SAR and QSAR in Environmental Research,
Vol. 15 (5–6), October–December 2004, pp. 481–499

Taylor & Francis
Taylor & Francis Group

THREE-BLOCK BI-FOCAL PLS (3BIF-PLS) AND ITS APPLICATION IN QSAR*

L. ERIKSSON[a,†], J. DAMBORSKY[b], M. EARLL[c], E. JOHANSSON[a], J. TRYGG[d] and S. WOLD[a,d]

[a]*Umetrics AB, POB 7960, SE-907 19, Umeå, Sweden;* [b]*National Centre for Biomolecular Research, Masaryk University, Kotlarska 2, 611 37 Brno, Czech Republic;* [c]*Umetrics UK Ltd, Woodside House, Woodside Road, Winkfield, Windsor SL4 2DX, UK;* [d]*Department of Chemistry, Umeå University, SE-90187 Umeå, Sweden.*

(Received 9 May 2004; In final form 20 June 2004)

When **X** and **Y** are multivariate, the two-block partial least squares (PLS) method is often used. In this paper, we outline an extension addressing a special case of the three-block (**X/Y/Z**) problem, where **Z** sits "under" **Y**. We have called this approach three-block bi-focal PLS (3BIF-PLS). It views the **X/Y** relationship as the dominant problem, and seeks to use the additional information in **Z** in order to improve the interpretation of the **Y**-part of the **X/Y** association. Two data sets are used to illustrate 3BIF-PLS. Example I relates to single point mutants of haloalkane dehalogenase from *Sphingomonas paucimobilis* UT26 and their ability to transform halogenated hydrocarbons, some of which are found as organic pollutants in soil. Example II deals with soil remediation capability of bacteria. Whole bacterial communities are monitored over time using "DNA-fingerprinting" technology to see how pollution affects population composition. Since the data sets are large, hierarchical multivariate modelling is invoked to compress data prior to 3BIF-PLS analysis. It is concluded that the 3BIF-PLS approach works well. The paper contains a discussion of pros and cons of the method, and hints at further developmental opportunities.

Keywords: QSAR; PLS; Two-block PLS; Three-block PLS; Hierarchical modelling

INTRODUCTION

In traditional QSAR analysis, it is a common practice to analyse data with multiple linear regression (MLR) [1]. MLR is used to uncover the relationship between chemical descriptors in a data table **X** and one biological response variable, **Y**. The investigated problems are either one (descriptor)-to-one (response) or few-to-one in nature. Nowadays, however, QSAR data are increasingly stored in large matrices **X** and **Y**, which may even contain more columns than rows (i.e., they are short and wide). The QSAR problems therefore involve many-to-many relationships. The investigation of such complex QSAR relationships requires powerful data analytical tools, which can extract and display the systematic structure in the data.

*Presented at the 11th International Workshop on Quantitative Structure–Activity Relationships in the Human Health and Environmental Sciences (QSAR2004), 9–13 May 2004, Liverpool, England.

†Corresponding author. E-mail: lennart.eriksson@umetrics.com

ISSN 1062-936X print/ISSN 1029-046X online © 2004 Taylor & Francis Ltd
DOI: 10.1080/10629360412331297452

When the **X**- and **Y**-matrices contain inter-related columns ("correlated variables"), the method of projections to latent structures by means of partial least squares (PLS) is a viable alternative [2]. PLS takes advantage of the multicollinearity among variables and uses this to stabilise the QSAR modelling. PLS has been used in a number of QSAR investigations both in drug design [2–9] and environmental sciences [10–16]. The method exists in a number of extended formats including implementations for non-linear modelling [17–22], discriminant analysis [23], hierarchical analysis [24,25], and resolution of three-way toxicity data structures [26].

Here we describe yet another modification of the basic two-block (**X/Y**) PLS method, namely a special variant of the three-block (**X/Y/Z**) extension (Fig. 1). We have called this variant three-block bi-focal PLS (3BIF-PLS). The objective of this contribution was to discuss 3BIF-PLS as applied to QSAR. In this context, we view the relationship between the multivariate matrix **X** and **Y** as the dominant problem. The additional data—available in terms of the multivariate matrix **Z**—may enhance the understanding of the **X/Y** relationship. Formally, this implies that **X** contains information about the *columns* in **Y**, and **Z** contains information about the *rows* in **Y** (Fig. 1).

It should be noted that the present three-block data structure can be analysed using the conventional two-block PLS approach, i.e. the **Y**-matrix can be investigated from two "directions", "horizontally" ($\mathbf{X} \Rightarrow \mathbf{Y}$) and "vertically" ($\mathbf{Z}' \Rightarrow \mathbf{Y}'$). Although 3BIF-PLS modelling might be the overall goal, the dual two-block models are still warranted as they can be used both for pre-processing and to gain an initial feeling of the complexity of the problem at-hand.

The utility of 3BIF-PLS is demonstrated using two data sets. The first illustration relates to single point mutants of haloalkane dehalogenase from *Sphingomonas paucimobilis* UT26

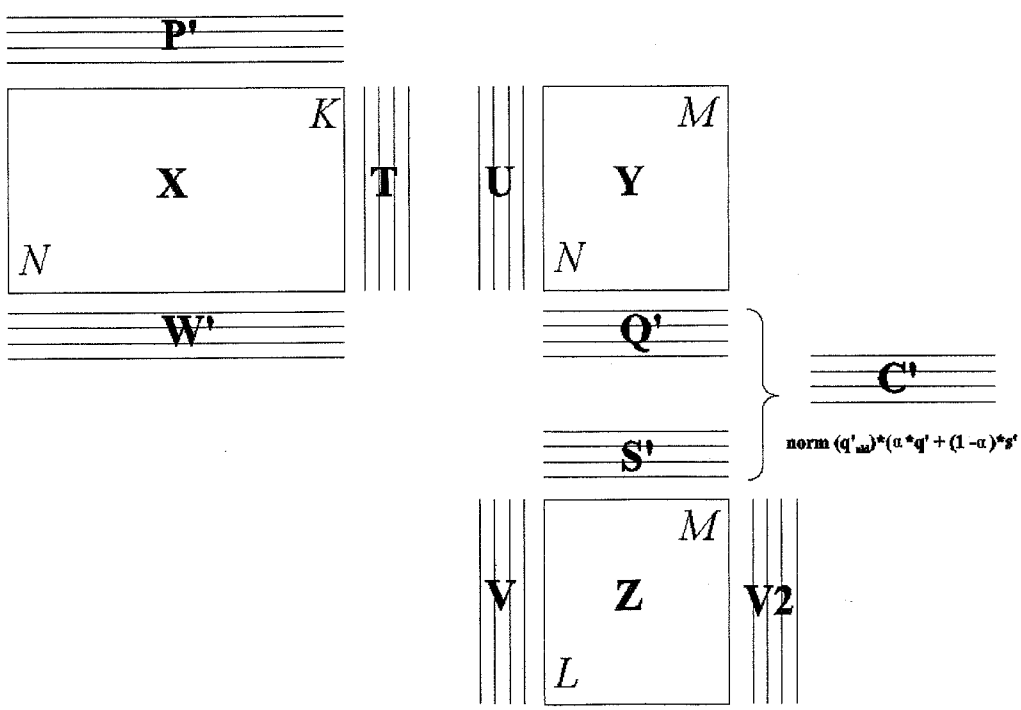

FIGURE 1 Schematic overview of the **X/Y/Z** three-block problem. We see the **X/Y** relationship as the dominant question. Information is extracted from **Z** that can enrich the interpretation of the association between **X** and **Y**. Data structure of example I (single point mutants) is: $N = 17$ mutants, $K = 402$ sequence descriptors (**X**), $M = 9$ disappearance rate constants (**Y**), and $L = 92$ chemical descriptors (**Z**). Data arrangement of example II (bacterial population) is: $N = 21$ observations (days), $K = 1018$ DNA-fingerprint variables (**X**), $M = 10$ concentration readings (**Y**), and $L = 1121$ chemical descriptors (**Z**).

and their ability to transform halogenated hydrocarbons, some of which are found as organic pollutants in soil. The second data set deals with soil remediation capability of bacteria. Whole bacterial communities are monitored over time using "DNA-fingerprinting" technology to see how pollution affects population composition.

EXAMPLE DATA SETS

Data Set I: Single Point Mutants

For pollutants in soil, the haloalkane dehalogenase from *Sphingomonas paucimobilis* UT26 splices off a halogen and replaces it by an hydroxyl group [27]. Engineered enzymes may have increased catalytic activity and/or broadened specificity. Chaloupkova *et al.* changed the size and the shape of the entrance tunnel leading into the active site by modifying the surface amino acid residue (leucine) in position 177 [27]. Fifteen mutants were constructed and their ability to transform nine pollutants was investigated. The wild type was also measured twice giving a total of $N = 17$ observations. The following nine haloalkanes were tested [27]: (4) 1-chlorobutane; (6) 1-chlorohexane; (18) 1-bromobutane; (29) 1-iodobutane; (37) 1,2-dichloroethane; (47) 1,2-dibromoethane; (54) 1,3-diidopropane; (67) 1,2-dichloropropane; (80) 1,2,3-trichloropropane; (115) chlorocyclohexane; (117) bromocyclohexane; (209) 3-chloro-2-methylpropene.

To reach some understanding of this **X/Y/Z** problem (cf. Fig. 1), data were structured as follows: the **X**-matrix contains $K = 402$ amino acid property descriptors [28] for the $N = 17$ dehalogenases (15 mutants + two wild type runs), the **Y**-matrix contains $M = 9$ response variables with disappearance rate constants for nine haloalkanes [27], and the **Z**-matrix contains $L = 92$ physico-chemical descriptors of the $M = 9$ haloalkanes [27]. Thus, in principle, the horizontal relationship ($\mathbf{X} \rightarrow \mathbf{Y}$) is a quantitative sequence-function relationship (QSFR; although in just one position) problem, and the vertical relationship ($\mathbf{Z}' \rightarrow \mathbf{Y}'$) is a similar QSAR problem. More experimental and computational details are found in the original Refs. [27,28].

The **X** amino acid descriptors were used as compiled in Ref. [28], without any pre-selection of descriptor variables. Hence, it can be anticipated that some of the descriptors, such as beta-sheet (B) descriptors, might be less relevant in the current application. However, rather than going through an awkward and subjective variable pre-selection scheme, we decided to use hierarchical multivariate modelling (see discussion in next paragraph) to focus on the **X**-variables being most relevant for the **Y**-responses.

Data Set II: DNA-Fingerprinting of Bacterial Population

The second example relates to a soil remediation study. In polluted soils, organic compounds are metabolised and mineralised by means of microbial action. As part of an investigation of the mechanism of this microbial action, ten organic substances were added to a tank of artificial soil [29,30] and observed over time. For each day during the three week long investigation period in October 1998, the decay curves (concentrations) of these $M = 10$ compounds were measured using toluene extraction and gas chromatographic analysis [29,30]. These concentration readings form the $M = 10$ responses. DNA fingerprinting

methodology [29,30] was used to characterise the microbial biomass of the changing bacterial population at each sampling occasion.

Thus, there are $N = 21$ observations (days) with the **X**-matrix having $K = 1018$ variables (DNA fingerprint variables) and the **Y**-matrix having $M = 10$ variables (concentration readings) [29,30]. The following 10 halogenated hydrocarbons were examined: ethylbenzene, *o*-xylene, decalin (decahydro-naphthalene), naphthalene, tridecane, heptamethylnonane, tetradecane, phenol, *p*-cresol and pristane (2,6,10,14,-tetramethyl-pentadecane). We used the DRAGON software (www.disat.unimib.it/chm/dragon.htm)[‡] to parameterise the molecular properties of the organic pollutants. The **Z**-matrix comprises $L = 1121$ chemical descriptors for the $M = 10$ compounds.

DATA ANALYSIS METHODS

Two-Block PLS

In this paper we have used two-block (**X/Y**) PLS [2] as implemented in the software SIMCA-P, version 10.5[¶]. PLS summarises **X** and **Y** and models the relationship between them. The information in the predictor block (**X**) is summarised by the A **X**-scores, collected in the matrix **T**, and the corresponding variation in the response block (**Y**) is described by the A **Y**-scores, collected in the matrix **U**. These score matrices (**T** and **U**) express the relationships among the observations.

Basically, PLS maximises the covariance between **T** and **U** [31]. For each model dimension, a weight vector $\mathbf{w}'$, is computed, which reflects the partial contribution of each **X**-variable to the modelling of Y. The resulting $(A*K)$ **X**-weight matrix, **W**, hence reflects the structure in **X** that maximises the covariance between **T** and **U**. The corresponding matrix of **Y**-weights is designated **C** (sometimes also denoted as **Q**). Additionally, a matrix of **X**-loadings, **P**, is calculated in order to deflate **X** appropriately. This matrix expresses the covariance structure among the **X**-variables with regard to the score vector matrix **T**.

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

$$\mathbf{X} = \mathbf{TP}' + \mathbf{E};\ \mathbf{Y} = \mathbf{TC}' + \mathbf{F} \tag{1}$$

To estimate the number of PLS components, cross-validation with seven exclusion groups was used [32]. The data were pre-processed by means of column-wise mean-centring and scaling to unit variance, unless otherwise stated.

The two-block PLS method is here used as a pre-processing tool of the **X**- and $\mathbf{Z}'$-data of both data sets. Thus, the **X**-block of example I, containing the sequence property description of the single point mutants, was divided in six logical blocks according to the nature of the descriptor variables. Using two-block PLS each such smaller **X**-block was related to the **Y**-matrix. A few summary scores were extracted for each of the six smaller **X**-blocks. The extracted local **X**-score vectors were then united to form a new, compressed **X**-block containing only the most relevant sequence description. The **X**- and $\mathbf{Z}'$-matrices of both examples were pre-processed in this way (see further details in the "Results" sections).

[‡]www.disat.unimib.it/chm/Dragon.htm. Accessed 2004-03-24.
[¶]www.umetrics.com. Accessed 2004-03-24.

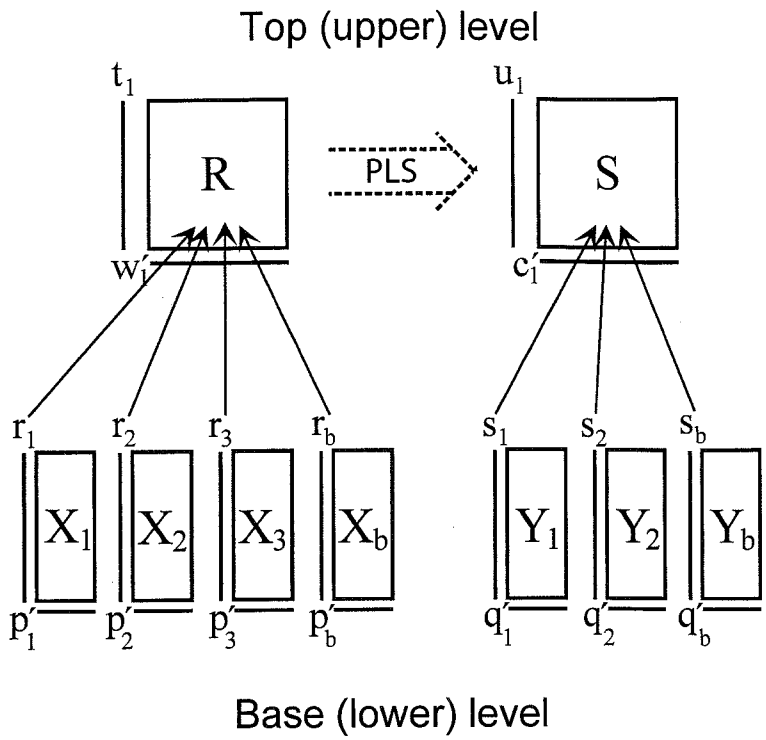

FIGURE 2 Schematic overview of hierarchical modelling. In the lower (base) level, each data block is locally modelled by either a PCA- or a PLS-model. Each block is summarised by one or more score vectors ("super variables"). Note that different numbers of "super variables" can be used for each block. The computed "super variables" are then merged together to new "**X**" and "**Y**" data matrices, here denoted **R** and **S**, respectively. All conventional multivariate statistics and diagnostics are retained within each level.

Conceptually, this means we are using a hierarchical modelling of the system [24,25,35,36]. The two-block PLS models form the lower level structure (Fig. 2). The resulting lower level score vectors are then concatenated on the upper level to form compressed versions of the **X**- and **Z**′-matrices, which are later analysed using the three-block approach.

Three-Block Bifocal PLS (3BIF-PLS)

In the two-block PLS framework, the **X**-matrix contains information about the *rows* of the **Y**-matrix. Finding such **X/Y** relationships is crucial in the analysis of environmental data. Here, we are concerned with the case when additional data are available, which can be used to "colour" the interpretation of the **X/Y**-relationship. This may occur in terms of a matrix **Z**, which contains information about the *columns* of **Y**. Hence, as seen from Fig. 1, the three matrices **X/Y/Z** altogether share no matrix size dimension, but are connected via the **Y**-matrix.

Interestingly, a PLS-based approach to the **X/Y/Z** problem was outlined as early as 1987 [35]. Although the initial algorithm experienced some convergence problems, it still provided encouraging results. Recently, another approach was published by Martens *et al.* [36]. Here, principal component analysis (PCA) was used to extract latent variables summarising the covariance matrix (**X′YZ**). The resulting scores and loadings were then used as loading vectors for the **X**- and **Z**-matrices, respectively, for subsequent prediction of the **Y**-matrix.

The approach presented here (3BIF-PLS) departs from the conventional two-block PLS algorithm, as reviewed in [2]. In the 3BIF-PLS approach, each model dimension comprises parameters similar to the two-block approach, notably an **X**-score vector **t**, a **Y**-score vector **u**,

an **X**-loading vector $\mathbf{p}'$, an **X**-weight vector $\mathbf{w}'$, and a **Y**-weight vector $\mathbf{c}'$ (sometimes also denoted $\mathbf{q}'$, Fig. 1). Additionally, the following "new" parameters are used, a **Z**-score vector **v** and a **Z**-weight vector $\mathbf{s}'$. In the deflation step of **Z**, the **Z**-score vector $\mathbf{v}_{\mathrm{TWO}}$ is utilised (Fig. 1). It expresses the covariance structure between the rows in **Z** with regard to the weight matrix $\mathbf{S}'$.

The decomposition in 3BIF-PLS of **X**,**Y** and **Z** can be described as:

$$\mathbf{X} = \mathbf{T}\mathbf{P}' + \mathbf{E};\ \mathbf{Y} = \mathbf{T}\mathbf{Q}' + \mathbf{F};\ \mathbf{Z} = \mathbf{V}_{\mathrm{TWO}}\mathbf{C}' + \mathbf{G} \tag{2}$$

The steps of the 3BIF-PLS algorithm are summarised in the "Appendix" section. As seen, the algorithm computes a weight vector $\mathbf{c}'$ as a weighted average of the **Y**-matrix weight vector $\mathbf{q}'$ and the **Z**-matrix weight vector $\mathbf{s}'$. The derivation of $\mathbf{c}'$ contains an adjustable parameter (a scalar) α, which will regulate the extent to which the variation in **Z** influences the **X**/**Y** projection. When $\alpha = 1$, **Z** is not allowed to influence the **X**/**Y** projection, and we thus arrive at the conventional two-block PLS **X**/**Y** solution. Conversely, the more α approaches 0, the more the variation in **Z** is allowed to enter the model of **X** to **Y**, i.e., a **Z**/**Y** PLS model. A more detailed account of the 3BIF-PLS algorithm and some related variants thereof, including different approaches to pre-processing and deflation procedures, is given in a separate communication [37].

Prior to the 3BIF-PLS calculations, data were pre-processed column-wise by means of mean-centring and scaling to unit variance, unless otherwise stated. All computations were carried out in SIMCA-P[¶] and MATLAB[§]. Cross-validation [32] was not used in the 3BIF-PLS computations. Results from related two-block PLS models were used to approximate the appropriate number of components in the 3BIF-PLS models.

RESULTS FOR DATA SET I (SINGLE POINT MUTANTS)

Two-Block "Horizontal" Modelling (X → Y)

The **X**-matrix was divided in six blocks:

- A (alpha helix) descriptors, $K_{\mathrm{A}} = 118$;
- B (beta sheet) descriptors, $K_{\mathrm{B}} = 37$;
- C (coil) descriptors, $K_{\mathrm{C}} = 24$;
- H (hydrophobicity) descriptors, $K_{\mathrm{H}} = 149$;
- O (other) descriptors, $K_{\mathrm{O}} = 28$;
- P (polarity) descriptors, $K_{\mathrm{P}} = 46$.

A detailed account of these descriptors is found in the original reference [28].

PLS was used to relate each sub-set of amino acid **X**-descriptors to the **Y**-block (with degradation rate constants of the nine haloalkanes). These six models (M1–M6) are summarised in Table I. We can see that alpha helix and hydrophobicity amino acid **X**-descriptors have the highest explained variances for the **Y**-responses. A total of 30 score vectors were extracted across the six base models. This assessment was not based on cross-validation performance. Instead, we required that the explained **X**-variance in each block model should exceed 80%

[§]www.mathworks.com. Accessed 2004-03-24.

TABLE I Overview of two-block PLS models

DataSet	*Model*	*Mode*	*Hier*	*A*	R^2X	R^2Y	Q^2Y	*Type*
Mutant	M1	X → Y	Base	8	0.82	0.92	0	A (alpha helix)
Mutant	M2	X → Y	Base	4	0.84	0.49	0	B (beta sheet)
Mutant	M3	X → Y	Base	3	0.89	0.4	0	C (coil)
Mutant	M4	X → Y	Base	6	0.82	0.77	0.22	H (hydrophobicity)
Mutant	M5	X → Y	Base	5	0.85	0.61	0.04	O (other)
Mutant	M6	X → Y	Base	4	0.86	0.51	0.05	P (polarity)
Mutant	M7	X → Y	Top	4	0.57	0.77	0.4	Hierarchical
Mutant	M8	Z′ → Y′	Base	2	0.96	0.5	0.18	Misc
Mutant	M9	Z′ → Y′	Base	3	0.83	0.59	0	TSAR
Mutant	M10	Z′ → Y′	Base	2	0.78	0.67	0.31	VAMP
Mutant	M11	Z′ → Y′	Base	5	0.89	0.94	0.35	VolSurf
Mutant	M12	Z′ → Y′	Base	2	0.84	0.46	0	MOPAC
Mutant	M13	Z′ → Y′	Top	4	0.87	0.87	0.21	Hierarchical
Population	M14	X → Y	Base	2	0.63	0.86	0.77	1–102
Population	M15	X → Y	Base	2	0.68	0.82	0.68	103–204
Population	M16	X → Y	Base	2	0.49	0.87	0.65	205–306
Population	M17	X → Y	Base	2	0.44	0.85	0.65	307–408
Population	M18	X → Y	Base	2	0.4	0.87	0.63	409–510
Population	M19	X → Y	Base	2	0.53	0.86	0.75	511–612
Population	M20	X → Y	Base	2	0.47	0.87	0.71	613–714
Population	M21	X → Y	Base	2	0.16	0.84	0.54	715–816
Population	M22	X → Y	Base	2	0.62	0.87	0.78	817–918
Population	M23	X → Y	Base	2	0.69	0.83	0.69	919–1018
Population	M24	X → Y	Top	3	0.89	0.91	0.82	Hierarchical
Population	M25	Z′ → Y′	Base	2	0.82	0.36	0.01	Constitutional
Population	M26	Z′ → Y′	Base	4	0.88	0.94	0.13	Topological
Population	M27	Z′ → Y′	Base	1	0.84	0.06	0.02	Molecular walk
Population	M28	Z′ → Y′	Base	1	0.85	0.14	0.08	BCUT
Population	M29	Z′ → Y′	Base	2	0.92	0.09	0	Galvez topol
Population	M30	Z′ → Y′	Base	4	0.82	0.87	0	2Dautocorrelation
Population	M31	Z′ → Y′	Base	1	0.97	0.08	0.04	Aromaticity indices
Population	M32	Z′ → Y′	Base	1	0.97	0.09	0.05	Randic
Population	M33	Z′ → Y′	Base	3	0.88	0.65	0	Geometrical
Population	M34	Z′ → Y′	Base	2	0.82	0.11	0.05	RDF
Population	M35	Z′ → Y′	Base	3	0.88	0.32	0.1	3D-MORSE
Population	M36	Z′ → Y′	Base	3	0.8	0.86	0.46	WHIM
Population	M37	Z′ → Y′	Base	3	0.8	0.89	0.45	Getaway
Population	M38	Z′ → Y′	Base	3	0.83	0.53	0	Functional groups
Population	M39	Z′ → Y′	Base	2	0.81	0.51	0	Atom-centered
Population	M40	Z′ → Y′	Base	2	0.85	0.52	0	Misc.
Population	M41	Z′ → Y′	Top	3	0.68	0.97	0.69	Hierarchical

Data Set refers to example I or II. *Model* refers to model number. *Mode* refers to "horizontal" or "vertical" direction for two-block PLS. *Hier* refers to whether the model is a base model or a top model in the hierarchical framework. *A* refers to the number of PLS-components. R^2X refers to the explained **X**-variation. R^2Y refers to the explained **Y**-variation. Q^2Y refers to the predicted **Y**-variation. *Type* refers to type of variables used, which is further described in the text.

($R^2X > 0.8$). This is because at this stage the main objective was to get a good approximation of the respective **X**-block. Predictive power is in focus on the top level.

We then calculated the top level PLS model based on the 30 lower level score vectors. The top level model contained 4 components, significant according to cross-validation. The performance statistics of this model are $R^2X = 0.57, R^2Y = 0.71$ and $Q^2Y = 0.40$ (Table I, model M7). It is noteworthy that the explained and predicted **Y**-variances are higher for the top level model than any of the six base level PLS models. The relatively low Q^2Y of 0.40 depends largely on two somewhat extreme mutants (177A and 177G), and when these are excluded in the cross-validation rounds their predictions are affected. This degrades Q^2Y. It is possible to increase the $Q^2Y(\approx 0.45)$ by using 18 rather than 30 score variables, but

since variable deletion might jeopardise other diagnostic procedures we refrained from doing so.

Two-Block "Vertical" Modelling ($\mathbf{Z}' \rightarrow \mathbf{Y}'$)

In the vertical two-block PLS modelling, the $\mathbf{Z}'$ matrix (here acting as the "**X**-matrix") was partitioned in five blocks:

- Miscellaneous descriptors, $L_{\text{Misc.}} = 7$;
- TSAR descriptors, $L_{\text{TSAR}} = 17$;
- VAMP descriptors, $L_{\text{VAMP}} = 7$;
- VolSurf descriptors, $L_{\text{VolSurf}} = 37$;
- MOPAC descriptors, $L_{\text{MOPAC}} = 24$.

A detailed account of these descriptors is found in the original Ref. [27].

As above, the complexity of the five local PLS models (M8–M12, Table I) was dictated by the requirement that a predictor block explained variance exceeded 0.8. This was fulfilled for all but one model, the VAMP model (M9), because this $\mathbf{Z}'$-block only contained seven predictor variables and an inclusion of a third PLS-component would bring the predictor block explained variance unrealistically close to unity. Hence, only two PLS components were computed for this model. As seen from Table I, these five models altogether resulted in 14 score vectors.

The ensuing top level PLS model (Table I, M13) had the following model statistics (using four PLS components): $R^2X = 0.87, R^2Y = 0.87$ and $Q^2Y = 0.21$. The low Q^2Y is in this case not so problematic, since in the three-block modelling we will extract from **Z** only what is of direct relevance to the **Y**-part of the **X/Y** association.

Three-Block Modelling ($\mathbf{X} \rightarrow \mathbf{Y} \leftarrow \mathbf{Z}$)

As a result of the initial two-block PLS modelling the sizes of the **X**- and **Z**-matrices are now changed (compare with Fig. 1). The size of **X** is now 17×30 (mutants $\times$ score vectors), the size of **Y** 17×9 (mutants $\times$ disappearance rate constants of substrates), and the size of **Z** 14×9 (score vectors $\times$ substrates).

As no cross-validation routine was invoked, 3BIF-PLS modelling complexity was based on the corresponding results of the two-block models. Since both the $\mathbf{X} \rightarrow \mathbf{Y}$ top level model (Table I, M7) and the $\mathbf{Z}' \rightarrow \mathbf{Y}'$ top level model (Table I, M13) comprised four components, we used four components also in 3BIF-PLS.

In total, we calculated 11 3BIF-PLS models, the results of which are tabulated in Table II (models MM1–MM11). These 11 models were obtained by allowing the α scalar (see "Appendix" section) to vary between unity and zero in steps of 0.1. And, as pointed out above, the use of $\alpha = 1$ provides us with the two-block PLS **X/Y** solution and hence the **X/Y** part of model MM1 (Table II) is identical to model M7 (Table I).

Table II shows that R^2X is little affected by the choice of α, and R^2Y also remains stable down to $\alpha = 0.8$. Below that level of α, however, the explained **Y**-variance drops rather quickly as α approaches zero. Not unexpectedly, the performance parameter being most sensitive to the setting of α is R^2Z and it varies between 0.41 (Table II, model MM1) and 0.88 (Table II, model MM11).

TABLE II Overview of three-block bi-focal PLS models

Data Set	*Model*	*Alpha*	*A*	R^2X	R^2Y	R^2Z
Mutant	MM1	1	4	0.57	0.77	0.41
Mutant	MM2	0.9	4	0.57	0.76	0.47
Mutant	MM3	0.8	4	0.57	0.74	0.52
Mutant	MM4	0.7	4	0.57	0.72	0.57
Mutant	MM5	0.6	4	0.57	0.69	0.62
Mutant	MM6	0.5	4	0.56	0.65	0.66
Mutant	MM7	0.4	4	0.56	0.61	0.69
Mutant	MM8	0.3	4	0.55	0.55	0.75
Mutant	MM9	0.2	4	0.53	0.43	0.82
Mutant	MM10	0.1	4	0.52	0.37	0.87
Mutant	MM11	0	4	0.51	0.34	0.88
Population	MM12	1	3	0.89	0.91	0.24
Population	MM13	0.9	3	0.89	0.9	0.31
Population	MM14	0.8	3	0.89	0.87	0.37
Population	MM15	0.7	3	0.89	0.84	0.43
Population	MM16	0.6	3	0.89	0.78	0.49
Population	MM17	0.5	3	0.89	0.71	0.54
Population	MM18	0.4	3	0.89	0.63	0.58
Population	MM19	0.3	3	0.89	0.53	0.62
Population	MM20	0.2	3	0.89	0.41	0.66
Population	MM21	0.1	3	0.88	0.28	0.69
Population	MM22	0	3	0.88	0.13	0.7

Data Set refers to example I or II. *Model* refers to model number. *Alpha* refers to the scalar used. *A* refers to the number of PLS-components. R^2X refers to the explained **X**-variation. R^2Y refers to the explained **Y**-variation. R^2Z refers to the explained **Z**-variation.

Interpretation of 3BIF-PLS Model Using $\alpha = 1$

In the following discussion we will be concentrating on models MM1 ($\alpha = 1$), MM6 ($\alpha = 0.5$), and MM11 ($\alpha = 0$). Primary attention will be paid to the first two components, as these are much larger than the last two. Figure 3(a)–(i) show scatter plots of the **X** scores **t**, the **X** weights **w**, the **Y** weights **c**, and the **Z** scores **v** of these three models.

The score plot in Fig. 3a shows the relationship among the observations (single point mutants) when $\alpha = 1$ and the solution is equivalent to the two-block PLS model. Observations (mutants) close to each other have similar dehalogenation profiles. Mutants L177W, L177F, L177M and L177V are most similar to the wild type and are grouped to the left in the score plot. These mutants contain a large and non-polar amino acid in position 177, and the wild type itself here carries a leucine.

Other interesting groupings of the mutants involve, e.g. the mutants in which the residue **177** is replaced by an amino acid with a charged side chain (H, K, R and D). These observations are predominantly located in the central and bottom left region. Also, mutants with small-sized amino acids like A and G, and mutants with small-sized OH-containing amino acids like T and S, are grouped pair-wise near each other.

Thus, inspection of Fig. 3a seems to indicate that the first 3BIF-PLS **t**-score vector separates mutants containing large and small amino acids in residue 177. The second score vector is related partly to side chain polarity by placing mutants containing polar and charged amino acids in the lower part of the score plot, and mutants with non-polar amino acids in the upper part.

A simultaneous inspection of the corresponding wc-weights (Fig. 3d) indicates that the compounds are grouped in smaller sub-groups. Generally, the most reactive compounds are

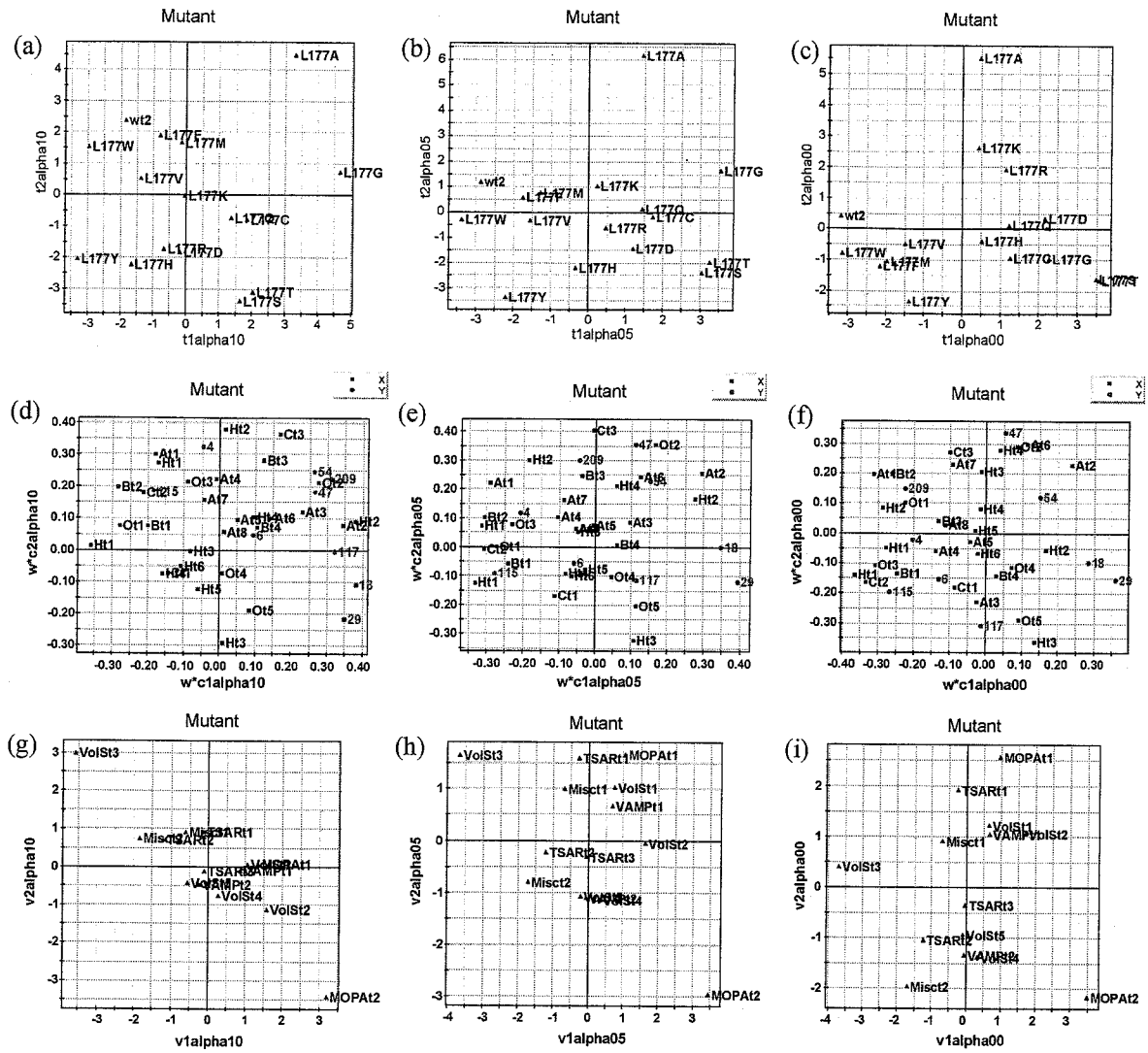

FIGURE 3 3BIF-PLS results of mutant example. Names of compounds and variables are found in the text. (a) Score plot t_1/t_2, $\alpha = 1$, (b) Same as in (a), $\alpha = 0.5$, (c) Same as in (a), $\alpha = 0$, (d) Weight plot w^*c_1/w^*c_2, $\alpha = 1$, (e) Same as in (d), $\alpha = 0.5$, (f) Same as in (d), $\alpha = 0$, (g) Score plot v_1/v_2, $\alpha = 1$, (h) Same as in (g), $\alpha = 0.5$, (i) Same as in (g), $\alpha = 0$.

found in the right-hand part of the weight plot and there they split in two sub-sets. One sub-set is made up of compounds **18**, **29**, and **117** and the other sub-set comprises compounds **47**, **54**, and **209**. The L117G mutant appears particularly apt at degrading these compounds. The first cluster contains mono-bromo and -iodinated compounds, and the second cluster di-bromo and di-iodocompounds together with a propene-derivative.

The third cluster is made up of the least reactive compounds **4**, **6**, and **115**, which are only chlorinated. The current experience reveals that transformation rates for brominated and iodinated compounds are clearly different from transformation rates of chlorinated ones, but propene-structures behave like compounds with bromine and iodine substituents under these and similar experimental circumstances [J. Damborsky *et al.*, unpublished results].

Among the amino acid **X**-descriptors, we can see that the first score vector of each of the six local models are highly correlated, (Fig. 3(d)–(f)). This is interesting and means that there is overlapping information between the different classes of **X**-descriptors. A more in-depth interpretation of the mutants sequence property descriptors is outside the main scope of this paper.

A more intriguing question is whether the **v**-scores (Fig. 3g) of the bi-focal modelling procedure are able to highlight some information in the **Z**-matrix of relevance for the **X/Y**

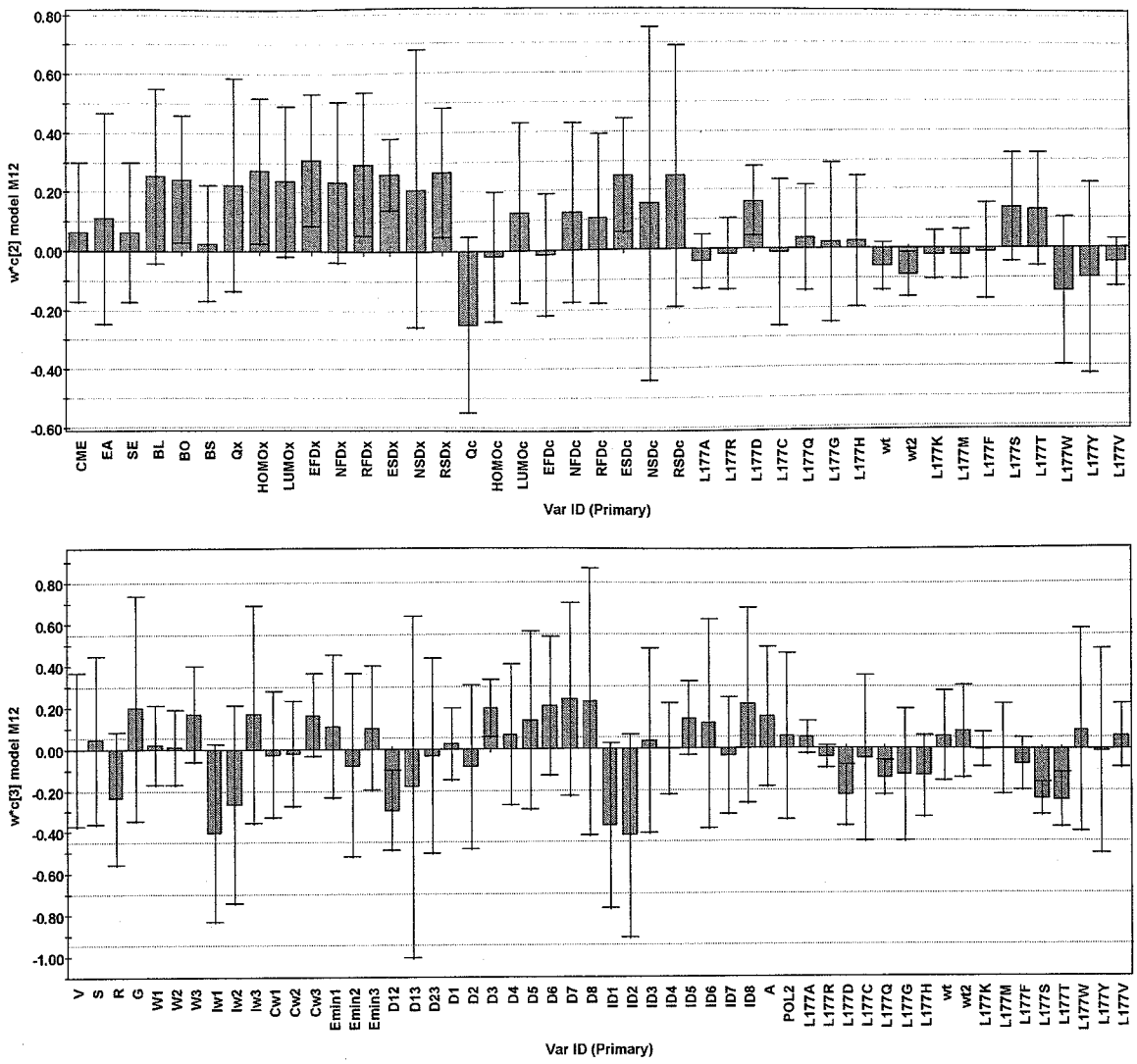

FIGURE 4 Focused interpretation of the score variables **(a)** MOPACt2 and **(b)** VolSurft3.

relationship. Figure 3(g) indicates that the lower level score variables MOPACt2 (t2 of model M12) and VolSurft3 (t3 of model M11) are most important. A numerically high value of MOPACt2 and low value of VolSurft3 reflect molecular properties that facilitate rapid transformation by the mutant proteins. Figure 4(a) and (b) provide the detailed interpretative basis of these two score variables.

With regards to the MOPCAt2 score variable, long bond length (BL) and high bond order (BO) of the C–X bond that undergo cleavage are favourable for fast transformation. This is consistent with the S_N2 reaction mechanism anticipated in this context. Additionally, other variables being important in the model, and correlating to the former two, are partial charge (Qx), HOMO density (HOMOx), LUMO density (LUMOx), electrophilic, nucleophilic, and radical frontier density (EFDx, NFDx, RDFx), and electrophilic, nucleophilic and radical superdelocalisability (ESDx, NSDx, RSDx), all on the leaving halogen atom. Also ESDc, RSDc and Qc of the carbon are important.

It is seen that for some of these descriptors, the confidence limits are quite wide. The confidence limits were estimated using cross-validation and jack-knifing [2]. Wide confidence limits point to variables whose model impact change a lot during the cross-validation and jack-knifing phase, and their interpretation should therefore be done with caution. Using many variables, although they might not be "statistically significant"

according to one criterion, will remedy this situation by stabilising the projection to latent structures. The reader is also reminded that we are looking into the second and third components of two base layer PLS models. It is quite common that confidence limits estimated via cross-validation and jack-knifing are narrower for variables dominating in the first PLS component.

There are fewer chemical descriptors that contribute to the VolSurft3 score variable, notably Iw1, Iw2, ID1, ID2, and D12. In turn and order these variables are integy moments 1 and 2, hydrophobic integy moments 1 and 2, and local interaction energy minimum distance 1–2. Low numerical values in these five variables encode fast reacting chemical species.

Integy moments reflect the imbalance between the centre of mass of a molecule and the position of the hydrophilic regions around it. If the integy moment is high, there is an abundance of hydrated regions in only one part of the molecular surface. If the integy moment is small, the polar moieties are either close to the centre of mass or they are at opposite "ends" of the molecule.

The Impact of the Changing α-Value

When examining the series of t_1/t_2 score plots obtainable for models MM1–MM11 (not all plots shown, cf. Fig. 3a–c) it appears that the changing value of α brings about a counter-clockwise rotation pattern of the mutants. The impact of L177A is concentrated to the second score vector with decreasing levels of α. The structure around the wild type is preserved and is little affected by the choice of α. The mutants being most sensitive to the value of the scalar are those which contain a charged amino acid (H, K, R, D, and Q).

The corresponding series of weight plots (plots only shown for α 1, 0.5, and 0, Fig. 3(d)–(f)) indicate a remarkable stability in the correlation structure among the nine haloalkanes. Hence, there is much information overlap between the three matrices (**X/Y/Z**). A close inspection of such plots suggests that compound 117 is most prone to migration at α-values slightly below unity. It moves from the high-reactive region (at $\alpha = 1$, Fig. 3d) to a position somewhere between slow and fast dehalogenation (at $\alpha = 0$, Fig. 3f).

Moreover, it is interesting to compare the structure of the v-scores in Fig. 3(g), (i). It is clear that the pattern detected at $\alpha = 1$ also dominates for $\alpha = 0$. However, the $\alpha = 0$ model seems to confirm that the VolSurft3 score variable is best able to differentiate between slow and fast metabolised compounds. The MOPACt2 score variable reflects this segregation too, but carries additional structure of relevance for the second model component.

RESULTS FOR DATA SET II (DNA-FINGERPRINTING OF BACTERIAL POPULATION)

Two-Block "Horizontal" Modelling (X → Y)

Similar to the previous example the **X**-matrix was divided in to sub-blocks, in this case 10 blocks. The spectral range was simply divided in 10 sub-sets. The numbers seen in plots (Fig. 5) refer to the variable numbers. More experimental details are found in Refs. [29,30].

Ten lower level PLS models were then calculated relating each of the **X** sub-blocks to the complete **Y**-matrix (Table I, models M14–M23). The related top level model was then based on the score variables of the lower level models (Table I, model 24).

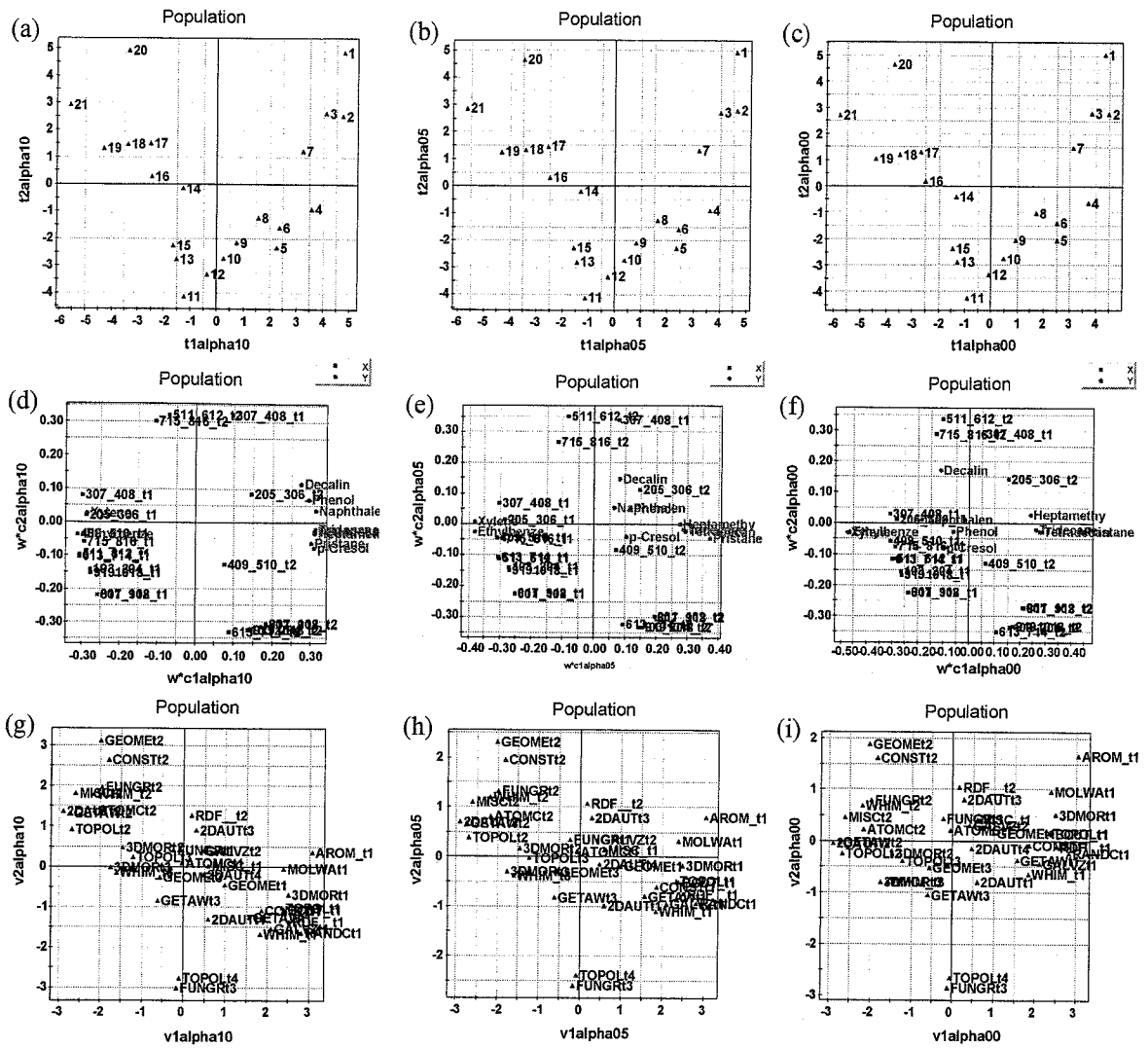

FIGURE 5 3BIF-PLS results of population example. Names of compounds and variables are found in the text, (a) Score plot t_1/t_2, $\alpha = 1$, (b) Same as in (a), $\alpha = 0.5$, (c) Same as in (a), $\alpha = 0$, (d) Weight plot w^*c_1/w^*c_2, $\alpha = 1$, (e) Same as in (d), $\alpha = 0.5$, (f) Same as in (d), $\alpha = 0$, (g) Score plot v_1/v_2, $\alpha = 1$, (h) Same as in (g), $\alpha = 0.5$, (i) Same as in (g), $\alpha = 0$.

Initially, we used the requirement that $R^2X > 0.8$, which gave four PLS components in all cases. However, the top level model revealed that only two PLS components from each lower level model were meaningful. Thus, the final statistics of M24 were: $R^2X = 0.89, R^2Y = 0.91, Q^2Y = 0.82$ using $A = 3$ components and 20 lower level score variables. Again, the predicted **Y**-variance is higher for the top level model than any of the ten individual lower level models.

Two-Block "Vertical" Modelling ($\mathbf{Z}' \rightarrow \mathbf{Y}'$)

The **Z**-matrix of example II is substantially larger than in example I and comprises 1121 physico-chemical descriptors (here: rows) in comparison with 92 in the first example. The 1121 DRAGON descriptors were divided in 16 blocks according to their nature and underlying generation principle:

- Constitutional (Const.) descriptors, $L_{\text{Const.}} = 27$;
- Topological (Topol.) descriptors, $L_{\text{Topol.}} = 213$;
- Molecular walk (Molwa.) descriptors, $L_{\text{Molwa.}} = 15$;

- BCUT descriptors, $L_{\text{BCUT}} = 64$;
- Galvez topological (Galvz.) descriptors, $L_{\text{Galvz.}} = 21$;
- 2D autocorrelation (2Daut.) descriptors, $L_{\text{2Daut.}} = 96$;
- Aromaticity indices (Arom.) descriptors, $L_{\text{Arom.}} = 4$;
- Randic (Randc.) descriptors, $L_{\text{Randc.}} = 41$;
- Geometricial (Geome.) descriptors, $L_{\text{Geome.}} = 22$;
- Radial distribution function (RDF) descriptors, $L_{\text{RDF}} = 150$;
- 3D-MORSE (3Dmor.) descriptors, $L_{\text{3DMor.}} = 160$;
- WHIM (Const.) descriptors, $L_{\text{WHIM}} = 99$;
- Getaway (Getaw.) descriptors, $L_{\text{Getaw.}} = 197$;
- Functional group (Fungr.) descriptors, $L_{\text{Fungr.}} = 10$;
- Atom-centered fragment (Atomc.) descriptors, $L_{\text{Atomc.}} = 10$;
- Miscellaneous (Misc.) descriptors, $L_{\text{Misc.}} = 27$.

More details around these descriptors are found in www.disat.unimib.it/chm/Dragon.htm. Accessed 2004-03-24.

The requirement for the explained predictor block variance (to exceed 80%) was maintained throughout the 16 lower level PLS models (Table I, models M25–M40). This resulted in a total of 37 lower level score variables to be used at the upper level. When computing the ensuing top level model (Table I, model M41), the performance statistics were: $R^2X = 0.68, R^2Y = 0.97, Q^2Y = 0.69$ with $A = 3$ components.

Three-Block Modelling ($\mathbf{X} \rightarrow \mathbf{Y} \leftarrow \mathbf{Z}$)

The two-block PLS analyses summarise the data as intermediate score variables as follows: the size of the regularised **X** is 21×20 (days × score vectors), the size of **Y** is 21×10 (days × concentration readings of pollutants), and the size of the regularised **Z** is 37×10 (score vectors × pollutants). A total of 11 3BIF-PLS models were derived by changing the scalar between 1 and 0, and using $A = 3$ components all the time (Table II, models MM12–MM22). Figure 5(a)–(i) display scatter plots of the key model parameters.

The results in Table II demonstrate a stable projection structure in the **X**-matrix as R^2X hardly changes. Apart from this insensitivity to the change of the α-value, the pattern is similar to the mutant data set, i.e. R^2Y decreases and R^2Z increases with decreasing value of α. The explanatory power of **Y** is not noticeably penalised until α falls below 0.7–0.8.

Interpretation of 3BIF-PLS Models

Figure 5(a)–(c) give the t_1/t_2 score plot for α equal to 1, 0.5 and 0. As seen, the correlation among adjacent observations and the time series trajectory do not change appreciably. There is a V-shaped pattern in the time series and some kind of mechanistic phase shift occurs in the middle of the sampling campaign (around day 11). It is encouraging to see that there are no distinct outliers. This indicates that there are no sudden, abrupt, and discontinuous changes in the composition of the microbial population.

In the matching **wc**-weight plots (Fig. 5(d)–(f)), we find, somewhat surprisingly, that eight concentration variables are grouped to the right and two to the left. This means that only for the eight species to the right are the concentrations decreasing over time. The amounts of xylene and ethylbenzene actually increase over time. This might indicate that, as the other

eight compounds are metabolised by microbial action, xylene and ethylbenzene are formed. It should be pointed out that there may be other factors, abiotic in nature (e.g., binding, solvation, adsorption, ...) which may influence the degradation pathways of these compounds [29,30]. The important finding is that the ten responses are highly correlated, and hence a single PLS-model for all *Y*-variables is warranted. There are no strong outliers in the concentration data.

Figure 5e shows that with a decreasing α-value, thus allowing more information from the physico-chemical descriptors (**Z**) to come into play, the eight correlated concentration readings in the right-hand area tend to divide into two sub-sets. Four responses start to migrate in a westerly direction along t_1. These are the concentration readings of decalin, naphthalene, *p*-cresol and phenol, which are all cyclical molecules and they move towards the ring-containing substances *o*-xylene and ethylbenzene. The concentration readings of the four non-cyclical compounds tridecane, tetradecane, heptamethylnonane, and pristane are stationary. In conclusion, it appears that elements from the physico-chemical description that mirror ring-formation is picked-up by 3BIF-PLS when lowering α.

The triplet of PLS weight plots, Fig. 5(d)–(f) highlights that concerning the fingerprint data, the first score vector of each fingerprint window, except window 307–408, are correlated internally and also relate to the discrimination between cyclical and non-cyclical pollutants. The second component on the other hand is much influenced by most of the $\mathbf{t}_2$-score variables and the $\mathbf{t}_1$-vector of the interval 307–408.

Finally, the v-score plots in Fig. 5(g)–(i) suggest that the first and second score vectors of approximately 10 of the 16 ($\mathbf{Z}' \rightarrow \mathbf{Y}'$) lower level models express the noted differential behaviour between compounds with or without ring-structure. We therefore conclude there is a high degree of overlap and redundancy among the 1121 DRAGON parameters. These descriptors (and summaries thereof) indicate that decalin is a special compound see Fig. 5(f). It is the only compound which initially decreases in the polluted soil, but from day 11–12 an onwards the percentage of decalin reaches a plateau and does not change anymore. This fact could actually be one determinant behind the phase change noted above in the middle part of the sampling period. The score variables TOPOLt4 (topological descriptors) and FUNGRt3 (functional group descriptors) dominate in the second component.

DISCUSSION

General Considerations

The objective of using two-block PLS is to find relations between two tables of data, e.g. in QSAR between chemical descriptors (**X**) and biological responses (**Y**). This approach works well when the training set is homogenous and the response profile is similar for all observations. From time to time extraneous information (**Z**) may be available which can be included to deepen the understanding of the **Y**-part of the **X/Y** relationship. The 3BIF-PLS procedure outlined in this paper addresses this situation, by seeing **X/Y** as the main question and **Z** as supplementary information.

We have here reported the application of 3BIF-PLS to two problems of environmental relevance, which both contain elements of structure–activity relationship modelling. In both cases the main issue is one of understanding bacterial ability to degrade environmental

pollutants, either in terms of single point mutants or whole bacterial populations. The structure–activity "direction" is primarily used for the understanding of what it takes for a molecule to become easily degraded.

When 3BIF-PLS is of Relevance and when it is not

The 3BIF-PLS procedure is of relevance when there is some diversity among the compound properties, and the **Y**-variables are not too correlated. It is preferable that there exists a latent structure in the **Y**-matrix comprising more than one latent variable. Otherwise it is likely that the latent variable dimensionality in **X** and in the extraneous **Z** will be higher than in Y, which may cause convergence problems and render the solution unstable. Thus, if the **Y**-matrix contains few and strongly correlated variables, we believe there is little gain in replacing conventional two-block PLS with 3BIF-PLS.

Moreover, it is desirable that the respective two-block relationships ($\mathbf{X} \rightarrow \mathbf{Y}$ and $\mathbf{Z}' \rightarrow \mathbf{Y}'$) are strong and of similar nature. One easy check of such model conformity, and, thus indirectly an indication of whether the 3BIF-PLS approach will be warranted, consists of computing the two two-block PLS solutions and then comparing them. Particularly, we think here of creating a scatter plot of the C-weights of the "horizontal" $\mathbf{X} \rightarrow \mathbf{Y}$ model and comparing it with a scatter plot of the T-scores of the "vertical" $\mathbf{Z}' \rightarrow \mathbf{Y}'$ model. If these two plots are similar, then attempting 3BIF-PLS is a logical continuation. We note that such plots (not provided) were initially created for both data sets, and results did indeed show this kind of plot similarity. The main structure between the various two-block PLS models were preserved.

Pre-Processing of Data

Like any other projection method (PCA, PCR, PLS, ...), the outcome of 3BIF-PLS is sensitive to scaling and centring. The problem formulation (**X/Y** main question) dictates the manner in which data were scaled and centred. Here, the matrices **X**, **Y**, and **Z** were centred and scaled column-wise. It should be observed that, due to the hierarchical modelling approach, also the rows of the compressed **Z**-matrix are centred since they originate from score vectors which are centred in their derivation.

Martens *et al.* [36] invoke a double-centring and double-scaling procedure of **Y** and discusses its advantages and disadvantagtes. That implies not only subtraction of the column and row means but also the subsequent addition of the grand average. We have not explored the effect of double-centring and double-scaling of **Y**, since we opted for a pre-processing procedure that would harmonise with our view that the **X**-to-**Y** problem is the main data analytical issue.

Another aspect is that of data-laundering prior to 3BIF-PLS modelling. Since it is not unlikely that both **X** and **Z** may store **Y**-orthogonal variation, that is, systematic structure which is not of relevance to **Y**, filtering of **X** and **Z** may facilitate model derivation, interpretation and use. One approach that can be used to pre-process **X** and **Z** in this context is O-PLS [38], the effect of which will be reported in the near future [37].

Method Extensions

Regarding further extensions of the outlined approach, one point concerns the prudent use of cross-validation methodology. In the current work, we let the dimensionality of

the two-block PLS models guide the complexity of the 3BIF-PLS models. We thus used four components in the first data set and three components in the second data set.

One may also consider modifications of the algorithm presented (see "Appendix"). The deflation step (Step 14) is of particular relevance, and some alterations are conceivable. Such studies are underway and will be dealt with in a separate publication [37].

Additionally, the modelling infrastructure around 3BIF-PLS opens up possibilities for a 4BIF-PLS counterpart. That extension would correspond to a situation when a fourth matrix of data is available and which contains a description about the columns of **X**.

CONCLUDING REMARKS AND FUTURE OUTLOOK

We have here discussed a special variant of the multivariate three-block (**X/Y/Z**) problem. The scalar (α), which is an adjustable parameter, ranging between 0 and 1, gives the user a way of regulating the extent to which the external matrix **Z** is allowed to influence the projection of $\mathbf{X} \rightarrow \mathbf{Y}$. Our experience so far seems to suggest that $\alpha \approx 0.8$ provides a good trade-off between too much focus on the **X/Y** relation ($\leftrightarrow \alpha = 1$) and the risk of letting **Z** get too high an influence on the model ($\leftrightarrow \alpha$ approaching 0).

The presented methodology is currently being evaluated on other application data. An interesting application area concerns the development of new functional foods, where **X** represents chemical, physical and sensory data of the new products, **Y** contains the likings of hundreds of interrogated consumers, and **Z** provides background demographical data on the consumers.

Additionally, regarding the single point mutant data set, experimental work is currently going on to gather equivalent data for new proteins. This will present an excellent opportunity for external predictive verification of the proposed 3BIF-PLS models for this data set. We look forward to seeing such data with great interest.

Acknowledgements

We are indebted to Professor Roberto Todescini for allowing us to use and test the DRAGON software.

References

[1] Hermens, J.L.M. (1989) "Quantitative structure–activity relationships of environmental pollutants", In: Hutzinger, O., ed., Handbook of Environmental Chemistry Reactions and Processes (Springer-Verlag, Berlin), Vol. 2E, pp 111–162.

[2] Wold, S., Sjöström, M. and Eriksson, L. (2001) "PLS-regression: a basic tool of chemometrics", *Chemometr. Intell. Lab. Syst.* **58**, 109–130.

[3] Dunn, W.J., III (1989) "Quantitative structure–activity relationships (QSAR)", *Chemometr. Intell. Lab. Syst.* **6**, 181–190.

[4] Cocchi, M. and Johansson, E. (1993) "Amino acids characterization by GRID and multivariate data analysis", *Quant. Struct. Act. Rel.* **12**, 1–8.

[5] Hansson, L.O., Ennis, M.D. and Stjernlöf, P. (1997) "Quantitative structure–activity relationships in the 8-amino-6,7,8,9-tetrahydro-3H-benzeindole ring system. Analysis of serotonin 5-HT1A effects *in vivo* and *in vitro* via partial least squares regression", *Eur. J. Med. Chem.* **32**, 571–582.

[6] Sandberg, M., Eriksson, L., Jonsson, J., Sjöström, M. and Wold, S. (1998) "New chemical dimensions relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids", *J. Med. Chem.* **41**, 2481–2491.

[7] Damborsky, J. (1998) "Quantitative structure–function and structure–stability relationships of purposely modified proteins", *Protein Eng.* **11**, 21–30.

[8] Giraud, E., Luttmann, C., Lavelle, F., Riou, J.F., Mailliet, P. and Laoui, A. (2000) "Multivariate data analysis using D-optimal designs, partial least squares, and response surface modeling: a directional approach for the analysis of farnesyltransferase inhibitors", *J. Med. Chem.* **43**, 1807–1816.
[9] Eriksson, L., Arnhold, T., Beck, B., Fox, T., Johansson, E. and Kriegl, J. (2004) "Onion design and its application in a pharmaceutical QSAR problem", *J. Chemometr.*, **18**, 188–202.
[10] Miyashita, Y., Ohsako, H., Takayama, C. and Sasaki, S. (1992) "Multivariate structure–activity relationships analysis of fungicidal and herbicidal thiolcarbamates using partial least squares method", *Quant. Struct. -Act. Rel.* **11**, 17–22.
[11] Tysklind, M., Lundgren, K., Rappe, C., Eriksson, L., Jonsson, J., Sjöström, M. and Ahlborg, U.G. (1992) "Multivariate characterization and modeling of polychlorinated dibenzo-*p*-dioxins and dibenzofurans", *Environ. Sci. Technol.* **26**, 1023–1030.
[12] Lindgren, Å., Sjöström, M. and Wold, S. (1996) "QSAR modeling of the toxicity of some technical non-ionic surfactants towards fairy shrimps", *Quant. Struct. Act. Rel.* **15**, 208–218.
[13] Eriksson, L., Verboom, H. and Peijnenburg, W. (1996) "Multivariate QSAR modeling of the rate of reductive dehalogenation of haloalkanes", *J. Chemometr.* **10**, 483–492.
[14] Lindgren, F., Sjöström, M., Berglind, R. and Nyberg, B. (1996) "Modeling of the biological activity for a set of ceramic fibre materials: a QSAR study", *SAR QSAR Environ. Res.* **5**, 229–230.
[15] Andersson, P.L., van der Burght, A.S.A.M., van den Berg, M. and Tysklind, M. (2000) "Multivariate modeling of polychlorinated biphenyl-induced Cyp1A activity in hepatocytes from three different species: ranking scales and species differences", *Environ. Toxicol. Chem.* **19**, 1454–1463.
[16] Eriksson, L., Jaworska, J., Worth, A.P., Cronin, M.T.D., McDowell, R.M. and Gramatica, P. (2003) "Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs", *Environ. Health Perspect.* **111**, 1361–1375.
[17] Wold, S., Kettaneh-Wold, N. and Skagerberg, B. (1989) "Non-linear PLS modeling", *Chemometr. Intell. Lab. Syst.* **7**, 53–65.
[18] Wold, S. (1992) "Non-linear partial least squares modeling. II. Spline inner relation", *Chemometr. Intell. Lab. Syst.* **14**, 71–84.
[19] Qin, S.J. and McAvoy, T.J. (1992) "Non-linear PLS Modeling using neural networks", *Comput. Chem. Eng.* **16**, 379–391.
[20] Kimura, T., Miyashita, Y., Funatsu, K. and Sasaki, S. (1996) "Quantitative structure–activity relationships of the synthetic substrates for elastase enzyme using non-linear partial least squares regression", *J. Chem. Inf. Comput. Sci.* **36**, 185–189.
[21] Berglund, A. and Wold, S. (1997) "INLR, implicit non-linear latent variable regression", *J. Chemometr.* **11**, 141–156.
[22] Eriksson, L., Johansson, E., Lindgren, F. and Wold, S. (2000) "GIFI-PLS: modeling of non-linearities and discontinuities in QSAR", *Quant. Struct.-Act. Rel.* **19**, 345–355.
[23] Nouwen, J., Lindgren, F., Hansen, B., Karcher, W., Verhaar, H.J.M. and Hermens, J.L.M. (1997) "Classification of environmentally occurring chemicals using structural fragments and PLS discriminant analysis", *Environ. Sci. Technol.* **31**, 2313–2318.
[24] Eriksson, L., Johansson, E., Lindgren, F., Sjöström, M. and Wold, S. (2002) "Megavariate analysis of hierarchical QSAR data", *J. Comput.-Aided Mol. Des.* **16**, 711–726.
[25] Eriksson, L. and Earll, M. (2002) "Multi and Megavariate data analysis of hierarchical biological data", In: Ford, M., Livingstone, D., Dearden, J. and Van de Waterbeemd, H., eds, Euro QSAR 2002 Designing Drugs and Crop Protectants: Processes, Problems and Solutions (Blackwell Publishing), pp 276–280.
[26] Eriksson, L., Gottfries, J., Johansson, E. and Wold, S. (2004) "Time-resolved QSAR: an approach to PLS modeling of three-way biological Data", *Chemometr. Intell. Lab. Sys.*, in press.
[27] Chaloupkova, R., Sykorova, J., Prokop, Z., Jesenka, A., Monincova, M., Pavlova, M., Tusda, M., Nagata, Y. and Damborsky, J. (2003) "Modification of activity and specificity of haloalkane dehalogenase from *Sphingomonas paucimobilis* UT 26 by engineering of its entrance tunnel", *J. Biol. Chem.* **278**, 52622–52628.
[28] Nakai, K., Kidera, A. and Kaneshi, M. (1988) "Cluster analysis of amino acid indices for prediction of protein structure and function", *Protein Eng.* **2**, 93–100.
[29] Wikström, P.B., Andersson, A.C. and Forsman, M. (1999) "Biomonitoring complex microbial communities using random amplified polymorphic DNA and PCA", *FEMS—Microbiol. Ecol.* **28**, 131–139.
[30] Wikström, P.B., (2001) "Biomonitoring of complex microbial communities that biodegrade aromatics" Ph.D. Thesis (Umeå University, Umeå, Sweden).
[31] Trygg, J. (2001) "Parsimonious multivariate models", Ph.D. Thesis (Umeå University Umeå, Sweden).
[32] Wold, S. (1978) "Cross-validatory estimation of the number of components in factor and principal components models", *Technometrics* **20**, 397–405.
[33] Wold, S., Kettaneh, N. and Tjessem, K. (1996) "Hierarchical multiblock PLS and PC models for easier model interpretation and as an alternative to variable selection", *J. Chemometr.* **10**, 463–482.
[34] Janné, K., Pettersen, J., Lindberg, N.O. and Lundstedt, T. (2001) "Hierarchical principal component analysis (PCA) and projection to latent structure (PLS) technique on spectroscopic data as a data pretreatment for calibration", *J. Chemometr.* **15**, 203–213.
[35] Wold, S., Hellberg, S., Lundstedt, T., Sjöström, M. and Wold, H. (1987) PLS Modeling with latent variables in two or more dimensions. (Proceedings Frankfurt PLS-meeting, September, 1987) pp 1–21.

[36] Martens, H., Anderssen, E., Flatberg, A., Gidskehaug, L.H., Höy, M., Westad, F., Thybo, A. and Martens, M. (2004) "Regression of a data matrix on descriptors of both its rows and of its columns via latent variables: L-PLSR", *Comput. Stat. Data Anal.*, in press.
[37] Trygg, J., Eriksson, L., Johansson, E. and Wold, S. (2004) "PLS approaches to the bi-focal three-matrix (**X-Y-Z**) problem in chemometrics", Manuscript in preparation.
[38] Trygg, J. and Wold, S. (2002) "Orthogonal projections to latent structures", *J. Chemometr.* **16**, 119–128.

APPENDIX:

STEPS OF THE 3BIF-ALGORITHM

(1) $w' = u'\mathbf{X}/u'u$
(2) $\|w'\|$
(3) $\mathbf{t} = \mathbf{X}w$
(4) $q' = t'\mathbf{Y}/t't$
(5) $\|q'\|$
(6) $v = \mathbf{Z}q$
(7) $s' = v'\mathbf{Z}/v'v$
(8) $\mathbf{v} = \mathbf{v}^* \|s'\|$
(9) $\|s'\|$
(10) $c' = \|q'_{old}\|(\alpha q' + (1 - \alpha)s')$
(11) $u = \mathbf{Y}c/c'c$
(12) $p' = t'\mathbf{X}/t't$
(13) $\mathbf{v}2 = \mathbf{Z}s/s'\mathbf{s}$
(14) Deflation step: $E = \mathbf{X} - tp', F = \mathbf{Y} - tc', G = \mathbf{Z} - v2c'$