Modeling Robust QSAR. 2. Iterative Variable Elimination Schemes for CoMSA: Application for Modeling Benzoic Acid pK_a Values

Rafal Gieleciak and Jaroslaw Polanski*

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

Received July 14, 2006

A variety of issues decide the efficiency of 3D QSAR methods, and their practical importance for drug design is still controversial. This refers both to the predictive ability and the possibility for the indication of these areas within 3D molecular representations that are responsible for biological or chemical effects. Technically, the latter comes down to the selection or elimination of the reliable variables during 3D QSAR modeling using the Partial Least-Squares (PLS) method. In this paper we used a series of benzoic acids to test the dependence between the predictive ability and variable selection performance of PLS with Iterative Variable Elimination (IVE-PLS) in the Comparative Molecular Surface Analysis (CoMSA) modeling of Hammett constant which correlates with the pK_a values. Modeling this chemical effect allowed us to select the IVE-PLS variant that plots the contour maps indicating a carboxylic function, i.e., the region including the dissociation reaction center that determines the respective pK_a values. In fact, it appeared that a novel robust IVE version is capable of the indication of the proper contour plots independent of the method used for the calculation of partial atomic charges (AM1 or Gasteiger−Marsili).

INTRODUCTION

Basically, 3D QSAR in its classical form, e.g., CoMFA and related approaches, has been developed for the simplified modeling of the interactions of drug molecules with biological receptors. A basic assumption of such an approach is that we can reconstruct and explain biological effects on the basis of the series of the receptor ligands and the activity data alone without the detailed description of the receptor itself. Since the interactions take place somewhere outside the molecules, CoMFA constructs a molecular field grid to calculate the potential interactions in the molecular neighborhood. Next, the calculated values are processed by the PLS analysis in an attempt to find the best relationship between the activity and potentials at the grid points. Accordingly, the volume areas indicated by CoMFA as these contributing largely to the modeled activity can be considered as pharmacophoric-like elements.

However, we can also explain 3D QSAR as the comparison scheme that enables us to compare a series of molecules in an attempt to identify similarity between the molecules investigated. Bonding atoms into molecules implies atomic interactions, which means that a certain molecular atom or moiety influences the remaining system. This also means that we can recognize its appearance within the different part of the molecule. As a consequence, molecular descriptors applied in 3D QSAR can be highly intercorrelated. This effect can apply to each molecular moiety and molecular grid location used for the construction of the CoMFA field. Therefore, it is not quite clear if we can extract from the 3D QSAR data these variables that are at the origin of the activity and not those that intercorrelate but do not map a real pharmacophore. Doweyko observed that in 3D-QSAR, in particular in CoMFA, different models provide similar statistical performance.[1,2] In fact, it has been observed that extracting reliable variables in PLS poses important modeling problems for collinear variables.[3] Similar effect can also be indicated in classical QSAR where different parameters can be used for the model construction. This of course brings a problem of model interpretation.

Modeling chemical reactivity can be an interesting 3D QSAR variant that allows for distinguishing between the above-mentioned viewpoints. Although usually we cannot indicate any pharmacophore decided by the molecular receptor-like environment that is responsible for the chemical reaction, it is somewhere near the reaction center where we would expect 3D QSAR analysis to focus. Although 3D QSAR schemes have been used relatively rarely for modeling chemical reactions, at least several such examples have been reported.[4,5] Modeling pK_a values of organic acids or bases are probably the most often investigated example of such studies.[6−8]

In the current work we used the Iterative Variable Elimination method[9] as well as its alternative robust versions for the selection of such areas that provides the best 3D QSAR models. In an attempt to evaluate modeling performance both in the context of predictive ability and the selection of the pharmacophoric-like sites, i.e., volume areas involved in the reaction, we model the Hammett constants that correlate to the pK_a values of a series of benzoic acids that has been investigated previously.[6] Modeling chemical reactivity allowed us to compare such sites that are indicated by the 3D QSAR method with those expected by the chemical common sense.

Shape analysis is a powerful tool in chemistry and drug design. Therefore, for modeling 3D QSAR we used the Comparative Molecular Surface Analysis (CoMSA) a method that enables the comparison of the surfaces of the series of

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

10.1021/ci600295z CCC: $37.00
Published on Web 01/11/2007

Table 1. Hammett Constants for Substituted Benzoic Acid

	substituent	Hammett constants		substituent	Hammett constants
		Training Set			
1	H	0.00	**26**	4-Br	0.23
2	3-Br	0.39	**27**	4-CF_3	0.54
3	3-CF_3	0.54	**28**	4-CH_3	−0.17
4	3-CH_3	−0.07	**29**	4-Cl	0.23
5	3-Cl	0.37	**30**	4-CN	0.66
6	3-CN	0.56	**31**	4-F	0.06
7	3-F	0.34	**32**	4-I	0.18
8	3-I	0.35	**33**	4-NH_2	−0.66
9	3-NH_2	−0.16	**34**	4-NO_2	0.78
10	3-NO_2	0.71	**35**	4-OCF_3	0.35
11	3-OCF_3	0.38	**36**	4-OH	−0.37
12	3-OH	0.12	**37**	4-OCH_3	−0.27
13	3-OCH_3	0.12	**38**	4-SH	0.15
14	3-SH	0.25	**39**	4-SCH_3	0.00
15	3-SCH_3	0.15	**40**	4-SCF_3	0.50
16	3-SCF_3	0.40	**41**	4-$C(CH_3)_3$	−0.20
17	3-$C(CH_3)_3$	−0.10	**42**	4-C_2F_5	0.52
18	3-C_2F_5	0.47	**43**	4-CH_2Br	0.14
19	3-CH_2Br	0.12	**44**	4-CH_2Cl	0.12
20	3-CH_2Cl	0.11	**45**	4-CH_2I	0.11
21	3-CH_2I	0.10	**46**	4-C_2H_5	−0.15
22	3-C_2H_5	−0.07	**47**	4-SO_2CF_3	0.93
23	3-SO_2CF_3	0.79	**48**	4-SO_2F	0.91
24	3-SO_2F	0.80	**49**	4-SO_2CH_3	0.72
25	3-SO_2CH_3	0.60			
		Test Set			
50	3-$CH{=}CH_2$	0.05	**62**	4-$CH{=}CH_2$	−0.02
51	3-CH_2CN	0.16	**63**	4-CH_2CN	0.01
52	3-CHO	0.35	**64**	4-CHO	0.42
53	3-CH_2OCH_3	0.02	**65**	4-CH_2OCH_3	0.03
54	3-$COCH_3$	0.38	**66**	4-$COCH_3$	0.50
55	3-$CONH_2$	0.28	**67**	4-$CONH_2$	0.36
56	3-NCS	0.48	**68**	4-NCS	0.38
57	3-$NHCH_3$	−0.30	**69**	4-$NHCH_3$	−0.84
58	3-$N(CH_3)_2$	−0.15	**70**	4-$N(CH_3)_2$	−0.83
59	3-$OCOCH_3$	0.39	**71**	4-SCN	0.52
60	3-SCN	0.41	**72**	4-SO_2NH_2	0.57
61	3-SO_2NH_2	0.46			

molecules.[2,10] Several different CoMSA protocols have been published, recently in which the molecular surface zones can be ordered by the SOM neural network[10,11,12] or rectangular sector grid.[13] Technically, PLS protocol is used then to build a final analytical QSAR model.

EXPERIMENTAL

Model Builders. All the experimental data, i.e., **1**−**72**, were extracted from ref 6 and are given in Table 1. All the molecules were superimposed before the calculation of the molecular surfaces. The superimposition was performed by covering all non-hydrogen atoms of benzoic acid. We used the Match3D program for performing this operation.[14] The calculation of the molecular surface descriptors was based on CoMSA according to the procedure detailed elsewhere.[10,15]

CoMSA Method. The competitive Kohonen strategy was used to construct a 2D topographic map from the signals of points sampled randomly at the molecular surface. A projection of the electrostatic potential value (MEP) from the surface points into such a 2D arrangement of neurons, after calculating the average MEP value within this particular neuron results in the so-called feature map. A feature map illustrates the (MEP) of a single molecule (template). As the weights of the Kohonen network contain information on the shape of the particular molecular surface, the network can be used to compare molecular surface geometries of other molecules. The coordinates of other molecule(s) can be sent to such a network, and surface vectors, e.g., the electrostatic potential, can be projected on this network. The resulting comparative feature map is a kind of superimposition of the molecule and the template.

The competitive training of the network was based on the rule that each point, s, of the molecular surface was projected into that neuron, sc, that has weights, w_{ci}, that come closest to the Cartesian coordinates, x_{si}, of this point, s (eq 1).

$$\mathrm{out}_{\mathrm{sc}} \leftarrow \min\left[\sum_{i=1}^{3} (x_{si} - w_{ji})^2\right] \qquad (1)$$

The size of the networks amounts to *20*x*20*, and the resulting feature maps were transformed to respective 20^2 element vectors.

PLS Analysis. Vectors obtained were processed by the PLS analysis with a leave-one-out cross-validation procedure.[16] The PLS procedures were programmed within the MATLAB environment (MATLAB).

Molecules **1**−**72** were divided into two sets, molecules **1**−**49** form a series used for modeling PLS, and **50**−**72** were used only as a test set for the evaluation of this model. The PLS model was constructed for the centered data, and its complexity was estimated based on 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 based on the model with one, two, three, etc. factors. The Root-Mean-Square Error of CV additionally corrected for the number of PLS components is defined as

$$\mathrm{RMSECV}_j = \sqrt{\frac{\sum (\mathrm{obs} - \mathrm{pred}_j)^2}{m - j - 1}} \qquad (2)$$

where obs denotes the assayed value; pred denotes the predicted value of dependent variable; m is the number of objects; and j refers to the numbers of PLS factors, which ranges from 1 to $A_{\max}$. 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_{cv}

$$q^2_{\mathrm{cv}} = 1 - \frac{\sum (\mathrm{obs} - \mathrm{pred})^2}{\sum (\mathrm{obs} - \mathrm{mean(obs)})^2} \qquad (3)$$

where obs is the assayed values; pred is the predicted values; and mean is the mean value of obs.

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

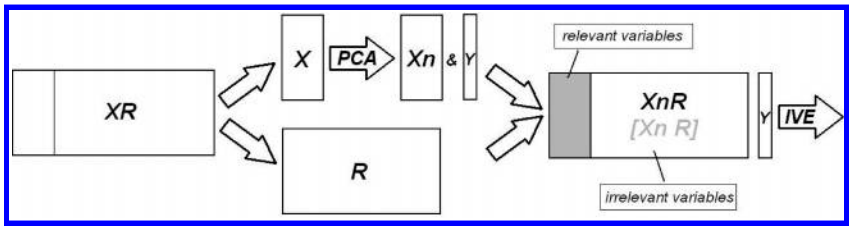

Figure 1. The schematic illustration explaining the construction of the simulated data, details in text.

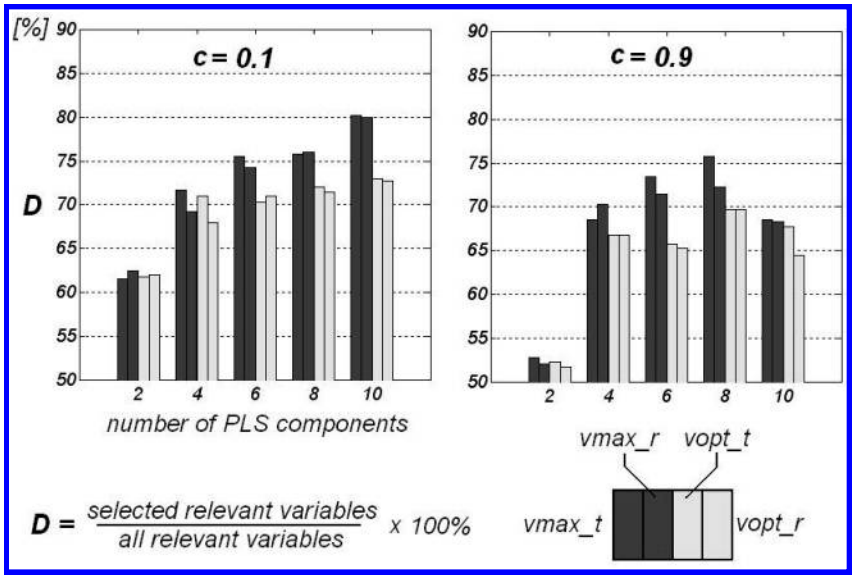

Figure 2. The variable elimination performance given by the number of the reliable variable surviving the IVE processing to the total number of variable. Details in text.

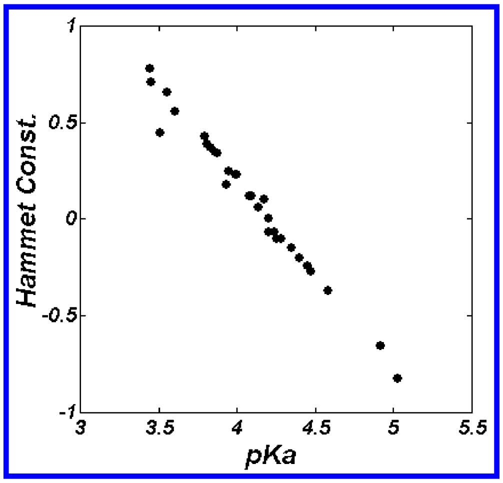

Figure 3. The dependence between the pK_a value and Hammett constant for the benzoic acid series according to data available in ref 28.

The quality of external predictions was measured by the $q^2{}_{\text{test}}$ parameter

$$q^2{}_{\text{test}} = 1 - \frac{\sum (\text{obs}_{\text{test}} - \text{pred}_{\text{test}})^2}{\sum (\text{obs}_{\text{test}} - \text{mean}(\text{obs}_{\text{test}}))^2} \qquad (4)$$

where $\text{pred}_{\text{test}}$ is the predicted value for external test set object; and obs_{test} is the observed value for external test set object.

Data Elimination. In order to indicate these parts on the molecular surface that contribute mostly to the activity we used the modified procedure of the PLS with Uninformative Variable Elimination (UVE-PLS), namely the Iterative Variable Elimination PLS (IVE-PLS) procedure.[9] The UVE algorithm was originally proposed by Centner et al.[17] as a possible improvement of the PLS procedure. The main idea of UVE-PLS is to reduce the number of the redundant variables included in the final model. The UVE algorithm based on the analysis of the regression coefficients calculated by the PLS method. The PLS method allows presenting the relation between the $\boldsymbol{Y}_{(m,1)}$ answer and the $\mathbf{X}_{(m,n)}$ predictors in a form of

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{b} + \boldsymbol{e} \qquad (5)$$

where $\boldsymbol{b}_{(n,1)}$ is a vector of the regression coefficients; and e is the vector of the errors.

Thus, the UVE algorithm analyzes a value of ($\boldsymbol{t}_{(1,n)}$) called stability that is calculated on the basis of the $\boldsymbol{b}_{(n,1)}{}^T$ coefficients of the PLS eq 5. The $\boldsymbol{t}_{(1,n)}$ score for the variables is given by eq 6

$$t = \frac{\text{mean}(\mathbf{B})}{\text{std}(\mathbf{B})} \qquad (6)$$

where $\mathbf{B}_{(m,n)}$ is a matrix of b coefficients obtained during the leave-one-out cross-validation procedure.

Then, only the variables of the "relative" high $\boldsymbol{t}$-value are included in the final PLS model. In order to estimate the cutoff level, the artificial random number noise is created (the level of the noise is 10^{-10} of the original variable order) and put (as additional columns) into the matrix of the original variables. PLS analysis of such a matrix is performed, and

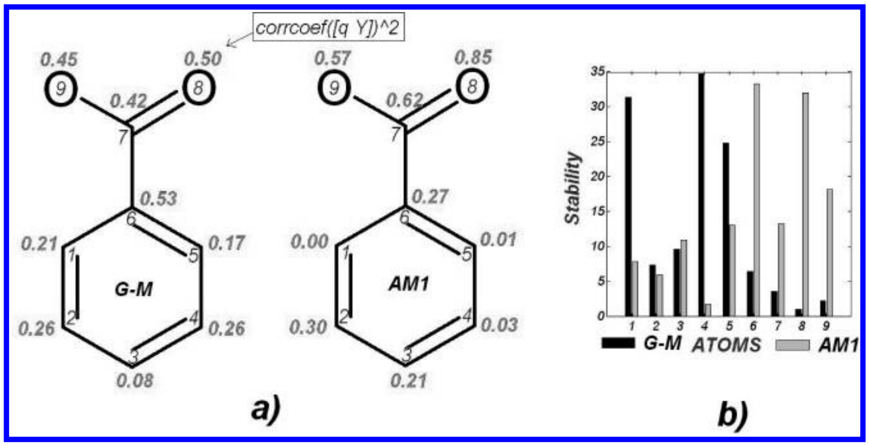

Figure 4. Correlation between partial atomic charges and Hammett constant for the benzene ring and carboxylic function atoms, as calculated by GM and AM1 (a) methods. The respective values of stability for PLS modeling are shown in (b).

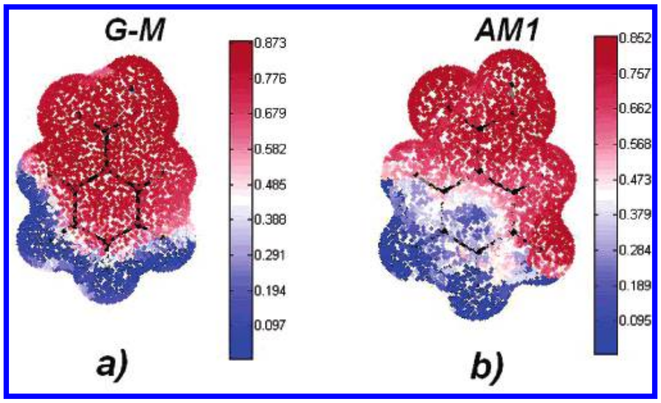

Figure 5. Correlation coefficients between the potential values at the molecular surface and the Hammett constant for the partial atomic charges calculated by GM (a) and AM1 (b) methods, respectively.

the ***t*** parameter is analyzed for each column. The highest absolute value, abs(***t***), for the noisy column determines the cutoff level for the original variables.

Iterative Variable Elimination, IVE1 (IVE-max). We described the IVE-PLS method in our previous publications.[9] Instead, of a single-step procedure we used here an iterative algorithm based on the abs(***t***) criterion to find variables to be eliminated. In order to distinguish this procedure, we named this Modified Uninformative Variable Elimination with the iterative leave-one-out cross-validation (IVE-PLS). In the current publication this procedure is labeled IVE1. This includes the following:

1. PLS modeling with leave-one-out cross-validation is performed for the $\mathbf{X}_{(m,n)}$ matrix with the in advance fixed number of PLS components ($A_{\max}$). This gives matrix $\mathbf{B}_{(m,n)}$ constructed of the m vectors of coefficients $\boldsymbol{b}_{(n,1)}{}^{T}$, as given by eq 7[16]

$$\boldsymbol{b}_{(n,1)} = \mathbf{W}_{(n,A_{\max})}{*}(\mathbf{P}_{(n,A_{\max})}{}^{T}{*}\mathbf{W}_{(n,A_{\max})})^{-1}{*}\mathbf{Q}_{(A_{\max},1)} \quad (7)$$

where $\mathbf{W} - \mathbf{X}$ is the weight matrix; $\mathbf{P}$ is the loading matrix; and $\mathbf{Q} - \boldsymbol{Y}$ is the weight matrix.

2. Determination of the stability parameter $\boldsymbol{t}_{(1,n)}$ based on matrix $\mathbf{B}_{(m,n)}$, calculated by eq 6.

3. The elimination from matrix $\mathbf{X}_{(m,n)}$ a column of the lowest abs(***t***) value.

4. Standard PLS analysis of the new matrix $\mathbf{X}_{(m,n-1)}$ without the column eliminated in the step 3.

5. Iterative repetition of steps 1−4 to maximize the q^2_{cv} parameter.

Although the stability value in IVE1 is calculated for a fixed number of latent variables, the final model (step 4) is always corrected to the optimal complexity.

Below we describe modified IVE procedures, namely, IVE2−IVE4.

IVE2 (IVE-max-robust). The difference to IVE1 is that criterion given by eq 6 takes a form of eq 8

$$t(\text{robust}) = \frac{\text{median}(\mathbf{B})}{\text{iqr}(\mathbf{B})} \quad (8)$$

where iqr is the interquartile range.

IVE3 (IVE-opt). In this version during the IVE procedure the optimal number of PLS components (Aopt) is always precisely estimated just during the calculation of the original PLS model (given b coefficients) according to eq 9

$$\boldsymbol{b}_{(n,1)} = \mathbf{W}_{(n,\text{Aopt})}{*}(\mathbf{P}_{(n,\text{Aopt})}{}^{T}{*}\mathbf{W}_{(n,\text{Aopt})})^{-1}{*}\mathbf{Q}_{(\text{Aopt},1)} \quad (9)$$

where $\mathbf{W} - \mathbf{X}$ is the weight matrix; $\mathbf{P}$ is the loading matrix; and $\mathbf{Q} - \boldsymbol{Y}$ is the weight matrix.

IVE4 (IVE-opt-robust). IVE4 is a subversion of IVE3 but uses the robust stability criterion t(robust) calculated by eq 8.

The meaning of $A_{\max}$ parameter indicated for all methods is the depth of the PLS component space to which modeling is tested. At this level modeling is truncated even if the optimal number exceeds the $A_{\max}$ value.

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

Testing IVE1−IVE4 Procedures. The IVE1−IVE4 procedures were tested using the artificial simulated data set and Selwood benchmark series, as described below.

Artificial data set generation is explained in Figure 1.

1. Matrix **XR** with dimensionality **(20,100)** was generated from multivariate normal distribution with zero mean vector and the assumed covariance matrix **C**. The elements of matrix **C** were chosen as given by eq 10. This procedure follows

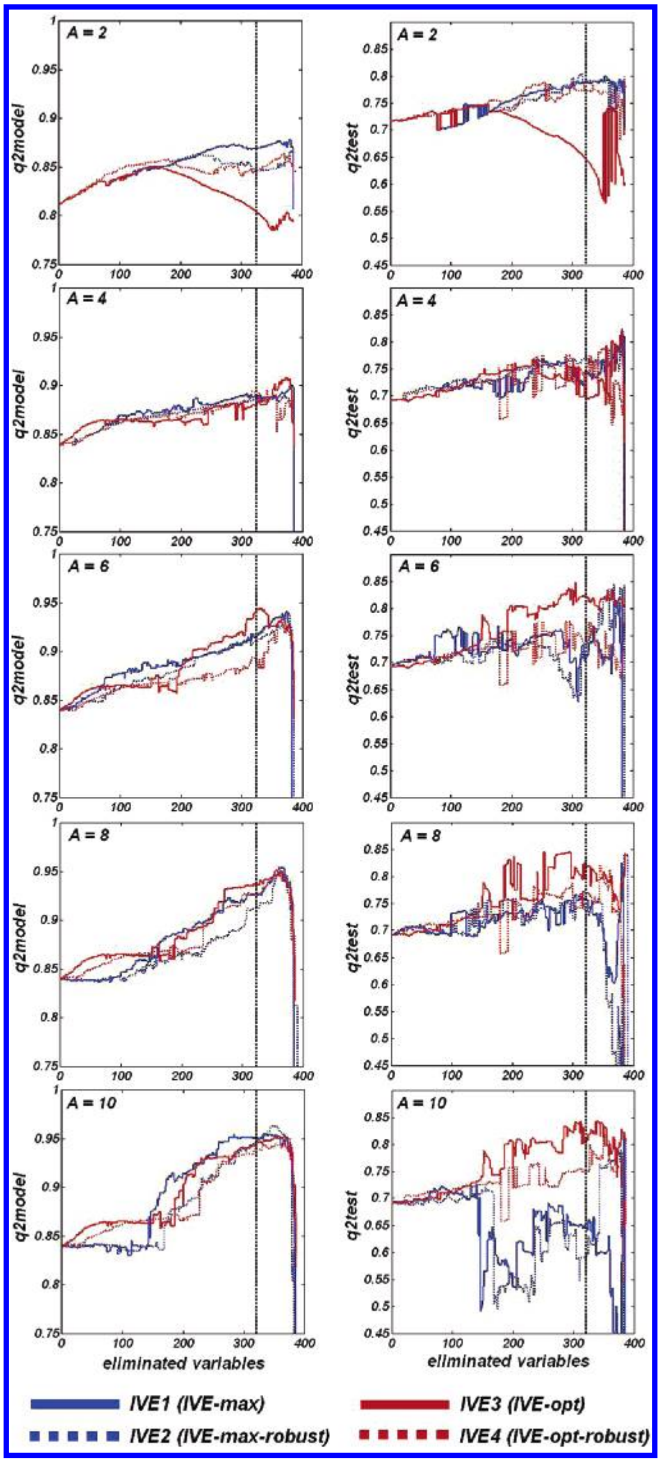

Figure 6. The performance of the PLS CoMSA (GM charges) modeling coupled with the IVE1−IVE4 tested by q^2_{test} and q^2_{model} performance measures, details in text. Vertical line indicates a variable range that appears as red dots in Figure 8.

the protocol reported in ref 3. In this way we simulated the correlation appearing between 3D QSAR descriptors

$$C_{ij} = c^{|i-j|} \tag{10}$$

where c is a fixed level of correlations between predictors; and i is a row and j is a column of matrix **C**.

We generated two families of matrix **C** with the values of c amounting to 0.1 and 0.9, respectively.

2. The matrix **XR** was divided randomly in two matrices **X(20,20)** and **R(20,80).**

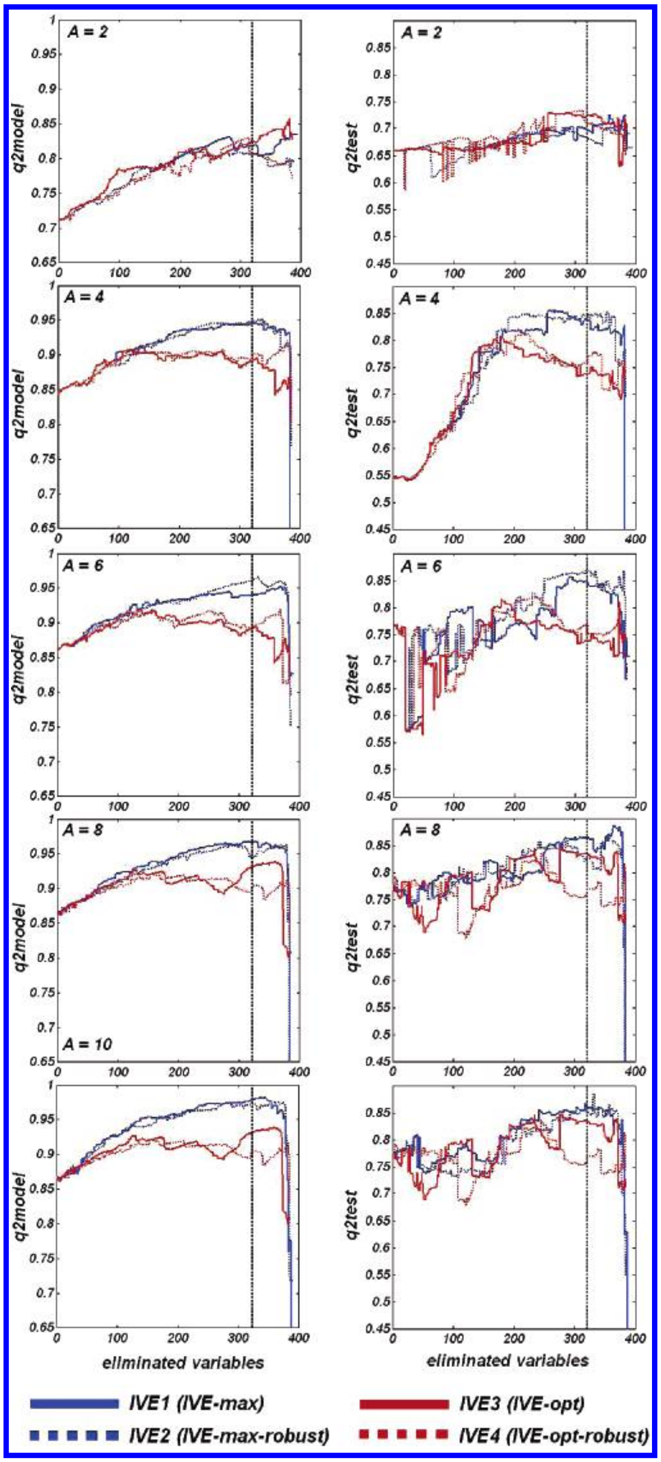

Figure 7. The performance of the PLS CoMSA (AM1 charges) modeling coupled with the IVE1−IVE4 tested by q^2_{test} and q^2_{model} performance measures, details in text. Vertical line indicates a variable range that appears as red dots in Figure 9.

3. PCA on the matrix **X(20,20)** is performed, yielding scores and loadings.

4. The multiplication of the first five score vectors **(20,5)** by the first five loading vectors **(5,20)** gives a simulated denoised data matrix **Xn(20,20)**.

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

6. New matrix **XnR(20,100)** incorporates the relevant noisefree variables **Xn** and irrelevant variables **R(20,80)**. We

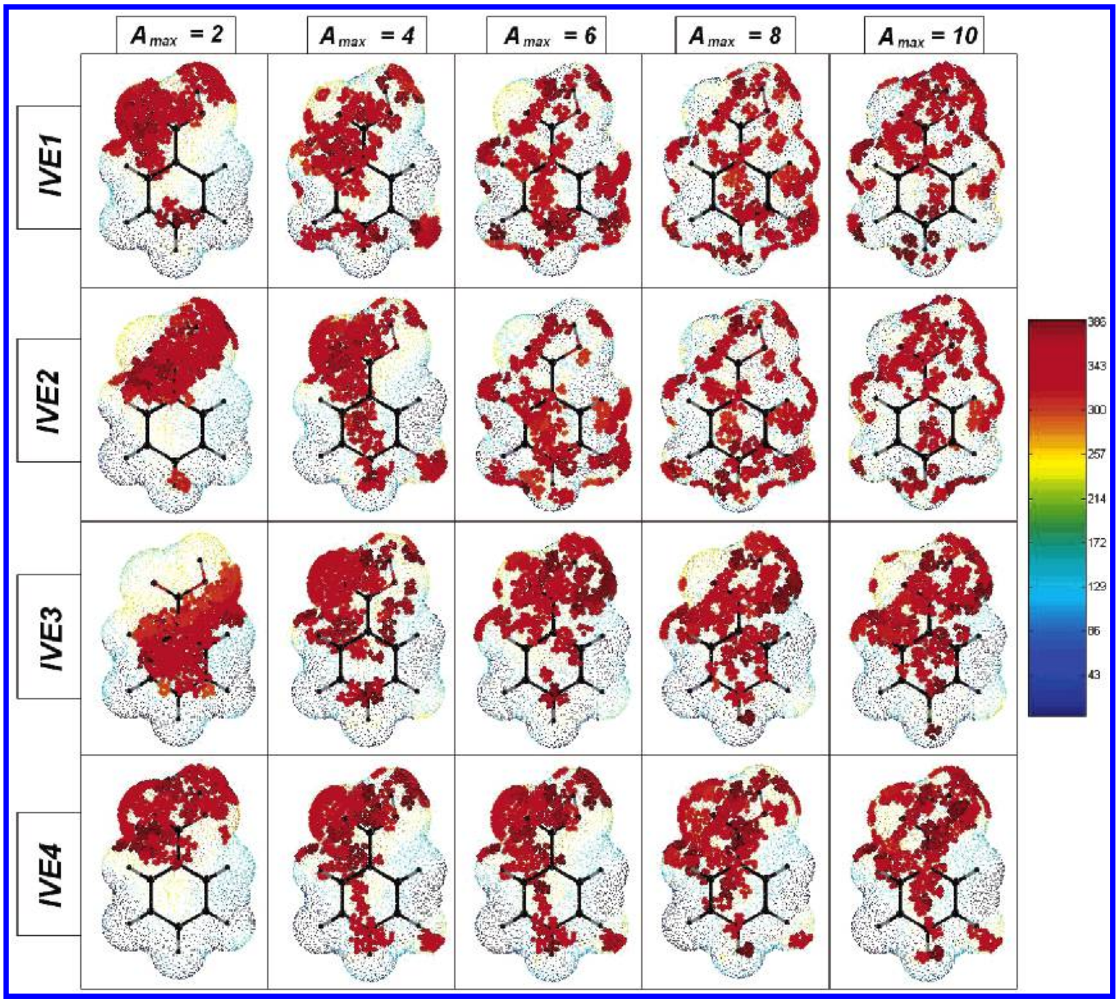

Figure 8. Interaction contour plots revealed by IVE1−IVE4 for the different model complexity, using the GM charges. Colors indicate the order of eliminated variables, where blue means the early eliminated variables and maroon means these that tend to survive the process.

must remember that correlation between variables in matrix **XnR** can be fixed on level 0.1 or 0.9.

All four variable selection (IVE1−IVE4) procedures were performed in order to extract the columns that include information needed to calculate the ***Y*** answers. The results obtained are shown in Figure 2 which plots the percentage of the rate of the relevantly selected variables to all relevant variables in the IVE1−IVE4 procedures for 100 replicated simulations. Although all the methods provide similar results, the IVE-max versions generally provides slightly better D ratios.

SELWOOD Test. Additionally, we used the Selwood data set[18] consisting of 31 compounds and 53 descriptors for testing the IVE versions. This data set has become a literature benchmark for evaluating variable selection procedures, e.g., refs 19 and 20. The results indicate that in this particular case IVE1 and IVE2 provides the results which resemble literature selections more closely than IVE3 and IVE4. The detailed plots are shown in the Supporting Information.

RESULTS AND DISCUSSION

An analysis of the 3D QSAR literature may suggest that predicting the activity of novel compounds is a basic objective of this method. This is not true even though the so-called predictive ability is a basic parameter calculated in QSAR modeling. However, current QSAR concentrates on modeling relationships in a factual chemical compounds space. It means that both modeling QSAR relationships and testing predictive ability or validating this is performed by the application of the compounds of the known activity that were synthesized apriori. Thus, QSAR predictability does not take into account novel compounds (virtual chemical space) that are to be synthesized in the next step of the investigation and then how to define the basic objective of 3D QSAR strategy. In such a strategy we extend the selected atom property (usually partial charges) outside the atoms or even molecules investigating the molecular surface or fields. The information on the relative orientation of the molecules should support the original atom property data type. In order to better understand a strategy of the grid methods we can cite here a number of other methods that use atom property data type to evaluate molecular similarity and consequently to predict the activity without the construction of the molecular field. These are for example 4-D QSAR,[21,22] Carbo indexes,[23] RLNN,[24] etc. Technically, the calculation of the molecular field or surface property significantly increases the amount of variables to be included into the analysis, which complicates also the modeling procedure. This allows us, however, to indicate the molecular field volumes that are apparently involved in the drug receptor interactions. Accordingly, the basic goal of the molecular field type modeling is to find the grid points denoting the variables

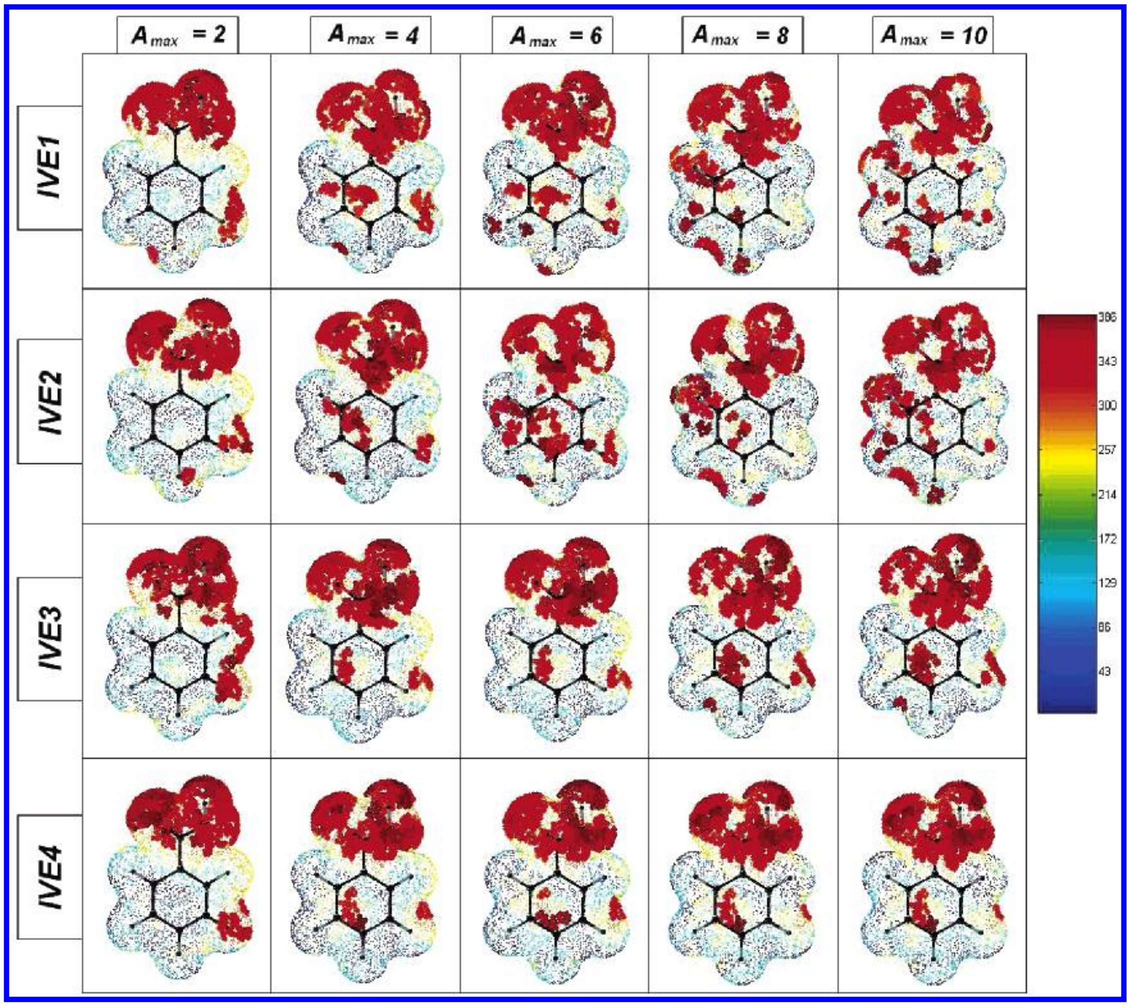

Figure 9. Interaction contour plots revealed by IVE1−IVE4 for the different model complexity, using the AM1 charges. Colors indicate the order of eliminated variables, where blue means the early eliminated variables and maroon means these that tend to survive the process.

that contribute mostly to the activity. Cross-validation in multidimensional QSAR or model testing using the external molecular set of the known activity is performed not in the direct search for new molecules, i.e., for molecular design, but for model validation. In standard CoMFA the interaction surfaces are determined by filtering regression weights that decide a contribution of original variables (potential values in the grid points) into a final PLS model. The points of the highest standard deviation for the whole molecular series form the space sectors of positive and negative influence for the activity. Although in CoMFA the results of PLS analysis are given in a form of regression, they cannot be interpreted directly because of the large number of variables. Instead, the generation of the contour maps that illustrates the molecular field regions and volumes contributing negatively or positively into the activity is a standard way in which we visualize the results.[25] Then, we can compare the contour plots for the individual molecules trying to find molecular features common for the high or low activity compounds, i.e., these that increase or decrease the activity.[26]

Figure 3 plots the Hammett constants vs the pK_a values for the series of *meta* and *para* substituted benzoic acids. As expected this plot proves a nice correlation. The colinearity shown in Figure 3 makes the identity between these problems, and we report further the results for modeling Hammett constants to be fully comparable to the results described previously in the literature.[7]

Figure 4 illustrates the results of PLS modeling of the Hammett constant values using the atomic type data, i.e., the partial charges calculated by the Gasteiger−Marsili (G−M) or AM1 method. PLS models calculated for such data provides similar predictive power for both methods, and at the same time AM1 gives slightly better results. Thus, G−M modeling provides $q^2_{\text{model}} = 0.80$ (6); $q^2_{\text{test}} = 0.72$ vs AM1 - $q^2_{\text{model}} = 0.86$ (5); $q^2_{\text{test}}= 0.77$. This compares well to the q^2_{model} of 0.89 (6) reported for the series in ref 27 using the AM1 derived charges. The reported data indicate that PLS modeling using the atomic representation alone gives models of the comparable predictive ability to the CoMFA one.

Although a carboxylic function is evidently a site distinguished by high correlation to the Hammett constant as shown in Figure 4a, the AM1 method seems to be slightly more selective. However, when we modeled an analytical PLS equation for the whole atomic molecular representation and compared the individual atomic contribution by the stability value as shown in Figure 4b, it was clear that both methods indicate completely different atoms as these contributing mostly to the modeled values. Now, it is the AM1 method that clearly tells apart the COOH group. Since the carboxylic function is a place where ionization reaction of benzoic acids takes place, it seems that AM1 information is more relevant for the real process investigated.

Modeling pK_a values or Hammett constant if related to the molecular structure of the acid illustrates nicely a problem of

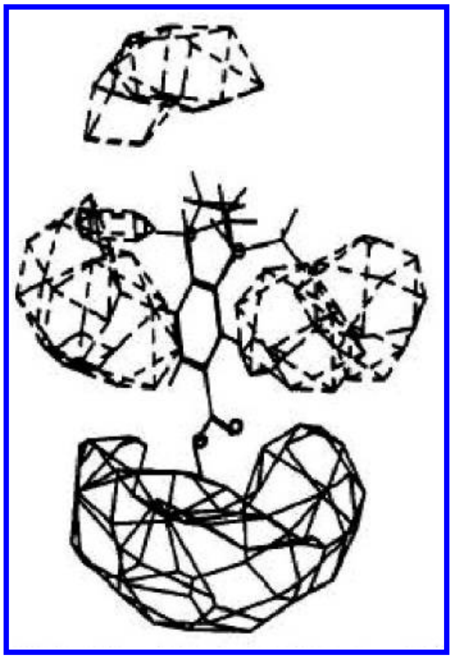

Figure 10. Interaction contour plots revealed for the same series by CoMFA. Modified after 6 ACS.

variable intercorrelation. The interactions of the substituents located at the different molecular area focus at the carboxylic function which can be observed in the regulation of the dissociation constant of the respective acid. The substituents can also affect other molecular moieties correlating with the activity analyzed. It would be interesting to observe the correlation coefficient between electrostatic potential of the molecular surface and the Hammett constant. This is shown in Figure 5 both for the G−M and AM1 derived atomic charges. A large red colored molecular surface area indicates the location of the variables calculated for the surface points that are highly correlated with the Hammett constant. This illustrates better the complex nature of the problem of PLS modeling aimed at the selection of the proper molecular surface points under variable multicolinearity.

Figures 6 and 7 report the performance during the investigated CoMSA modeling using the different variable elimination procedures. Unlike, the atomic data involving only 9 variables, the CoMSA data include more than 3500 surface points compressed to 400 neurons by the SOM neural network. We used PLS with four different IVE protocols to find a final model and illustrate the pharmacophoric-like sites. For the G−M calculated atomic charge variable elimination curves (Figure 6) proceed similarly for all four procedures IVE1−IVE4 within the same component numbers A_{max}, where $A_{\mathrm{max}} = 2$ is the only exception. In the latter case IVE3 provides maximal values for the lower number of eliminated variables. For the AM1 method the curves of IVE1/IVE2 and IVE3/IVE4 proceed slightly different, namely, IVE1/IVE2 after the first 100 eliminated variables usually outperforms IVE3/IVE4. It is worth mentioning that testing modeling performance for the test set by q^2_{test} proves the predictive power of the models. Although the analysis of the test set predictions reveals higher instability (indicated by q^2_{test} oscillations) in comparison to those provided for the modeling set, q^2_{test} is also well above a value of 0.5 indicating the predictive rate.

It is interesting to observe the molecular areas surviving variable elimination, as indicated by the different IVE methods. This is illustrated in Figures 8 and 9, respectively. The red color surface points are included in the last ca. 80 neurons eliminated, which corresponds approximately to the vertical line indicated in Figures 6 and 7. It can be observed that the robust IVE4 is the method pointing more clearly for the carboxylic function as the pharmacophoric-like area although the G−M method that fails to provide similar plots for the higher model complexity IVE4 for the lower complexity (A_{max}=2) evidently points for the carboxylic function. The robust IVE4 provides also the most reliable model for the AM1 type charges shown in Figure 9. This is preserved also for the high complexity of $A = 10$ that is additionally supported by the high modeling and testing q^2 performance values. The respective contour CoMSA plots

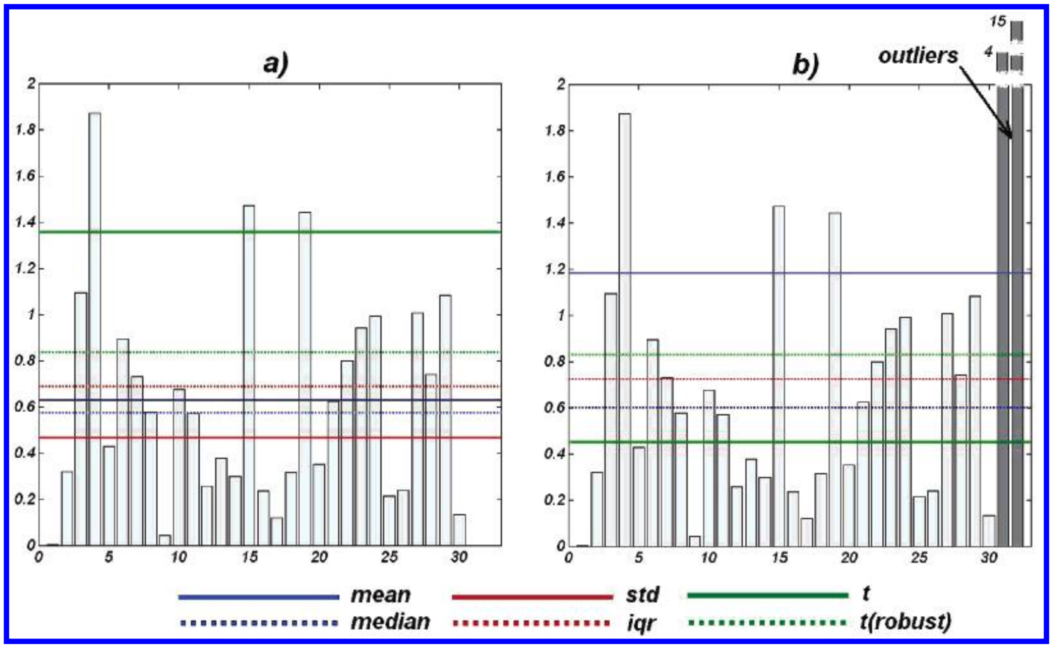

Figure 11. The schematic illustration comparing the performance of the traditional (a) and robust (b) stability parameter, ***t***, respectively, carried on the simulated data, details in text.

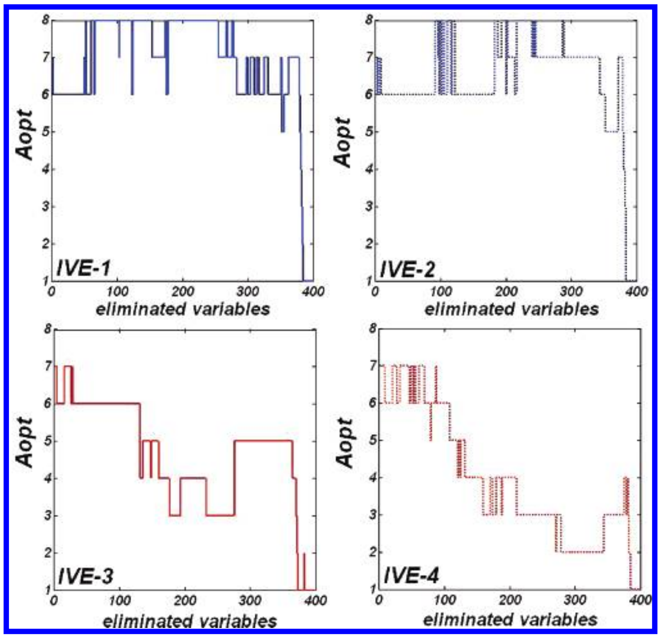

Figure 12. An optimal number of the latant PLS components Aopt (model complexity) during variable elimination using the IVE1-IVE4-PLS protocols (AM1 charges).

compare advantageously to the CoMFA contour maps (calculated for the AM1 partial charges) shown in Figure 10 reproduced after ref 6 which indicates the large molecular area other than the carboxylic function. In the Supporting Information, we reported additionally the results obtained for the PM3 partial charge data. This indicates that the results obtained by the AM1 method compares advantageously also to the PM3 method.

A question arises as to what decides the efficiency of IVE1/IVE4 protocols. Individual protocols differ in the way in which we calculate the stability parameter ***t*** by criterion (6). Thus, in Figure 11 we compare a value of ***t*** yielded by criterion (6) for two different sample inputs. A stable homogeneous input illustrated in Figure 11a is disturbed in Figure 11b by two very different signals. It can be compared that unlike the standard stability ***t*** calculated after criterion (6) including the *mean* value, which significantly differs in parts a and b in Figure 11, the robust stability ***t*(robust)** value remains unaffected by the outlier noise. In Figure 12 we can observe the decreasing model complexity with the decreasing number of variables included in the IVE-PLS model for the robust IVE versions (IVE3 and IVE4). Although we illustrated only the plots resulting for the AM1 models, this appeared true for the all tested models (G−M, PM3). Thus, the robust stability ***t*(robust)** allows us to denoise the IVE-PLS model. It is also interesting to compare two robust protocols IVE3 and IVE-4 by themselves. If so it appears that the IVE-4 protocol that carefully monitors the exact number of complexity provided also the best results.

CONCLUSIONS

A variety of issues decides the efficiency of QSAR methods, and their practical importance for drug design is still controversial. The possibility of the extraction of the reliable contour maps that points for the pharmacophoric-like sites is an important factor that limits the efficiency of QSAR. In this paper we used a series of benzoic acids to test the dependence between the predictive and variable selection performance of PLS in the Comparative Molecular Surface Analysis (CoMSA) modeling of the Hammett constant which correlates with the acidic pK_a values. Modeling this chemical effect allowed us to select the Iterative Variable Elimination (IVE) variant that indicates contour maps located near carboxylic function, i.e., the region including the dissociation reaction center determining the respective pK_a values. In fact, it appeared that the robust IVE version is capable of the indication of the proper contour plots independent of the method used for the calculation of partial atomic charges (AM1 or Gasteiger−Marsili). The comparison of the performance of the individual IVE1-IVE4-PLS protocols allowed us to reveal that the robust stability ***t*** applied in IVE3 and IVE4 works as the denoising tool. In consequence, the PLS model complexity during robust variable elimination decreases steadily with the number of variables eliminated. Among robust protocols IVE3 and IVE4, the IVE-4 one that carefully monitors the exact number of complexity provided also the best results.

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 grants no. 3 T09A 011 27 is gratefully acknowledged. R.G. thanks the Foundation for Polish Science for an individual grant.

Supporting Information Available: Performance of IVE1-4 variable elimination for the Selwood data set and the data on CoMSA modeling the Hammett constant for the benzoic acid series using the PM3 method. This information is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Doweyko, A. 3D-QSAR Illusions. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 587−596.

(2) Polanski, J.; Bak, A.; Gieleciak, R.; Magdziarz, T. Modeling Robust QSAR. *J. Chem. Inf. Model.* **2006**, *46*, 2310−2318.

(3) Chong, I.-G.; Jun, Ch-H. Performance of Some Variable Selection Methods When Multicollinearity is Present. *Chemom. Intell. Lab. Syst.* **2005**, *78*, 103−112.

(4) Vrtacnik, M.; Voda, K. HQSAR and CoMFA Approaches in Predicting Reactivity of Halogenated Compounds with Hydroxyl Radicals. *Chemosphere* **2003**, *52*, 1689−1699.

(5) Lipkowitz, K. B.; Pradhan, M. Computational Studies of Chiral Catalysts: A Comparative Molecular Field Analysis of An Asymmetric Diels-Alder Reaction with Catalysts Containing Bisoxazoline or Phosphinooxazoline Ligands. *J. Org. Chem.* **2003**, *68*, 4648−4656.

(6) Kim, K. H.; Martin, Y. C. Direct Prediction of Linear Free Energy Substituent Effects from 3D Structures Using Comparative Molecular Field Analysis. 1. Electronic Effects of Substitued Benzoic Acids. *J. Org. Chem.* **1991**, *56*, 2723−2729.

(7) Hollingsworth, Ch. A.; Seybold, P. G.; Hadad, Ch. M. Substituent Effects on the Electronic Structure and pK_a of Benzoic Acid. *Int. J. Quantum. Chem.* **2002**, *90*, 1396−1403.

(8) Gross, K C.; Seybold, P. G. Comparison of Quantum Chemical Parameters and Hammett Constants in Correlating pK_a Values of Substituted Anilines. *J. Org. Chem.* **2001**, *66*, 6919−6925.

(9) Polanski, J.; Gieleciak, R. The Comparative Molecular Surface Analysis (CoMSA) with Modified Uninformative Variable Elimination-PLS (UVE-PLS) Method: Application to The Steroids Binding the Aromatase Enzym. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 656−666.

(10) Polanski, J.; Walczak, B. The Comparative Molecular Surface Analysis (CoMSA): A Novel Tool for Molecular Design. *Comput. Chem.* **2000**, *24*, 615−625.

(11) 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.

(12) 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.

(13) Polanski, J.; Gieleciak, R.; Magdziarz, T.; Bak, A. 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.

(14) Gasteiger, J. *Match3D*; Molecular Networks GmbH. http://www.molecular-networks.com/software/overview/index.html (accessed Nov 13, 2006).

(15) 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.

(16) Wold, S.; Sjöström, M.; Eriksson, L. PLS-regression: A Basic Tool of Chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109−130.

(17) 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.

(18) Selwood, D. L.; Livingstone, D. J.; Comley, J. C.W.; O'Dowd, A. B.; Hudson, A. T.; Jackson, P.; Jandu, K. S.; Rose, V. S.; Stables, J. N. Structure-Activity Relationships of Antifilarial Antimycin Analogues: A Multivariate Pattern Recognition Study. *J. Med. Chem.* **1990**, *33*, 136−142.

(19) Rogers, D.; Hopfinger, A. J.; Application of Genetic Function Approximation (GFA) to Quantitative Structure-Activity Relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854−866.

(20) Kubinyi, H. Variable Selection in QSAR Studies. I. An Evolutionary Algorithm. *Quant. Struct.-Act. Relat.* **1994**, *13*, 285−294.

(21) Hopfinger, A.; Wang, S.; Tokarski, J.; Jin, B.; Albuquerque, M.; Madhav, P.; Duraiswami, C. Construction of 3D-QSAR Models Using the 4D-QSAR Analysis Formalism. *J. Am. Chem. Soc.* **1997**, *119*, 10509−10524.

(22) Polanski, J.; Bak, A. Modeling Steric and Electronic Effects in 3D and 4D-QSAR Schemes: Predicting Benzoic pK_a Values and Steroic CBG Binding Affinities. *J. Chem. Inf. Comput. Sci*. **2003**, *43*, 2081−2092.

(23) Carbo, R.; Leyda, L.; Arnau, M.; How Similar is A Molecule to Another? An Electron Density Measure of Similarity Between Two Molecular Structures. *Int. J. Quant. Chem.* **1980**, *17*, 1185−1189.

(24) Polanski, J. The Receptor-Like Neural Network for Modeling Corticosteroid and Testosterone Binding Globulins. *J. Chem. Inf. Comp. Sci.* **1997**, *37*, 553−561.

(25) Kubinyi, H. Comparative Molecular Field Analysis (CoMFA). In *Handbook of Chemoinformatics. From data to knowledge*; Gasteiger, J., Ed.; Wiley VCH: BRD, Weinheim, 2003, Vol. 4, pp 1555−1574.

(26) Cramer, R., III; Patterson, D.; Bunce, J. Comparative Molecular Field Analysis (CoMFA). *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.

(27) Martin, Y. C.; Lin, T. C.; Hetti, Ch.; DeLazzer, J. PLS Analysis of Distance Matrices to Detect Nonlinear Relationships Between Biological Potency and Molecular Properties. *J. Med. Chem.* **1995**, *38*, 3009−3015.

(28) Physical and Chemical Data Compendium. *Poradnik fizykochemiczny*; WNT: Warsaw, 1974; pp 347−351.

CI600295Z