Modeling Steric and Electronic Effects in 3D- and 4D-QSAR Schemes: Predicting Benzoic pK_a Values and Steroid CBG Binding Affinities

Jaroslaw Polanski* and Andrzej Bak

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

Received June 16, 2003

We conducted a systematic study of the performance of the 3D- and 4D-QSAR schemes in modeling steric and electronic effects. In particular, we compared the CoMFA and Hopfinger's 4D-QSAR schemes, which apply completely different concepts for the generation of the molecular data used for modeling QSAR. Hence, we attempted to predict the pK_a values of (*o*-, *m*-, and *p*-)benzoic acids which were divided into three subseries in order to simulate different levels of steric and electronic control. The steroids binding to CBG were used as a benchmark series where biological activity is limited by shape factors. Although individual models differ depending upon the individual scheme, generally, both CoMFA and 4D-QSAR appeared to provide comparable results, irrespective of the differences in the coding schemes used for the description. Moreover, a new 4D-QSAR scheme involving a self-organizing neural network was designed. Generally, the SOM scheme that we designed performs comparably to the grid scheme; however, it provides better results for the charge type descriptors, and the robust neuron architecture allows for the decrease of the influence of the molecular superimposition mode.

INTRODUCTION

Modeling chemical and biological effects are important objectives of the present day chemistry and pharmacology. Therefore, the expansion of rational techniques, in particular of the Quantitative Structure Activity Relationships (QSAR) and all its variants, i.e., QSPR, QSRR, and finally of the three-dimensional (3D) approaches,[1−4] is obviously necessary. 3D-QSAR strategies, especially CoMFA, have notably contributed to our ability to forecast the activity of potential bioeffectors. Although theoretically 3D methods should offer better efficiency than 2D techniques, in reality this conclusion does not always prove true. However, a number of investigations have shown that CoMFA protocols can be modified to provide models with better predictive quality. This includes the improvements in the description of molecules and rules for the alignment of molecules[5] as well as the statistics used.[6−8] On the other hand, multidimensional QSARs, namely, Vedani's 4D- and 5D-QSAR concepts[9] or Hopfinger's 4D-QSAR,[10] are new ideas that can improve modeling performance.

A uniform spatial grid constructed over molecular configurations in Hopfinger's 4D-QSAR resembles the CoMFA method. Basically, the fourth dimension of conformational freedom substantially differentiates 4D- from the most of 3D-QSAR methods. In fact, however, there are also some important differences in the way the molecular structures are encoded both in the 4D-QSAR and CoMFA methods. While in CoMFA the nodes of this grid define the points in which the electrostatic or steric potential is calculated, in 4D-QSAR the occupancy of the grid cells (unit cubes) by individual atoms or even atom groups composing each molecule describes the molecule itself. Accordingly, if we consider a molecule as a certain volume contained within a space enclosed in the limits of the molecular surface, then Hopfinger's 4D-QSAR can be interpreted as a variant of the Molecular Shape Analysis,[11] in which molecular surfaces are represented not by spherelike boundaries but by polycube structures formed from a set of cubic domains (cells). The substitution of the molecular representation with spherelike atoms using cubic cells of different resolution creates a fuzzy picture of molecular objects.[12] It is worth noticing that a similar description of the molecular shape, namely that a molecule is divided into spatial regions either filled or unfilled by atoms or groups of atoms of certain volumes, has also been used previously by Purcel and Testa, Motoc, or Allinger.[13] Neural networks,[14,15] in particular, a self-organizing neural network,[16−18] can also be an interesting option for a similar fuzzification of molecular representations. In 3D-QSAR among many modifications suggested, molecular flexibility (improving molecular superimposition) is a feature that can provide a better description of the binding to the putative receptor, hence, the fuzzification offered by these methods seems to be of special interest.

The aim of the current investigation is 2-fold. First, we would like to compare the influence of the coding system on the efficiency of QSAR modeling of the CoMFA and Hopfinger's 4D-QSAR schemes. At the same time we do not concentrate here on the details of each method, e.g., the additional conformational freedom of 4D-QSAR, but on the philosophy of the description of molecular objects. Thus, we applied these methodologies to model molecular effects that are controlled by steric and electronic factors. In particular, we used different series of (*o*-, *m*-, and *p*-)benzoic acids[19] and the CoMFA steroids[20] to simulate different levels of steric and electronic control of the pK_a values and biological activity. We endeavored to verify if modeling performances can be used as a measure for estimating the

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

10.1021/ci034118l CCC: $25.00
Published on Web 09/18/2003

relative influence of steric vs electronic control and to check the interrelations between model qualities. Furthermore, we also attempted to implement a coupled neural network and PLS system to design a new 4D-QSAR scheme.

THEORETICAL BACKGROUND

4D-QSAR Strategy. Hopfinger's 4D-QSAR[10,21−23] scheme is the extension of the 3D-QSAR that includes conformational analysis as the fourth dimension. Molecular dynamic simulations provide conformers for further comparative analysis. The alignment of the molecules is optimized, and conformations generated are superimposed. Each conformation from the conformational ensemble profile is then placed in the reference grid, and a series of so-called occupancy descriptors is calculated on the basis of the pattern in which the unit grid cubes are loaded by individual atoms of the molecule. Formally, grid cell occupancy descriptors (GCODs) are given by the following equation

$$A_o(c,i,j,k,N) = \sum_{t=0}^{T} O_t(c,i,j,k) \tag{1}$$

where A_o means the absolute occupancy of grid cell (i, j, k) at time t during the MDs for the chosen atoms of compound c; $O_t(c, i, j, k) = m$ if m atoms of c are present in the cell (i, j, k) at time t, and N is the number of sampling steps (T/t). Joint occupany J_o and self-occupancy S_o for compound c with a reference compound R is defined as

$$J_o(c,i,j,k,N) = \sum_{t=0}^{T} O_t(c,i,j,k) \cap O_t(R,i,j,k) \tag{2}$$

$$S_o(c,R,i,j,k,N) = \sum_{t=0}^{T} \{O_t(c,i,j,k) - [\sum_{t=0}^{T} O_t(c,i,j,k) \cap O_t(R,i,j,k)]\} \tag{3}$$

Using these descriptors a final QSAR model can be then calculated in the PLS procedure or using variables selected by genetic algorithms.[10,21−23] A big advantage of 4D-QSAR analysis is that it makes it possible to indicate the bioactive conformation.

Self-Organizing Neural Network. A self-organizing neural network[16−18] is an unsupervised architecture consisting only of a single layer—usually a two-dimensional grid of neurons. Two-dimensional topology of the grid means that we can distinguish the neighborhood relations between the nodes by defining the distances between them. N-dimensional data vectors presented into such a network are distributed between these neurons in such a way that those that are similar are put closer (into the neurons that are closer neighbors) to each other than those that differ. This operation can be performed by using a few algorithms. Classical competitive Kohonen learning is based on the strategy according to which each N-dimensional input vector is compared with the n-element ($n = N$) weight vectors describing each neuron to detect the one into which the individual input vector will be projected. At the beginning of the learning the weight vectors are set (e.g., randomly, but in newer versions more effective initialization have been developed), and the network is presented with the first input vector. A formal criterion for the selection of the winner (out_c) can be based for example on the Euclidean distance between a vector (x) and a weight (w).

$$out_c \leftarrow \min[\sum_{i=1}^{m} (x_{ai} - w_{ji})^2] \tag{4}$$

Then the weight of the winner is corrected to decrease this distance. This means the neuron "learns" to recognize this individual input. Moreover, the weights of the neighboring neurons are also corrected to attract similar inputs. Now the network is presented with the second vector and so on.

In such a process, a self-organizing network works as a clustering diagram grouping similar inputs from the input space into similar neurons of the output space. Moreover, we usually project a large number of the input data into a two-dimensional neural space of much smaller size. Thus it is also a compression device that enables one to substantially reduce the amount of data. Eventually, the network that has learned a topology of the input forms a base for the projection of the selected feature describing input data. If this feature is color-coded we obtain a two-dimensional pattern called a self-organizing map (SOM) or feature map. Thus, SOM is also a visualization tool. Bioinformatics is probably one of the hottest issues among recent SOM applications.

Zupan and Gasteiger designed a scheme for the application of the SOM algorithm for the projection of the molecular surface properties into the two-dimensional feature map, e.g. the electrostatic potential map.[18] In such a technique the SOM network is fed with the Cartesian coordinates (x, y, z) of the points sampled from the molecular surfaces. As a result 10 000−30 000 points sampled at the surface are mapped into ca. 400−2500 neurons arranged into a form of a two-dimensional map of electrostatic potential or other molecular features. A similar scheme was used in a complex method for the comparison of molecular surfaces, namely, Comparative Molecular Surface Analysis that uses a coupled neural network and PLS analysis to model 3D-QSARs.[20,24] We have also shown that self-organizing neural networks can be used for the comparison of molecules represented by atomic coordinates and superimposed by moments of inertia. In such a version the spherical volume, defined in the space by the SOM neuron, constructs a fuzzy spherelike molecular surface defining the volumes projected into individual neurons.[25]

EXPERIMENTAL SECTION

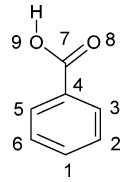
Model Builders. All the experimental data, i.e., pK_a for the benzoic acids and the biological activities for the CoMFA steroids, were extracted from refs 19 and 20 and are given in Tables 1 and 2, respectively.

CoMFA. All modeling work was performed using the Sybyl 6.2 software package run on Silicon Graphics O2 workstation. The initial geometry was optimized using the standard Tripos force field (POWELL method) with 0.005 kcal/mol energy gradient convergence criterion and a distant-dependent dielectric constant. Charges were calculated using the Gasteiger−Marsilli method implemented in Sybyl. Further geometry optimization was performed by MOPAC with the AM1 Hamiltonian. Alternatively, in supporting materials we presented CoMFA results for charges calculated

Table 1. pK_a Data for Benzoic Acids **1a**−**41a** According to Ref 19

compound	benzene ring-substituent	p$K_{a(exp)}$	compound	benzene ring-substituent	p$K_{a(exp)}$
1a	3-NH_2	4.78	**22a**	3-NO_2	3.45
2a	4-NH_2	4.85	**23a**	4-NO_2	3.44
3a	3-Br	3.90	**24a**	H	4.18
4a	4-Br	3.97	**25a**	2-OH, 3-OH	2.94
5a	4-$C(CH_3)_3$	4.40	**26a**	2-OH, 4-OH	3.29
6a	3-Cl	3.82	**27a**	2-OH, 5-OH	2.97
7a	4-Cl	3.99	**28a**	2-OH, 6-OH	1.30
8a	3-CN	3.60	**29a**	2-OH, 4-OH, 6-OH	1.68
9a	4-CN	3.55	**30a**	2-NO_2, 4-NO_2, 6-NO_2	0.65
10a	3-OH, 4-OH	4.48	**31a**	2-NH_2	4.95
11a	3-OH, 5-OH	4.04	**32a**	2-$C(CH_3)_3$	3.54
12a	4-C_2H_5	4.35	**33a**	2-Br	2.85
13a	3-F	3.87	**34a**	2-Cl	2.94
14a	4-F	4.14	**35a**	2-C_2H_5	3.79
15a	4-OH	4.58	**36a**	2-F	3.27
16a	3-I	3.86	**37a**	2-OH	3.00
17a	4-I	3.93	**38a**	2-I	2.86
18a	3-OCH_3	4.09	**39a**	2-OCH_3	4.09
19a	4-OCH_3	4.47	**40a**	2-CH_3	3.92
20a	3-CH_3	4.27	**41a**	2-NO_2	2.17
21a	4-CH_3	4.36			

with AM1 MOPAC method. We used the FIT option of the Sybyl to align the compounds analyzed. Parent benzoic acid, i.e., a common fragment for all molecules, was chosen as a template.



In CoMFA analyses we used the lowest conformational state that originated MD simulations. Four alternative alignments were carried out by superimposing the following atoms: 1,3,5 (model A); 2,4,6 (model B); 4,7,8 (model C); and 7,8,9 (model D); respectively.

The steric (Lennard-Jones) and electrostatic (Columbic) fields around the set of compounds were sampled with the probe atoms: sp^3 carbon (charge +1 and 0) and hydrogen (charge +1), on the rectangular grid that encompasses all aligned molecules (with margin of 3.0−4.0 Å). The CoMFA grid spacing was set to 2.0 Å in all dimensions. We kept a convention to truncate the steric and electrostatic values at the level of 30.0 kcal/mol.

4D-QSAR. The molecules after AM1 optimization were used as the initial structures in the molecular dynamic simulations (MDs). Each 3D structure is the starting point in generating conformational ensemble profile (CEP). Molecular dynamics was performed using the Sybyl software with standard Tripos force field. The temperature was set at 300 K. The atomic 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 the CEP of 100 000 trajectory states. Partial atomic charges were calculated using the semiempirical AM1 Hamiltonian. In practice, we used HYPERCHEM package for this purpose.

The alignment of the molecules was the next step of the 4D-QSAR analysis. Four different alignments were considered which were identical with those defined in CoMFA study. Each conformation is then placed in the grid cell space surrounding the aligned compounds. In this study space was divided into cubic grid lattice of 20 Å on each side containing a grid cell with a resolution of 1, 2, and 0.5 Å. Different types of grid cell occupancy descriptors (GCODs) were considered and calculated for the indicated atoms referred as interaction pharmacophore elements (IPE). Apart from, the occupancy GCODs actually used by Hopfinger,[21−23] in our current work we applied the absolute charge occupancy (A_q) defined as

$$A_q(c,i,j,k,N) = \sum_{t=0}^{T} O_t(c,i,j,k)q/m \qquad (5)$$

where m means the number of the atoms of compound c which are present in the cell (i, j, k) at time t, and q means the sum of partial atomic charges present in the cell cell −(i, j, k) at time t.

The joint J_q and self-charge occupancy S_q with the most active reference compound R were defined according to the following equations:

$$J_q(c,i,j,k,N) = \sum_{t=0}^{T} O_t(c,i,j,k) \cap O_t(R,i,j,k)q/m \qquad (6)$$

$$S_q(c,R,i,j,k,N) = \sum_{t=0}^{T}\{O_t(c,i,j,k) - [\sum_{t=0}^{T} O_t(c,i,j,k) \cap O_t(R,i,j,k)]\}q/m \qquad (7)$$

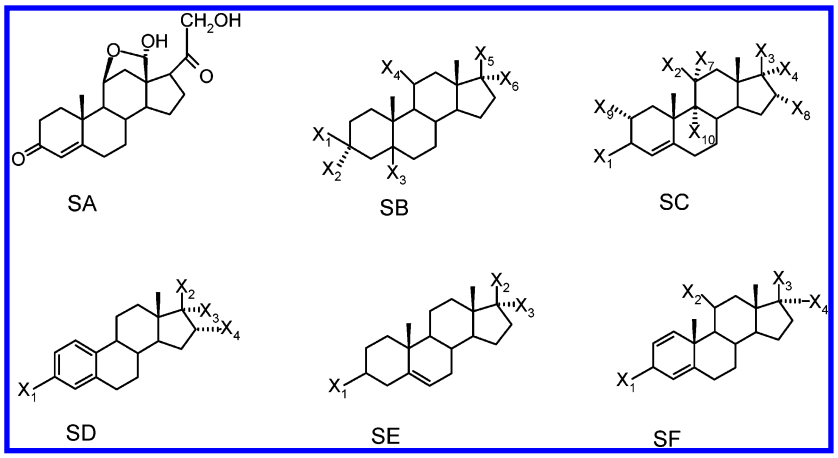
Each alignment produces a unique grid cell occupancy distribution for a given CEP of a molecule.

We used the MATLAB environment to program the calculation of the above-mentioned descriptors.

Partial Least Squares (PLS) method was used to estimate the relationship between independent variables (GCODs) and pK_a values or the corticosteroid binding globulin (CBG) affinity.

SOM-4D-QSAR. Figure 1 compares a single butane conformer defined by the van der Waals surface and a surface formed by a polycubic grid used in 4D-QSAR, respectively.

Table 2. Steroid Structures and the CBG Affinity According to Ref 20



no.	molecular formula	X^1	X^2	X^3	X^4	X^5	X^6	X^7	X^8	X^9	X^{10}	CBG affinity
1s	SA											−6.279
2s	SB	OH	H	H[a]	H	OH	H					−5.000
3s	SE	OH	OH	H								−5.000
4s	SC	=O	H	=O				H	H	H	H	−5.763
5s	SB	H	OH	H[a]	H	=O						−5.613
6s	SC	=O	OH	$COCH_2OH$	H			H	H	H	H	−7.881
7s	SC	=O	OH	$COCH_2OH$	OH			H	H	H	H	−7.881
8s	SC	=O	=O	$COCH_2OH$	OH				H	H	H	−6.892
9s	SE	OH	=O									−5.000
10s	SC	=O	H	$COCH_2OH$	H			H	H	H	H	−7.653
11s	SC	=O	H	$COCH_2OH$	OH			H	H	H	H	−7.881
12s	SB	=O		H[a]	H	OH	H					−5.919
13s	SD	OH	OH	H	H							−5.000
14s	SD	OH	OH	H	OH							−5.000
15s	SD	OH	=O		H							−5.000
16s	SB	H	OH	H[b]	H	=O						−5.255
17s	SE	OH	COMe	H								−5.255
18s	SE	OH	COMe	OH								−5.000
19s	SC	=O	H	COMe	H			H	H	H	H	−7.380
20s	SC	=O	H	COMe	OH			H	H	H	H	−7.740
21s	SC	=O	H	OH	H			H	H	H	H	−6.724
22s	SF	=O	OH	$COCH_2OH$	OH							−7.512
23s	SC	=O	OH	$COCH_2OCOMe$	OH			H	H	H	H	−7.553
24s	SC	=O	=O	COMe	H				H	H	H	−6.779
25s	SC	=O	H	$COCH_2OH$	H			OH	H	H	H	−7.200
26s	SC[c]	=O	H	OH	H			H	H	H	H	−6.144
27s	SC	=O	H	COMe	OH			H	OH	H	H	−6.247
28s	SC	=O	H	COMe	H			H	Me	H	H	−7.120
29s	SC[c]	=O	H	COMe	H			H	H	H	H	−6.817
30s	SC	=O	OH	$COCH_2OH$	OH			H	H	Me	H	−7.688
31s	SC	=O	OH	$COCH_2OH$	OH			H	H	Me	F	−5.797

[a] 5-α. [b] 5-β. [c] H instead of Me at the C_{10}.

By changing the grid resolution we obtain a series of fuzzy molecular representations. As we have shown previously the SOM network can be used for fuzzification of molecular surfaces;[12,20,25,26] therefore, it can also be applied to generate a fuzzy 4D-QSAR-like representation of the molecular conformational space. In such an approach, a sphere defined in space by a single SOM neuron substitutes a unit cube of the Hopfinger method.

Kohonen Mapping. The competitive Kohonen strategy was used to construct a two-dimensional topographic map obtaining the signals from atoms of the conformational states generated by the MD simulations. We used neural network with toroidal neighborhood, as described in details in our previous publications.[20,25,26] Hence, each neuron, j, was defined by three weights, w_{ji}. The competitive training of the network was based on the rule that each atomic coordinate, a, of the conformational state was projected into the neuron, c, with the weights, w_{ci}, that come closest to the Cartesian coordinates, x_{ai}, of this point, a (e.g. 4).

The Cartesian coordinates (x,y,z) of all atoms generated during MD simulation are used as the input vectors of the SOM network. During training these vectors are distributed among output neurons, and such a network is used to form an atomic occupancy SOM. Accordingly, the output neurons obtain a value of 1 (any input signal is projected to those neurons) or 0 (no input signal is projected to those neurons). Alternatively, the addition of the number of incoming vectors is performed in the output neurons of the sum_occupancy maps. The projection of the atomic partial charges to the SOM network, i.e., partial atomic charges of the respective atoms, are projected and averaged in the respective SOM maps, providing the charge type maps.

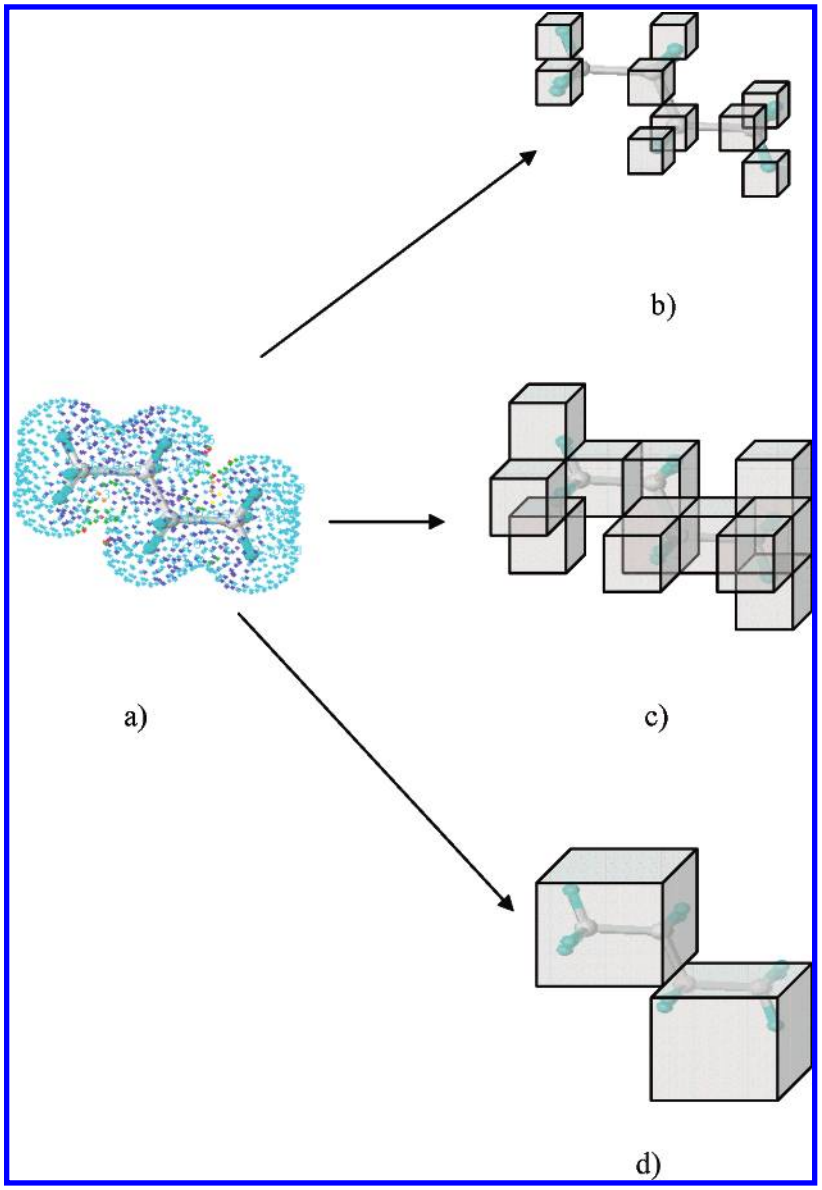


Figure 1. The comparison of the molecular representation defined by the van der Waals molecular surface (a) and that specified by a set of unit cubes of the different resolution (b−d).

Comparative Kohonen Mapping. In fact, such a map illustrates the property (e.g., the occupancy or partial atomic charges) of the conformational ensemble of a single molecule. As however, the weights of the Kohonen network also contain the knowledge on the conformational ensemble profile of the molecule used during training (template molecule), then it can be used to compare the data from another molecule. In such a method the trained Kohonen network processes the signals coming from other molecule(s), i.e., the occupancy or charges are projected through the network to the output neurons to obtain a series of comparative maps characterizing conformational space of the compounds analyzed. Molecules were superimposed before processing through the network using the same alignment procedure as in the grid 4D-QSAR. Figure 2 illustrates a scheme of such an application of the comparative Kohonen network for the SOM-4D-QSAR method.

PLS Analysis. The data generated by the methods discussed above were processed by the PLS analysis with a leave-one out cross-validation procedure. We used the performance metrics that are accepted and widely exercised in COMFA analyses, i.e., cross-validated q^2_{cv}

$$q^2_{cv} = 1 - \frac{\sum(\text{obs}_i - \text{pred}_i)^2}{\sum(\text{obs}_i - \text{mean(obs)})^2} \tag{8}$$

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

$$s = \sqrt{\frac{\sum(\text{obs}_i - \text{pred}_i)^2}{m - k - 1}} \tag{9}$$

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

$$\text{SDEP} = \sqrt{\frac{\sum(\text{pred}_i - \text{obs}_i)^2}{n}} \tag{10}$$

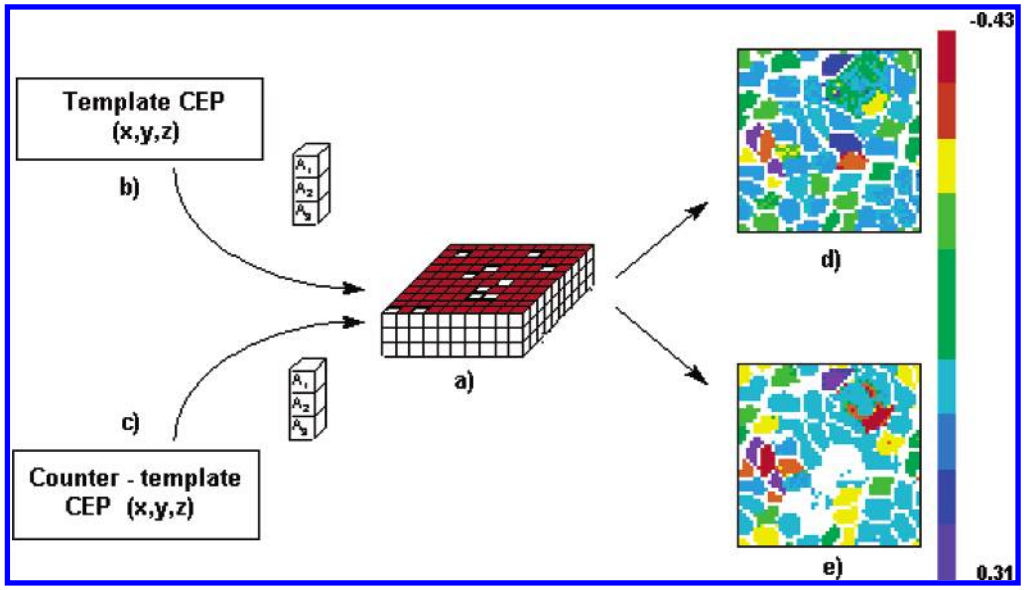


Figure 2. The comparative Kohonen network (a) designed to learn and process the atomic coordinates (*x*,*y*,*z*) of each conformation state generated for the template molecule (b) and countertemplate (c). A pair of comparative patterns (d) and (c) resulted. CEP means conformational ensemble profile and the colors code the mean atomic partial charge.

where pred represents the predicted value, obs represents the observed value, and *n* represents the number of compounds included in the test series.

RESULTS AND DISCUSSION

Modeling Steric and Electronic Control of the Benzoic pK_a Values by CoMFA Analysis. CoMFA is a general method that allows us to analyze the influence of the molecular shapes upon molecular effects. Basically, the assumption that chemical or biological behavior can be explained by the analysis of the noncovalent interactions is the main postulate of this method, and the spatial pattern of the molecular field is a basis for such analysis. However, molecular fields generated in CoMFA depend on electrostatic and shape factors. This fact is well-known, and the relative importance of the electrostatic vs shape interactions in the molecular environment can be, for example, controlled in CoMFA by the selection of the proper atomic probe. Essentially, the distribution of the potential values on the grid points depends on the atomic partial charges and atomic coordinates within the molecule, and steric and electronic effects within the chemical molecule determine this.

The Hammett equation scaled on the ionization of *p*- and *m*-substituted benzoic acids is an approach providing a measure of electronic influence of a certain group substituting the parent benzoic acid. Although thermodynamically pK_a is a bulk property, it has been found that the application of 3D-QSAR that includes conformational information allows for better prediction of both pK_a or Hammett constants than 2D-QSAR and Kim and Martin modeled Hammett constants using CoMFA method.[27−29] Although reliable models were described by these authors, Hammett constants are defined for *m*- and *p*-benzoic acids only. In our previous publication[30] we reported the results on the application of Comparative Molecular Surface Analysis (CoMSA) for predicting pK_a values for a series of the *o*-, *m*-, and *p*-benzoic acids. Although we also performed CoMFA analyses to compare the CoMSA results, we did not investigate factors that limit the modeling ability of steric and electronic effects. In fact, CoMFA literature lacks a systematic direct analysis of modeling efficiency of steric and electronic effects in a sense that these terms are used in organic chemistry. It is clear that generally a type of molecular field and superimposition applied in CoMFA should influence this efficiency. A question appears, however, what are the decisive factors here and what are interrelations between intramolecular patterns directly affected by electronic and steric effects and the molecular fields generated by CoMFA.

Below we present the results of such a systematic study. Since the steric and electronic control of the pK_a values of benzoic acids is probably the effect that is described most precisely in the literature, we attempted to model these effects using the benzoic acid series.[19] This series has been sampled randomly without any effort to control pK_a values of the compounds or any other molecular features of these compounds. Actually, in real 3D-QSAR the situation usually encountered is that it is the compound availability rather than rational basis that determines their embodiment in the model. To mimic different levels of steric and electronic control we divided the series into three subseries, using the substitution pattern (*p*- and *m*- or *o*-) as the obvious selection criterion. Hence, the first set (**series 1:** compounds **1a**−**41a**) was formed of all the molecules, then a second one (**series 2**: compounds **1a**−**24a**) contained *p*- and *m*-substituted analogues, and the third (**series 3** compounds **1a**−**30a**) included molecules that are not monosubstituted *o*-analogues. We formed the latter series by the assumption that the increase of the share of polysubstituted analogues should also increase the steric influences.

We conducted the systematic CoMFA analyses observing the q^2 performances while changing the molecular probes, i.e., H(+), CH_3(+), and CH_3(0). We applied two basic modes of superimposition, i.e., by covering the ring atoms (modes: **A**, **B**) or carboxylic group atoms (modes: **C**, **D**), respectively. The results are shown in Table 3. It can be observed that both the type of atomic probe and the molecular superimposition mode significantly affects the q^2 values. However,

Table 3. CoMFA Analysis for the Selected Series of Benzoic Acids **1a**−**41a**

		molecular probe					
		$CH_3(+1)$		H(+1)		$CH_3(0)$	
superimposition mode	acid series/model	q^2(onc)[a]	*s*	q^2(onc)[a]	*s*	q^2(onc)[a]	*s*
ring (A)	**1/3D−A1**	0.78(3)	0.45	0.81(2)	0.41	0.47(3)	0.70
	2/3D−A2	0.61(2)	0.25	0.52(2)	0.29	−0.02(1)	0.40
	3/3D−A3	0.89(4)	0.35	0.86(3)	0.38	0.75(2)	0.51
ring (B)	**1/3D−B1**	0.78(3)	0.45	0.81(2)	0.41	0.50(3)	0.69
	2/3D−B2	0.61(2)	0.25	0.62(2)	0.39	−0.02(2)	0.40
	3/3D−B3	0.89(4)	0.35	0.85(3)	0.35	0.75(2)	0.50
carboxylic (C)	**1/3D−C1**	0.70(5)	0.54	0.70(5)	0.53	0.24(5)	0.86
	2/3D−C2	0.61(2)	0.25	0.62(2)	0.25	−0.02(1)	0.40
	3/3D−C3	0.63(4)	0.63	0.83(4)	0.42	0.59(5)	0.69
carboxylic (D)	**1/3D−D1**	0.71(5)	0.53	0.69(6)	0.56	0.20(5)	0.89
	2/3D−D2	0.61(2)	0.25	0.62(3)	0.29	−0.02(1)	0.40
	3/3D−D3	0.64(4)	0.63	0.84(4)	0.42	0.58(5)	0.70

[a] (onc) − optimal number of the PLS components.

the change of the atoms to be covered in the ring or carboxylic superimposition modes does not change the quality of the models. This can be expected from the analysis of the conformations generated that follows. In particular, the planarity of the benzene ring makes both ring superimpositions identical.

The values of the CoMFA q^2 performances with the $CH_3(0)$ probe clearly differentiate subseries **1**, **2**, and **3**. The $CH_3(0)$ probe should account for the molecular shapes and it in fact differentiates the relative importance of steric effects. Therefore, this influence sequentially increases as follows **series 2** ($q^2 = -0.02$) < **series 1** ($q^2 = 0.20 \div 0.50$) < **series 3** ($q^2 = 0.58 \div 0.75$). This means that the CoMFA with the $CH_3(0)$ probe does not allow for predictive modeling ($q^2 < 0$) of the pK_a values for **series 2**, which indicates that pK_a values are controlled by electrostaticity, implying consequently the significant dominance of electronic effects. In this particular case, regardless of the superimposition mode the $CH_3(0)$ probe does not provide predictive models. The best models for **series 2** are obtained using H(+) and $CH_3(+)$ probes. In such a situation the q^2 performances do not depend on molecular superimposition mode, i.e., **a carboxylic superimposition** performed by covering the atoms of carboxylic group, or **a ring superimposition**—performed by covering the benzene ring atoms. This is quite clear, because a carboxylic group can take a conformation fully coplanar with the ring. Thus, both superimposition modes provide indistinguishable patterns of the molecular fields generated for the compound series.

It is, however, not the case for the two series containing *o*-analogues, i.e., series **1** and **3**. In Figure 3 we show the superimposition of the molecules for all three series **1**−**3**. This documents that steric hindrance can prevent coplanarity of the aromatic ring and carboxylic function. Such a distortion can be observed in the series **1** and **3** which include *o*-analogues but not in the series **2** containing *p*- and *m*-analogues only. Consequently, in the series **1** and **3** both superimposition modes result in different q^2 performances. Now a question of how in such cases a superimpositon should be performed to obtain optimal pK_a models needs to be answered. Since the carboxylic group is of key importance for ionization, it may look as if it is this group that should be covered. There are no general rules solving such problems, and it is the CoMFA experiment that provided us with an answer. It can be found in Table 3 that ring superimposition is favored and provides better results, e.g. models **3D**−**A1** or **3D**−**B1** are superior to **3D**−**C1** or **3D**−**D1**. Thus, the maximum cover of the carboxylic function does not lead to the optimal CoMFA model. The important conclusion is that functional group superimposition is not the best solution for the situation discussed.

In addition, if we compare modeling ability for individual series, it is quite surprising that the highest q^2 performances, i.e., the best models, can be obtained for the series **3** that include the analogues for which steric effects are relatively important (models **3D**−**A3** and **3D**−**B3**). This indicates that in CoMFA the variance of the pK_a values resulting from the modification of the molecular shape can be modeled more efficiently than the one resulting exclusively from the distribution of charges. Moreover, also series **1** that includes all *o*-analogues together with *m*- and *p*-acids allows for better modeling of pK_a values (models **3D**−**A1** and **3D**−**B1**) than *m*- and *p*-analogues alone (models **3D**−**A2** and **3D**−**B2**). However, regardless of the series analyzed, the best models can be obtained in CoMFA for the charge loaded probes. This implies the importance, if not the dominance, of the charge distribution and consequently electronic effects even for o-analogues.

4D-QSAR Formalism for Modeling Benzoic Acid pK_a Values and Steroid CBG Affinity. Table 4 illustrates the results of Hopfinger's 4D-QSAR analysis performed for benzoic acid series **1**, **2**, and **3**. It is worth mentioning that 4D-QSAR is an extremely time-consuming process. Although we did not register the precise CPU time, a single series of simulations need a few days on our SGI O2 workstation. Unlike in CoMFA, the individual superimposition mode (the indication of a certain atom triplet to be covered) in the ring or carboxylic superimposition type does influence the modeling ability. We can observe that generally the absolute type descriptors perform much better than the joint type ones.

The comparison of the influence of the ring and carboxylic superimposition modes on the modeling ability for the best models given by the A_o descriptors indicates that within series **3** controlled by steric effects the influence of the alignment mode is much less significant than in corresponding CoMFA models (compare models **4D**−**A3**$_{Ao}$, **4D**−**B3**$_{Ao}$, **4D**−**C3**$_{Ao}$, and **4D**−**D3**$_{Ao}$ to the respective **3D** counterparts). Thus, in

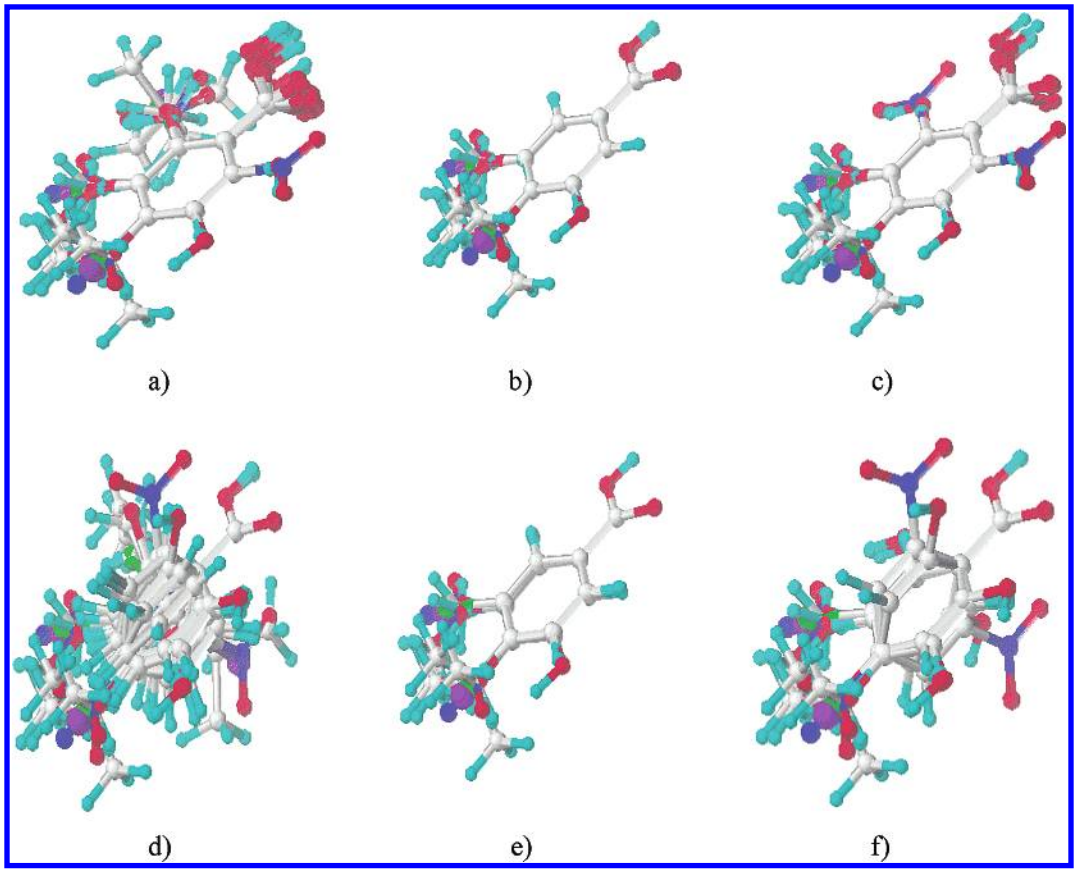


Figure 3. The ring (upper line) and carboxylic (lower line) superimpositions modes of the bezoic acid series **1** (a, d), **2 (**b, e), and **3 (**c, f). Details in text.

Table 4. 4D-QSAR Models of the pK_a Values for the Selected Series of Benzoic Acids **1a**−**41a**

superimposition mode	acid series/model	4D-QSAR descriptors[a]											
		A_o		A_q		J_o [b]		J_q [b]		S_o [b]		S_q [b]	
		q^2(onc)[c]	s	q^2(onc)[c]	s	q^2(onc)[c]	s	q^2(onc)[c]	s	q^2(onc)[c]	s	q^2(onc)[c]	s
ring (A)	**1/4D−A1**	0.40(9)	0.81	0.45(5)	0.73	0.04(5)	0.73	0.11(7)	0.96	0.47(9)	0.77	0.44(5)	0.74
	2/4D−A2	0.67(10)	0.30	0.45(2)	0.30	−0.14(1)	0.43	0.24(1)	0.35	0.47(10)	0.37	0.44(2)	0.32
	3/4D−A3	0.91(10)	0.37	−0.18(2)	1.10	0.07(4)	1.01	0.46(10)	0.90	0.89(10)	0.41	−0.18(2)	1.12
ring (B)	**1/4D−B1**	0.58(7)	0.66	0.47(10)	0.78	−0.02(6)	1.01	0.33(3)	0.86	0.62(8)	0.65	0.41(10)	0.83
	2/4D−B2	0.58(6)	0.29	0.35(1)	0.32	−0.12(1)	0.43	0.33(2)	0.33	0.42(6)	0.35	0.37(1)	0.32
	3/4D−B3	0.89(10)	0.39	0.05(6)	1.07	−0.10(5)	1.13	0.53(10)	0.83	0.81(8)	0.53	0.01(7)	1.14
carboxylic (C)	**1/4D−C1**	0.61(9)	0.65	0.40(4)	0.75	0.32(3)	0.79	0.16(3)	0.88	0.64(10)	0.64	0.40(4)	0.76
	2/4D−C2	0.26(3)	0.36	0.38(2)	0.32	−0.13(1)	0.43	0.34(2)	0.32	0.45(3)	0.32	0.40(2)	0.32
	3/4D−C3	0.79(8)	0.52	−0.13(6)	1.17	0.34(3)	0.84	0.14(2)	0.94	0.81(10)	0.54	−0.22(3)	1.16
carboxylic (D)	**1/4D−D1**	0.17(1)	0.85	0.67(9)	0.59	0.06(3)	0.93	0.22(10)	0.94	0.16(1)	0.87	0.65(5)	0.62
	2/4D−D2	0.08(3)	0.40	0.41(2)	0.38	−0.11(1)	0.42	0.22(2)	0.36	0.19(4)	0.39	0.39(1)	0.32
	3/4D−D3	0.87(10)	0.44	−0.13(9)	1.17	0.09(3)	0.98	0.15(8)	1.06	0.88(8)	0.40	−0.12(6)	1.18

[a] Box 20 Å:20 Å:20 Å, grid size 1 Å. [b] Compound **24a** was used as a reference compound. [c] (onc) − optimal number of the PLS components.

this particular case (steric control) 4D-QSAR with A_o descriptor is quite independent of the molecular superimposition mode.

Although we simulated the data for benzoic acids in such a way that steric effects were of different importance for the three series, it is evident that electrostatic interactions evidently control pK_a values. The CBG steroids are compounds for which shape is a dominating factor that determines the binding affinity.[20,24,26,31] Thus, we tried to apply 4D-QSAR method to model CBG affinity. Table 5 shows the results of such an attempt. The comparison of the CoMFA and 4D-QSAR approaches shows that for the best model (including molecules **1s**−**21s)** the 4D-QSAR performance ($q^2 = 0.84$, $s = 0.50$) slightly outperforms the modeling ability reported for the corresponding series in CoMFA ($q^2 = 0.73$, $s = 0.64$).[32]

Although the comparison of the occupancy descriptors (A_o, J_o, S_o) with those containing charge information (A_q, J_q, S_q) indicates that the occupancy descriptors always explain the affinity more efficiently, these differences are much smaller than in the benzoic acid series.

SOM-4D-QSAR. Hopfinger's 4D-QSAR provides results that compare well to the CoMFA for the steroid series analyzed. However, the performance was inferior to that of CoMSA ($q^2_{cv} = 0.88$, $s = 0.42$) where we used a self-organizing map for generating a fuzzy molecular representation.[20] Thus below we report the results of the procedure that includes the SOM neural network as a central element

Table 5. 4D-QSAR Models for the CBG Binding Steroids

	4D-QSAR descriptors[a]											
	A_o		A_q		J_o [b]		J_q [b]		S_o [b]		S_q [b]	
model	q^2(onc)[c]	s	q^2(onc)[c]	s	q^2(onc)[c]	s	q^2(onc)[c]	s	q^2(onc)[c]	s	q^2(onc)[c]	s
A[d]	0.69(2)	0.63	0.63(1)	0.67	0.65(2)	0.66	0.60(1)	0.70	0.66(2)	0.64	0.60(1)	0.68
B[d]	0.81(2)	0.50	0.72(2)	0.60	0.71(2)	0.60	0.68(1)	0.63	0.80(2)	0.50	0.70(2)	0.61
C[d]	0.84(2)	0.50	0.71(2)	0.68	0.80(2)	0.56	0.63(1)	0.73	0.80(2)	0.54	0.69(2)	0.67

[a] Box 20 Å:20 Å:20 Å, grid size 1 Å. [b] Compound **6s** was used as a reference compound. [c] (onc) − optimal number of the PLS components. [d] Model A: compounds **1s**−**31s**, model B: compounds **1s**−**30s**, model C: compounds **1s**−**21s**.

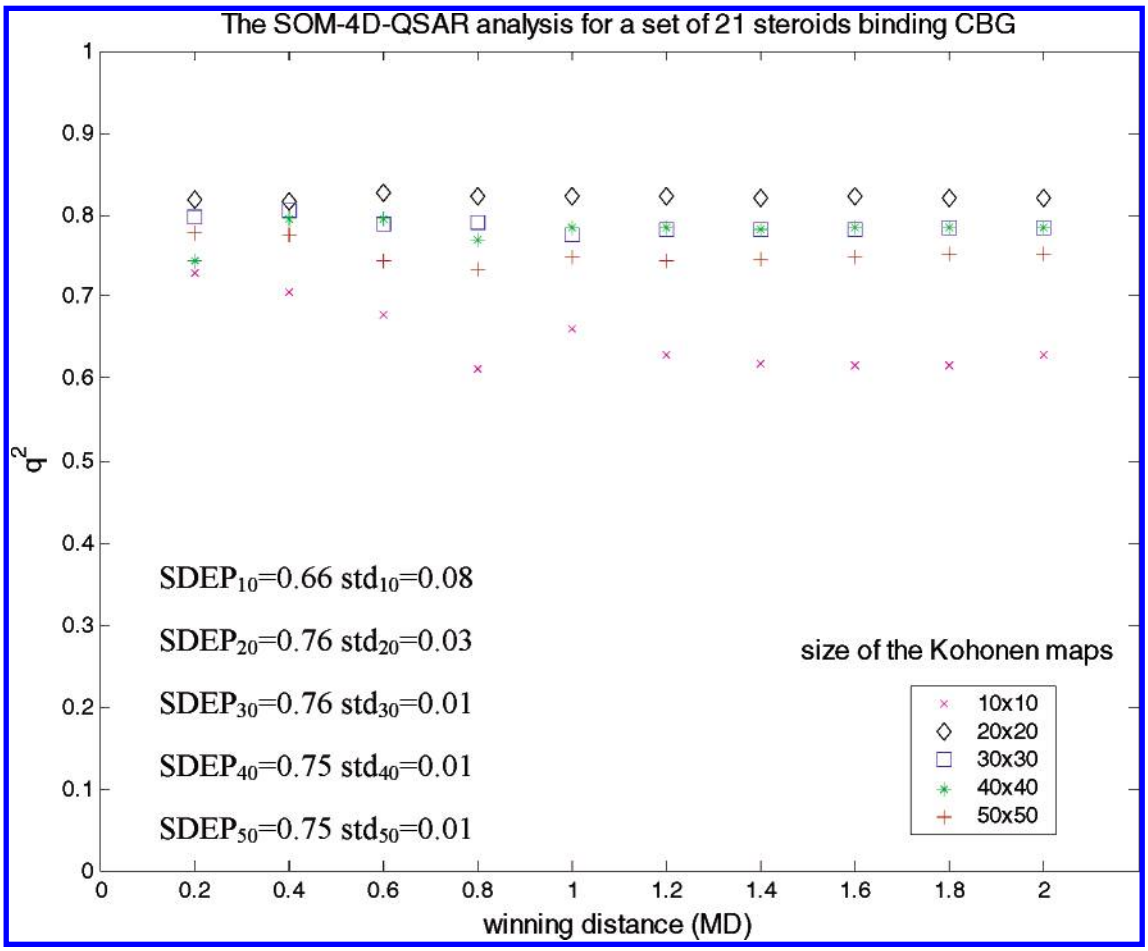


Figure 4. The dependence of the 4D-QSAR q^2 performances on the winning distance and the Kohonen map size while modeling CBG affinity of steroids **1s**−**21s** with the charge type maps. The SDEP errors given for each map size refer to mean values (10 simulations, std − standard deviation) and are for compounds **22s**−**31s**. The comparison indicates that these values are slightly better then CoMFA SDEP for the same series (0.837); compare ref 35 for this and further SDEP error values referring to the other 3D-QSAR models of this series.

of the coupled SOM-4D-QSAR system. The description of the method is given in the Experimental Section. Formally, we can draw an analogy between this scheme and 4D-QSAR, by the geometrical interpretation of the SOM network. Thus, a sphere defined in space by the weights of a single SOM neuron substitutes a unit cube of the Hopfinger method.[10,21−23] A diameter of this sphere (neuron winning distance) is a parameter that resembles the grid resolution. Figure 4 illustrates the performance of the SOM-4D-QSAR as a function of the winning distance (MD) and the size of the self-organizing map (M). For the maps larger than 10 × 10 the results are stable. As can be also compared in Figure 4, the external prediction validated by the SDEP estimator for the CBG series compounds **22s**−**31s** in SOM-4D-QSAR are also better than that calculated for CoMFA method.[33] Table 6 specifies the results obtained with the application of the SOM-4D-QSAR for modeling pK_a values of the acid series **1**−**3**. Within the series of benzoic acids the SOM-4D-QSAR performs better than the grid technique for the whole compound set—series **1** (q^2 = 0.72, s = 0.55 (**SOM-4D-C1$_o$**) vs q^2 = 0.61, s = 0.65 (**4D-C1$_{Ao}$**)). Both methods provide similar results for series **3** (q^2 = 0.91, s = 0.31 (**SOM-4D-A3$_o$**) vs q^2 = 0.91, s = 0.37 (**4D-A3$_{Ao}$**)), and **4D-QSAR** is more effective for the series **2** (q^2 = 0.59, s = 0.32 (**SOM-4D-C2$_o$**) vs q^2 = 0.67 s = 0.30 (**4D-A2$_{Ao}$**)). However the SOM-4D-QSAR provides significantly better results if one includes charges, i.e. q^2 = 0.55, s = 0.28 (**SOM-4D-C2$_q$**) vs q^2 = 0.45, s = 0.30 (**4D-A2$_{Aq}$**) for series **2** and q^2 = 0.60, s = 0.65 (**SOM-4D-C3$_q$**) vs q^2 = 0.05, s = 1.07 (**4D-B3$_{Aq}$**) for series **3**.

The Efficiency of the 3D and 4D-QSAR Modeling of Steric and Electronic Effects. Modeling molecular effects

Table 6. SOM-4D-QSAR Models of the pK_a Values for the Selected Series of Benzoic Acids **1a**−**41a** Obtained for the Kohonen Maps Containing 400 (20 × 20) Neurons

		SOM-4D-QSAR$_o$ [a]			SOM-4D-QSAR$_q$ [b]			SOM-4D-QSAR$_r$ [c]		
superimposition mode	model	q^2(onc)[d]	*s*	md	q^2(onc)[d]	*s*	md	q^2(onc)[d]	*s*	md
ring (A)	**4D−A1**	0.63(4)	0.59	1.8	0.68(4)	0.54	2.0	0.57(3)	0.63	1.4
	4D−A2	0.33(8)	0.39	1.0	0.53(2)	0.29	0.8	0.36(5)	0.35	1.8
	4D−A3	0.91(5)	0.31	1.4	0.44(3)	0.77	1.2	0.85(6)	0.41	1.8
ring (B)	**4D−B1**	0.60(5)	0.62	1.8	0.52(4)	0.71	0.6	0.58(3)	0.62	0.6
	4D−B2	0.42(2)	0.31	1.2	0.41(2)	0.31	0.6	0.03(1)	0.39	2.0
	4D−B3	0.87(5)	0.39	1.6	0.46(3)	0.76	1.8	0.83(4)	0.44	0.8
carboxylic (C)	**4D−C1**	0.72(8)	0.55	1.0	0.64(6)	0.60	1.2	0.57(4)	0.64	1.0
	4D−C2	0.59(9)	0.32	1.0	0.55(4)	0.28	0.8	0.15(1)	0.36	1.0
	4D−C3	0.73(7)	0.58	1.0	0.60(3)	0.65	1.4	0.61(2)	0.62	2.0
carboxylic (D)	**4D−D1**	0.57(4)	0.64	1.6	0.63(5)	0.60	2.0	0.29(1)	0.79	1.2
	4D−D2	0.37(3)	0.33	1.2	0.33(2)	0.33	0.6	0.01(1)	0.40	1.0
	4D−D3	0.80(6)	0.48	1.4	0.57(7)	0.74	2.0	0.62(3)	0.64	1.6

[a] sum_occupancy (with the addition of vectors incoming to output neurons). [b] Mean charges. [c] Occupancy (0 or 1). [d] (onc) − optimal number of the PLS components.

Table 7. Comparison of the 4D-QSAR and SOM-4D-QSAR Analysis for the CBG Binding Steroids

	A_o [a]		J_o [a]		SOM-4D-QSAR$_o$ [a]		A_q [b]		J_q [b]		SOM-4D-QSAR$_q$ [b]	
model	q^2(onc)[c]	*s*	q^2(onc)[c]	*s*	q^2(onc)[c]	*s*	q^2(onc)[c]	*s*	q^2(onc)[c]	*s*	q^2(onc)[c]	*s*
A[d]	0.69(2)	0.63	0.65(2)	0.66	0.68(4)	0.65	0.63(1)	0.67	0.60(1)	0.70	0.64(3)	0.68
B[d]	0.81(2)	0.72	0.71(2)	0.60	0.80(4)	0.52	0.72(2)	0.60	0.68(1)	0.63	0.78(4)	0.54
C[d]	0.84(2)	0.71	0.80(2)	0.56	0.86(3)	0.47	0.72(2)	0.68	0.63(1)	0.73	0.83(4)	0.54

[a] Occupancy (with the addition of vectors incoming to output neurons). [b] With mean charge values in the output neurons. [c] (onc) − optimal number of the PLS components. [d] Model A: compounds **1s**−**31s**, model B: compounds **1s**−**30s**, model C: compounds **1s**−**21s**.

or biological activity evolved by the series of effector or ligand structures is still a challenge. It is not surprising because a molecule interacting with the environment forms a complicated system, and any quantitative description that must do without information on that environment is often an oversimplification. There are many issues that contribute to this, but focusing on the congeneric effector structures, the separation of steric, electronic, and hydrophobic factors constitute a large barrier. Accordingly, this fact is concluded by Hansch: *too often little attention is paid to choosing congeners that will as much as possible resolve these three major substituent effects.*[34]

Generally it is believed in QSAR that modeling steric effects is much more complicated than the electronic ones.[34] This can be explained by the fact that even a minute change in a molecule shape can completely change its interactions with the environment, especially with biological receptors. Even a 2D approach aimed at modeling benzoic acid pK_a values requires a special trick of including the so-called proximity electronic effect.[35] However, it is not a case that we observed in this study for the benzoic acid series. The high efficiency of modeling steric control of benzoic pK_a values, that we observed, results probably from the fact that acid ionization is determined by the interactions of the molecule with a solvent system that is much less shape restrictive than biological receptors. On the other hand, the shape changes induce larger differences in the molecular field which can be easier described in quantitative terms.

If we compare the modeling efficiency of 3D- and 4D-QSAR for the benzoic acid of series **1**, irrespective of the descriptor type, the q^2 performances in 4D-QSAR are generally slightly lower than those obtained in CoMFA. However, the q^2 performance of the $\mathbf{A_o}$ descriptor reaches a value of 0.91 (Table 4: model **4D−A3$_{Ao}$**) and 0.67 (model **4D−A2$_{Ao}$**) for series **2** and **3**, respectively. This means that in this particular case both 3D- and 4D-QSAR modeling provide at least comparable results. Since, however, more PLS components are needed to construct the optimal models, namely Table 4: models **4D−A3$_{Ao}$** and **4D−A2$_{Ao}$** than the respective CoMFA models (Table 3: **3D−A3** and **3D−A2**) preference should rather be given to the latter ones. The comparison of the **3D** and **4D B-**row models indicates similar regularities. Contrary to CoMFA, in the 4D-QSAR modeling of pK_a the influence of the alignment mode for series **2** (electronic control) is quite clear. At the same time the ring superimposition is more efficient for 4D-QSAR pK_a modeling. This can be explained probably by the conformational lability of the carboxylic group. Thus, in a collection of conformations which are generated during the molecular dynamics simulations there are also a large number of such conformers that differ from those with benzene coplanar COOH, i.e., those that are inadequate for the explanation of pK_a controlled in this case by the electronic effect. As the benzene ring includes more atoms than the carboxylic group, then superimposition by carboxylic atoms puts a larger part (ring atoms) of individual conformers in unreliable positions. Although the same applies to the ring superimposition, here only the spatial occupancy of the minority of atoms (carboxylic ones) are far from the "active" pattern.

Moreover, quite surprisingly the inclusion of the charges in 4D-QSAR (A_q parameter) does not improve the model quality even in **series 2**. However, this makes the models much less dependent upon superimposition modes (compare models **4D−A2$_{Aq}$**, **4D−B2$_{Aq}$**, **4D−C2$_{Aq}$**, and **4D−D2$_{Aq}$** in Table 4) in comparison to the occupancy descriptors (A_o). Even though it might have seemed that the 4D-QSAR scheme with the occupancy descriptors would be especially suitable for modeling steric effects, it also performs quite

well for series **2** that is limited by electronic effects. The investigations of the CBG 4D-QSAR models reveal that although in this particular case the occupancy GCODs are also superior, quite surprisingly, these differences are much smaller than in the benzoic acid series, where we expected to observe the biggest advantage of the inclusion of charge information. We cannot provide a conclusive explanation for this fact. However, we believe that the quite rigid steroid skeleton during MD simulations prevents significant distortions from the reliable active conformational states. Consequently, the distribution of charges is also always reliable. On the other hand, this result indicates explicitly that electrostaticity (electronic factors) can also account for the CBG affinity.

Table 7 compares the models obtained while using 4D-QSAR and SOM-4D-QSAR to model the CBG affinity of steroid compounds. The best SOM-4D-QSAR steroid model is described by the performances of $q^2 = 0.86$, $s = 0.47$ with 3 PLS components (sum_occupancy type map) or $q^2 = 0.83$, $s = 0.54$ with 3 PLS components for a charge type map. For the comparison the best grid 4D-QSAR model yields $q^2 = 0.84$ and $s = 0.50$ for 2 PLS components, respectively, for the A_o descriptor. This means that in this particular case both methods provide very similar results that slightly outperform CoMFA ($q^2 = 0.73$, $s = 0.64$).[32] However, if we compare the performance for the J_q and J_o descriptors that are the closest SOM counterpart in the grid method, then the SOM-4D-QSAR provides better results.

We hope that the modification of the SOM scheme to give the descriptors similar to the absolute type 4D-QSAR descriptors will allow us to improve further the efficiency of this method. Moreover, the big advantage of SOM-4D-QSAR is that although different alignment modes influence model quality, this effect is much less pronounced than in CoMFA or 4D-QSAR.

CONCLUSIONS

We conducted a systematic study of the performance of the 3D- and 4D-QSAR schemes in modeling steric and electronic effects. In particular, we compared the influence of the coding system on the efficiency of QSAR modeling of the CoMFA and Hopfinger's 4D-QSAR schemes. Hence, we attempted to predict the pK_a values of (*o*-, *m*-, and *p*-)benzoic acids which were divided into three subseries in order to simulate different levels of steric and electronic control. The steroids binding to CBG were used as model series where biological activity is limited by shape factors. Although individual models differ depending upon the individual scheme, generally, both CoMFA and 4D-QSAR appeared to provide comparable results, irrespective of the differences in the coding scheme used for the description. Moreover, a new 4D-QSAR scheme involving a self-organizing neural network was designed. Generally, the SOM scheme that we designed performs comparably to the grid scheme.

ACKNOWLEDGMENT

The authors would like to thank Prof. Johann Gasteiger of the University of Erlangen-Nuernebrg, BRD for facilitating access to his program package including CORINA, PETRA, SURFACE, and KMAP. Financial support of the KBN Warsaw: Grant no. PBZ KBN - 040 P04/08 and 4T09A 088 25 is gratefully acknowledged.

Supporting Information Available: Data concerning Tables 3−5 obtained by alternative calculations using different parameters. 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 approach. In *Methods and principles in medicinal chemistry*; Mannhold, R., Krokgsgaard-Larsen, P., Timmerman, H., Eds.; VCH: Weinheim, 1993.
(4) Katrizky, A. R.; Maran, U.; Labanov, V. S.; Karelson, M. Structurally diverse quantitative structure−property relationship correlations of technology relevant physical properties. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1−18.
(5) Kim, K. H.; Greco, G.; Novellino, E. A critical review of the recent CoMFA applications. *Perspect. Drug Discov. Des.* **1998**, *12/13/14*, 257−315.
(6) Kroemer, R. T.; Hecht, P.; Guessregen, S.; Liedl, K. R. Improving the quality of CoMFA models. *Perspect. Drug Discov. Des.* **1998**, *12/13/14*, 41−56.
(7) Martin, Y. C. 3D QSAR: Current state, scope and limitations. *Perspect. Drug Discov. Des.* **1998**, *12/13/14*, 3−23.
(8) Norinder, U. Recent progress in CoMFA methodology and related techniques. *Perspect. Drug Discov. Des.* **1998**, *12/13/14*, 25−39.
(9) Vedani, A.; Dobler, M. 5D-QSAR: The key for simulating induced fit? *J. Med. Chem.* **2002**, *45*, 2139−2149.
(10) Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B.; Albuquerque, M.; Madhav, P. J.; Duraiswami, C. Construction of 3D-QSAR models using the 4D-QSAR analysis formalism. *J. Am. Chem. Soc.* **1997**, *119*, 10509−10524.
(11) Hopfinger, A. J. A QSAR investigation of dihydrofolate reductase by Baker triazines based upon molecular shape analysis. *J. Am. Chem. Soc.* **1980**, *102*, 7196−7206.
(12) Polanski, J. Molecular Shape Analysis. In *Handbook of chemoinformatics From data to knowledge*; Gasteiger, J., Ed.; Wiley-VCH: Weinheim, BRD, 2003; pp 302−319.
(13) Motoc. J. *Molecular shape descriptors.* In *Steric effects in drug design*; Charton, M., Motoc, J., Eds.; Akademie Verlage: Berlin, 1983; pp 93−105.
(14) Jain, A. N.; Koile, K.; Chapman, D. Compass-predicting biological activities from molecular-surface properties-performance comparisons on a steroid benchmark. *J. Med. Chem.* **1994**, *37*, 2315.
(15) Manallack, D. T.; Livingstone, D. J. Neural networks in drug discovery- have they lived up to their promise. *Eur. J. Med. Chem.* **1999**, *34*, 195−208.
(16) Polanski, J. Self-organizing neural networks for pharmacophore mapping. *Adv. Drug Delivery Rev.* **2003**, *55*, 1149−1162.
(17) Anzali, S.; Gasteiger, J.; Holzgrabe, U.; Polanski, J.; Sadowski, J.; Teckentrup, A.; Wagener, M. The use of self-organizing neural network in drug design. *Perspect. Drug Discov. Des.* **1998**, *9/10/11*, 273−299.
(18) Zupan, J.; Gasteiger, J. *Neural network and drug design for Chemists*, 2nd ed.; VCH: Weinheim, 1999.
(19) *Physical and chemical data compendium. Poradnik fizykochemiczny*; WNT: Warsaw, 1974; pp 347−351.
(20) Polanski, J.; Walczak, B. The comparative molecular surface analysis (CoMSA): a novel tool for molecular design. *Comput. Chem.* **2000**, *24*, 615−625.
(21) Ravi, M.; Hopfinger, A. J.; Hormann, R. E.; Dinan, L. 4D-QSAR analysis of a set ecdysteroids and a comparison to CoMFA modeling. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1587−1604.
(22) Krasowski, M. D.; Hong, X.; Hopfinger, A. J.; Harrison, N. L. 4D-QSAR analysis of a set propofol analogues: Mapping binding sites for an anestheric phenol on the $GABA_A$ receptor. *J. Med. Chem.* **2002**, *45*, 3210−3221.
(23) Hong, X.; Hopfinger, A. J. 3D-pharmacophores of flavonoid binding at the benzodiazepine $GABA_A$ receptor site using 4D-QSAR analysis. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 324−336.
(24) Polanski, J.; Gieleciak, R. The comparative surface analysis (CoMSA) with modified uninformative variable elimination − PLS (UVE-PLS) method: application to the steroids binding the aromatase enzyme. *J. Chem. Inf. Comput. Sci.* **2002**, *43*, 656−666.

(25) Polanski, J.; Gasteiger, J.; Jarzembek, K.; Self-organizing neural networks for screening and development of novel artificial sweetener candidates. *Comb. Chem. High Throughput Screening* **2000**, *3*, 553−561.

(26) 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. Applications to the analysis of corticosteroid globulin activity of steroids. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 521−540.

(27) Kim, K. H.; Martin, Y. C. Direct Prediction of Dissociation Constants (pK_a's) of Clonidine-like Imidazolines, 2-Substituted Imidazoles, and 1-Methyl-2-substituted-imidazoles from 3D Structures Using a Comparative Molecular Field (CoMFA) Approach. *J. Med. Chem.* **1991**, *34*, 2056−2060.

(28) 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.

(29) Martin, Y. C.; Lin, T. C.; Hetti, Ch.; DeLazzer, J. PLS analysis of distance matrices to detect nonlinear relationships between biological potency and molecular properties. *J. Med. Chem.* **1995**, *38*, 3009−3015.

(30) 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 for benzoic and alkanoic acids. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 184−191.

(31) Polanski, J. The receptor-like neural network for modeling corticosteroid and testosterone binding globulins. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 553−561.

(32) Coats, E. U. The CoMFA steroids as a benchmark dataset for development of 3D QSAR methods. *Prospect. Drug Discov. Des.* **1998**, *12/13/14*, 199−213.

(33) Robinson, D. D.; Winn, P. J.; Lyne, P. D.; Richards. W. G. Self-organizing molecular field analysis (SOMFA): a tool for structure−activity studies. *J. Med. Chem.* **1999**, *42*, 573−583.

(34) Hansch, C.; Leo, A. *Exploring QSAR Fundamentals and Applications in Chemistry and Biology*; American Chemical Society: Washington, DC, 1995; pp 88−93.

(35) Fujita, T. Nishioka, T. The analysis of the ortho effect. *Prog. Phys. Org. Chem*. **1976**, *12*, 49−89.

CI034118L