Modeling Robust QSAR. 1. Coding Molecules in 3D-QSAR − from a Point to Surface Sectors and Molecular Volumes

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

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

Received April 26, 2005

Shape analysis is a powerful tool in chemistry and drug design. In the current work, we compare the results of CoMFA and Comparative Molecular Surface Analysis (CoMSA), the 3D-QSAR method, for a series of hypolipidemic and antiplatelet asarones and antifungal N-myristoyltransferase inhibitors. In this publication we show that a sector CoMSA formalism enables an analysis of the biological activity that is more directly related to the molecular shape and individual molecular functionalities than the traditional uniform and directionless CoMFA field. Iterative Variable Elimination allowed us to identify the potential pharmacophoric sites. We modeled QSARs for both series and demonstrate that sector-based molecular descriptors give very predictive models and allow one to generate a spatial interpretation of the QSAR models. In particular, we identified the central aromatic ring and carbonyl functions as the moieties determining the activity of the asarones series, while the pattern of substitution of the aromatic ring determines the activity of N-myristoyltransferase inhibitors.

INTRODUCTION

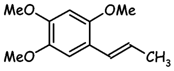
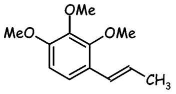
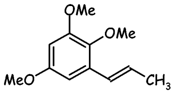
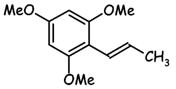
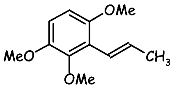
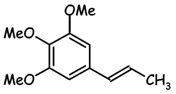
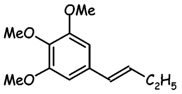
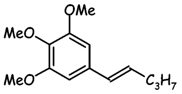
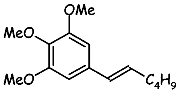
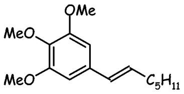
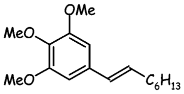
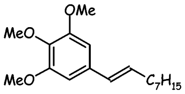
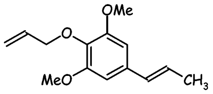
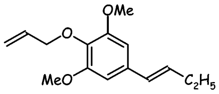
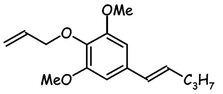
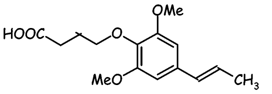
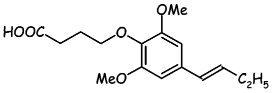
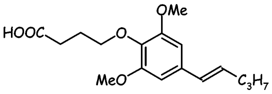
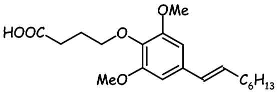
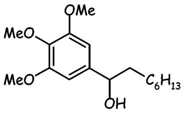
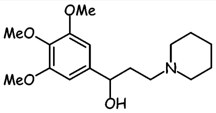
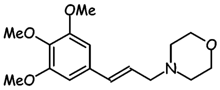
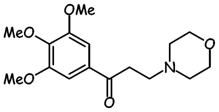
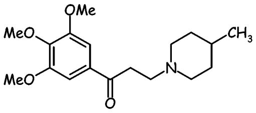
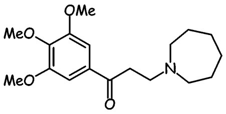
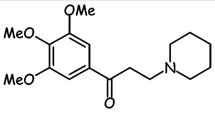
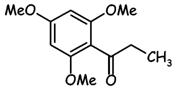
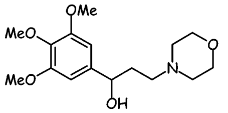
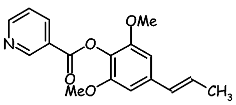
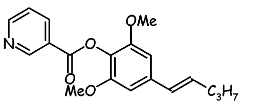
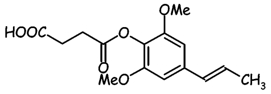
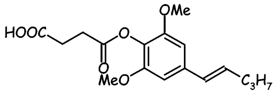
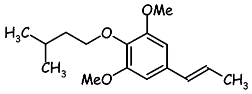
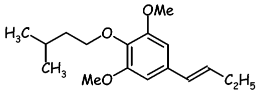
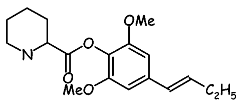
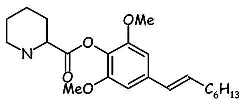
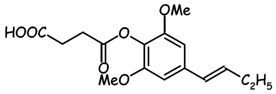
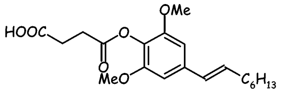
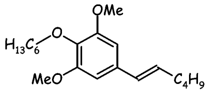
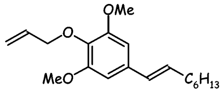
It may appear to be a paradox, but *the most fundamental and lasting objective of* (chemical) *synthesis is not the* production of new compounds but the production of properties.[1] Molecular design is a computational tool for screening virtual chemical compound space in a search for novel properties, and QSAR should function like a dictionary between molecular structures and properties. This clearly makes it an essential and irreplaceable method in molecular design. However, more and more sophisticated tools are needed for the efficient and robust transformation of molecular structure space into compound property space. A variety of issues decide the efficiency of the QSAR methods,[2] and their practical importance for drug design is still controversial.[3] Although it is common to think of QSAR as a drug design method, in fact, this technique is more an a posteriori data analysis process than real design. First, it is not the same to start from a molecular structure and attempt to predict its properties as to begin from a property and find the molecule having this particular property. Second, the term 'design' anticipates the extrapolation of the data to novel objects. However, the fact that a very minute molecular modification can evoke a substantial activity change makes such extrapolation extremely risky. Molecular superposition and conformational flexibility present further problems for QSAR modeling. In this context, a number of possible molecular arrangements and configurations can be generated before multidimensional QSAR modeling. This makes QSAR a highly data dependent operation and introduces substantial noise into QSAR results. Can we decrease data dependency in QSAR, making it less sensitive to the variation in inputs using novel, more robust systems. Different superimposition rules provide completely different activity contour plots in CoMFA.[4] Recently, several improvements in structure overlay have appeared that allow for more flexible or sophisticated superimposition.[5] Hopfinger's 4D-QSAR investigating molecular conformational space has been supplemented by architecture that uses a fuzzy self-organizing neural neuron.[6−8] Data handling can improve QSAR robustness, and the application of new computational methods including neural networks, data elimination, genetic algorithms, and novel model validation schemes have often been reported in this field. Essentially, the efficiency of data handling largely depends on the descriptors that are used for the characterization of the molecular objects analyzed. Thus, the method that is used for coding molecules in 3D- or 4D-QSAR is an important factor that does influence final model robustness. In the CoMFA-like fields a molecule is represented by a set of points determined in space by a 3D grid.[9] Different smooth and box CoMFA-like fields have recently been thoroughly tested.[10] A surface can serve as the base for the molecule description in several methods, e.g., Compass, CoRSA, SURFCOMP, or CoMSA, which are based on sampling points on the molecular surface or near such a surface.[11−14] Various algorithms have been developed for the comparison of the surface sectors. Thus, the lattice generated in CoRSA on the molecular surface defines the nodes that are further compared for the series of individual molecules. Alternatively, to compare surface sectors for a series of molecules, molecular volumes can be defined in analyzed molecules. Different algorithms can be used to generate such volumes, which can take the form of a rectangular cube (s-CoMSA)[15] or a sphere generated by self-organizing neural networks (SOM-CoMSA)[16−22] or a supervised neural network (Compass).[11] Similarly, in Hopfinger's 4D-QSAR a molecule is coded openly by the descriptors defining the pattern in which atoms occupy volume sectors.[23] Alternatively, a self-organizing neural network can be used for the generation of molecular volumes in SOM-4D-QSAR.[24]

In the present work we investigate the influence of the way in which molecules are coded on the efficiency and

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

10.1021/ci0501488 CCC: $30.25
Published on Web 08/31/2005

Table 1. α-Asarone Compounds and Their Atherogenic Index − $I_{TG/LDL}$

		$I_{TG/LDL}$
a1		2.10
a2		1.27
a3		1.56
a4		1.34
a5		2.35
a6		1.31
a7		0.36
a8		0.45
a9		0.16
a10		0.27
a11		0.13
a12		0.14
a13		0.47
a14		0.15
a15		0.10
a16		0.45
a17		0.22
a18		0.13
a19		0.85
a20		0.77
a21		1.74
a22		2.05
a23		2.00
a24		1.38
a25		1.48
a26		1.71
a27		1.90
a28		1.26
a29		3.15
a30		1.52
a31		1.79
a32		3.28
a33		0.45
a34		0.56
a35		1.98
a36		2.58
a37		1.92
a38		1.96
a39		1.28
a40		0.80

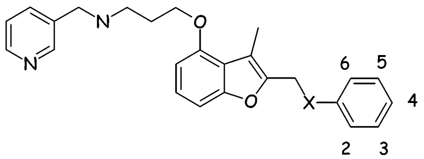
robustness of 3D-QSAR modeling. In particular, we compare the traditional molecular point field generated by the CoMFA and volume-based descriptors used by the CoMSA method. Molecules are rich in analogies, and different molecular descriptors can provide similar QSAR models. Although it is believed that the method used for the calculation of partial atomic charges highly influences resultant QSAR, Doweyko[4] showed that CoMFA provides models of a similar statistical quality independent of the calculation method used. In a context of QSAR, this problem introduces further noise, namely, *molecular similarity noise*. In our investigations we analyzed the application of the Iterative Variable Elimination[19] for the indication of the molecular areas that are specific to biological activity, which should reduce similarity noise and indicate possible pharmacophoric sites.

We investigated asarones **a1**−**a40** and benzofuranes **b1**−**b29**, the same compound series that have previously been investigated with 3D-QSARs.[25,26]

EXPERIMENTAL SECTION

Model Builders. All of the experimental data, i.e., **a1**−**a40** and **b1**−**b29**, were extracted from refs 25 and 26 and are given in Tables 1 and 2, respectively.

Table 2. NMT Inhibitory Activity of Benzofurans



no.	X	R2	R3	R4	R5	R6	log(1/IC$_{50}$)
b1	O	F	H	F	H	H	8.12
b2	O	H	CF$_3$	H	H	H	6.72
b3	O	H	H	H	H	H	7.14
b4	O	H	H	Cl	H	H	7.14
b5	S	H	H	H	H	H	6.21
b6	S	H	H	Cl	H	H	5.71
b7	O	F	H	H	H	H	8.08
b8	O	H	F	H	H	H	6.95
b9	O	F	H	H	H	F	7.47
b10	O	F	H	H	F	H	8.36
b11	O	F	F	H	H	H	8.44
b12	O	F	F	H	F	H	8.18
b13	O	F	H	F	F	H	8.03
b14	O	F	F	F	H	H	8.24
b15	O	F	F	H	H	F	7.48
b16	O	F	H	F	H	F	7.09
b17	O	F	F	F	F	F	5.85
b18	O			2-Py[a]			5.53
b19	O			3-Py[a]			7.24
b20	O			4-Py[a]			5.81
b21	O	CN	H	H	H	H	7.78
b22	O	H	CN	H	H	H	7.03
b23	S	H	H	F	H	H	5.78
b24	O	F	F	H	F	F	6.24
b25	O	F	H	Br	H	H	7.55
b26	O	H	Br	H	H	H	6.40
b27	O	H	H	Br	H	H	6.06
b28	O	2-Py[a]	Cl	H	H	H	6.49
b29	O	2-Py[a]	H	Cl	H	H	6.64

[a] Py − pyridine.

All of the molecules were superimposed prior to calculation of the molecular surfaces. The superimposition was performed by covering the central benzene ring of molecules (**a1**−**a40**) and the central benzofurane ring of molecules (**b1**−**b29**). We used the program Match3D for performing this operation.[27] Calculation of the molecular surface descriptors was based on molecular volumes (SOM-CoMSA and s-CoMSA).

The SOM-CoMSA formalism based on the self-organizing neural network has been previously described[15−22] and will not be detailed here. Hasegawa described a similar but slightly modified procedure.[28,29] For the calculation of shape s-CoMSA descriptors we used formalism similar to Hopfinger's 4D-QSAR grid coding system.[23] 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 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.

Formally, the descriptors are defined as follows. Each molecule m is represented by a set of points sampled on the molecular surface $P_m(x, y, z, \upsilon)$ where x, y, and z are coordinates in three-dimensional space and υ represents a surface property in a given point, e.g., υ is an electrostatic potential value.

Let $P_{mi}(x, y, z, \upsilon)$ be a subset of $P_m(x, y, z, \upsilon)$ such that all its elements are included in sector i. This is satisfied when

$$\begin{gathered} x_k \geq x_{l_i} \wedge x_k < x_{h_i} \wedge \\ y_k \geq y_{l_i} \wedge y_k < y_{h_i} \wedge \\ z_k \geq z_{l_i} \wedge z_k < z_{h_i} \end{gathered} \qquad (1)$$

where x_{l_i}, y_{l_i}, z_{l_i} and x_{h_i}, y_{h_i}, z_{h_i} are the elements of vectors l_i and h_i which define the highest and lowest sector borders, respectively, and where i indexes a sector and k indexes the points sampled within the individual sectors. P_{mi} is an empty set if no k satisfies condition (1).

The s-CoMSA calculation lies in the generation of matrix D_A of size $g \times n$ where g is a number of analyzed molecules and n is the total number of sectors, and where d_{mi} is given by

$$d_{mi} = \begin{cases} \dfrac{\sum_{k=1}^{k_{\max}} (\nu_{mi})_k}{k_{\max}} & , \quad \text{when } k_{\max} \neq 0 \\ 0 & , \quad \text{when } k_{\max} = 0 \end{cases} \qquad (2)$$

and $k_{\max}$ is the number of points on the surface P of the molecule m enclosed in sector i.

Matrix D defined by eq 2 describes each molecule independently of all others. After 4D-QSAR nomenclature, we defined such a descriptor absolute occupancy, D_A.

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).[30]

A PLS model was constructed for the centered data, and its complexity was estimated based on the leave-one-out cross-validation (CV) procedure. 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 based on the model with one, two, three, etc. factors. The Root Mean Square Error of CV for the model with j factors is defined as

$$\mathrm{RMSECV}_j = \sqrt{\frac{\sum(\mathrm{obs} - \mathrm{pred}_j)^2}{m}} \qquad (3)$$

where *obs* denotes the assayed value and *pred* denotes the predicted value of the dependent variable. A model with k factors, for which RMSECV reaches a minimum, is considered as an optimal one.

For the construction of the individual model reported in this work, the optimal number of latent PLS variables was truncated not to exceed the value of 1−10, respectively, which is clearly indicated in the figures.

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

$$q_{\mathrm{cv}}^{2} = 1 - \frac{\sum(\mathrm{obs} - \mathrm{pred})^2}{\sum(\mathrm{obs} - \mathrm{mean(obs)})^2} \tag{4}$$

where *obs* is the assayed values, *pred* is the predicted values, mean is the mean value of obs, and the cross-validated standard error *s*

$$s = \sqrt{\frac{\sum(\mathrm{obs} - \mathrm{pred})^2}{m - k - 1}} \tag{5}$$

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

Before 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

$$\mathrm{SDEP} = \sqrt{\frac{\sum(\mathrm{pred} - \mathrm{obs})^2}{n}} \tag{6}$$

where *pred* is the predicted value, *obs* is the observed value, and *n* is the number of measurements.

Data Elimination. To identify the parts of the molecular surface that contribute the most to activity, we used a modified procedure of the PLS with Uninformative Variable Elimination (UVE-PLS), namely the Iterative Variable Elimination PLS (IVE-PLS) procedure.[19] The UVE algorithm was originally proposed by Centner et al.[31] as a possible improvement to the PLS models. The main idea of the method is to reduce the number of 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 the presentation of the relation between the **Y** answer and **X** predictors in the form of

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

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 in the final PLS model. To estimate the cutoff level artificial random number noise is generated (the level of the noise is 10^{-10} of the original variable order) and included (as additional columns) in the matrix of the original variables. 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*)), observed in the noisy column determines the cutoff level for the original variables.

Modified Uninformative Variable Elimination based on Iterative Leave-One-Out Cross-Validation (IVE-PLS). Below we describe a modified procedure for Uninformative Variable Elimination (UVE-PLS). Instead of a single step procedure, we used here an iterative algorithm based on the abs(mean(*b*)/s(*b*)) criterion to identify variables to be eliminated. To distinguish this procedure, we named this Modified Uninformative Variable Elimination with the iterative leave-one-out cross-validation (IVE-PLS). This procedure includes the following: 1. standard PLS analysis applied to analyze the matrices yielded from the s-CoMSA procedure with the leave-one-out cross-validation to estimate the performance of the PLS model (q^2), 2. elimination of the matrix column of the lowest abs(mean(*b*)/s(*b*)) value, 3. standard PLS analysis of the new matrix without the column eliminated in step 2, and 4. iterative repetition of the steps 1−3 to maximize the LOO CV q^2 parameter.

The UVE and IVE procedures were programmed within the MATLAB environment (MATLAB).[30] All MATLAB functions and m-scripts are available from the authors upon request.

RESULTS AND DISCUSSION

Separating the molecule into partitions of spatial regions of certain volumes, either filled or unfilled by atoms or groups of atoms, can provide an interesting method for calculation of the molecular descriptors for efficient QSAR modeling.[32] Hopfinger’s 4D-QSAR method uses similar formalism for the description of the molecular conformational space. Thus, in Figure 1 we compare some of the methods using the sector formalism (CoMSA, receptor-like neuron network, 4D-QSAR) to the conventional CoMFA point formalism. We show below that such formalism not only increases the fuzziness of the molecular description[13] but also significantly influences the performance of the QSAR model generated. It also changes the visual pattern in which such a QSAR model can be presented.

Hypolipidemic and Antiplatelet Asarones. Excess food intake makes lipid metabolism disorders a common plague of contemporary society. The therapy of such disorders is based on drugs having hypolipidemic activity, such as fibric acid derivatives, the inhibitors of 3-hydroxymethyl-conzyme A (3-HMG Co-A) reductase, an enzyme involved in de novo sterols synthesis, probucol, lifibrol, and others.[33,34] α-Asarones are compounds having hypolipidemic activity[35,36] and have been investigated in many pharmacological,[37,38] toxicological,[39,40] and QSAR studies.[41,42] We have recently applied CoMSA analysis for the design of potential asarone drugs. Cross-validated q^2_{cv} values range from 0.49; $s = 0.66$ (CoMFA) to $q^2_{\mathrm{cv}} = 0.69$; $s = 0.54$ (SOM-CoMSA). This demonstrates a correlation between the descriptors and the hypolipidemic activity of the asarone series.[43] In Figure 2, we analyzed a model performance during the IVE-PLS variable elimination. This procedure was performed in such a way that the optimal number of PLS latent components was always estimated. This number, however, was truncated not to exceed the values of 1−10, respectively. We conclude this section with a few general remarks. First, CoMFA and s-CoMSA provide initial models of similar q^2_{cv} performance of ca. $q^2_{\mathrm{cv}} = 0.5$. The neural SOM-CoMSA technique allowed us to increase this value to ca. $q^2_{\mathrm{cv}} = 0.7$, even without data elimination. IVE-PLS enabled an increase in model performance measured by q^2_{cv} to ca. 0.7 (CoMFA) or 0.8 (SOM-CoMFA); 0.9 (s-CoMSA). For both CoMSA methods, model quality clearly depends on the maximal number of the PLS latent variables that can be included in the model. In contrast, the q^2_{cv} value in IVE-PLS-CoMFA does not depend on this number. We would also like to stress

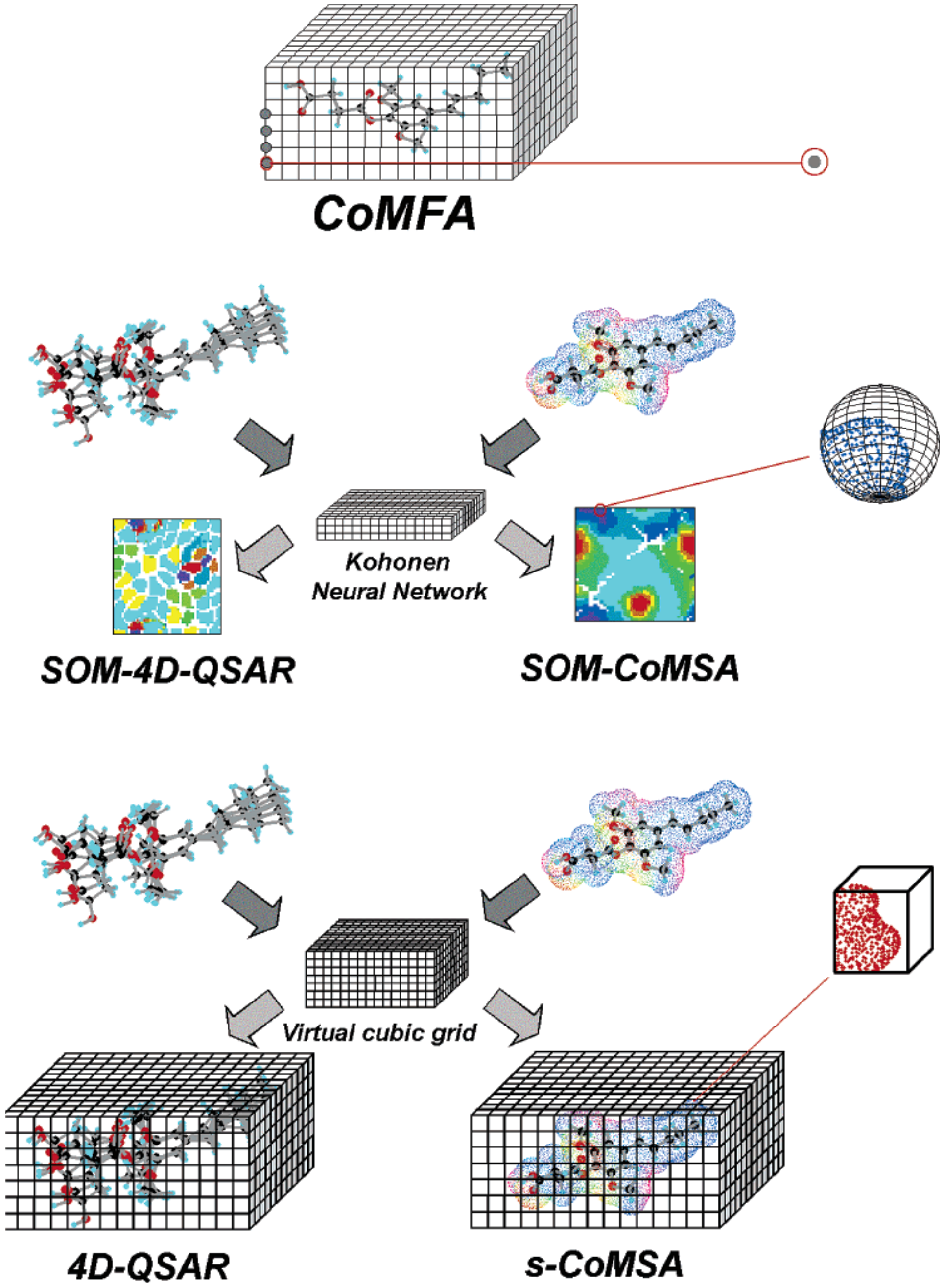


Figure 1. A schematic illustration comparing a uniform CoMFA grid (a sharply defined point defines molecular feature) with the neural SOM-CoMSA and SOM-4D-QSAR (a fuzzy SOM neuron defined defines molecular feature) and sector (Hopfinger's) 4D-QSAR and s-CoMSA (cubic sector defines molecular feature).

that external predictability does not change during IVE-PLS data elimination if tested by the SDEP parameter (results not shown). A second conclusion is that although the neural method provides a better starting model, data elimination is more efficient for the s-CoMSA. This indicates that the additional model improvement due to neural network fuzziness cannot be easily *extrapolated* to other models, and IVE-PLS does not enhance this extra q^2_{cv} gain.

Figure 3 compares the visualization of the compound's activity by the CoMFA and CoMSA methods. Thus, the uniform CoMFA field filtered by either a standard deviation value (Figure 3a) or further transformed by data elimination (IVE-PLS − Figure 3b) gives a uniform illustration for all molecules. Unlike CoMFA, both CoMSA methods explain the compounds' activity by the indication of the areas that can be easily differentiated for active and inactive molecules. This effect results from the dissimilarity of the molecular surfaces of the individual molecules. The respective color-coding allows us to identify the influence of the points sampled on the molecular surface by the combination of the electrostatic value and a value of the b weight in the PLS model. Points contributing to the activity on a level close to 0 (near 90% of the points sampled) were omitted. Such an illustration suggests the key pharmacophore for the asarones investigated. Thus, an area near the central aromatic ring substituted with alkoxyl functionality provides a negative contribution (blue-colored sections). Since lower values indicate higher activity, a *negative contribution* increases a compound's activity. A negatively charged carbonyl oxygen in the side chain generally causes lower activity, clearly decreasing the activity as illustrated by the yellow and red molecular surface areas. This rule can be proved by

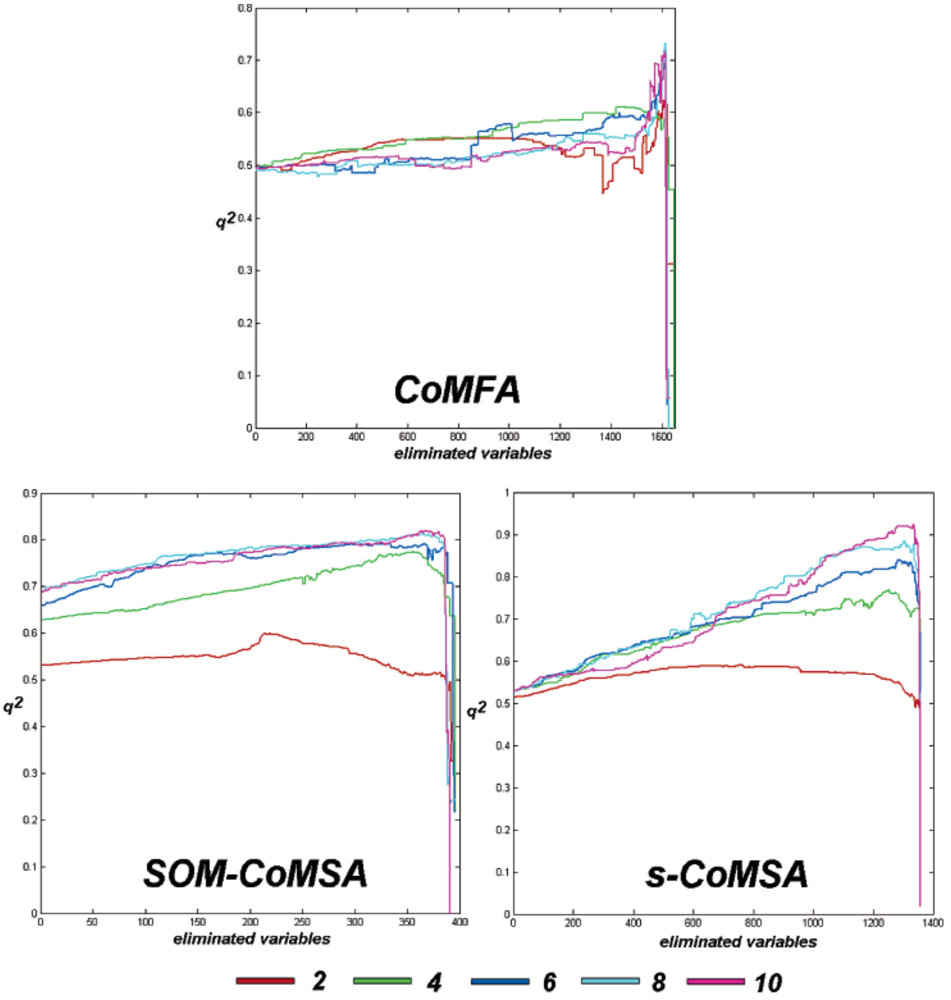


Figure 2. Profiles reporting the IVE-PLS modeling processes for the CoMFA and CoMSA for the asarones series, details in text.

examination of the structures given in Table 1. It is worth noticing that if a hydroxyl function replaces a carbonyl oxygen, e.g. compound **a28**, we do not observe an activity decrease similar to that affected by a carbonyl group, e.g., compound **a26**.

Antifungal N-Myristoyltransferase Inhibitors. N-Myristoyltransferase (NMT) is an enzyme that catalyzes the transfer of myristic acid, from myristoyl-CoA to the N-terminal glycine's amine. This process is common for a variety of eukaryotic organisms.[26] Benzofuran compounds have been found to be selective C albicans NMT selective inhibitors. Since C albicans are organisms causing systemic fungal infections in immunocompromised patients, the compounds can be important drug candidates in AIDS therapy.[26]

In Figure 4 we illustrate the profiles describing IVE-PLS data elimination for the SOM and s- CoMSA methods, respectively. This allowed us to obtain final SOM-CoMSA ($q^2_{\mathrm{cv}} = 0.84$) and s-CoMSA ($q^2_{\mathrm{cv}} = 0.96$) models, which compare well with the Hasegawa SOM-CoMSA with a genetic algorithm (GA-SOM-CoMSA) model characterized by $q^2_{\mathrm{cv}} = 0.81$.[26] Similar, to the models discussed above for asarones the sector method defined on the basis of the cubic grid appeared superior, providing slightly better final models. Figure 5 indicates the key surface sectors that are important for compound activity. Thus, a blue and red colored area indicates a positive contribution, which in this particular case increases the activity, while cyan and magenta sections generally decrease the activity. The Hasegawa model identifies for the electron-withdrawing substituents at the 2-, 3-, and 5-positions as the critical factors for activity. Generally, our model reveals a similar effect. Thus, the distribution of the electrostatic potential within the benzene ring decides the activity. For example, a large blue area determines the high activity of the compound **11b** in comparison to the low activity of **18b** (Figure 5).

From a theoretical point of view an interesting point of the Hasegawa series is the fact that individual compounds change only in the region of the aromatic ring. Thus, this region differs the most between compounds and should also be indicated as specific for the interaction in data elimination during QSAR analysis. In fact, this is true for the IVE-PLS models performed for the lower PLS latent variable numbers; however, the inclusion of a higher number of latent variables can also reveal possible interactions in the sectors that are common for all molecules. This is clearly illustrated by Figures 1 and 2 in the Supporting Information that analyze the areas of specific interactions as a function of the maximal number of latent variables that can be included in the model.

CONCLUSIONS

Shape analysis is a powerful tool in chemistry and drug design. In the current work, we compare the results of CoMFA and Comparative Molecular Surface Analysis

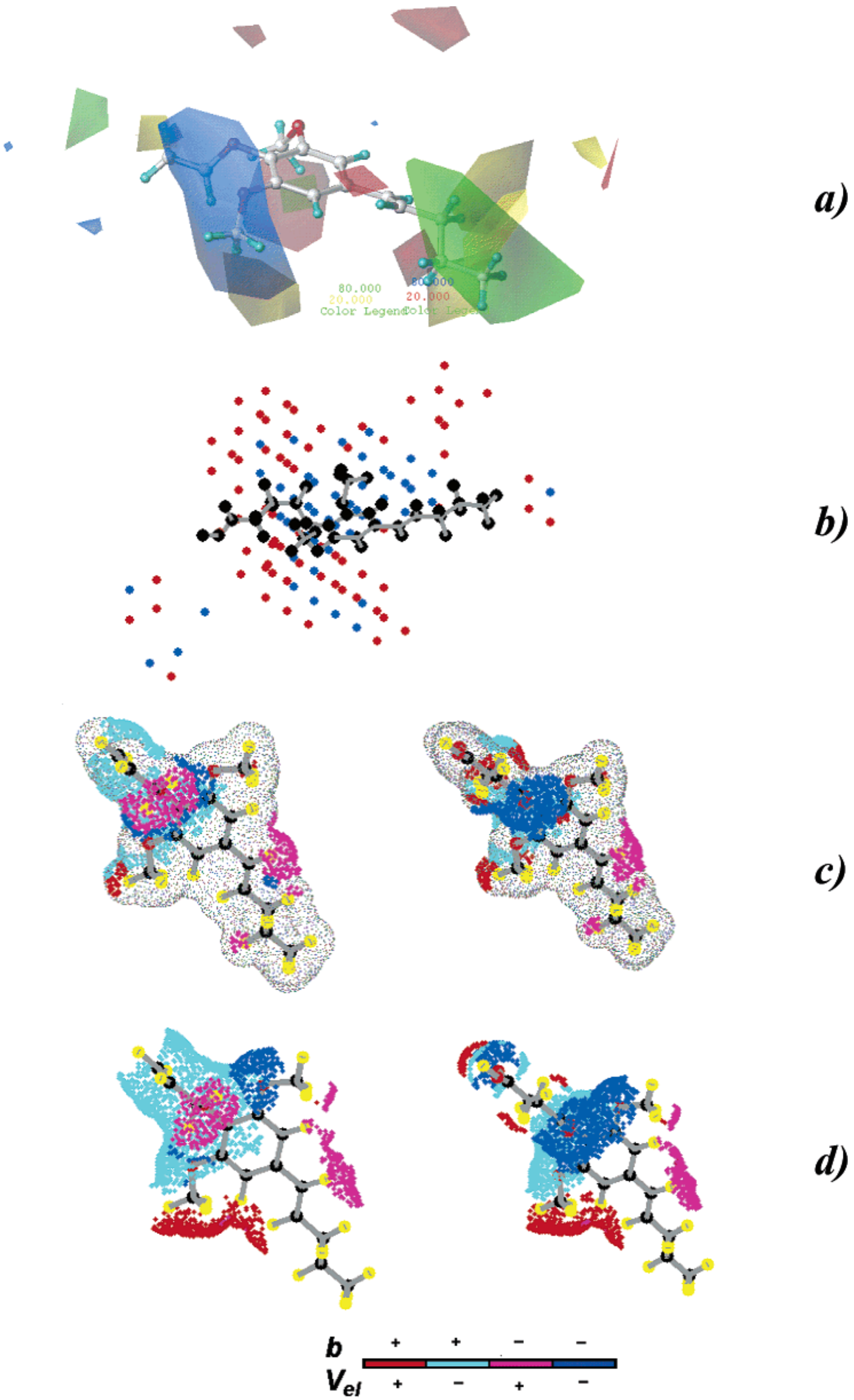


Figure 3. An illustration of the molecular areas that are key for the antipletelet asarones **a1**−**a40** activity, in a form of the traditional CoMFA profiles (a), CoMFA-IVE-PLS (b), SOM-CoMSA-IVE-PLS (c), and s-CoMSA-IVE-PLS (d). Compounds of the highest and lowest activity are shown for the CoMSA plots. Color codes indicate the following: for figure a − standard CoMFA coding; for figure b − blue − 50 points of the highest contribution to the model; red − next 100 points (out of 1650 points); figure c and d − a combination of the electrostatic potential sign and the sign of the b weight in the model, as shown by the colorbar: +/+ (red − decreases the activity), −/+ (cyan − increases the activity), +/− (magenta − increases the activity), −/− (blue − decreases the activity).

(CoMSA), the 3D QSAR method, for a series of hypolipidemic and antiplatelet asarones and antifungal N-myristoyltransferase inhibitors. In this publication we show that a sector CoMSA formalism enables an analysis of the biological activity that is more directly related to the molecular shape and individual molecular functionalities than the traditional uniform and directionless CoMFA field. Iterative Variable Elimination allowed us to identify the potential pharmacophoric sites. We modeled QSARs for both series and demonstrate that sector-based molecular descriptors give

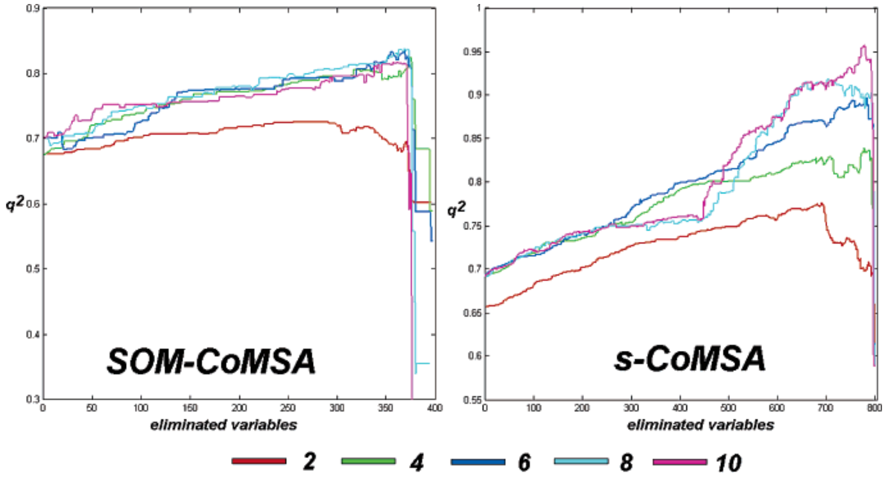


Figure 4. Profiles reporting the IVE-PLS modeling processes for the CoMFA and CoMSA, for the Hasegawa benzenefurans, details in text.

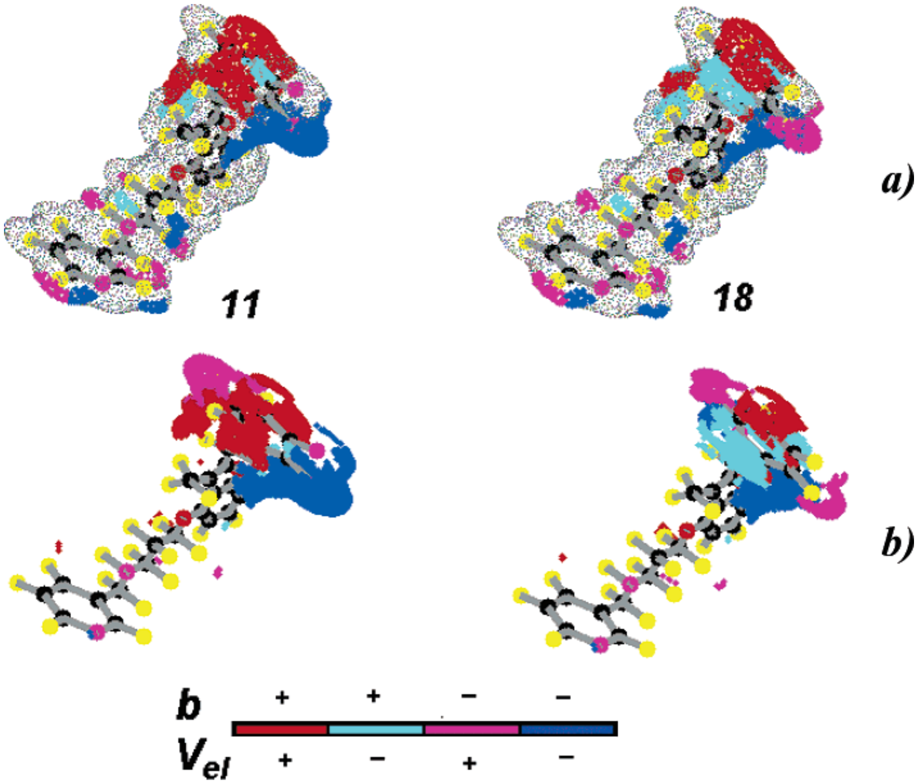


Figure 5. An illustration of the molecular areas that are key for the antifungal activity of compounds **b1**−**b29** (two compounds of the highest and lowest activity are shown): SOM-CoMSA-IVE-PLS (a) and s-CoMSA-IVE-PLS (b). Color codes indicate a combination of the electrostatic potential sign and the sign of the b weight in the model: +/+ (red − increases the activity), −/+ (cyan − decreases the activity), +/− (magenta − decreases the activity), −/− (blue − increases the activity), as shown by the colorbar.

very predictive models and allow one to generate a spatial interpretation of the QSAR models. In particular, we identified the central aromatic ring and carbonyl functions as the moieties determining the activity of the asarones series, while the pattern of substitution of the aromatic ring determines the activity of N-myristoyltransferase inhibitors.

ACKNOWLEDGMENT

The authors thank Professor Johann Gasteiger of the University of Erlangen-Nürnberg, BRD for facilitating access to the CORINA, PETRA, SURFACE, and KMAP programs. The financial support of the KBN Warsaw under grant nos. KBN 3T09 A01127 and PBZ 040 P04/08 is gratefully acknowledged. R.G. thanks Foundation for Polish Science for an individual grant.

Supporting Information Available: Illustrations of the molecular areas that are key for the antifungal activity as a function of the maximal number of the PLS latent variables included in the model for compounds **b1**−**b29** SOM-CoMSA-IVE-PLS (Figure 1) and s-CoMSA-IVE-PLS (Figure 2). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Kolb, H. C.; Finn, M. G.; Sharpless, K. B. Click Chemistry: Diverse Chemical Function from a Few Good Reactions. *Angew. Chem., Int. Ed.* **2001**, *40*, 2004−2021.

(2) Kubinyi, H. QSAR: Hansch Analysis and Related Approaches. In *Methods and Principles in Medicinal Chemistry*; Mannhold, R., Krogsgaard-Larsen, P., Timmerman, H., Eds.; VCH: Weinheim, 1993.

(3) Wermuth, C. G. The impact of QSAR and CADD methods on drug discovery. In *Rational Approaches to Drug Design* − *Proceedings of the 13th European Symposium on Quantitative Structure*−*Activity Relationships*; Holtje, H.-D., Sippl, W., Eds.; Prous: Barcelona, 2001; pp 3−20.

(4) Doweyko, A. M. 3D-QSAR illusions. *J. Comput.-Aided. Mol. Des.* **2004**, *18*, 587−96.

(5) Korhonen, S.-P.; Tuppurainen, K.; Laatikainen, R.; Peräkylä, M. FLUFF-BALL, A template-based grid-independent superposition and QSAR technique: Validation using a benchmark steroid data set. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1780−1793.

(6) 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.

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

(8) Polanski, J.; Bak, A.; Gieleciak, R.; Magdziarz, T. Self-organizing neural networks for modeling robust 3D and 4D QSAR: Application to dihydrofolate reductase inhibitors. *Molecules* **2004**, *9*, 1148.

(9) 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.

(10) Melville, J.; Hirst, J. D. On the stability of CoMFA models. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1294−1300.

(11) 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*, 2315−2327.

(12) 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.

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

(14) Hofbauer, C.; Aszodi, A. SH2 Binding site comparison: A new application of the SURFCOMP method. *J. Chem. Inf. Model*. **2005**, *45*, 414−421.

(15) Polanski, J.; Gieleciak, R.; Magdziarz, T. The grid formalism for the comparative molecular surface analysis: application to the CoMFA benchmark steroids, azo dyes and HEPT derivatives. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1423−1435.

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

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

(18) 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.

(19) 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.

(20) 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.

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

(22) Polanski, J.; Gieleciak, R.; Bak, A. Probability issues in molecular design: Predictive and modeling ability in 3D-QSAR schemes. *Comb. Chem. High Throughput Screening* **2004**, *7*, 793−807.

(23) 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.

(24) 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.

(25) Chilmonczyk, Z.; Siluk, D.; Kaliszan, R.; Lozowicka, B.; Poplawski, J.; Filipek, S. New chemical structures of hypolipidemic and antiplatelet activity. *Pure Appl. Chem.* **2001**, *73*, 1445−1458.

(26) 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.

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

(28) 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.

(29) 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.

(30) MATLAB 5.0 program, available from: The Mathworks Inc., Natick, MA. http://www.mathworks.com.

(31) Centner, V.; Massart, D. L.; de Noord, O. E.; de Jong, S.; Vandeginste, B. M. V.; Sterna, C. Elimination of uninformative variables for multivariate calibration. *Anal. Chem.* **1996**, *68*, 3851−3858.

(32) 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.

(33) Eghdamian, B.; Ghose, K. Mode of action and adverse effects of lipid lowering drugs. *Drugs Today* **1998**, *34*, 943−956.

(34) Farmer, J. A.; Gotto, A., Jr. Current and future therapeutic approaches to hyperlipidemia. *Adv. Pharmacol.* **1996**, *35*, 79−114.

(35) Chamorro, G.; Garduno, L.; Sanchez, A.; Labarrios, F.; Salazar, M.; Martinez, E.; Diaz, F.; Tamariz, J. Hypolipemic activity of dimethoxy unconjugated propenyl side-chain analogues of alpha-asarone in mice. *Drug Dev. Res.* **1998**, *43*, 105−108.

(36) Labarrios, F.; Garduno, L.; Vidal, M.; Garcia, R.; Salazar, M.; Martinez, E.; Diaz, F.; Chamorro, G.; Tamariz, J. Synthesis and hypolipidaemic evaluation of a series of alpha-asarone analogues related to clofibrate in mice. *J. Pharm. Pharmacol.* **1999**, *51*, 1−7.

(37) Dandiya, P. C.; Sharma, J. D. Studies on Acorus calamus. V. Pharmacological actions of asarone and beta-asarone on central nervous system. *Indian J. Med. Res.* **1962**, *50*, 46−60.

(38) Dandiya, P. C.; Menon, M. K. Effects of asarone and beta-asarone on conditioned responses, fighting behaviour and convulsions. *Br. Pharm. Chemother.* **1963**, *20*, 436−442.

(39) Belova, L.; Alibekov, S.; Baginskaya, A.; Sokolov, S.; Pokrovskaya, G.; Stikhin, V.; Trumpe, T.; Gorodnyuk, T. Asarone and its biological properties. *Farmak. Toksikol.* **1985**, *48*, 17−20.

(40) Salazar, M.; Salazar, S.; Ulloa, V.; Mendoza, T.; Pages, N.; Chamorro, G. Teratogenic action of alpha-asarone in the mouse. *J. Toxicol. Clin. Exp.* **1992**, *12*, 149−154.

(41) Filipek, S.; Lozowicka, B. Alpha-asarone congeners as hypolipidemic agents. Pseudoreceptor versus minireceptor modeling. *Acta Pol. Pharm.* **2000**, *57*, 106−109.

(42) Cruz, M.; Salazar, M.; Garciafigueroa, Y.; Hernandez, D.; Diaz, F.; Chamorro, G.; Tamariz, J. Hypolipidemic activity of new phenoxyacetic derivatives related to alpha asarone with minimal pharmacophore features *Drug. Dev. Res.* **2003**, *60*, 186−190.

(43) Magdziarz, T.; Lozowicka, B.; Gieleciak, R.; Bak, A.; Polanski, J.; Chilmonczyk, Z. The 3D QSAR study of hypolipidemic asarones by comparative molecular surface analysis. *Bioorg. Med. Chem.* submitted.

CI0501488