The Comparative Molecular Surface Analysis (CoMSA) with Modified Uniformative Variable Elimination-PLS (UVE-PLS) Method: Application to the Steroids Binding the Aromatase Enzyme

Jarosław Polanski* and Rafał Gieleciak

Department of Organic Chemistry, Institute of Chemistry, University of Silesia, PL-40-006 Katowice, Poland

Received June 9, 2002

The application of the CoMSA method to analyze 3D QSAR of 50 steroid aromatase inhibitors is described. The 3D QSAR model obtained, reaching a value of cross-validated $q^2 = 0.96$ ($s = 0.31$), significantly outperforms those reported in the literature for the CoMFA or CoSA (CoSASA). It is shown that the Uniformative Variable Elimination UVE-PLS or modified iterative UVE procedure (IVE-PLS) can be used for indicating the regions contributing to the binding activity. Thus, after separating the series into two groups of the training and test molecules quite correct external predictions result from the processing of the training set. We proved that the procedure of the data elimination provides stable results, if tested in 50 random runs of the IVE-PLS-CoMSA with different training/test sets. Depending upon the procedure used the quality of the predictions for 25 test molecules is given by SDEP $= (\Sigma(y_{pred}-y_{obs})^2/n)^{1/2} = 0.321 - 0.782$.

INTRODUCTION

Steroids are important for many effects that take place in human organism, in particular, aromatase is one of the enzymes affected by these compounds. Aromatase is capable of producing estradiol by the conversion of testosterone. As this process is important during the development of the some types of tumors, the inhibition of the aromatase can be significant for controlling breast cancer. It has been shown that such drugs that can inhibit this enzyme can be effective in advanced breast cancer in postmenopausal women. Aromatase inhibitors are also tested as promising drugs in early stage breast cancer in hopes that they could lower recurrence and mortality.

On the other hand, the rigid skeleton makes steroids an interesting object for similarity studies, e.g. a series of 31 steroids complexed by CBG and TBG globulin is a kind of benchmark allowing the evaluation of drug design methods. Modeling chemical or biological effects and predicting molecular properties are one of the most challenging aims of present day chemistry and pharmacology. Quantitative Structure Activity Relationship (QSAR) and all its variants, i.e., QSPR, QSRR, and finally three-dimensional (3D) approaches,[1−4] are possible strategies that can be followed in such cases. 3D QSAR methods, especially CoMFA and related approaches, have notably contributed to our ability to forecast the activity of potential bioefectors. In previous papers we have described the alternative method of the Comparative Molecular Surface Analysis (CoMSA) that provides a flexible technique for the comparison of the molecular surface and modeling 3D QSARs.[5−7] In our current paper we discuss the application of the CoMSA to the series of 50 steroids inhibiting aromatase. The QSAR and QSDAR analyses of this series have been described in ref 8.

* Corresponding author e-mail: Polanski@us.edu.pl.

Moreover, we discuss possible technical improvements in the CoMSA strategy. Regardless of the method used, modeling 3D QSAR demands a method capable of the multivariative data analysis. The quality of the PLS models depends on the quality of the data input. We will show below that a slightly modified Uniformative Variable Elimination (UVE) procedure,[9] namely, Iterative Variable Elimination based on leave-one-out cross-validation (IVE-PLS), can significantly improve the CoMSA strategy.

Theoretical Background. Self-organizing neural network (SOM) is a technique designed to reduce the dimensionality of the data while preserving topology. Bioinformatics is probably one of the most important recent implementations illustrating the importance of this technique.[10−12] The method has also been applied in chemistry,[13,14] in particular, for two-dimensional mapping of the electrostatic potential on the three-dimensional molecular surfaces [14,15] or partial atomic charges for the atomic molecular representation.[16] The ability to compress the size of the data and to reconstruct the 3D object from the 2D representation makes this procedure an interesting tool for molecular design.[13,14,17] Such maps were used for the visualization of the interactions of individual molecules with biological receptors or designing drugs.[13,14]

In our previous publications we described the use of the Kohonen neural network in QSAR investigations, in particular we designed a scheme of 3D QSAR method by a coupled neural network and PLS system which we called the Comparative Molecular Surface Analysis (CoMSA).[5−7,18]

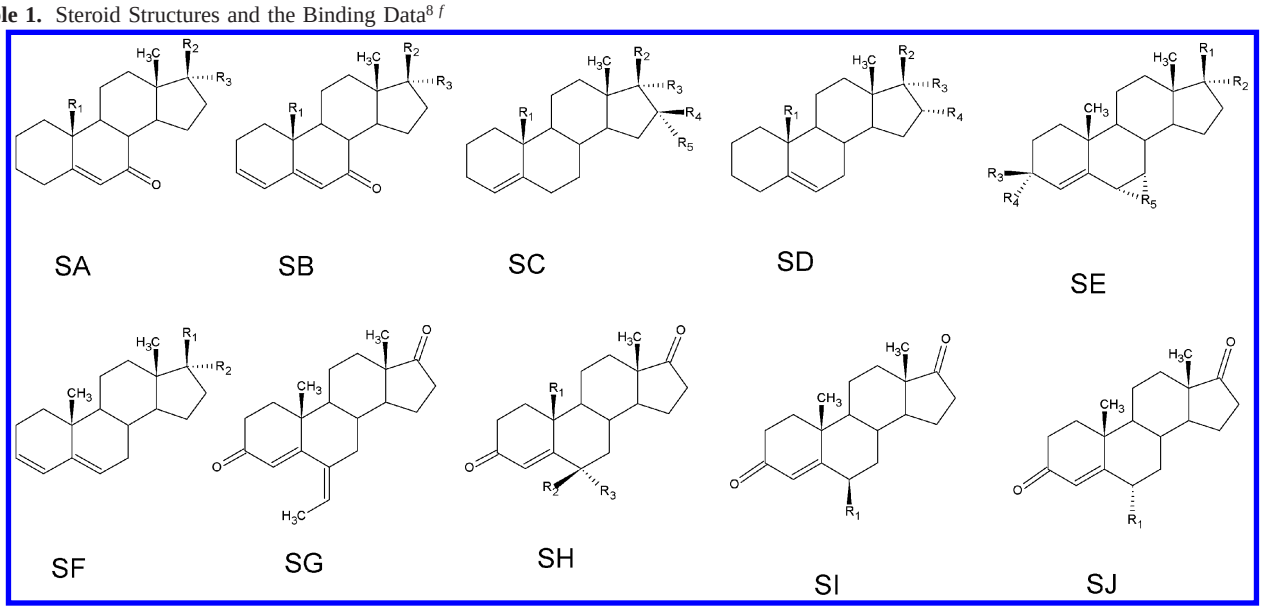
EXPERIMENTAL SECTION

Model Building. All the experimental data for the steroid inhibitors of the aromatase enzyme are extracted from ref 8 and are given in Table 1.

We used Gesteiger's software package for modeling purposes. The 3D-coordinates of all molecules were obtained

10.1021/ci020038q CCC: $25.00
Published on Web 02/08/2003

Table 1. Steroid Structures and the Binding Data[8] [f]



no.	structure	R_1	R_2	R_3	R_4	R_5	BIND[a]	BIND[b]	BIND[c]	BIND[d]
1	SA	CH_2OH	=O				−2.92	−2.35	−3.06	−2.73
2	SA	CH_2OH	OH	H			−3.54	−3.38	−3.71	*e*
3	SA	CHO	=O				−3.00	−2.91	−2.79	−2.58
4	SA	H	=O				−3.26	−2.71	−3.18	*e*
5	SA	Me	OH	H			−2.62	−2.31	−2.59	−2.86
6	SB	CH_2OH	=O				−3.06	−0.73	−2.96	*e*
7	SB	CHO	=O				−2.14	−2.26	−2.20	−2.05
8	SB	H	=O				−2.36	−2.86	−2.50	*e*
9	SD	CH_2OH	=O		H		−1.89	−2.19	−2.15	−2.13
10	SD	CH_2OH	OH	H	H		−2.88	−2.60	−2.64	*e*
11	SD	CHO	=O		H		−2.03	−2.42	−2.48	−2.33
12	SD	Me	=O		H		−0.97	−1.56	−1.18	*e*
13	SD	Me	=O		Br		−2.93	−1.28	−2.09	−2.04
14	SA	Me	=O				−1.28	−2.20	−1.70	*e*
15	SB	Me	=O				−1.23	−1.94	−0.82	−0.68
16	SB	Me	OH	H			−2.61	−3.24	−2.45	*e*
17	SD	Me	OH	H	H		−2.36	−1.63	−2.40	−2.22
18	SF	=O					−0.65	−1.18	−0.58	*e*
19	SF	OH	H				−2.19	−1.39	−2.14	−2.29
20	SH	H	H	H			−1.03	−0.01	−0.78	*e*
21	SC	Me	=O		H	H	0.00	−0.32	−0.12	−0.19
22	SC	CH_2OH	=O		H	H	0.46	−0.25	0.02	*e*
23	SH	CH_2OH	H	H			−0.84	0.16	−0.70	−0.61
24	SH	Me	=O				0.15	−0.32	−0.20	*e*
25	SE	=O		=O		CF_2	−0.13	−0.15	0.01	−0.11
26	SE	=O		H	H	CH_2	0.87	0.15	0.57	*e*
27	SE	OH	H	H	H	CH_2	−0.51	−1.15	−0.78	−0.89
28	SC	Me	OH	H	H	H	−1.35	−1.33	−1.33	*e*
29	SC	CH_2OH	OH	H	H	H	−0.67	−1.65	−0.74	−0.54
30	SC	$MeC(O)OCH_2$	=O		H	H	−0.89	−0.55	−0.61	*e*
31	SC	Me	=O		H	Br	−0.79	−0.93	−0.97	−0.98
32	SC	Me	=O		H	H	−1.09	0.03	−1.09	*e*
33	SC	CF_3	=O		H	H	−1.08	−0.59	−1.27	−0.98
34	SI	Me					0.56	0.44	0.53	*e*
35	SJ	Me					0.87	0.58	0.81	1.05
36	SI	C_2H_5					1.56	0.79	1.16	*e*
37	SJ	C_2H_5					0.94	0.63	0.79	0.69
38	SI	C_3H_7					0.94	0.98	0.93	*e*
39	SJ	C_3H_7					0.78	0.66	0.44	0.69
40	SI	C_4H_9					0.65	1.07	0.48	*e*
41	SJ	C_4H_9					0.53	0.72	0.48	0.64
42	SI	$CH(CH_3)_2$					0.21	0.48	0.39	*e*
43	SJ	$CH(CH_3)_2$					0.04	0.37	0.36	0.22
44	SI	C_6H_5					−0.04	0.22	−0.22	*e*
45	SJ	C_6H_5					0.24	0.16	0.78	0.94
46	SI	$CH_2C_6H_5$					−0.24	0.23	0.40	*e*
47	SJ	$CH_2C_6H_5$					0.61	0.33	0.47	0.48
48	SI	$CH{=}CH_2$					0.91	0.53	1.47	*e*
49	SI	C≡CH					−0.32	0.58	−0.09	−0.06
50	SG						0.96	0.45	0.55	*e*

[a] Experimental values. [b] CoMSA without variable elimination (comp. Table 2, model CoMSA 1). [c] CoMSA with variable elimination (comp. Table 2, model CoMSA 4). [d] CoMSA − external predictions, details in text. [e] Training series, details in text. [f] Structures SA−AJ used with Table 1 for the enzyme aromatase steroid series.

by the 3D structure generator CORINA[19] from the constitution of the respective molecules.[20,21] Partial atomic charges were calculated by the PEOE method,[22,23] and the SURFACE program was used for the calculation of the Coulomb electrostatic potential on the molecular surface.

Data Analysis. Kohonen Mapping. The competitive Kohonen strategy[24] was used to construct a two-dimensional topographic map obtaining the signals from the points sampled randomly at the molecular surface. As molecular surfaces are continuous the plane of projection was also selected to be a continuous surface. Thus we used a torus for this purpose, which was cut along two perpendicular lines and then spread into a plane. Each neuron, *j*, was then defined by three weights, w_{ji}. The competitive training of the network was based on the rule that each point, *s*, of the molecular surface was projected into that neuron, sc, that has weights, w_{ci}, that come closest to the Cartesian coordinates, x_{si}, of this point, *s* (eq 1).

$$\text{out}_{sc} \leftarrow \min[\sum_{i=1}^{m}(x_{si} - w_{ji})^2] \tag{1}$$

A projection of the electrostatic potential value (MEP) from the surface points, *s*, into such a two-dimensional arrangement of neurons, after calculating the average MEP value within this particular neuron and scaling this values into the respective colors results in the so-called feature map.

Comparative Kohonen Mapping. In fact, such a map illustrates the property (MEP) of a single molecule. As however, the weights of the Kohonen network contain the shape of the certain molecular surface, it can be used to compare the geometries of molecular surfaces of other molecules. In such a method the trained Kohonen network is processing the signals coming from the surface of other molecule(s), i.e., the electrostatic potential of each input vector was projected through the network to obtain a series of comparative maps both for the template molecule and each analyzed molecule. The respective electrostatic potential values from the surfaces of the processed molecules were then projected into such a network allowing us to compare these parts of the molecule surfaces that can be superimposed. If the surfaces cannot be superimposed on the reference molecule (template), then the respective output neurons get no signal from the molecules processed. We used either molecule 1 or unsubstituted steroid skeleton or steroid ring D as the templates.

All the molecules were superimposed before the calculation of the molecular surfaces indicating all (17) atoms of the steroid skeleton to be superimposed. In practice, we used Match3D program[25] for performing this operation. The KMAP 3.0 program[25] was used for the simulation of Kohonen networks. The size of the Kohonen networks amounts to 30×30 neurons. The output of this program was used for the calculation of the mean electrostatic potential values within each neuron and respective feature maps were transformed to a respective 30^2 element vectors.

PLS Analysis. Vectors obtained were processed by the PLS analysis with a leave-one-out cross-validation procedure. The PLS procedures were programmed within the MATLAB environment (MATLAB).

A PLS model was constructed for the centered data, and its complexity was estimated based on the leave-one-out cross-validation procedure (CV). The date was recentered (but not rescaled) for each cross-validation run. In the leave-one-out CV one repeats the calibration *m* times, each time treating the *i*th left-out object as the prediction object. The dependent variable for each left-out object is calculated based on the model with one, two, three, etc. factors. The Root Mean Square Error of CV for the model with *j* factors is defined as

$$\text{RMSECV}_j = \sqrt{\frac{\sum(\text{obs}_i - \text{pred}_{i,j})^2}{m}} \tag{2}$$

where obs denotes the assayed value, pred is the predicted value of dependent variable, and *i* refers to the object index, which ranges from 1 to *m*. Model with *k* factors, for which RMSECV reaches a minimum, is considered as an optimal one.

We used the performance metrics that are accepted and widely used in CoMFA analyses, i.e., cross-validated q^2_{cv}

$$q^2_{cv} = 1 - \frac{\sum(\text{obs}_i - \text{pred}_i)^2}{\sum_i(\text{obs}_i - \text{mean(obs)})^2} \tag{3}$$

where obs is the the assayed values, pred is the predicted values, mean is the mean value of obs, and *i* refers to the object index, which ranges from 1 to *m*, and the cross-validated standard error *s* is denoted as

$$s = \sqrt{\frac{\sum(\text{obs}_i - \text{pred}_i)^2}{m - k - 1}} \tag{4}$$

where *m* is the number of objects and *k* is the number of PLS factors in the model.

The quality of external predictions was measured by the SDEP parameter

$$\text{SDEP} = \sqrt{\frac{\sum(\text{pred}_i - \text{obs}_i)^2}{n}} \tag{5}$$

where pred is the predicted value, obs is the observed value, mean is the mean value, *n* is a number of measurements, and opt is a number of the PLS factors used in the model.

Univformative Variable Elimination−PLS Method (UVE-PLS). The UVE algorithm was originally proposed by Centner et al.[9] as a possible improvement of the PLS models. The main idea of the method is to reduce the number of the variables included in the final PLS model. The UVE algorithm based on the analysis of the regression coefficients calculated by PLS method. PLS method allows presenting the relation between the Y answer and X predictors in a form of

$$\mathbf{Y} = \mathbf{X}b + e$$

where *b* is a vector of the regression coefficients and *e* is the vector of the e errors.

Thus, the UVE algorithm analyzes the reliability of the mean(b)/s(b) ratio (where s(b) means standard deviation of b). Then, only the variables of the "relative" high mean(b)/s(b) ratio are included into the final PLS model. To estimate

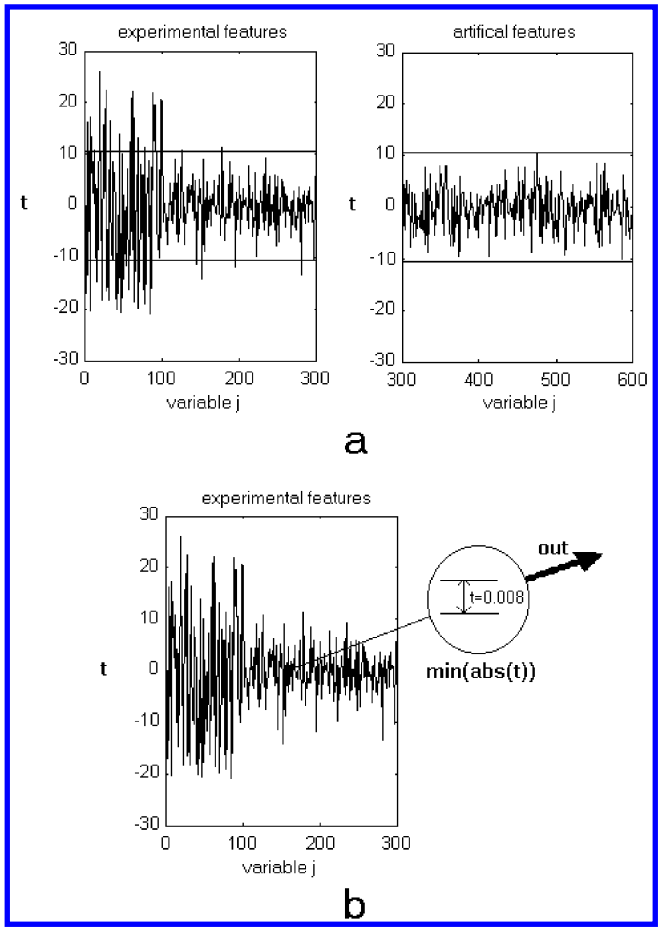


Figure 1. Variable elimination using a/ the robust UVE-PLS − elimination is a single step operation and the region defining the level of surviving variables is given by the *t* value illustrated by two horizontal lines; b/ IVE-PLS − elimination is an iterative procedure; during a single iteration survive all variables but that taking the lowest value.

the cutoff level artificial random number noise is created (the level of the noise is 10^{-10} of the original variable order) and put (as additional columns) into the matrix of the original variables. PLS analysis of such a matrix is performed and mean(b)/s(b) parameter is analyzed for each column. The highest absolute value, abs(mean(b)/s(b)), for the noisy column determines the cutoff level for the original variables as shown in Figure 1a.

Modified Univformative Variable Elimination based on the Iterative Leave-One-Out Cross-validation (IVE-PLS). Below we describe a modified procedure for the Uniformative Variable Elimination (UVE-PLS). Instead, of a single step procedure we used here an iterative algorithm to find variables to be eliminated. To distinguish this procedure, we named this Modified Uniformative Variable Elimination with the iterative leave-one-out cross-validation **(**IVE-PLS). This procedure includes the following:

1. Standard PLS analysis applied to analyze the data yielded from Kohonen mapping with the leave-one-out cross-validation to estimate the performance of the PLS model (q^2),

2. The elimination of the matrix column of the lowest *b* value (Figure 1b),

3. Standard PLS analysis of the new matrix without the column eliminated in the step 2,

4. Iterative repetition of the steps 1−3 to maximize the LOO CV q^2 parameter.

The UVE procedures were programmed within the MATLAB environment (MATLAB). All MATLAB functions and m-scripts are available from the authors on request.

RESULTS AND DISCUSSION

Simulated Data for the Estimation of the UVE-PLS and IVE-PLS. We used the procedure reported by Centner[9] to simulate the noise free data matrix for the validation of the methods proposed. Thus,

1. An XS matrix of the (20,100) size is constructed from the random numbers,

2. PCA on the centered matrix XS is performed, yielding scores and loadings,

3. Multiplication of the first five score vectors (20,5) by the first five loading vectors (5,100) gives a simulated pure data matrix SIM (20,100), that does not contain any noise,

4. The Y vector of the answers is calculated as Y = 5*PC1 + 4*PC2 + 3*PC3 + 2*PC4 + 1*PC5; where PCs are vector of scores on PCs,

5. SIMUIN_20 incorporates the noise free data matrix SIM and additionally an uninformative variable matrix UIN. The SIMIUN_20(20,120) including matrix UIN(20,20) random numbers and the classical UVE and IVE procedures were performed in order to extract the columns that includes information needed to calculate the Y answers. The results obtained by the analysis of the SIMUIN_20 matrix are shown in Figure 2a.

6. The SIM matrix is supplemented by 200 columns SIMUIN_200(20,300) including random numbers and the classical UVE and IVE procedures were performed in order to extract the columns that includes information needed to calculate the Y answers. The results obtained are shown in Figure 2b.

Below we discuss the main conclusions for this part.

Clearly, the noise free SIM matrix allows us calculation of the Y answers with the 100% performance of $q^2 = 1$. The supplementation of the SIM by the further 20 columns still makes possible to extract (during both UVE-PLS and IVE-PLS) the relevant 100 columns from the SIMUIN_20 matrix with the q^2 parameter reaching a value $q^2 = 1$. This is shown in Figure 2a which illustrates the q^2 parameter as a function of the number of columns that have been eliminated during these procedures. In the IVE procedure this number corresponds to the number of iterations. Similar results were obtained during the analysis of the SIMUIN_200 matrix (Figure 2b). However, the analysis of the relevant columns surviving to be incorporated into the final model after the UVE-PLS and IVE-PLS (Figure 3) indicates that it is the IVE-PLS that provides better results, especially for the matrices to which more noisy columns were added (Figure 3b,c).

IVE-PLS for the Analysis of the Simulated Activity Data. For a better evaluation of the precision of variable elimination during CoMSA we performed the analysis of the simulated activity data. Therefore, a series of the comparative Kohonen maps (the size of 30 × 30) was simulated (Figure 4a). In the next step we arbitrarily supposed that the group of neurons selected within the maps (Figure 4b) describes the activity. Such a situation will describe a real ligand that interacts with the receptor only within a single

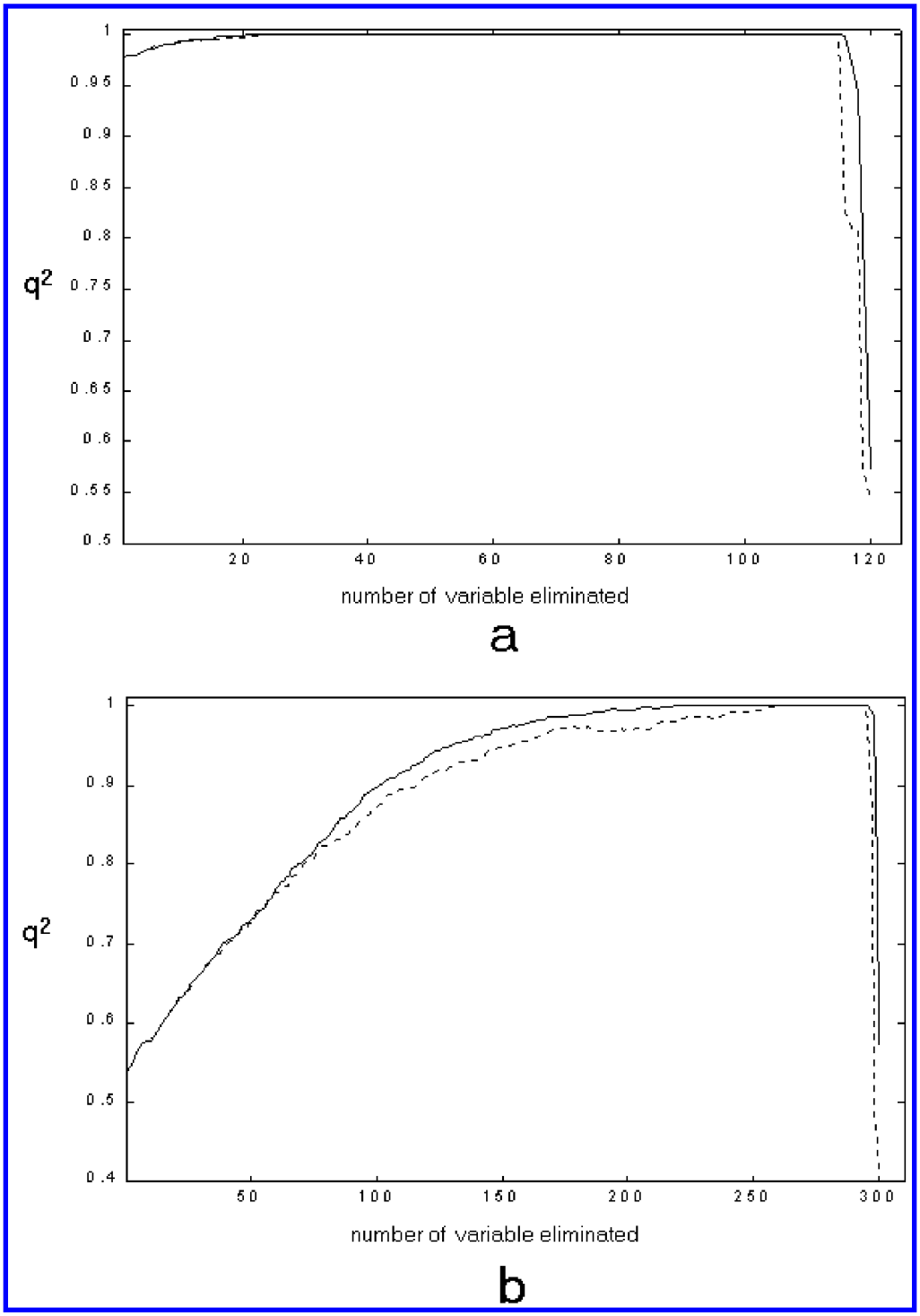


Figure 2. The relationship between the q^2 performance (LOO-*cross*-validation) and the number of the variable eliminated during the UVE-PLS (dotted line) and IVE-PLS (solid line) performed as illustrated in Figure 1, within a/ the SIM + 20 columns and b/ the SIM +200 columns.

part of the molecule. Then, by the selection of random coefficients we simulated the activity of the individual molecules as a polynomial function of the neurons selected in the first step. In the next step we used the UVE-PLS and IVE-PLS procedures to extract variables to be included in a final PLS model. Both procedures were performed to reach the optimal LOO CV (leave-one-out cross-validated) q^2. The results are shown in Figure 4 (parts c and d, respectively). It can be compared that during IVE-PLS more neurons survived that are originally used in Figure 4b for the simulation of the activity.

The Steroid Data. Table 2 compares the q^2 performance of the standard CoMSA (Table 1 − entry 4) procedure with that obtained in previous 3D QSAR investigations (Table 1 − entries 1−3). To perform a fair comparison we limit the number of the possible PLS components that can be included into models to 5. As can be observed, the standard CoMSA model is better than the CoMFA one. The CoMSA LOO CV q^2 variance reaches a value of 0.77 that compares well to the best q^2 value that was obtained by the use of the CoSA method.[9] Moreover, we compared the model obtained by the use of different modeling approaches and software available,

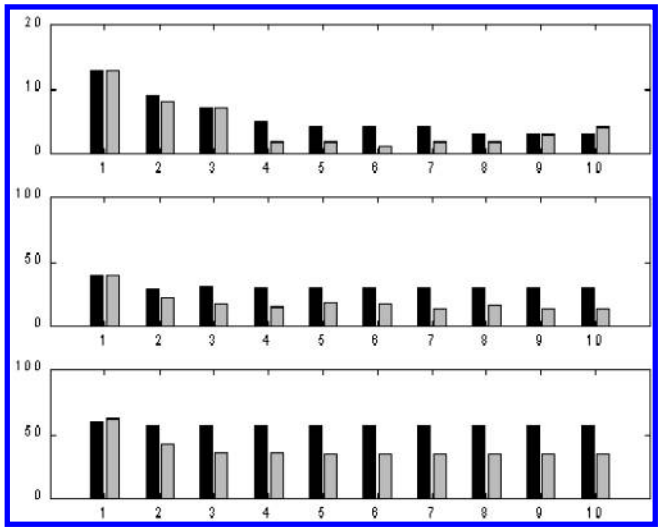


Figure 3. The number of the reliable columns improperly eliminated in the UVE-PLS (black bars) and IVE-PLS (gray bars) from the a/ SIM + 20 columns, a/ SIM + 200 columns, and a/ SIM +900 columns, respectively. The numbers (from 1 to 10) across the *x*-axis illustrate the number of the latent PLS components used.

e.g., AM1/ HYPERCHEM, (Table 2 − entry 5). Although the differences are not very large, this comparison shows that the best model results from Gasteiger's program package (Table 2 − entry 4).

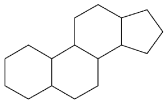
Figure 5 shows the relationship between the experimental activity and that predicted from model 6. The accurate LOO CV values are given in Table 1 (columns 9). The predictions for the compounds **6** and **13** are clearly worse than for the other steroids.

Although the elimination of variables is always risky, this can both improve the performance of PLS model and indicate

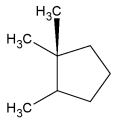
Table 2. Comparison of the Best Performance for the Models Reported in the Literature[9] and Those Obtained as the Result of CoMSA Analysis

no.	model	q^2	*s*	components used
1	CoMFA	0.72	*a*	5
2	QSAR	0.66	*a*	5 parameters
3	CoSA	0.77	*a*	5
4	CoMSA 1[b]	0.77	0.71	5
5	CoMSA 2[c]	0.74	0.72	1
6	CoMSA 3[d]	0.75	0.75	5
7	CoMSA 4[e]	0.96	0.31	5

[a] Not reported. [b] Gasteiger package: CORINA → PETRA → SURFACE → KMAP. Template (see below):



[c] HYPERCHEM (AM1) → SURF(MATLAB) → Kohonen SOM_PACK for MATLAB. Template (compound no. 1). [d] Gasteiger package. Template (see below):



[e] HYPERCHEM (AM1) → SURF(MATLAB) → Kohonen SOM_PACK for MATLAB. Template (compound no.1), →IVE-PLS.

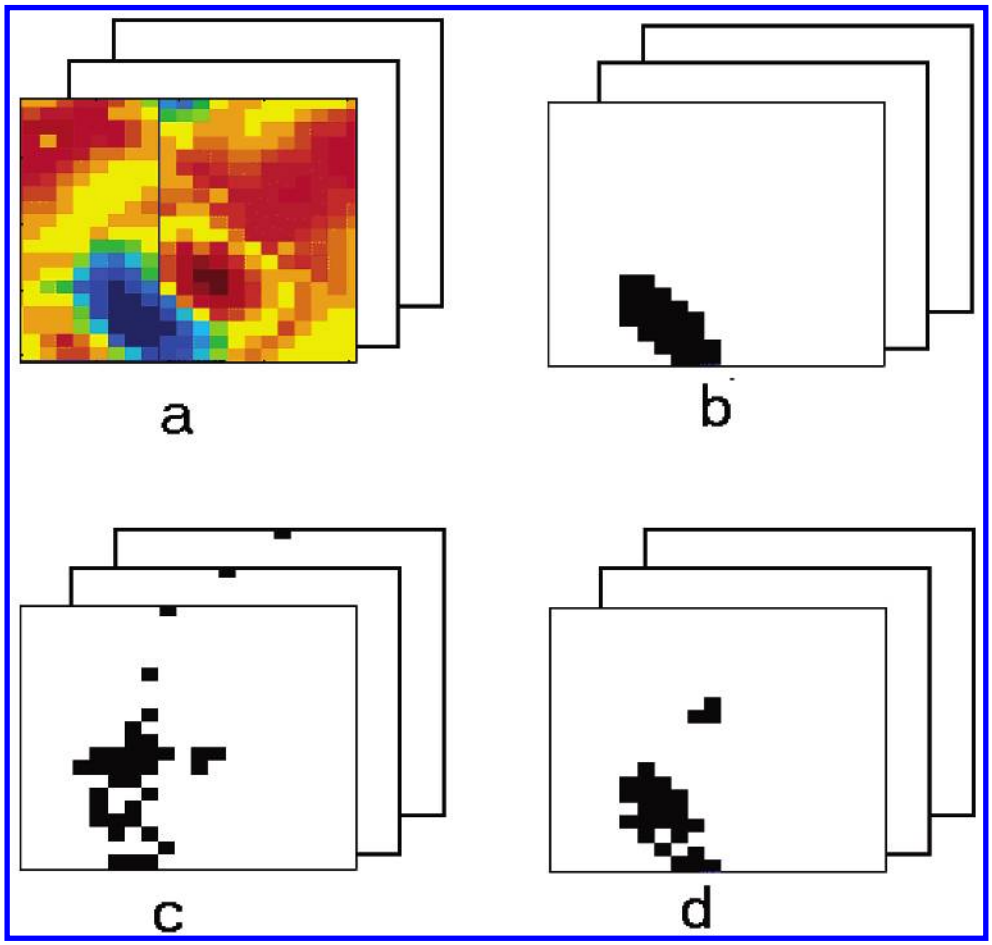


Figure 4. The interpretation of the results obtained during modeling of the Y answer generated by the series of molecules mapped by the comparative Kohonen maps a/. It is assumed that Y is a polynomial function of the arbitrarily selected neurons b/. The neurons extracted in the UVE-PLS c/ and IVE-PLS d/ procedures into a final, optimal PLS model.

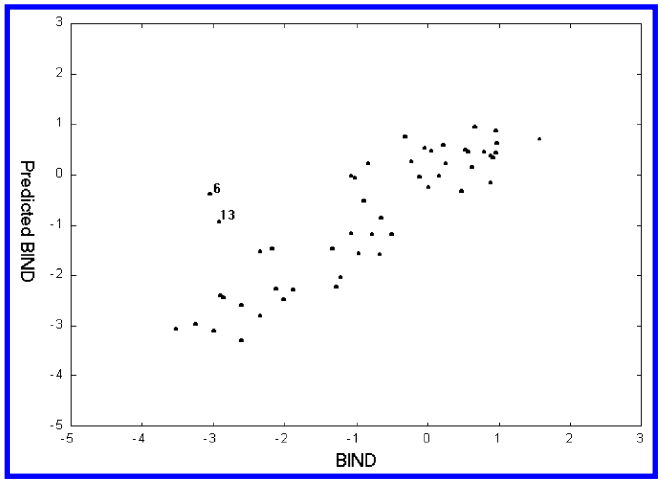


Figure 5. The 3D QSAR model of the aromatase binding steroids: CoMSA without variable elimination.

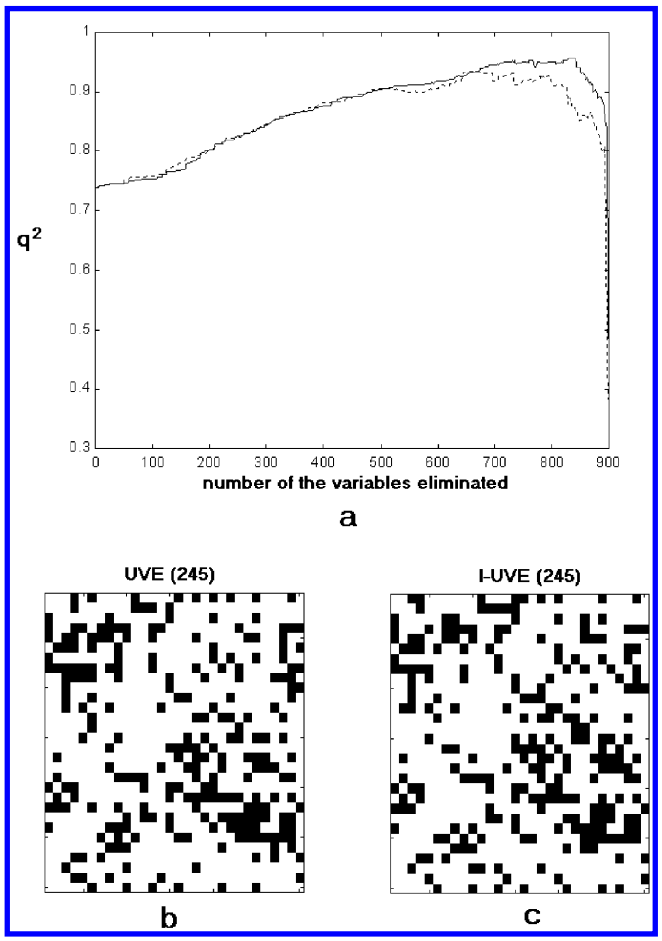


Figure 6. The results of CoMSA modeling of the aromatase binding activity of steroids: a/ the relationship between the q^2 performance (LOO-*cross*-validation) and the number of the variables eliminated during the CoMSA UVE-PLS (dotted line) and CoMSA IVE-PLS (solid line), b/ the neurons surviving after CoMSA UVE-PLS stopped at the q^2 value optimal for UVE-PLS (245 neurons) and c/ the neurons surviving after CoMSA IVE-PLS stopped at the q^2 value optimal for UVE-PLS (245 neurons), details in text.

such areas of the molecules that particularly contribute to the activity. Figure 6a compares the relationship between the q^2 performance during PLS-UVE (dotted line) and PLS-IVE (solid line) procedures and the number of the neuron columns eliminated. It can be compared that both methods allow improvements in the q^2 performance as much as to more than 0.91 ($s = 0.35$). Both methods give parallel results until about 650 eliminated neurons. The q^2 UVE-PLS reaches its maximum in this region. The neurons extracted for the final PLS models until this time are given in Figure 6b,c, respectively. As can be observed the patterns of neurons marked are very similar for both methods. Further elimination of variables by IVE-PLS can still improve the q^2 performance giving a value of 0.96 for the 839 neuron columns eliminated. Figure 7a illustrates the pattern of the neurons surviving the IVE-PLS elimination until its maximal q^2 value. Two main clusters of the neurons can be indicated within the pattern. These neurons, if back-projected onto the molecular surface, allow identification of two areas that influences mostly the binding activity (Figure 7b). One is indicated by the red color, corresponding to the atoms of the A and B steroid ring. This area was also indicated as the important one in the previous 3D QSAR models. However, the largest cluster (encircled with the blue color) corresponds to the atoms of the steroid ring D. This was not indicated in the previous models. Figure 7c shows the final CoMSA model, and the accurate cross-validated values can be found in Table 1 (column 10). It can be observed that now the activity of all compounds was predicted quite accurately (also for compounds **6** and **13** that are mispredicted by the previous model). The quality of the CoMSA IVE-PLS is characterized by the $q^2 = 0.96$ ($s = 0.31$), Table 2 − entry 7.

Because we observed that according to the model (Table 2 − entry 7) the steroid D ring is an important area determining the binding of the compounds, we simulated the additional CoMSA model using the template that visualizes only this area, i.e., the 1,2-dimethylcyclopentane (footnote *d* to the Table 2) was used as a template. In fact, we obtained a very predictive model (Table 2 − entry 6). It must be indicated here that the CoMSA method used includes both steric and electrostatic effects.

Moreover, for all the CoMSA models we tested the stability and reliability by testing the PRESS errors while performing the leave-two- (L-2-O), leave-three-out (L-3-O) cross-validations. The PRESS errors estimated now do not differ significantly from those estimated for the LOO method.

Estimation of the Model Predictivity. The experience of 3D-QSAR indicates that generally the q^2 performance cannot be a proof of the model predictivity. Golbraikh and Tropsha[26] have recently carefully analyzed this problem, postulating some further criteria for the predictive QSAR models. In the current work we used the standard deviation of errors of prediction (SDEP) for the estimation the model predictivity. Thus, we divided the steroids into two groups each consisting of 25 compounds. We distinguished these groups by even and odd numbers because such a separation enables placement of the molecules of the different stereochemical series into each group. Then, we used the first group as a training set to predict the activity for the second group. The results are shown in Figure 8. The accurate prediction values can be found in Table 1 (column 11). These values correspond quite well to the actual values of the binding constants.

Figure 8 illustrates the performance of the method assessed by the SDEP parameter. This value amounts to 0.782 for

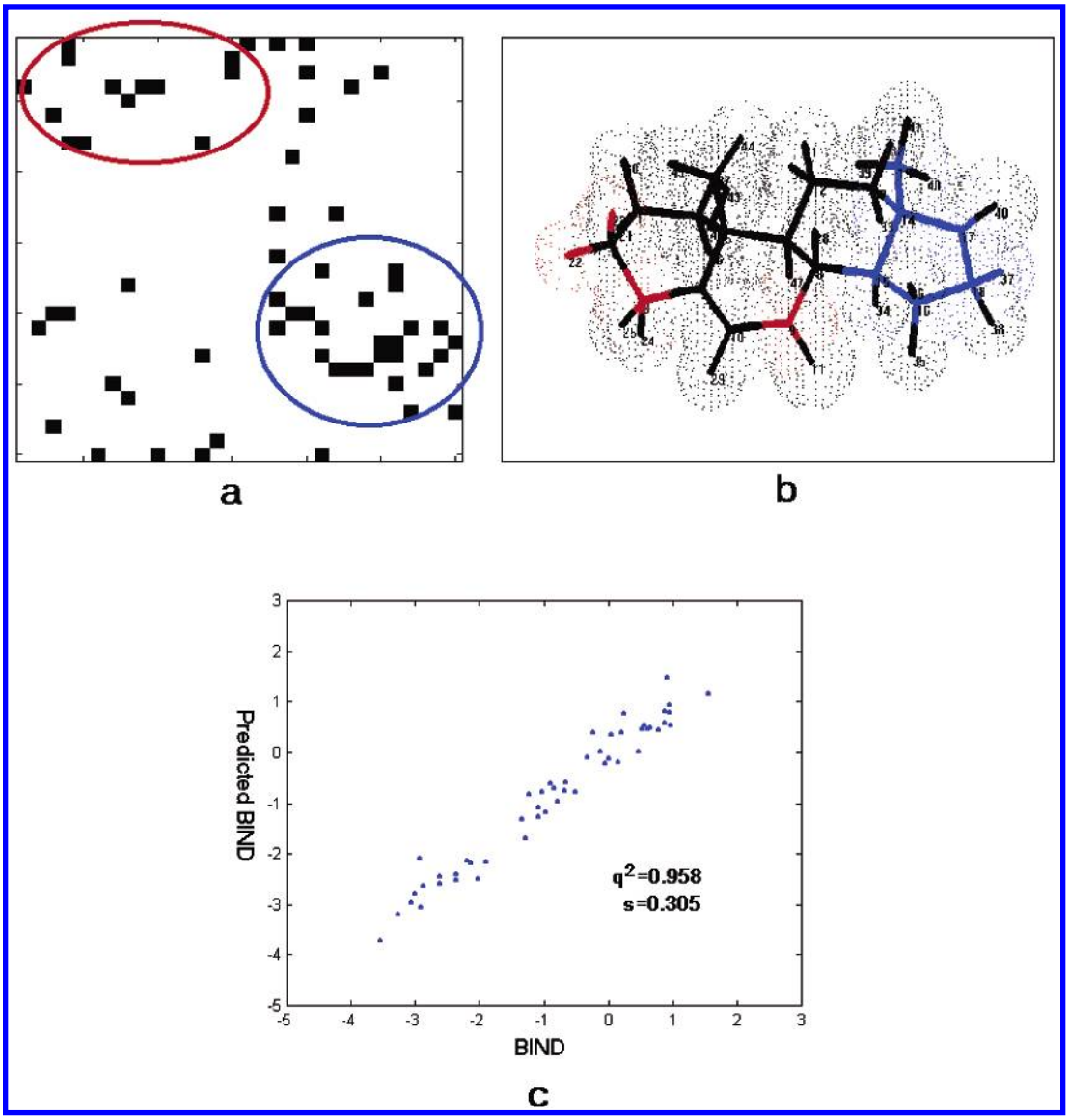


Figure 7. The results of the CoMSA modeling of the aromatase binding activity of steroids. The map a/ illustrates the columns surviving after CoMSA IVE-PLS stopped at the q^2 value optimal for IVE-PLS (61 neurons) and the molecular surface areas, b/ projected into two main clusters of neurons encircled by the line of the same color, respectively, and c/ CoMSA model obtained including only the neurons selected by IVE-PLS, details in text.

standard CoMSA (Figure 8a). Figure 8b illustrates the results obtained for the procedure including the IVE-PLS variable elimination. Thus, first we performed CoMSA involving the IVE-PLS for the whole set of molecules. Next the compounds, as previously, were divided into the training and test sets, and the variables that survived after previous elimination were used to obtain the QSAR model for the even numbered compounds in the CoMSA IVE-PLS. This model if applied for the prediction of the activity of the odd number compounds yields the activity values plotted in Figure 8b. This scheme provides significantly better predictions (SDEP amounts to 0.321). It should be, however, clearly indicated that this procedure does not guarantee a full separation of the modeling and prediction step, because the compounds present in the test set have also been used during the first IVE procedure. In fact, the SDEP value closely resembles this of the CoMSA4 model-describing s ($s = 0.31$) − entry 7 in Table 2. Therefore, to estimate reliable model predictivity we performed a series of further experiments controlling the SDEP parameter during the generation of the IVE-PLS-CoMSA models by the use of the similar procedure as described above. However, the compounds were divided into training and test sets (25/25 compounds) randomly. In the Supporting Information we record full data for such 50 random runs of IVE-PLS-CoMSA procedures. These results are summarized in Figure 9a,b which plots the relationships of the mean value of the SDEP, mean(SDEP), vs mean value of the q^2, mean(q^2), and SDEP standard deviation, std(SDEP), vs q^2 standard deviation, std(q^2), respectively. The most important conclusion is that, generally, IVE-PLS does not deteriorate the SDEP performance. In particular, the predictivity defined by the SDEP error does not depend on the q^2 value. It also appeared from the plot shown in Figure 9 b that a large region exists in which QSAR predictivity is very stable. This situation changes, however, dramatically after reaching the maximal q^2 value. Further data elimination results in the substantial deterioration of the model predictivity. We would like to indicate here that the results

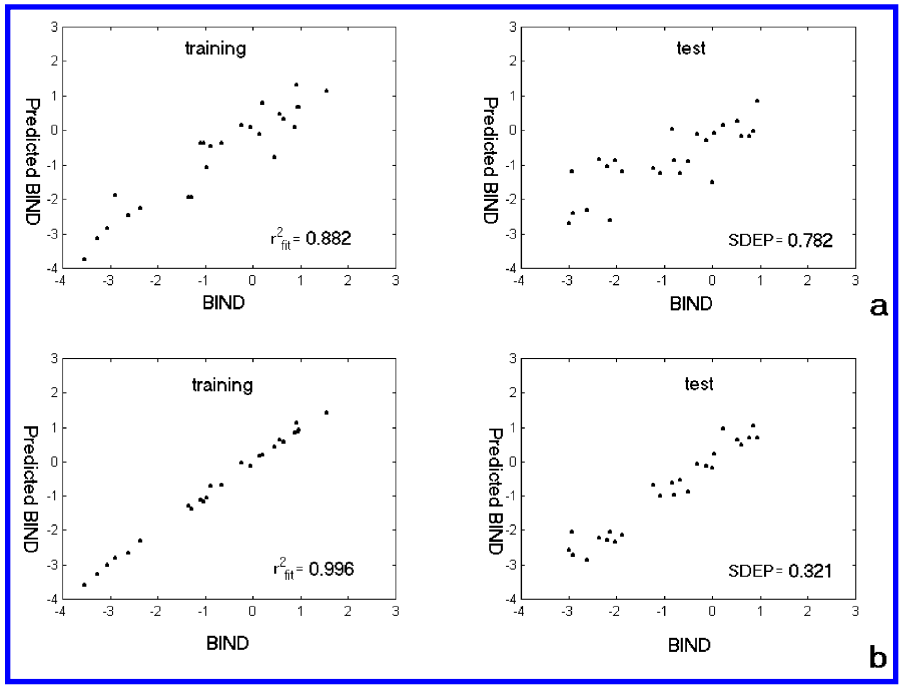


Figure 8. The estimation of the predictivity of the CoMSA models. The steroid series is divided into the training (25 molecules) and test (25 molecules) sets. Figure a/ shows the results of the CoMSA without variable elimination for the training and test sets, and b/ CoMSA with IVE-PLS procedure for the training and test sets, respectively, details in text.

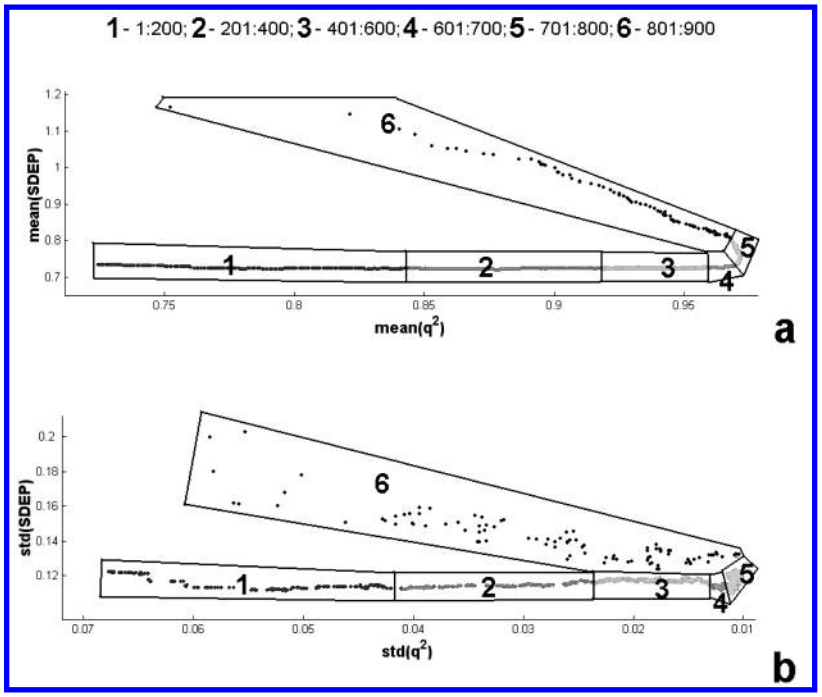


Figure 9. Validation of 50 random runs during variable elimination by IVE-PLS-CoMSA. a/ SDEP mean value vs q^2 mean value. b/ SDEP standard deviation vs q^2 standard deviation. The numbers indicate the ranges (from 1 to 6) of the variables eliminated.

illustrated in Figure 9 summarizes the performance for 44 950 different models. This number results from 50 random CoMSA runs with different variable sets (variable elimination started from 900 variables and in each iteration one variable was eliminated), that gives in total $50 \times 899 = 44\,950$ models.

Generally, the conclusion that external predictivity does not depend on a value of q^2 corresponds well with the results reported by Golbraikh and Tropsha, who proposed some new criteria for the model validation. Although we were quite satisfied with a fact that q^2 guided IVE-PLS variable elimination does not deteriorate SDEP error, we attempted to use these criteria for the validation of the predictivity of one among the best models among 50 analyzed in Figure 9. In this particular case the equation was modeled while using a following distribution of the compounds in the training/test set: training − 30, 1, 22, 39, 36, 20, 13, 50, 35, 26, 24, 49, 7, 21, 11, 42, 37, 5, 40, 4, 33, 44, 16, 19, 17; test − 28, 12, 43, 25, 15, 23, 31, 48, 45, 9, 2, 47, 27, 29, 10, 14, 18, 34, 3, 6, 38, 8, 32, 46, 41. The results are given in Figure 10. Similarly, to Golbraikh and Tropsha we were able to indicate the model of the high predictivity (SDEP = 0.428).

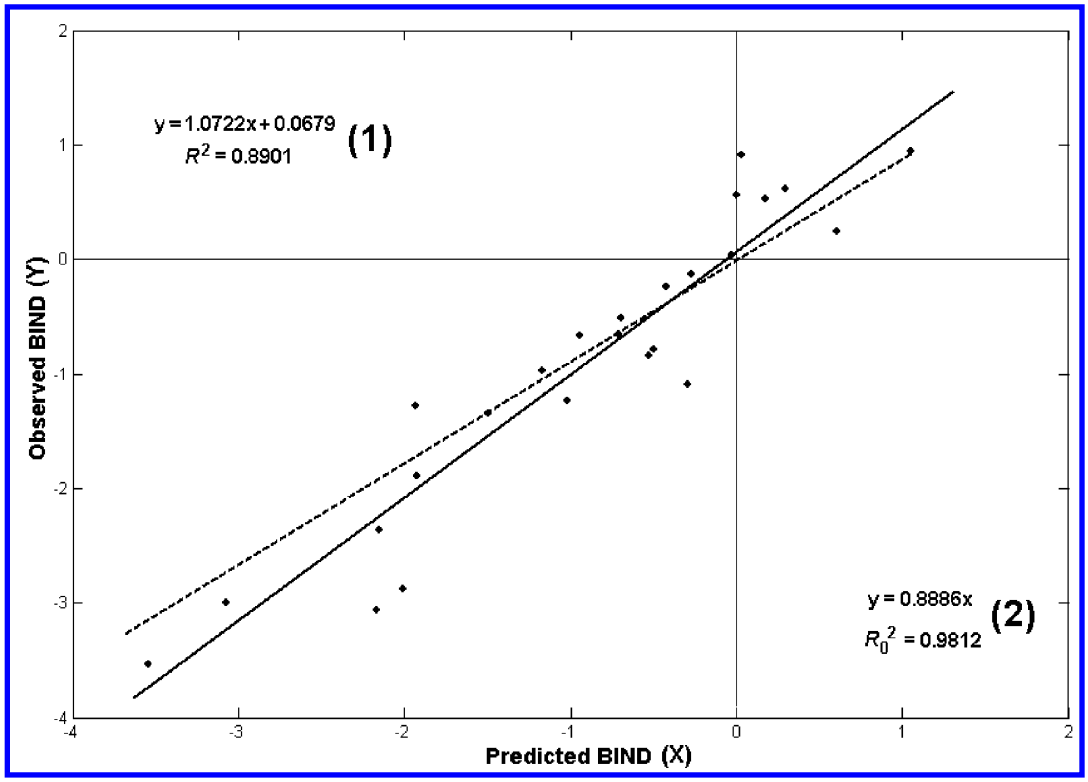


Figure 10. Validation of the best model by the Golbraikh-Tropsha criteria. We tried to keep the style of the presentation of the authors.[26] The regression between observed (Y) and predicted (X) activity values for the test set. The solid line shows the regressional equation given by (1). The dotted line illustrates the regression without the bias (2). The closest these linear plots the better model predictivity. Calculations after:

$$\mathrm{pred}_i^{\,0} = k^*\mathrm{pred}_i$$

$$k = \frac{\sum \mathrm{obs}_i \mathrm{pred}_i}{\sum \mathrm{pred}_i^{\,2}}$$

$$R_0^{\,2} = 1 - \frac{\sum(\mathrm{pred}_i - \mathrm{pred}_i^{\,0})^2}{\sum(\mathrm{pred}_i - \mathrm{mean(pred)})^2}$$

where upper index 0 relates to regression observed (Y) vs predicted (X), k is a slope of the regression through the origin (2), and R_0^2 is the correlation coefficient for the regression of observed (Y) vs predicted (X) without bias. $[(R^2 - R_0^2)/R^2] < 0.1$ and $0.85 \le k \le 1.15$ as recommended by Golbraikh and Tropsha.[26]

It is worth mentioning that this was achieved for the 50%/50% training/test set distribution.

CONCLUSIONS

The application of the CoMSA method to analyze 3D QSAR of 50 steroid aromatase inhibitors is described. The 3D QSAR model obtained, reaching a value of cross-validated $q^2 = 0.96$ ($s = 0.31$), significantly outperforms those reported in the literature for the CoMFA or CoSA (CoSASA). It is shown that the Uniformative Variable Elimination UVE-PLS or modified iterative UVE procedure (IVE-PLS) can be used for indicating the regions contributing to the binding activity. Thus, after separating the series into two groups of the training and test molecules quite correct external predictions result from the processing of the training set. We proved that the procedure of the data elimination provides stable results, if tested in 50 random runs of the IVE-PLS-CoMSA with different training/test sets. Depending upon the procedure used the quality of the predictions for 25 test molecules is given by SDEP $= (\Sigma(y_{\mathrm{pred}} - y_{\mathrm{obs}})^2/n)^{1/2} = 0.321 - 0.782$.

ACKNOWLEDGMENT

The authors thank Professor Johann Gasteiger of the University of Erlangen-Nürnberg, BRD both for his valuable discussion and for facilitating the programs of CORINA, PETRA, SURFACE, and KMAP. The financial support of the KBN Warsaw (grant numbers: PBZ KBN − 040 P04/08 and T08E02820) is gratefully acknowledged.

Supporting Information Available: Data concerning 50 random CoMSA runs. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Kubinyi, H. QSAR and 3D QSAR in drug design. Part 1: Methodology. *Drug Discovery Today* **1997**, *2*, 457−467.
(2) Kubinyi, H. QSAR and 3D QSAR in drug design. Part 2: Applications and problems. *Drug Discovery Today* **1997**, *2*, 538−546.
(3) Kubinyi, H. QSAR: Hansch analysis and related approaches. In *Methods and principles in medicinal chemistry*; Mannhold, R., Krokgsgaard-Larsen, P., Timmerman, H., Eds.; VCH: Weinheim, 1993.
(4) Katrizky, A. R.; Maran, U.; Lobanov, V. S.; Karelson, M. Structurally diverse quantitative structure − property relationship correlations of technologically relevant physical properties. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1−18.

(5) Polanski, J. The mapping of the molecular surfaces by means of self-organizing neural networks within MATLAB 5.2 for WINDOWS-95. *Acta Pol. Pharm.* **1999**, *56*, 80−84.
(6) Polanski, J.; Walczak, B. The comparative molecular surface analysis (CoMSA): a Novel tool for molecular design. *Comput. Chem.* **2000**, *24*, 615−625.
(7) Polanski, J.; Gieleciak, R.; Bak, A. The comparative Molecular Surface Analysis (CoMSA) − a nongrid 3D QSAR method by a coupled neural network and PLS system: Predicting pK_a values of benzoic and alkanoic acids. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 184−191,
(8) Beger, D. R.; Buzatu, A. D.; Wilkes, G. J.; Lay, O. J. ^{13}C NMR quantitative spectrometric data − activity relationship (QSDAR) models of steroids binding the aromatase enzyme. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1360−1366.
(9) Centner, V.; Massart, D. L.; de Noord, O. E.; de Jong, S.; Vandeginste, B. M. V.; Sterna, C. Elimination of uninformative variables for multivariate calibration. *Anal. Chim. Acta* **1996**, *330*, 1−17.
(10) Gerstein, M.; Greenbaum, D.; Luscombe, M. N. What is bioinformatics? A proposed definition and overview of the field. *Method Inform. Med.* **2001**, *40*, 346−358.
(11) Brazma, A.; Vilo, J. Gene expression data analysis. *FEBS Lett.* **2000**, *480*, 17−24.
(12) Toronen, P.; Kolehmainen, M.; Wong, G.; Castren, E. Analysis of gene expression data using self − organizing maps. *FEBS Lett.* **1999**, *451*, 142−146.
(13) Anzali, S.; Gasteiger, J.; Holzgrabe, U.; Polanski, J.; Teckentrup, A.; Wagener M. The use of self-organizing neural networks in drug design. *Perspect. Drug Discov. Design* **1998**, *9*/*10*/*11*, 273−299.
(14) Zupan, J. and Gasteiger, J. *Neural Networks and drug design for Chemists*, 2nd ed.; VCH: Weinheim, 1999.
(15) Gasteiger, J., Li, X., Rudolph, Ch., Sadowski J. and Zupan, J. The representation of molecular electrostatic potentials by topological feature maps. *J. Am. Chem. Soc.* **1994**, *116*, 4608−4620.
(16) Polanski, J. The receptor-like neural network for modeling corticosteroid and testosterone binding globulins. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 478−484.
(17) Anzali, S.; Barnickel, G.; Krug, M.; Sadowski, J.; Wagener, M.; Gasteiger, J. and Polanski, J. The comparison of geometric and electronic properties of molecular surfaces by neural networks: Application to the analysis of corticosteroid globulin activity of steroids. *J. Comput.-Aided Mol. Design* **1996**, *10*, 521−540.
(18) Polanski, J.; Gasteiger, J.; Jarzembek, K. Self − organizing neural networks for screening and development of novel artificial sweetener candidates. *Combin. Chem. High Throughput Screen.* **2000**, *3*, 481−495.
(19) Gasteiger, J. CORINA for the information see: http://www.mol-net.de.
(20) Sadowski, J.; Gasteiger, J. From atoms and bonds to three-dimensional atomic coordinates: Automatic model builders. *Chem. Rev.* **1993**, *93*, 2567−2581.
(21) Sadowski, J.; Gasteiger, J.; Klebe, G. Comparison of automatic three-dimensional model builders using 639 X-ray structures. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1000−1008.
(22) Gastaiger, J.; Saller, H. Calculation of the charge distribution in conjugated systems by a quantification of the resonance concept. *Angew. Chem.* **1985**, *97*, 699−701.
(23) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity − a rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219−3228.
(24) Kohonen, T. *Self-organization and associative memory*, 3rd ed.; Springer: Berlin, 1989.
(25) Gasteiger, J. Match3D; KMAP for the information see: http://www2.ccc.uni-erlangen.de.
(26) Golbraikh, A.; Thropsha, A. Beware of q^2 ! *J. Mol. Graph. Mod.* **2002**, *20*, 269−276.

CI020038Q