

Multilinear PLS Analysis

Application to a 3D QSAR Data Set

This chapter is based on the article: Nilsson, J.; De Jong, S.; Smilde, A. K. Multiway Calibration in 3D QSAR. *J. of Chemometrics* **1997**,11, 511-524.

Summary

The multilinear PLS method has been employed for the analysis of a set of benzamides with affinity for the dopamine D₃ receptor subtype, synthesised 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 accounts for clearly defined spatial regions, 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 better predicted with multilinear PLS ($Q^2 = 31\%$) as compared with bilinear PLS ($Q^2 = 25\%$). Additionally, with multilinear PLS one loses in fit, but gains in stability and simplicity due to the smaller number of 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 stabilises the 3D QSAR model.

5.1 Introduction

Since Cramer *et al.*¹ presented the Comparative Molecular Field Analysis (CoMFA)^{1,2} procedure in 1988, it has frequently been utilised by medicinal³⁻⁵ and environmental chemists,⁶ as implemented in the SYBYL molecular modelling package.⁷ Today, other similar approaches are available, e.g., the GRID⁸ program in combination with GOLPE variable selection⁹ (see also Chapter 4). Rational drug design with 3D QSAR comprises several subsequent steps: conformational analyses, alignment of the molecules, generation of molecular descriptors and regression analysis. Optionally, one or more biological response(s) can be used as the independent variable(s).

First, low energy conformations of the molecules are aligned by superimposition of mutual and possible interaction points with the target receptor protein (Chapters 4 and 6). This is by far the most crucial step in order to achieve reliable 3D QSAR models.

A molecular field is a three dimensional grid, large enough to enclose all the aligned molecules, where in each grid point interactions between a probe atom and each molecule are calculated (see Chapter 1). The interaction values in the grid points are thus utilised as variables in the subsequently following regression analysis.

Since multicollinearity among the descriptor variables may affect the regression analysis detrimentally, PLS¹⁰ is frequently used as the regression method in 3D QSAR. Recently, Bro¹¹ presented the multilinear PLS algorithm (N-PLS, Chapter 2) and demonstrated some additional advantages; multilinear PLS was less influenced by noise, more stable, increased the predictability¹² and improved the interpretation of the result as compared to other methods applied to his data set. Accordingly, the multilinear PLS algorithm was implemented for the analysis of a 3D QSAR data set comprising a set of benzamides and naphthamides^{13,14} (Figure 5.1, Tables 4.1–4.3) characterised in the GRID program. The same data set was also analysed in the previous chapter, utilising the GRID/GOLPE approach. In this chapter, the performances of the N-PLS¹¹ and the bilinear PLS^{10,15} methods have been scrutinised and compared.

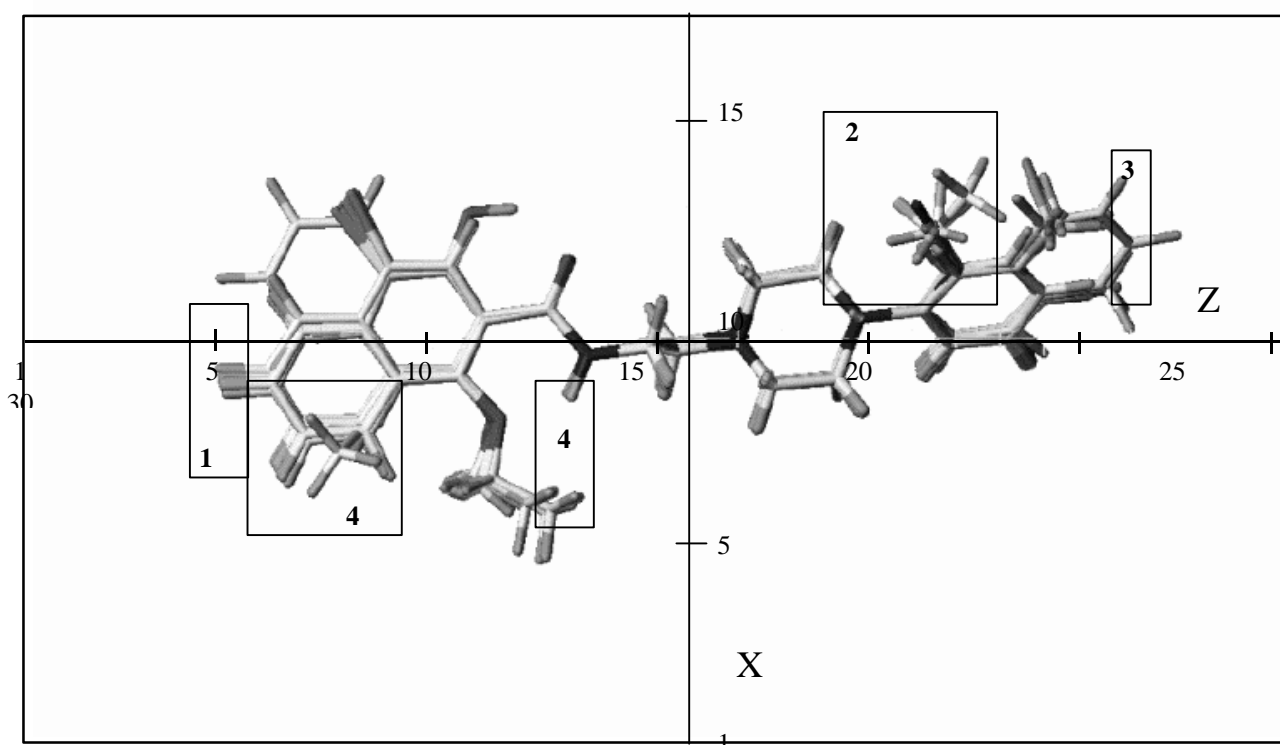


Figure 5.1 The 30 aligned molecules included in the training set viewed in the x and the z mode. The squares indicates the regions where the first four N-PLS components are focused.

It is well known that redundant variables may affect the regression analysis detrimentally and, consequently, several methods to reduce the number of variables^{9,16,17} have been proposed. Multilinear PLS has been employed for the variable reduction of the present data set and the performance of the reduced model was compared with that of the complete model.

5.2 Theory

The multilinear PLS algorithm^{11,18} and the bilinear PLS algorithm^{10,15} have both been, thoroughly, described in Chapter 2. Accordingly, the theory described below are tools, *i.e.*, partial PLS coefficients and leverages, used for the interpretation of the multilinear PLS models. In addition to the data set analysed in the previous chapter, a test set consisting of 21 compounds has been added, to the present investigation, for validation purposes.

: Partial PLS Coefficients

For the purpose of interpretation, the results from CoMFA studies are often presented with contour plots of the partial regression coefficients \mathbf{b}_{PLS} .⁷ Basically, the coefficients \mathbf{b}_{PLS} are needed for predictions of new samples, but since the sizes and the signs of the coefficients reveal the relative importance of the variables, they are also suitable for the interpretation.

A direct relationship between $\mathbf{X}^{(0)}$ and $\bar{\mathbf{y}}$ is searched for:

$$\bar{\mathbf{y}} = \mathbf{T}\mathbf{b}_A = \mathbf{X}^{(0)}\mathbf{b}_{\text{PLS}} \quad (5.1)$$

where $\mathbf{X}^{(0)}$ ($I \times R$) is the unfolded original $\underline{\mathbf{X}}$, $\bar{\mathbf{y}}$ ($I \times 1$) is the fitted \mathbf{y} , \mathbf{b}_A ($A \times 1$) are the regression coefficients as defined in Equation 5.2 and \mathbf{T} ($I \times A$) is the score matrix. The derivation of the full and closed predictions with multilinear PLS has been presented by Smilde,¹⁸ but since the PLS coefficients are frequently utilised in 3D QSAR, it is essential to repeat the derivation also in this context.

Since the scores from different components are not orthogonal the regression coefficients \mathbf{b}_A , in Equation 5.1, have to be calculated taking all the score vectors into account:

$$\mathbf{b}_A = (\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T\mathbf{y} \quad (5.2)$$

Additionally, the weights obtained with multilinear PLS are also not orthogonal and need to be taken into account when the \mathbf{b}_{PLS} coefficients are derived (below).

For clarity, \mathbf{X} is updated after the a th component with $\mathbf{X}^{(a)} = \mathbf{X}^{(a-1)} - \mathbf{t}_a\mathbf{w}_a^T$, as in Martens' non-orthogonalized PLS algorithm.¹⁵ If $\underline{\mathbf{X}}$ is three-way, $\mathbf{w}_a = \mathbf{w}_k^K \otimes \mathbf{w}_j^J$, where \otimes represents the Kronecker product then

$$\mathbf{t}_1 = \mathbf{X}^{(0)}\mathbf{w}_1 \quad (5.3)$$

$$\mathbf{t}_2 = \mathbf{X}^{(1)}\mathbf{w}_2 = (\mathbf{X}^{(0)} - \mathbf{t}_1\mathbf{w}_1^T)\mathbf{w}_2 = (\mathbf{X}^{(0)} - \mathbf{X}^{(0)}\mathbf{w}_1\mathbf{w}_1^T)\mathbf{w}_2 = \mathbf{X}^{(0)}(\mathbf{I} - \mathbf{w}_1\mathbf{w}_1^T)\mathbf{w}_2 \quad (5.4)$$

...

$$\mathbf{t}_A = \mathbf{X}^{(0)}(\mathbf{I} - \mathbf{w}_1\mathbf{w}_1^T)\dots(\mathbf{I} - \mathbf{w}_{A-1}\mathbf{w}_{A-1}^T)\mathbf{w}_A \quad (5.5)$$

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

$$\mathbf{T} = \mathbf{X}^{(0)} \left[\mathbf{w}_1 \left| (\mathbf{I} - \mathbf{w}_1\mathbf{w}_1^T)\mathbf{w}_2 \right| \dots \left| (\mathbf{I} - \mathbf{w}_1\mathbf{w}_1^T)(\mathbf{I} - \mathbf{w}_2\mathbf{w}_2^T)\dots(\mathbf{I} - \mathbf{w}_{A-1}\mathbf{w}_{A-1}^T)\mathbf{w}_A \right. \right] \quad (5.6)$$

Insertion of Equation 5.6 in Equation 5.1 followed by rearrangement gives:

$$\mathbf{b}_{\text{PLS}} = \left[\mathbf{w}_1 \left| \left(\mathbf{I} - \mathbf{w}_1 \mathbf{w}_1^T \right) \mathbf{w}_2 \right| \dots \left| \left(\mathbf{I} - \mathbf{w}_1 \mathbf{w}_1^T \right) \left(\mathbf{I} - \mathbf{w}_2 \mathbf{w}_2^T \right) \dots \left(\mathbf{I} - \mathbf{w}_{A-1} \mathbf{w}_{A-1}^T \right) \mathbf{w}_A \right| \right] \mathbf{b}_A \quad (5.7)$$

When the number of variables is large, as in 3D QSAR, computing the outer product of the weights can be a problem. However, computational shortcuts are possible (see below).

If $\mathbf{w}_i^T \mathbf{w}_j = 0$ ($i \neq j$) then Equation 5.7 reduces to:

$$\mathbf{b}_{\text{PLS}} = [\mathbf{w}_1 | \mathbf{w}_2 | \dots | \mathbf{w}_A] \mathbf{b}_A = \mathbf{W} \mathbf{b}_A \quad (5.8)$$

which resembles the solution obtained with Martens' non-orthogonalized PLS algorithm.¹⁵

: Leverages

In order to determine which variables that have influenced the model most, the variables were ranked by their leverages¹⁵ (\mathbf{h}). The leverages are determined by first calculating an overall weight matrix, $\mathbf{W} = (\mathbf{w}_1 | \mathbf{w}_2 | \dots | \mathbf{w}_A)$, in which \mathbf{w}_a ($R \times 1$; $R = JKLM$) combines the weights from the different modes as

$$\mathbf{w}_a = \mathbf{w}_a^M \otimes \mathbf{w}_a^L \otimes \mathbf{w}_a^K \otimes \mathbf{w}_a^J \quad (a = 1, \dots, A) \quad (5.9)$$

The \otimes sign represents the Kronecker product and a denotes the component number. The leverage vector¹⁵ \mathbf{h} ($R \times 1$) after A components is then expressed as

$$\mathbf{h} = \text{diag}(\mathbf{W} \mathbf{W}^T) \quad (5.10)$$

A variable with a leverage h_r close to zero has not affected the model very much while a variable with a h_r close to one is very important for the model. The average h_r is A/R and variables with leverage exceeding $h_{\text{cut}} \times A/R$ (h_{cut} being an integer, normally 2 or 3) may, according to Martens and Næs,¹⁵ be considered significant.

: Model Validation

In the present investigation crossvalidation and external predictions have been utilised for the validation of the obtained models. The results from the validation experiments are quantified with the crossvalidated Q^2 and the predicted Q^2 as calculated in Equation 5.11. The quality of the calibrations are given by the multiple regression coefficient R^2 in Equation 5.12:

$$Q^2 = \left[1 - \left(\frac{\sum_{i=1}^I (y_i - \bar{y}_{(i)})^2}{\sum_{i=1}^I (y_i - \bar{y})^2} \right) \right] \times 100 \quad (5.11)$$

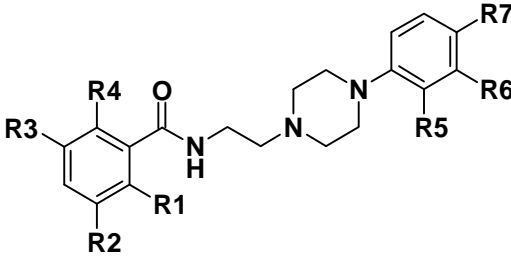
$$R^2 = \left[1 - \left(\frac{\sum_{i=1}^I (y_i - \bar{y}_i)^2}{\sum_{i=1}^I (y_i - \bar{y})^2} \right) \right] \times 100 \quad (5.12)$$

The predicted y in Equation 5.11 is denoted $\bar{y}_{(i)}$, *i.e.*, in the case of crossvalidation an estimation of y_i using a model with the i th object excluded. In the case of external predictions, y_i is the response of the i th test object estimated with the complete calibration model. The fitted y from the calibration in Equation 5.12 is denoted \bar{y}_i .

: The Test Set

The molecules analysed in this investigation were synthesised by Glase *et al.*¹³ In addition to the 30 compounds from the previous chapter a test set, consisting of 21 compounds, was introduced for validation purposes (Tables 5.1–5.3).

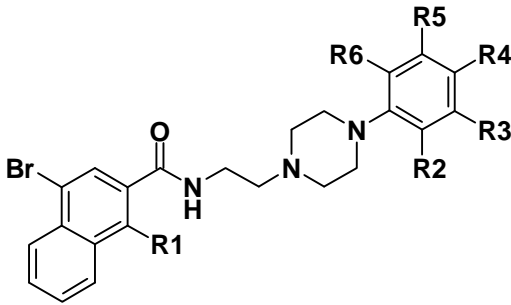
Table 5.1 The benzamides included in the test set used for the validation of the models obtained in this chapter



Compd	R1	R2	R3	R4	R5	R6	R7	$\log_{10}(K_i)^a$
t1								2.5
t2	-OMe	-Cl	-Cl	-OH	-OMe			2.9
t3	-OMe	-Cl	-Cl	-OH		-Cl	-F	3.1
t4	-OMe	-Cl	-Cl	-OH		-CF ₃		3.3

^a \log_{10} was performed on the K_i (nM)

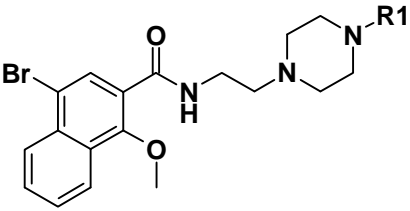
Table 5.2 Naphthamides included in the test set used for the validation of the models obtained in this chapter.

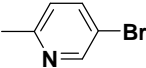



Compd	R1	R2	R3	R4	R5	R6	$\log_{10}(K_i)^a$
t5	-OMe	-Cl	-Cl				0.9
t6	-OH	-Cl	-Cl				1.7
t7	-OMe	-Cl			-Cl		2.4
t8	-OMe			-Cl			2.1
t9	-OMe		-Cl	-F			2.9
t10	-OMe	-F	-F				1.7
t11	-OMe	-F			-F		1.5
t12	-OMe		-F				2.1
t13	-OMe		-F	-F			2.6
t14	-OMe	-Br					2.3
t15	-OMe		-Br				1.6
t16	-OMe			-Br			3.1
t17	-OMe		-CN				0.9
t18	-OMe			-CN			3.2
t19	-OMe	-Me				-Me	2.7

^a \log_{10} was performed on the K_i (nM)

Table 5.3 Naphthamides included in the test set used for the validation of the models obtained in this chapter.



Compd	R1	$\log_{10}(K_i)^a$
t20		2.4
t21		3.0

^a \log_{10} was performed on the K_i (nM)

Low energy conformations of all the molecules were initially aligned as described in Chapter 4 and subsequently surrounded by a three dimensional grid large enough to enclose all the aligned molecules with four Å in all directions (Figure 5.1). The directions x, y and z in the grid were divided into 31, 15 and 18 steps of 1 Å, respectively, yielding a total of 8370 grid points. The surroundings of each molecule were mapped by calculating the interactions between probe atoms and each molecule at each grid point. The resulting grid, filled with interaction values, is called a molecular field. Three different probes⁸ were used, 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, the differences in these fields are correlated with, *e.g.*, the affinities for a certain receptor subtype. The complete model is described graphically in Figure 5.2.

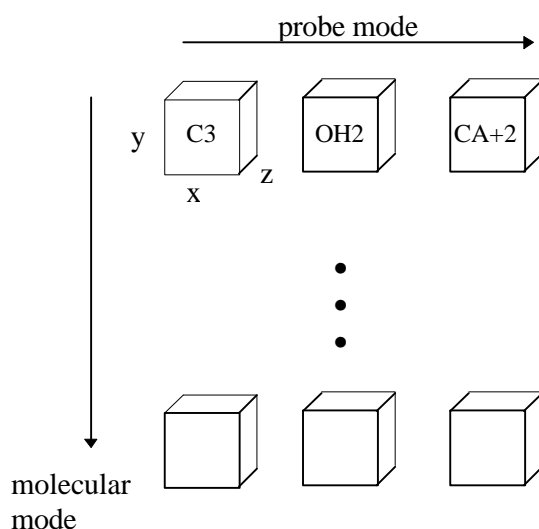


Figure 5.2 The data set comprises five different modes. The molecular, x, y, z and the probe modes consist of 30, 31, 15, 18 and 3 dimensions, respectively.

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 2.4. With multilinear PLS, however, the unfolding step is omitted and the one-component decomposition 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

$$\underline{\mathbf{X}} = \begin{array}{c} \mathbf{w}^J \\ \mathbf{w}^K \\ \mathbf{w}^L \\ \mathbf{w}^M \\ \mathbf{t} \end{array} + \underline{\mathbf{E}}$$

Figure 5.3 The multiway decomposition of $\underline{\mathbf{X}}$ ($I \times J \times K \times L \times M$) into 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 ($M \times 1$). $\underline{\mathbf{E}}$ is the part of $\underline{\mathbf{X}}$ not accounted for by the model.

in Figure 5.3. The vectors \mathbf{t} , \mathbf{w}^J , \mathbf{w}^K , \mathbf{w}^L and \mathbf{w}^M correspond directly to the molecular, x, y, z and the probe mode, respectively, as described in Figure 5.2.

5.3 Results

: Model I

The only data pre-processing applied was mean-centering in the molecular mode. In bilinear PLS, scaling is often performed column-wise, *e.g.*, auto-scaling¹⁰ whereas in multilinear PLS scaling is not that straightforward.¹⁹

The objective of this investigation is to introduce the multilinear PLS method in 3D QSAR modelling and compare its solution with the bilinear PLS solution. Accordingly, the complete model (Model I) was calibrated and validated with both regression methods, presented in Tables 5.4 and 5.5, respectively. With multilinear PLS (Table 5.4) maximum crossvalidated Q^2 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 crossvalidated Q^2 was found after only one component ($Q^2 = 48\%$), where 22% of the variation in \mathbf{X} explained 62% of the variation in \mathbf{y} . The weights from the different modes obtained with multilinear PLS are useful for the interpretation of the result. The weights from the first four components are plotted in Figure 5.4. For comparison, the weight vector from the first component with the reduced bilinear PLS model is plotted in Figure 5.5.

Table 5.4 Calibration and validation of Model I (30×25110) with multilinear PLS.^a

#LV	$R^2(\underline{\mathbf{X}})$	$R^2(\mathbf{y})$	$Q^2(\text{LOO})$	$Q^2(\text{Pred})^b$
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 the external test set (21×25110)

Table 5.5 Calibration and validation of Model I (30×25110) with bilinear PLS.^a

#LV	$R^2(\mathbf{X})$	$R^2(\mathbf{y})$	$Q^2(\text{LOO})$	$Q^2(\text{Pred})^b$
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 the external test set (21×25110)

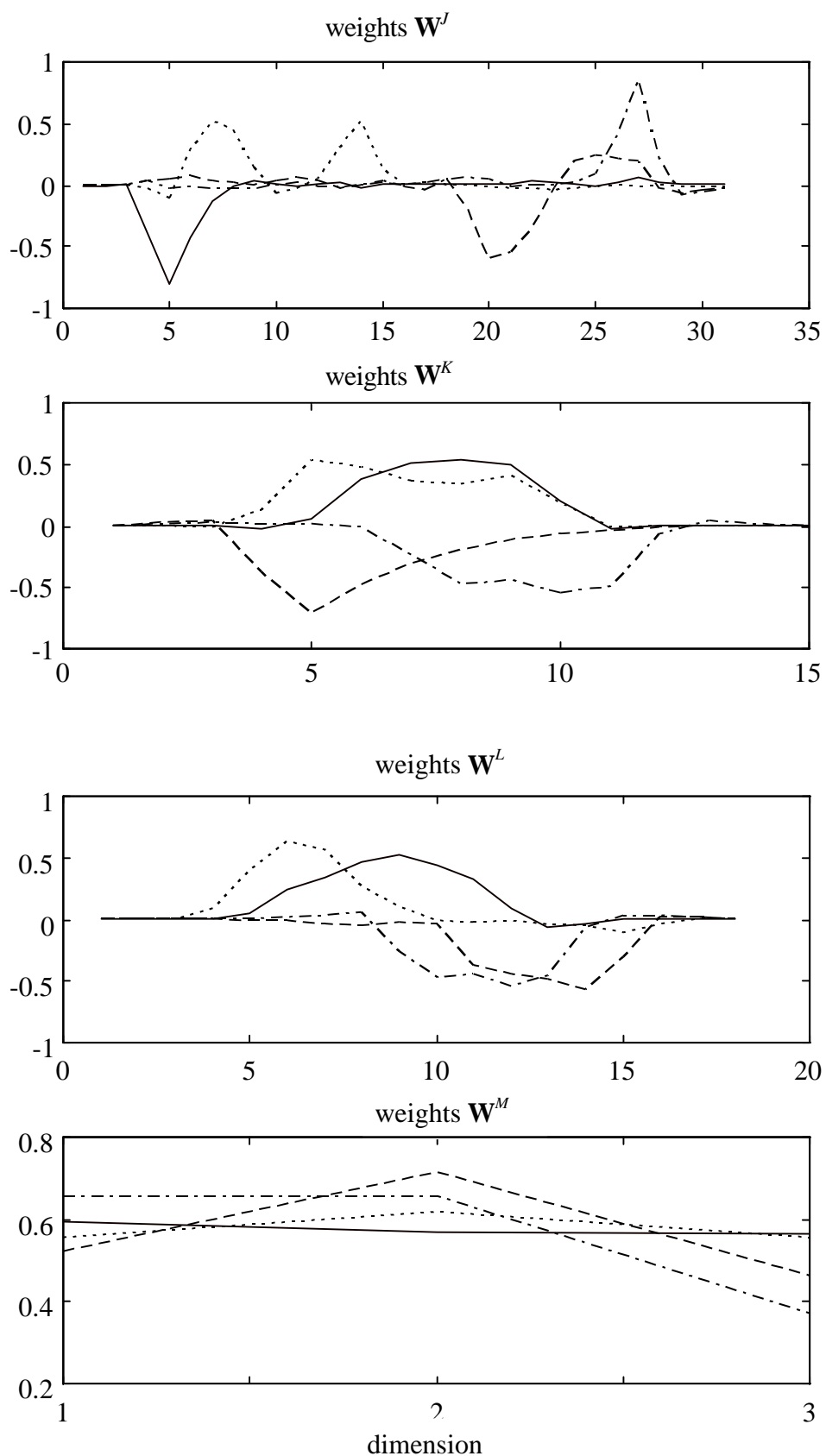


Figure 5.4 The weights W^J , W^K , W^L and W^M : ———, component one; ·····, component two; - - - - -, component three; - · - · - ·, component four.

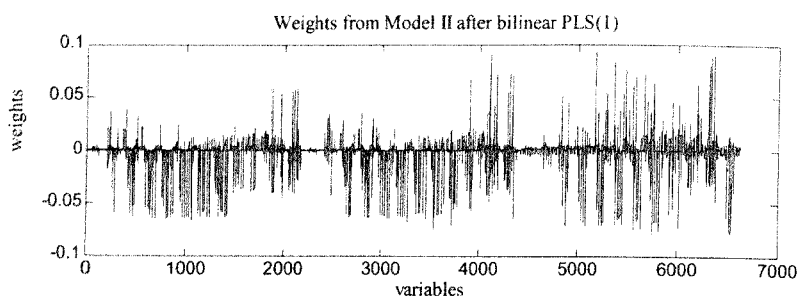


Figure 5.5 Weight vector (w_i) after first component with bilinear PLS.

The number of significant components was estimated by leave-one-out (LOO) crossvalidation and maximum crossvalidated Q^2 was found after four components (Table 5.4) with multilinear PLS. In order not to lose information during the variable reduction step, the variable reduction was performed from a model one component more complex than optimal. Accordingly, the absolute sum of the weights, after 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 number of 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 summarised in Tables 5.6 and 5.7 which was validated thoroughly (Table 5.7) with crossvalidation and external predictions. In addition to traditional 'leave-one-out' crossvalidation also 'leave-three-out' and 'leave-five-out' crossvalidations 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 crossvalidation experiments.^{20,21}

In order to simplify the interpretations of a PLS model in 3D QSAR, the partial PLS coefficients b_{PLS} in Equation 5.8 are often presented as comprehensive iso-contour plots. That is, each b_{PLS} is transferred back to its original position in the grid, where grid points with similar coefficients are connected. In Figure 5.6, the b_{PLS} contour is plotted in stereo from the C3 probe after the fourth multilinear PLS component.

Table 5.6 Crossvalidations and external predictions of Model II (30×6624).

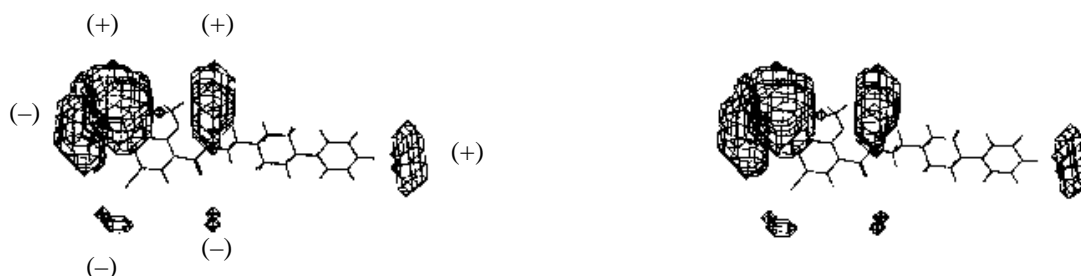
# LV	N-PLS ^a				PLS ^a			
	LOO ^b	L3O ^b	L5O ^{b,c}	Pred. ^d	LOO ^b	L3O ^b	L5O ^{b,c}	Pred. ^d
1	39	43	42	19	50	50	50	25
2	43	44	41	18	48	46	47	23
3	45	43	41	29	48	46	48	33
4	51	53	49	31	44	41	42	30
5	43	42	38	34	39	37	40	31

^a all values in percentage; ^b LOO is short for leave-one-out, L3O for leave-three-out and L5O for leave-five-out; ^c average from 20 Q^2 s; ^d predictions of the external test-set (21×6624)

Table 5.7 Calibration of Model II with bilinear PLS and multilinear PLS for the first five components.^a

# LV	N-PLS		PLS	
	$R^2(\underline{X})$	$R^2(\underline{y})$	$R^2(\underline{X})$	$R^2(\underline{y})$
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 5.6** The b_{PLS} coefficients from the final multilinear PLS model and C3 probe after four components.

5.4 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 compounds. The solution should also be simple and straightforward, since also the non-expert must be able to interpret the model.

The initial complete model (Table 5.4) indicated four significant components with leave-one-out crossvalidation. With help from Figure 5.4 it can be determined, with good precision, which regions are accounted for by the components. The full lines in Figure 5.4 represent the weights from the first component, the broken curves the second component, the chain curves the third and the dotted curves the fourth component. For clarity, the weights \mathbf{W}^J and \mathbf{W}^L correspond to the x and z modes

in Figure 5.1, respectively. The first component has high weights w^j in position 5 and high weights w^l between positions 5 and 10 (Figure 5.4), which correspond to the region where the naphthalene moiety protrudes (Figure 5.1). Thus the first component accounts for the differences 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 (Figures 5.4), the weights from bilinear PLS (Figure 5.5) are difficult to interpret.

Striking is the significantly lower percentage variance explained with multilinear PLS method (Table 5.4) as compared with the bilinear PLS (Table 5.5) method. A speculative explanation for this is the fewer parameters that need to be estimated with multilinear PLS.^{11,22} Additionally, 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 predictability detrimentally. From Figure 5.4 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 above, a reduced model with 6624 variables was obtained which was validated with crossvalidation and external predictions. Variable selection must be performed very carefully, otherwise problems with overfitting may occur. Norinder¹⁷ and Cho¹⁶ reported that the crossvalidated Q^2 increased in their models, while the ability to predict external test sets decreased when the number of variables was reduced. In the present investigation the number of variables were reduced from 25110 to 6624, which speeded up further calculations, but with no improvement in the predictability (nor decrease) as the result.

From Model II (Table 5.6) it can be concluded that the model is homogenous and stable, since the crossvalidated Q^2 was not affected very much when larger groups of molecules were left out each time. Each crossvalidation experiment was repeated 20 times²⁰ and, accordingly, reported as the average Q^2 .

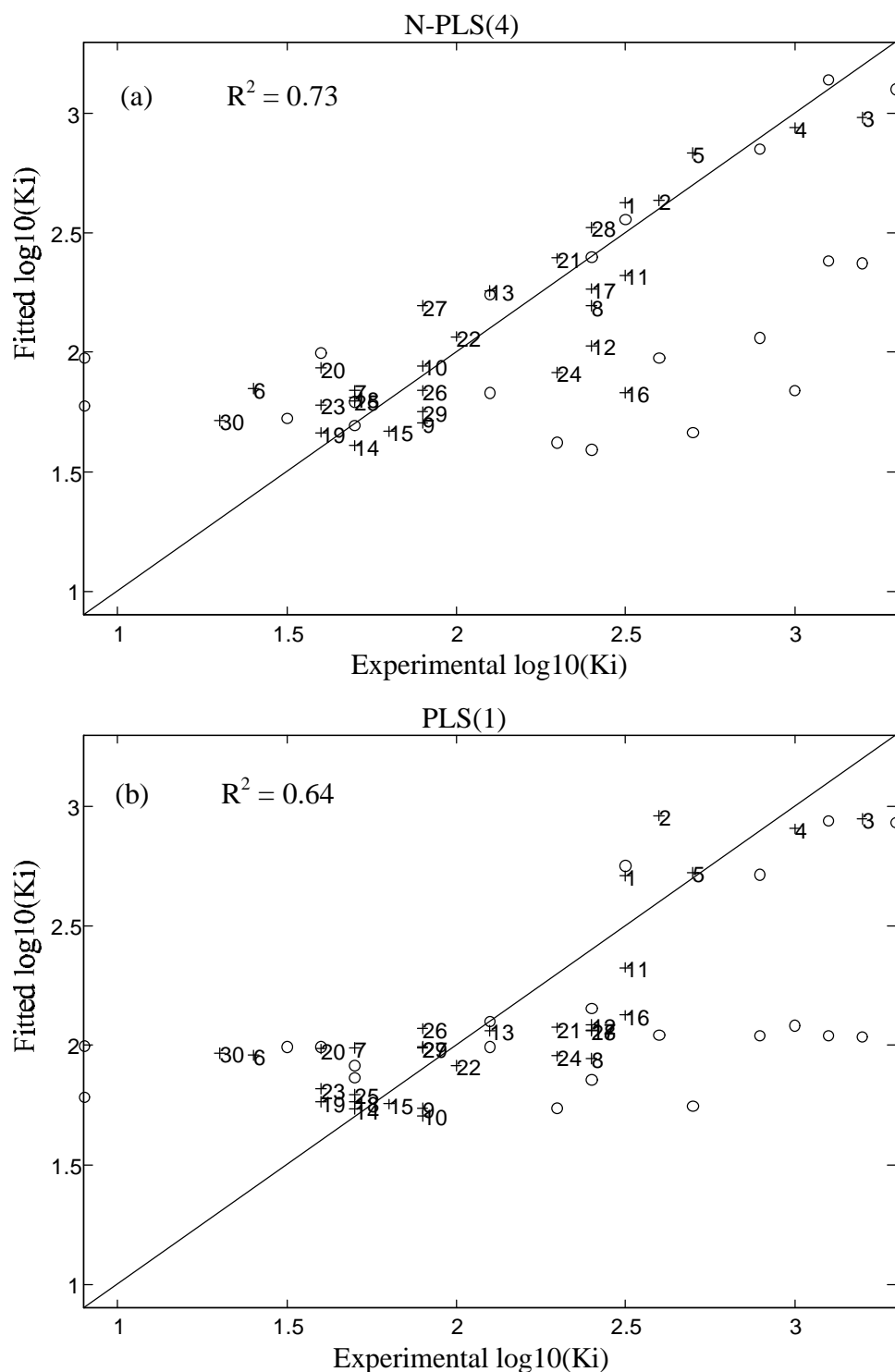


Figure 5.7 Experimental $\log_{10}(K_i)$ versus fitted $\log_{10}(K_i)$ after (a) four component with multilinear PLS and (b) one component with bilinear PLS. The small rings represents the predictions of the external test set (21×6624).

In Figure 5.7(a) and 5.7(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. The four-component model with multilinear PLS ($R^2 = 73\%$) explains more of the variation in y as compared to the 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 5.7(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 5.7(a)).

The iso-contour plot of the \mathbf{b}_{PLS} coefficients after the fourth component, in Figure 5.6, is probably the most comprehensible tool for the interpretation of the model:

$$y = x_1b_1 + \dots + x_ib_i + \dots + x_Rb_R + e \quad (5.13)$$

If a novel molecule is designed with a substituent protruding in a negative \mathbf{b}_{PLS} region then x_i in Equation 5.13 will be positive and consequently x_ib_i will be negative. This substituent will thus have a negative effect on y . If low values of y is desirable, new substituents must be added in regions where the \mathbf{b}_{PLS} for the C3-probe (steric-field) is negative, and *vice versa*. For a more elaborated explanation of how to interpret the iso-contour plots the SYBYL-manual⁷ or Chapter 4 in this thesis are recommended.

In Figure 5.8 the 6624 variables are ranked by their leverages. Even after variable reduction a lot of variables with low influence on the model are present.

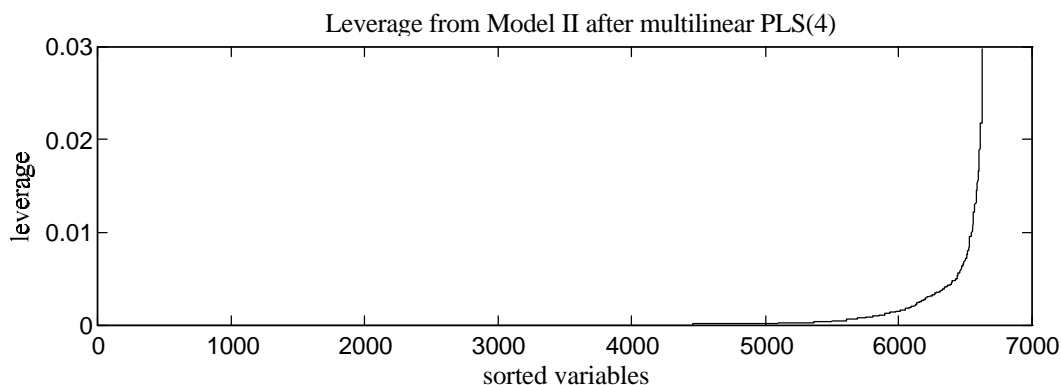


Figure 5.8 Leverage from Model II after four multilinear PLS components ordered in increasing order of size.

5.5 Conclusions

The multilinear PLS method has successfully been introduced 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 were effectively reduced with help from the multilinear PLS weights. The variable selection did not improve the predictability but speeded up the calculations significantly. The number of high leverage variables was quite low even after variable reduction.

5.6 Matlab Code for Regression Coefficients in Multilinear PLS

Smilde¹⁸ gives the following explicit expression for the regression coefficients in multilinear PLS1 calibration based on A components:

$$\mathbf{b}_{\text{PLS}} = \mathbf{W}^* \mathbf{b}_A \quad (5.14)$$

where

$$\mathbf{W}^* = \left[\mathbf{w}_1 \left| \left(\mathbf{I}_R - \mathbf{w}_1 \mathbf{w}_1^T \right) \mathbf{w}_2 \right| \dots \left| \left(\mathbf{I}_R - \mathbf{w}_1 \mathbf{w}_1^T \right) \left(\mathbf{I}_R - \mathbf{w}_2 \mathbf{w}_2^T \right) \dots \left(\mathbf{I}_R - \mathbf{w}_{A-1} \mathbf{w}_{A-1}^T \right) \mathbf{w}_A \right] \quad (5.15)$$

In Equation 5.15 \mathbf{w}_a is the vectorized (unfolded) form of the rank-1 N -way tensor product obtained from the mode-specific weight vectors \mathbf{w}^J , \mathbf{w}^K , etc. that define the a th PLS component.

Equation 5.15 is not suitable for implementation in predictive CoMFA computations using N-PLS regression since it involves very large matrices $\mathbf{I}_R - \mathbf{w}_a \mathbf{w}_a^T$ ($R \times R$). For example, in the current application ($R = JKLM = 31 \times 15 \times 18 \times 3 \approx 25000$) 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 $(\mathbf{I}_R - \mathbf{w}_1 \mathbf{w}_1^T) \mathbf{w}_2$ 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 $\mathbf{w}_2 - (\mathbf{w}_1^T \mathbf{w}_2) \mathbf{w}_1$. The same approach can be used recursively in each of the subsequent columns, starting from the back. MATLAB²³ code implementing this procedure is given below as Algorithm I. It requires little additional storage and involves $2A^2R$ flops.

The speed may be increased even further by starting at the last column of \mathbf{W}^* , *i.e.*, computing $\mathbf{b}_A \mathbf{w}_A$, projecting this onto \mathbf{w}_{A-1}^\perp , adding the result to $\mathbf{b}_{A-1} \mathbf{w}_{A-1}$, projecting this onto \mathbf{w}_{A-2}^\perp , adding the result to $\mathbf{b}_{A-2} \mathbf{w}_{A-2}$, and so forth. In this way an alternative Algorithm II is obtained. It requires $(4A-3)R$ flops; hence Algorithm II is about $A/2$ times faster than Algorithm I.

Other approaches to compute N-PLS regression coefficients for prediction purposes are discussed elsewhere.²⁴

ALGORITHM I

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

% function bPLS = getbpls1(W, b)
% gives explicit b_PLS in trilinear
% PLS
% (i.e. y^hat = X * b_PLS )
% from W(JKxA) and b(Ax1)

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
```

ALGORITHM II

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

% function bPLS = getbpls1(W, b)
% gives explicit b_PLS in trilinear
% PLS
% (i.e. y^hat = X * b_PLS )
% from W(JKxA) and b(Ax1)

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

5.7 References

1. Cramer III, R.D.; Patterson, D.E.; Bunce, J.D. Comparative Molecular Field Analyses (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959-5967.
2. Cramer III, R.D. and Wold, S. inventors. Comparative Molecular Field Analyses (COMFA). 5025388. United States. Date Filed: **1988/08/26**.
3. Agarwal, A.; Pearson, P.P.; Taylor, E.W.; Li, H.B.; Dahlgren, T.; Herslof, M.; Yang, Y.; Lambert, G.; Nelson, D.L.; Regan, J.W.; et al Three-Dimensional Quantitative Structure-Activity Relationships of 5-HT Receptor Binding Data for Tetrahydropyridinylindole Derivatives: a Comparison of the Hansch and CoMFA Methods. *J. Med. Chem.* **1993**, *36*, 4006-4014.
4. Raghavan, K.; Buolamwini, J.K.; Fesen, M.R.; Pommier, Y.; Kohn, K.W.; Weinstein, J.N. Three-Dimensional Quantitative Structure-Activity Relationship (QSAR) of HIV Integrase Inhibitors: a Comparative Molecular Field Analysis (CoMFA) Study. *J. Med. Chem.* **1995**, *38*, 890-897.
5. Oprea, T.I.; Waller, C.L.; Marshall, G.R. 3D-QSAR of Human Immunodeficiency Virus (I) Protease Inhibitors. III. Interpretation of CoMFA Results. *Drug Des. Discov.* **1994**, *12*, 29-51.
6. Briens, F.; Bureau, R.; Rault, S.; Robba, M. Applicability of CoMFA in Ecotoxicology: a Critical Study on Chlorophenols. *Ecotoxicol. Environ. Saf.* **1995**, *31*, 37-48.
7. SYBYL- Molecular Modeling Software, 6.3, Tripos Incorporated, 1699 S. Hanley Rd, St. Louis, Missouri 63144-2913, USA,
8. GRID, Goodford, P.J. Molecular Discovery Ltd, University of Oxford, England, SGI.
9. Baroni, M.; Costantino, G.; Cruciani, G.; Riganelli, D.; Valigi, R.; Clementi, S. Generating Optimal Linear PLS Estimations (GOLPE): An Advanced Chemometric Tool for Handling 3D-QSAR Problems. *Quant. Struct. -Act. Relat.* **1993**, *12*, 9-20.
10. Geladi, P. and Kowalski, B.R. Partial Least Squares: A Tutorial. *Anal. Chem. Acta* **1986**, *185*, 1-17.
11. Bro, R. Multiway Calibration. Multilinear PLS. *J. of Chemometrics* **1996**, *10*, 47-61.
12. Bro, R. and Heimdahl, H. Enzymatic Browning of Vegetables. Calibration and Analysis of Variance by Multiway Methods. *Chemom. and Intell. Lab. Syst.* **1996**, *34*, 85-102.
13. Glase, S.; Akunne, H.C.; Heffner, T.G.; Johnson, S.J.; Kesten, S.R.; MacKenzie, R.G.; Manley, P.J.; Pugsley, T.A.; Wright, J.L.; Wise, L.D. 4-Bromo-1-methoxy-N-[2-(4-aryl-1-piperazinyl)ethyl]-2-naphthalenecarboxamides: Selective Dopamine D₃ Receptor Partial Agonists. *Bioorg. &Med. Chem. Lett.* **1996**, *6*, 1361-1366.
14. Nilsson, J.; Wikström, H.; Smilde, A.K.; Glase, S.; Pugsley, T.A.; Cruciani, G.; Pastor, M.; Clementi, S. A GRID/GOLPE 3D-QSAR Study on a Set of Benzamides and Naphthamides, with Affinity for the Dopamine D₃ Receptor Subtype. *J. Med. Chem.* **1997**, *40*, 833-840.
15. Martens, H. and Næs, T. Multivariate Calibration. John Wiley & Sons: New York, **1989**
16. Cho, S.J. and Tropscha, A. Crossvalidated R²-guided Region Selection for Comparative Molecular field Analyses: A Simple Method to Achieve Consistent Results. *J. Med. Chem* **1995**, *38*, 1060-1066.
17. Norinder, U. Single Domain Mode Variable Selection in 3D QSAR Applications. *J. of Chemometrics* **1996**, *10*, 95-105.
18. Smilde, A.K. Comments on Multilinear PLS. *J. of Chemometrics* **1997**, *11*, 367-377.
19. Smilde, A.K. Three-way Analyses. Problems and Prospects. *Chemom. and Intell. Lab. Syst.* **1992**, *15*, 143-157.
20. Cruciani, G.; Baroni, M.; Costantino, G.; Riganelli, D.; Skagerberg, B. Predictive Ability of Regression Models. Part I: Standard Deviation of Prediction Errors (SDEP). *J. of Chemometrics* **1992**, *6*, 335-346.
21. Baroni, M.; Clementi, S.; Cruciani, G.; Costantino, G.; Riganelli, D.; Oberrauch, E. Predictive Ability of Regression Models. Part II: Selection of the Best Predictive PLS Model. *J. of Chemometrics* **1992**, *6*, 347-356.
22. Smilde, A.K. and Doornbos, D.A. Three-way Methods for the Calibration of Chromatographic Systems: Comparing PARAFAC and Three-way PLS. *J. of Chemometrics* **1991**, *5*, 345-360.
23. Matlab, 4.2c, Simulink Inc.
24. De Jong, S. Regression Coefficients in Multilinear PLS. *submitted* **1997**