Multiway Calibration in 3D QSAR

Nilsson, J.; de Jong, S.; Smilde, A.K.

Published in:
Journal of Chemometrics

JOURNAL OF CHEMOMETRICS, VOL. 11, 511–524 (1997)

MULTIWAY CALIBRATION IN 3D QSAR

JONAS NILSSON,[1]* SIJMEN DE JONG[2] AND AGE K. SMILDE[3]
[1] *Department of Medicinal Chemistry, University Centre for Pharmacy, Antonius Deusinglaan 1, NL-9713 AV Groningen, Netherlands*
[2] *Unilever Research Laboratorium, PO Box 114, NL-3130 AC Vlaardingen, Netherlands*
[3] *Laboratory for Analytical Chemistry, University of Amsterdam, Nieuwe Achtergracht 166, NL-1018 WV Amsterdam, Netherlands*

SUMMARY

We have introduced multilinear PLS in 3D QSAR and applied it to GRID descriptors from a set of benzamides with affinity to the dopamine D_3 receptor subtype, synthesized as potential drugs against schizophrenia. The key issue in 3D QSAR modelling is to obtain a predictive model that is easy to interpret. Each component in the multilinear PLS model explains clearly defined details, e.g. substituent positions, while the bilinear PLS solution is general and more difficult to interpret. The best models were obtained after four components with multilinear PLS ($Q^2=51\%$) and after only one component with bilinear PLS ($Q^2=50\%$). The external test set was predicted better with multilinear PLS ($Q^2=31\%$) than with bilinear PLS ($Q^2=25\%$). With multilinear PLS one loses in fit and gains in stability and simplicity owing to the fewer parameters that need to be estimated as compared with bilinear PLS. Finally, multilinear PLS is also less influenced by insignificant variation in the descriptor block, which is an advantage in 3D QSAR modelling

Journal of Chemometrics, Vol. **11,** 511–524 (1997) (No. of Figures: 8 No. of Tables: 4 No. of References: 22)

KEY WORDS multilinear PLS; multiway calibration; leverage; 3D QSAR; PARAFAC

INTRODUCTION

Since Cramer *et al.*[1] presented the CoMFA (comparative molecular field analysis) procedure in 1988, it has frequently been used by medicinal[2–4] and environmental[5] chemists, as implemented in the SYBYL molecular modelling package.[6] Today, other similar approaches are available, e.g. the GRID[7] program in combination with GOLPE variable selection.[8] Rational drug design with 3D QSAR may be divided into three parts: alignment of the molecules, generation of the molecular fields and regression analysis with one or more biological activity parameters as the response.

First, low-energy conformations of the molecules are aligned by superimposition of mutual and possible interaction points. e.g. atoms on the molecules, with a target receptor protein. This is by far the most crucial step in order to achieve a reliable 3D QSAR model.

A molecular field is a three-dimensinal grid large enough to enclose all the aligned molecules, where at each grid point the interactions between a probe atom and each molecule are calculated. Thus the grid points in the grid are the variables.

Since multicollinearity among the descriptor variables may affect the regression analysis detrimentally, PLS[9] is traditionally used as the regression method in 3D QSAR. However, Bro[10] recently presented the multilinear PLS algorithm and demonstrated additional advantages as compared with bilinear PLS. Bro and Heimdal[11] showed that multilinear PLS is less influenced by noise, more stable, increases the predictive ability and improves the interpretation of the result as compared with

Correspondence to: J. Nilsson, Department of Medicinal Chemistry, University Centre for Pharmacy, Antonius Deusinglaan 1, NL-9713 AV Groningen, Netherlands

CCC 0886–9383/97/060511–14 $17.50

Received 15 November 1996
Accepted 28 April 1997

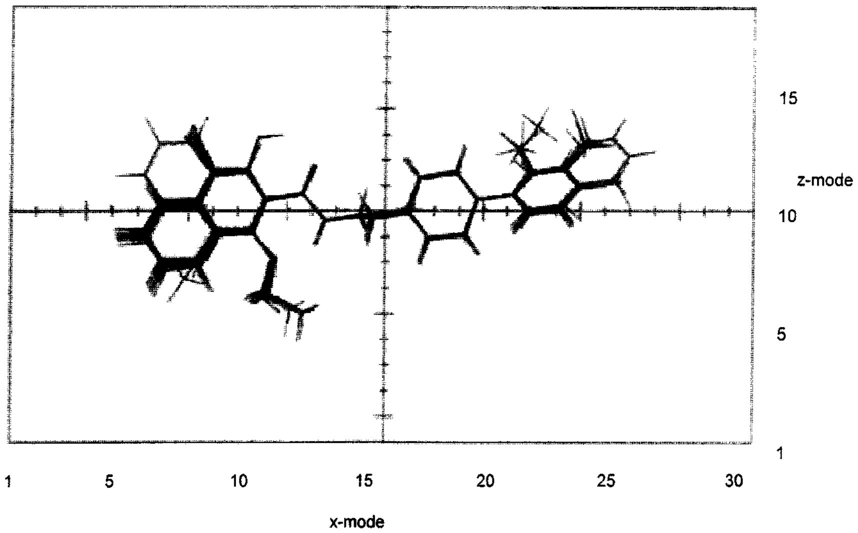

Figure 1. Thirty aligned training molecules, enclosed in grid, viewed in x- and z-mode

other methods applied to their data set. We have applied multilinear PLS regression analysis to GRID descriptors from a set of benzamides (Figure 1), synthesized as potential drugs against schizophrenia,[12] and compared the result with that obtained with bilinear PLS.

It is well known that redundant variables may affect the regression analysis detrimentally and in the literature several methods to reduce the number of variables[8, 13, 14] have been presented. We have used multilinear PLS for variable reduction of our data set and investigated whether the performance of the reduced model was improved as compared with the complete model.

Throughout this paper, lowercase characters represent scalars, boldface lowercase characters represent vectors, boldface uppercase characters represent matrices and underlined boldface uppercase characters represent multiway matrices.

THEORY AND METHODS

The multilinear PLS algorithm

The algorithm for multilinear PLS is actually an extension of traditional bilinear PLS. In PLS one wants to build a regression model between an independent (**X**) and a dependent (**y**) variable block. In bilinear PLS the **X**-block is a matrix where each row contains the variables measured for each object. Prior to regression analysis the **X**-block is decomposed into scores **t** (=**Xw**) and weights **w** (Figure 2(a)), where **w** is chosen such that **t** has the property of maximum covariance with **y**. Analogously, if $\underline{\mathbf{X}}$ is a threeway matrix ($I \times J \times K$) and **y** is a univariate ($I \times 1$), with typical elements x_{ijk} and y_i respectively, $\underline{\mathbf{X}}$ is decomposed into one score vector **t** ($I \times 1$) and two weight vectors $\mathbf{w}^J$ ($J \times 1$) and $\mathbf{w}^K$ ($K \times 1$), i.e. one vector per mode. The general idea, expressed in (1), is to find $\mathbf{w}^J$ and $\mathbf{w}^K$ so that the covariance between **t** and **y** is maximized:

$$\max_{\mathbf{w}^J\mathbf{w}^K}\left[\sum_{i=1}^{I} t_i y_i \,\middle|\, t_i = \sum_{j=1}^{J}\sum_{k=1}^{K} x_{ijk} w_j^J w_k^K \,\middle|\, \|\mathbf{w}^J\| = \|\mathbf{w}^K\| = 1\right] \tag{1}$$

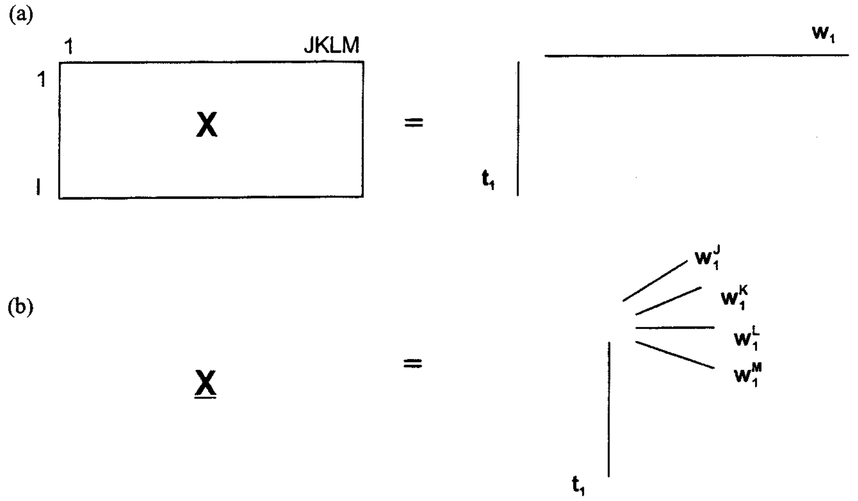

Figure 2. Graphical description of regression methods: (a) bilinear PLS; (b) multilinear PLS

Consequently, $\mathbf{t}$, $\mathbf{w}^J$ and $\mathbf{w}^K$ are a least squares approximation of $\underline{\mathbf{X}}$. However, the least squares property is valid for given $\mathbf{w}^J$ and $\mathbf{w}^K$ but not for general $\mathbf{w}^J$ and $\mathbf{w}^K$, hence (2) is a *partial* least squares model:

$$x_{ijk} = t_i w_j^J w_k^K + e_{ijk} \tag{2}$$

In order to solve $\mathbf{w}^J$ and $\mathbf{w}^K$, we rearrange (1) into

$$\max_{\mathbf{w}^J\mathbf{w}^K} \left[\sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} y_i \, x_{ijk} w_j^J \mathbf{w}_k^K \,\middle|\, \|\mathbf{w}^J\| = \|\mathbf{w}^K\| = 1 \right] \tag{3}$$

and since $\underline{\mathbf{X}}$ and $\mathbf{y}$ already are known, we can define $\mathbf{Z}$ $(J \times K)$ with typical element $z_{jk} = \Sigma\, y_i x_{ijk}$. Now we can rewrite (3) as

$$\max_{\mathbf{w}^J\mathbf{w}^K} \left[\sum_{j=1}^{J} \sum_{k=1}^{K} z_{jk} w_j^J w_k^K \,\middle|\, \|\mathbf{w}^J\| = \|\mathbf{w}^K\| = 1 \right] \tag{4}$$

and determine $\mathbf{w}^J$ and $\mathbf{w}^K$ by a singular value decomposition (SVD) of $\mathbf{Z}$ as in (5):

$$\max_{w^J\mathbf{w}^K} [(\mathbf{w}^J)^{\mathrm{T}} \mathbf{Z} \mathbf{w}^K] \quad \Rightarrow \quad (\mathbf{w}^J, \mathbf{w}^K) = \mathrm{SVD}(\mathbf{Z}) \tag{5}$$

Further, $\underline{\mathbf{X}}$ is updated by subtracting the contribution from the first component with typical element $\hat{x}_{ijk}^{(1)} = t_i w_j^J w_k^K$. This results in $\underline{\mathbf{E}}_1$ with typical element $e_{ijk} = x_{ijk} - \hat{x}_{ijk}^{(1)}$, where superscript (1) indicates the component subtracted. Accordingly, $\underline{\mathbf{E}}_1$ replaces x_{ijk} in (1) and the weights and scores from the next following component can be determined.

Since the scores from different components are not orthogonal, the regression coefficients $\mathbf{b}_A$ in (6) have to be calculated taking all the score vectors into account:

$$\mathbf{b}_A = (\mathbf{T}^{\mathrm{T}}\mathbf{T})^{-1}\mathbf{T}^{\mathrm{T}}\mathbf{y} \tag{6}$$

The score matrix **T** has the dimension $I \times A$, where the ath column represents the ath score vector.

In the case of 3D QSAR data, five different modes may be defined: the molecular mode, the grid x-direction, the grid y-direction, the grid z-direction and finally the probe mode (see Figure 3). Thus we have a data set in five modes with one dependent variable which requires a pentalinear PLS1 algorithm.

Analogously to the three-way problem, the five-way solution is obtained by finding the weight vectors $\mathbf{w}^J$, $\mathbf{w}^K$, $\mathbf{w}^L$ *and* $\mathbf{w}^M$ (Figure 2(b)). Since $\underline{\mathbf{X}}$ is of order higher than three, the solution cannot be accomplished by an SVD, but similarly the weight vectors are now obtained by a one-component PARAFAC[10, 15, 16] decomposition of $\underline{\mathbf{Z}}$ as in (7) with typical element $z_{jklm}=\sum_i y_i x_{ijklm}$:

$$\max_{\mathbf{w}^J\mathbf{w}^K\mathbf{w}^L\mathbf{w}^M}\left(\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{m=1}^{M} z_{jklm} w_j^J w_k^K w_l^L w_m^M\right) \tag{7}$$

The multilinear PLS algorithm discussed above has been thoroughly scrutinized by Bro[10] and Smilde.[17]

Partial PLS coefficients

For the purpose of interpretation the results from CoFMA studies are presented as contour plots of the partial regression coefficients $\mathbf{b}_{\mathrm{PLS}}$.[6] Basically, the coefficients $\mathbf{b}_{\mathrm{PLS}}$ are needed for predictions of new

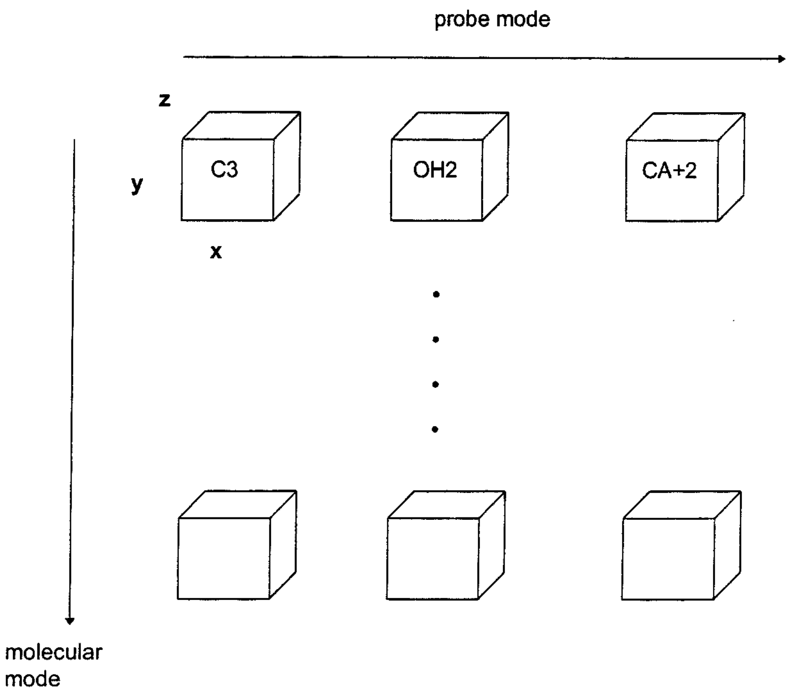

Figure 3. Complete data set defining five modes. The object mode is of 30 dimensions, the x-mode of 30 dimensions, the y-mode of 15 dimensions, the z-mode of 18 dimensions and the probe mode of three dimensions

samples, but since the size and sign of the coefficients reveal the relative importance of the variables, they are also suitable for the interpretation.

What we are looking for is a direct relationship between $\mathbf{X}^{(0)}$ and $\hat{\mathbf{y}}$ as in (8):

$$\hat{\mathbf{y}} = \mathbf{T}\mathbf{b}_A = \mathbf{X}^{(0)}\mathbf{b}_{\mathrm{PLS}} \tag{8}$$

where $\mathbf{X}^{(0)}$ $(I \times R)$ is the unfolded original $\underline{\mathbf{X}}$, $\hat{\mathbf{y}}$ $(I \times 1)$ is the fitted $\mathbf{y}$, $\mathbf{b}_A$ $(A \times 1)$ are the coefficients calculated as in (6) and $\mathbf{T}$ $(I \times A)$ is the scorfe matrix. The derivation of the full and closed predictions for multilinear PLS is presented by Smilde,[17] but since the PLS coefficients are frequently utilized in 3D QSAR, we find it essential to repeat the derivation also in this context.

Since the weights obtained with multilinear PLS are not orthogonal, we need to take this into account when we derive the $\mathbf{b}_{\mathrm{PLS}}$.

For clarity, $\mathbf{X}$ is updated after the ath component with $\mathbf{X}^{(a)} = \mathbf{X}^{(a-1)} - \mathbf{t}_a\mathbf{w}_a^{\mathrm{T}}$. If $\underline{\mathbf{X}}$ is three-way, $\mathbf{w}_a = \mathbf{w}_k^K \otimes \mathbf{w}_j^J$, where $\otimes$ is the Kronecker product. Then

$$\mathbf{t}_1 = \mathbf{X}^{(0)}\mathbf{w}_1 \tag{9}$$

$$\mathbf{t}_2 = \mathbf{X}^{(1)}\mathbf{w}_2 = (\mathbf{X}^{(0)} - \mathbf{t}_1\mathbf{w}_1^{\mathrm{T}})\mathbf{w}_2 = (\mathbf{X}^{(0)} - \mathbf{X}^{(0)}\mathbf{w}_1\mathbf{w}_1^{\mathrm{T}})\mathbf{w}_2 = \mathbf{X}^{(0)}(\mathbf{I} - \mathbf{w}_1\mathbf{w}_1^{\mathrm{T}})\mathbf{w}_2 \tag{10}$$

$$\cdots$$

$$\mathbf{t}_A = \mathbf{X}^{(0)}(\mathbf{I} - \mathbf{w}_1\mathbf{w}_1^{\mathrm{T}}) \cdots (\mathbf{I} - \mathbf{w}_{A-1}\mathbf{w}_{A-1}^{\mathrm{T}})\mathbf{w}_A \tag{11}$$

With $\mathbf{T} = (\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_A)$ the following holds:

$$\mathbf{T} = \mathbf{X}_{(0)}[\mathbf{w}_1 | (\mathbf{I} - \mathbf{w}_1\mathbf{w}_1^{\mathrm{T}})\mathbf{w}_2 | \ldots | (\mathbf{I} - \mathbf{w}_1\mathbf{w}_1^{\mathrm{T}})(\mathbf{I} - \mathbf{w}_2\mathbf{w}_2^{\mathrm{T}}) \ldots (\mathbf{I} - \mathbf{w}_{A-1}\mathbf{w}_{A-1}^{\mathrm{T}})\mathbf{w}_A] \tag{12}$$

Insertion of (12) in (8) followed by rearrangement gives

$$\mathbf{b}_{\mathrm{PLS}} = [\mathbf{w}_1 | (\mathbf{I} - \mathbf{w}_1\mathbf{w}_1^{\mathrm{T}})\mathbf{w}_2 | \ldots | (\mathbf{I} - \mathbf{w}_1\mathbf{w}_1^{\mathrm{T}})(\mathbf{I} - \mathbf{w}_2\mathbf{w}_2^{\mathrm{T}}) \ldots (\mathbf{I} - \mathbf{w}_{A-1}\mathbf{w}_{A-1}^{\mathrm{T}})\mathbf{w}_A]\mathbf{b}_A \tag{13}$$

When the number of variables is large, as in 3D QSAR, computing the outer product of the weight can be a problem. However, computational short-cuts are possible (see Appendix).

If $\mathbf{w}_i^{\mathrm{T}}\mathbf{w}_j = 0$ $(i \neq j)$, then (13) is reduced to

$$\mathbf{b}_{\mathrm{PLS}} = [\mathbf{w}_1 \;\; \mathbf{w}_2 \;\; \ldots \;\; \mathbf{w}_A]\mathbf{b}_A = \mathbf{W}\mathbf{b}_A \tag{14}$$

which resembles the solution obtained with Martens and Naes' non-orthogonalized PLS algorithm.[18]

Leverages

In order to determine which variables have affected the model most, we rank the variables by their leverages, determined by first calculating an overall weight matrix $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_a, \ldots, \mathbf{w}_A)$, in which $\mathbf{w}_a$ $(R \times 1;\ R = JKLM)$ combines the weights from the different modes as

$$\mathbf{w}_a = \mathbf{w}_a^M \otimes \mathbf{w}_a^L \otimes \mathbf{w}_a^K \otimes \mathbf{w}_a^J \quad (a = 1, \ldots, A) \tag{15}$$

The $\otimes$ sign represents the Kronecker product and a denotes the component number. The leverage[18] h $(R \times 1;\ R = JKLM)$ after A components is then expressed as

$$h = \mathrm{diag}(\mathbf{W}\mathbf{W}^{\mathrm{T}}) \tag{16}$$

A variable with a leverage h_T close to zero does not affect the model very much, while a variable with an h_r close to one is very important for the model. The average h_r is A/R and variables with leverage exceeding $h_{\mathrm{cut}} \times A/R$ (h_{cut} being an integer, normally two or three) may, according to Martens and Naes,[18] be considered significant.

Validation

The cross-validation (CV) experiments and the external predictions (Pred) are quantified with Q^2(CV) and Q^2(Pred) in (17), while the quality of the calibrations is given by R^2 in (18):

$$Q^2 = \left[1 - \left(\sum_{i=1}^{I} (y_i - \hat{y}_{(i)})^2 / \sum_{i=1}^{I} (y_i - \bar{y})^2\right)\right] \times 100 \tag{17}$$

$$R^2 = \left[1 - \left(\sum_{i=1}^{I} (y_i - \hat{y}_i)^2 / \sum_{i=1}^{I} (y_i - \bar{y})^2\right)\right] \times 100 \tag{18}$$

The predicted y in (17) is denoted $\hat{y}_{(i)}$, i.e. in the case of cross-validation an estimation of y_i using a model with the ith object excluded. In the case of external predictions, y_i is the response of the ith test object estimated with the complete calibration model. The fitted y from the calibration in (18) is denoted $\hat{y}_i$

THE DATA SET

The molecules considered in this paper were synthesized by Glase *et al.*[12] and later modelled with traditional 3D QSAR methods by Nilsson *et al.*[19] In addition to the 30 compounds modelled by Nilsson *et al.*,[19] we have introduced a test set consisting of 21 compounds for validation purposes.

Low-energy conformations of all the molecules were initially aligned as described by Nilsson *et al.*[19] and subsequently surrounded by a three-dimensional grid large enough to enclose all the aligned molecules with a border of 4 Å in all directions (Figure 1). The directions x, y and z in the grid were divided into 31, 15 and 18 steps of 1 Å each respectively, yielding a total of 8370 grid points. The surroundings of each molecule were mapped by calculating the interactions between a probe atom and each molecule at each grid point. The resulting grid, filled with interaction values, is called a molecular field. Different types of probe atoms yield different types of fields. We used three different probes,[7] a carbon atom (the C3 probe), a water molecule (the OH2 probe) and a plus two charged calcium ion (the CA+2 probe), reflecting the steric field, the hydrogen-bonding field and the electrostatic field respectively. In CoMFA one wants to correlate the differences in these fields with e.g. the affinities for a certain receptor subtype. The complete model is described graphically in Figure 3.

Prior to bilinear PLS analysis the data set is unfolded to form a two-way matrix which is decomposed into scores $\mathbf{t}$ $(I \times 1)$ and loadings $\mathbf{p}$ $(JKLM \times 1)$ as described in Figure 2(a). In multilinear PLS, however, the unfolding step is omitted and the one-component decompostion consists of a score vector $\mathbf{t}$ $(I \times 1)$ and four weight vectors $\mathbf{w}^J$ $(J \times 1)$, $\mathbf{w}^K$ $(K \times 1)$, $\mathbf{w}^L$ $(L \times 1)$ and $\mathbf{w}^M$ $(M \times 1)$ as in Figure 2(b). The vectors $\mathbf{t}$, $\mathbf{w}^J$, $\mathbf{w}^K$, $\mathbf{w}^L$ and $\mathbf{w}^M$ in Figure 2(b) correspond directly to the molecular mode, the grid x-direction, the grid y-direction, the grid z-direction and the probe mode respectively as described in Figure 3.

RESULTS

Model I

The only data preprocessing applied was mean-centering in the object (I) direction. In bilinear PLS, scaling is often performed column-wise, e.g. autoscaling,[9] whereas in multilinear PLS, scaling is not that straightforward.[16]

Table 1. Calibration and validation of Model I (30×25, 110) with multilinear PLS[a]

#LV	$R^2(\underline{\mathbf{X}})$	$R^2(\mathbf{y})$	Q^2(LOO)	Q^2(Pred)[b]
1	7	48	39	19
2	12	58	43	18
3	15	64	45	29
4	17	73	51	31
5	18	76	34	34

[a] All values in percentage.
[b] Predictions of external test set (21×25, 110).

The objective of this paper is to introduce multilinear PLS in 3D QSAR modelling and compare the solution with the traditional bilinear PLS solution. Accordingly, the complete model (Model I) was calibrated and validated with both regression methods, presented in Tables 1 and 2 respectively. With multilinear PLS (Table 1), maximum Q^2(CV) was obtained after four components ($Q^2=51\%$), where 17% of the variation in $\underline{\mathbf{X}}$ explained 73% of the variation in $\mathbf{y}$. With bilinear PLS, however, maximum Q^2(CV) was found after only one component ($Q^2=48\%$), where 22% of the variation in $\underline{\mathbf{X}}$ explained 62% of the variation in $\mathbf{y}$. The weights from the different modes obtained with multilinear PLS contain information useful for the interpretation of the result and in Figure 4 the weights from the first four components are plotted. For comparison the weight vector from the first component with bilinear PLS is plotted in Figure 5

The number of significant components was estimated by leave-one-out cross-validation and we found maximum Q^2(CV) after four components (Table 1) with multilinear PLS. In order not to lose information during the variable reduction step, we performed the variable reduction starting from a model one component more complex than optimal. Accordingly the absolute sum of the weights from the first five components was calculated for each mode separately. A position in a mode was considered significant and selected only if it exceeded a lower cut-off value. An arbitrary cut-off value of 0·2 generated a reduced data set with 6624 variables, called Model II. Stated differently, only variables with high weights from Model I were selected and included in Model II. The probe mode was left intact, hence variables from all three probes were included in the reduced data set.

Model II

The results from Model II are summarized in Tables 3 and 4 respectively. Model II was validated thoroughly (Table 4) with cross-validation and external predictions. In addition to traditional 'leave-

Table 2. Calibration and validation of Model I (30×25, 110) with bilinear PLS[a]

#LV	$R^2(\underline{\mathbf{X}})$	$R^2(\mathbf{y})$	Q^2(LOO)	Q^2(Pred)[b]
1	22	62	48	26
2	34	76	47	21
3	43	86	46	32
4	53	89	42	31
5	59	93	37	32

[a] All values in percentage.
[b] Predictions of external test set (21×25, 110).

one-out' cross-validation, also 'leave-three-out' and 'leave-five-out' cross-validations were performed, where in each experiment objects were left out randomly but only once. The results are reported as the average Q^2 of 20 cross-validation experiments.[20, 21]

In order to simplify the interpretations of a PLS Model in 3D QSAR, the partial PLS coefficients $\mathbf{b}_{\mathrm{PLS}}$ in (13) are often presented as comprehensive iso-contour plots. That is, each $\mathbf{b}_{\mathrm{PLS}}$ is transferred back to its original position in the grid, where grid points with similar coefficients are connected. In Figure 6 the $\mathbf{b}_{\mathrm{PLS}}$ contour is plotted in stereo from the C3 probe after the fourth multilinear PLS component.

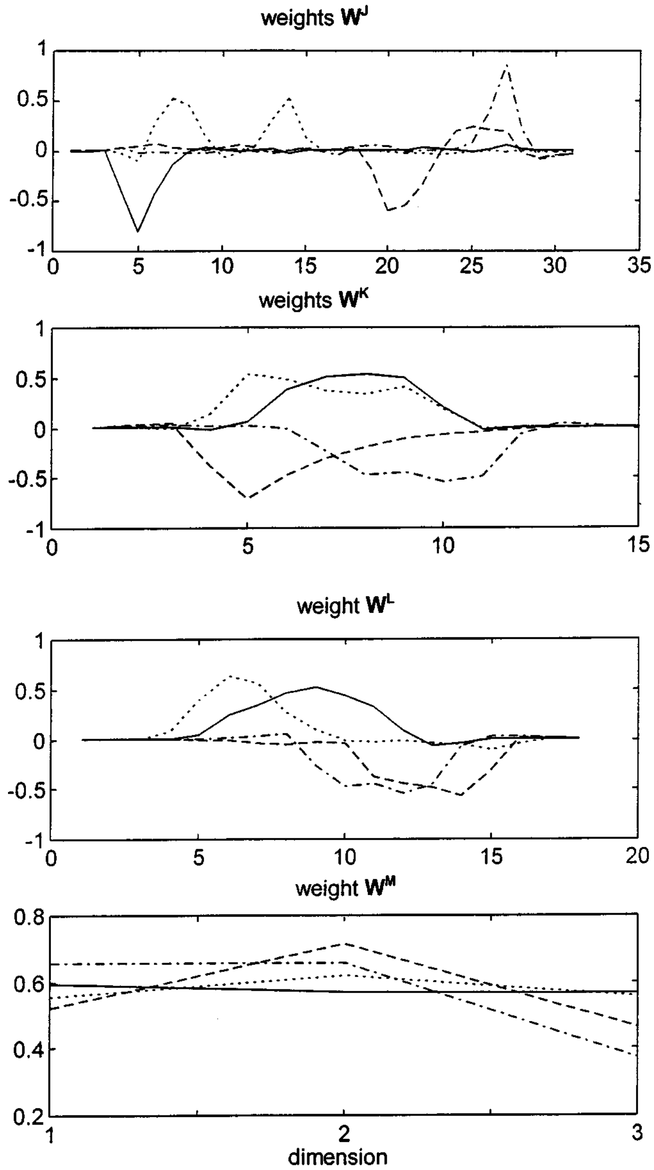

Figure 4. Weights $\mathbf{W}^J$, $\mathbf{W}^K$, $\mathbf{W}^L$ and $\mathbf{W}^M$: ——, component one; – – –, component two; – · –, component three; · · · ·,component four

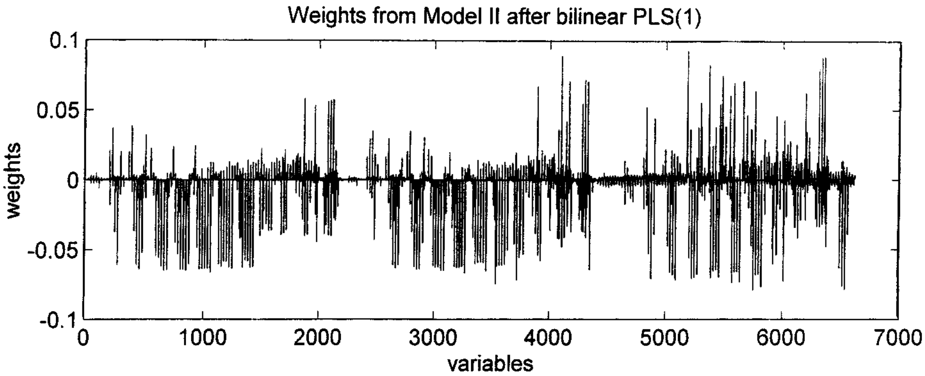

Figure 5. Weight vector after first component with bilinear PLS

Table 3. Cross-validation[a] and external predictions of Model II (30×6624)

	N-PLS[b]				PLS[b]			
@LV	LOO	L3O[c]	L5O[c]	Pred.[d]	LOO	L3O[c]	L5O[c]	Pred.[d]
1	39	43	42	19	50	50	50	25
2	43	44	41	18	48	46	47	23
3	45	43	41	29	48	46	48	33
4	51	53	49	31	44	41	42	30
5	43	42	38	34	39	37	40	31

[a] LOO is short for leave-one-out, L3O for leave-three-out and L5O for leave-five-out.
[b] All values in percentage.
[c] Average from 20 Q^2.
[d] Predictions of external test set (21×6624).

Table 4. Calibration of Model II with bilinear PLS and multiliner PLS for first five components

	N-PLS		PLS	
#LV	$R^2(\underline{\mathbf{X}})$	$R^2(\mathbf{y})$	$R^2(\underline{\mathbf{X}})$	$R^2(\mathbf{y})$
1	8	48	22	64
2	13	58	32	79
3	16	64	41	86
4	19	73	51	90
5	20	76	58	93

[a] All values in percentage.

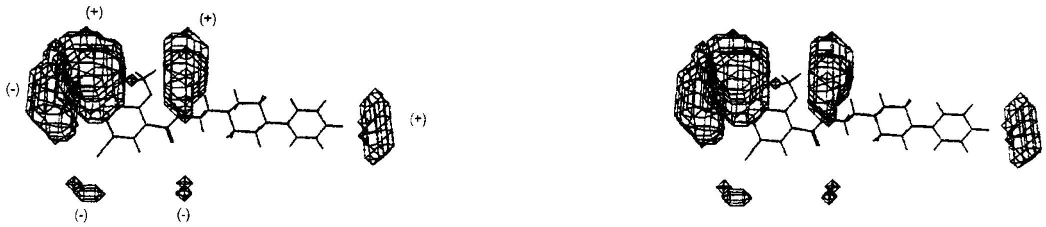

Figure 6. Coefficients $\mathbf{b}_{PLS}$ from final multilinear PLS model and C3 probe after four components

DISCUSSION

The key issue in 3D QSAR modelling is to find a predictive model which can be used as a tool in the design of new potent compounds. The solution should also be simple and straightforwrd, since also the non-expert must be able to interpret the model.

The initial complete model (Table 1) indicted four significant components with leave-one-out cross-validation. With help from Figure 4 we can determine, with good precision, the region accounted for by each component. The full curves in Figure 4 represent the weights from the first component, the broken curves the second component, the chain curves the third component and the dotted curves the fourth component. For ease of interpretation the weights $\mathbf{W}^J$ and $\mathbf{W}_L$ correspond to the x-mode and z-mode respectively in Figure 1. The first component has high weight $\mathbf{w}^J$ at position 5 and high weights $\mathbf{w}^L$ between positions 5 and 10 (Figure 4), which correspond to the region where the naphthalene moiety protrudes (Figure 1). Thus the first component explains the difference between naphthamides and benzamides. Similarly, it can be concluded that the second component mainly deals with the *ortho* and *meta* positions on the arylpiperazine phenyl ring, the third component the *para* position and, finally, the fourth component with substituents on the benzamide phenyl ring.

In contrast with the weights from multilinear PLS (Figure 4), the weights from bilinear PLS (Figure 5) are difficult to interpret.

Striking is the much less explained variance obtained with multilinear PLS (Table 1) as compared with bilinear PLS (Table 2). A speculative explanation for this is the fewer parameters that need to be estimated with multilinear PLS.[10, 15] Also, each component in multilinear PLS focuses on small specific items, e.g. regions in the grid, while bilinear PLS searches for more general directions for its components and is more flexible.

It is well known that many of the variables in a 3D QSAR model are more or less redundant and may affect the predictive ability detrimentally. From Figure 4 it is clear that positions corresponding to grid points in the periphery of the grid have low weights and also limited influence on the model. By omitting these variables, as described in the theory part, a reduced model with 6624 variables was obtained which was validated with cross-validation and external predictions. Variable selection must be performed very carefully, otherwise problems with overfitting may occur. Norinder[14] and Cho and Tropscha[13] reported increased cross-validated Q^2 but decreased ability to predict an external test set when the number of variables was reduced. We reduced the number of variables in our model from 25, 110 to 6624, which speeded up further calculations, but the predictive ability did not improve.

From Model II (Table 3) it can be concluded that our model is homogeneous and stable, since the cross-validated Q^2 was not affected very much when larger groups of molecules were left out each time. Each cross-validation experiment was repeated 20 times[20] and, accordingly, reported as the average Q^2.

In Figures 7(a) and 7(b) the experimental $\log_{10}(K_i)$ are plotted against the fitted $\log_{10}(K_i)$ from Model II for the training set with multilinear PLS and bilinear PLS respectively. The 21 test compounds have been predicted and plotted on the same figures as small circles. A four-component

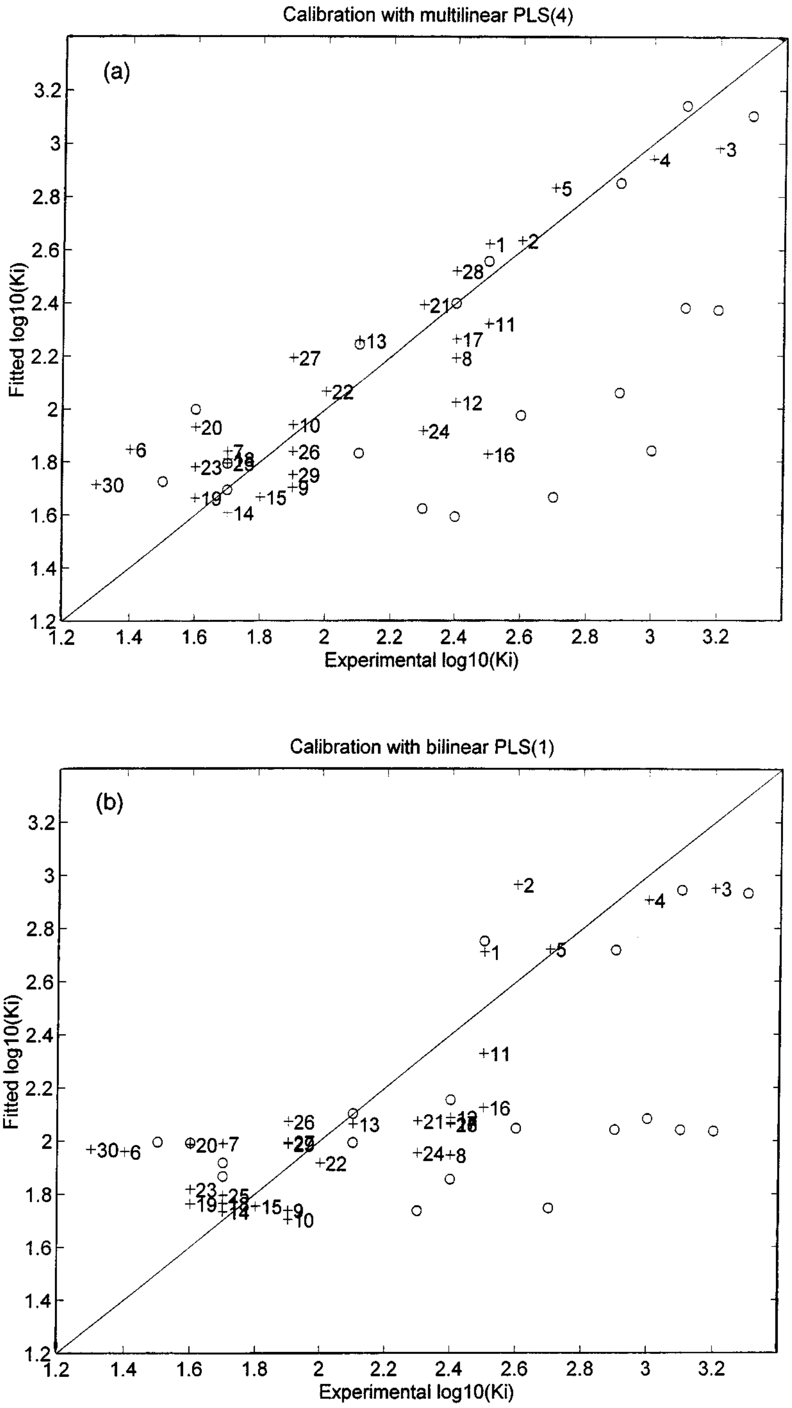

Figure 7. Experimental $\log_{10}(K_i)$ versus fitted $\log_{10}(K_i)$ after (a) fourth component with multilinear PLS and (b) first component with bilinear PLS

model with multilinear PLS ($R^2=73\%$) explains more of the variation in **y** as compared with a one-component bilinear PLS model ($R^2=64\%$). The test compounds were also better predicted with multilinear PLS ($Q^2=31\%$) than with bilinear PLS ($Q^2=25\%$). In fact, the bilinear PLS model (Figure 7(b)) more or less distinguishes between two groups of compounds, i.e. between benzamides and naphthamides, while the multilinear PLS model is much better fitted (Figure 7(a)).

The iso-contour plot of the coefficients b_{PLS} from the fourth component in Figure 6 is probably the most comprehensible tool for the interpretation of the model:

$$\text{Bio-Act}=x_1b_1+\cdots+x_ib_i+\cdots+x_kb_k+e \tag{19}$$

If a novel molecule is designed with a substituent protruding in a negative b_{PLS} region, then x_i in (19) will be positive and consequently x_ib_i will be negative. This substituent will thus have a negative effect on Bio-Act. If low Bio-Act is desirable, new substituents must be added in regions where b_{PLS} for the C3 probe (steric field) is negative and *vice versa*. For more specifics about how to interpret the iso-contour plots, we refer to the SYBYL manual[6] and Reference 19.

In Figure 8 we rank the 6624 variables by their respective leverage. Even after variable reduction a lot of variables with low influence on the model are present.

CONCLUSIONS

We have successfully introduced multilinear PLS as regression method in 3D QSAR. The main improvement lies in the interpretation of the result and the slightly better predictive ability as compared with bilinear PLS. The multilinear PLS model is also superior to bilinear PLS with regard to simplicity and stability, since fewer parameters need to be estimated.

The number of variables can effectively be reduced with help from multilinear PLS. Although the predictive ability did not improve, the speed of the calculations did. The number of high-leverage

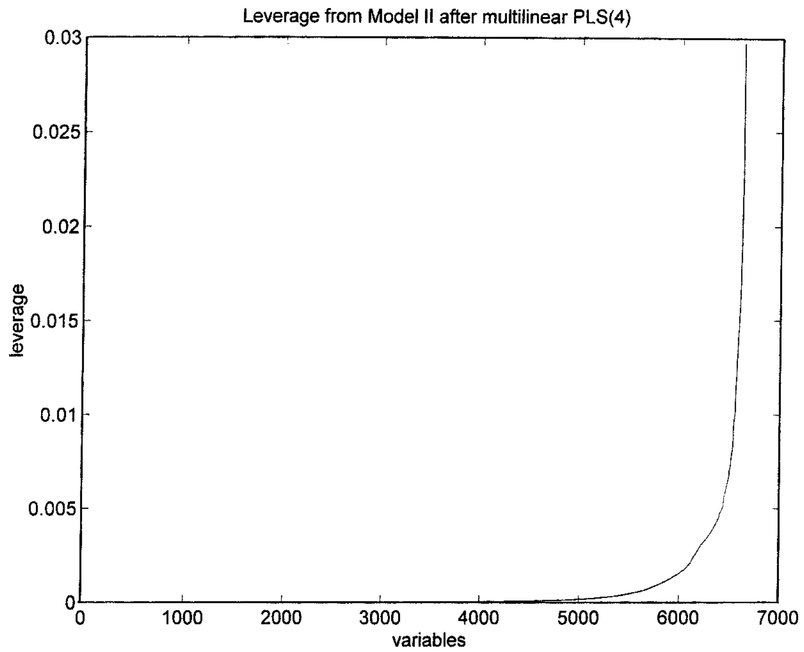

Figure 8. Leverages from Model II after multilinear PLS(4) in increasing order of size

variables was quite low, even after variable reduction, and the idea of omitting low-leverage variables is tempting. This will be dealt with in future research.

APPENDIX: MATLAB CODE FOR REGRESSION COEFFICIENTS IN MULTILINEAR PLS

Smilde[17] gives the following explicit expression for the regression coefficients in multilinear PLS1 calibration based on A components:

$$\mathbf{b}_{\mathrm{PLS}} = \mathbf{W}^*\mathbf{b}_A \tag{20}$$

where

$$\mathbf{W}^* = [\mathbf{w}_1|(\mathbf{I}_P - \mathbf{w}_1\mathbf{w}_1^{\mathrm{T}})\mathbf{w}_2|\ldots|(\mathbf{I}_P - \mathbf{w}_2\mathbf{w}_2^{\mathrm{T}})(\mathbf{I}_P - \mathbf{w}_1\mathbf{w}_1^{\mathrm{T}})\ldots(\mathbf{I}_P - \mathbf{w}_{A-1}\mathbf{w}_{A-1}^{\mathrm{T}})\mathbf{w}_A] \tag{21}$$

In equation (21), $\mathbf{w}_a$ is the vectorized (unfolded) form of the rank-one N-way tensor product obtained from the mode-specific weight vectors $\mathbf{w}^J$, $\mathbf{w}^K$, etc. that define the ath PLS component.

Equation (21) is not suitable for implementation in predictive CoMFA computations using N-PLS regression, since it involves very large matrices $\mathbf{I}_P - \mathbf{w}_a\mathbf{w}_a^{\mathrm{T}}$ $(P \times P)$. For example, in our current application ($P = JKLM = 31 \times 15 \times 18 \times 3 \approx 25\,000$) one such matrix occupies 5 Gb. Merely multiplying two such matrices takes 31 Tflops!

Let us consider the second column of $\mathbf{W}^*$. The expression $(\mathbf{I}_P - \mathbf{w}_1\mathbf{w}_1^{\mathrm{T}})\mathbf{w}_2$ represents the projection of $\mathbf{w}_2$ onto $\mathbf{w}_1^{\perp}$, the orthogonal complement of $\mathbf{w}_1$. It is more efficient, with respect to both space and time, to compute this as $\mathbf{w}_2 - (\mathbf{w}_1^{\mathrm{T}}\mathbf{w}_2)\mathbf{w}_1$. The same approach can be used recursively in each of the subsequent columns, starting from the back. The MATLAB code implementing this procedure is given below as Algorithm I. It requires little additional storage and involves $2A^2P$ flops.

Algorithm I

```
function bPLS=getbpls1(W,b)

% function bPLS=getbpls1(W,b)
% gives explicit b_PLS in trilinear PLS
% (i.e. yhat=X*b_PLS)
% from W(JK×A) and b(A×1)

A=size(W, 2);
bPLS=0;
for a=1:A
   v=W(:,a);
   for j=a-1:-1:1
      v=v-(v'*W(:,j))*W(:,j);
   end
   bPLS=bPLS+b(a)*v;
end
```

We may increase the speed further by starting at the last column of $\mathbf{W}^*$, i.e. computing $b_A\mathbf{w}_A$, projecting this onto $\mathbf{w}_{A-1}^{\perp}$, adding the result to $b_{A-1}\mathbf{w}_{A-1}$, projecting this onto $\mathbf{w}_{A-2}^{\perp}$, adding the result to $b_{A-2}\mathbf{w}_{A-2}$ and so forth. In this way we obtain the alternative Algorithm II. It requires $(4A\text{-}3)P$ flops; hence Algorithm II is about $A/2$ times faster than Algorithm I.

Algorithm II

```
function bPLS=getbpls1 (W,b)

% function bPLS=getbpls1 (W, b)
% gives explicit b_PLS in trilinear PLS
% (i.e. yhat=X*b_PLS)
% from W (JK×A) and b(A×1)

A=length(b);
bPLS=b(A)*W(:,A);
for a=A-1:-1:1
   bPLS=bPLS+(b(a)-bPLS'*W(:,a))*W(:,a);
end
```

Other approaches to computing N-PLS regression coefficients for prediction purposes are discussed in Reference 22.

REFERENCES

1. R. D. Cramer III, D. E. Patterson and J. D. Bunce, *J. Am. Chem. Soc.* **110**, 5959–5967 (1988).
2. A. Agarwal, P. P. Pearson, E. W. Taylor, H. B. Li, T. Dahlgren, M. Herslof, Y. Yang, G. Lambert, D. L. Nelson, J. W. Regan *et al. J. Med. Chem.* **36**, 4006–4014 (1993).
3. K. Raghavan, J. K. Buolamwini, M. R. Fesen, Y. Pommier, K. W. Kohn and J. N. Weinstein, *J. Med. Chem.* **38**, 890–897 (1995).
4. T. I. Oprea, C. L. Waller and G. R. Marshall, *Drug Des. Discov.* **12**, 29–51 (1994).
5. F. Briens, R. Bureau, S. Rault and M. Robba, *Ecotoxicol. Environ. Safety*, **31**, 37–48 (1995).
6. *SYBYL—Molecular Modeling Software, 6.3*, Tripos, St. Louis (1996).
7. P. Goodford, *GRID*, Molecular Discovery, Oxford (1995).
8. M. Baroni, G. Costantino, G. Cruciani, D. Riganelli, R. Valigi and S. Clementi, *Quant. Struct.–Act Relat.* **12**, 9–20 (1993).
9. P. Geladi and B. R. Kowalski, *Anal. Chim. Acta*, **185**, 1–17 (1986).
10. R. Bro, *J. Chemometrics*, **10**, 47–61 (1996).
11. R. Bro and H. Heimdal, *Chemometrics and Intell. Lab Syst.* **34**, 85–102 (1996).
12. S. Glase, H. C. Akunne, T. G. Heffner, S. J. Johnson, S. R. Kesten, R. G. Mackenzie, P. J. Manley, T. A. Pugsley, J. L. Wright and L. D. Wise, *Bioorg. Med. Chem. Lett.* **6**, 1361–1366 (1996).
13. S. J. Cho and A. Tropscha, *J. Med. Chem.* **38**, 1060–1066 (1995).
14. U. Norinder, *J. Chemometrics*, **10**, 95–105 (1996).
15. A. K. Smilde and D. A. Doornbos, *J. Chemometrics*, **5**, 345–360 (1991).
16. A. K. Smilde, *Chemometics Intell. Lab. Syst.* **15**, 143–157 (1992).
17. A. K. Smilde, *J. Chemometrics*, **11**, (1997).
18. H. Martens and T. Naes, *Multivariate Calibration*, Wiley, Chichester (1989).
19. J. Nilsson, H. Wilkström, A. K. Smilde, S. Glase, T. Pugsley, G. Cruciani, M. Pastor and S. Clementi, *J. Med. Chem.* **40**, 833–840 (1997).
20. G. Cruciani, M. Baroni, G. Costantino, D. Riganelli and B. Skagerberg, *J. Chemometrics*, **6**, 335–346 (1992).
21. M. Baroni, S. Clementi, G. Cruciani, G. Costantino, D. Riganelli and E. Oberrauch, *J. Chemometrics*, **6**, 347–56 (1992).
22. S. de Jong, in preparation (1997).

