GRID Formalism for the Comparative Molecular Surface Analysis: Application to the CoMFA Benchmark Steroids, Azo Dyes, and HEPT Derivatives

Jaroslaw Polanski,* Rafal Gieleciak, Tomasz Magdziarz, and Andrzej Bak

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

Received January 21, 2004

Shape analysis is a powerful tool in chemistry and drug design, and molecular surface defines shape in the molecular scale. In the current publication we presented a novel formalism for the comparative molecular surface analysis (s-CoMSA). The method enables both quantitative modeling of 3D-QSAR and finding possible pharmacophoric sites. The method provides very predictive models for the CBG activity of the benchmark steroid series, tinctorial properties of the heterocyclic azo dyes and anti-HIV activity of the HEPT series.

INTRODUCTION

Shape is one of the fundamental categories used by the human brain for the perception and description of 3D objects. It is the particular way in which the external edges or boundaries of objects are connected with each other that determines their shapes. Molecular surfaces model shapes of molecular objects. Although molecular surfaces are only a conventional imitation of the molecular boundaries, it has been shown that such models can often explain fundamental chemical or pharmacological effects; therefore; molecular shape is an important property that determines molecular recognition and drug−receptor interactions.

On the other hand, a number of three-dimensional Quantitative Structure Activity Relationships (3D-QSAR) methods do not use direct shape descriptors. In particular, the Comparative Molecular Field Analysis (CoMFA),[1] the first technique developed for modeling and analyzing 3D-QSARs is probably the most typical example here. This method constructs a spatially uniform 3D field around a series of superimposed molecules to investigate molecular environment, therefore, masking explicit shape information by the regularity of the cubic grid used. On the contrary, other approaches, that include explicit shape information, are well-known in molecular design. Hopfinger's Molecular Shape Analysis,[2] Receptor Surface Models,[3−5] Comparative Receptor Surface Analysis (CoRSA),[6] and Compass[7] are some of the recent methods. Direct comparison of the molecular objects usually needs a special tool that normalizes individual shapes of a series of molecules. Different techniques have been suggested to achieve this. Thus, a virtual receptor represented by the van der Waals spheres is constructed around an entire set of molecules superimposed in the receptor surface model method. Shape grids are constructed in Receptor Surface Analyses, e.g. in the CoRSA method, while the Compass uses a supervised neural network to fit two nonidentical points near the molecular surfaces. Similarly, in the Comparative Molecular Surface Analysis (CoMSA)[8] the unsupervised SOM neural network[9] is applied, that compares two (unnecessarily identical) slices of the molecular surfaces of two different molecules using the ability of the SOM neuron to group the patterns (vectors) located near the template attractor. This method enabled us to model 3D-QSAR by the analysis of various surface properties, e.g. the electrostatic potential. We proved that this technique could be an efficient tool for modeling 3D-QSAR or exploring molecular diversity.[10−16] Generally, modeling and predictive ability of SOM-CoMSA, including the comparative Kohonen neural network[17] coupled with the Partial Least Squares (PLS) method,[18] outperforms this of the CoMFA.[10−14] Recently, Hasegawa has also proved the efficiency of similar SOM-CoMSA schemes in 3D-QSARs.[19−21]

Although the SOM network has certain advantages, as it can both model nonlinear systems and preserve the topology of the objects projected,[22] the indeterministic behavior of neural architecture and the need for the special software packages that realize SOM algorithm can cause some problems. In the current research we developed a nonneural, sector version of the CoMSA (s-CoMSA) based on the grid formalism similar to that of the Hopfinger's 4D-QSAR.[2] This method has been applied for modeling 3D-QSAR of the CoMFA steroid benchmark series,[8] heterocyclic azo dye series,[13,23] and HEPT HIV blocking agents.[24]

EXPERIMENTAL SECTION

Model Builders. All the experimental data, i.e., biological activities for the CoMFA steroids, HEPT analogues, and cellulose affinities for the heterocyclic azo dyes, were extracted from the refs 8, 24, and 23 and are given in Tables 1−3, respectively.

All the molecules were superimposed before the calculation of molecular surfaces. The superimposition was performed by covering:

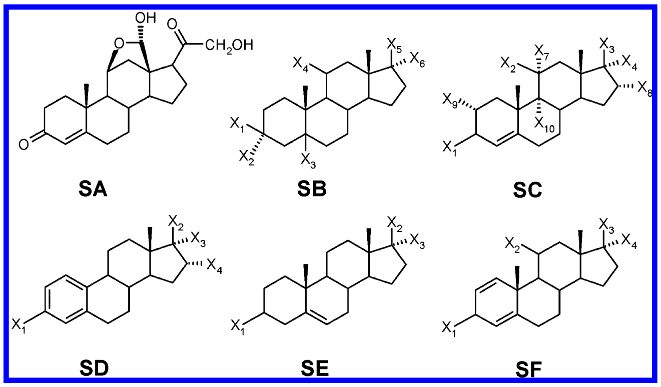
- for molecules (**s1**−**s31**) all non-hydrogen atoms of four steroid rings − molecule s6 (see Table 1),

- for molecules (**d1**−**d21**) different superimposition modes were tested, as discussed in detail further,

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

10.1021/ci049960l CCC: $27.50
Published on Web 07/08/2004

Table 1. Steroid Structures and the CBG Affinity Data[8]



no.	S	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	CBG
s1	SA											−6.279
s2	SB	OH	H	H[a]	H	OH	H					−5.000
s3	SE	OH	OH	H								−5.000
s4	SC	=O	H	=O				H	H	H	H	−5.763
s5	SB	H	OH	H[a]	H	=O						−5.613
s6	SC	=O	OH	$COCH_2OH$	H			H	H	H	H	−7.881
s7	SC	=O	OH	$COCH_2OH$	OH			H	H	H	H	−7.881
s8	SC	=O	=O	$COCH_2OH$	OH				H	H	H	−6.892
s9	SE	OH	=O									−5.000
s10	SC	=O	H	$COCH_2OH$	H			H	H	H	H	−7.653
s11	SC	=O	H	$COCH_2OH$	OH			H	H	H	H	−7.881
s12	SB	=O		H[a]	H	OH	H					−5.919
s13	SD	OH	OH	H	H							−5.000
s14	SD	OH	OH	H	OH							−5.000
s15	SD	OH	=O		H							−5.000
s16	SB	H	OH	H[b]	H	=O						−5.255
s17	SE	OH	COMe	H								−5.255
s18	SE	OH	COMe	OH								−5.000
s19	SC	=O	H	COMe	H			H	H	H	H	−7.380
s20	SC	=O	H	COMe	OH			H	H	H	H	−7.740
s21	SC	=O	H	OH	H			H	H	H	H	−6.724
s22	SF	=O	OH	$COCH_2OH$	OH							−7.512
s23	SC	=O	OH	$COCH_2OCOMe$	OH			H	H	H	H	−7.553
s24	SC	=O	=O	COMe	H				H	H	H	−6.779
s25	SC	=O	H	$COCH_2OH$	H			OH	H	H	H	−7.200
s26	SC[c]	=O	H	OH	H			H	H	H	H	−6.144
s27	SC	=O	H	COMe	OH			H	OH	H	H	−6.247
s28	SC	=O	H	COMe	H			H	Me	H	H	−7.120
s29	SC[c]	=O	H	COMe	H			H	H	H	H	−6.817
s30	SC	=O	OH	$COCH_2OH$	OH			H	H	Me	H	−7.688
s31	SC	=O	OH	$COCH_2OH$	OH			H	H	Me	F	−5.797

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

- and for molecules (**h1−h107**) all non-hydrogen atoms of pyrimidine ring.

We used Match3D program[25] for performing this operation.

4D-QSAR Calculation. We used Hopfinger's spatial grid system[2] for coding molecules. The molecules after AM1 optimization were used as initial structures in the molecular dynamic simulation (MDs). Each 3D structure is the starting point in generating conformational ensemble profile (CEP). Molecular dynamics was performed using the Sybyl software[26] with standard Tripos force field. 2500 conformations were sampled for each analogue. Partial atomic charges were calculated using the semiempirical AM1 Hamiltonian (HYPERCHEM package[27]). The alignment of the molecules was the next step of the 4D-QSAR analysis. We aligned the molecules according to the previous rules of the CoMFA study.[1] Individual conformers are placed in the grid cell space surrounding the aligned compounds. We applied cubic grid lattice of 20 Å on each side with grid cell resolution of 1, 2, or 0.5 Å, respectively. Different types of grid cell occupancy descriptors (GCODs) were considered and calculated for the indicated atoms referred to as interaction pharmacophore elements (IPE).[2] Apart from, the GCODs used by Hopfinger,[2] we applied in our current work the absolute charge occupancy (A_q) for the chosen IPE atoms of compound c defined as

$$A_q(c,i,j,k,N) = \sum_{t=0}^{T} O_t(c,i,j,k) \times q/m \qquad (1)$$

where m means the number of the atoms of compounds, c present in the cell (i,j,k) at time t, q means the sum of partial atoms of charges present in some cell at time t, and T is the length of the time in MDs. N is the number of sampling

MDs steps. The joint (J_q) and self-charge occupancy (S_q) with the most active reference compound R defined after following equations:

$$J_q(c,i,j,k,N) = \sum_{t=0}^{T} O_t(c,i,j,k) \cap O_t(R_q,i,j,k) \times q/m \quad (2)$$

$$S_q(c,R,i,j,k,N) = \sum_{t=0}^{T}\{O_t(c,i,j,k) - [\sum_{t=0}^{T} O_t(c,i,j,k) \cap O_t(R,i,j,k)]\} \times q/m \quad (3)$$

We used the MATLAB[28] environment to program the calculation of the above-mentioned descriptors. The Partial Least Squares (PLS) method with variable elimination was used to estimate the relationship between independent variables (GCODs) and corticosteroid binding globulin (CBG) affinity.

Calculation of the Molecular Surface (s-COMSA) Descriptors based on Virtual Cubic Grid. For the calculation of shape descriptors we applied a formalism similar to Hopfinger's 4D-QSAR grid coding system using the absolute type descriptors, as given by the above-mentioned equations. However, unlike in 4D-QSAR our method compares single conformers. Thus, each 3D molecular representation is placed in its own virtual cubic grid, and the molecular surface is calculated, respectively. The electrostatic potential is calculated for the points randomly sampled on the molecular surface and a mean value of the electrostatic potential corresponding to the respective points found in each grid cell is used to describe this cell. Grid cells are unfolded into vectors and vectors describing all molecules of the series are aligned into a matrix. Grid cells that are empty for all molecules in the series analyzed are eliminated, and the resulting matrix was used for further calculations using the PLS method.

PLS Analysis. Vectors obtained were processed by the PLS analysis with a leave-one-out cross-validation procedure. The PLS procedures were programmed within the MATLAB environment (MATLAB).[28]

A PLS model was constructed for the centered data, and its complexity was estimated on the basis of the leave-one-out cross-validation procedure (CV). In the leave-one-out CV one repeats the calibration m times, each time treating the ith left-out object as the prediction object. 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

$$\mathrm{RMSECV}_j = \sqrt{\frac{\sum_i(\mathrm{obs}_i - \mathrm{pred}_{i,j})^2}{m}} \quad (4)$$

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. The model with k factors, for which RMSECV reaches a minimum, is considered as an optimal one.

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

$$q^2 = 1 - \frac{\sum(\mathrm{obs}_i - \mathrm{pred}_i)^2}{\sum(\mathrm{obs}_i - \mathrm{mean(obs)})^2} \quad (5)$$

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

$$s = \sqrt{\frac{\sum(\mathrm{obs}_i - \mathrm{pred}_i)^2}{m-k-1}} \quad (6)$$

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

Before the PLS analysis was performed the descriptors were centered, and this operation was repeated for each cross-validation run.

The quality of external predictions was measured by the SDEP parameter

$$\mathrm{SDEP} = \sqrt{\frac{\sum_i(\mathrm{pred}_i - \mathrm{obs}_i)^2}{n}} \quad (7)$$

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

Data Elimination. To find these parts on the molecular surface that contribute to activity we used the modified procedure of the PLS with Uninformative Variable Elimination (UVE-PLS), namely the Iterative Variable Elimination PLS (IVE-PLS) procedure.[11] The UVE algorithm was originally developed by Centner et al.[29] as a possible improvement of PLS models. The purpose of the method is to reduce the number of the variables included in the final PLS model. The UVE algorithm is based on the analysis of the regression coefficients calculated by the PLS method. The PLS method allows for presenting the relation between the Y answer and the X predictors in a form of

$$Y = Xb + e$$

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

Thus, the UVE algorithm analyzes the reliability of the mean(b)/s(b) ratio (where s(b) means standard deviation of b). Then only the variables of the "relative" high mean(b)/$s(b)$ ratio are included into the final PLS model. To estimate the cutoff level artificial random number noise is created (the level of the noise is 10^{-10} of the original variable order) and put (as additional columns) into the matrix of the original variables. The PLS analysis of such a matrix is performed and the mean(b)/s(b) parameter is analyzed for each column. The highest absolute value, abs(mean(b)/s(b)) for the noisy column, determines the cutoff level for the original variables. In the current publication we used both typical UVE-PLS[29] and its modified iterative version IVE-PLS procedure (for the detailed description see ref 11).

RESULTS AND DISCUSSION

The description of the molecular shape by spatial sectors was originally suggested by Purcell and Testa[30] and then

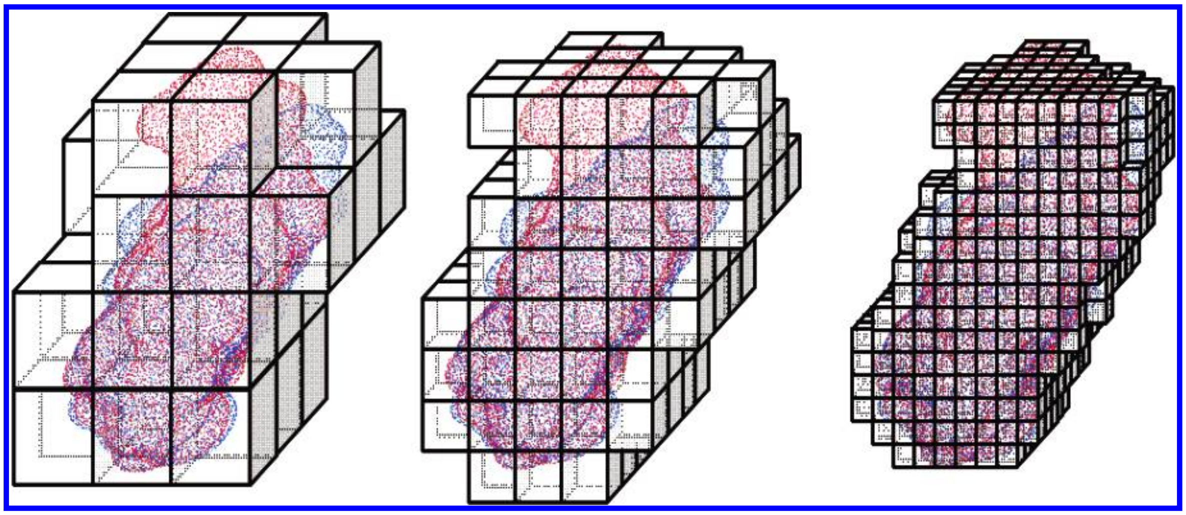


Figure 1. The molecular surface formed by a polycube grid of the different resolution.

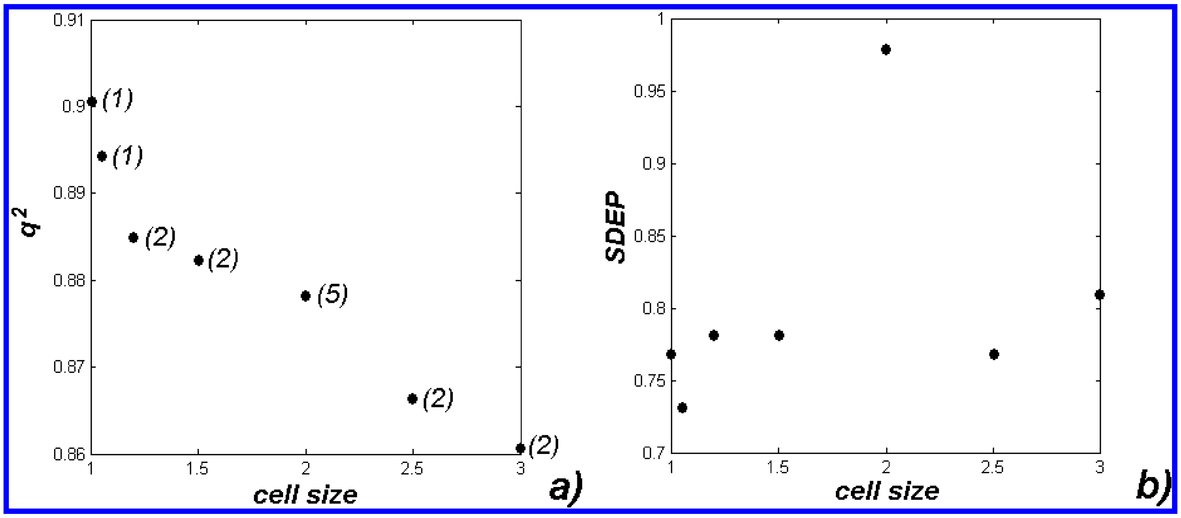


Figure 2. The dependence of the q^2 (a) and SDEP (b) performances on the grid mesh. The numbers shown indicates the complexity (an optimal number of components) in the PLS models, respectively.

improved by Motoc.[31] In this method a molecule is separated into partitions of spatial regions either filled or unfilled by atoms or groups of atoms of certain volumes. A similar method was used by Hopfinger to develop 4D-QSAR formalism. An interesting feature of the sector models is their fuzziness.[32] In Figure 1 we show a single-molecule-defined molecular surface that is subsequently replaced for the cubic-like molecular boundary. In such a representation molecular configuration is represented not by entities with sphere-like boundaries but by a polycube formed of a set of cubic domains (cells). Now a shape of the molecule depends on the grid resolution. This also means that two molecules can be compared with the different tolerance, which resembles fuzziness of the SOM neurons of the adjustable tolerance.[32]

Steroids that are complexed by the corticosteroid binding globulin (CBG) are used as a benchmark series in many publications aimed at 3D-QSAR modeling. A number of errors in the steroid structures that can be found in early publications have been corrected in recent years. For a review see ref 33. Similarly to previous studies, we validated the performance of the new method by the leave-one-out cross-validation (LOO CV). Moreover, the steroid series are split into two subseries as previously. In the current study we used the sampling scheme that is usually reported in the literature, i.e., training set: **s1**−**s21**; test set: **s22**−**s31**. Within the grid mesh of 1−1.5 Å the performance of the s-CoMSA only slightly depends on the resolution, as shown in Figure 2, while the LOO CV q^2 takes a value of ca. 0.90−0.88 (for compounds **s1**−**s21**; SDEP = 0.78−0.73 for compounds **s22**−**s31**). This compares advantageously to the CoMFA performances (q^2 = 0.73 for compounds **s1**−**s21**,[33] SDEP − compounds **s22**−**s31** − 0.837[34]). On the other hand, this resembles the performance values obtained in the neural version of the SOM-CoMSA (q^2 = 0.88 for all compounds; SDEP for compounds **s22**−**s31** − 0.69)[8] or this reported for Quasar calculations (q^2 = 0.90 − for compounds **s1**−**s21**).[35]

Although the elimination of variables is always risky, this can both improve the performance of the PLS model and indicate such areas of the molecules that particularly contribute to the activity. Figure 3a,b shows the q^2 (for the training set molecules **s1**−**s21**) and SDEP (for the test set molecules **s22**−**s31**) performances during the IVE-PLS procedure as a function of the number of variables eliminated. The q^2 and SDEP performances depend not only on the number of variables eliminated but also on the complexity, i.e., a number of the PLS components included in the respective models. This effect can be observed especially for the SDEP values, for example, in Figure 2 illustrating the performances for the different models resulted from the different grid resolution. It is evident that a model of the

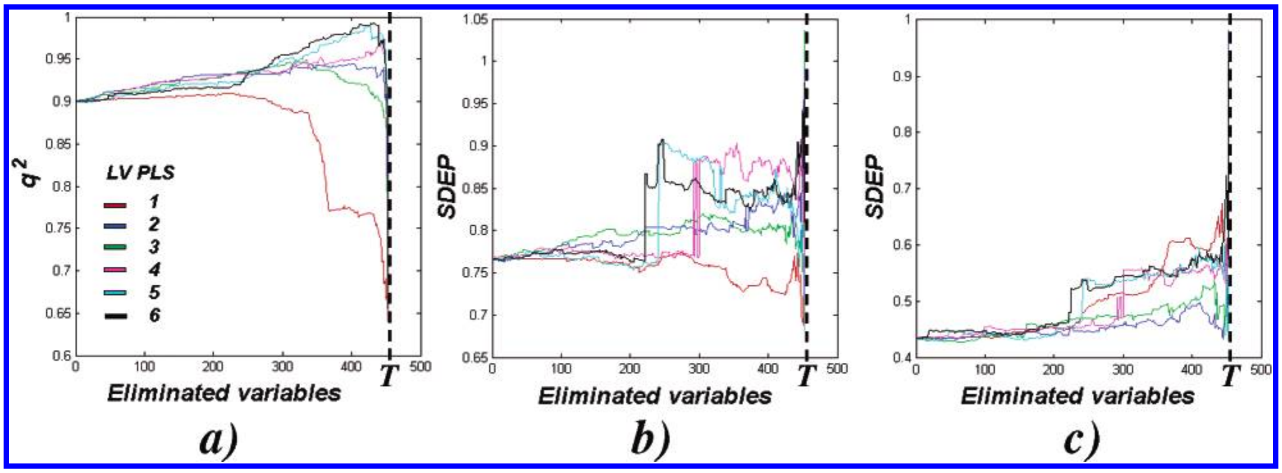


Figure 3. The dependence of the q^2 for the training set **s1−s21** (a) and SDEP performances for the test set **s22−s31** (b) or **s22−s30** (c) during IVE-PLS variable elimination upon the number of variable eliminated. Different colors illustrate the models of the different complexities (details in text), and T indicates a total number of original variables (nonempty grid cells).

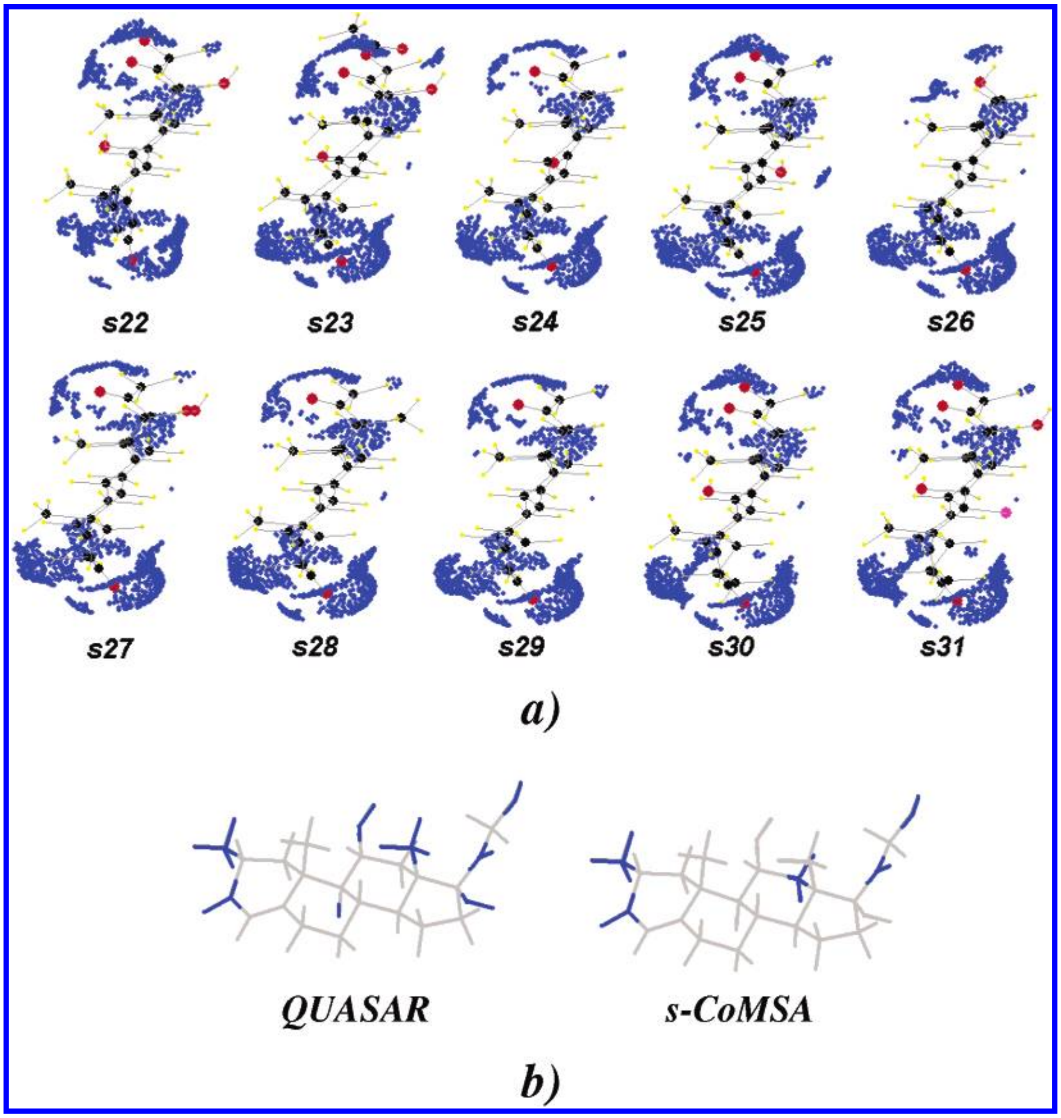


Figure 4. The surface areas (illustrated by the points sampled) of the highest contribution to the CBG activity, as indicated by the IVE-PLS performed for the training set **s1−s21**, illustrated for the test set molecules **s22−s31** (a), and compared to the respective Quasar results (b).[35]

highest complexity (5) provides also the lowest SDEP predictivity. Therefore, the IVE-PLS procedure was performed in such a way that a number of the PLS latent components were always optimized. This number, however, was truncated not to exceed a value of 1−6, respectively. This means that model complexity cannot exceed 1 in the plot shown in red, while e.g. the plot shown in black (6 components) can include the models of the different

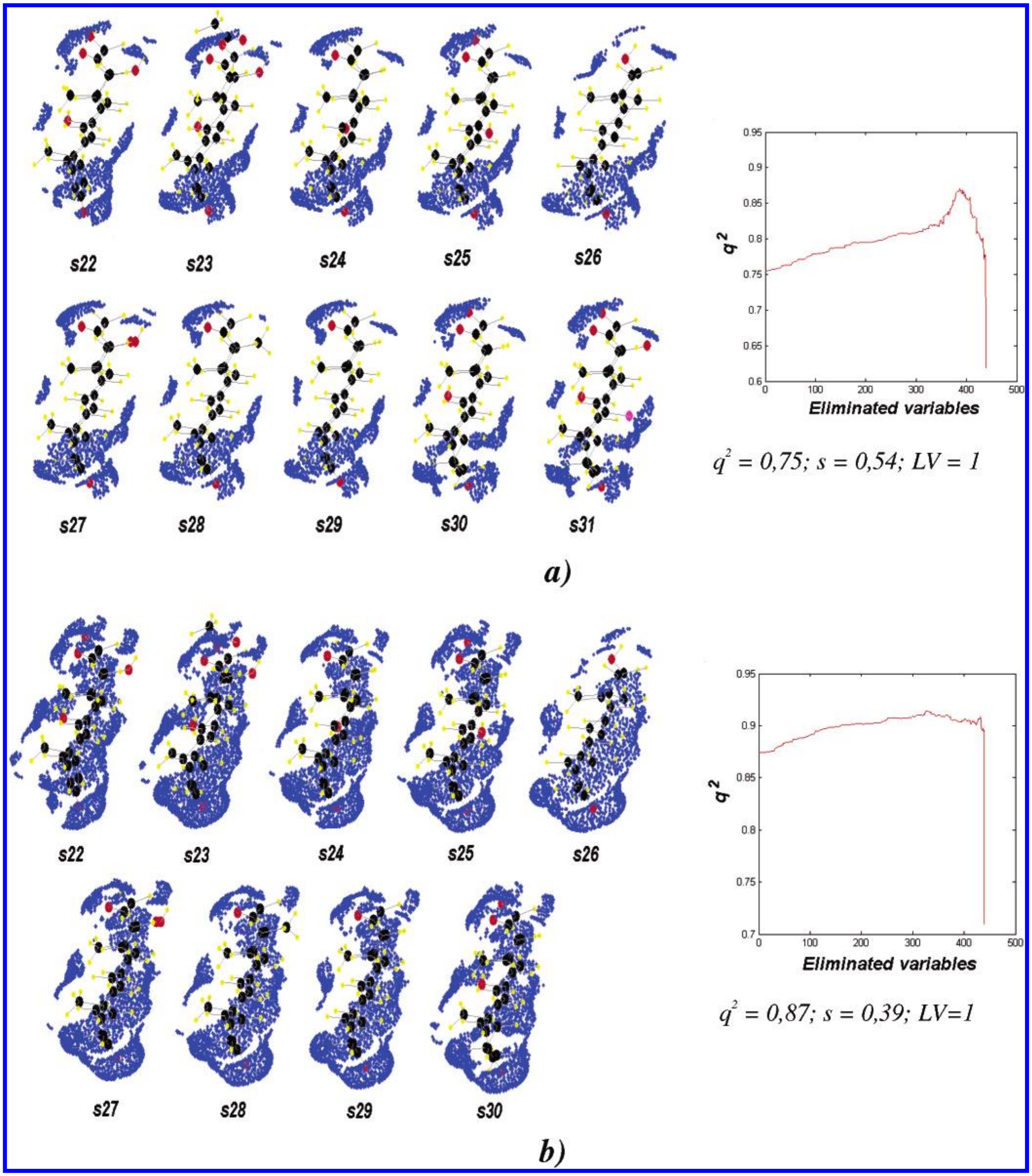


Figure 5. The surface areas (illustrated by the points sampled) of the highest contribution to the CBG activity, as indicated by the IVE-PLS performed for the whole series **s1**−**s31** (a) or molecules **s1**−**s30** (b). The plots shown report the q^2 IVE-PLS performances during variable elimination, respectively. The q^2 and s performances refer to the initial PLS model (before IVE-PLS) and LV indicates a final model complexity. The maximal complexity of the model was truncated to 3.

complexities (each time the optimal value is sought after) from 1 to 6. The complexity of the initial PLS models (elimination was always started from the same model) before variable elimination amounts to 3. Figure 3c illustrates a fact that the elimination of the molecule **s31** from the test set significantly improves the SDEP values. Similarly, to the number of the other methods[33] s-CoMSA evidently misclassifies this compound.

Although for the models of the highest possible complexity (6), variable elimination enables to achieve an impressive value of the q^2 higher than 0.99, the most stable and lowest SDEP values are given for a PLS model with a single component (SDEP = 0.69). This indicates that the models of lower complexity, even if this complexity is lower than the optimal one, are more flexible in handling external objects. However, the evaluation of variable elimination

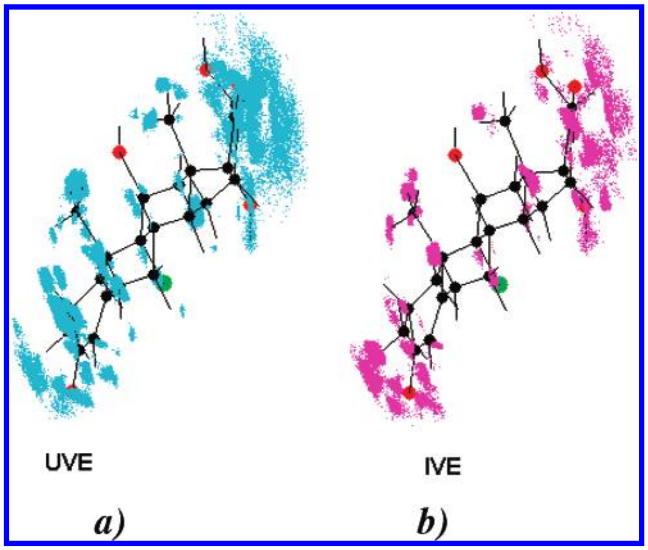


Figure 6. Atomic coordinates of the highest contribution to the CBG activity as indicated by 4D-QSAR−PLS simulations with UVE (a) and IVE (b) data elimination, respectively (details in text). The calculation were conducted using the absolute descriptors occupancy (A_0).

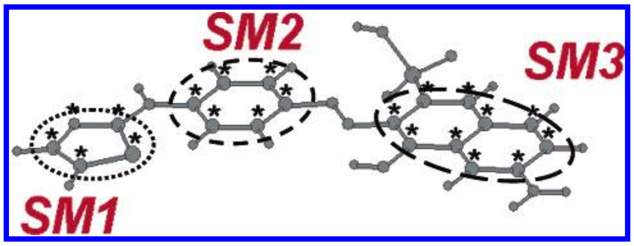


Figure 7. Different superimposition modes SM1−SM3 tested for the dye molecules. The circles indicate the molecule areas covered in the individual superimposition, and the asterisks show individual atoms specified for covering.

without **s31** (Figure 3c) indicates that a 2- or 3-component model gives better predictivity (lower SDEP values). Thus, we used such a model for pointing the possible pharmacophoric sites on the molecular surface of the test set molecules. In Figure 4a we compare the structures of the steroids **s22**−**s31** showing the points sampled originally on the molecular surface and located within the grid volumes that survived after IVE-PLS variable elimination in the training set **s1**−**s21**, as reported in Figure 3. Figure 4b indicates clear similarities between these regions and the ones visualized by the Quasar method.[35]

Alternatively, Figure 5 illustrates similar variable elimination procedure performed, however, for the whole compound series **s1**−**s31** (Figure 5a) or **s1**−**s30** (Figure 5b). The plots shown report the q^2 values during variable elimination. Unlike previously, now the procedure cannot be additionally monitored by the calculation of the SDEP, because all compounds are used during the modeling step. Only compounds **s22**−**s31** or **s22**−**s30** are shown in Figure 5 in order for the easier comparison with Figure 4. The comparison of Figures 4 and 5 illustrates that a change of the compound series used in the IVE-PLS affects the results obtained, i.e., different surface areas are selected, which emphasizes the stochastic nature of this procedure. Moreover, it can be observed that the molecular basis of **s1**−**s30** approves the largest number of variables into a final model. This effect results from a fact that in this case the IVE-PLS starts from a relatively high q^2 level. Moreover, the q^2 increase is much less steep, and a large plateau can be observed near the maximal q^2 value. Thus, in this particular case the procedure is much more stable, and even if we go beyond the maximal q^2 value any sharp decline in the model q^2 cannot be observed. However, the maximal q^2 complies with quite a large number of original variables as shown in Figure 5b.

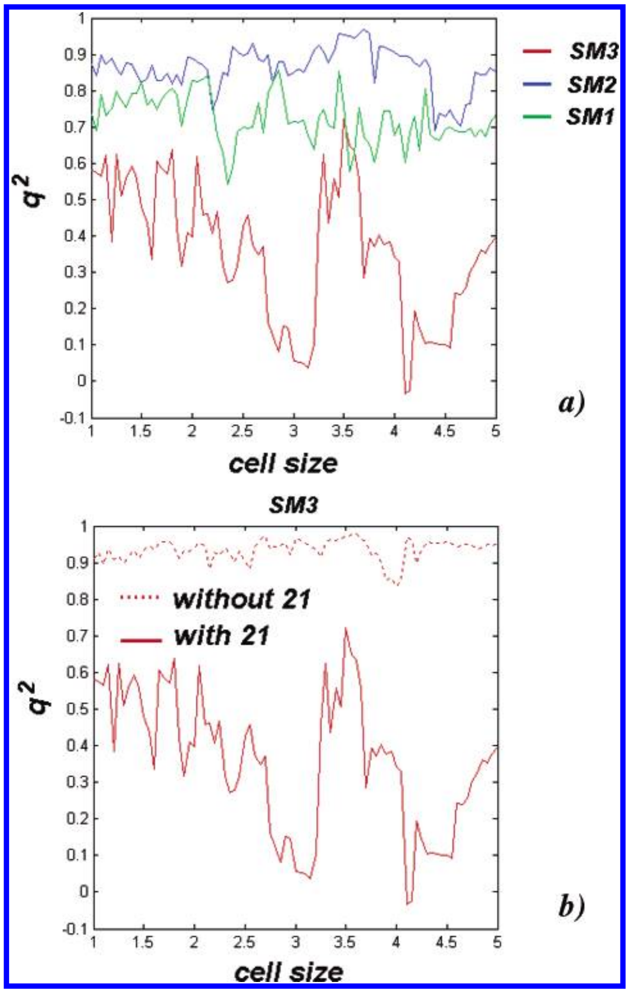


Figure 8. The dependence of the LOO−CV q^2 performance for the s-CoMSA models of the tinctorial properties (**d1**−**d21**) upon the grid resolution for three different superimpostion modes SM1−SM3 tested (a). The exclusion of the molecule **d21** can significantly improve the q^2 performance (b).

A question on the relative performance of the s-CoMSA and 4D-QSAR methods may arise. Steroid series seems to be a proper object for the comparison. Therefore, we performed 4D-QSAR simulations with UVE (version modified according to ref 11) and IVE-PLS. The best performance ($q^2 = 0.94$, $s = 0.38$, SDEP = 0.77, including **s31**) compares well to that of the s-CoMSA and are significantly better than this reported for 4D-QSAR without data elimination ($q^2 = 0.84$, $s = 0.50$, SDEP = 0.83, including **s31**).[36] Similarly, the molecular areas (atomic coordinates) indicated as important for the activity in 4D-QSAR, as illustrated in Figure 6, resemble those suggested by s-CoMSA (surfaces defined by respective atoms). This fact seems to show that for the relatively rigid steroid structures the investigations of the conformational space by 4D-QSAR calculation does not improve the performance, and these calculations are extremely time-consuming processes.

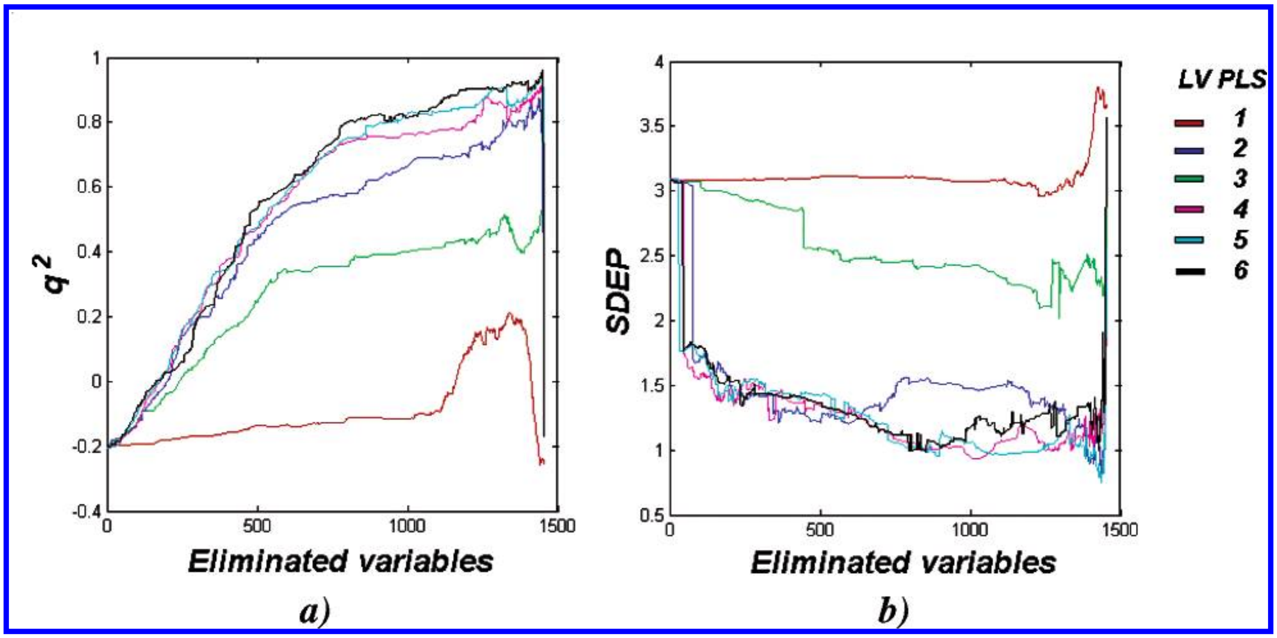


Figure 9. The dependence of the q^2 for the training set **d1**:2:**d21** (a) and SDEP performances for the test set **d2**:2:**d20** (b) during IVE-PLS variable elimination upon the number of variable eliminated. Different colors illustrate the models of the different complexities (details in text).

The interaction between a dye molecule and cellulose is a complicated phenomenon, which can be described by the Langmuir isotherm.[37,38] An isotherm does not, however, provide a molecular description of the process. Moreover, we cannot use such an approach for the optimization of the molecular structure of dye. The influence of the electrostatic, van der Waals, or hydrogen bonding as well as hydrophobic forces on the dyeing process has been investigated.[37] On the other hand, it has been speculated that specific binding sites exist on the crystalline region of the supramolecular cellulose structure that forms holes and cavities capable of incorporating a dye molecule.[39] Does this mean that a similarity between the drug−receptor and dye−fiber interactions makes possible to extend a pharmacophore concept and develop an idea of *tinctophore* in dye chemistry to predict tinctorial properties of dyes by the use of QSAR or related methods? Although it is not clear whether we can treat it similarly to the contacts taking place during targeting a receptor by a drug molecule, several QSAR studies have been published recently[13,40−47] that make use of this concept in investigations of cellulose dyeing. Both 2D- and 3D-QSAR modeling have been applied, including the Hansch, MTD, and Comparative Molecular Field Analysis (CoMFA) methods that appeared to provide quite satisfactory models for different compound series.[23,44−46] In particular, the results of the CoMFA and CoMSA methods indicated that the electrostatic field predominates. On the other hand, dye−cellulose interactions seemed to be less specific than drug−receptor interactions.[12,13,38]

Below the results of the application of s-CoMSA for the analysis of the heterocyclic dye series are reported. Figure 7 indicates three different superposition modes SM1−SM3 tested. In Figure 8 we show the LOO CV q^2 value as a function of the grid resolution. Independent of the grid size, superposition SM3 gives lower q^2 performances than the SM1 and SM2 ones. On the other hand, it can be observed that the exclusion of a single compound **d21** can improve the SM3 q^2 value. This gives similar values to those of the SM1 or SM2 (Figure 8b). In Figure 9 we reported the results of the data reduction experiment that is performed under the control of both the values of q^2 − for the training set (compounds **d1**, **d3**, **d5**, **...**, **d21**) and SDEP − for the test set (compounds **d2**, **d4**, **d6**, **...**, **d20**). The experiment started from a very low initial q^2 value (−0.2). However, data reduction with the IVE-PLS procedure can improve this value up to more then 0.9. Unlike for the benchmark steroid series, now the models of higher complexities provide better models both in respect to the q^2 and SDEP values. Figure 10 indicates the molecular areas indicated by the IVE-PLS variable elimination as important for the tinctorial properties of the dye series. The comparison of these results to those obtained with the SOM-CoMSA version indicates the neural method is better at giving a q^2 value of 0.98.[13] Moreover, the SOM-CoMSA indicates for the SM2 region (best models are obtained for the templates based on central part of the

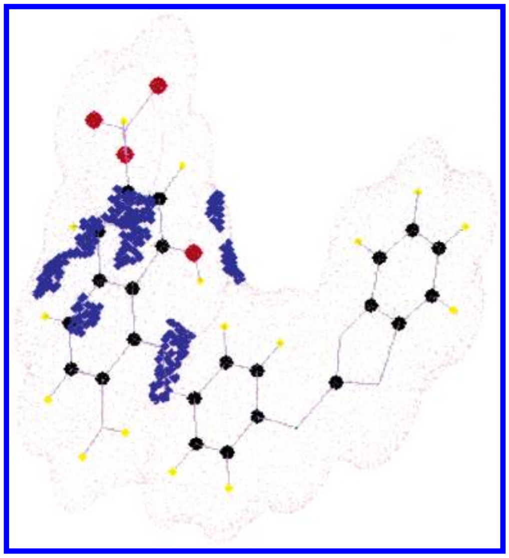


Figure 10. The surface areas (illustrated by the points sampled) of the highest contribution to the tinctorial properties, as indicated by the IVE-PLS procedure.

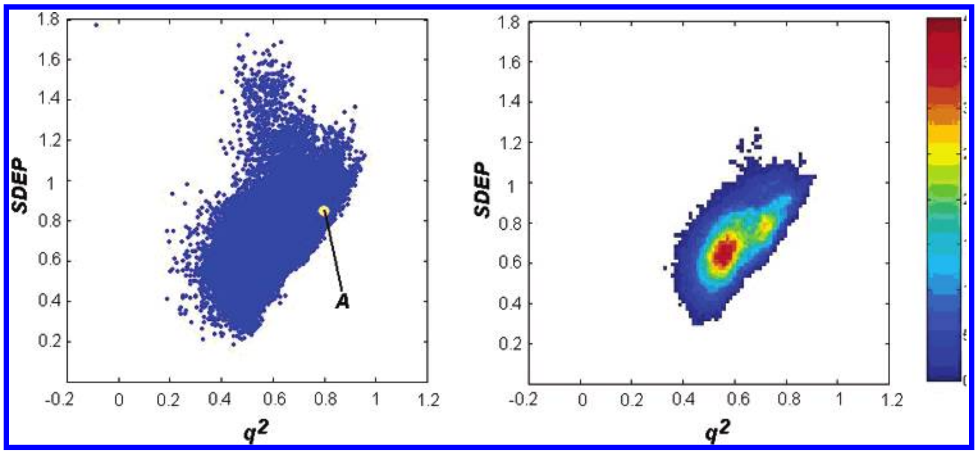


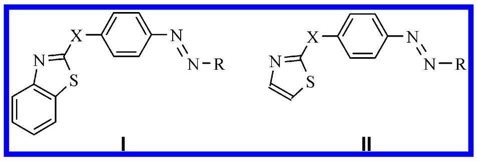
Figure 11. The influence of the sampling of the compounds into the training and test set for the CoMFA calculations of the steroid CoMFA series. The relationships between the LOO−CV q^2 performance estimated for all possible 21-molecule-containing training sets sampled randomly from **s1**−**s31** and SDEP calculated for the 10-molecule-conteining test sets, respectively. The results are shown as a standard plot (a) or a density colormap (b). A indicates the results for the literature sampling—training: **s1**−**s21**; test: **s22**−**s31**.[48]

molecule),[13] i.e., similar areas to those that survive IVE-PLS variable elimination in the s-CoMSA method.

We bear in mind a fact that very low q^2 performance was obtained if we sampled only a half of the dye molecules (**d1**−**d21**) into the training set. We used here a random sampling with the molecules **d1**, **d3**, **d5**, **...**, **d21** (training) and **d2**, **d4**, **d6**, **...**, **d20** (test). Thus an attempt to model 3D-QSAR for such a training set gives a q^2 value of ca. −0.2, while for the whole series a value of 0.9 was obtained. Indeed, we have carefully analyzed the influence of the sampling of the compounds into the training and test series in 3D-QSAR modeling. It was discovered that q^2 significantly depends on such sampling. Although the detailed discussion of this effect is beyond the scope of this paper we would like to indicate here an example of the CoMFA calculation for the steroid series **s1**−**s31**. In the experiment performed we sampled all 31 molecules into the training series including 21 molecules and test series (10 molecules).[48] The total amount of different sampling in such an experiment amounts to 44 352 165 (31!/21!*10!). As it is technically impossible to verify all these models we tested each 1 of 500 possible, which makes 8 870 433 different combinations (Figure 11). In this experiment the q^2 performance ranges from −0.16 to 0.95. Similar SOM-CoMSA calculations provide significantly better performances of $q^2 = 0.42-0.98$. It is worth mentioning that the effect described does not indicate the instability of our 3D-QSAR calculations; we obtained a standard q^2 value of 0.79 and SDEP = 0.83 for the CoMFA model with literature sampling of training/test **s1**−**s21**/**s22**−**s31** but emphasizes the stochastic nature of the 3D-QSAR calculations performed for a few molecular objects using highly multidimensional molecular descriptor data.

To further test the s-CoMSA method we performed a 3D-QSAR study of a series of HEPT derivatives, inhibitors of the reverse transcriptase enzyme of the HIV virus. 3D-QSAR investigations for a large group of 107 HEPT analogues or different subgroups have been recently reported in many papers.[24,49−55] The reported q^2 performances ranged from $q^2 = 0.92$, $s = 0.36$ for 12 compounds[49] to $q^2 = 0.78$;[24] $q^2 = 0.82$ (in both studies the s value has not been given)[51] or $q^2 = 0.86$.[52] The CoMFA analysis for a series of 101 molecules gives for a training set a performance of $q^2 = 0.86$, $s = 0.53$ (80 compounds, one outlier eliminated).[50] The s-CoMSA model obtained in our study significantly outperformed these values. Thus, the performance of the optimal s-CoMSA IVE-PLS LOO−CV model obtained for the grid mesh of 1.5 Å amounts to $q^2 = 0.86$, s = 0.58 for the whole series of 107 compounds **h1**−**h107**. To validate the model predictivity we divided the series into the training (**h1**−**h80**) and test set (**h81**−**h107**), i.e., using similar rules as those reported

Table 2. Heterocyclic Dye Structures and the Affinity Data to Cellulose[23]



no.		X	R	$-\Delta\mu^0$ (kJ/mol)
d1	I	-NH-	A	6.78
d2	I	-NH-	B	9.20
d3	I	-NH-	D	12.60
d4	I	-NH-	E	15.30
d5	I	O	A	3.26
d6	I	O	B	5.27
d7	I	O	D	7.61
d8	I	O	E	10.30
d9	I	O	G	10.20
d10	I	S	A	1.26
d11	I	S	B	3.56
d12	I	S	D	5.02
d13	I	S	E	8.45
d14	I	S	G	8.12
d15	II	-NH-	E	15.33
d16	II	-NH-	D	12.60
d17	II	-NH-	B	9.24
d18	II	-NH-	A	6.80
d19	I	S	C	5.86
d20	I	S	F	10.33
d21	I	S	E (acid coupled)	9.75

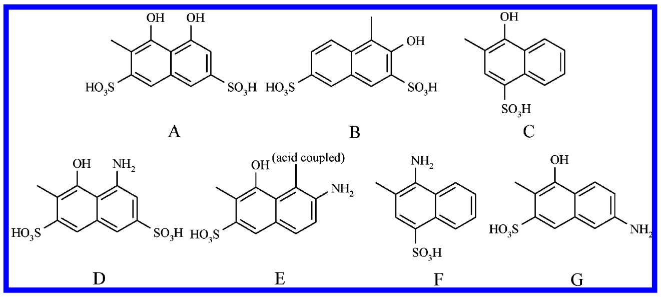
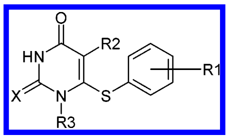


Table 3. Chemical Structures with the Observed[24] and Calculated Values of Anti-HIV Activity for the HEPT Derivatives



no.	R1	R2	R3	X	obs	pred^{a}	pred^{b}	pred^{c}	pred^{c}	pred^{d}
h1	2-Me	Me	$CH_2OCH_2CH_2OH$	O	4.15	4.75	4.47^{e}	4.47	5.15	3.84
h2	2-NO_2	Me	$CH_2OCH_2CH_2OH$	O	3.85	5.03	3.74	3.76	3.83	4.10
h3	2-OMe	Me	$CH_2OCH_2CH_2OH$	O	4.72	4.11	3.40	4.65	5.03	4.94
h4	3-Me	Me	$CH_2OCH_2CH_2OH$	O	5.59	5.40	4.70^{e}	4.83	4.97	5.44
h5	3-Et	Me	$CH_2OCH_2CH_2OH$	O	5.57	5.33	4.73^{e}	5.11	4.94	5.65
h6	3-t-Bu	Me	$CH_2OCH_2CH_2OH$	O	4.92	5.30	5.03^{e}	5.36	4.97	4.93
h7	3-CF_3	Me	$CH_2OCH_2CH_2OH$	O	4.35	5.21	4.71	4.09	4.39	4.62
h8	3-F	Me	$CH_2OCH_2CH_2OH$	O	5.48	4.67	5.11	4.54	4.80	5.29
h9	3-Cl	Me	$CH_2OCH_2CH_2OH$	O	4.89	5.08	5.00^{e}	4.86	4.98	5.26
h10	3-Br	Me	$CH_2OCH_2CH_2OH$	O	5.24	5.09	4.81^{e}	5.08	4.99	5.24
h11	3-I	Me	$CH_2OCH_2CH_2OH$	O	5.00	5.29	4.80	5.64	5.23	5.26
h12	3-NO_2	Me	$CH_2OCH_2CH_2OH$	O	4.47	4.79	5.37	4.62	4.33	4.57
h13	3-OH	Me	$CH_2OCH_2CH_2OH$	O	4.09	5.34	5.28	4.72	4.71	4.93
h14	3-OMe	Me	$CH_2OCH_2CH_2OH$	O	4.66	5.02	5.55	5.50	5.07	5.23
h15	3,5-Me_2	Me	$CH_2OCH_2CH_2OH$	O	6.59	5.99	5.90^{e}	6.27	6.44	6.42
h16	3,5-Cl_2	Me	$CH_2OCH_2CH_2OH$	O	5.89	5.88	6.07^{e}	6.25	6.21	6.28
h17	3,5-Me_2	Me	$CH_2OCH_2CH_2OH$	S	6.66	6.37	6.21	6.60	6.33	6.50
h18	3-COOMe	Me	$CH_2OCH_2CH_2OH$	O	5.10	4.83	3.77	5.37	4.83	4.63
h19	3-COMe	Me	$CH_2OCH_2CH_2OH$	O	5.14	4.70	4.69	5.55	5.12	4.36
h20	3-CN	Me	$CH_2OCH_2CH_2OH$	O	5.00	4.77	5.59	5.27	5.08	4.72
h21	H	$CH_2CH{=}CH_2$	$CH_2OCH_2CH_2OH$	O	5.60	5.61	5.47	5.60	5.18	5.68
h22	H	Et	$CH_2OCH_2CH_2OH$	S	6.96	6.24	6.27	6.35	6.92	6.74
h23	H	Pr	$CH_2OCH_2CH_2OH$	S	5.00	6.71	6.4^{e}	6.79	5.88	6.01
h24	H	i-Pr	$CH_2OCH_2CH_2OH$	S	7.23	7.58	7.88	6.75	6.15	7.32
h25	3,5-Me_2	Et	$CH_2OCH_2CH_2OH$	S	8.11	7.32	6.97^{e}	7.48	7.69	7.76
h26	3,5-Me_2	i-Pr	$CH_2OCH_2CH_2OH$	S	8.30	8.51	8.28	8.45	8.26	8.30
h2	3,5-Cl_2	Et	$CH_2OCH_2CH_2OH$	S	7.37	7.26	6.54	8.12	7.84	7.64
h28	H	Et	$CH_2OCH_2CH_2OH$	O	6.92	6.21	5.50^{e}	5.97	6.85	6.66
h29	H	Pr	$CH_2OCH_2CH_2OH$	O	5.47	6.43	5.52^{e}	6.07	5.43	5.93
h30	H	i-Pr	$CH_2OCH_2CH_2OH$	O	7.20	7.11	6.27	6.79	6.83	7.24
h31	3,5-Me_2	Et	$CH_2OCH_2CH_2OH$	O	7.89	7.52	6.41	7.24	7.79	7.68
h32	3,5-Me_2	i-Pr	$CH_2OCH_2CH_2OH$	O	8.57	8.30	7.45	8.20	8.55	8.22
h33	3,5-Cl_2	Et	$CH_2OCH_2CH_2OH$	O	7.85	7.25	6.32	7.51	7.84	7.56
h34	4-Me	Me	$CH_2OCH_2CH_2OH$	O	3.66	4.90	4.52^{e}	5.36	3.71	5.39
h35	H	Me	$CH_2OCH_2CH_2OH$	O	5.15	5.10	4.62^{e}	4.90	5.04	5.30
h36	H	Me	$CH_2OCH_2CH_2OH$	S	6.01	5.42	5.92	5.26	5.27	5.38
h37	H	I	$CH_2OCH_2CH_2OH$	O	5.44	5.64	5.11^{e}	5.68	5.41	5.66
h38	H	$CH{=}CH_2$	$CH_2OCH_2CH_2OH$	O	5.69	6.20	5.84	4.80	5.32	6.53
h39	H	$CH{=}CHPh$	$CH_2OCH_2CH_2OH$	O	5.22	5.35	6.18	5.40	5.22	4.75
h40	H	CH_2Ph	$CH_2OCH_2CH_2OH$	O	4.37	5.43	5.68	5.28	5.08	4.81
h41	H	$CH{=}CPh_2$	$CH_2OCH_2CH_2OH$	O	6.07	5.11	4.22	4.92	6.10	6.06
h42	H	Me	$CH_2OCH_2CH_2OMe$	O	5.06	4.64	4.71	5.26	5.14	5.35
h43	H	Me	$CH_2OCH_2CH_2OAc$	O	5.17	4.98	4.71	5.24	5.05	4.62
h44	H	Me	$CH_2OCH_2CH_2OCOPh$	O	5.12	4.59	4.78	5.53	5.24	5.66
h45	H	Me	CH_2OCH_2Me	O	6.48	5.76	5.88	5.82	5.75	5.75
h46	H	Me	$CH_2OCH_2CH_2Cl$	O	5.82	5.99	6.02	5.87	5.84	5.73
h47	H	Me	$CH_2OCH_2CH_2N_3$	O	5.24	5.71	4.99	6.07	5.18	4.74
h48	H	Me	$CH_2OCH_2CH_2F$	O	5.96	4.94	6.27	5.49	5.39	5.23
h49	H	Me	$CH_2OCH_2CH_2Me$	O	5.48	5.72	6.08^{e}	5.60	5.26	5.67
h50	H	Me	CH_2OCH_2Ph	O	7.06	6.25	6.92	6.88	6.96	6.42
h51	H	Et	CH_2OCH_2Me	O	7.72	6.96	6.73^{e}	6.72	6.80	7.14
h52	H	Et	CH_2OCH_2Me	S	7.58	6.92	7.16	6.91	6.79	7.22
h53	3,5-Me_2	Et	CH_2OCH_2Me	O	8.24	8.35	8.03	8.12	8.24	8.17
h54	3,5-Me_2	Et	CH_2OCH_2Me	S	8.30	8.27	7.64	8.29	8.21	8.26
h55	H	Et	CH_2OCH_2Ph	O	8.23	7.78	7.73^{e}	6.63	7.41	7.64
h56	3,5-Me_2	Et	CH_2OCH_2Ph	O	8.55	9.74	9.21	8.54	8.36	8.51
h57	H	Et	CH_2OCH_2Ph	S	8.09	7.06	7.38^{e}	7.53	8.09	7.72
h58	3,5-Me_2	Et	CH_2OCH_2Ph	S	8.14	9.00	7.66	8.75	8.17	8.59
h59	H	i-Pr	CH_2OCH_2Me	O	7.99	7.48	7.48	7.67	8.13	7.72
h60	H	i-Pr	CH_2OCH_2Ph	O	8.51	8.44	8.49	8.53	8.19	8.16
h61	H	i-Pr	CH_2OCH_2Me	S	7.89	8.07	8.32	7.84	7.79	7.80
h62	H	i-Pr	CH_2OCH_2Ph	S	8.14	7.78	8.52	8.31	8.12	8.24
h63	H	Me	CH_2OMe	O	5.68	5.85	5.87^{e}	5.90	5.68	6.08
h64	H	Me	CH_2OBu	O	5.33	5.89	6.07^{e}	5.73	5.64	5.58
h65	H	Me	Et	O	5.66	6.17	5.06	6.29	5.51	6.64
h66	H	Me	Bu	O	5.92	5.42	4.84	6.41	5.61	5.98
h67	3,5-Cl_2	Et	CH_2OCH_2Me	S	7.89	8.30	7.18	8.00	7.81	8.13
h68	H	Et	$CH_2O-i\text{-}Pr$	S	6.66	7.06	7.19	6.79	6.64	6.87

Table 3 (Continued)

no.	R1	R2	R3	X	obs	pred[a]	pred[b]	pred[c]	pred[c]	pred[d]
h69	H	Et	CH_2O-c-Hex	S	5.79	5.85	7.00	6.63	5.95	6.06
h70	H	Et	CH_2OCH_z-c-Hex	S	6.45	7.13	7.53	6.57	7.06	6.01
h71	H	Et	$CH_2OCH_2C_6H_4$(4-Me)	S	7.11	7.56	7.45	6.97	7.63	7.57
h72	H	Et	$CH_2OCH_2C_6H_4$(4-Cl)	S	7.92	7.01	7.33	7.44	7.99	7.63
h73	H	Et	$CH_2OCH_2CH_2Ph$	S	7.04	6.75	7.50	6.61	6.82	7.60
h74	3,5-Cl_2	Et	CH_2OCH_2Me	O	8.13	8.41	8.22	8.03	8.21	8.05
h75	H	Et	CH_2O-i-Pr	O	6.47	6.46	6.01	6.79	6.71	6.78
h76	H	Et	CH_2O-c-Hex	O	5.40	6.09	6.46	5.80	5.18	5.98
h77	H	Et	CH_2OCH_z-c-Hex	O	6.35	7.32	6.93[e]	6.44	6.45	5.93
h78	H	Et	$CH_2OCH_zCH_2Ph$	O	7.02	7.66	6.88	7.10	7.16	7.52
h79	H	c-Pr	CH_2OCH_2Me	S	7.02	7.32	6.61	7.39	6.85	6.61
h80	H	c-Pr	CH_2OCH_2Me	O	7.00	6.95	6.28[e]	7.16	7.01	6.53
h81	H	Me	$CH_2OCH_2CH_2OC_5H_{11}$-n	O	4.46	5.05	4.84[e]	6.01	5.24	5.12
h82	2-Cl	Me	$CH_2OCH_2CH_2OH$	O	3.89	4.37	4.27	4.64	5.14	4.31
h83	3-CH_2OH	Me	$CH_2OCH_2CH_2OH$	O	3.53	6.62	5.26	4.39	4.83	4.60
h84	4-F	Me	$CH_2OCH_2CH_2OH$	O	3.60	5.00	3.80	4.50	3.58	5.10
h85	4-Cl	Me	$CH_2OCH_2CH_2OH$	O	3.60	5.33	4.00[e]	5.04	3.80	5.45
h86	4-NO_2	Me	$CH_2OCH_2CH_2OH$	O	3.72	6.29	3.91	4.43	4.22	5.04
h87	4-CN	Me	$CH_2OCH_2CH_2OH$	O	3.60	5.80	3.18	5.65	4.17	4.98
h88	4-OH	Me	$CH_2OCH_2CH_2OH$	O	3.56	5.08	4.20	4.95	3.57	4.74
h89	4-OMe	Me	$CH_2OCH_2CH_2OH$	O	3.60	5.83	4.15	4.98	3.61	5.28
h90	4-COMe	Me	$CH_2OCH_2CH_2OH$	O	3.96	7.04	3.54	4.91	3.68	5.00
h91	4-COOH	Me	$CH_2OCH_2CH_2OH$	O	3.45	6.19	4.31	4.69	4.56	4.67
h92	3-$CONH_2$	Me	$CH_2OCH_2CH_2OH$	O	3.51	5.09	5.33	4.99	4.68	4.14
h93	H	COOMe	$CH_2OCH_2CH_2OH$	O	5.18	8.25	4.76	4.22	5.01	5.28
h94	H	CONHPh	$CH_2OCH_2CH_2OH$	O	4.74	7.15	5.26	5.29	5.13	4.60
h95	H	SPh	$CH_2OCH_2CH_2OH$	O	4.68	7.27	6.10	5.01	5.57	4.97
h96	H	CCH	$CH_2OCH_2CH_2OH$	O	4.74	6.90	4.70[e]	5.95	5.23	6.30
h97	H	CCPh	$CH_2OCH_2CH_2OH$	O	5.47	6.67	4.38	5.86	5.26	4.71
h98	3-NH_2	Me	$CH_2OCH_2CH_2OH$	O	3.60	4.89	4.81	4.76	5.02	4.57
h99	H	$COCHMe_2$	$CH_2OCH_2CH_2OH$	O	4.92	7.80	5.78	5.78	4.58	5.27
h100	H	COPh	$CH_2OCH_2CH_2OH$	O	4.89	6.44	4.00	5.08	5.15	5.11
h101	H	CCMe	$CH_2OCH_2CH_2OH$	O	4.72	6.80	4.68	6.09	5.28	5.57
h102	H	F	$CH_2OCH_2CH_2OH$	O	4.00	6.04	4.61	5.25	5.11	5.00
h103	H	Cl	$CH_2OCH_2CH_2OH$	O	4.52	5.84	4.51[e]	5.46	5.16	5.36
h104	H	Br	$CH_2OCH_2CH_2OH$	O	4.70	5.40	4.66	5.49	5.16	5.47
h105	H	Me	$CH_2OCH_2CH_2OCH_2Ph$	O	4.70	4.61	4.68	5.61	5.63	5.93
h106	H	Me	H	O	3.60	5.80	6.18	6.06	5.52	5.63
h107	H	Me	Me	O	3.82	4.64	4.90	6.16	5.49	.75

[a] s-CoMSA. [b] s-CoMSA for the Kennard−Stone training/test set sampling. [c] Activity values according to the ref 24. [d] Activity values according to the ref 51. [e] Test set.

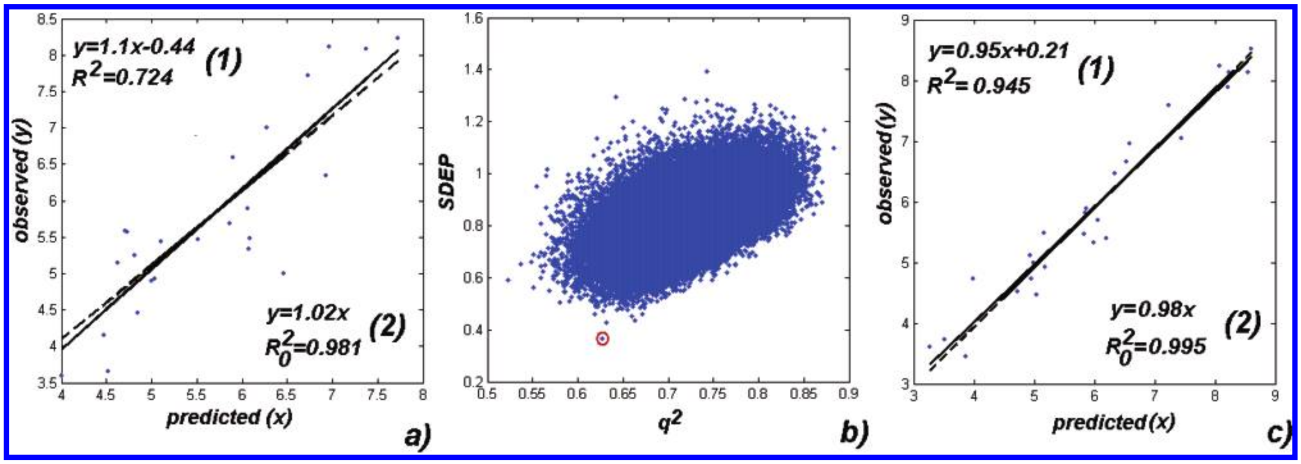


Figure 12. The GT model validation for the Kennard−Stone training/test set sampling (a), the dependence of the SDEP vs q^2 for 110 000 random training/test set samplings (b), and the GT model validation for the highly predictive model selected among these models (SDEP = 0.37) (c). We tried to keep the style of the presentation of the authors.[56] The regression between observed (Y) and predicted (X) activity values for the test set. The solid line shows the regressional equation given by (1). The dotted line illustrates the regression without the bias (2). The closer are these linear plots, the better is the model predictivity. Calculations after $\text{pred}_i^0 = k^*\text{pred}_i$, $k = \sum \text{obs}_i\text{pred}_i/\sum \text{pred}_i^2$, and $R_0^2 = 1 - \sum(\text{pred}_i - \text{pred}_i^0)^2/\sum(\text{pred}_i - \text{mean(pred)})^2$ where the upper index 0 relates to regression observed (Y) vs predicted (X), k is a slope of the regression through the origin (2), and R_0^2 is the correlation coefficient for the regression of observed (Y) vs predicted (X) without bias. $[(R^2 - R_0^2)/R^2] < 0.1$ and $0.85 \leq k \leq 1.15$ as recommended by Golbraikh and Tropsha.[56]

previously.[24,51] The predicted values are given in Table 3, respectively. This procedure results in the relatively high SDEP value of 2.01. Moreover, we validated this model performing the calculation of the Golbraikh−Tropsha (GT)

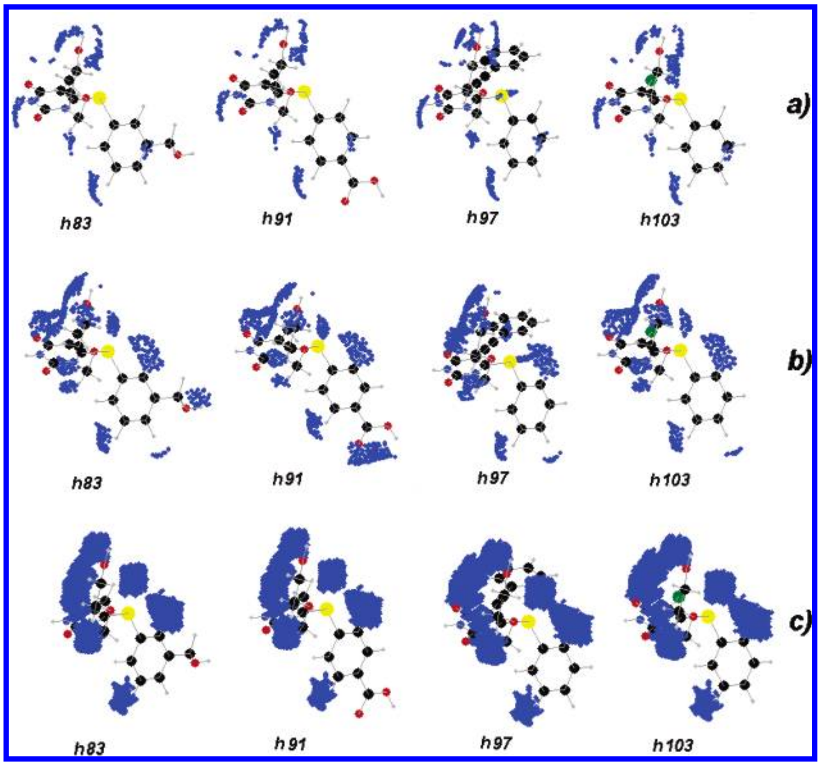


Figure 13. The surface areas (illustrated by the points sampled) of the highest contribution to the anti-HIV activity, as indicated by the IVE-PLS procedure for the training series **h1**−**h80** (a) and the whole series **h1**−**h107** (b). Alternatively, the areas surviving IVE-PLS for the whole series **h1**−**h107** were aggregated for all molecules, and these (25%) of the lowest population were eliminated (c). The plots illustrate only selected molecules.

criterion,[56] as shown in Supporting Information. Clearly, our model does not satisfy this criterion. However, similar results were obtained for the models given in the literature. We think the reason for such behavior is a careless training and test subset selection. Thus, we used the Kennard−Stone (KS) algorithm[57] for sampling the compounds to the training/test subsets (test set: **h1**, **h4**, **h5**, **h6**, **h9**, **h10**, **h15**, **h16**, **h23**, **h25**, **h28**, **h29**, **h34**, **h35**, **h37**, **h49**, **h51**, **h55**, **h57**, **h63**, **h64**, **h77**, **h80**, **h81**, **h85**, **h96**, **h103**). In Figure 12a we illustrated the results of model validation by the GT calculations. Due to the lack of the independent variable (*X*) data used in investigations reported in the literature we could not repeat this procedure to compare our performance to those represented by the literature data. Model quality is much better now, and SDEP value amounts to 0.69. By random testing (Figure 12b) of the relatively small fraction (110 000 models) of all possible 107!/(80!*27!) samplings of the 107 HEPT molecules into the training and test sets (test set: **h3**, **h6**, **h8**, **h11**, **h16**, **h17**, **h22**, **h29**, **h38**, **h44**, **h46**, **h52**, **h53**, **h58**, **h60**, **h61**, **h62**, **h64**, **h73**, **h75**, **h76**, **h81**, **h86**, **h87**, **h91**, **h101**, **h103**; 80/27 molecules) we were able to find models even better fulfilling the GT criterion (Figure 12c). This clearly indicates that the s-CoMSA method provides reliable and highly predictive models. Figure 13 illustrates the molecular plots indicating the regions important for the activity of some representative molecules.

CONCLUSIONS

Shape analysis is a powerful tool in chemistry and drug design. It seems that the approaches including explicit shape representations should gain more attention, being more adequate for the analysis of molecular phenomena that are usually more of less spatially oriented processes. Thus, for example, the investigations of the molecular recognition or receptor−ligand interactions should probably focus *somewhere near a molecular surface* that is still more precise information than *anywhere near a molecule*. In the current publication we have presented a novel formalism for the comparative molecular surface analysis (s-CoMSA). The method enables both quantitative 3D-QSAR modeling and finding possible pharmacophoric sites. The method provides very predictive models for the CBG activity of the benchmark steroid series, tinctorial properties of the heterocyclic azo dyes and anti-HIV activity of the HEPT series.

ACKNOWLEDGMENT

The authors thank Professor Johann Gasteiger of the University of Erlangen-Nürnberg, BRD both for his valuable discussion and for facilitating access to the programs of CORINA, PETRA, SURFACE, and KMAP. The financial support of the KBN Warsaw, grants no. T08E02820, PBZ KBN - 4T09A 088 25, and 4 T09A 034 24, is gratefully acknowledged.

Supporting Information Available: Validation of different models by the Golbraikh−Tropsha criteria for HEPT analogues. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Cramer, III, R. D.; Patterson, D. E.; Bunce, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.
(2) 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.
(3) Hahn, M.; Rogers, D. Receptor surface models. *Perspect. Drug Discov. Des.* **1998**, *12/13/14*, 117−133.

(4) Hahn, M. Receptor surface models: 1. Definition and construction. *J. Med. Chem.* **1995**, *38*, 2080−2090.

(5) Hahn, M.; Rogers, D. Receptor surface models: 2. Application to QSAR. *J. Med. Chem.* **1995**, *38*, 2091−2102.

(6) Ivanciuc, O.; Ivanciuc, T.; Cabrol-Bass, D. 3D QSAR with CoRSA. Comparative receptor surface analysis. Application to calcium channel agonists. *Analusis* **2000**, *28*, 637−642.

(7) Jain, A. N.; Koile, K.; Bauer, B.; Chapman, D. Compass: Predicting biological activities from molecular surface properties. Performance comparisons on a steroid benchmark. *J. Med. Chem.* **1994**, *37*, 2315−2327.

(8) Polanski, J.; Walczak, B. The comparative molecular surface analysis (CoMSA): a novel tool for molecular design. *Comput. Chem.* **2000**, *24*, 615−625.

(9) Anzali, S.; Gasteiger, J.; Holzgrabe, U.; Polanski, J.; Teckentrup, A.; Wagener M. The use of self-organizing neural networks in drug design. *Perspect. Drug Discov. Des.* **1998**, *9/10/11*, 273−299.

(10) Polanski, J.; Gieleciak, R.; Bak, A. The comparative molecular surface analysis (CoMSA) − a nongrid 3D QSAR method by a coupled neural network and PLS system: Predicting pK_a values of benzoic and alkanoic acids. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 184−191.

(11) 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.

(12) 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.

(13) Polanski, J.; Gieleciak, R.; Wyszomirski, M. Mapping dye pharmacophores by the comparative molecular surface analysis (CoMSA): application to heterocyclic monoazo dyes. *Dyes Pigm.* **2004**, *62*, 63−78.

(14) 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*, 481−495.

(15) Polanski, J.; Gieleciak, R. Comparative molecular surface analysis: a novel tool for drug design and molecular diversity studies. *Mol. Diversity* **2003**, *7*, 45−59.

(16) Polanski, J. Self-organizing neural networks for pharmacofore mapping. *Adv. Drug Deliv. Rev.* **2003**, *55*, 1149−1162.

(17) Kohonen, T. *Self-Organization and Associative Memory*, 3rd ed.; Springer: Berlin, 1989.

(18) Wold. S.; Sjöström, M.; Eriksson, L. PLS-regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109−130.

(19) Hasegawa, K.; Matsuoka, S.; Arakawa, M.; Funatsu, K. New molecular surface-based 3D-QSAR method using Kohonen neural network and 3-Way PLS. *Comput. Chem.* **2002**, *26*, 583−589.

(20) Hasegawa, K.; Morikami, K.; Shiratori, Y.; Ohtsuka, T.; Aoki, Y.; Shimma, N. 3D-QSAR study of antifungal N-myristoyltransferase inhibitors by comparative molecular surface analysis. *Chemom. Intell. Lab. Syst.* **2003**, *69*, 51−59.

(21) Hasegawa, K.; Matsuoka, S.; Arakawa, M.; Funatsu, K. Multi-way PLS modeling of structure−activity data by incorporating electrostatic and lipophilic potentials on molecular surface. *Comput. Biol. Chem.* **2003**, *27*, 381−386.

(22) Zupan, J.; Gasteiger, J. *Neural Networks and drug design for Chemists*, 2nd ed.; VCH: Weinheim, 1999.

(23) Timofei, S.; Fabian, W. M. F. Comparative molecular field analysis (CoMFA) of heterocyclic monoazo dye-fiber affinities. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1218−1222.

(24) Jalali-Heravi, M.; Parastar, F. Use of artificial neural networks in a QSAR study of anti-HIV activity for a large group of HEPT derivatives. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 147−154.

(25) Match3D program package, available from Professor J. Gasteiger, Computer-Chemie-Centrum, University Erlangen-Nurnberg, Germany. See: http://www2.ccc.uni-erlangen.de.

(26) Sybyl 6.5. program, available from the Tripos Inc., St. Louis, MO, U.S.A. http://www.tripos.com.

(27) HyperChem 5.0 program, available from the HyperCube Inc., Gainesville, FL, U.S.A. http://www.hyper.com.

(28) MATLAB 6.5. program, available from The Mathworks Inc., Natick, MA. http://www.mathworks.com.

(29) Centner, V.; Massart, D. L.; de Noord, O. E.; de Jong, S.; Vandeginste, B. M. V.; Sterna, C. Elimination of uninformative variables for multivariate calibration. *Anal. Chim. Acta* **1996**, *330*, 1−17.

(30) Testa, B.; Purcell, W. P. A QSAR study of sulfonamide binding to carbonic anhydrase as test of steric models. *Eur. J. Med. Chem.* **1978**, *13*, 509−514.

(31) Motoc, I. *Molecular Shape Descriptors, in Steric Effects in Drug Design*; Charton, M., Motoc, I., Eds.; Akademie: Berlin, 1983; pp 93−105.

(32) Polanski, J. Molecular shape analysis. In *Handbook of chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH: Verlag: Weinheim, 2003; pp 302−319.

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

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

(35) User and Reference Manual Quasar 4.0. http://www.biograf.ch.

(36) Polanski, J.; Bak, A. Modeling steric and electronic effects in 3D and 4D-QSAR schemes: Predicting benzoic pK_a values and steroid CBG binding affinities. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2081−2092.

(37) Peters, R. H. *Textile chemistry. The physical chemistry of dyeing*; Elsevier: Amsterdam, 1975; Vol. III.

(38) Timofei, S.; Schmidt, W.; Kurunczi, L.; Simon, Z. A Review of QSAR for dye affinity for cellulose fibres. *Dyes Pigm.* **2000**, *47*, 5−16.

(39) French, A. D.; Battista, O. A.; Cuculo, J. A.; Gray, D. G. In *Kirk-Othmer Encyclopedia of Chemical Technology*, 4th ed.; Wiley: New York, 1993; Vol. 5, p 476.

(40) Timofei, S.: Schmidt, W.; Kurunczi, L.; Simmon, Z.; Sallo, A. A QSAR study of the adsorption by cellulose fibre of antraquinone vat dyes. *Dyes Pigm.* **1994**, *24*, 267−279.

(41) Timofei, S.; Kurunczi, L.; Schmidt, W.; Fabian, W. M. F.; Simon, Z. Structure-affinity binding relationships by principal component regresion analysis of antraquinone dyes. *Quant. Struct. Act. Relat.* **1995**, *14*, 444−449.

(42) Timofei, S.; Kurunczi. L.; Schmidt, W.; Simon, Z. Structure-affinity binding relationships of some 4-aminobenzene derivatives for cellulose fibre. *Dyes Pigm.* **1995**, *29*, 251−258.

(43) Timofei, S.; Kurunczi. L.; Schmidt, W.; Simon, Z. Lipophilicity in dye-cellulose fibre binding. *Dyes Pigm.* **1996**, *32*, 25−42.

(44) Fabian, W. M. F. Timofei, S.; Kurunczi. L. Comparative molecular field analysis (CoMFA), semiempirical (AM1) molecular orbital and multiconformational minimal steric difference (MTD) calculation of antraquinone dye-fibre affinities. *J. Mol. Struct. THEOCHEM* **1995**, *340*, 73−81.

(45) Fabian, W. M. F.; Timofei, S. Comparative molecular field analysis (CoMFA) of dye-fibre affinities II: symmetrical bisazo dyes. *J. Mol. Struct. THEOCHEM* **1996**, *362*, 155−162.

(46) Oprea, T. I.; Kurunczi, L.; Timofei, S. QSAR studies of disperse azo dyes towards the negation of the pharmacophore theory of dye − fibre interaction? *Dyes Pigm*. **1997**, *33*, 41−64.

(47) Funar-Timofei, S.; Schüürmann, 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.

(48) Polanski, J.; Gieleciak, R.; Bak, A. Probability issues in molecular design: Predictive and modeling ability in 3D-QSAR schemes. *Comb. Chem. High Throughput Screening* in print.

(49) Kireev, D. B.; Chretien, J. R.; Grierson, D. S.; Monneret, C. A 3D QSAR study of a series of HEPT analogues: the influence of conformational mobility on HIV-1 reverse transctriptase inhibition. *J. Med. Chem.* **1997**, *40*, 4257−4264.

(50) Hannongbua, S.; Nivesanond, K.; Lawtrakul, L.; Pungpo, P.; Wolschann, P. 3D-Quantitative structure−activity relationships of HEPT derivatives as HIV-1 reverse trascriptase inhibitors, based on ab initio calculations. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 848−855.

(51) Luco, J. M.; Ferretii, F. H. QSAR based on multiple linear regression and PLS methods for the anti-HIV activity of a large group of HEPT derivatives. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 392−401.

(52) Douali, L.; Villemin, D.; Charquaoui, D. Comparative QSAR based on neural networks for the anti-HIV activity of HEPT derivatives. *Curr. Pharm. Des.* **2003**, *9*, 1817−1826.

(53) Mager, P. P. Hybrid canonical-correlation neural-network approach applied to nonnucleoside HIV-1 reverse transcriptase inhibitors (HEPT derivatives). *Curr. Med. Chem.* **2003**, *10*, 1643−1659.

(54) Douali, L.; Villemin, D.; Charquaoui, D. Neural networks: accurate nonlinear QSAR model for HEPT derivatives. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1200−1207.

(55) Gayen, S.; Debnath, B.; Samanta, S.; Jha, T. QSAR study on some anti-HIV HEPT analogues using physicochemical and topological parameters. *Bioorg. Med. Chem.* **2004**, *12*, 1493−1503.

(56) Golbraikh. A.; Thropsha, A. Beware of q^2 ! *J. Mol. Graph. Mod.* **2002**, *20*, 269−276.

(57) Kennard, R. W.; Stone, L. A. Computer aided design of experiments. *Technometrics* **1969**, *11*, 137−148.

CI049960L