhaps 2D-NMR.

Three-way *Y*-data is also rare in QSAR, a notable exception being the data set of Pirselova et al. [6,7]. Such data provide important information in QSAR. Much in the same way as bivariate data provide more information than univariate data, and multivariate data more information than bivariate data, three-way data structures amplify our knowl-

edge in comparison with two-way multivariate data tables. The reasons for this are several. First of all, a multivariate set of "two-way" measurements realized on a multitude of intelligently selected observations tend to stabilize the latent variables in a data set and make them appear more distinctly. Secondly, the richness of data in a three-way array will enhance our ability to reliably detect clusters, trends, outliers, aberrant measurements, and other anomalies. Thirdly, three-way data offer different ways of unfolding to produce two-way arrays of data, which will help out to elucidate the dominant patterns in a data set from different angles.

6.2. Benefits of the outlined approach

The outlined approach to dealing with three-way *Y*-data offers several interesting diagnostic and interpretative options. First and foremost, a score line plot of the architecture presented in Figs. 11a and 13a allows a rapid identification of compounds for which the toxic potency is susceptible to changes in external factors like time, temperature and pH. Not only is it possible to identify how toxic potency is modified, but the shape of the trajectory of each compound will perhaps also reveal information of mechanistic relevance. It is noted that a score line plot of this type is often used in modelling of time-influenced metabonomics data—where each line would correspond to an animal—to identify, e.g., slow responders or fast metabolizers [22,23]. This type of assessment is sometimes termed "metabolic trajectory deconvolution" [22], because "it allows effects of drugs and their metabolites to be separated for mechanistic purposes" [22].

Secondly, the lower level modelling provides an opportunity to focus on the individual observations (measurements) rather than the whole string of data for one and the same molecule. This is an advantage since it facilitates better understanding of the nature of the data set, as well as "debugging" by means of outlier elimination.

Last, but certainly not least, an important consequence of this approach is that targeted QSAR modelling is possible. Conventional QSAR analysis only allows a "static" look at the structure–activity relationship that prevails in a given data set in a given application. With the present approach, however, biological data can be regularized ("resolved") in terms of how they relate to external factors, such as time, temperature and pH. Other sources of variation in toxicity data that may influence the "static" look on the structure–activity relationships may then be peeled-off and temporarily disregarded.

Because of the strong associations between toxicity data and time (Model 2), and between toxicity data and pH (Model 5), we see that such time and pH-filtering is indeed warranted. The interpretation of the respective resulting upper level QSAR model (Models 3 and 6) will then point to which chemical properties correlate with the time- and pH-influenced changes in the structure–activity relationships (see Section 5).

6.3. Interpretation of QSAR models

One interpretation of the time-dependency would be that the cells adapt to the compounds, e.g., via conjugation to glutathion or glucoronic acid, or that the compounds undergo any alternative detoxification by enzymatic metabolism (see Fig. 10).

The pH-dependent behaviour has some obvious and natural correlates to the chemical properties of the compounds, since only the compounds having a pK_a close to the range of pH variation (i.e., phenolic pK_a between 6 and 8 resulting in anions upon dissociation) in the experiments showed decreased toxicity with increased pH.

There appears to be at least two plausible explanations to such behaviour, first and most obvious would be that the dissociated form of the acidic compounds is less toxic than the neutral form, or, secondly, that the dissociated form may be eliminated more efficiently than the neutral species. Of course combinations of the two or parallel interactions might add complexity to the picture, e.g., that the acidic function is toxic per se and thus contributes toxicity in an additive or synergistic fashion in comparison to a similar but aprotic compound.

Applying a holistic perspective on the pH-dependent trajectories in Fig. 13 would perhaps suggest that the latter explanation ("elimination") is the most reasonable. This because also aprotic compounds appear toxic with relatively uniform scores, i.e., t_1 values between -3.5 and 1, while those revealing pH-dependence approach t_1 scores of 4 in experiments using the highest pH. Some exception to this pattern constitutes, e.g., compounds 1 and 5, which both display a pK_a around 8, which is in the high end for the pH-dependent compounds, and some pH dependency was expected. However, these are also two of the least toxic compounds within the measured 4-pyranone series. Without any obvious correlation, oddities in their chemical structure and simple explanations like lack of assay resolution might be indicated. Compounds 12 and 35 also appear in the low toxicity range; however, they are the only two molecules comprising an aliphatic ester functionality and thus might have been rapidly hydrolysed by esterases (as corroborated by the time dependency in Fig. 10) with pursuing detoxification via elimination.

In conclusion, both the time and pH dependencies, analysed in the present study, appear to have possible and reasonable explanations. However, both toxicity and metabolism are complex in their nature, and structure–activity linearity is often limited to relatively narrow structural classes, due to, e.g., altered metabolism routes, as indicated by compounds 12 and 35. Thus, the above interpretations should be seen as hypotheses and further studies would be needed to fully understand the molecular mechanism of the 4-pyranone's toxicity.

6.4. Concluding remarks and future outlook

We have here outlined an approach that simplifies the interpretation of complex three-way *Y*-data in QSAR. We see this methodology as a promising pinpointing tool with which compounds that have time- or pH-sensitive toxicological profiles are rapidly identified. Using this approach it is possible to overview the major trends in toxicity and to zoom-in onto more detailed phenomena. For instance, the M1 and M4 lower level models manifested very clearly that many compounds under scrutiny have toxicity data that are sensitive to changes in time and pH. Hence, the QSAR analysis on the upper level could be directed towards these and thereby exploring which physico-chemical properties relate to time- and pH-induced changes in the toxicity data.

Three-way *Y*-data can be easily formed by acquiring toxicity data as a function of time and pH. In environmental QSAR applications also other external physical variables can be imagined as having a decisive impact on the biological response data. When investigating, e.g., soil sorption of environmental pollutants, or toxic potency of such chemicals to, say, microorganisms living in the soil, parameters like soil temperature and organic content will affect the response data being measured. Moreover, this methodology has applicability in the pharmaceutical industry, e.g., for following time-dependent changes in animals or patients subject to drug exposure.

Finally, we note that it is of relevance to further this study by comparing current results with results obtained when employing more conventional and "static" QSAR-analysis. In this context, the unfolding of the *Y*-data would result in a two-way array of toxicity data, having the compounds as rows and the 30 toxicity variables (5 time points*6 pH-values) as columns. Actually, this could be understood as using the approach to batch process analysis pioneered by MacGregor et al. [18,19], where three-way batch data are unfolded preserving the observation direction. The two-way **Y**-matrix could then be related to the battery of structure descriptors (*X*-variable) used in this paper. Such comparative studies are underway [26].

Acknowledgements

Figs. 3–7 were originally published in Ref. [14] and are used with permission.

References

[1] W.J. Dunn III, Quantitative structure–activity relationships (QSAR), Chemometrics and Intelligent Laboratory Systems 6 (1989) 181–190.
[2] L. Eriksson, J.L.M. Hermens, A multivariate approach to quantitative structure–activity relationships and structure–property relationships, in: J. Einax (Ed.), Chemometrics in Environmental Chemistry, The Handbook of Environmental Chemistry, vol. 5, Springer-Verlag, Berlin, Germany, 1995, pp. 135–168.
[3] M. Sjöström, Å. Lindgren, L.L. Uppgård, Joint multivariate quantitative structure–property and structure–activity relationships for a series of technical nonionic surfactants, in: G. Schüürmann, F. Chen (Eds.), Quantitative Structure–Activity Relationships in Environmental Sciences—VII, Proceedings of the 7th International Workshop on QSAR in Environmental Sciences, June 24–28, 1996, Elsinore, Denmark. SETAC Press, Pensacola, Florida, 1997, pp. 435–449.
[4] L. Eriksson, E. Johansson, F. Lindgren, M. Sjöström, S. Wold, Megavariate analysis of hierarchical biological data, Journal of Computer-Aided Molecular Design 16 (2002) 711–726.
[5] S. Wold, N. Kettaneh, H. Fridén, A. Holmberg, Modelling and diagnostics of batch processes and analogous kinetic experiments, Chemometrics and Intelligent Laboratory Systems 44 (1998) 331–340.
[6] K. Pirselova, S. Balaz, R. Ujhelyova, E. Sturdik, M. Veverka, M. Uher, J. Brtko, Quantitative structure–time–activity relationships (QSTAR): Part I. Growth inhibition of *Escherichia coli* by nonionizable Kojic acid derivatives, Quantitative Structure–Activity Relationships 15 (1996) 87–93.
[7] K. Pirselova, S. Balaz, E. Sturdik, R. Ujhelyova, M. Veverka, M. Uher, J. Brtko, Quantitative structure–time–activity relationships (QSTAR): Part II. Growth inhibition of *Escherichia coli* by ionizable and nonionizable Kojic acid derivatives, Quantitative Structure–Activity Relationships 16 (1997) 283–289.
[8] T.I. Oprea, J. Gottfries, Toward minimalistic modelling of oral drug absorption, J. Mol. Graph. Mod. 17 (1999) 261–274.
[9] T.I. Oprea, J. Gottfries, Chemography: the art of navigating in chemicals space, Journal of Combinatorial Chemistry 3 (2001) 157–166.
[10] T.I. Oprea, J. Gottfries, V. Sherbukhin, P. Svensson, T.C. Kühler, Chemical information management in drug discovery: optimizing the computational and combinatorial chemistry interfaces, Journal of Molecular Graphics and Modelling 18 (2000) 512–524.
[11] O.A. Raevsky, V.S. Skvortsov, 3D hydrogen bond thermodynamics (HYBOT) potentials in molecular modelling, Journal of Computer-Aided Molecular Design 16 (2002) 1–10.
[12] SIMCA-P+, version 10, http://www.umetrics.com.
[13] J.E. Jackson, A User's Guide to Principal Components, Wiley, New York, 1991, ISBN 0-471-62267-2.
[14] L. Eriksson, E. Johansson, N. Kettaneh-Wold, S. Wold, Multi- and Megavariate Data Analysis—Principles and Applications, Umetrics, Umeå, Sweden, 2001, ISBN 91-973730-1-X.
[15] A. Höskuldsson, A combined theory for PCA and PLS, Journal of Chemometrics 9 (1995) 91–123.
[16] S. Wold, C. Albano, W.J. Dunn III, U. Edlund, K. Esbensen, P. Geladi, S. Hellberg, E. Johansson, W. Lindberg, M. Sjöström, Multivariate data analysis in chemistry, in: B.R. Kowalski (Ed.), Chemometrics—Mathematics and Statistics in Chemistry, D. Reidel Publishing, Dordrecht, The Netherlands, 1984, pp. 1–81.
[17] K. Esbensen, P. Geladi, Strategy of multivariate image analysis (MIA), Chemometrics and Intelligent Laboratory Systems 7 (1989) 67–86.
[18] J.F. MacGregor, P. Nomikos, Monitoring Batch Processes, NATO ASI for Batch Processing Systems, May 29–June 7, Antalya, Turkey.
[19] S. Rännar, J.F. MacGregor, S. Wold, Adaptive batch monitoring using hierarchical PCA, Chemometrics and Intelligent Laboratory Systems 41 (1998) 73–81.
[20] R. Bro, PARAFAC. Tutorial and applications, Chemometrics and Intelligent Laboratory Systems 38 (1997) 149–171.
[21] H.A.L. Kiers, Some procedures for displaying results from three-way methods, Journal of Chemometrics 14 (2000) 151–170.
[22] J.K. Nicholson, J. Connelly, J.C. Lindon, E. Holmes, Metabonomics: a platform for studying drug toxicity and gene function, Nature Reviews 1 (2002) 153–162.
[23] H. Antti, M.E. Bollard, T. Ebbels, H. Keun, J.C. Lindon, J.K. Nicholson, E. Holmes, Batch statistical processing of ^{1}H-NMR-derived urinary spectral data, Journal of Chemometrics 16 (2002) 461–468.
[24] G. Cruciani, K.A. Watson, Comparative molecular field analysis using GRID force-field and GOLPE variable selection methods in a

study of inhibitors of glycogen phosphorylase b, Journal of Medicinal Chemistry 37 (1994) 2589–2601.
[25] M. Cocchi, E. Johansson, Amino acids characterization by GRID and multivariate data analysis, Quantitative Structure–Activity Relationships 12 (1993) 1–8.
[26] L. Eriksson, J. Gottfries, E. Johansson, S. Wold, Conventional and time-resolved QSAR approaches for three-way biological data, A comparative study, Manuscript in preparation.
[27] S. Wold, Cross-validatory estimation of the number of components in factor and principal components models, Technometrics 20 (1978) 397–405.