Modeling Robust QSAR†

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

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

Received August 5, 2005

Quantitative Structure Activity Relationship (QSAR) is a term describing a variety of approaches that are of substantial interest for chemistry. This method can be defined as indirect molecular design by the iterative sampling of the chemical compounds space to optimize a certain property and thus indirectly design the molecular structure having this property. However, modeling the interactions of chemical molecules in biological systems provides highly noisy data, which make predictions a roulette risk. In this paper we briefly review the origins for this noise, particularly in multidimensional QSAR. This was classified as the data, superimposition, molecular similarity, conformational, and molecular recognition noise. We also indicated possible robust answers that can improve modeling and predictive ability of QSAR, especially the self-organizing mapping of molecular objects, in particular, the molecular surfaces, a method that was brought into chemistry by Gasteiger and Zupan.

INTRODUCTION

It may look like a paradox but *the most fundamental and lasting objective of* (chemical) *synthesis is not a production of new compounds but the production of properties*.[1] However, a problem is that property production is still more of a dream than a reality. Despite the fact that the recent decade has brought forth a number of revolutionary ideas in molecular design, e.g. combinatorial chemistry, genomics, chemogenomics, the number of newly registered drugs decreases.[2] What is the place of QSAR in a novel molecular design landscape?

Basically, QSAR should work like a dictionary between the chemical compound space and the property space. This method includes a variety of procedures that usually starts from searching mathematical models describing a biological answer in a given property space. Next, the compound space is screened in a search for the virtual molecules of the required property. Virtual objects having promising properties can be synthesized in a hope for the optimization of the molecular structure. Therefore, we can define QSAR as an indirect molecular design by the iterative sampling of the chemical compound space to optimize a certain property and thus indirectly design the molecular structure having this property. Although generally QSAR realizes a strategy from molecules to property, the so-called inverse strategy from property to molecules has also been investigated.[3]

QSAR Is Highly Data Dependent. QSAR relates chemical or biological answers to the molecular structure offering a very different level of the explanation of the mechanisms controlling real processes. The Hammett equation was a first quantitative structure−reactivity approach in chemistry. Generally, it fails however to model biological activity. Hansch developed the first successful QSAR model that describes biological responses by including a hydrophobicity (log P) term.[4] A linear equation which takes the form of $\log(1/C) = a \log(P) + C$ is the simplest function that can relate structure, or more accurately a calculable property manifested by such a structure, to activity connecting the activity, log(1/C), and property, log(P) spaces. In fact, a number of QSAR models described by a similar equation can be found in the literature. However, generally log(P) is a parameter that describes biological transport phenomena. From the theoretical point of view this implies that there is some optimal log(P) value, and a linear function cannot be used for modeling such a phenomenon. Figure 1 illustrates a data dependency problem in a situation when linear and parabolic models can describe hypothetical activity. The effect presented in Figure 1 brings a problem of the predictive credibility. Neither the linear model **a** nor the linear model **b** is capable of the proper Y* predictions in space X, for the nonlinear Y vs X relationship. Figure 1a,b shows also a fact that the extrapolation of linear models is the most dangerous for the prediction credibility. We should however keep in mind that the biological activity profile in the molecular structure landscape is *controlled by the similarity paradox,* i.e., even a minute change in the compound structure can result in a substantial activity change. This makes any QSAR prediction for **a real external object** (a virtual molecule that has not been synthesized before the modeling step) an extrapolation beyond the well-explored borders rather than an interpolation. In this context, making predictions, which seems to be a main objective of QSAR, is a roulette risk operation.

Robust Answers for QSAR Data Dependence. QSAR is not the only example of data dependency in modeling. Weather forecasts can be an illustrative example of the highly risky predictions under extremely unstable conditions originated from a large number of hard to control variables. The reduction of the sensitivity of the modeled output to the variation of inputs that should decrease data dependency and improve predictions in such conditions is called *robust modeling*. *The term robustness in signal processing applications usually refers to approaches that are not degraded*

† Dedicated to Professor Johann Gasteiger.

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

10.1021/ci050314b CCC: $33.50 © 2006 American Chemical Society
Published on Web 12/08/2005

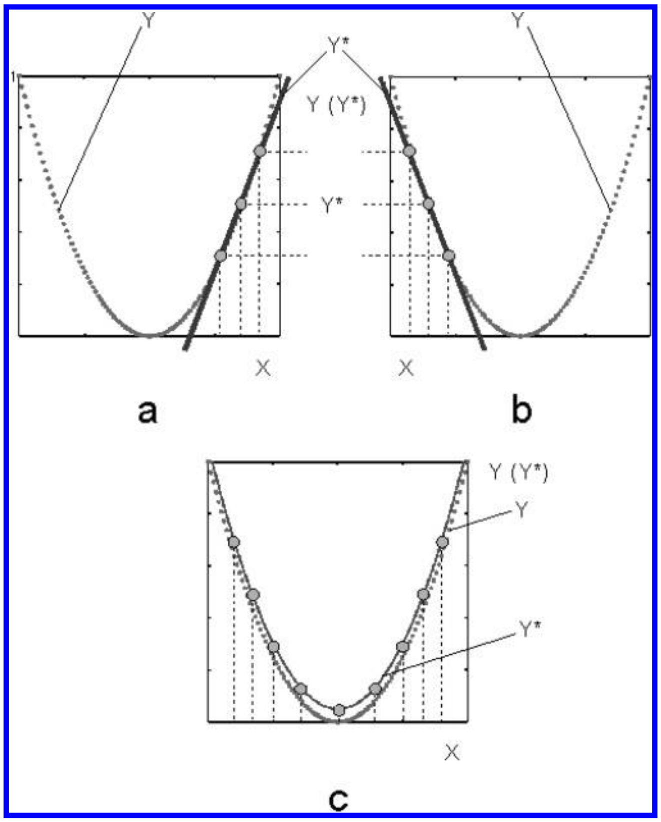

Figure 1. A schematic view of data dependency. In this particular case a difference between the modeled (Y*) and real (Y) answers are controlled by the discrepancy between a mathematical form of the equation used. If we assume that Y is given by a parabola in the X space, then a linear equation can be approximated by the assumed parabolic Y answer only in some limited X ranges (a) or (b). However, we need to include a nonlinear term to describe a parabola (c). Models (a) or (b) will be highly data dependent, and their application for the predictions of Y* will be restricted only to the certain X data range.

significantly when the assumptions that were invoked in defining the processing algorithm are no longer valid.[5]

Data dependency is well realized in QSAR, and a controversy about reliability of similar approaches in the context of predictability and practical applications among the medicinal chemistry audience is alive and well. The most robust answer would be almost completely data independent. Recently, several approaches have appeared in drug design that can be included in such a trend. Thus, the Lipinski rule of five,[6] druglikeness,[7] comparative QSAR, combined QSAR databases,[8,9] or the ADMET approach[10] are the methods that are based on the search for common features that connect all drugs. Oprea tested more than 12 000 compounds of the different biological activity types attempting to find the clusters of active, middle-activity, and inactive molecules. In a similar approach the relationship of the regression coefficients for more than 400 QSAR have been investigated.[11] Formally, in similar approaches we search for a certain property range, rather than a discrete property value, that would describe a hypothetical virtual drug space.

The *extensive data independence* implies, however, qualitative and not quantitative solutions. What are the solutions for quantitative modeling? The replacement of the linear Hansch QSAR model by the incorporation of the squared (log P)[2,4] or bilinear[12] terms indicates possible directions of the improvement of the model robustness, in this particular case, by eliminating the discrepancy between real interaction mechanism and mathematical formula we can model a single equation relating broader X range to Y. New computational methods including neural networks, data elimination, genetic algorithms, and novel model validation schemes are other examples in this field.[8,13−15] A combination of different data handling schemes as neural networks and genetic algorithms or neural networks and PLS analysis appeared to be especially effective. PLS analysis is a default method used in a number of 3D QSAR methods.[16] Moreover, modeling equations can be supported by more flexible data handling methods. Thus, for example, Support Vector Machine (SVM) is a new promising method for data classification and regression that has recently gained special attention in many fields of chemistry and medicine.[17,18] It has been introduced as a robust and highly accurate intelligent classification technique, well suited for QSAR applications.[19] Recent SVM modifications i.e., Support Vector Regression (SVR), offer a good prediction performance and can be used as the PLS replacement.[20] An interesting comparison between this technique and other QSAR tools can be found in ref 21. This method appeared especially useful in QSAR when data are not linearly separable. In this particular case, SVM can classify properly in a linear way by constructing the Optimal Separating Hyperplane (OSH) in a space of higher dimensionality. To avoid time-consuming calculations space mapping is realized implicit by so-called kernel functions.[22] Instead of minimizing the error on the training data the SVM maximizes the margin, i.e, the largest possible distance from the hyperplane to the closest objects of the two classes. The solution to the optimization problem is a global minimum, whereas other machine learning methods sometimes terminate in a local minima. In comparison to other QSAR techniques SVM gives the similar results of analysis but requires significantly smaller training times, which is increasingly important when learning large numbers of chemical compounds.[23]

Robustness and Data Noise. Usually, in QSAR data describing no more than 100 compounds are available. In traditional QSAR the activity of these molecules are related to several molecular descriptors. This situation dramatically changes in multidimensional QSAR where the number of molecular descriptors increases to thousands. Standard regression fails to deal with such data structure due to a number of cross-dependencies and chance correlations.[24] A robust answer for the data flood condition, i.e., data noise, can be Principal Component Analysis (PCA) or Partial Least Squares (PLS) methods that attempt to get better insight into the molecular descriptor space by data projection to so-called latent variables which are linear combinations of original variables. This problem will not be discussed here further, and a reader is referred to a number of monographs available.[25−27] Data noise demands special model validation schemes since fitted statistics applied for standard regression fail to provide a reliable analysis of the model quality. Cross-validation (CV) that includes several techniques such as k-fold leave-one-out (LOO), jackknife, delete-d, or bootstrapping[28] can be applied for the PLS model validation. Essentially, as shown in Figure 2, CV describes a procedure in which we divide the data into a number of groups to develop a number of parallel models, while one of the groups is deleted.[29−31] The activity predicted for this group is then

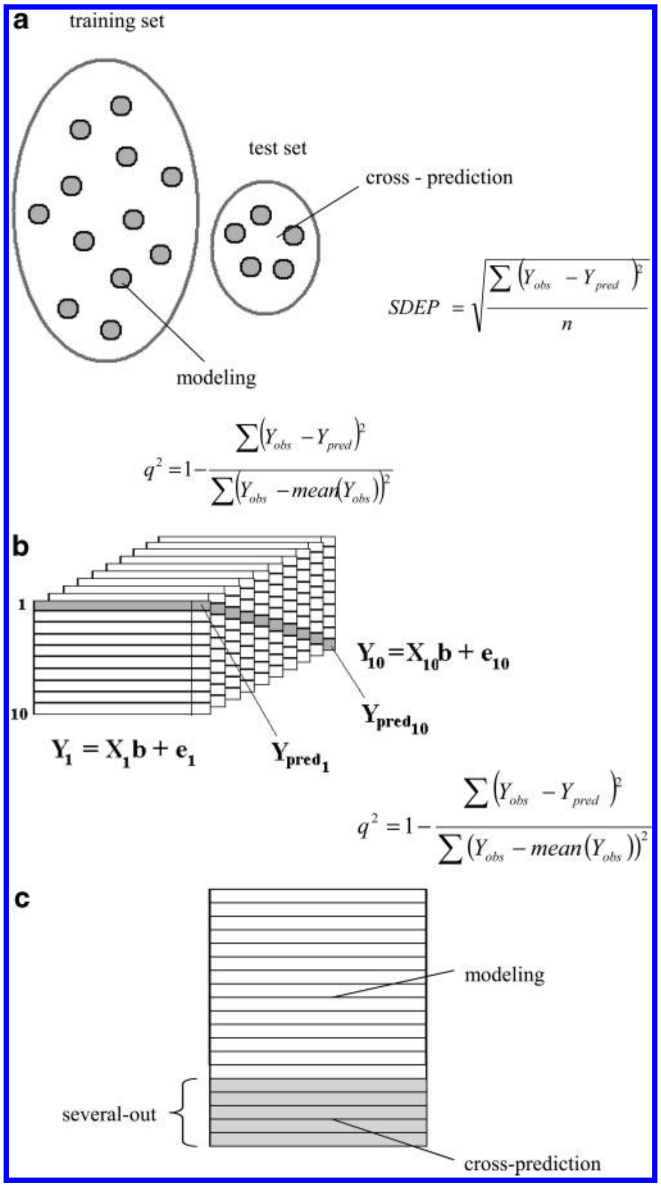
a
training set
test set
cross - prediction
$SDEP = \sqrt{\frac{\sum(Y_{obs} - Y_{pred})^2}{n}}$
modeling
$q^2 = 1 - \frac{\sum(Y_{obs} - Y_{pred})^2}{\sum(Y_{obs} - mean(Y_{obs}))^2}$
b
1
10
$Y_{10} = X_{10}b + e_{10}$
$Y_{pred_{10}}$
$Y_1 = X_1 b + e_1$
Y_{pred_1}
$q^2 = 1 - \frac{\sum(Y_{obs} - Y_{pred})^2}{\sum(Y_{obs} - mean(Y_{obs}))^2}$
c
modeling
several-out
cross-prediction

Figure 2. A validation of data noisy multidimensional QSAR models by cross-prediction. The sampling within molecular objects forms the *training* and *test*, respectively (a). Usually, an iterative leave-one-out cross-validation LOO CV (b) or cross-prediction by a single LSO/LOO CV iteration (c) is used for the estimation of the model relevance. Thus, in a series of experiments a model calculated for *all but one* molecule is used for the calculation of the activity of the eliminated molecule (b) or a model calculated for the *training* set (e.g., molecules 1−7) is to be cross-predicted into the *test* set (e.g., molecules 8−10) (c). The q^2 or SDEP (r^2) parameters are used for measuring model quality, respectively. $q^2 = 1 - (\sum(\text{obs}_i - \text{pred}_i)^2/\sum(\text{obs}_i - \text{mean(obs)})^2)$ SDEP = $(\sum(\text{pred}_i - \text{obs}_i)^2/n)$ where obs is the assayed values; pred is the predicted values, mean is the mean value of obs, and *i* refers to the object index, which ranges from 1 to *m*, *n* − the number of compounds included in the test series.

used for validating final model quality. Often, the deleted group, or so-called test set, contains a single molecule. Iterative manyfold cross-prediction for all molecules using the models derived from *all with the exception of this one molecule* is called leave-one-out (LOO) CV (Figure 2b). In practice, a q^2 parameter calculated during LOO CV is a basic measure for the performance of multidimensional QSAR models. Alternatively, if a larger test set is defined among molecular objects, the model quality is tested by activity predictions within this set by a model optimized using all but test set molecules, i.e, so-called training set molecules. As a rule, a single test set combination is tested in the latter process, and an SDEP parameter probes model quality, as illustrated in Figure 2c. Since, in this particular case, modeling within the training set must also be performed using the robust LOO CV protocol, the latter procedure can be formally described as a single iteration of the leave-some (or several)-out (LSO) CV coupled with LOO CV (LSO/LOO). Tropsha observed that, if we test several LSO/LOO iterations, no correlation between the quality of the model (given by a q^2 value) and the quality of prediction in the test set (SDEP values) exists.[32−34] In other words, the high quality model generated in the training set does not guarantee a proper prediction in the test set. The reasons for that can be easily explained.[35] Thus, during model validation in 3D-QSAR we are processing a strictly limited set of molecules for which an activity has been measured a priori. All of them should be active compounds. In fact, none of the predictions are real external predictions toward virtual molecules of really unknown properties. Thus, we propose in this publication to use for such a condition a term cross-prediction. Intuitively, during cross-prediction we can indicate at least two types of molecular objects: these of higher congenericity (hypothetical similarity level implying similar ligand−receptor interaction mechanism) that can fit easily a common mathematical model, and those of the lower fit into such a model. If we select into the training set preferentially the objects that can be easier fitted into the model, then the chance that the remaining group would provide a worse fit is higher since only the *worse* molecules are available for the test set. Because we have not made any provisions for the division of the molecules into the training and test sets any correlation between modeling ability in the training set and cross-predictions in the test set would be rather a surprise.

Stochastic Model Validation (SMV) for the Data Noise Conditions. To explore further cross-prediction we generalized model validation by the investigations into all possible combinations of the molecular objects into the training/test containing *t* (training) and *all-t* (test) set molecular objects, respectively. This will give *all!/(all-t)!t!* such combinations. Technically, the scheme is constructed by probing all LSO/LOO iterations, as shown in Figure 3. Formally, this method is an extension of the single iterations analyzed by Tropsha or Doweyko.[32−34] The application of SMV for the analysis of the modeling and cross-predictive ability of QSAR schemes has been discussed thoroughly in ref 35. Figure 4 provides a few illustrative examples of this issue. Since the SMV scheme analyzes the influence of all inputs upon the quality of the cross-predictive models, it should be a nice measure of the modeling robustness, and we will use it below for such purposes. The main conclusion from the SMV is *that the q^2 estimator measures modeling rather than predictive ability*. Moreover, *modeling ability in the training set and predictive capability in the test set are of the dichotic nature, i.e., the higher modeling ability of the training group the lower predictability in a test set.*[28]

If we test a given QSAR data by a single training/test set sampling, then we obtain two single values of the SDEP and q^2 estimators characterizing the prediction/modeling ability within these two sets. However, because both these estimators depend on the way in which we sample the data, the

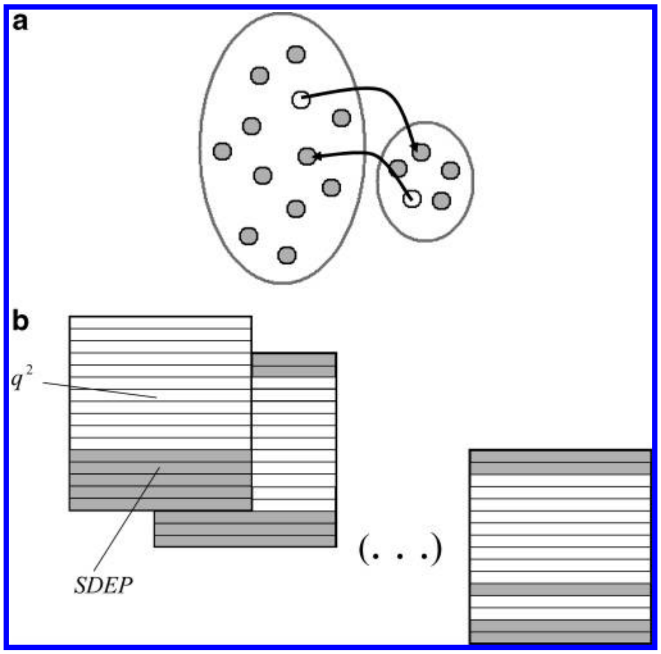

Figure 3. Stochastic model validation. The molecules are iteratively shuffled between the training and test sets until all combinations are probed (a) and a coupled LSO/LOO CV is iteratively repeated (b).

quality of the model can only be evaluated by the comparison of these two estimators. In contarary, SMV can be interpreted as a data probing technique. Thus, by changing the training/test set data sampling we are disturbing the statical QSAR modeling technique and observing how the answer of the model looks like when we change the modeling basis. For a given bioactive compound series the higher the robustness of the model is the lower a diffusion of an ideal single point answer is (Figure 4a).

The analysis of the QSAR data by cross-prediction in stochastic schemes probing intensively larger model populations, i.e., of the protocols similar to SMV, can be found in the literature only very rarely. To estimate the predictive ability more precisely Clark suggested a method he called *boosted leave-many-out CV*.[36] This method estimates model quality by the calculation of the external predictive error of the models obtained by several divisions of the analyzed series into smaller training sets and larger test sets. Training and test sets are selected using the OptiSim procedure which can generate representative or diverse test sets. The analysis of the external and the internal prediction ability of the obtained models confirms its dichotic nature. Sheridan who investigated the dependence of the selection of the training set on the prediction accuracy developed so-called retrospective CV.[37] Thus, a user-specified number of molecules are randomly selected from the original data set or a subset therefrom. These molecules form a training set. Molecules that are not included in the training set form the test set. Then, a QSAR model is generated within the training set, and the model is used to predict the activity of all molecules in the original data and a note is made of which were in the training set. Each molecule in the original data set is assigned an extrapolation measure of how close it is to the training set. Then, the procedure is repeated 10 times. The results were displayed as the predicted vs observed activity plots. The main conclusion is that similarity to molecules in the training set is a good discriminator for the accuracy of the QSAR cross-predictions.

Robust Predictions in Virtual Chemical Compound Space. In fact, cross-predictions in multidimensional QSAR are performed not in the direct search for new molecules, i.e., for molecular design, but for model validation. It is naive nowadays to expect a reliable prediction for a single **virtual molecule** on the basis of multidimensional QSAR. Instead, in this method we use the equations optimized by robust model validation for the illustration of the so-called interaction contour plots. 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. Basically, an extension of the molecules into such regions should increase or decrease compound activity. Compare refs 38 and 39 for the examples of such a molecular

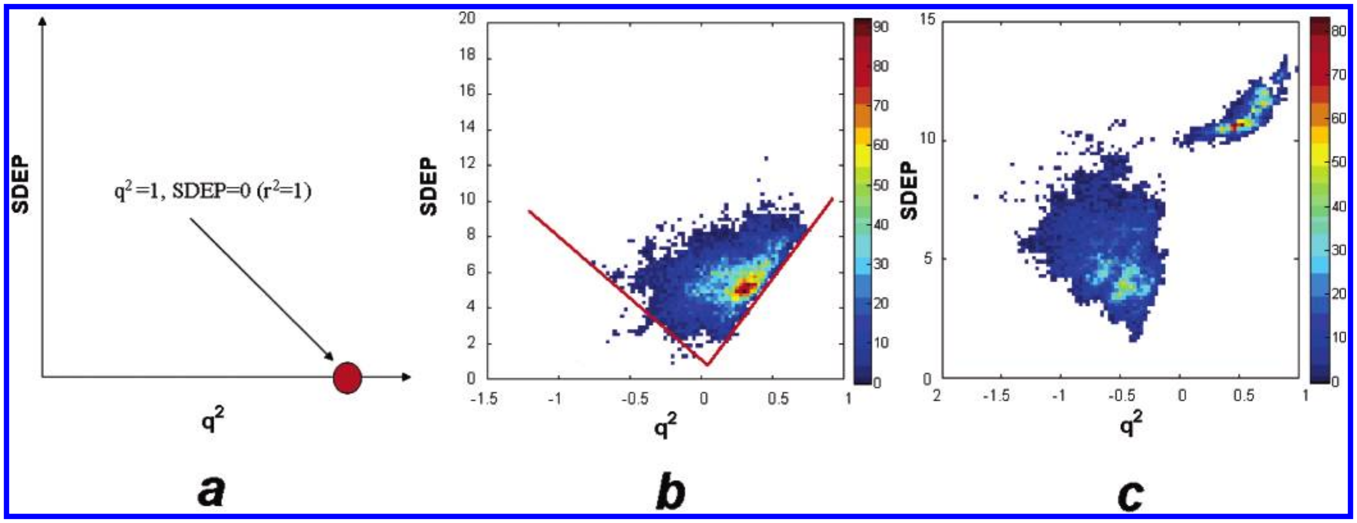

Figure 4. Stochastic model validation for simulated data. If we assume an ideal system in which the Y answer can always be given by modeled Y* without any error (Y*=Y), then a single point reports the SMV experiment ($q^2 = 1$, SDEP = 0) (a). The inclusion of the noise (Y*=Y + noise) results in the explosion of the SMV probes into the q^2, SDEP plot (b). If two different models operate for Y*, i.e., $Y_1* = f(X)$ or $Y_2* = g(X)$, then binomial plot forms a response in SMV. Adapted from ref 35.

design (prediction) approaches procedure.

Robust Answers for Molecular Superimposition Noise. Two different points are the only systems that can be put together without any ambiguity. Generally, for two systems that are not identical a number of covering modes exists. As a result superimposition generates an important noise that comes into the modeling. In consequence, in 3D-QSAR, and in CoMFA in particular, molecular superimposition significantly influences final results, e.g., the interaction contour plots can be completely different for two different superimposition modes (see ref 34 for further discussion).

The uncertainty in this context is due to a question if we can cover certain atoms within two molecules, respectively. The most radical solution that completely reduces the uncertainty due to molecular superimposition is passing over this operation.[40] Formally, this can be achieved by using some superimposition invariants as *distance geometry*,[41] *autocorrelation vectors*,[42] *3D MoRSE or RDF codes*,[43] or by the application of some default covering modes, e.g. covering along the molecular inertial axis in the CoMSIA,[44] Receptor-like Neuron Network,[45] GRIND,[46] or CoMMA[47] methods. Although we usually do not realize this, in fact, default superimposition is also performed by a variety of 2D-QSARs, e.g. in Free-Wilson analysis we are comparing certain molecular fragments, which means we are assuming they can interact similarly with the receptor or their atoms can be covered.[48] Eventually, QSAR implies a comparison, which means that molecular objects must be arranged in a certain *orientation* during such an operation. Most often a user must define this orientation. In this context an additional problem appears, namely, what is the relevance between the molecular superimposition modes for an individual molecule and a template and molecular recognition phenomena. In typical QSAR this question cannot be answered, but some novel approaches address such issues, as will be discussed below.

Although it is convenient to avoid superimposition this operation is extremely important for the results. In other words, if we get rid of this operation we can lose or modify information analyzed. Therefore, in the majority of 3D-QSAR we are controlling superimposition. In traditional approaches two possible responses (yes or no) answer the question of superimposition likelihood. We can improve model robustness by including some tolerance into the molecular superimposition. This can be interpreted as a fuzzy logic approach that forces a possibility for superimposition even if formally atoms cannot be covered providing a third additional answer, namely *atom congruence is acceptable*. Several improvements in the structure overlay have appeared that allow for more flexible or sophisticated superimposition.[49] In particular, neural networks have been used for flexible superimposition in Compass,[50] CoMSA,[51,52] or SOM-4D-QSAR.[53] Below we discuss the application of self-organizing neural network for the construction of fuzzy molecular representations that enable a control of the tolerance of molecular superimposition.[54]

Self-Organizing Neural Networks for Robust Superimposition. A transformation of 3D objects, e.g. the molecular surfaces, to 2D images always needs some projection or data transformation in order to reduce the dimension of the data to be visualized. Standard projections usually deform the topology of the objects to be transformed. In the early 1990s Zupan and Gasteiger developed a method for the Kohonen topographic mapping[55] of the molecular electrostatic potential to transform the 3D molecular surface data into a form of the 2D map.[56] For this transformation 3D coordinates sampled from the molecular surface are input directly to the competitive Kohonen neurons. Then, each output neuron of the 2D map is colored by the mean electrostatic potential value of the points projected from the 3D surface into this neuron. The preservation of the surface topology is among the most important issues of the transformation that is originally highlighted. Distinguishing between active and inactive molecules using Kohonen patterns was one of the first applications suggested.[57,58] Although there are some problems concerning a precise comparison of a pair of such maps, the method was used also for a so-called bioisosteric design.[59−61] However, if a single network is used for the processing of two different molecules in so-called *comparative mapping*, we obtain a kind of a precise superimposition tool.[58−62] The performance of such architecture in molecular design is thoroughly described in previous publications[45,51−54,58,59,61−64] and will not be further discussed here. The ability to generate fuzzy molecular surface representations which is of the importance for the explanation of the possible improvements in QSAR robustness is illustrated in Figure 5.

Comparative mapping if coupled with PLS analysis provided us a 3D-QSAR method, the Comparative Molecular Surface Analysis (CoMSA),[51] that is analogous to CoMFA but bases on the comparison of the surface sectors. The performance of CoMSA measured by single q^2 or SDEP parameters is generally better than CoMFA.[65,66] Figure 6 analyzes the robustness of CoMFA and CoMSA modeling of the CBG steroid activity by the SVM scheme.[35] It is clear that CoMSA gives higher q^2 and lower SDEP values. More recently Hasegawa improved the performance of CoMSA, additionally improving the method by coupling 3-way PLS,[67,68] and the non-neural CoMSA version has been developed.[52]

Data Reduction as a Robust Answer for Molecular Similarity Noise. Molecules are full of similarities, and it is not quite clear if in QSAR we are capable of extracting from the molecular structure only these aspects of similarity that is important for a certain activity. In fact, we lack the systematic investigations of this problem in QSAR. In classical QSAR different models can be constructed for a given molecular series, which of course brings a problem of model interpretation. For a discussion of this problem compare ref 69. It is observed in 3D-QSAR, namely in CoMFA, that different molecular descriptors give the final models of a similar statistical performance.[34] This indicates an effect that can be called molecular similarity noise, i.e., a fact that different analyses reveal different aspects of molecular correspondences that are however intercorrelated. Of course, similarity noise is unfavorable in 3D-QSAR because we can find molecular areas that are only coincidentally correlated with the activity and not the areas of the specific ligand−receptor interactions.

The CoMFA variables (describing individual molecular fields) included into a final model are weighted by the PLS analysis. Therefore, a final model contains information on the whole molecular field. We can indicate such an approach as the analysis of aggregated molecular similarity. If we were

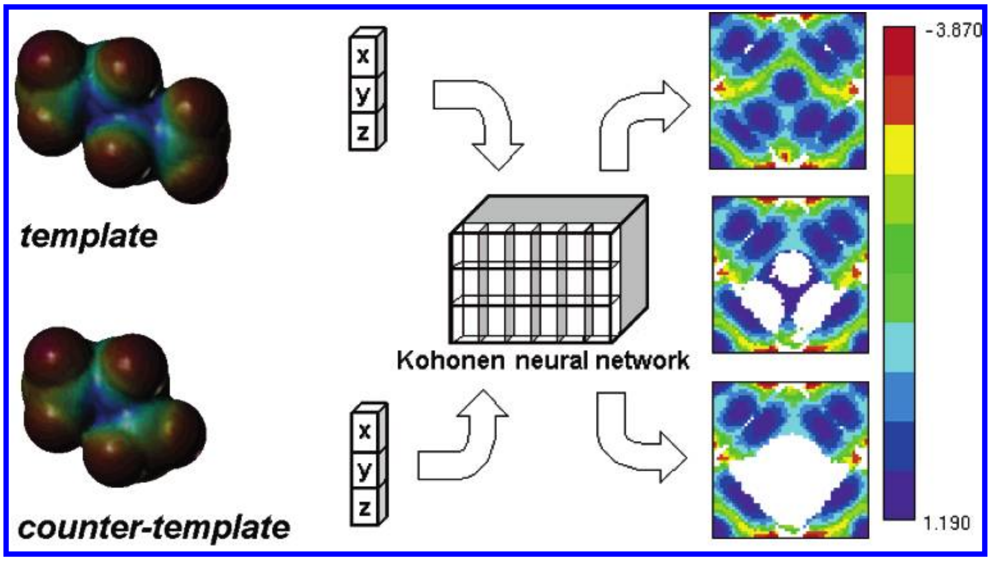

Figure 5. Fuzzy molecular surface representations by the comparative mapping of the butane and propane molecules. The molecules can be superimposed (no-empty, white, neurons are observed on two-dimensional propane map) or cannot be superimposed if the comparison is less tolerant (a large area of white neurons is observed on a two-dimensional propane map). Adapted from ref 54.

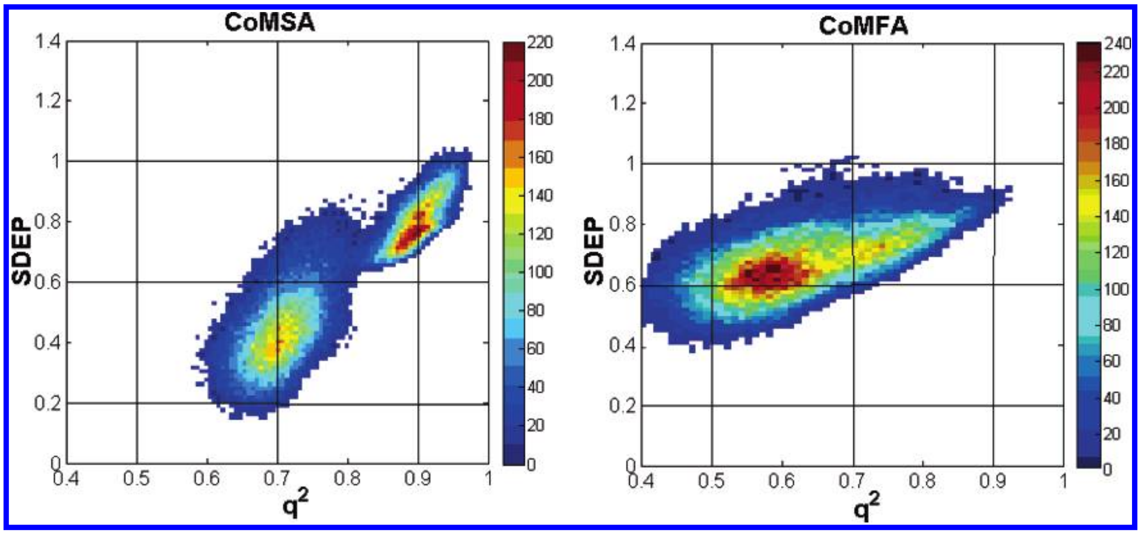

Figure 6. The SMV schemes for the validation of the CoMFA and CoMSA modeling of the CBG steroid series. Adapted from ref 35.

however capable of the indication of some among initial variables, then we could shift from a molecular similarity model to a pharmacophore based model. Generally, PLS only rarely combines with data elimination because this method works by *extracting* from each variable a right share to contribute for a total value.[29] However, CoMFA modeling coupled with data elimination can provide better results then the traditional CoMFA.[70,71] We have also shown that PLS combined with variable elimination, e.g. UVE or modified UVE versions, can be a powerful tool significantly improving both the predictive power of the CoMSA model and its illustrative ability. This in turn can bring an increased understanding of the molecular basis for the compounds biological interactions, e.g. steroid aromatase inhibitors.[72] Illustrative ability is especially worth mentioning in the aspect of pharmacophore mapping.[73] For better understanding Figure 7 shows the interaction contour plots for the series of hypolipidemic asarones.[74] Unlike CoMFA plots that are identical for all molecules, CoMSA provides a different illustration for each molecule. The incongruencies of molecular surfaces in individual molecules result in much more clear contour plots that are much easier for the interpretation. Thus, for example, we can observe that a carbonyl function is unfavorable for the activity of asarones.

Robust Answer for Conformational Noise. Biological response results from a certain atomic configuration (conformation). Since generally a variety of such molecular representations are available for a single molecule, this generates an effect that can be described by the term conformational noise. This raises the question if various conformational modes can produce statistically valid QSAR. 4D-QSAR addresses this issue by the investigations into the conformational space of molecules. In Hopfinger's 4D-QSAR molecular dynamic simulations provide *conformational ensemble profile* describing each molecule. Then, descriptors defining the pattern in which atoms occupy volume sectors are calculated.[75−80] Alternatively, a self-organizing neural network can be used for the generation of molecular volumes in SOM-4D-QSAR.[81] 4D-QSAR demands the variable elimination or selection step mounted as the integral filtering unit. Technically, Hopfinger's method uses the genetic algorithm for this purpose. In SOM-4D-QSAR we applied IVE-PLS.[52,72] 4D-QSAR can significantly improve model robustness even for such rigid molecules as

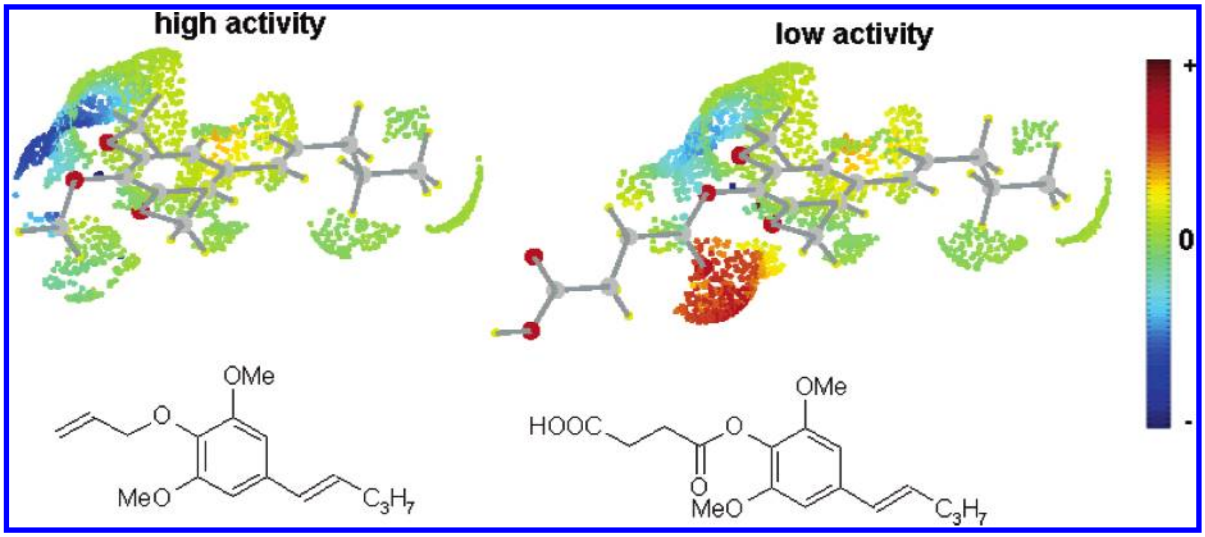

Figure 7. CoMSA interaction contour plots indicating the areas of the positive (activity decrease) and negative (activity increase) contribution for the activity in some arbitrarily selected high and low activity asarones. Adapted from ref 74.

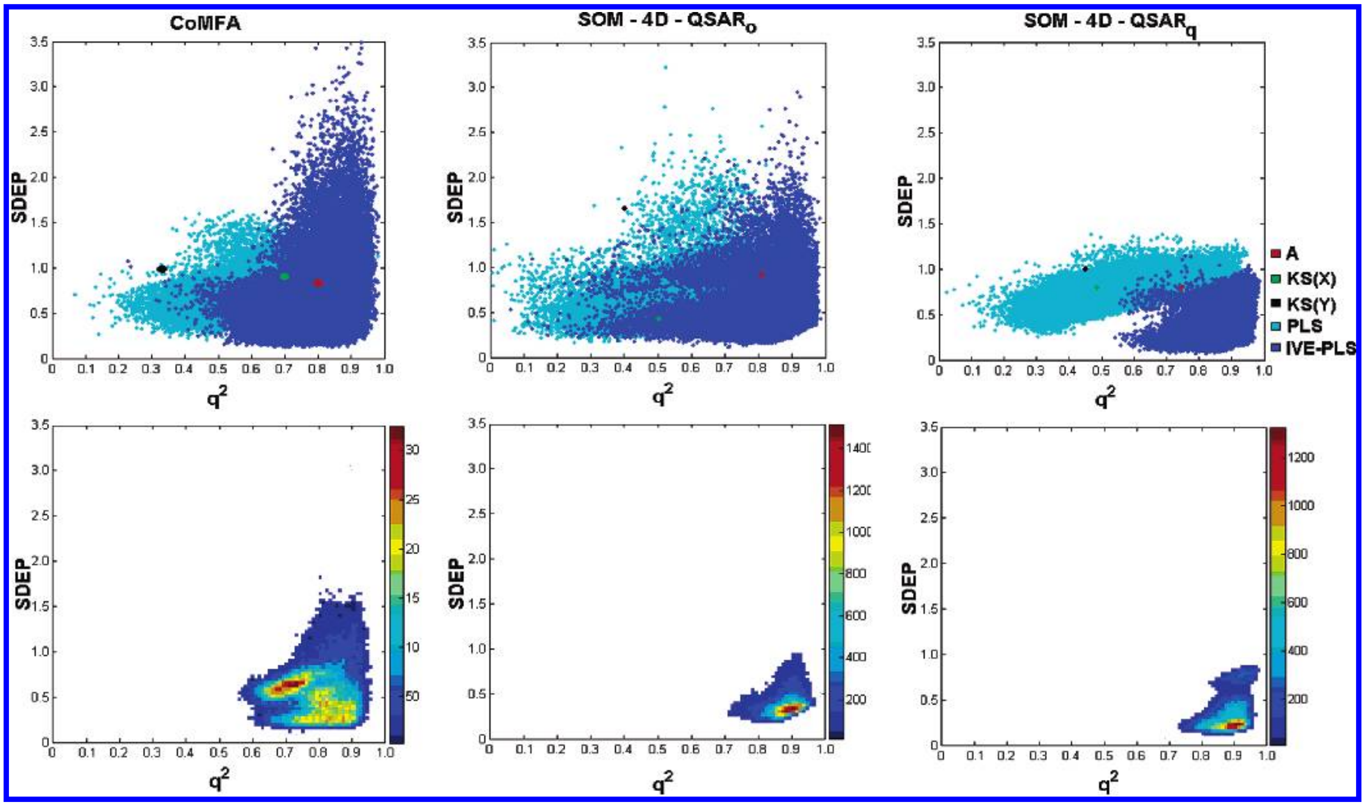

Figure 8. The SMV schemes for the estimation of the robustness of the CoMFA (a), SOM-4D-QSAR with occupancy descriptors (b) and SOM-4D-QSAR with charge descriptors (c) modeling of the CBG steroid series. Standard plots (upper line) illustrate the difference in the PLS and IVE-PLS modeling. Single points within the maps presented in the upper line indicate the following: A − a single validation usually reported in the literature, i.e., the model calculated for molecules 1−21 (training set) is cross-predicted to molecules 22−31 (test set); KS(X) is a single validation for the training and set sampled by the Kennard Stone protocol using the molecular descriptor (X) data; and KS(Y) is a single validation for the training and set sampled by the Kennard Stone protocol using the activity (Y) data. The density maps for the IVE-PLS models (bottom) are shown to illustrate the real distribution of the models.

steroid CBG series. This is illustrated in Figure 7 that shows SMV profiles for the CoMFA and SOM-4D-QSAR with occupancy and charge descriptors, respectively. Clearly, the latter model is also the most robust one. We have shown recently that SOM-4D-QSAR coupled with IVE-PLS is capable of the proper illustration of the HEPT inhibitor interaction with HIV-1 reverse transcriptase.[82] This not only gave a high quality model but also (as the only QSAR model reported) indicated a conformational mode that was consistent with real interactions determined by X-ray analysis.[83,84]

Robust Answer for Molecular Recognition Noise. QSAR is usually considered as a method that analyzes the data in a receptor-independent mode. This is not fully true, because biological activity data obviously depend on ligand−receptor interactions, but in fact generally a single number i.e., the activity value, accounts for the whole receptor. This forms a large imbalance in comparison to a data number describing a ligand molecule. Each molecule within the series analyzed in QSAR can interact within the receptor space receiving a specific ligand−receptor orientation. In a receptor independent approach we are assuming that an orientation optimized during the molecular superimposition step describes also a relative orientation during molecular recognition phenomena. If the real ligand−receptor orientation differentiates individual molecules and does not comply with that optimized during the superimposition step, then a clear discrepancy between modeled and real data appears. Moreover, additional specific effects can be stimulated by ligand−receptor interactions during binding. This generates an important noise that can be described by the term *molecular recognition noise*. Therefore, an important improvement can be achieved by the incorporation of the receptor structure data into QSAR modeling. Thus, binding affinities can be calculated in silico for a series of molecules and a receptor structure and then modeled into an equation. The COMBINE is an example of such a method[85,86] that is developed for the calculation of ligand−receptor binding energies. The analysis of the HIV-1 RT inhibitor series is one of the latest applications of this method.[87] Alternatively, ligand−receptor data can also be used in QSAR modeling.[88,89]

CONCLUSIONS

QSAR is a term describing a variety of approaches that are of substantial interest for chemistry. This method can be defined as indirect molecular design by the iterative sampling of the chemical compounds space to optimize a certain property and thus indirectly design the molecular structure having this property. However, property production and modeling the interactions of chemical molecules in biological systems provides highly noisy data, which makes predictions a roulette risk. In this paper we briefly review the origins for this noise, particularly in multidimensional QSAR. This was classified as the data, superimposition, molecular similarity, conformational, and molecular recognition noise. We also indicated possible robust answers that can improve modeling and predictive ability of QSAR, especially self-organizing mapping of the molecular objects, in particular, the molecular surfaces, a method that was brought into chemistry by Gasteiger and Zupan.

REFERENCES AND NOTES

(1) Kolb, H.; Finn, G.; Sharpless, B. Click chemistry: Diverse chemical function from a few good reactions. *Angew. Chem., Int. Ed.* **2001**, *40*, 2004−2021.

(2) Mullin, R. Recalibrating the clinic. High-tech tools and streamlined business processes are making their way to the far reaches of the drug development pipeline. *C&EN* **2005**, *83*, 29−39.

(3) De Julian-Ortiz, J. Virtual Darwinian drug design: QSAR inverse problem. *Comb. Chem. High Throughput Screening* **2000**, *4*, 295−310.

(4) Hansch, C.; Leo, A. *Exploring QSAR: Fundamentals and applications in chemistry and biology*; American Chemical Society: Washington, DC, 1995.

(5) Cox, H.; Heaney, K. Approaches to robustness, *J. Acoust. Soc. Am.* **2003**, *113*, 2262−2262.

(6) Lipinski, A. Drug-like properties and the cause of poor solubility and poor permeability. *J. Pharmacol. Toxicol. Methods* **2004**, *44*, 235−249.

(7) Hann, M.; Oprea, T. Pursuing the leadlikeness concept in pharmaceutical research. *Curr. Opin. Chem. Biol.* **2004**, *8*, 255−263.

(8) Oprea, T. Current trends in lead discovery. Are we looking for the appropriate properties? *J. Comput.-Aided Mol. Des.* **2002**, *16*, 325−334.

(9) Hansch, C.; Hoekman, D.; Leo, A.; Weininger, D.; Selassie, C. Chem-bioinformatics: Comparative QSAR at the interface between chemistry and biology. *Chem. Rev.* **2002**, *102*, 783−812.

(10) Hodgson, J. ADMET − turning chemicals into drugs. *Nat. Biotechnol.* **2001**, *19*, 722−726.

(11) Oprea, T. 3D-QSAR modeling in drug design. In *Computational Medicinal Chemistry and Drug Discovery*; Tolleneare, J., De Winter, H.; Langenaeker; W., Bultinck, P., Eds.; Marcel Dekker: New York, 2004.

(12) Kubinyi, H. Quantitative structure−activity relationships. The bilinear model, a new model for nonlinear dependence of biological activity on hydrophobic character. *J. Med. Chem.* **1977**, *20*, 625−629.

(13) Anzali, S.; Gasteiger, J.; Holzgrabe, U.; Polanski, J.; Sadowski, J.; Teckentrup, A.; Wagener, M. The use of self-organizing neural networks in drug design. Kubinyi, H., Folkers, G., Martin, Y., Eds.; Kluwer: Dordrecht, The Netherlands, 1998.

(14) Xu, L.; Zhang, W. Comparison of different methods for variable selection, *Anal. Chim. Acta* **2001**, *446*, 477−483.

(15) Leardi, R. Genetic algorithms in chemometrics and chemistry: A review. *J. Chemom.* **2001**, *15*, 559−569.

(16) Saxena, K.; Prathipati, P. Comparison of MLR, PLS and GA-MLR in QSAR. *SAR QSAR Environ. Res.* **2003**, *14*, 433−445.

(17) Vapnik, V. N. *The nature of statistical learning theory*; Verlag Springer: New York, 1999.

(18) Furey, T.; Cristianini, N.; Duffy, N.; Bednarski, D.; Schummer, M.; Haussler, D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* **2000**, *16*, 906−914.

(19) Norinder, U. Support vector machine models in drug design: application to drug transport processes and QSAR using simplex optimizations and variable selection. *Neurocompiuting* **2003**, *55*, 337−346.

(20) Demiriz, A.; Bennet, K.; Breneman, C.; Embrechts, M. Support vector machine regression in chemometrics. *Comput. Sci. Stat.* **2001**, *33*, 289−296.

(21) Corne, D. W.; Martin A. C. Artificial intelligence in bioinformatics. *Comput. Chem.* **2002**, *26*, 1−3.

(22) David, V.; Sanchez, A. Advanced support vector machine and kernel methods. *Neurocomputing* **2003**, *55*, 5−20.

(23) Burbidge, R.; Trotter, M.; Buxton, B. Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Comput. Chem.* **2001**, *26*, 5−14.

(24) Varmuza, K. Multivariate data analysis in chemistry. In *Handbook of chemoinformatics*; Wiley VCH: Verlag: Weinheim, 2003.

(25) Esbensen, S.; Geladi, P. Principal component analysis. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37−52.

(26) Geladi, P.; Kowalski, B. Partial least squares: A tutorial. *Anal. Chim. Acta* **1986**, *185*, 1−17.

(27) Helland, I. Some theoretical aspects of partial least squares regression. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 97−107.

(28) Good, P. *Resampling methods: A practical guide to data analysis*; Birkhauser: 1999.

(29) Wold, S.; Sjöström, M.; Eriksson, L. In *The Encyclopedia of Computational Chemistry*; Wiley and Sons: Chichester, U.K., 1999.

(30) Wakeling, N.; Morris, J. A test of significance for partial squares regression. *J. Chemom.* **1993**, *7*, 291−304.

(31) Clark, M.; Crammer III, R. The probability of chance correlation using partial least squares (PLS). *Quant. Struct.−Act. Relat.* **1993**, *12*, 137−145.

(32) Tropsha, A.; Gramatica, P.; Gombar, K. The importance on being earnest: Validation is the absolute essential for successful application and interpretation of QSAR models. *QSAR* **2003**, *22*, 69−77.

(33) Golbraikh, A.; Tropsha, A. Beware of q^2! *J. Mol. Graphics Modell.* **2002**, *20*, 269−276.

(34) Daweyko, A. 3D-QSAR illusions. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 587−596.

(35) 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.

(36) Clark, R. Boosted leave-many-out cross-validation: The effect of training and test set diversity on PLS statistics. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 265−275.

(37) Sheridan, R.; Feuston, B.; Maiorov, V.; Kearsley, S. Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1912−1928.

(38) Cramer III, R.; Patterson, D.; Bunce, J. Comparative molecular field analysis (CoMFA). *J. Am. Chem. Soc.* **1998**, *110*, 5959−5967.

(39) Kubinyi, H. Comparative molecular field analysis (CoMFA). In *Handbook of Chemoinformatics. From data to knowledge*; Gasteiger, J., Ed.; Wiley VCH: BRD, Weinheim, 2003.

(40) Melani, F.; Gratteri, P.; Adamo, M.; Bonaccini, C. Field interaction and geometrical overlap: A new simplex and experimental design based computational procedure for superposing small ligand molecules. *J. Med. Chem.* **2003**, *46*, 1359−1371.

(41) Lemmen, C.; Lengauer, T. Computational methods for the structural alignment of molecules. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 215−232.

(42) Wagener, M.; Sadowski, J.; Gasteiger, J. Autocorrelation of molecular surface properties for modeling corticosteroid binding globulin and cytosolic Ah receptor activity by neural networks. *J. Am. Chem. Soc.* **1995**, *117*, 7769−7775.

(43) Zupan, J.; Gasteiger, J. *Neural Networks in Chemistry and Drug Design*, 2nd ed.; Wiley VCH: BRD, Weinheim, 1999.

(44) Klebe, G.; Abraham, U.; Mietzner, T. Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. *J. Med. Chem.* **1994**, *37*, 4130−4136.

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

(46) Pastor, M.; Cruciani, G.; McLay, I.; Pickett, S.; Clementi, S. GRID-independent descriptors (GRIND): A novel class of alignment-independent three-dimensional molecular descriptors. *J. Med. Chem.* **2000**, *43*, 3233−3243.

(47) Silverman, B.; Platt, D. Comparative molecular field moment analysis (CoMMA). *J. Med. Chem.* **1996**, *39*, 2129−2140.

(48) Free, S.; Wilson, J. A mathematical contribution to structure -activity studies. *J. Med. Chem.* **1964**, *7*, 395−399.

(49) Korhonen, S. P.; Tuppurainen, K.; Laatikainen, R.; Peräkyla, 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.

(50) Jain, A.; Koile, K.; Champan, D. Compass: Predicting biological activities from molecular surface properties. Performance comparison on a steroid benchmark. *J. Med. Chem.* **1994**, *37*, 2315−2327.

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

(52) 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.

(53) Polanski, J.; Bak, A.; Gieleciak, R.; Magdziarz, T. Self-organizing neural network for modeling robust 3D and 4D QSAR: Application to dihydrofolate reductase inhibitors. *Molecules* **2004**, *9*, 1148−1159.

(54) Polanski, J. Molecular Shape Analysis. In *Handbook of Chemoinformatics. From data to knowledge*; Gasteiger, J., Ed.; Wiley VCH: BRD, Weinheim, 2003.

(55) Kohonen, T. *Self-organizing and associate memory*, 3rd ed.; Springer; Berlin, 1989.

(56) Gasteiger, J.; Li, X.; Rudolph, C.; Sadowski, J.; Zupan, J. Representation of molecular electrostatic potentials by topological feature maps. *J. Am. Chem. Soc.* **1994**, *116*, 4608−4620.

(57) Gasteiger, J.; Zupan, J. Neural networks in chemistry. *Angew. Chem., Int. Ed. Engl.* **1993**, *32*, 503−512.

(58) Polanski, J.; Gasteiger, J.; Wagener, M.; Sadowski, J. The comparison of molecular surfaces by neural networks and its application to quantitative structure activity studies. *Quant. Struct.−Act. Relat.* **1998**, *17*, 27−36.

(59) Anzali, S.; Barnickel, G.; Krug, M.; Sadowski, J.; Wagener, M.; Gasteiger, J.; Polanski, J. The comparison of geometric and electronic properties of molecular surfaces by neural networks: Applications to the analysis of corticosteroid binding globulin activity of steroids. *J. Comput-Aided Mol. Des.* **1996**, *10*, 521−534.

(60) Anzali, S.; Maderski, W.; Osswald, M.; Dorsch, D., Endothelin antagonists: Search for surrogates of methylendioxyphenyl by means of a Kohonen neural network. *Bioorg. Med. Chem. Lett.* **1998**, *8*, 11−16.

(61) 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.

(62) Barlow, T. Self-organizing maps and molecular similarity. *J. Mol. Graphics* **1995**, *13*, 24−27.

(63) Polanski, J.; Gasteiger, J. The comparison of molecular surface by assembly of self-organizing neural network, In proceedings of the III-th International Conference "Computers in Chemistry '94", Technical University of Wroclaw, Wroclaw, Poland, 1994, p 88.

(64) Livingstone, D.; Manallack, D. Neural networks in 3D QSAR. *QSAR Comb. Sci.* **2003**, *22*, 510−518.

(65) Polanski, J.; Gieleciak, R.; Bak, A.; Jarzembek, K.; Wyszomirski, M. The comparative molecular surface analysis (CoMSA). A novel efficient technique for drug design. *Acta Pol. Pharm.* **2002**, *59*, 459−461.

(66) 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.

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

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

(69) Wermuth, C. The impact of QSAR and CADD methods in drug discovery In *Rational approach to drug design*; Höltji, H., Sippl, W., Eds.; Prous Science: Barcelona, 2001.

(70) Cho, S.; Tropsha, A. Cross-validated r^2-guided region selection for comparative molecular field analysis: A simple method to achieve consistent results. *J. Med. Chem.* **1995**, *38*, 1060−1066.

(71) Cho, S.; Tropsha, A.; Suffnes, M.; Cheng, Y.; Lee, K. Antitumor agents. 163. Three-dimensional quantitative structure activity relationship study of 4′-O-dimethylepipodophyllotoxin analogues using the modified CoMFA/q^2-GRS approach. *J. Med. Chem.* **1996**, *39*, 1383−1395.

(72) 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 enzyme. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 656−666.

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

(74) Gieleciak, R.; Magdziarz, T.; Bak, A.; Polanski, J. Modeling robust QSAR 1: Coding molecules in 3D QSAR − from a point to surface sectors and molecular volumes. *J. Chem. Inf. Model.* **2005, *45*,** 1447−1455.

(75) 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.

(76) Albuquerque, M.; Hopfinger, A.; Barreiro, E.; De Alencastro, R. Four-dimensional quantitative structure−activity relationship analysis of a series of interphenylene 7-oxabicycloheptane oxazole thromboxane A$_2$ receptor antagonists. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 925−938.

(77) Santos-Filho, O.; Hopfinger, A. A search for sources of drug resistance by the 4D-QSAR analysis of a set of antimalarial dihydrofolate reductase inhibitors. *J. Comput-Aided Mol. Des.* **2001**, *15*, 1−12.

(78) Ravi, M.; Hopfinger, A.; Hormann, R.; Dinan, L. 4D-QSAR analysis of a set of ecdysteroids and a comparison to CoMFA modeling. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1587−1604.

(79) Krasowski, M.; Hong, X.; Hopfinger, A.; Harrison, N. 4D-QSAR analysis of a set of propofol analogues: Mapping binding sites for an anesthetic on the GABA$_A$ receptor. *J. Med. Chem.* **2002**, *45*, 3210−3221.

(80) Hong, X.; Hopfinger, A. 3D-pharmacophores of flavonoid binding at the benzodiazepine GABA$_A$ receptor site using 4D-QSAR analysis. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 324−336.

(81) 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.

(82) Bak, A.; Polanski, J. The 4D-QSAR study on anti-HIV HEPT analogues. *Bioorg. Med. Chem.* **2006**, *14*, 273−279.

(83) Kireev, D.; Chrétien, J.; Grierson, D.; Monneret, C. A 3D-QSAR study of a series of HEPT analogues: The influence of conformational mobility on HIV-1 reverse transcriptase inhibition. *J. Med. Chem.* **1997**, *40*, 4257−4264.

(84) 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.

(85) Murcia, M.; Ortiz, A. Virtual screening with flexible docking and COMBINE-based models. Application to a series of factor Xa inhibitors. *J. Med. Chem*. **2004**, *47*, 805−820.

(86) Wang, T.; Wade, R. Comparative binding energy (COMBINE) analysis of OppA-peptide complexes to relate structure to binding thermodynamics. *J. Med. Chem.* **2002**, *45*, 4828−4837.

(87) Rodriguez-Barrios, F.; Gago, F. Chemometrical identification of mutations in HIV-1 reverse transcriptase conferring resistance or enhanced sensitivity to arylsulfonylbenzonitriles. *J. Am. Chem. Soc.* **2004**, *126*, 2718−2719.

(88) Rondeau, J. M.; Schreuder, H. Protein Crystallography and Drug Discovery, In *The Practice of Medicinal Chemistry*; Wermuth, C. G., Ed.; Academic Press: 2003; pp 417−443.

(89) Sippl, W.; Contreras, M.; Parrot, I.; Rival, M.; Wermuth, G. Structure-based 3D-QSAR and design of novel acetylcholinesterase inhibitors. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 395−410.

CI050314B