Modeling Robust QSAR 3: SOM-4D-QSAR with Iterative Variable Elimination IVE-PLS: Application to Steroid, Azo Dye, and Benzoic Acid Series

Andrzej Bak and Jaroslaw Polanski*

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

Received January 24, 2007

In the current paper we present a receptor-independent 4D-QSAR method based on self-organizing mapping (SOM-4D-QSAR) and in particular focus on its pharmacophore mapping ability. We use a novel stochastic procedure to verify the predictive ability of the method for a large population of 4D-QSAR models generated. This systematic study was conducted on a series of benzoic acids, azo dyes, and steroids that bind aromatase. We show that the 4D-QSAR method coupled with IVE-PLS provides a very stable and predictive modeling technique. The method enables us to identify the molecular motifs contributing the most to the fiber-dye affinity and the aromatase enzyme binding activity of the steroid. However, the method appeared much less effective for the benzoic acid series, in which the efficacy was limited by electronic effects strictly correlated to a single conformer.

1. INTRODUCTION

In essence, properties of a compound, and not the compound itself, are the focus of chemistry. However, available technologies do not provide us with an obvious and efficient method for property design. Thus, we usually need to synthesize a large number of compounds to find a single one having the required qualities. For example, Merck researchers had to investigate more than 250 000 structures to find the diketoacid-like HIV integrase inhibitors.[1] Despite the application of sophisticated drug design tools, these compounds are still far from being approved to appear on the market. The quantitative structure−activity relationships technique is a strategy that employs a search of the function relating molecular properties to their structures on the basis of a comparison of the molecular structures themselves. Thus, the compounds obtained during a more or less intuitive property screening of the compound libraries that form the so-called chemical compound space can be used for modeling QSAR. Multidimensional QSAR (m-QSAR) is a variation that samples data from the modeled 3D molecular structures. In fact, a variety of methods have evolved from this concept. 3D- and 4D-QSAR focus exclusively on receptor ligands. The expansion of dimension allows us to not only analyze the conformational profile of the ligand (4D-QSAR) but also to select the apparent receptor-induced fit mode (5D-QSAR) and consider different solvation models (6D-QSAR).[2−4]

A variety of issues decide the efficacy of m-QSAR methods, and their practical importance for drug design is still controversial. What is the place of m-QSAR, and in particular 3D- and 4D-QSAR, in the current drug design landscape? Figure 1 defines the QSAR problem using the concept of chemical space (CS).[5] This space can be divided into the factual (FCS) and virtual (VCS) spaces, where FCS defines the portion of chemical compounds that has already been synthesized and described. A 3D-QSAR model relating

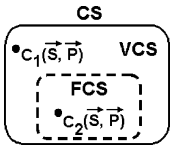


Figure 1. The concept of the chemical space CS formed by virtual (VCS) and factual (FCS) subspaces (CS = VCS ∪ FCS) which are occupied by molecular objects described by structural S and physicochemical P property vectors.

activity to structure gives the illusion that we can easily extend this relation to predict the activity of novel compounds in VCS. In fact, the predictive power of m-QSAR is generally discussed in m-QSAR. However, this parameter is a measure of statistical reliability of the respective model and rely only to the simulated within FCS. In fact, a real drug design approach that transforms m-QSAR data into drugs, for example by coupling QSAR with the design of a VCS molecule, is practically unprecedented in the field. Thus, quantitative VCS prediction is not the goal of current state of the art research on m-QSAR. Instead, we use m-QSAR for visualization of the spatial SAR requirements. The resulting pattern should be a valuable drug design tool that indicates promising areas for synthetic modification and/or molecular areas apparently involved in biological interactions.

However, this operation is also a complex issue.[6,7] From the technical point of view, during m-QSAR the complex multidimensional data are generated. Thus, we are to relate a single activity value (Y block data) to a large number of molecular descriptors (X block data) coding a 3D molecular structure having a data for a relatively small number of molecules. Actually, this cannot be done directly by a standard regression. Instead, we generally use the PLS (Partial Least-Squares) method for modeling a final equation. Essentially, this method is a mathematical trick that allows us to generate novel variables, the so-called latent PLS components, that are a linear combination of the original ones formed in such a way that the portions of the individual

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

10.1021/ci700025m CCC: $37.00
Published on Web 06/14/2007

original X block variables are optimized to provide the highest description of the Y (activity) block data variance. This enables the replacement of the original set of, let us say, several hundred or even several thousand X block variables by a few PLS components that, in some sense, *compress* the initial data preserving however a level of their influence. QSAR visualization needs to eliminate or at least sort according to their importance the contribution of the original variables into the PLS model. This should make it possible to reveal a few spatial moieties contributing to the activity. In this respect, we have described recently the application of the iterative variable elimination (IVE-PLS), a version of the uninformative variable elimination (UVE-PLS) [8] in 3D-QSAR.[9]

The main objective of this publication is to test the applicability of the SOM-4D-QSAR method coupled with the IVE-PLS method. We should be aware that the applied 4D-QSAR method does not take into account the influence of the receptor on ligand or ligand solvation, the effects incorporated in 5D- and 6D-QSAR modeling. However, in our investigations we used three different series of compounds interacting with the biological or chemical environment in modes that do not always allow us to take into account these effects. First, typical steroid drugs interact with a biological receptor. Second, it is believed that tincturing properties of dyes depend on an apparent receptor-like cavity on the cellulose surface. This receptor-like cavity differs, however, from a typical drug receptor. Third, while modeling the chemical behavior of a series of benzoic acids, no receptor-like moiety can be identified. Through a comparison of these different modes of interactions, we would like to show that the method can be used for modeling and visualization of receptor-independent QSAR.

2. EXPERIMENTAL SECTION

2.1. Self-Organizing Maps (SOM). A self-organizing neural architecture is an unsupervised method based on Kohonen neural network.[10] A variety of SOM applications in drug design has been described recently, which also included pharmacophore mapping.[11] SOM mapping is a nonlinear projection tool which reduces the dimensionality of the input object, e.g., converts 3D objects to 2D, while maintaining the topological relationships between the input and output data. Thus, the method can be applied to compare molecular surfaces and their properties (i.e., electrostatic potential) or molecular representations with atomic partial charges.[12] In such application each 3-dimensional input vector is compared to a 3-element weight vector describing each neuron to find the closest neighbor and then project a signal into this particular neuron.

Formally, the winning neuron ($\mathrm{out_c}$) is selected by the optimization of the Euclidean distance between a vector (x) and a weight (w):

$$\mathrm{out_c} = \min\left[\sum_{i=1}^{m} (x_i - w_{ij})^2\right] \tag{1}$$

The weights of the winning neuron and the neighboring neurons are then adjusted to attract similar input vectors. A SOM of single molecular representation can form a template that if used for the processing of the other molecular data provides a series of the comparative SOM maps. We have described recently a series of QSAR methods based on such comparative Kohonen strategy. This includes CoMSA (Comparative Molecular Surface Analysis)[13] and SOM-4D-QSAR.[14]

In practice, we applied the KMAP 3.0[15] for the simulations of 20×20 or 30×30 SOMs. The output maps were then transformed to a 400 or 900 element vector, respectively. The obtained vectors were processed by the PLS analysis.

2.2. PLS Analysis. The PLS model was constructed for the centered data, and its complexity was estimated using the leave-one-out (LOO) cross-validation procedure (CV). In the LOO-CV one repeats the calibration m times, each time treating the ith left-out object as the prediction object.[16] The dependent variable for each left-out object is calculated on the basis of the model with one, two, three, etc. factors. The root-mean-square error of CV for the model with j factors is defined as

$$\mathrm{RMSECV}_j = \sqrt{\frac{\sum_{i}^{m} (\mathrm{obs}_i - \mathrm{pred}_{ij})^2}{m}} \tag{2}$$

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

Similarly to the other m-QSAR analyses, we used a cross-validated leave-one-out q^2_{CV} value for the estimation of the model performance

$$q^2_{\mathrm{CV}} = 1 - \frac{\sum_{i}^{m} (\mathrm{obs}_i - \mathrm{pred}_i)^2}{\sum_{i}^{m} (\mathrm{obs}_i - \mathrm{mean}(\mathrm{abs}_i))^2} \tag{3}$$

where obs is the assayed value; 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. The cross-validated standard error of prediction s is as follows

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

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

The quality of external predictions was measured by the standard deviation of error of prediction (SDEP) or q^2_{test} defined as

$$\mathrm{SDEP} = \sqrt{\frac{\sum_{i}^{n} (\mathrm{pred}_i - \mathrm{obs}_i)^2}{n}} \tag{5}$$

$$q^2_{\text{test}} = 1 - \frac{\sum_i^n (\text{obs}_i - \text{pred}_i)^2}{\sum_i^n (\text{obs}_i - \text{mean}(\text{obs}_i))^2} \quad (6)$$

where n is the number of objects in test set.

2.3. Variable Elimination. Although variable elimination in PLS modeling is a complex problem, several novel methods have been described recently.[9,17] In our previous paper we have reported that uninformative variable elimination (UVE-PLS) as well as its modifications, namely modified UVE (m-UVE) and iterative variable elimination (IVE-PLS), can be successfully applied both in 3D- and 4D-QSAR schemes. The above algorithms allowed us not only to increase the predictive ability of the standard PLS procedure but also to identify pharmacophoric elements or the apparent molecular areas important for the interactions with biological receptor or enzymes. In the current calculations we used IVE-PLS that is a modification of the UVE algorithm originally proposed by Centner.[8] This method is based on the analysis of the regression coefficient calculated by the PLS method. PLS modeling is given by the relation between the variable Y and a set of predictors X in the form given by equation

$$Y = Xb + e \quad (7)$$

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

The UVE and IVE variable elimination methods are based on the estimation of the so-called stability parameter given by a mean(b)/std(b) ratio, where std(b) is the standard deviation of the model weights b. Then, only variables of the relatively high stability values are included in the final PLS model. Instead of a single-step UVE procedure, we used here an iterative IVE-PLS algorithm based on the stability criterion. This was described in details in our previous publications.[9] Generally, this procedure includes the following:

Step 1. Standard PLS analysis applied to analyze the matrices yielded from SOM-4D-QSAR procedure with LOO-CV to evaluate the performance of the PLS model (q^2_{CV})

Step 2. Elimination of the matrix column of the lowest abs(mean(b)/std(b)) value

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

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

This was programmed in the MATLAB environment. All functions and m-scripts are available from the authors on request.

2.4. SOM-4D-QSAR Analysis. The 4D-QSAR approach was developed by Hopfinger to explore molecular objects optimized by molecular dynamics.[18] The 4D-QSAR method incorporates conformational freedom into the development of 3D-QSAR models by considering the multiple conformational states of a ligand molecule. The method has been successfully applied to a variety of molecular series and is especially well suited for the search of the active conformation and binding mode of the conformationally flexible molecules. Recently the method has also been extended to include the receptor data which is described as receptor dependent (RD) 4D-QSAR.[19] The SOM version of the 4D-QSAR has also been developed.[20] The current publication describes the application of the receptor independent SOM-4D-QSAR which algorithm includes 6 steps:

Step 1. Model Building. The initial step is analogous to the starting point of each 3D-QSAR analysis—a 3D structure of each molecule in the training set is generated. Although, in general, any 3D structure can start conformational ensemble sampling, in practice, each analog is energy-minimized using the Sybyl/Tripos 7.1 software.[21]

Step 2. Superimposition. The next step is the selection of the trial alignment of the molecule. This is usually performed by selecting in individual molecules 3 atoms to be covered.

Step 3. Interaction Pharmacophore Elements (IPE). The objective of this step is the partition of the molecules into groups of atoms apparently playing a privileged role in the interactions modeled, e.g., aromatic, hydrogen bond donors, hydrogen bond acceptors, polar positive or negative partial charge and unrestricted (all) atom type.

Step 4. Conformational Ensemble Profile (CEP). Molecular dynamic simulations (MDs) provide conformers for further comparative analysis. The energy-minimized structures were used as the initial step to create CEP of each compound. The partial atomic charges were calculated using the AM1 method implemented in the HyperChem 5.0 package.[22]

Step 5. Comparative Kohonen Mapping. Each of the Cartesian coordinates and partial atomic charges are then used to construct a two-dimensional SOM map. During training these data are distributed among neurons providing the sum_occupancy or mean_charge maps, respectively.

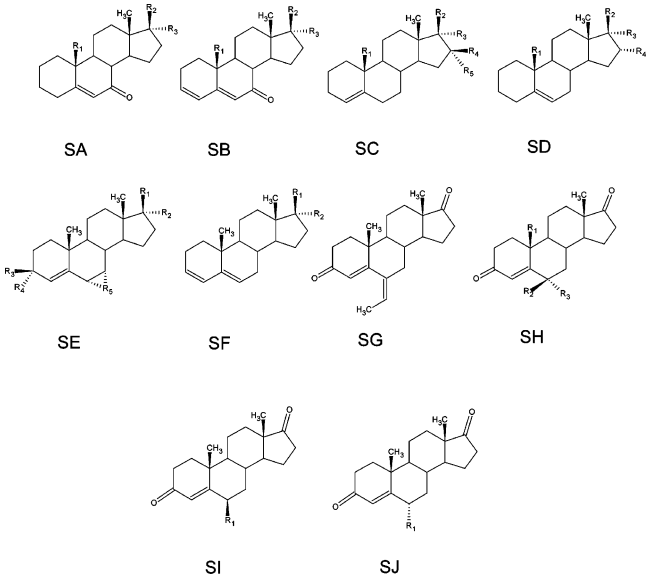
Step 6. Data Reduction and Model Validation. In a final step a 4D-QSAR relationship is modeled using the PLS algorithm with the LOO CV procedure supported by the IVE-PLS scheme. Model validation performed by LOO CV is additionally monitored by measuring predictive ability for the external test set as described in section 2.3. A variety of test/training sets samplings has been monitored by the iterative Stochastic Model Validation (SMV) scheme as described in detail in ref 23.

2.5. Model Builder. The chemical structures and experimental data, i.e., for the steroids binding the aromatase enzyme, azo dyes, and benzoic acids are extracted from refs 9, 24, and 25 and are introduced in Tables 1−3, respectively.

2.6. Molecular Modeling. All modeling studies were performed using the Sybyl/Tripos 7.1 software package running on Pentium IV 2.8 with the Fedora Core 4 operating system. The initial geometry was optimized using a standard Tripos force field (POWELL method) with 0.005 kcal/mol energy gradient convergence criterion and a distant dependent dielectric constant. Partial atomic charges were calculated using the Gasteiger−Marsili method implemented in Sybyl.

Energy-minimized molecules were used as the initial structures in the molecular dynamic simulations (MDs) with the standard Tripos force field. Each 3D structure is a starting point in generating the conformational ensemble profile (CEP). The CEP is generated from a MDs run of 100 ps generated at intervals of 0.001 ps time step. The temperature for the MDs was normally set to 300 K. The atomic

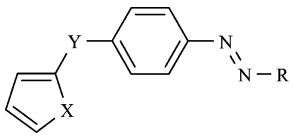
Table 1. Steroid Structures and the Binding Data



no.	structure	R_1	R_2	R_3	R_4	R_5	BIND[a]
1s	SA	CH_2OH	=O				−2.92
2s	SA	CH_2OH	OH	H			−3.54
3s	SA	CHO	=O				−3.00
4s	SA	H	=O				−3.26
5s	SA	Me	OH	H			−2.62
6s	SB	CH_2OH	=O				−3.06
7s	SB	CHO	=O				−2.14
8s	SB	H	=O				−2.36
9s	SD	CH_2OH	=O		H		−1.89
10s	SD	CH_2OH	OH	H	H		−2.88
11s	SD	CHO	=O		H		−2.03
12s	SD	Me	=O		H		−0.97
13s	SD	Me	=O		Br		−2.93
14s	SA	Me	=O				−1.28
15s	SB	Me	=O				−1.23
16s	SB	Me	OH	H			−2.61
17s	SD	Me	OH	H	H		−2.36
18s	SF	=O					−0.65
19s	SF	OH	H				−2.19
20s	SH	H	H	H			−1.03
21s	SC	Me	=O		H	H	0.00
22s	SC	CH_2OH	=O		H	H	0.46
23s	SH	CH_2OH	H	H			−0.84
24s	SH	Me	=O				0.15
25s	SE	=O		=O		CF_2	−0.13
26s	SE	=O		H	H	CH_2	0.87
27s	SE	OH	H	H	H	CH_2	−0.51
28s	SC	Me	OH	H	H	H	−1.35
29s	SC	CH_2OH	OH	H	H	H	−0.67
30s	SC	$MeC(O)OCH_2$	=O		H	H	−0.89
31s	SC	Me	=O		H	Br	−0.79
32s	SC	Me	=O		H	H	−1.09
33s	SC	CF_3	=O		H	H	−1.08
34s	SI	Me					0.56
35s	SJ	Me					0.87
36s	SI	C_2H_5					1.56
37s	SJ	C_2H_5					0.94
38s	SI	C_3H_7					0.94
39s	SJ	C_3H_7					0.78
40s	SI	C_4H_9					0.65
41s	SJ	C_4H_9					0.53
42s	SI	$CH(CH_3)_2$					0.21
43s	SJ	$CH(CH_3)_2$					0.04
44s	SI	C_6H_5					−0.04
45s	SJ	C_6H_5					0.24
46s	SI	$CH_2C_6H_5$					−0.24
47s	SJ	$CH_2C_6H_5$					0.61
48s	SI	$CH=CH_2$					0.91
49s	SI	C≡CH					−0.32
50s	SG						0.96

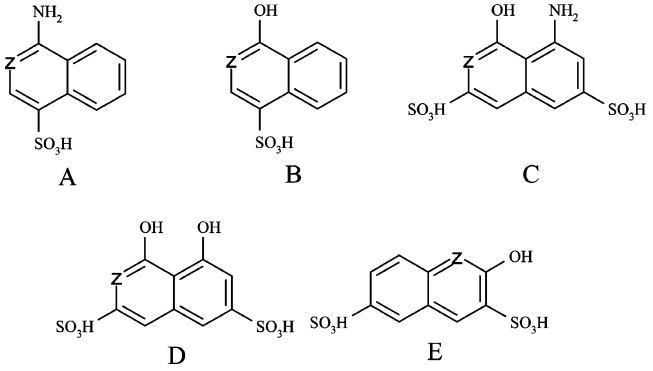
[a] Experimental binding data according to ref 9.

Table 2. Experimental Dye Affinity



no.	X	Y	R	$-\Delta\mu^{\circ}$ [a] (kJ/mol)
1a	-S-	-CH=CH-	A	15.80
2a	-CH=CH-	-CH=CH-	A	14.25
3a	-S-	-CONH-	A	13.08
4a	-CH=CH-	-CONH-	A	12.00
5a	-S-	-CH=CH-	B	9.66
6a	-S-	-CH=CH-	C	9.45
7a	-CH=CH-	-CH=CH-	B	9.20
8a	-S-	-CONH-	C	9.03
9a	-S-	-CO-	A	8.78
10a	-CH=CH-	-CH=CH-	C	8.40
11a	-CH=CH-	-CONH-	C	8.28
12a	-S-	-CONH-	B	7.15
13a	-S-	-CH=CH-	D	7.06
14a	-CH=CH-	-CO-	A	7.02
15a	-CH=CH-	-CONH-	B	6.52
16a	-S-	-CH=CH-	E	6.27
17a	-S-	-CONH-	D	6.23
18a	-CH=CH-	-CH=CH-	D	6.02
19a	-CH=CH-	-CH=CH-	E	5.81
20a	-CH=CH-	-CONH-	D	5.18
21a	-S-	-CONH-	E	5.10
22a	-S-	-CO-	C	4.64
23a	-CH=CH-	-CONH-	E	4.26
24a	-S-	-CO-	B	4.22
25a	-CH=CH-	-CO-	C	4.10
26a	-CH=CH-	-CO-	B	4.05
27a	-S-	-CO-	D	3.85
28a	-CH=CH-	-CO-	D	3.43
29a	-S-	-CO-	E	3.22
30a	-CH=CH-	-CO-	E	2.84

[a] Experimental affinities according to ref 24. Z − attachment point.



coordinates of each conformation and its total energy were recorded every 0.1 ps. One thousand conformations were sampled for each analogue that resulted in CEP of 100 000 trajectory states. Partial atomic charges were calculated using the semiempirical AM1 Hamiltonian implemented in HyperChem software. Before processing by neural network, the molecules were arbitrarily superimposed according to different 3-atom alignment by covering the atoms as illustrated in Figure 2. We used the most active analog in each series of compounds (**36s**, **1a**, and **48b**, respectively) to form the template molecule training the 20 × 20 SOM network.

3. RESULTS AND DISCUSSION

3.1. Steroids Binding the Aromatase Enzyme. Steroids, an important class of natural compounds, have appeared as

Table 3. Hammett Constants for Substituted Benzoic Acid

no.	substituent	Hammett constants	no.	substituent	Hammett constants
1b	H	0.00	**37b**	p-OCH_3	−0.27
2b	m-Br	0.39	**38b**	p-SH	0.15
3b	m-CF_3	0.54	**39b**	p-SCH_3	0.00
4b	m-CH_3	−0.07	**40b**	p-SCF_3	0.50
5b	m-Cl	0.37	**41b**	p-$C(CH_3)_3$	−0.20
6b	m-CN	0.56	**42b**	p-C_2F_5	0.52
7b	m-F	0.34	**43b**	p-CH_2Br	0.14
8b	m-I	0.35	**44b**	p-CH_2Cl	0.12
9b	m-NH_2	−0.16	**45b**	p-CH_2I	0.11
10b	m-NO_2	0.71	**46b**	p-C_2H_5	−0.15
11b	m-OCF_3	0.38	**47b**	p-SO_2CF_3	0.93
12b	m-OH	0.12	**48b**	p-SO_3F	0.91
13b	m-OCH_3	0.12	**49b**	p-SO_2CH_3	0.72
14b	m-SH	0.25	**50b**	m-$CH{=}CH_2$	0.05
15b	m-SCH_3	0.15	**51b**	m-CH_2CN	0.16
16b	m-SCF_3	0.40	**52b**	m-CHO	0.35
17b	m-$C(CH_3)_3$	−0.10	**53b**	m-CH_2OCH_3	0.02
18b	m-C_2F_5	0.47	**54b**	m-$COCH_3$	0.38
19b	m-CH_2Br	0.12	**55b**	m-$CONH_2$	0.28
20b	m-CH_2Cl	0.11	**56b**	m-NCS	0.48
21b	m-CH_2I	0.10	**57b**	m-$NHCH_3$	−0.30
22b	m-C_2H_5	−0.07	**58b**	m-$N(CH_3)_2$	−0.15
23b	m-SO_2CF_3	0.79	**59b**	m-$OCOCH_3$	0.39
24b	m-SO_2F	0.80	**60b**	m-SCN	0.41
25b	m-SO_2CH_3	0.60	**61b**	m-SO_2NH_2	−0.02
26b	p-Br	0.23	**62b**	p-$CH{=}CH_2$	0.01
27b	p-CF_3	0.54	**63b**	p-CH_2CN	0.01
28b	p-CH_3	−0.17	**64b**	p-CHO	0.42
29b	p-Cl	0.23	**65b**	p-CH_2OCH_3	0.03
30b	p-CN	0.66	**66b**	p-$COCH_3$	0.50
31b	p-F	0.06	**67b**	p-$CONH_2$	0.36
32b	p-I	0.18	**68b**	p-NCS	0.38
33b	p-NH_2	−0.66	**69b**	p-$NHCH_3$	−0.84
34b	p-NO_2	0.78	**70b**	p-$N(CH_3)_2$	−0.83
35b	p-OCF_3	0.35	**71b**	p-SCN	0.52
36b	p-OH	−0.37	**72b**	p-SO_2NH_2	0.57

[a] Experimental Hammett constants according to ref 25.

an interesting target for drug design methods due to their broad spectrum of biological interactions and rigid chemical structure, which facilitates m-QSAR comparison of the molecules. A series of CBG and TBG complexing steroids was used during development of the CoMFA method.[26] Recently, we demonstrated that 4D-QSAR can also be useful in investigations of these rigid molecules.[27] The group of steroids that binds the aromatase enzyme has become another benchmark series used for testing new methods aimed at 3D-QSAR modeling. Aromatase belongs to the superfamily of cytochrome P450 hemeproteins whose biological function consists of androgen aromatization, which produces estrogens. Modeling ligand−receptor interactions in cytochromes presents many problems due to the influence of the charge-transfer phenomena thermodynamics and kinetics of this process. An example of a sophisticated m-QSAR approach to the prediction of the small molecules binding to CYP P450 3A4 can be found in ref 28. Aromatase can be important in the development of some types of tumors. In particular, it has been shown that inhibitors of this enzyme can be effective in advanced breast cancer in postmenopausal women.[29] Below we describe a SOM-4D-QSAR model for a series of steroids. As in our previous studies, steroids were divided into two series, namely, a training set (**2s, 4s, 6s,..., 50s**) and a test set (**1s, 3s, 5s, ..., 49s**), each including 25 compounds, respectively. SOM-4D-QSAR analysis was then performed as described in the Experimental Section.

The key idea behind IVE-PLS is that the number of PLS latent variables, referred to as model complexity (A_{opt}), is forced to take the lowest possible value in this algorithm. Thus, during the search for the optimal IVE-PLS model complexity, the value of A_{opt} is always set not to exceed a certain value. In this procedure, for example, if the optimal complexity is 4, and we are truncating at 2, we assume 2 as the final A_{opt} value. Although such a procedure formally decreases the q^2_{CV} predictive ability, it provides more stable QSAR visualization.[30]

Table 4 reports the predictive power of some selected 4D-QSAR and SOM-4D-QSAR models as measured by the LOO-CV procedure and evaluated further by their prediction ability for the external test set. Entries 1−4 were obtained by separating the series of **1s**−**50s** into two groups: the training (even numbers) and test molecules (odd numbers), whereas entries 5−8 report modeling results for the whole compound data set. The best model is given by entry 2: $q^2_{CV} = 0.74$, $s = 0.82$, $A_{opt} = 3$. This compares well to the corresponding CoMFA ($q^2_{CV} = 0.72$, $s = 0.77$, $A_{opt} = 4$)[29]

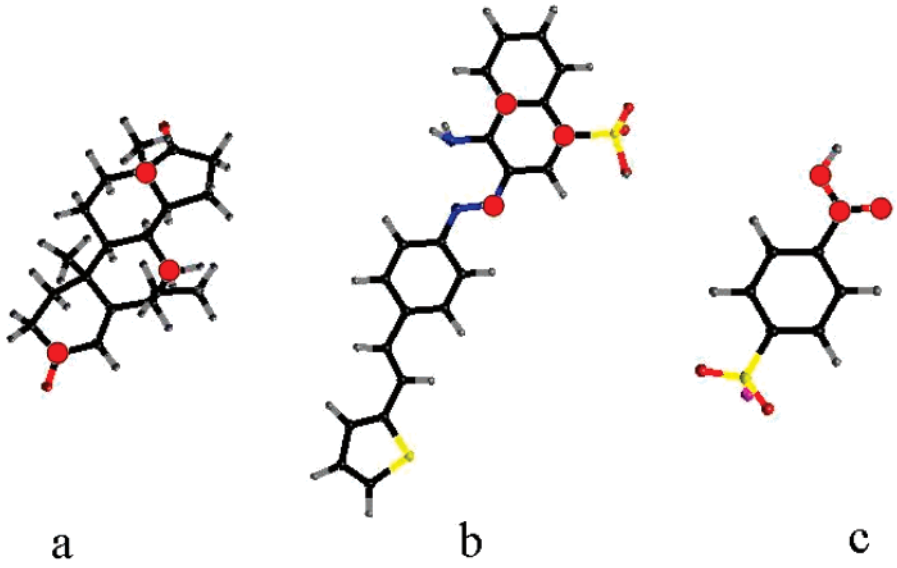


Figure 2. The molecular series analyzed with the indicated superimposed atoms.

Table 4. Comparison of 4D-QSAR Models for the Steroid Analogues

entry	model	$q^2_{CV}(A_{opt})$	s	SDEP	q^2_{test}
1.	SOM-4D-QSAR$_{\mathbf{o}}$[a,c,d]	0.53(3)	0.79	0.79	0.61
2.	SOM-4D-QSAR$_{\mathbf{q}}$[a,c,e]	0.74(3)	0.82	0.72	0.67
3.	4D-QSAR-**J**[a,b,c]	0.61(4)	1.03	0.76	0.65
4.	4D-QSAR-**J**[a,b,e]	0.72(3)	0.84	0.70	0.70
5.	SOM-4D-QSAR$_{\mathbf{o}}$[c,d]	0.69(3)	0.80		
6.	SOM-4D-QSAR$_{\mathbf{q}}$[c,e]	0.76(4)	0.71		
7.	4D-QSAR-**J**[b,c]	0.72(3)	0.76		
8.	4D-QSAR-**J**[b,d]	0.80(4)	0.65		

[a] Model: training set (even numbers) and test set (odd numbers). [b] Box 30 Å:30 Å:30 Å. [c] Compound **36s** was used as reference compound **R**. [d] Sum_occupancy descriptor. [e] Mean_charge descriptor.

and CoMSA models ($q^2_{CV} = 0.77$, $s = 0.71$, $A_{opt} = 5$).[9] It is worth noting that charge descriptors (**q**) give better models (higher q^2_{CV} values) than do occupancy descriptors (**o**). Additionally, the IVE-PLS elimination procedure has been applied in entries 1−4 and monitored by the calculation of the SDEP or q^2_{test} parameter based on the external (test) data sets. In fact, a comparison of the obtained IVE-PLS models to the previous models indicates the improvement in their predictive ability, e.g., from $q^2_{CV} = 0.53$, SDEP = 0.79, $q^2_{test} = 0.61$ (SOM-4D-QSAR$_{\mathbf{o}}$) to $q^2_{CV} = 0.81$, SDEP = 0.92, $q^2_{test} = 0.48$ (SOM-4D-QSAR$_{\mathbf{o}}$-IVE) and $q^2_{CV} = 0.74$, SDEP = 0.72, $q^2_{test} = 0.67$ (SOM-4D-QSAR$_{\mathbf{q}}$) to $q^2_{CV} = 0.93$, SDEP = 0.67, $q^2_{test} = 0.69$ (SOM-4D-QSAR$_{\mathbf{q}}$-IVE). This improvement is most notable for the charge descriptors where the large increase in q^2_{CV} is accompanied also by a slight increase in the q^2_{test} value. For the occupancy descriptors, the data reduction with IVE-PLS procedure improved the q^2_{CV} performance, while maintaining the predictive ability given by q^2_{test} at nearly the same level. Because the quality of the obtained models depends significantly on the distribution of the molecules into the training and test sets, we conducted an additional experiment. Here we iteratively repeated the sampling of 50 molecules into the training and test series in a proportion of 2/3 to 1/3, with each set containing 33 and 17 molecules, respectively. This procedure provides the data for the Stochastic Model Validation (SMV) scheme.[23] It is technically impossible to verify all possible 33/17 samplings of 50 molecular objects: thus, the total number of samplings was reduced to a relatively small fraction of 80 000 randomly generated distributions. This scheme carefully analyzes the influence of the training/test set sampling of the series on the final results of the 4D-QSAR modeling. Figure 3 illustrates the dependence between q^2_{CV} and q^2_{test} for 80 000 of the above-mentioned 4D-QSAR models. We also tested the sampling defined by the Kennard-Stone (KS) algorithm. Essentially, the q^2_{CV} performance ranges from 0.3 to 0.9 for occupancy_maps ($q^2_{CV} = 0.62$, $q^2_{test} = 0.72$ values for the KS sampling) and 0.55 to 0.9 for mean_charge_maps ($q^2_{CV} = 0.75$, $q^2_{test} = 0.68$ for the KS sampling), respectively. The population of models in Figure 3 is color coded, and red indicates the region of highest model density. This region is characterized by both relatively high q^2_{CV} and q^2_{test}, which suggests high modeling and predictive ability of the analyzed models. The histogram in Figure 3 illustrates the frequency of the individual compound selection into the test set as a function of a compound number, if tested within the region encircled with the red line. This reveals a relatively smooth distribution of each compound between the training and the test sets. Hence, we selected 100 random training/test set samplings within this region to illustrate the behavior of q^2_{CV} and q^2_{test} during the whole IVE-PLS variable elimination procedure. Figure 4a shows the dependence of q^2_{CV} for the training set and q^2_{test} performance on the number of variable eliminated. It can be observed that the elimination monitored by the performance of q^2_{CV} and q^2_{test} is a stable process. Moreover, the q^2_{test} value, which measures the predictive ability in the test set, is only slightly lower than q^2_{CV} for the training set.

The procedure illustrated in Figure 4 carefully monitors the statistical reliability of the SOM-4D-QSAR models in a

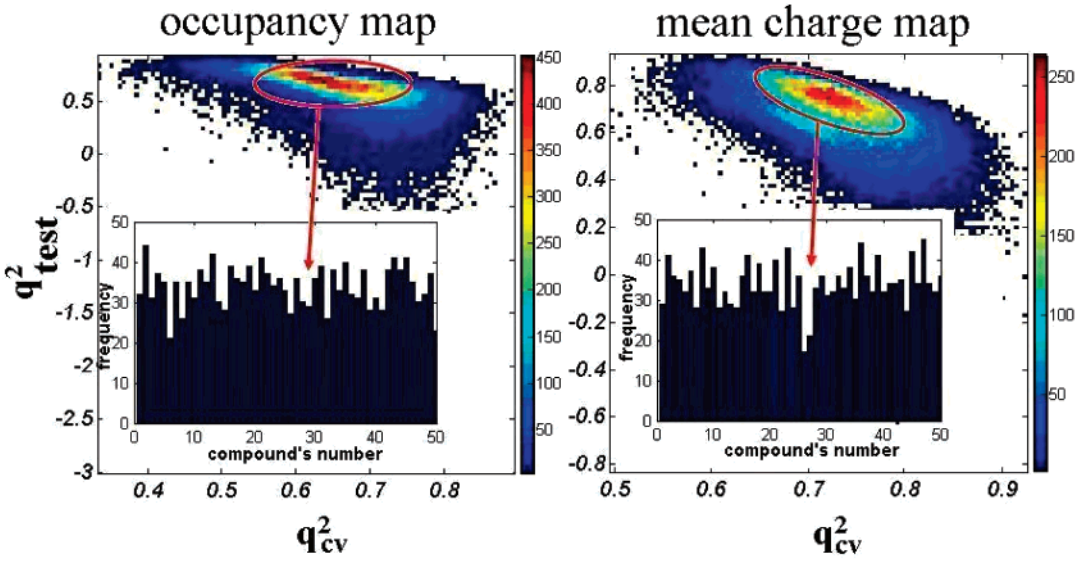


Figure 3. The density SMV plot illustrating relationships between a q^2_{CV} value evaluated with the LOO CV method for the 33 molecule training set sampled for all 50 molecules against a q^2_{test} value estimated by the application of the respective LOO CV model for the prediction of the activity for the remaining 17 molecules (test set), while using occupancy (a) and charge descriptors (b). Histograms specify a number of the individual compounds appearing in the test set within the most densely populated region which is encircled with red.

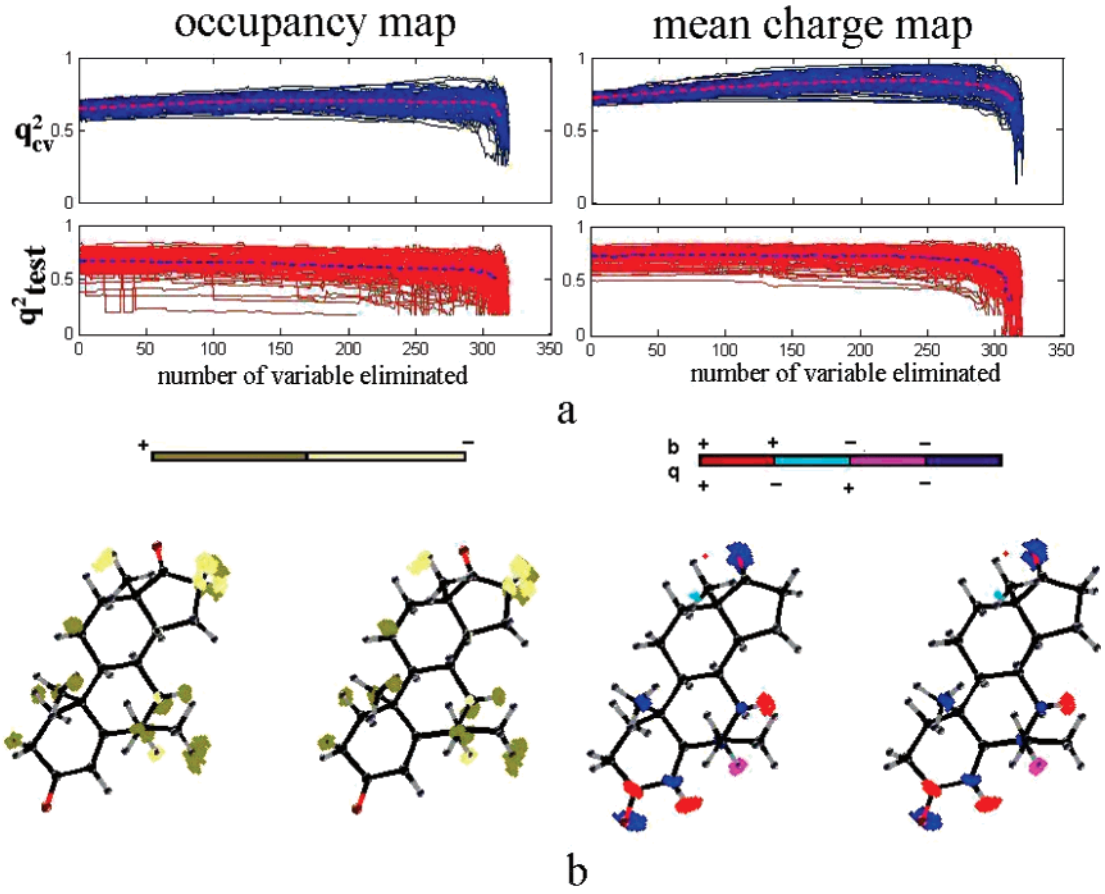


Figure 4. 4D-QSAR IVE-PLS monitored for 100 random 33/17 training/test set samplings for the occupancy (left) and charge (right) descriptors. The plots (a) illustrate the changes of q^2_{CV} and q^2_{test} as a function of the number of original variables eliminated. Molecular plots (b) visualize the molecular sectors of the largest contribution into the activity revealed from the 100 random models illustrated in (a), details in text. Colors code the sign of influence. For the charge descriptors colors code four possible combinations of the sign of charge (q) and the sign of the weight in the PLS model (b).

stochastic course of action. The stability observed during data elimination defined not only by q^2_{CV} but also by q^2_{test} inspired us to extend this procedure into the QSAR visualization step. Unlike standard procedures, which display such plots from a single training set, we attempted to identify variables that are important simultaneously for all models plotted in Figure 4a. Thus, 15% of the variables characterized by the highest stability included in each final PLS model were identified and normalized to the range [0−1]. Columns with a value higher than 0.5 were chosen for further analysis. Figure 4b illustrates the analyzed areas with the highest contribution to the activity. However, the pattern shown in this figure was obtained by further filtering of the variables. We included only the upper 50% of the variables having the largest contribution. In particular, it can be observed that the displayed regions correspond accurately to the substituents of the A and B steroid ring as was described in previous 3D-QSAR studies.[29] Moreover, these results reveal the importance of the D ring, which has also been previously observed.[9] Generally the obtained results compare well with those accessible in the literature.[29] Detailed inspection of the visual representations of the IVE models confirmed the significance of regions situated near rings A and B. Moreover, the positive coefficient characterizing these areas (Figure 4b, occupancy descriptor) suggests the need for a substituent that would increase the activity. For models with mean charge descriptors, a carbonyl group on ring A is identified as a site that enhances activity—an important region where negative charge is favorable. The negative regression coefficient in this area probably means that some polar or hydrogen-bonding acceptor occupies this site. The occupancy descriptors near the D ring area have generally negative coefficients, suggesting that, for a given alignment, this region cannot be occupied by any type of atom without a decrease in activity. This site seems to be sterically disallowed because of the sterical bulk of the receptor pocket. The negative regression coefficient for keto oxygen likely means that a polar or hydrogen-bonding acceptor is located in this position.

3.2. Azo Dyes Series. The interaction between a dye molecule and cellulose is a complex phenomenon, which can be described by Langmuir or Freundlich equations.[31] The adsorption isotherms describing cellulose dyeing suggest a specific interaction between dye and fiber but do not provide a molecular description of the process. Despite recent investigations aimed at modeling 2D and 3D-QSAR for dye molecules, it is still controversial whether a pharmacophore hypothesis can be applied for analysis of dye-cellulose affinities. On the other hand, the arrangement of dye molecules on the cellulose surface suggested that there could exist binding sites in the crystalline region that form cavities capable of incorporating dye molecules. The multiplicity of interactions taking place during targeting of cellulose by a dye molecule, for example electrostatic, van der Waals,

Table 5. Comparison of 4D-QSAR Models for the Azo Dyes

entry	model	$q^2_{CV}(A_{opt})$	*s*	SDEP	q^2_{test}
1.	SOM-4D-QSAR$_{\mathbf{o}}$[a,c,d]	0.79(5)	2.00	2.18	0.51
2.	SOM-4D-QSAR$_{\mathbf{q}}$[a,c,e]	0.43(3)	3.00	2.50	0.35
3.	4D-QSAR-**J**[a,b,c]	0.85(7)	1.92	2.73	0.58
4.	4D-QSAR-**J**[a,b,e]	0.66(3)	2.31	2.49	0.15
5.	SOM-4D-QSAR$_{\mathbf{o}}$[c,d]	0.74(5)	1.86		
6.	SOM-4D-QSAR$_{\mathbf{q}}$[c,e]	0.60(7)	2.59		
7.	4D-QSAR-**J**[b,c]	0.81(7)	1.98		
8.	4D-QSAR-**J**[b,c]	0.67(5)	2.16		

[a] Model: training set (odd numbers) and test set (even numbers). [b] Box 30 Å:30 Å:30 Å. [c] Compound **1a** was used as reference compound **R**. [d] Sum_occupancy descriptor. [e] Mean_charge descriptor.

hydrophobic or hydrogen bonding, indicates a similarity between the drug-receptor and dye-fiber complexes. On the other hand, dye-cellulose interactions seemed to be less specific. Several studies have been published recently that make use of the pharmacophore concept in the investigation of cellulose dying and which indicated that the electrostatic field significantly contributes to the dyeing affinity. A tinctophore concept has been developed to extend the idea of the pharmacophore into the dye chemistry.[32,33]

The objective of this study is a systematic analysis of the tinctorial properties of heterocyclic monoazo dyes using the SOM-4D-QSAR procedure. Table 5 illustrates the results obtained. As observed, the q^2_{CV} performance, which ranges from 0.43 to 0.81, is quite consistent with the best CoMFA ($q^2_{CV} = 0.44 - 0.73$)[32] and slightly below the CoMSA verified by the Golbreikh-Tropsha criterion ($q^2_{CV} = 0.93$).[33] Entries 1−4 present results obtained after dividing the set of compounds into training/test (odd/even numbers) subsets, whereas 5−8 include results for the whole data set. It is notable that the predictive ability for the test set is at precisely the same level as the q^2_{CV} value. Moreover, the occupancy descriptors (4D-QSAR$_{\mathbf{o}}$: $q^2_{CV} = 0.81$) provide better models than the charge descriptors (4D-QSAR$_{\mathbf{q}}$: $q^2_{CV} = 0.67$). This tendency is also observed in the case of the SOM-4D-QSAR$_{\mathbf{o}}$ ($q^2_{CV} = 0.74$) and SOM-4D-QSAR$_{\mathbf{q}}$ ($q^2_{CV} = 0.60$). This conclusion contrasts with results obtained in the 3D-QSAR analyses (CoMFA, CoMSA), which suggest that the electrostatic field (and not the steric) is dominant for binding.[32,33] This illustrates the fact that incorporation of the conformational flexibility into the QSAR analysis of these nonrigid structures can provide significantly different results than those observed for nonflexible molecules. As in the procedure described for steroids, we performed a SMV validation of the azo dye models with a training/test set sampling ratio of 2/3 to 1/3, as shown in Figure 5. In the region of highest density (q^2_{CV}/q^2_{test}) the q^2_{CV} performance ranges from 0.4 to 0.8 for the occupancy_maps ($q^2_{CV} = 0.68$, $q^2_{test} = 0.68$ for the KS sampling) and 0.2−0.65 for the mean_charge_maps ($q^2_{CV} = 0.43$, $q^2_{test} = 0.41$ for the KS sampling), respectively.

Figure 6a,b illustrates the stochastic procedure used for determination of the molecular areas with the largest contribution to dye-cellulose interactions, as revealed by a procedure similar to that described in chapter 3.1. The IVE-PLS modeling, tested for 100 models with the molecules sampled randomly into the test and training sets, indicates that this procedure is now much more stable for the occupancy_maps (Figure 6a). Unlike the steroid models (shown in Figure 3), which specify a sharp atomic representation, we can now observe a large molecular environment displayed in Figure 6b. This result suggests that the tinctophore pattern differs from the classical meaning of the drug pharmacophore and involves a large part of the molecule. On the other hand, this also suggests that the interactions are much less specific, and a large molecular environment is important for the modeling of dye-cellulose interactions.

3.3. Modeling Hammett Constants for Benzoic Acid Series. Modeling chemical reactions by the Hammett equation is a first example of a QSAR-like analysis. Despite the fact that there is no evidence for the existence of any pharmacophore built on the basis of a receptor-like environ-

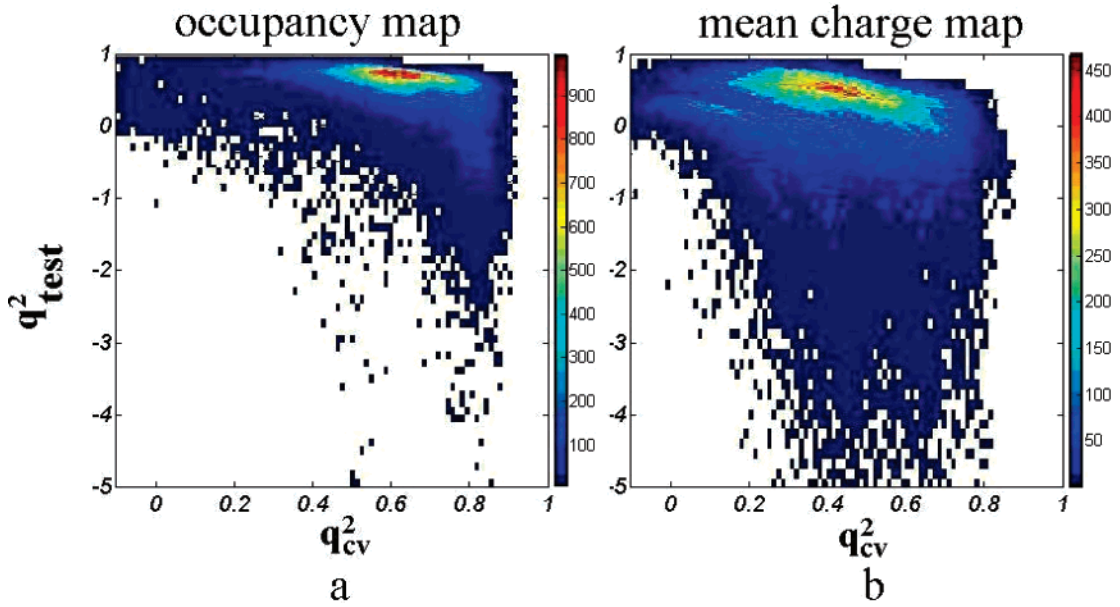


Figure 5. The density SMV plot illustrating relationships between a q^2_{CV} value evaluated with the LOO CV method for the 20/10 training/test set sampled of all molecules against a q^2_{test} value estimated by the application of the respective LOO CV model for the prediction of the activity for the remaining molecules (test set), while using occupancy (a) and charge descriptors (b).

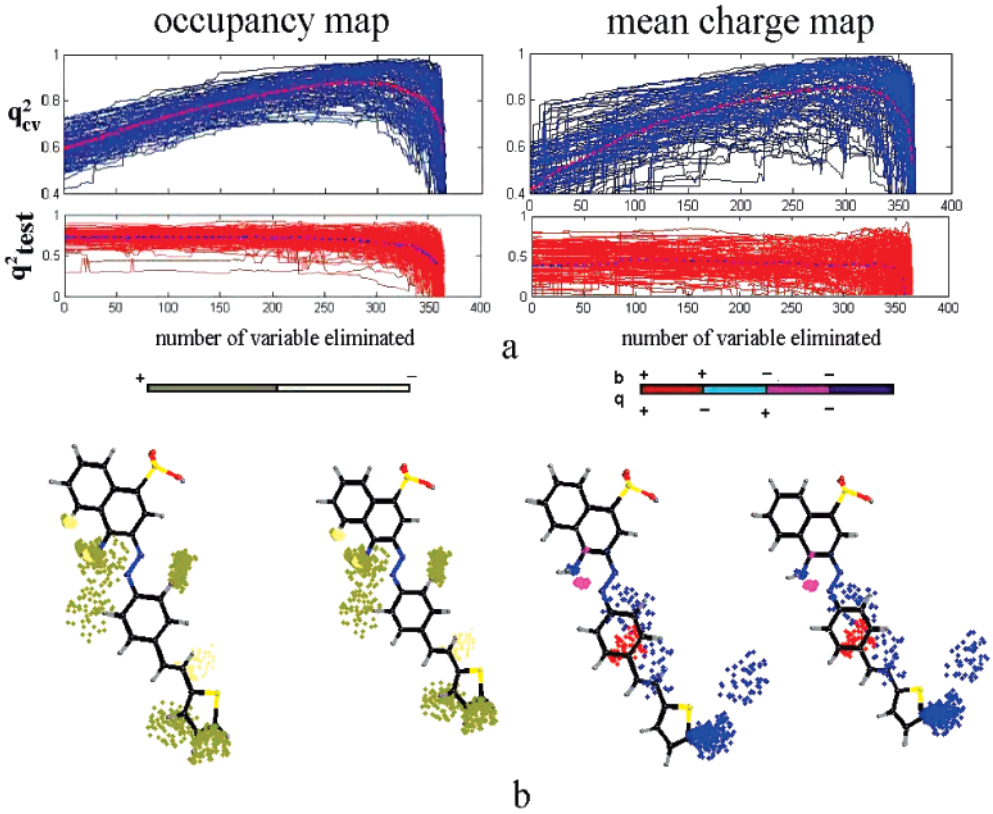


Figure 6. 4D-QSAR IVE-PLS monitored for 100 random 20/10 training/test set samplings for the occupancy (left) and charge (right) descriptors. The plots (a) illustrate the changes of q^2_{CV} and q^2_{test} as a function of the number of original variables eliminated. Molecular plots (b) visualize the molecular sectors of the largest contribution into the activity revealed from the 100 random models illustrated in (a), details in text. Colors code the sign of influence. For the charge descriptors colors code four possible combinations of the sign of charge (q) and the sign of the weight in the PLS model (b).

ment that is responsible for the chemical reaction, several applications of the 3D-QSAR method for modeling chemical reactions have been published recently.[29,34,35] Modeling Hammett constants for benzoic acids is probably the most often tested example. Below we present a SOM-4D-QSAR study of a series of 72 benzoic acids with the Hammett constant data that was previously analyzed by 3D-QSAR methods.[27,36] Similar to previous investigations, molecules **1b**−**72b** were arbitrarily divided into two subseries: the first containing **1b**−**49b**, which form a training set, and **50b**−**72b**, which were used for validation of the obtained models (test set). Kim and Martin modeled the Hammett constant within the **1b**−**49b** acid subset using the AM1 partial atomic charges, which provides a CoMFA model described by $q^2_{CV} = 0.89$.[24] Since all other previous calculations demonstrated that the AM1 method provides the best fit models,[29] we also applied this method in our calculations. The q^2_{CV} performance for the SOM-4D-QSAR$_q$ method reaches a value of $q^2_{CV} = 0.80$, SDEP = 0.22, $q^2_{test} = 0.65$ for a single training/test set sampling of **1b**−**49b**/**50b**−**72b** (as reported in previous publications) and $q^2_{CV} = 0.75$, $q^2_{test} = 0.70$ for the KS sampling. However, this was achieved when carboxylic group atoms were arbitrarily indicated as the IPE, i.e., only these atoms were taken analyzed while 4D QSAR modeling. Similarly, Figure 7 reports the SMV validation scheme, if focused exclusively on the carboxylic IPE atoms. In this particular case, the overall performance is comparable with that observed for two examples discussed in chapters 3.1 and 3.2. However, inclusion of the non-carboxylic atoms of

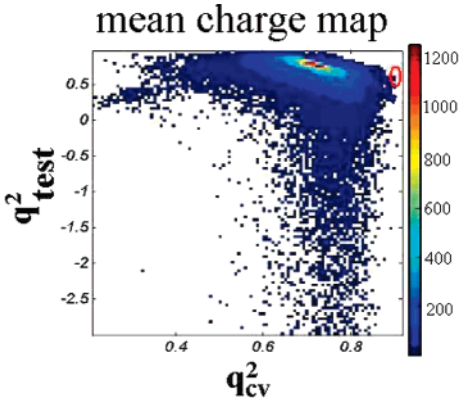


Figure 7. The density SMV plot illustrating a relationships between a q^2_{CV} value evaluated with the LOO CV method for the 49/23 training/test set sampled of all molecules against a q^2_{test} value estimated by the application of the respective LOO CV model for the prediction of the ionization ability for the remaining benzoic acid series molecules (test set).

the benzoic ring deteriorates the quality of the models. The single sampling performance for the initial test/training set distribution **1b**−**49b**/**50b**−**72b** are given by $q^2_{CV} = 0.75$, SDEP = 0.23, $q^2_{test} = 0.65$ (an anchoring C1 atom of the benzene ring included) and $q^2_{CV} = 0.50$, SDEP = 0.35, $q^2_{test} = 0.15$ for all atoms, respectively. Because the ionization reaction involves a carboxylic function, we would expect to observe this function displayed in the 4D QSAR method. Figure 8a shows the dependence of the q^2_{CV} for the training

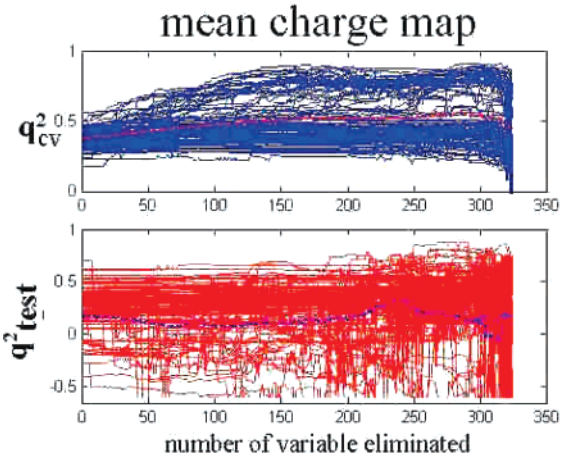


Figure 8. 4D-QSAR IVE-PLS monitored for 100 random 49/23 training/test set samplings using charge descriptors for the benzoic acid series. The plots illustrates the changes of q^2_{CV} and q^2_{test} as a function of the number of original variables eliminated.

set and q^2_{test} performance on the number of variables eliminated in the stochastic procedure for 100 randomly sampled models, including all atoms in each molecule, with different test and training set molecules sampled in a 2/3 to 1/3 ratio. The inclusion of all molecules brings further noise into the model and makes the elimination monitored by the q^2_{CV} and q^2_{test} performance much less stable than observed for the steroid and azo-dye molecules.

This makes the SOM-4D-QSAR procedures for the benzoic acid series much less efficient than those described for steroids and dyes. Figure 9 illustrates the cause of that effect. We observe that the Hammett constant is correlated with the standard deviation of the atomic charges of the atoms forming the carboxylic group, as plotted in Figure 9a for single conformers. Figure 9b plots the correlation coefficient for this relationship tested for different conformers sampled. This result indicates that the relationship revealed by Figure 9a remains true only for a tiny fraction of conformers; in fact, only for a single starting conformer. The Hammett constant and pK_a are strictly limited by the electronic effect, which cannot be described by the occupancy descriptors. On the other hand, any distortions of the main conformers can deteriorate the accuracy of calculations of the partial atomic charges, which also decreases the performance of 4D-QSAR modeling.

Based on this fact, the stochastic protocol for model visualization has been modified in this particular case. In Figure 10 we illustrate the molecular areas and visualize the atoms making the largest contribution to the model; however; unlike for the previous series, we included only the ten top models with the largest q^2_{CV} and characterized by relatively high values of q^2_{test} ($q^2_{CV} \geq 0.9$, $q^2_{test} \geq 0.6$), which are encircled by the red line in Figure 7. The best model selected by IVE-PLS is described by $q^2_{CV} = 0.93$, $q^2_{test} = 0.21$SDEP $= 0.33$ and includes the following compounds in the training

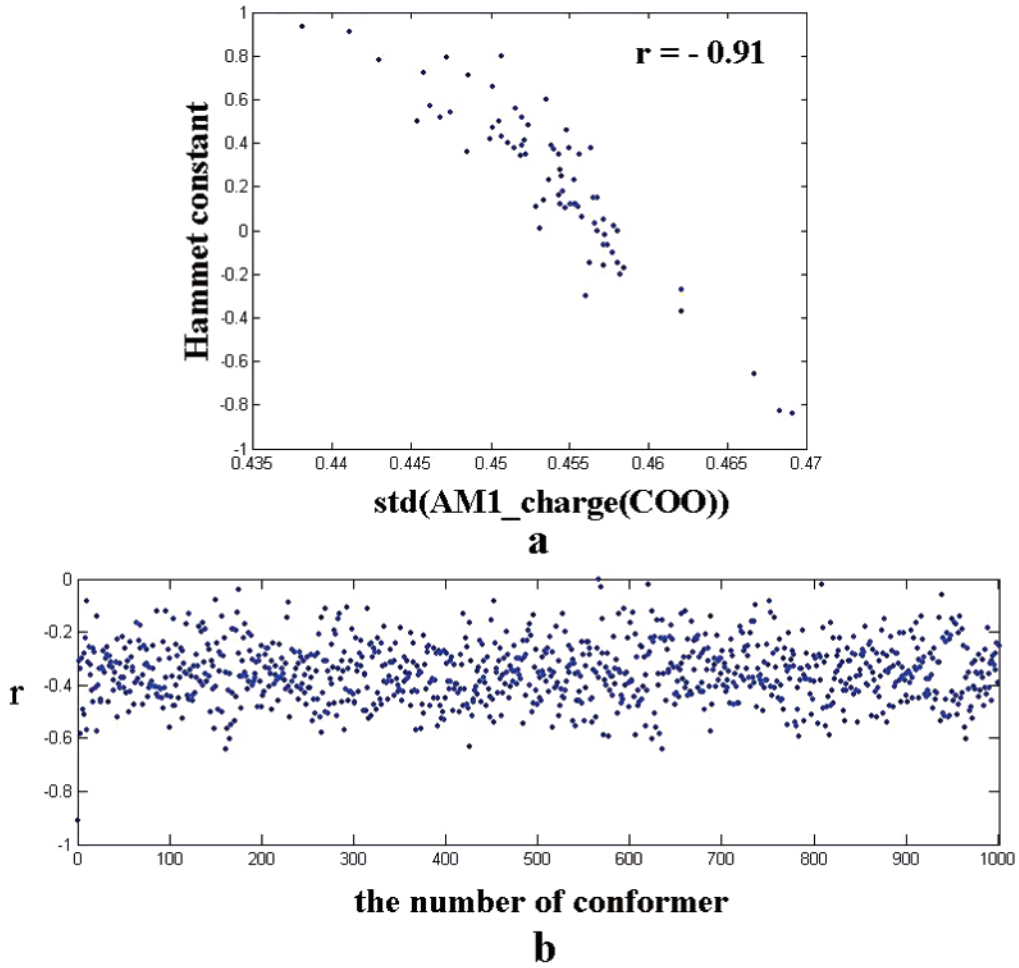


Figure 9. The correlation between the standard deviation of the charge on the carboxylic group atoms (single conformer) and the Hammett constant (a) and for different conformations (b).

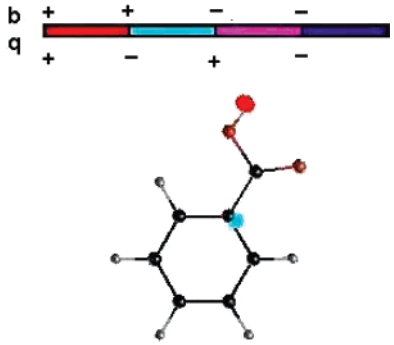


Figure 10. Molecular plot visualizing the molecular sectors of the largest contribution into the activity as revealed from ten models of the largest q^2_{CV} encircled by red line in Figure 7. Colors code four possible combinations of the sign of charge (q) and the sign of the weight in the PLS model.

set: 1−16, 18, 19, 21, 23, 25−30, 33, 37, 39−41, 43−47, 49, 50, 53, 57, 58, 61, 62, 64−66, 68, 69, 72 and with 17, 20, 22, 24, 31, 32, 34−36, 38, 42, 48, 51, 52, 54−56, 59, 60, 63, 67, 70, 71 in the test set. The contour plot shown in Figure 10 indicates the carboxylic as an ionization center. However, the low q^2_{test} value indicates the much lower predictive ability of the model.

5. CONCLUSIONS

In the current paper we present a receptor-independent 4D-QSAR method based on self-organizing mapping (SOM-4D-QSAR) and in particular focus on its pharmacophore mapping ability. We use a novel stochastic procedure to verify the predictive ability of the method for a large population of 4D-QSAR models generated. This systematic study was conducted on a series of benzoic acids, azo dyes, and steroids that bind aromatase. We show that the 4D-QSAR method coupled with IVE-PLS provides a very stable and predictive modeling technique. The method enables us to identify the molecular motifs contributing the most to the fiber-dye affinity and the aromatase enzyme binding activity of the steroid. However, the method appeared much less effective for the benzoic acid series, in which the efficacy was limited by electronic effects strictly correlated to a single conformer.

ACKNOWLEDGMENT

The authors thank Professor Johann Gasteiger of the University of Erlangen-Nurnberg, BRD for facilitating access to the programs KMAP. Partial financial support of the KBN Warsaw, grant no. 3 T09A 01127, is gratefully acknowledged. Dr. Andrzej Bak thanks the Foundation for Polish Science for his individual grant.

REFERENCES AND NOTES

(1) Pommier, Y.; Johnson, A. A.; Marchand, C. Integrase Inhibitors to Treat HIV/AIDS. *Nat. Rev. Drug Discovery* **2005**, *4*, 236−248.
(2) Vedani, A.; Dobler, M.; Lill, M. A. Combining Protein Modeling and 6D-QSAR. Simulating the Binding of Structurally Diverse Ligands to the Estrogen Receptor. *J. Med. Chem.* **2005**, *48*, 3700−3703.
(3) Vedani, A.; Briem, H.; Dobler, M.; Dollinger, H.; McMasters, D. R. Multiple-Conformation and Protonation-State Representation in 4D-QSAR: The Neurokinin-1 Receptor System. *J. Med. Chem.* **2000**, *43*, 4416−4427.
(4) Vedani, A.; Dobler, M. 5D-QSAR: The Key for Simulating Induced Fit? *J. Med. Chem.* **2002**, *45*, 2139−2149.
(5) Polanski, J. Drug Design Using Comparative Molecular Surface Analysis. *Expert Opin. Drug Discovery* **2006**, *1* (7), 693−707.
(6) Doweyko, A. 3D-QSAR Illusions. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 587−596.
(7) Polanski, J.; Bak, A.; Gieleciak, R.; Magdziarz, T. Modeling Robust QSAR. *J. Chem. Inf. Model.* **2006**, *46*, 2310−2318.
(8) 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. Chem.* **1996**, *68*, 3851−3858.
(9) Polanski, J.; Gieleciak, R. The Comparative Molecular Surface Analysis (CoMSA) with Modified Uninformative Variable Elimination-PLS (UVE-PLS) Method: Application to The Steroids Binding the Aromatase Enzym. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 656−666.
(10) Zupan, J.; Gasteiger, J. *Neural Networks and Drug Design for Chemists*, 2nd ed.; Wiley VCH: BRD, Weinheim, 1999
(11) Polanski, J. Self-Organizing Neural Networks for Pharmacophore Mapping. *Adv. Drug Deliverery Rev.* **2003**, *55*, 1149−1162.
(12) Anzali, S.; Barnickel, G.; Krug, M.; Sadowski, J.; Wagener, M.; Gasteiger, J.; Polanski, J. The Comparison of Geometric and Electronic Properties of Molecular Surfaces by Neural Networks: Application to the Analysis of Corticosteroid Binding Globulin Activity of Steroids. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 521−534.
(13) Polanski, J.; Walczak, B. The Comparative Molecular Surface Analysis (CoMSA):A Novel Tool for Molecular Design. *Comput. Chem.* **2000**, *24*, 615−625.
(14) Polanski, J.; Bak, A.; Gieleciak, R.; Magdziarz, T. Self-organizing Neural Networks for Modeling Robust 3D and 4D QSAR: Application to Dihydrofolate Reductase Inhibitors. *Molecules* **2004**, *9*, 1148−1159.
(15) Gasteiger, J. KMAP 3.0; Molecular Networks GmbH. http://www-.molecular-networks.com/software/overview/index.html (accessed Nov 13, 2006).
(16) Wold. S.; Sjöström, M.; Eriksson, L. PLS-regression: A Basic Tool of Chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109−130.
(17) Xu, L.; Zhang, W. Comparison of Different Methods for Variable Selection. *Anal. Chim. Acta* **2001**, *446*, 477−483.
(18) Hopfinger, A.; Wang, S.; Tokarski, J.; Jin, B.; Albuquerque, M.; Madhav, P.; Duraiswami, C. Construction of 3D-QSAR models using the 4D-QSAR analysis formalism. *J. Am. Chem. Soc.* **1997**, *119*, 10509−10524.
(19) Santos-Filho O.; Hopfinger A. Structure-Based QSAR Analysis of a Set of 4-Hydroxy-5,6-dihydropyrones as Inhibitor of HIV-1 Protease: An Application of the Receptor Dependent (RD) 4D-QSAR Formalism. *J. Chem. Inf. Model.* **2006**, 345−354
(20) Bak, A.; Polanski, J. The 4D-QSAR Study on Anti-HIV HEPT Analogues. *Bioorg. Med. Chem.* **2006**, *14*, 273−279.
(21) Sybyl 7.1 program, available from Tripos Inc., St. Louis, MO, U.S.A. http://www.tripos.com./
(22) HyperChem 5.0 program, available from HyperCube Inc., Gaineswile, FL, U.S.A. http://www.hyper.com./
(23) Polanski, J.; Gieleciak, R.; Bak, A. Probability Issues in Molecular Design. Predictive and Modeling Ability in 3D-QSAR Schemes. *Comb. Chem. High Throughput Screening* **2004**, *7*, 793−807.
(24) Funar-Timofei, S.; Schrüümann, G. Comparative Molecular Field Analysis (CoMFA) of Anionic Azo Dye-Fiber Affinities I: Gas-Phase Molecular Orbital Descriptors. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 788−795.
(25) Kim, K. H.; Martin, Y. C. Direct Prediction of Linear Free Energy Substituent Effects from 3D Structures Using Comparative Molecular Field Analysis. 1. Electronic Effects of Substitued Benzoic Acids. *J. Org. Chem.* **1991**, *56*, 2723−2729.
(26) Cramer, R., III; Patterson, D.; Bunce, J. Comparative Molecular Field Analysis (CoMFA). *J. Am. Chem. Soc.* **1998**, *110*, 5959−5967.
(27) Polanski, J.; Bak, A. Modeling Steric and Electronic Effects in 3D and 4D-QSAR Schemes: Predicting Benzoic p*K*a Values and Steroic CBG Binding Affinities. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2081−2092.
(28) Lill, M. A.; Dobler, M.; Vedani, A. Prediction of Small-Molecule Binding to Cytochrome P450 3A4: Flexible Docking Combined with Multidimensional QSAR. *ChemMedChem* **2006**, *1*, 73−81.
(29) Beger, R.; Buzatu, D.; Wilkes, J.; Lay 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.
(30) Gieleciak, R.; Polanski, J. Modeling Robust QSAR 2: Iterative Variable Elimination Schemes for CoMSA: Application for Modeling Benzoic Acid pKa Values. *J. Chem. Inf. Comput. Sci.* **2007**, *47*, 547−556.
(31) Timofei, S.; Schmidt, W.; Kurunczi, L.; Simon, Z. A Review of QSAR for Dye Affinity for Cellulose Fibres. *Dyes Pigm.* **2000**, *47*, 5−16.

(32) Timofei, S.; Fabian, W. M. F. Comparative Molecular Field Analysis (CoMFA) of Heterocyclic Monoazo Dye-Fibre Affinities. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1218−1222.
(33) Polanski, J.; Gieleciak, R.; Wyszomirski, M. Comparative Molecular Surface Analysis (CoMSA) for Modeling Dye-Fiber Affinities of the Azo and Antraquinone Dyes. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1754−1762.
(34) Lipkowitz, K. B.; Pradhan, M. Computational Studies of Chiral Catalysts: A Comparative Molecular Field Analysis of An Asymmetric Diels-Alder Reaction with Catalysts Containing Bisoxazoline or Phosphinooxazoline Ligands. *J. Org. Chem.* **2003**, *68*, 4648−4656.
(35) Vrtacnik, M.; Voda, K. HQSAR and CoMFA Approaches in Predicting Reactivity of Halogenated Compounds with Hydroxyl Radicals. *Chemosphere* **2003**, *52*, 1689−1699.
(36) Hollingsworth, Ch. A.; Seybold, P. G.; Hadad, Ch. M. Substituent Effects on the Electronic Structure and pK_a of Benzoic Acid. *Int. J. Quantum Chem.* **2002**, *90*, 1396−1403.

CI700025M