

How to Recognize and Workaround Pitfalls in QSAR Studies: A Critical Review

T. Scior*¹, J.L. Medina-Franco², Q.-T. Do³, K. Martínez-Mayorga², J. A. Yunes Rojas¹ and P. Bernard³

¹Department of Pharmacy, Benemérita Universidad Autónoma de Puebla, C.P. 72570 Puebla, Mexico; ²Torrey Pines Institute for Molecular Studies, 11350 SW Village Parkway, Port St. Lucie, Florida 34987, USA; ³Greenpharma S.A., 3 allée du Titane, 45 100 Orléans, France

Abstract: Quantitative Structure-Activity Relationships (QSAR) are based on the hypothesis that changes in molecular structure reflect proportional changes in the observed response or biological activity. In order to successfully conduct QSAR studies certain conditions have to be met that are not frequently reported in the literature. This suggests that some authors are not aware of the principle flaws, occasional shortcomings, and circumstantial downsides of QSAR methods. The present paper focuses on prerequisites to set up correct models and on limitations of model applications. Their implications are systematically described and illustrated as pitfalls that have strong implications in QSAR, and possible solutions are suggested. The paper is focused on small scale 2D- and 3D-QSAR studies for lead optimization. The work is enriched with comprehensive comments and non-mathematical explanations for the computer practitioner in Medicinal Chemistry.

Keywords: Chemoinformatics, quantitative structure-activity relationships, problems, solutions, regression analysis, virtual screening.

Dedicated to Prof. Dr. Joachim Karl Seydel, Borstel, Germany, on the occasion of his birthday, 13.01.2010.

INTRODUCTION

Quantitative Structure-Activity Relationships (QSAR) is a widely accepted predictive and diagnostic used for finding associations between chemical structures and biological activity (Table 1). QSAR has emerged and has evolved trying to fulfill the Medicinal chemists' need and desire to predict biological response [1]. In a seminal paper *Kubinyi* describes the history of QSAR [2].

At the beginning of modern science during the 18th century, *Auguste Comte*, a French philosopher, made a comment regarding the progress of early chemistry: "Every attempt to employ mathematical methods in the study of chemical questions must be considered profoundly irrational and contrary to the spirit of chemistry." Furthermore: "If mathematical analysis should ever hold an important place in chemistry (...) it would occasion a swift and general degeneration of that science."

Intuitive Medicinal chemist's thinking is based on the assumption that there is an inherent association between chemical structure and biological activity. This is the starting point to use mathematical tools to correlate structural descriptors (predictors, regressors) and biological activity (response or target variable). Although the researcher uses the fundamental similarity principle that states that similar structures have similar activities [3], and despite significant advances in computer-based similarity searching and similarity-based virtual screening [4-6], it is still a challenge to devise meaningful concepts and uses of structural similarity [7-13].

There are different types of computational methods in QSAR for different levels of data complexity [14]: two-dimensional (2D), three-dimensional (3D) and higher dimensional methods.

2D-QSAR is insensitive to the conformational arrangements of atoms in space, while 3D-QSAR needs information on the position of the atoms in three spatial dimensions [15]. In 4D-QSAR for each molecule a set of automatically docked orientations and conformations are devised by genetic algorithms [16-20]. Induced-fit scenarios of ligands upon binding to the active site and solvation models can be thought of as the fifth (protein flexibility) and sixth (entropy) dimensions in 5D- and 6D-QSAR, respectively.

Three major categories of descriptors in QSAR or QSPR work exist: (i) the simplest type (one-dimensional information) provides constitutional information such as molecular weight, number and types of atoms and bonds in a condensed formula, i.e. counts for a whole molecule or its chemical fragments and groups like the number of aliphatic ethers, or tertiary amines, etc.; (ii) 2D-descriptors are based on two-dimensional properties of either fragments (substituent constants (σ , π , MR) or whole molecules (logP, reactivity). Topological indices are based on atoms and their bond connectivities; (iii) 3D-descriptors reflect the three-dimensional nature of molecular structures (conformations, isomerism) and the surrounding space (stereochemistry). Another way to consider descriptor classification is according to fragmental or whole molecule properties [15].

Today, two major tenets exist in QSAR that are opposed to one another. On one hand, the mathematical-minded adept and experts in statistics solve correlation problems in a formally correct manner [21,22]. Their correlations tend to be statistically sound but mechanistically spurious because the parameters in the equations are difficult to interpret. In contrast, other researchers follow a QSAR procedure (Table 1) guided by an established protocol of some QSAR software [23-25]. Here the choice of parameters is often taken on intuitive grounds, and amenable to draw conclusions for a next generation cycle of candidate design, synthesis and testing.

*Address correspondence to this author at the Department of Pharmacy, Benemérita Universidad Autónoma de Puebla, C.P. 72570 Puebla, Mexico; E-mail: tscior@siu.buap.mx

Table 1. Typical Steps in QSAR Modeling

1	Response data collection	Experimental measurements (imperfections) and biological material (variations) are error-prone but data thereof should be normally distributed. Systematic errors should be absent.
2	Selection of congeners	Congeners are similar enough to guarantee both, the same interaction mechanism and a wide potency range of several log units.
3	Clustering	Divide the series into chemical groups of more specific homologous variations.
4	Data sets	Split the series into a training set and smaller representative test set.
5	3D-QSAR	Build molecule models and perform conformational analysis and alignment. Alignment can be performed using SBA or PBA, i.e. either atom-wise superposition of common substructures (scaffold) or superposition by means of docking into a target binding site.
6	Descriptor selection	Calculating parameters to numerically represent structural features of the compounds. The numbers can also be of categorical nature (yes/no, 1/0).
7	Model generation	Applying statistical means: PCA for complexity reduction, SLR and MLR, PLS for linear regression; clustering, and factorial design.
8	Internal model validation	Using LOO - crossvalidation to improve the Q^2 criterion.
9	Control loop	If necessary, go back to step 5 for modifications and rerun 6 and 8; else an acceptable model is reached.
10	Test set	Evaluate preliminary equations in the test set.
11	Control loop	If necessary, go back to step 5 for modifications and redo steps 6 to 10.
12	Interpretation	Interpret the final model and start a new cycle of drug development.
13	External model validation	Predict hitherto unseen compounds to test the predictive power of the final model.

They are grouped into three blocks: (i) input data preparation and pre-processing (steps 1 to 4); (ii) generation of models and equations (steps 5 to 7) in combination with their immediate validation to control model quality (steps 8 to 11) or external validation (13); and (iii) the interpretation of results (12). The preliminary QSAR models are tested in an iterative manner along with synthesis and bioassays. Conditions for step 13 are rarely met. If available use different programs and tools for comparison.

- Word used as "algorithm" is the model equation.

Table 2. Listing of Keywords from the OECD Guidelines for QSAR: The Use of the Word "Algorithm"* is Inappropriate, it is not a Sequence of Instructions but a Model Equation [29]

Defined endpoint
Unambiguous model "algorithm"* in MLR, PCA, PCR, PLR, PLS
Defined application domain
Mechanistic interpretation
Appropriate validation
Model quality
Number of compounds
Number of descriptors (not mentioned but added by the authors to estimate the overfitting risk)
Coefficient of determination r^2 as a goodness-of-fit index
SEE (standard error of the estimate)
F (variance ratio)
CV Q^2 (internal)

QSAR is a vast research area and several methods are available [14,26-29] that deal with very large, diverse and frequently noisy data sets with thousands of compounds [9,30]. The scope of the present paper, however, addresses small-scale studies that are conducted in the medicinal chemistry community in lead optimization projects. There is an ongoing discussion regarding the benefits and downsides of QSAR [15,26,28,31-34]. In view of such controversy, OECD guidelines for QSAR (Table 2) have been released [29] next to a monograph to standardize QSAR on behalf of regulatory

authorities, industries, and health institutions in charge of risk management and environment protection concerning chemicals [27]. Important prerequisites and pitfalls in QSAR have been discussed in the literature with a non-mathematical flavor [6,28,33-35]. There are also papers more focused on mathematical argumentations that may not reach the wide medicinal chemistry audience [26,27,36-38].

This review is directed to medicinal chemists with experience or at least literature knowledge in the field of compu-

tational molecular simulations, and statistics (Table 3). We explain why small-scale QSAR studies may fall short of expectation and show the need of caution (Table 3). The work is based on the authors' experience and, to the best of our knowledge, constitutes a more comprehensive and systematic presentation thereof [23,24,30,25,51-57].

The paper is organized in two major sections. In the first section the specific pitfalls and more general flaws of QSAR studies are described and potential workarounds are suggested in the following section.

In principle several statistical approaches exist to relate descriptors based on chemical structures (x-variables) with their biological response or activity (y-variable) [14,27].

Simple Linear Regressions

SLR is the simplest regression technique, i.e. a least squares procedure to produce a straight line in a two dimensional Cartesian plot (with 2 orthogonal axes x and y) fol-

lowing the equation $y = b x + c + e$; where y is the target or response variable, b is the slope, x is the parameter or predictor variable, c the intercept, and e the sampling error. The latter is reduced by the least squares procedure. In regression y and x are called dependent variable (DV) or independent variable (IV), respectively.

Multiple Linear Regressions

MLR is an extension of SLR dealing with more than one IV. In a hyper-dimensional data space (with nonorthogonal axes representing redundant information) just one DV (univariate) and many more IVs numerically describe the biological and chemical properties in QSAR. In rare cases there are more than one DV (multivariate model). DVs are random variables representing the measured values concerning the biological activities of the drug candidates. Note: MLR is a full least squares procedure and the standard implementation in QSAR [58]. It does not perform well in presence of correlated descriptors [27].

Table 3. Issues in QSAR. Recommended Textbooks to Expand the Reader's Knowledge About QSAR Techniques are Marked with an Asterisk (*)

Reported limitations and shortcomings	Ref.
Q^2 is no reliable proof of predictive ability of a QSAR model. High Q^2 is a necessary but not sufficient condition of predictability.	[39]
External validation (by predicting unseen examples with the final QSAR equations) is the only way to establish a reliable QSAR model.	[36]
Y-scrambling (to decrease the probability of chance correlations), multiple LOO-CV, external validation (to improve reliability) and definition of the applicability domain are recommended.	[26]
Discussion on variable selection and reduction of its sheer number, external model validation, assessment of model reliability and applicability borders (what others call model domain) is presented. Moreover, the authors provide values of statistical criteria to compare QSAR performance.	[27]
In silico ADME(T) models fail because of wrong expectations and interpretations or the QSAR models show invalid assumptions. Recommendations are given to develop, interpret, and use models.	[28]
Presentation of pitfalls focusing on QSAR models for toxicity. Using examples, the work discusses practices that lead to problems if they are not avoided. The problems are classified into three different sources: biological activity, numerical descriptors, and statistics.	[31]
QSAR modeling constitutes only an indirect design method and introduces noise by numerically describing the structural features of the compounds to quantify/optimize the biological response. Five different types of noise are presented: data, superimposition, molecular similarity, conformational, and molecular recognition.	[40]
A list of criteria to evaluate the predictive power of QSAR models is presented. QSAR models developed with LOO-CV are not reliable because no correlation between Q^2 of the training set and accuracy of prediction (R^2) for the test set exists. In addition, several strategies to divide the original data set into training and test sets are proposed. Sets for external validation are used to establish a reliable QSAR model.	[41]
In order to assess the reliability of QSAR models based on MLR, the authors present an automated solution (scoring function) to assess how well a new compound will be predicted.	[35]
New methods to identify "structure-activity cliffs" are presented. In particular, the articles evaluate pairs of molecules that are most similar but surprisingly show larger changes in activity compared to others with structural similarities.	[33,42,43,]
A textbook of basic concepts in computer-aided drug design. It introduces a vast variety of methods in molecular modeling to the newcomer. The fundamentals of QSAR are explained in detail.	[44] (*)
A reference handbook of a monograph-like systematic collection of quantitative descriptors for QSAR, contains listings of acronyms and descriptor abbreviations, tables and examples of calculations as references.	[45] (*)
Description of molecular properties relevant for the pharmacokinetic behavior of drugs and drug-like molecules, such as acidity, lipophilicity, water solubility, cell permeation. Also includes studies of conformational states.	[46] (*)
The textbook is intended for practical work with QSAR and provides hints on how to use and analyze data with analytical tools, but does not go into the details of statistical theory. Many examples come from pharmaceutical applications, especially QSAR. It includes factorial design.	[47] (*)
A textbook presenting an in-between approach when compared to [47] and [49]. It is less a practical guide than the book cited in [47] but introduces more of mathematical and statistical concepts without discussing algorithms like [49].	[48] (*)
A guide with practical recommendations. how to select compounds, deal with response data, descriptors, data dimensionality reduction, variable eliminations and selection, model testing and interpretation, a broad look on methods including artificial neural networks to extend QSAR.	[14] (*)
An in-depth look into the development of statistical methods and their implementation into algorithms for Fortran and C languages under different operation systems, e.g. least squares fitting, linear and polynomial regression. Factorial design is not treated.	[49]
A general introduction to statistics for pharmaceutical sciences providing a vast overview of fields of application. However, precise definitions of statistical terms and fundamental concepts are missing in the introductory part of each chapter (one - way or two-way ANOVA etc.). They would be very helpful for the novice.	[50]

Partial Least Squares

PLS is an extension to MLR especially designed to treat large amounts of noisy, multicollinear x data found in 3D-QSAR like CoMFA due to thousands of grid points loaded with computed x values.

Free-Wilson

Analyses are parameter-free additive correlation procedures, i.e. physicochemical properties (descriptors, IVs). Instead, they directly relate structural fragments (categorical variables) with biological properties [15]. The biological activity is thought of to be the sum of the independent increment of each present substitution fragment and the contribution of the unsubstituted scaffold [59]. Historically, this constant additivity assumption found its adepts for QSPR on small organic compounds. Today's challenge in the field of drug research faces more complex structural interactions with target receptors, probably due to nonadditive or synergistic effects (entropy, enthalpy). Nevertheless, the approach lends valuable insight to assist the elaboration of working hypotheses [58].

Factorial Analyses

FA reduces the explanatory complexity of numerical information (matrix of redundant variables x), i.e. as a result a much smaller number of variables (factors) is needed to explain the tendencies in the data. It requires a controlled (designed) experiment (factorial design) during which a response is measured under the documentation of controlled variables (assigned factor levels) and conditions (known factors) which influence the observed effects (response value). Despite a higher interpretability and predictability than MLR and better identification of variable interactions, this method cannot be applied to QSAR because the latter are undesigned experiments in search of hitherto unknown causal variables (structural contributions) and unassigned values. (So, one cannot speak of factors and levels here).

Principal Component Analyses

The aim of PCA is equivalent to FA, i.e. extracting data covariance in a matrix of variables x to achieve data reduction. FA and PCA are not identical in their implementations and applicability but their results come close if the input data of FA were homoscedastic, i.e. if the errors of all the measured response data share the same variance. In contrast to FA, PCA can be used to preprocess QSAR input data.

One at a Time Approach

An often-misleading simplification in drug design consists in that the chemical substituents always form independent groups with additive properties on a common scaffold (cf. assumptions of *Free-Wilson* and FA). Popular one-at-a-time decisions to synthesize and test drug candidate based on previous observation cycles are neither efficient nor their predictions reliable. In fact, they are comparable to a walk following an arbitrarily chosen straight line (extrapolation) without knowing the landscape (entropic and enthalpic to-

pology). Spurious results should therefore be no real surprise.

With time passed and experience gathered the computational practitioner may pose the question: is QSAR a sort of *random walk* similar to one-at-a-time research but situated on a higher level of sophistication through more complex multiterm linear equations?

PITFALLS IN QSAR STUDIES

QSAR modeling involves three mayor steps each of which contains its own group of pitfalls: (1) input data preparation and preprocessing, (2) model generation and validation, and (3) analysis of results, i.e., output interpretation (Table 1).

i. Pitfalls Concerning input Data Preparation and Preprocessing

Pitfall: Incompatible Concepts and Contraindications for QSAR

Multiconditionality

Drug action is based on a sequence of complicated physicochemical events (delivery, targeting, metabolism, and excretion) that are either still unknown or not fully understood on a molecular level. For this reason and because of hardware and software limitations *in silico* studies can only fragmentally reproduce real world observations. QSAR and QSPR are used to describe quantitatively ADMET processes in living cells, e.g. protein binding (plasma enzymes etc.), active and inactive metabolites, diffusion processes through tissues or entire body in cell tests or whole organism [26,28,29,31,60].

Common Action Mechanism and Multiple Binding Modes

An important prerequisite of QSAR is the use of a series of congeners with a common target structure and shared mechanism of action therein. Chemical similarity is not a foolproof guaranty for a common action mechanism of all congeners. A complication is the occurrence of various binding modes (MBM) of the very same ligand to its target molecule. Examples are guanidinium benzoate [61] and so-called flip-flop of xanthenes and symmetrical purines [52]. The most hitherto known crystal complexes show single modes but to what extend other MBM exists either under other crystallization conditions or in true physiological solutions can only be speculated. This way, QSAR is conducted under the silent assumption that no MBM is present when comparing molecular similarities with LBA or PBA docking techniques.

Multiple Targets and Multipotency

Normal QSAR work with cell-free data is not affected by drug binding to multiple targets. Such multipotencies occur *in vivo* when a molecule in lower doses binds to a biomolecule with higher affinity, while in higher doses the same ligand may bind to other targets with lower affinity.

Prodrug Function and Instable Products of Synthesis

The molecules considered in QSAR studies are not necessarily the ones responsible for the biological response. Such is the case of inactive prodrugs. Only after activation

by biotransformation their *in vivo* products become responsible for the measured potency. Any attempt to perform a sensible QSAR study must fail because of the hidden structural changes. When the experimental conditions of syntheses and bioassays are incompatible *in vitro* activity measurements may lead to the decay of unstable compounds generating new unknown structures. In both cases if not detected, completely wrong SAR assumptions are made [53].

Pitfall: Experimental Errors and Inappropriate Bioassays

At best compounds are tested under the same experimental conditions in the same place so that the measured potencies are directly comparable [31]. Any literature data for QSAR development, should be treated with caution, especially if they are a compilation of values from different sources [53]. In particular, when laboratory, instrumentation or personnel changes modifications in the preparation or assay protocol occur - whether intended or not - and an internal standard is needed. A reference compound (sometimes taken from the literature) provides a record about sampling errors. The value may be single one or a mean. With tens of repeat measurements (replicates) it should become a random variable showing signs of normal distribution as a proof that no systematic errors (bias) had jeopardized the measurements. In most cases bioassays are too expensive to be repeated but the modeler should ask if the value is a mean or not. QSAR should not be developed with statistical fits greater than the observed variability in the biological response and physicochemical data (cf. overfitting) [31]. A typically tolerated experimental imprecision of about 2-fold applied to 19 compounds is equivalent to a standard error of 0.2–0.3 log units, which limits the model to a R^2 not higher than 0.88 [32].

Whole cell tests include transport and diffusion processes through lipid barriers of cellular membranes and media prior to target binding. Often the former (enzyme-based) processes are slower or are saturated in such a way that the measured response is a mere function of the media and cell permeation performance of that molecule. This means activity depends on molecular lipophilicity or solubility but not binding affinity. It is wise to get a feedback from the laboratory running the assays to learn about what experimenter belief are minor changes to the protocol like addition of cosolvents, salts, or amounts used.

Pitfall: Data Size and Variety

The chemical space covered by a series of drug candidates tends to be rather small due to synthetic bias, i.e. similar substituents, absence of systematic changes because of unexplored routes of organic synthesis. Small data sets with 10 to 30 compounds, however, tend to show linear relationships between the predictor variables and the response. With some luck, even in larger data sets a straight line can be still fitted to apparently linear portions of the nonlinear data (Fig. (1)). In more general terms, if the spread in data points (molecules) is limited, often the series (molecules) is more likely to be homogeneous and break down of linear relationships between predictors and biological response is less probable [27]. Particularly, if the x data lie close together and only a few points cover the entire range then a good R^2 must not lead to the assumption that a linear relationship exists between main group and extreme points.

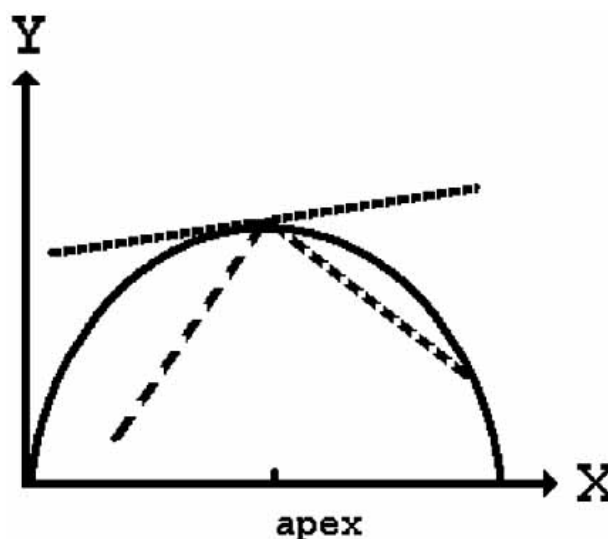


Fig. (1). Schematic explanation of apparently linear (discontinuous lines) but truly nonlinear data (continuous curve). Segments of the curve flanking the apex (maximum) in a distance do not show pronounced curvatures. Hence they can be idealized by a straight line through MLR (dotted and dashed lines). In practice the linear appearance is a computed result due to imprecise response measurements without sampling, outliers, and irregular spread of data points (x,y). Less ambiguity exists around the apex as a clear zone of non-linearity. All apparent linear relationships will rapidly break down when the data range encompasses both flanks and the apex. A larger range, however, is a rarely seen case, because drug research faces limited resources and focuses on promising candidates and not systematic data collection (bias).

Composition of Test Sets

Prior to model generation the data is usually divided into a study or training set and a test set, which is typically much smaller (*approx.* one tenth the size of the training set). To validate the model, the activities of each member of the test set are predicted by the QSAR equation. If not inspected, the few members in the test set may be not representative in terms of the activity and chemical variability. Sometimes the test set is modified or “filtered” to avoid outliers [31]. In this context *Polanski* comments “... 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. ...” [38]. *Gerald M. Maggiora*, who was involved in revising earlier versions of this work, stresses that “some of the outliers may, in fact, be activity cliffs. Thus, removing such points would severely prejudice models’ predictive capabilities” [52].

DV of categorical types: As the number of active molecules is, in general, significantly smaller than the number of inactive ones, the latter tends to be over-represented. A random partitioning of the training and test sets may result in a test set which is not representative of the whole ensemble (e.g. test set with only inactive products). When used for validation, the number of correctly predicted molecules tends to be artificially high leading to a misinterpretation of the robustness of the model, whereas the model is probably biased to the inactive molecules.

ii. Pitfalls Concerning Model Generation and Validation

Pitfall: Selection of Predictor Variables

Meaningless Descriptor Selection

Contrary to its widespread application dipole moment is not at all a useful parameter to describe electronic effects. It cannot be used satisfactorily as a measure of polarity, nor is it highly correlated with lipophilicity due to the geometric and electronic symmetry of the molecules, e.g. 1,4-dichlorobenzene. It does not possess a DM ($=0$), although its ortho- (1,2-) or meta- (1,3-) isomers do. However, solvent partition experiments revealed that their corresponding logP values are essentially the same (+3.4). Polarizability can be important only in reactions in gas phase but is surely not as important in aqueous solutions of biological systems. These findings suggest that their inclusions in QSAR studies should be deprecated. A listing of either commonly computed or experimentally determined descriptors used in VS has been published elsewhere [30]. Partition coefficients logP and logD, as well as MR and *Hansch* π constants (of lipophilic fragments) are popular variables which are more suitable to predict drug activity in tests where cell permeation and membrane diffusion are involved than in receptor affinity tests.

Pitfall: Collinearity

The number of IDVs in a final QSAR model should be as low as possible. The reintroduction of redundant (collinear) variables improves greatly R^2 or Q^2 during LOO examinations but deteriorates future model applicability (cf. overfitting) [27]. Molecular properties tend to be collinear with MW regardless their units or scales if they are sums of all atomic contributions in the molecule. For instance, the sum of all atomic electronegativities and polarizabilities yield values in their respective units that reflect the number of atoms encoding their element types. In more statistical terms, high correlation with MW can be expected for descriptors, which are based on atom counts in which case it is strongly recommended to inspect the covariance matrices to detect collinearities. In conducting QSAR on compound series, one needs to be careful when using properties that are highly dependent of the side chain size. For instance, logP parallels MW when adding homologous lipophilic/aliphatic fragments due to a limited variety in building blocks, i.e. the larger the molecules the higher their values of LogP.

Errors of Descriptor Calculations

Poor correlation results are eventually due to experimental or computational errors: incorrect program handling, erroneous algorithms like empiric estimation of physico-chemical properties (log P, solubility etc) derived from experimental data [31]. Software bugs are reported in release notes of the latest software versions. They may also come from a "wrong" representation of the chemical structure (e.g. tautomerism, aromatization).

Raw Data Transformation

Values might be unduly reshaped, i.e. pretreatment of measured data through mathematical operations like centering or scaling [27]. The former can be achieved by subtracting the mean whereas the latter helps us in case of variables with large numerical range which would otherwise dominate

over variables with small numerical range [27]. Log(arithmetic) transformation is carried out to reshape so-called lognormal data (e.g. biol. signals) to facilitate the application of parametric methods in statistics, which assume that all data is to some degree normal distributed.

Not Constant "Constants"

Certain electronic properties associated with molecular or fragmental descriptors may change their values in different chemical environments. For instance, protonation states, electrophilicity, or acidity can shift the pK_a : (i) increase pK_a values of weaker basic amino acids in the catalytic site in the presence of other basic groups, or (ii) decrease pK_a values of weaker acidic amino acids with the help of other acidic side chains. Another shift occurs when the substrate docks in a hydrophobic site: A decrease of the dielectric constant leads to stronger electrostatic attraction of oppositely charged atoms or groups within hydrophobic regions. Therefore, a manual inspection or a computer routine is necessary to properly account for the shifted magnitudes at the site of interaction.

Pitfall: Robust Statistical Procedures and "Black Boxes"

The user-friendly interfaces of software have the downside that the programs may be used as "black boxes" that is the user does not have to know much about its functions and limitations. As a result QSAR models can be developed without a detailed understanding of the underlying theories and statistics.

Pitfall: Over- and Under-Determined Equations

In QSAR studies, overfitting occurs if too many IVs, relative to the number of data points, are included in a regression equation. In such cases, regression equations tend to fit the "noise" or errors in the data and, in general, do not yield robust predictions [31]. Sometimes to avoid the presence of undesired descriptors, which would jeopardize a theory, essential descriptors are discarded from the final model leading to underfitting. The decrease of model performance is due to equations describing interrelatedness by insufficient quantities.

Pitfall: Linearity Assumption

Linear relationship between IVs and DV is the *a priori* requirement and can never be inferred from QSAR results. The linearity assumption makes QSAR an outspokenly vulnerable tool. Moreover, it is very difficult to prove with statistical means. Doubts remain and nonlinearity of complex biological data cannot be ruled out in principle [31] (Fig. (1)).

Pitfall: Model Quality and Outliers

Frequently it is assumed that the correlation coefficient R (we prefer its squared form R^2) guarantee or prove excellent QSAR model quality (descriptor choice, predictability of new compounds) or even linearity. R^2 and Q^2 show values between 0 and 1 (unity, 100%), where values greater than 0.5 are interpreted as good correlations and lower one would indicate that there is no link between the chemical structures and their corresponding biological activities. To some extent all this is not correct. Properly speaking R^2 indicates the proportion of the variation in the variables that is explained by

the regression equation. It is said to measure goodness-of-fit. By convention its cross-validated equivalent is called Q^2 . It is obtained through successive removal of data points to predict their response values by the remainder (LOO) [62]. It is frequently reported as a measure of the goodness-of-prediction. However, there exists no correlation between Q^2 in the training set and predictive power (R^2) for the test set in general [41]. To our understanding Q^2 qualifies only the internal accuracy of prediction, i.e. it is limited to the set of used compounds. Experience shows there is no guaranty when applied to new data that is performing an external validation with hitherto unseen compounds [31]. Some authors even speak of a “misuse of Q^2 as a measure of predictivity” [32]. Hence, better criteria for evaluation of predictive ability of QSAR models have been suggested [26,36].

Unusually behaving data points disturb the smooth linear treatment of the data and are known as outliers. They belong to two kinds: (i) true outliers due to systematic errors in the measured response variable or the computed predictor variables. They are undesired but can be avoided (feedback to the experimenter or bug report to programmer); (ii) apparent outliers due to wrong linear equations or true curvature, i.e. unmodeled nonlinearity. They should not be eliminated because they are a most valuable piece of information serving as counter-evidence to explain rules and their limitations. Many studies attempt to obtain the sharpest possible correlation with lowest standard deviations possible. To this end, after identifying outliers they are removed from the data set [31]. The attempt to reach better correlation comes at the expense of predictive capacity because structural diversity is lost or at least reduced. Conversely, data sets with more chemical variation tend to yield results with poorer correlation but improved predictive ability (cf. pitfall about data size).

Pitfall: Starting Geometries in 3D-QSAR

2D-QSAR does not need aligned compounds (superposition) because 2D descriptors are calculated without conformational considerations. In contrast 3D-QSAR is based on a correct alignment of equivalent parts of all molecules under investigation. The active conformation, however, is not necessarily identical to the compound's crystal structure if available. A highly active and rigid molecule could be taken as structural template to fit the others (active analog approach). *Philippe Bernard* and coworkers showed that protein-based alignment of the ligands is a better choice if a 3D target model is available [55-57].

Pitfall: 3D-QSAR Blindness

Several 3D- approaches like CoMFA require meaningfully aligned molecules to map their three-dimensional properties onto a precalculated grid cell or lattice in space. Where the ligands in superposition show structural variations the PLS regression analysis of the grid points weight the highest standard deviations and projects favorable and unfavorable steric and electronic fields in space. In certain molecule collections “binding-relevant” fragments (pharmacophore) may be identical throughout the series and coincide after superposition. However, where chemical variation is absent no fields will be calculated, that is what we call CoMFA blindness (Fig. (2)). Conversely, the PBA approach leads to an odd

superposition of common substructures introducing what we call PBA noise.

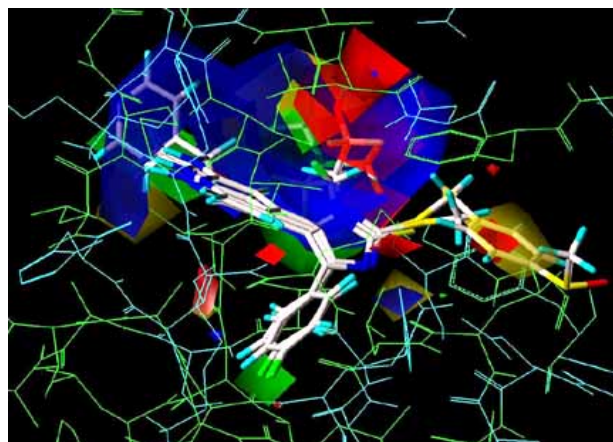


Fig. (2). Illustration of 3D-QSAR blindness: A CoMFA study was achieved after docking of ligands to obtain their protein-based alignment. All ligands possess a common scaffold with its atoms placed in the same location (superposition). Three docked ligands are depicted belonging to three different substitution patterns. Each pattern class folds into distinct side chain geometry. Green and yellow field colors indicate that an increase in bulk favors or disfavors inhibitor activities, respectively. Blue and red fields indicate that an increase in positive charge favors or disfavors inhibition, respectively whereas common substructures do not contribute to the field projection (CoMFA blindness on pyridine nitrogen in blue, leftmost ring). But when they are not meticulously superimposed by LBA or their positions are taken from the PBA procedures, the model mistakenly infers a favorable projection for more volume to enhance activity: here a 4-fluorophenyl ring (lowermost green field). In general terms, aligned ligands from docking may introduce noise to the 3D-grid calculations to produce mere chance fields.

iii. Pitfalls Concerning Model Interpretation

Pitfall: Unrelatedness, also known as the “Correlation Problem”

A good correlation is often mistakenly interpreted as a proof of causality and, unfortunately, this confusion is widespread in QSAR. On the one hand, the basis for the predictive rule of QSAR models lays only on the pharmacological knowledge. On the other hand, achieving significance through MLR or PLS means only “significance on a statistical level” and nothing more. Hence, a significant variable or model can be completely irrelevant on pharmacological grounds.

Pitfall: Chance Correlation

Having some IVs correlating with activity does not necessarily mean that the corresponding feature is directly involved in explaining SAR. Even with highly diverse, very large training sets, apparently correct correlations may occur (by chance) in which case the final QSAR model is seriously wrong because the relevant features only appear in molecules that also contain the wrong features. Subsequently mechanistically irrelevant terms enter the QSAR equations

and jeopardize the study. During backwards MLR starting with very large sets of descriptors chance correlations between biological activity and descriptors may happen unexpectedly for the MLR analyst. To her or his surprise the model is wrong despite its acceptable R^2 coefficient. The larger the IV set to be analyzed by ML [32] and the more trials are performed on the same data set, the greater is the risk to find such random results [27], i.e. the proposed variables apparently correlate but in reality should not. The risk is statistically expressed as significance level p . For instance, the computer practitioner finds by chance an average of up to five false positive in every 100 proposed correlation coefficients at the conventional significance level of $p \leq .05$. In addition, the larger the pool of available variables, the higher is the probability of chance correlations. This is of particular importance in QSAR studies with extremely larger numbers of IVs than the amount of target data [38]. The method of choice is PLS. It reduces chance correlation even under extremely large numbers of values and is therefore implemented in 3D-QSAR, like CoMFA.

Pitfall: Multiple Solutions

MLR (whether preprocessed with PCA or not) leads to a model that does not describe all levels of explanatory complexity in nature. The computed model is a simplification that intends to approximate to reality and perform predictions of new data points along the regression line. Hence, various solutions exist with different dimensionality. If there are tens or even hundreds of IVs at hand, they may be dependent from the others (overlap, redundancy) despite their assuring name. Multiple solutions are always possible and are not necessarily an indicator of wrong models. To this end the so-called kNN-QSAR approach has been developed [24].

Pitfall: Extrapolation and Interpolation

In theory, activities can be predicted either by interpolation, i.e. between the observed data points, or by extrapolation to areas outside the variable levels (value range in the experiments). The activity of the new test molecules is predicted by the established equations but their structures behave differently and are poorly described by the chemical properties of the equations. In more general terms, data points may not be predicted due to missing observations, discontinuous data spread, or nonlinear behavior. The awareness about inappropriate extrapolation into not covered areas of structural information, as well as too low or high an activity, is quite high among the practitioners although rough activity surfaces - or cliffs according to *Maggiore* [33] - may cause unforeseeable Y- response values for smooth value transitions in X- variables. This could make some QSAR predictions for a real external object (a molecule not included during the modeling) 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 risky operation [31,38].

Pitfall: Too Short a Life Cycle of QSAR Study without Validation

The problem with performing QSAR is that sometimes a study may take a long time to complete. Actually, this in

itself is not a problem, but such QSAR studies will never enjoy improvements while feedback about the reliability (percentage of correctly predicted unseen examples) will be missing. Model validation would always be a comforting feature to grasp but cannot be achieved under such circumstances. Few teams decide to test the predictability with hitherto unseen test compounds.

Pitfall: Biased Validation by Unfair Choice of Control Compounds

The unseen examples can be biased: if they had been synthesized according to the conclusions drawn from the QSAR study, then they inherently possess a higher probability to be correct than others not being profiled on this study. The difference is noticeable and should be clearly documented when getting reported.

SUGGESTED WORKAROUNDS TO AVOID THE PITFALLS

Now that we have seen in more detail the potential problems that can arise when carrying out standard techniques in QSAR on a small set of compounds, we can take a sharp look at what actually prevents such problems. The previous section elucidated many, if not all, possible areas of potential errors and downsides of QSAR studies. At this point the reader is aware of their existence and is able to identify them - albeit some of which are unavoidable in principle as outlined in the previous section. For this reason, and to serve as an independent checklist, the proposed solutions are described separately in the forthcoming section.

i. Input Data Preparation and Preprocessing

The researchers can detect some communalities or tendencies by eyesight to corroborate an already existing working hypothesis with known linearity. To this end they can list the congeners with falling activities, and explain why some lie on top and others further down to the bottom of the listing exploiting concepts like steric hindrance, bulk, flexibility, electronic effects, chemical groups, or ionization states. In case of docked molecules that are bound to amino acids certain properties may change, most probably dissociation.

Data Quality

Ideally, all the observations, the measured data points, should come from the same source: experiment protocol, laboratory, material and personnel. At best a reference compound is included in the tests to enable the comparison of measured values. For QSAR studies, it is assumed that the response is a clearly defined observable criterion (endpoint), like half concentration of inhibitions in pharmacology or toxicology [31]. The variation of repeated measurements of compounds and reference during the tests should be kept very small. The modeler should ask the experimenter(s) not only for the mean but rather all individual values of sampled biological tests. The fluctuation between repeated measurements should be at most a third of its total magnitude. Note: controlling the normal distribution of biological response variable is no option because of the limited repetitions of expensive bioassays.

Sampling of Observations and Data Range

The probability of gathering non-representative compounds decreases with the increase in data size. The collection of data should be sufficiently large (> 20 molecules). Another rule of thumb states the difference between highest and lowest biological activity should be 3 to 4 orders of magnitude (Fig. (3)).

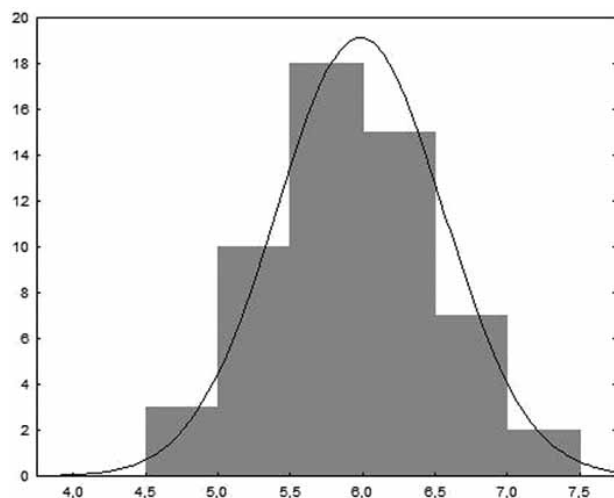


Fig. (3). Typical activity histogram (vertical bars) and the ideal normal distribution (bell-shaped curve). It is less likely that very potent drug candidates are found than weaker ones. Hence the histogram is not ideally symmetrical but is shifted towards the left side. Here, the activity range (horizontal x-axis) spans 3 log units of measured negative logarithmic values (pIC₅₀) while the total number of observations (55 compounds represented on the y-axis) is divided into six intervals (groups) of unit halves. It is known that the normal (Gaussian) distribution of measured values (random variable) occurs as a natural phenomenon under chance conditions with the mean as the most frequent observation. Intriguingly, rational drug research which intends to avoid random hits (trial and error tests) by exploiting knowledge-based decisions shows the same normal distribution behavior. In QSAR, however, it can be distorted by systematic errors or biased events (here: ligand - target knowledge). Fortunately, so-called parametric methods in statistics (based on the normality assumption) are much less vulnerable to such deviations. The reason why statistical studies under chance and more rational conditions show normal distribution alike is that results of both are due to multifactorial experimental influences. They are totally or partially uncontrolled, i.e. not sufficiently understood in its causal existence to be under complete control.

CoMFA Studies

The CoMFA fields should be compatible with existing pharmacophore models. *Ulrich Thibaut* of the former research team of *Gerd Folkers* with colleagues [50] reports on some typical problems with 3D-QSAR analyzing molecular fields. When equivalent or identical substructures or fragments do not match perfectly, unnecessary geometrical variance (noise) enters the field generation process and reflects a wrong assumption of variations in geometries. Fig. (2) illustrates how a volume increase (trifluoromethyl moiety) is wrongly attested to the 4-fluorophenyl side chain. To avoid this artifact, identical parts of docked ligands should lie in

exact superposition. Alternatively, more sophisticated 4D-QSAR could be supplied as successful solutions [13]. Another well-known pitfall associated with CoMFA is the translational and rotational dependence of the molecules with respect to the grid. This problem was tackled by a cross-validated selection of regions guided by an r-square-criterion (Q²) [63] and by GRid-INdependent Descriptors (GRIND) [64]. Topomer CoMFA is also an interesting alternative to avoid alignment caveats with results close to standard CoMFA [65].

Selecting the Members of the Test Set

The training and test sets should represent the entire activity range, as well as the known set of interacting atoms (pharmacophore). A fivefold set of “exclusion/inclusion” rules with parametric, structural, mechanistic, metabolic aspects has been proposed to assess the similarity between chemicals in the training and prediction sets and thusly resolve the model validity and applicability dilemma [37]. This means if the test set is chosen at random, the high quality model generated in the training set does not guarantee a proper prediction in the test set because the latter may not reflect the former. Random selection of test set candidates works well only for homogenous groups with not much variability in structure or activity. Four more strategies are outlined in the following:

- (1) Composition of training and test set by *Polanski*. We refer to Polanski's paper in [40]. “... Changing the training / test set data sampling we are disturbing the statistical 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.”
- (2) Solution by *Clark*. The author devises 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 [66].
- (3) Commonly used solution. Hand selection of one of the best and one of the worst and combining them with randomly chosen candidates forms a test set with good representation and low bias (user preference to avoid outliers).
- (4) Rational selection of training and test sets. To separate a data set into a training and test set innovative solutions are based on sphere exclusion algorithms [41]. They aim at the three following characteristics: All representative points in the test set must be in the proximity of compounds in the training set; all representative points in the training set must be close to those of the test set; the representative points of the training set must be distributed within the entire area occupied by the entire data set. Many solutions have been suggested and the issue is still unsatisfactorily answered for some authors who depreciate the use of a test set in favor of completely different CV methods [67].

- (5) In case of categorical DVs, it is wise to use a stratified random sampling. This procedure will randomly sample each category subset to have representative training and test sets.

ii. Model Generation and Validation

Activity concentrations must be based on molarity units because compounds vary in their MW. Biological signaling and magnitudes of natural response span a large range of values and should be expressed on a log scale. In QSAR studies negative log value (pIC_{50} etc.) are used so that lower concentrations show higher values reflecting the more active compounds.

Descriptor Selection

Descriptors are selected that support the hypothetical mechanism or known pharmacophore model. Several strategies for variable subset selection exist: a stepwise selection of a smaller number out of a larger one (stepwise regressions by backward elimination) or inversely forward selection. The latter preselects and combines potentially important descriptors on intuitive grounds. It may increment their number or keep them constant (but not their combination) in rigid regression analyses [30]. An in-between solution would be to select just one representative of every class of descriptors and go stepwise down or up to find a satisfactory model (cf. correlation below). Other methods are PCA, simulated annealing, Bayesian ranking and automated relevance determination (ARD) as well as evolutionary and genetic algorithms [15,27] next to a most useful listing of general pattern recognition techniques [14].

Certain descriptors are more suitable for cellular or *in vivo* experiments because of pharmacokinetic processes (diffusion, transport, $\log P$, pK_a , $\log D$, MW, V, PSA). Others are more amenable to drug-receptor affinity studies (H-bonds, partial charges, π values, reactive, steric and electronic items). Lists of molecular descriptors commonly used in QSAR and VS have been published ($-\text{pK}_a$, $\log P$, MR, Hammett sigma, Taft's Es, solubility, atomic charges, HOMO, LUMO, H-bonds, MW, MV, PSA, polarizability, DM) [29,30]. Mesomeric, tautomeric or dissociation systems have to be inspected. This advice applies to cases when no significant descriptor terms can be found.

Stereoselectivity and Tautomeric Forms

Problems between structure and activity arise if ligand binding is stereoselective and if some or all of the available data involves racemic mixtures. In extreme cases, the molecular mechanism of action may differ significantly. In addition, the pharmacodynamic properties of the enantiomers are not necessarily the same. Inversions of the asymmetric center(s) may occur during metabolism. Passive diffusion is not stereoselective, but the active transport is. When the biological activity is stereospecific but activity data is available only for the racemate, then the activity of the racemate should not be considered in 3D-QSAR studies. A specific tautomeric form of a compound may also be responsible for the activity in which case strategies have been developed to deal with tautomerism. In a recent study, the tautomeric form, which had the better prediction by a 3D-QSAR model, was considered as the putative biological active form [51].

The tautomeric equilibrium depends on the chemical environment. From an experimental determination in some aprotic organic solvent it cannot be concluded that the preferred tautomer is the same in water. Even the most sophisticated *ab-initio* calculation of tautomeric equilibrium states is useless if conducted in vacuum and not in aqueous solution. Nevertheless, the proper tautomeric form can be crucial for drug action. For instance, a tertiary nitrogen atom can act as a strong H-bond acceptor. In its protonated form, however, it functions as a very strong H-bond donor.

Clustering

Splitting the original set of compounds into several subclasses (clusters) generates local models with narrower similarities and better correlation. However, clustering and exhaustive representation of all possible influential factors is difficult to reconcile in practice: the former enhances accuracy while on the other hand the applicability domain to predict new molecules will narrow [37]. Each cluster should clearly contain more than ten or twelve members to explore the chemical space. The introduction of indicator variables allows the insertion of all clusters within a final equation. All descriptor terms are multiplied by $i = 1$ or 0 in order to turn them on or off. This way they apply only for certain groups of compounds.

Dendogram analyses in statistics can provide a graphical control over automated clustering. These tools are particularly helpful when the selection criteria are not detectable by eyesight. This is the case when subdividing (clustering) very large original sets of molecules into smaller groups with similar features to form series of more homogeneous congeners with less chemical variability, and thereby reduce the number of influential factors [27,30]. Developing satisfactory linear equation systems by automated clustering becomes rather a matter of noncontinuity.

Principle Component Analysis

Once the molecules aligned in rows and the variables stored in columns the table data can be preprocessed by PCA prior to studying correlation with MLR or PLS. PCA eliminates multi-collinearity among predictor variables (redundancy). To this end it suppresses PCs and indirectly reduces the number of descriptors (loaded on suppressed PCs) which in turn decreases complexity to better understand the trends. PCA can also test whether larger numbers of PCs would improve the performance of the model (increasing R values).

Indicator Variables

Besides the elimination of redundant variables by PCA a second solution exists by introduction of an indicator variable to distinguish chemical classes (cf. clustering, or FA above). For instance, upon clustering congeners according to chain lengths, the partial correlation between MW and $\log P$ disappears.

Multiple Linear Regression

After a successful refinement by PCA there is still need to change the combinations of remaining variables by MLR. This is because only MLR looks for the best (least squares) fit to the target values (DV). In consequence, a single final best answer is not very likely. A solution based on a set of a few preselected descriptors (forward stepwise MLR) will

probably look different from one started with all possible descriptors (backward stepwise MLR). The objective is to find those linear equations with the smallest number of predictor variables possible by maintaining a high R^2 retaining as much as possible significant variables. To this end beta as well as R^2 are inspected in preliminary models.

Beta Analysis

During the search of descriptor combinations the beta coefficients of the linear regression equations may vary within a series of compounds due to multifactorial dependency on the selected descriptors. The estimated regression coefficients (beta or β) representing the relative contribution of each IV in the prediction of the DV are compared. Beta is the normalized regression coefficient (b) of an IV in a linear equation. In its simplest form of SLR the equation is composed of just one descriptor term (bx): $y = a + bx$. In MLR most of, if not all of the IVs included in the final model should be statistically significant otherwise the exploitation of MLR or PLS in QSAR is senseless, mathematically speaking. It is of utmost importance, however, to understand that statistical significance is only a mere numerical property. Thus, whether a statistically significant linear equation would also reflect relevance in drug design is a question that needs expertise in Medicinal chemistry. It is imperative, that the selected variables should provide a biochemical sound interpretation in accordance to some theory. In order to avoid spurious results, models from literature should be compared to the present study.

Examination of R^2 or Q^2

They indicate how well the model fits the data. For instance, an R^2 of 0.9 indicates that we have accounted for 90% of the original data (variability) with the variables specified in the QSAR model and with a probability by the significance level p.

The *internal validation* of the QSAR model is conventionally performed through a test set. Ideally, *external validation* with hitherto unseen, new molecules would follow the study [31]. If internal (lateral) model validation is not possible, the fact should be stated explicitly. As much more effective validation strategies *Tropsha*, *Gramatica*, and *Gumbar* propose rigorous measures like leave-many-out (LMO) and other robust validation techniques, like y-randomization, test set prediction, or a concept of applicability domain for QSAR developers together with a benchmark dataset [26]. Several metrics were suggested to estimate the application domain by [68].

Significance Test

Assuming a normal distribution a statistical significance test (representativeness of the observed response for the entire population) can be performed with the following criteria: sum of squares (SS), Means of squares (MS), F-test, and p-level estimating error probabilities. The more IVs, the better R^2 of the present QSAR-model even if the variables are not significant. An acceptable solution shows a combination of IV which avoids the presence of various descriptors of the same type. Upon insertion of additional and relevant (new, independent) descriptors it always yields better correlation. However, a compromise has to be arranged between the descriptor number and R^2 since the desired lowering of the

former also lowers the value of the latter. Inversely, too many descriptors increase R^2 values further and the model becomes overfitted.

Chance Correlation Test

An appropriate test assigns random numbers to the columns (variables) and rows (response) to see if high R^2 or Q^2 of LOO are still obtained despite the fact that in reality there is no relationship, trend or rule based on such random values for that particular QSAR model. Alternatively, y-scrambling can be applied (with y values randomized). If either test results in high R^2 or Q^2 then the "final" QSAR model must be rejected [26]. To circumvent the limitation of CV / LOO that it provides no statement of the statistical significance of the estimated predictive ability the response permutation tests can be useful, i.e. random shuffling of original y-values. The final QSAR model is then fitted to various permuted y-data in order to evaluate the real Q^2 in light of a distribution of Q^2 values of randomly reordered response data [27].

Standard Plot or Results

Graphical representation of model performance for the entire data set. Ideally, all corresponding points form pairs of observed and predicted values lying on a straight line that goes through the origin (0,0) under an angle of 45° in a x,y diagram.

Residual Test

Outliers can be pinpointed in so-called normal probability plots using standardized (to unity 1) values of residuals [27]. The diagnostic tool is specifically designed to examine the scatter of residuals. The scatter plot verifies the deviations of the data points from the regression line. Residuals are the calculated differences between the predicted and observed values of the response variable. The smaller the variability of the residual values around the regression line relative to the overall variability, the better the predictive power is. Recently, the test was found to be unreliable especially in higher dimensional models [67].

Inspection of SEE

The prediction errors or deviations ($Y-Y' = d$) can be drawn as lines from the x,y points to the x,y' points on that line. The standard error of estimate is the "average" length of aforementioned prediction error lines on all x,y points (residuals). It should be kept small regarding the absolute values of typical y (cf. Residual test).

Verification of Normal Distribution

Because of limited means response measurements (individual values) are normally not replicated, and no sample distribution is at hand. Intriguingly, activity histograms of drug candidates show normal distributions when their groups are sufficiently large (Fig. (3)).

Outlier Test

It deals with examination of the predicted and residual values for each data point in the preliminary model. Knowledge about outliers and their properties help the understanding of the studied relationships. Some data points may appear as outliers because of an unlucky descriptor combination. Certain outliers reflect random errors during experi-

mental measurements of response and should not be eliminated in the model through so-called brushing methods which manage outliers in scatter plots. After fitting a regression equation, one could examine the predicted and residual scores (cf. Residual test). Particularly extreme outliers tend to jeopardize the results and lead to erroneous final conclusions [67].

Correlation coefficients can become substantially inflated or deflated if extreme values are present in the data; so one should examine the distribution through scatter plots. Alternatively, the normal distribution histogram can be examined if an outlier falls outside the mean of ± 3 times the standard deviation. Anyway, correlation data cannot conclusively prove causality. Thus, the major conceptual limitation of correlations is that you can only ascertain relationships, but never be sure about underlying causal mechanism (cf. step III below).

Guha and *Jurs* suggest that compounds that are very similar in a training set, generally exhibit smaller residuals and standard errors of prediction. The same authors suggest that the cut-off value must be selected so that the outliers are approximately 20 - 30% of the whole data set [69].

The computer practitioner should be aware of the potential activity cliffs in the data, and - if possible - make a distinction between activity cliffs and true outliers [33]. Note that apparent outliers may be a function of the particular representation of the molecules, i.e. set of descriptors to construct the models. In other words, "outliers in the data may not be due to statistical fluctuations or to measurement errors but rather may reflect the presence of activity cliffs." [33]. Computational strategies have been developed recently to identify and quantify activity cliffs and consensus activity cliffs in data sets [42,43].

Cluster Analysis

Group together typical congeners to confirm the hypothesis. Mark and keep the outliers until you can explain their positions as exceptions to the rule. Refinement of the final model to improve R by eliminating the outliers is legitimate, if reported in the paper. The modeler is advised to sacrifice R optimization in favor of a more chemically diverse training set by leaving the outliers in the final model. In turn she or he will be rewarded with a gain in predictive power. In physical chemistry an excellent correlation is expressed as $R^2 > 0.90$ but may be much lower in complex biological processes ($R^2 \geq 0.4$). Increasing standard deviations (sd) are due to missing descriptors inclusion of which would narrow again the range of uncertainty in the correlation. The modeler has to define for which structural cases the established QSAR rule(s) can be applied to avoid invalid applications.

Linearity Test and Cross-Validation Test

Because it is extremely difficult to prove mathematically based on first principles, a common practice is to visualize the linear relationships between response and descriptors by so-called bivariate scatter plots showing points of x,y variable pairs. Despite their convincing appearance they do not prove linearity (Fig. (1)). The linear models should also be subject to a cross-validation procedure such as LOO or bootstrapping to ensure that they are neither overfitted nor underfitted [69,70]. Notice if the "test set" contains a single mole-

cule then the automated iterative manifold cross-prediction for all molecules using the models derived from all with the exception of this one molecule is called a leave-one-out cross validation (LOO / CV). Logically, its coefficient Q^2 is slightly lower than R^2 (not validated regression) for the same data set.

Overfitting Test

The larger the number of IV in MLR, significant or not, the better R^2 [31]. Too large a number of variables yield poor predictive power despite an excellent R^2 very close to unit 1. The model is only useful for its data points, but not to unseen and new examples. Concerning MLR techniques, fitting an equation to correlate biological activities to structural data is only correctly executed when the number of experimental observations (molecules) is - at least - equal or - even better - greater than the number of chosen parameters. As a rule of thumb, the commonly accepted ratio to avoid over fitting is to have about five data points (or more, so the better) for each IV in the final equation.

Test of Uniqueness

The collinearity between the descriptors has to be studied by examination of the variable correlation matrices. Variables are highly correlated (covariance) when their cell coefficient (column intersection) is close(r) to -1 or +1. Particularly, those descriptor pairs with high collinear relationships ($R^2=0.9$) can be easily detected and eliminated upon inspection of the data matrix by any statistics program. Reports of unacceptable correlation coefficients between variables (in the covariance matrix) start from threshold 0.4 up to 0.9 in the literature [31].

iii. Model Interpretation (Output Analysis)

Causality

Whatever the result of the QSAR study will be, the conclusion about causality can be drawn neither from the correlation alone nor any statistical significance. On the contrary, the causal relationships between IDV and DV are the *a priori* bases *sine qua non* (that is their existence is beyond doubt and without them nothing goes). The QSAR models must be based on the knowledge and expertise in those particular fields of life science but not on pure statistics.

Non-Transparent Descriptors

Ideally, the terms in the final equation should allow a physicochemically transparent and mechanistically interpretable answer [31].

Predictive Power

Despite the claims of many authors considering a high R^2 (> 0.8) an indicator or even an ultimate proof that a highly predictive QSAR model has been found, the truth is in order to define the relative importance of the QSAR models, they should carefully reconsider other biological hints and scientific background information on the series under investigation.

External Validation

One of the ultimate goals in QSAR modeling is the prediction of new compounds. It has been demonstrated that

validation procedures, for example, producing high Q^2 are not necessarily predictive [36]. Hence, the true predictive quality of a QSAR model can be best achieved by external validation upon prediction on unknown compounds, i.e. those that were not used in either training or test sets.

3D-QSAR and CoMFA

Especially in 3D-QSAR models, when larger numbers of components (>10) are necessary to achieve good statistical results ($R^2 > 0.7$), the alignment rules (geometrical variation) and the training set (structural variation) should be revised. If the structural redundancy is poor / high, the Q^2 of LOO/CV may drop / rise and the models are rejected / accepted on unjustified grounds. If the data set contains many structurally similar compounds, the series could be halved to form two sets of equal proportion for training and testing in order to reduce redundancy in the training set [71]. Variable scaling is not recommended in CoMFA studies because the (desirable) reduction in variance (of even just one X variable) generates weight artifacts on distant grid points and strange CV results (Q^2). Negative Q^2 values may occur if the model depends on just one compound in which case the inspection of the individual CV residuals ($=0$) is recommended [71].

Multidisciplinary Work and Groups

Despite the ease to follow a protocol, QSAR remains a daunting task, and often conducted by a multidisciplinary group of co-workers. In a virtual world of computing, the procedure transforms chemistry into numerical language and hands it over to mathematics. Then after statistical analysis it returns numerical tendencies to some extent that in turn are interpreted as drug building rules on a biochemical level. According to the educational background, the chemistry, statistics, informatics or the biochemistry is more of a black box than desirable. In consequence unfamiliar limitations are more likely to be overlooked; for instance when the organic chemist synthesizes in non aqueous / low temperature milieu knowing about the hydrolytic / thermolytic decay of her or his products and the biologist performs *in vitro* tests exposing the compounds to water milieu / room temperature [53].

DISCUSSION AND CONCLUSION

Medicinal chemists often rely on knowledge-guided research in a way that their decisions are guided by some theory or working hypothesis. More statistically-minded researchers, however, undertake what they believe is a more unbiased approach by systematic (stepwise) forward or backward regression. This situation leads us directly to the question whether or not both approaches could deliver the same results, provided the use of the same data and tools by two independent groups. *Stephen Johnson* indirectly gives us a negative answer in his work "The trouble with QSAR" [34]. He describes QSAR as an approach to "... identify a few molecular properties critical for activity from a nearly infinite pool of detailed possibilities. ... The problem with this approach is that there are typically many possible solutions that yield approximately equal statistical measures of quality. ... Each of these equivalent solutions, however, represents a hypothesis regarding the underlying physical or biological phenomenon."

With experience gathered in model validation, it has turned out that QSAR equations often fail to predict the activities of compounds not considered during the model generation. The finding is even more intriguing because of the enormous effort to obtain optimal conditions in the quality of the input data and validation of the statistics. Some of these issues are also associated with the form of the response used. Sometimes large activity changes are observed quite unexpected from linear extrapolations. Outliers in a linear model might become completely well behaved if one uses a nonlinear model that is more appropriate for fitting the data. For this reason some authors question the underlying concepts of QSAR.

A related issue resides in the silently made assumption that considers the structure-activity landscape as a smooth function while - more likely - it is populated with activity cliffs [33,34,42,43]. If this is the case, the practitioner confounds valid data points and outliers as a consequence of unexpected but true changes in molecular behavior. "The common practice has been to select the model with the best fitness function score and predict a small group of observations that were withheld at the beginning. All too often, the model development process stops here or worse; the validation set is poorly predicted and models are iteratively tested until one predicts this set of compounds well." [34]. In fact, QSAR normally ends at that stage because waiting for the next generation candidates to serve as hitherto unseen test molecules (external validation of the final QSAR model) is not acceptable for many research teams because of the scheduling, funding, and ranking pressure (publish or perish). The temptation is there to react with iterative testing, as cited, or even worse: partial data omission for the sake of "getting published". QSAR studies are published prior to external validations (with unseen examples) that would prove their real potentials in R&D processes. On the other hand authors meet difficulties to get their work published showing bad linear correlation and unexplained outliers. This is in part due to the silently made assumptions and partly due to the misleading scope of some QSAR studies (Tables 4 and 5). Four valid categories have been well established amidst editorial boards of journals dealing with QSAR models [72]:

- (1) Methodological papers, describing new algorithms, and conceptual approaches, higher integration for VS of machine learning devices in the field of artificial neural networking.
- (2) Performance comparison and benchmarking analysis to compare the speed of different methods or outcome of different data sets.
- (3) Applications, where QSAR models support the decisions made by Medicinal chemists in their effort to find better drug candidates.
- (4) The input data comes from the same work group or an affiliated site. If this is not the case, the authors should present an added value to the original data set, which they took from the literature. For instance, show how and where the synthesized and tested compounds bind or why some are less active than others. In each of the four cases it is important that the authors mention the criteria to choose descriptors and

Table 4. Listing of Statistical Prerequisites for MLR and their Implications in QSAR Studies. Only the First Two Conditions are Fulfilled in Principle. The others Follow in Increasing Order of Inconsistencies in Practice [27]

X is fixed: IVs are computationally estimated; theoretically without random errors, i.e. repeated calculations yield the same values (rounding off and respecting precision limits).
Y is not fixed; DV is a random variable, it is experimentally determined and larger samples follow normal distribution.
Each data point <i>y</i> of the response variable (DV) is independent from others, i.e. the measured value of one data point <i>y</i> does not influence the measured values of the other <i>y</i> data.
X and Y are free of systematic errors, IVs: bias, wrong concepts (no single binding mode, tautomers, isomers), mechanisms (no common binding mechanism), work hypothesis (targets, prodrugs, partial a(NTA)gonism); DVs: wrong instrumentation, work protocols, interpretation.
Homogeneity assumption: the mechanism of influence of X on Y must be the same.
Dilemma of application domain and representativeness assumption: huge structural variety tends to increase predictive ability for new structures at the expense of good correlation (statistical quality) while more homogenous sets of molecules (with fewer outliers) decrease predictability but sharpen correlation (R^2 approaches to the optimal value of 1.0).
Dilemma of multicollinearity and assumption of uncorrelated variables to avoid the true loss of prediction power by gaining "apparent" prediction power (chance correlation). Adding more and more terms in the equation, the fit can be made arbitrarily close to unity 1.0 (overfitting).
Dilemma of insufficient data spread to see true nonlinearity: Inappropriate models arise when the response variable <i>y</i> (DV) appears to be a linear function of one or more <i>x</i> variables (IVs). Linearity must be assumed but cannot be proven by MLR itself, otherwise MLR and PLS cannot be applied.
Dilemma of outlier handling: how to distinguish between outliers generated by experimental errors, systematic errors, activity cliffs or unmodeled true nonlinearity? Removing outliers in linear models to enhance the fit (R^2) should be deprecated so that the model is not overtrained, i.e. deteriorates its performance with new compounds.
Dilemma of missing quality criteria: R^2 and Q^2 fail to characterize fit and predictive power.

Table 5. Guidelines for Presenting QSAR Work to Enhance Comparability of Data and Results [71]. Thibaut, Folkers, Klebe, Kubinyi, Merz and Rognan Suggested Publishing the Following Aspects

Methods to optimize geometry for 3D-model generation.
Methods used to generate atomic partial charges.
Methods used to generate the starting geometries or adopt active conformations at the interaction site.
The superposition criteria and structural alignment methods.
Variable scaling or weighting of fields.
The cross-validation to verify the validity of the predictive capabilities.
The number of CV groups and components.
The number of CV groups should be equal to the number of rows for LOO – CV methods.
An optimal number of components should be chosen as among those combinations of significant components which need the smallest number of components and still preserve satisfactorily good statistical results (R^2 , Q^2).
Describing outliers and justifying their exclusion in the final model.
Describing the selection of the final combination of variables.
Listing of all structures with observed and predicted response values.
Listing of statistically relevant data like R^2 , Q^2 , F-values, and if possible the number of cases (molecules) and standard errors of the response variable. In case of CoMFA studies: the grid box dimensions, grid point distances, charges and types of probe atoms to calculate the grid point values. The contour maps of the steric and electrostatic fields should be displayed or at least nongraphically discussed.
Offering the coordinates of aligned molecules.

interpret the model. This is especially relevant for the aforementioned approach of mere statistical maneuvering [21,22], and to a minor extent for research guided by Medicinal chemistry reasoning [23-25].

Since neither methodic difficulties nor negative results can easily be published, "me too success stories" avoiding words of criticism about data, methods, or linear equations possess more likelihood to enter the peer-reviewed scientific literature (acceptance bias). Another risk lies in that some works presenting excellent statistics reduce the importance

of interpreting the models (expectation bias). Besides, the established peer-review process of journal papers by no means guarantees that the invited reviewers' ethics and knowledge about the specific research field and applied methods are adequate.

In the early 90's *Thibaut et al.* [71] formulated a set of aspects authors should consider when publishing QSAR studies (Table 5). Ever since, their suggestion is necessary to standardize the contents of QSAR literature in what is also a concern of editorial policies (Table 6). Despite their efforts,

Table 6. Illustration of Some Prerequisites for Publishing QSAR Studies According to the Policies of Two Editorial Boards (Extracted from *Journal of Chemical Information and Modeling* and the *Journal of Medicinal Chemistry*) [73]

State what is novel about the present QSAR study with respect to methodology/theory and/or the findings derived from the experimental data set.
A new method/theory should be compared and validated against at least one other common method in concert with a published larger data set.
Use a supplemental material section as Supporting Information for the readers to make molecular structures available.
Model validation based on data not used in the training set must be provided.

there have still been insufficiently documented studies published.

On page 99 in *Brandt's* textbook on data analysis ([49] in Table 3) *D. J. De Solla - Price* is cited who discovered with statistical means that the number of scientific publications with practically identical results is distributed according to a Poisson distribution [73]. For instance, if the same discovery coincided twice (or four times), then the number of observed literature cases was found to be 179 (or 17), which comes close to 184 (or 15) predicted by a Poisson equation. *Brandt* states that *De Solla* concluded himself that scientists are more interested in getting their work published than reading those of others. In this regard, the current publishing and peer-reviewing is also at stake. It is beyond the scope to investigate the extent of such insufficiencies [74].

After five decades of QSAR development, some scientists in the field of drug development are skeptical [28,32] and others reluctant to exploit it because the pitfalls are so manifold (Table 7): inadequate descriptors, outliers due to either experimental errors, multiple binding modes or activity cliffs, non-linearity over a wider range, data manipulation, missing opportunities for external validation, limited predictability and, in extreme cases, the models show just self-explanatory elements: "... QSAR disappoints in cases where they are applied to data sets determined after the QSAR models were constructed ..." [33]. Yet, for others it is seemingly a virtuous means to combine mathematics and chemistry in a graceful way [21-25,74], and sometimes to develop innovative combinations of chemometric concepts [18,37,39,54,55,64,75,76].

Only in rare cases when the pharmacophore has sterically unconnected or electronically unconjugated parts the com-

puter practitioner is capable of identifying certain additive structure – activity patterns. Instead of all descriptors interacting together some of them become prominent factors of an *in situ* Factorial design. Their effects are independent of the presence and levels of the other factors leading to a (partial) response control and predictability similar to an *a priori* Factorial design (with full control).

QSAR is not flawless since it is based on the linearity assumption in order to apply MLR and PLS techniques which are implemented in its software. Other more reliable statistical techniques (e.g. FA identifies interaction of variables) cannot be applied because QSAR data constitute an undesigned experiment. In consequence various final QSAR models are always possible. In general terms the relationships between chemical and biological material is less satisfactorily described by empirical equations containing linear terms because linearity is a mathematical prerequisite but not a law of nature. Today, *Auguste Comte's* historical and philosophical speculation lies on more scientific grounds. QSAR is only indicated to study causation (cause and effect relationships) under one condition: linear relationship must exist a priori based on some external knowledge (Table 4). The early success and historic popularity of QSAR is partly due to the availability of few tested compounds and a handful of descriptors most of which have been experimentally traceable: a most lucky circumstance that explains why linearity and explanatory relevance were easily found in that tiny observation window. In cell culture studies, LogP, pK_a, logD, MW, or V were usually chosen to parallel biological effects in dependence of response-limiting diffusion and transport phenomena converting QSAR in a straightforward tool (few trials, good hits) [77]. Nowadays, the picture has

Table 7. Commonly Unsolved Problems Leading to the QSAR Dilemma

1) Data size and composition Huge structural variety tends to increase predictive ability for new structures at the expense of good correlation (statistical quality) while more homogenous sets of molecules decrease predictability but sharpen correlation (R^2 approaches unity 1).
2) Internal and external validation and the eternal loop of "final" models There is not <i>one</i> final best solution (descriptor combination) for the data treated with PCA, MLR or PLS in QSAR studies. In addition, internal validation (splitting of the original data set into a training and test set) is not as reliable or effective as testing of the final equations with hitherto unseen compounds (or truly inexistent). Internal validation is common practice to get immediately published, that is without waiting for obtaining truly new compounds for external validation. Conversely, a "final" unpublished model becomes a preliminary one (and modeling starts all over again) when the introduction of formerly "external" molecules lead to model improvements.
3) Linearity assumption It can neither be proven by excellent correlations nor inferred from statistical significance. The validity of the final QSAR equation is not a statistical matter but a question of external knowledge. QSAR applies it without proof. Luckily data can appear linear (cf. 2.)
4) Outlier handling Outliers can be either disturbing or helpful to find trends and limits. In principle the possibilities of outliers to exist can be manifold: unmodeled but true nonlinearity, rough linearity with activity cliffs, uneven data distribution (missing information), explanatory complexity (cf. 2.) and imprecise data determination due to unavoidable random errors or undetected systematic errors. Outliers of the latter should be discarded.

drastically changed: hundreds and thousands of cryptic properties have invaded the QSAR software, created to improve correlations but leaving serious doubts on their chemical feasibility or biological relevance for drug development.

All told, the results from this review extend the literature attesting to the known QSAR pitfalls further scrutiny with respect to possible problems and solutions.

ACKNOWLEDGEMENTS

The authors are very grateful to Prof. Dr. *Joachim K. Seydel*, Borstel, Germany, for discussing his QSAR work, and Prof. Dr. *G. M. Maggiora*, College of Pharmacy & BIO5 Institute, University of Arizona, Tucson, Arizona, USA, for most valuable advice, which we incorporated in many parts of the manuscript; to *Chemical Computing Group* for support, as well as *ChemAxon* for providing free academic modeling tools. We also would like to acknowledge Dr. *P. H. Hernandez*, VIEP – BUAP for supporting the temporal work of TS at Universität Tübingen, Germany. Thanks are also to CONACYT – Mexico City (2007/52639) for project support. J. L. M-F. and K. M-M. acknowledge the State of Florida, USA, for funding. We thank Kyle Kryak for reading the manuscript.

GLOSSARY AND LIST OF ABBREVIATIONS

ADMET	= Adsorption, distribution, metabolism, excretion, and toxicity; describing drug destiny in pharmacokinetics
ANOVA	= Analysis of variance
Beta	= Regression coefficient β is the normalized regression coefficient b in MLR
Bias	= Unfair treatment of data, or unfairness of their presentation
CV	= Cross validation; assessment of predictive power in PLS, e.g. LOO
DV	= Dependent variables, biological activity or response or target variable (on the Y-axes)
FA	= Factor analysis
Factor	= Assigned (fixed values), and independent predictor variable in FA
IV	= Independent variables or regressors, structural descriptors (on the X-axes)
Levels	= Values loaded on (assigned to) a factor
LOO	= Leave-one-out technique or jack-knifing; omission of one (test) compound to predict its known activity from the reminder (training set)
MLR	= Multiple linear regression, like SLR but with more than one x variable
MR	= Molecular Refraction, a surrogate for bulkiness
MW	= Molecular weight (mass)

Noise	= Useless data which the processing cannot get rid of
Outlier	= Data point not on the regression line; exception to the rule or trend of the rest of the data
PC	= Principal Component
PCA	= Principal Component Analysis
PLS	= Partial least squares (procedure) is an extension to MLR
Q, Q ²	= Cross-validated R or R ²
R, R ²	= Correlation coefficient, R ² is sometimes called determination coefficient, R ² expresses the amount of common variation between two variables.
Random	= error due to chance events i.e. natural diversity; as well as r. variables like DV following a normal distribution when sampled
Sample	= Repeated measurements of the same experiment (trial, run) showing natural variation (sampling and reading errors, material imperfections, biological diversity)
SD, SEE	= Standard deviation, or standard error of estimate
SLR	= Simple linear regression
SBA, PBA	= Structure based / protein based alignments (superposition) of compounds / ligands
Systematic	= errors due to defective experimental materials, instruments or protocols, as well as software bugs, wrong algorithms etc.
QSAR	= Quantitative-Structure Activity Relationships
QSPR	= Quantitative-Structure Property Relationships
SAR	= Structure-Activity Relationships
Undesigned	= Such experiments have data from uncontrolled combinations of descriptor variables; they are thought to be independent
VS	= Virtual screening
X	= Independent variable, descriptor, regressor or parameter
Y	= Dependent or target variable, biological activity, response

REFERENCES

- [1] Seidel, J.K.; Schaper, K.J. *Chemische Struktur und biologische Aktivität von Wirkstoffen*; Verlag Chemie Weinheim: New York, **1979**, pp. 1-11.
- [2] Kubinyi, H. *Quant. Struct.-Act. Relat.*, **2002**, *21*, 348-356.
- [3] Johnson, M.; Maggiora, G.M. *Concepts and applications of molecular similarity*, John Wiley & Sons: New York, **2006**.
- [4] Willett, P.; Barnard, J.M.; Downs, G.M. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983-996.
- [5] Andrew G.C.; Graham R.W. *Perspectives in Drug Discovery and Design*, Stevenage Netherlands: UK. **1998**, pp. 321-338.

- [6] Engel, T. *J. Chem. Inf. Model.* **2006**, *46*, 2267-2277.
- [7] Willett, P. *J. Med. Chem.*, **2005**, *48*, 4183-4199.
- [8] Willett, P. *Drug Discov. Today*, **2006**, *11*, 1046-1053.
- [9] Medina-Franco, J.L.; Maggiora, G.M.; Giulianotti, M.A.; Pinilla, C.; Houghten, R.A. *Chem. Biol. Drug Des.*, **2007**, *70*, 393-412.
- [10] Martínez-Mayorga, K.; Medina-Franco, J.L.; Giulianotti, M.A.; Pinilla, C.; Dooley, C.T.; Appel, J. R.; Houghten, R.A. *Bioorg. Med. Chem.* **2008**, *16*, 5932-5938.
- [11] Breneman C.M.; Bennett, K.P.; Embrechts, M.J.; Bi, J.; Demiriz, A.; Lockwood, L.; Momma, M.; Sukumar, N. *221st National Meeting, American Chemical Society, San Diego*, **2001**
- [12] Winkler, D.A. *Mol. Biotechnol.*, **2004**, *27*(2), 138-167.
- [13] Hopfinger, A.; Wang, S.; Tokarski, J.; Jin, B.; Albuquerque, M.; Madhav, P.; Duraiswami, C. *J. Am. Chem. Soc.*, **1997**, *119*, 10509-10524.
- [14] Livingstone, D.J. *Predicting Chemical Toxicity and Fate, CRC Press LLC: Boca Raton, FL*, **2004**, pp. 151-170.
- [15] Winkler, D.A. *Briefings Bioinform.*, **2002**, *3*(1), 73-86.
- [16] Albuquerque, M.; Hopfinger, A.; Barreiro, E.; De Alencastro, R. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 925-938.
- [17] Santos-Filho, O.; Hopfinger, A. *J. Comput-Aided Mol. Des.*, **2001**, *15*, 1-12.
- [18] Ravi, M.; Hopfinger, A.; Hormann, R.; Dinan, L. *J. Chem. Inf. Comput. Sci.*, **2001**, *41*, 1587-1604.
- [19] Krasowski, M.; Hong, X.; Hopfinger, A.; Harrison, N. *J. Med. Chem.*, **2002**, *45*, 3210-3221.
- [20] Hong, X.; Hopfinger, A. *J. Chem. Inf. Comput. Sci.*, **2003**, *43*, 324-336.
- [21] Ramírez-Galicia, G.; Garduno-Juarez, R.; Hemmateenejad, B.; Deeb, O.; Deciga-Campos, M.; Moctezuma-Eugenio, J.C. *Chem. Biol. Drug Des.*, **2007**, *70*, 53-64.
- [22] Ramirez-Galicia, G.; Garduno-Juarez, R.; Hemmateenejad, B.; Deeb, O.; Estrada-Soto, S. *Chem. Biol. Drug Des.*, **2007**, *70*, 143-153.
- [23] López-Vallejo, F.; Medina-Franco J.L.; Hernández-Campos, A.; Rodríguez-Morales, S.; Yépez, L.; Cedillo, R.; Castillo, R. *Bioorg. Med. Chem.*, **2007**, *15*(2), 1117-26.
- [24] Medina-Franco, J.L.; Golbraikh, A.; Oloff, S.; Castillo, R.; Tropsha, A. *J. Comput. Aided. Mol. Des.*, **2005**, *19*(4), 229-42.
- [25] Medina-Franco, J.L.; Rodríguez-Morales, S.; Juárez-Gordiano, C.; Hernández-Campos, A.; Castillo, R. *J. Comput. Aided. Mol. Des.*, **2004**, *18*(5), 345-60.
- [26] Tropsha, A.; Gramatica, P.; Gumbar, V.K. *QSAR Comb. Sci.*, **2003**, *22*, 69-77.
- [27] Eriksson, L.; Jaworska, J.; Worth, A.P.; Cronin, M.T.D.; McDowell, R. M.; Gramatica, P. *Environ. Health Perspect.*, **2003**, *111*(10), 1361-1375.
- [28] Stouch, T.R.; Kenyon, J.R.; Johnson S.R.; Chen, X.Q.; Doweiko A.; YiLi. *J. Comput.-Aided Mol. Des.*, **2003**, *17*, 83-92.
- [29] Zvinavashe, E.; Murk, A.J.; Rietjens, I.M.C.M. *Chem. Res. Toxicol.*, **2008**, *21*(12), 2229-2236.
- [30] Scior, T.; Bemard, P.; Medina-Franco, J.L.; Maggiora, G.M. *Mini-Rev. Med. Chem.*, **2007**, *7*, 851-860.
- [31] Cronin, M.T.D.; Schultz, T.W. *J. Mol. Struct. (Theochem.)*, **2003**, *622*, 39-51.
- [32] Doweiko, A.M. *J. Comput.-Aided Mol. Des.*, **2008**, *22*(2), 81-89.
- [33] Maggiora, G.M. *J. Chem. Inf. Model.*, **2006**, *46*, 1535-1535.
- [34] Johnson, S.R. *J. Chem. Inf. Model.*, **2008**, *48*(1), 25-6.
- [35] Guha, R.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.*, **2005**, *45*, 65-73.
- [36] Golbraikh, A.; Tropsha, A. *J. Mol. Graph. Model.*, **2002**, *20*, 269-276.
- [37] Dimitrov, S.; Dimitrova, G.; Pavlov, T.; Dimitrova, N.; Patlewicz, G. *J. Chem. Inf. Model.* **2005**, *45*, 839-849.
- [38] Polanski, J.; Bak, A.; Gieleciak, R.; Magdziarz, T. *J. Chem. Inf. Model.*, **2006**, *46*, 2310-2318.
- [39] Novellino, E.; Fattorusso, C.; Greco G. *Pharm. Acta Helv.*, **1995**, *70*, 149-154.
- [40] Polanski, J.; Bak, A.; Gieleciak, R.; Magdziarz, T. *J. Chem. Inf. Model.*, **2006**, *46*, 2310-2318.
- [41] Golbraikh, A.; Shen, M.; Xiao Z.; Xiao Y.D.; Lee K.H.; Tropsha, A. *J. Comput.-Aided Mol. Des.*, **2003**, *17*, 241-253.
- [42] Guha, R.; VanDrie, J. H. *J. Chem. Inf. Model.*, **2008**, *48*, 646-658.
- [43] Medina-Franco, J.L.; Martínez-Mayorga, K.; Bender, A.; Marín, R.M.; Giulianotti, M.A.; Pinilla, C.; Houghten, R.A. *J. Chem. Inf. Model.*, **2009**, *49*, 477-491.
- [44] Wermuth, C.G. *The Practice of Medicinal Chemistry*, Academic Press, **2008**.
- [45] Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*, 2 Volumes, Wiley-VCH, **2008**.
- [46] Mannhold, R.; Kubinyi, H.; Folkers, G. *Molecular Drug Properties Measurement and Prediction*, Wiley-VCH, **2007**.
- [47] Livingstone, D. *Data Analysis for Chemists: Applications to QSAR and Chemical Product Design*. Oxford University Press, **1996**.
- [48] Gilbert, N. *Statistics*. Saunders College Publishing; Philadelphia, PA, **1981**.
- [49] Brandt, S. *Data Analysis: Statistical and Computational Methods for Scientists and Engineers*, Springer, **1998**.
- [50] Bolton, S.; Bon, Ch. *Pharmaceutical Statistics, Practical and Clinical Applications (Drugs and the Pharmaceutical Sciences); Informa Health Care*, **2009**.
- [51] López-Vallejo, F.; Medina-Franco, J.L.; Hernández-Campos, A.; Rodríguez-Morales, S.; Yépez, L.; Cedillo, R.; Castillo, R. *Bioorg. Med. Chem.*, **2007**, *15*, 1117-1126.
- [52] Hauser, D. R.; Scior, T.; Domeyer, D.M.; Kammerer, B.; Laufer, S.A. *J. Med. Chem.*, **2007**, *50*(9), 2060-2066.
- [53] Scior, T.; Garcés-Eisele, J. *Curr. Med. Chem.*, **2006**, *13*(18), 2205-2219.
- [54] Bernard, Ph.; Kireev, D.B.; Chretien, J.; Fortier, P.L.; Coppet L. *J. Comput.-Aided Mol. Des.*, **1999**, *13*, 355-371.
- [55] Pintore, M.; Bernard, P.; Berthon, J.Y.; Chretien, J.R. *Eur. J. Med. Chem.*, **2001**, *36*(1), 21-30.
- [56] Bernard, P.; Pintore, M.; Berthon, J.Y.; Chretien, J.R. *Eur. J. Med. Chem.*, **2001**, *36*(1), 1-19.
- [57] Golbraikh, A.; Bernard, P.; Chretien, J.R. *Eur. J. Med. Chem.*, **2000**, *35*(1), 123-36.
- [58] Seidel, J.K.; Schaper, K.J. *Chemische. Struktur und biologische. Aktivität von Wirkstoffen*, Verlag Chemie, Weinheim. New York, **1979**, pp. 115-252.
- [59] Kubinyi, H. *Quant. Struct.-Act. Relat.*, **2006**, *7*(3), 121-133.
- [60] Tubic, M.; Wagner, D.; Spahn-Langguth, H.; Bolger, M.B.; Langguth, P. *Pharm. Res.*, **2006**, *23*(8), 1712-1720.
- [61] Böhm, H.J.; Klebe, G.; Kubinyi, H. *Wirkstoffdesign, Editorial Spektrum Akademischer Verlag, Heidelberg*, **1996**, pp. 339-340.
- [62] Leach, A.R.; Gillet, V.J. *An Introduction to Chemoinformatics*. Springer, New York, **2003**.
- [63] Cho, S.J.; Tropsha, A. *J. Med. Chem.* **1995**, *38*, 1060-1066.
- [64] Pastor, M.; Cruciani, G.; McLay, I.; Pickett, S.; Clementi, S. *J. Med. Chem.*, **2000**, *43*, 3233-3243.
- [65] Cramer, R.D. *J. Med. Chem.*, **2003**, *46*(3), 374-88.
- [66] Clark, R. *J. Comput.-Aided Mol. Des.*, **2003**, *17*, 265-275.
- [67] Hawkins, D.M.; Kraker, J.J. *Poster Abstract, 234th ACS National Meeting*, Boston, MA, United States, **2007**, *8*, 19-23.
- [68] Tetko, I.V.; Sushko I.; Pandey A.K.; Zhu H.; Tropsha A.; Papa E.; Oberg T.; Todeschini R.; Fourches D.; Varnek A. *J. Chem. Inf. Model.*, **2008**, *48*(9), 1733-46.
- [69] Guha R.; Jurs P.C. *J. Chem. Inf. Model.*, **2007**, *45*, 65-73.
- [70] Topliss J.G.; Edwards R.P. *J. Med. Chem.* **1979**, *22*, 1238 - 1244.
- [71] Folkers, G.; Klebe, G.; Kubinyi, H.; Merz, A.; Rognan, D. *Escom Editorial, Ulrich Thibaut*, **1993**.
- [72] Jorgensen, W.L. *J. Chem. Inf. Model*, **2006**, *46*(3), 937-937.
- [73] De Solla - Price, D.J. *Little Science, Big Science, Columbia University Press: New York*, **1965**, p. 67.
- [74] Kurup A. *J. Comput.-Aided Mol. Des.*, **2003**, *17*, 187-196.
- [75] Norinder, U. *J. Chemomet.*, **1996**, *10*, 95-105.
- [76] Kubinyi, H.; Hamprecht, F. A.; Mietzner T. *J. Med. Chem.*, **1998**, *41*, 2553-2564.
- [77] Hansch, C.; Leo, A. *Exploring QSAR. Volume 1*, American Chemical Society, Washington, DC, **1995**.