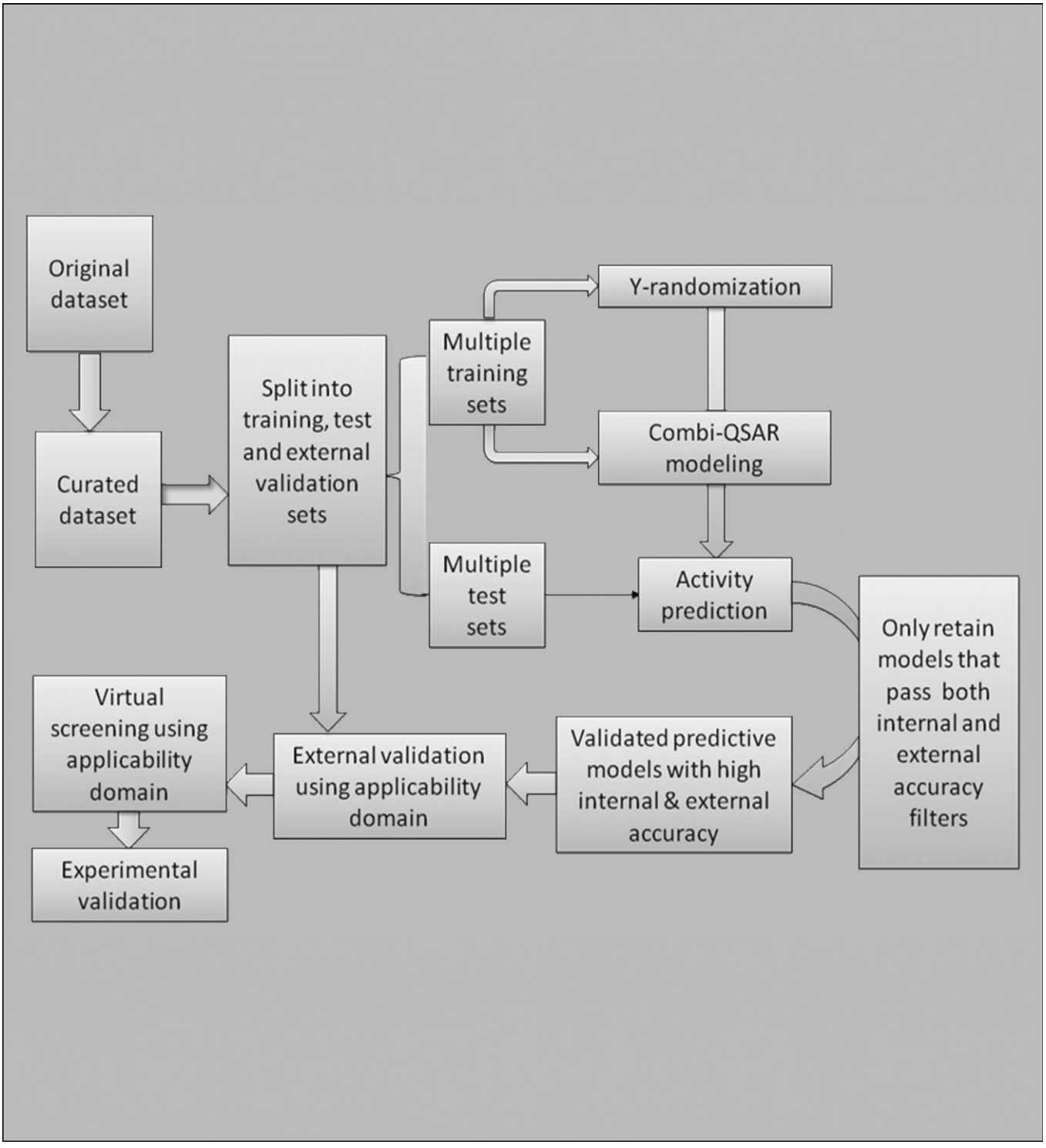
Review

DOI: 10.1002/minf.201000061

Best Practices for QSAR Model Development, Validation, and Exploitation

Alexander Tropsha*[a]



molecular informatics

Abstract: After nearly five decades "in the making", QSAR modeling has established itself as one of the major computational molecular modeling methodologies. As any mature research discipline, QSAR modeling can be characterized by a collection of well defined protocols and procedures that enable the expert application of the method for exploring and exploiting ever growing collections of biologically active chemical compounds. This review examines most critical QSAR modeling routines that we regard as best practices in the field. We discuss these procedures in the context of integrative predictive QSAR modeling workflow that is focused on achieving models of the highest statistical rigor and external predictive power. Specific elements of the workflow consist of data preparation including chemical structure (and when possible, associated biological data) curation, outlier detection, dataset balancing, and model validation. We especially emphasize procedures used to validate models, both internally and externally, as well as the need to define model applicability domains that should be used when models are employed for the prediction of external compounds or compound libraries. Finally, we present several examples of successful applications of QSAR models for virtual screening to identify experimentally confirmed hits.

Keywords: QSAR modeling · Model validation · Virtual screening · Drug discovery

1 Introduction: Basic Principles and Workflow of Predictive QSAR Modeling

Rapid development of information and communication technologies during the last few decades has dramatically changed our capabilities of collecting, analyzing, storing and disseminating all types of data. This process has had a profound influence on the scientific research in many disciplines, including the development of new generations of effective and selective medicines. Large databases containing millions of chemical compounds tested in various biological assays such as PubChem[1] are increasingly available as online collections (recently reviewed by Oprea and Tropsha;[2] see also recent commentary by Williams et al.[3]). In order to find new drug leads, there is a need for efficient and robust procedures that can be used to screen chemical databases and virtual libraries against molecules with known activities or properties. To this end, Quantitative Structure-Activity Relationships (QSAR) modeling provides an effective means for both exploring and exploiting the relationship between chemical structure and its biological action towards the development of novel drug candidates.

The QSAR approach can be generally described as an application of data analysis methods and statistics to developing models that could accurately predict biological activities or properties of compounds based on their structures. Any QSAR method can be generally defined as an application of mathematical and statistical methods to the problem of finding empirical relationships (QSAR models) of the form $P_i = k'(D_1, D_2,\ldots,D_n)$, where P_i are biological activities (or other properties of interest) of molecules, $D_1, D_2,\ldots,D_n$ are calculated (or, sometimes, experimentally measured) structural properties (molecular descriptors) of compounds, and k' is some empirically established mathematical transformation that should be applied to descriptors to calculate the property values for all molecules (Figure 1). The goal of QSAR modeling is to establish a trend in the descriptor values, which parallels the trend in biological activity. In essence, all QSAR approaches imply, directly or indirectly, a simple similarity principle, which for a long time has provided a foundation for the experimental medicinal chemistry: compounds with similar structures are expected to have similar biological activities. The detailed description of major tenets of QSAR modeling is beyond the scope of this paper; the overview of many popular QSAR modeling techniques including statistical and datamining techniques as well as approaches to descriptor calculations could be found in many reviews and monographs, e.g.,[4,5]). Here, we comment on most critical general aspects of model development and, most importantly, validation that are especially important in the context of using QSAR models for virtual screening. Most of our discussion captures trends that the author has either observed or contributed to in the last 20 years of active research in the field. Additional important information concerning both common errors as well as established practices in the QSAR modeling field can be found in other critical essays on the subject, e.g., by Stouch et al.[6] and Dearden et al.[7]

Our experience in QSAR model development and validation has led us to establish a complex strategy[8] that is summarized in Figure 2. It describes the predictive QSAR modeling workflow focused on delivering validated models and ultimately, computational hits that should be ultimately confirmed by the experimental validation. We start by carefully curating chemical structures and, if possible, associated biological activities to prepare the dataset for subsequent calculations. This issue of assessing and addressing data accuracy has not been properly addressed in the literature and we discuss some aspects of this critical component of the workflow below. Then, a fraction of compounds (typically, 10–20%) is selected randomly as an external evaluation set (a more rigorous *n*-fold external validation protocol can be employed when the dataset is randomly divided

[a] *A. Tropsha*
Laboratory for Molecular Modeling and Carolina, Center for Exploratory Cheminformatics Research, CB # 7568, UNC Eshelman School of Pharmacy, University of North Carolina at Chapel Hill Chapel Hill, NC 27599, USA
**e-mail: alex_tropsha@unc.edu*

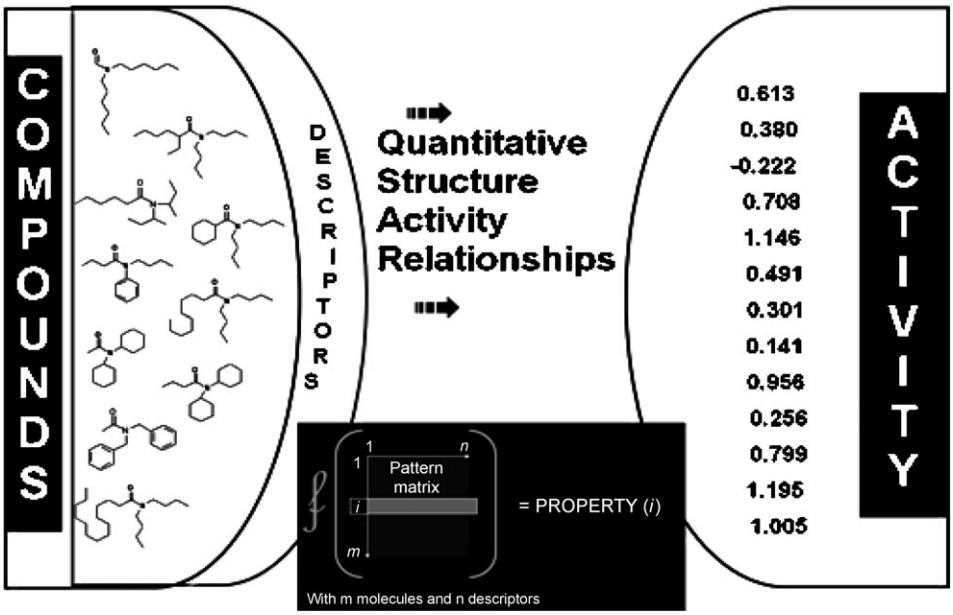


Figure 1. The process of QSAR model development.

into n nearly equal parts and then $n-1$ parts are systematically used for model development and the remaining fraction of compounds is used for model evaluation). The Sphere Exclusion protocol implemented in our laboratory[9,10] is then used to rationally divide the remaining subset of compounds (the modeling set) into multiple training and test sets that are used for model development and validation, respectively (alternative rational approaches for dividing the modeling set into diverse and representative training and test sets could be devised as well). We employ multiple QSAR techniques based on the combinatorial exploration of all possible pairs of descriptor sets and various supervised data analysis techniques (combi-QSAR) and select models characterized by high accuracy in predicting both training and test sets data. The model acceptability thresholds are typically characterized by the lowest acceptable value of the leave-one-out cross validated R^2 (q^2) for the training set and by conventional R^2 for the test set; our default values are 0.6 for both q^2 and R^2. All validated models are finally tested in an ensemble using the external evaluation set. The critical step of the external validation is the use of applicability domains (AD), which is defined uniquely for each model used in consensus (ensemble) prediction of the external set. If external validation demonstrates the significant predictive power of the models we employ them for virtual screening of available chemical databases (e.g., ZINC[11]) to identify putative active compounds and work with collaborators who could validate such hits experimentally.

Thus, models resulting from the predictive QSAR modeling workflow (Figure 2) can be used to prioritize the selection of chemicals for the experimental validation. In fact, it is increasingly critical to include experimental validation as the ultimate assertion of the model-based prediction. We note that the focus on experimental validation shifts the emphasis on ensuring good (best) statistics for the model that fits known experimental data towards generating testable hypotheses about purported bioactive compounds. Thus, the output of the modeling has exactly same format as the input, i.e., chemical structures and (predicted) activities making model interpretation and utilization completely seamless for medicinal chemists. Some of our application studies demonstrating the ability of models to identify computational hits that were subsequently validated experimentally are described below. We now discuss specific procedures (best practices) that should be followed within each individual component of the workflow.

Alexander Tropsha is K. H. Lee Distinguished Professor and Chair of the Division of Medicinal Chemistry and Natural Products in the Eshelman School of Pharmacy, UNC-Chapel Hill. He received PhD in Chemical Enzymology in 1986 from Moscow State University, Russia. He immigrated to the United States in 1989 and has been affiliated with UNC since then. His research interests are in the areas of Computer-Assisted Drug Design, Computational Toxicology, Cheminformatics, and Structural Bioinformatics. His research is supported by multiple grants from the NIH, NSF, EPA, and private companies. He is a member of several editorial boards of scientific journals, permanent member of the BDMA Study Section at the NIH and an elected member of the Board and vice-chair of the international Cheminformatics and QSAR Society in 2005–2010

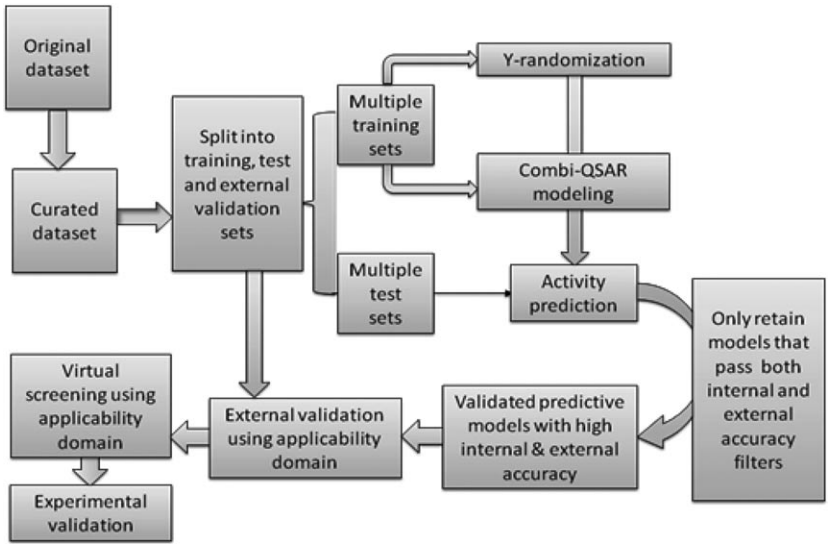


Figure 2. Predictive QSAR modeling workflow.

2 Best Practices for Key Elements of QSAR Modeling Workflow

In this section we discuss specific protocols and procedures that in our experience should be followed to enable the development of reliable and predictive QSAR models. The discussion follows the path of the workflow summarized in Figure 2, from data preparation to model development and validation to application of models for external prediction and virtual screening.

2.1 The Importance of Chemical Data Curation in QSAR Modeling

Molecular modelers typically analyze data generated by other (experimental) researchers. Consequently, when it comes to the experimental data quality they are always at the mercy of the data providers. Practically any cheminformatics study entails the calculation of chemical descriptors that are expected to accurately reflect intricate details of underlying chemical structures. Obviously, any error in the structure translates into either inability to calculate descriptors for erroneous chemical records or into erroneous descriptors; this outcome makes the models developed with such incomplete or inaccurate descriptors either restricted only to a fraction of formally available data or, what is even worse, making the models inaccurate. A recent study[12] showed that on average there are two structural errors per each medicinal chemistry publication with an overall error rate for compounds indexed in the WOMBAT database[13] as high as 8%. In another recent study,[14] the authors investigated several public and commercial databases to calculate their error rates: the latter were ranging from 0.1 to 3.4% depending on the database. As both data and data models as well as the body of scholarly publications in cheminformatics continue to grow it becomes increasingly important to address the issue of data quality that inherently affects the quality of models.

How significant is the problem of accurate structure representation as it concerns the adequacy and accuracy of cheminformatics models? There appears to be no systematic studies on the subject in the published literature. However, even a few recent reports indicate that this problem should be given very serious attention. For instance, recent benchmarking studies by a large group of collaborators from six laboratories[15,16] have clearly demonstrated that the type of chemical descriptors has much greater influence on the prediction performances of QSAR models than the nature of the model optimization techniques. Furthermore, in another recent seminal publication[14] the authors clearly pointed out the importance of chemical data curation in the context of QSAR modeling. They have discussed several illustrative examples of incorrect structures generated from either correct or incorrect SMILEs using commercial software. They also tried to determine the error rate in several known databases and evaluate the consequences of both random and systematic errors for the prediction performance of QSAR models. Their main conclusions were that small structural errors within a dataset could lead to significant losses in predictive abilities of QSAR models. At the same time they further demonstrated that manual curation of structural data leads to substantial increase in model predictivity.[14]

Although there are obvious compelling reasons to believe that chemical data curation should be given a lot of attention, it is also obvious that for the most part the basic steps to curate a dataset of compounds have been either considered trivial or ignored even by experts. For instance, in an effort to improve the quality of publications in the QSAR modeling field the Journal of Chemical Information

and Modeling published a special editorial highlighting the requirements to QSAR papers that should be followed should the authors consider publishing their results in the journal;[17] however, no special attention was given to data curation. There have been several recent publications addressing common mistakes and criticizing faulty practices in QSAR modeling field;[7,18,19] however, these papers have not explicitly described and discussed the importance of chemical record curation for developing robust QSAR models.

Generally speaking, since the models of chemical data may only be as good as the data itself there is a pressing need to develop and systematically employ standard chemical record curation protocols that should be helpful in the pre-processing of any chemical dataset. Recently, we have integrated several protocols in a standardized chemical data curation strategy[20] that in our opinion, should be followed at the onset of any molecular modeling investigation. The simple, but important, steps for cleaning chemical records in a database include the removal of a fraction of the data that cannot be appropriately handled by conventional cheminformatics techniques, e.g., inorganic and organometallic compounds, counterions, salts and mixtures; structure validation; ring aromatization; normalization of specific chemotypes; curation of tautomeric forms; and the deletion of duplicates. It is also critical to visualize and manually inspect at least a fraction of chemical data that go into model development to detect structures that for some reasons escaped the automatic curation steps described above.

It is important to realize that most of these structure curation steps do not depend on the level of chemical structure representation, i.e., 2D or 3D, with possible exception of instances when a dataset includes chiral compounds. Obviously, if standard descriptors are calculated from 2D representation of chemical structure, e.g., by chemical graphs, such as most of molecular connectivity indices,[21] then any pair of enantiomers or diastereoisomers will be formally recognized as duplicates. If specific chiralities for such pairs of compounds are known along with compounds' activities, descriptors taking chirality into account should be used, and all isomers should be retained in the dataset. If, however, chirality information is unavailable, only one compound, usually with the highest (or mean) activity should be retained, and chirality-sensitive descriptors should not be used.

There are different tools available for dataset curation. For example, Molecular Operating Environment (MOE) from CCG[22] includes Database Wash tool. It allows changing molecules' names, adding or removing hydrogen atoms, removing salts and heavy atoms, even if they are covalently connected to the rest of the molecule, and changing or generating the tautomers and protomers (cf. the MOE manual for more details). Various database curation tools are included in ChemAxon[23] as well. If commercial software tools such as MOE are unavailable (notably, the ChemAxon software is free to academic investigators), one can use standard UNIX/LINUX tools to perform some of the dataset cleaning tasks. It is important to have some freely available molecular format converters such as OpenBabel,[24] or MolConverter from ChemAxon.[23]

Figure 3 illustrates major elements of the data curation workflow discussed in more detail in our recent paper.[20] This protocol is enabled by accessible software tools; most of them are publicly available and free-for-academic-use (from ChemAxon,[23] OpenEye,[25] OpenBabel,[24] ISIDA,[26] HiT QSAR,[27] Hyleos[28]), but some are commercial (from Molecular Networks,[29] CCG,[22] CambridgeSoft[30]).

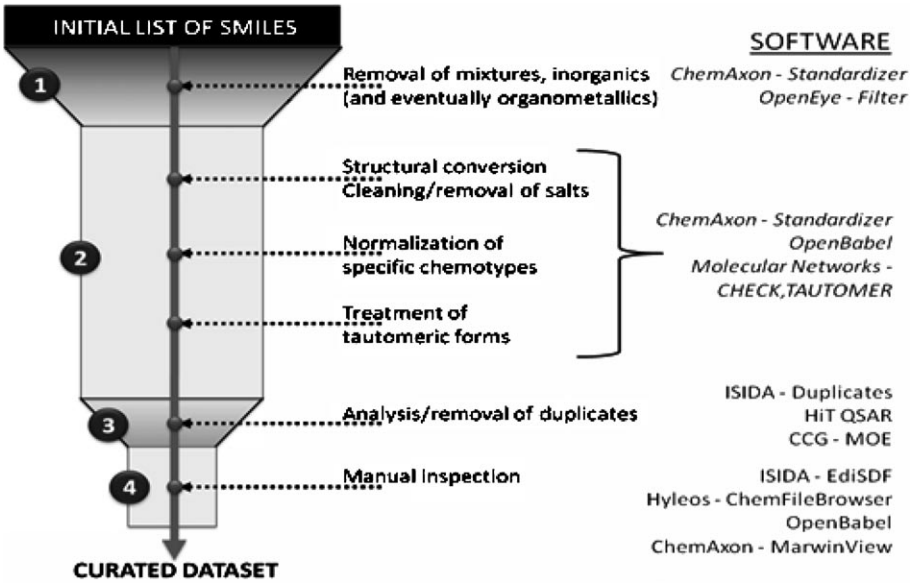


Figure 3. Workflow for chemical data curation.

It is more difficult to spot the errors in biological data since there are no obvious technical approaches similar to chemical record curation that can be used in this case. However, rigorously derived QSAR models could be indeed used to identify compounds for which predictions consistently disagree with experimental observations and that are likely to be annotated with erroneous biological testing results. Our recent studies provide specific examples demonstrating that the use of cheminformatics approaches helped spotting gaps or errors in biological annotations of toxic compounds.[20,31]

2.2 Dataset Size and Balancing

The number of compounds in the dataset for QSAR studies should not be too small, or, for practical reasons, too large. The upper limit is often defined by the computer and time resources available for building QSAR models using the selected methodologies. For example, for *k*-nearest neighbors (*k*NN) QSAR approach frequently practiced in our laboratory, the maximum number of compounds in the *training set* (i.e. compounds used to build QSAR models) may not exceed about ca. 2000 due to the inefficiency of the approach when processing large datasets. When a dataset includes more compounds, several approaches can be implemented: (i) select a diverse subset of compounds; (ii) cluster a dataset and build models separately for each cluster; (iii) sometimes, in case of classification or category QSAR when compounds belong to a small number of activity classes or categories (e.g., active and inactive), it is possible to exclude many compounds from model development. The difference between classes and categories is that in contrast to classes, categories can be ordered. Examples of classes are given by ligands of different receptors; and examples of categories are sets of compounds that are described as very active, moderately active, and inactive.

The lower limit of the number of compounds in the dataset is also defined by several factors. For example, in most cases as part of model validation schemes we divide the dataset into three subsets: training, test and external evaluation sets (see additional discussion below). Training sets are used in model development, and if they are too small, chance correlation and overfitting become major problems not allowing one to build truly predictive models. While it is impossible to give an exact minimum number of compounds in a dataset for which building reliable QSAR models is feasible, some simple ideas described here may help. We suggest that in case of continuous response variable (activity) the number of compounds in the training set should be at least 20, and about 10 compounds should be in each of the test and external evaluation sets, so the total minimum number of compounds should be no less than 40. In case of classification or category response variable, training set should contain at least about 10 compounds of each class, and test and external evaluation sets should contain no less than five compounds for each class. So, there should be at least 20 compounds of each class. The best situation is when the number of compounds in the dataset is between these two extremes: about 150–300 compounds in total, and in case of classification or category QSAR approximately equal number of compounds of each class or category.

There are also requirements for activity values. In case of continuous response variable, the total range of activities should be at least five times higher than the experimental error. No large gaps (that exceed 10%–15% of the entire range of activities) are allowed between two consecutive values of activities ordered by value. In case of classification or category QSAR, there should be at least 20 compounds of each class or category; preferably, the number of compounds in all classes or categories should be approximately the same. However, many existing datasets are imbalanced or biased (i.e. sizes of different classes or categories are different). In these cases, special approaches should be used to equalize the number of compounds in different classes or categories.

Indeed, in many datasets, the counts of compounds that belong to different classes or categories are significantly different (there could be several times and even orders of difference). Usually, active compounds constitute a smaller class and inactive compounds a larger class (which is practically always the case for datasets resulting from large scale HTS studies). Active compounds (typically binding to a certain biological target) belong to a relatively small number of structural classes. On the other hand, compounds included in the larger class (i.e. inactive compounds) can be very diverse: some of them can belong to the same structural classes as active compounds, while other compounds (often, the majority of them) have very different structures highly dissimilar from those included in the smaller class. So they cover a large area in the descriptor space relative to the active compounds which are much more similar to each other. In these cases, direct development of predictive QSAR models using entire datasets is difficult, if not impossible. Indeed, training and test sets reflect the composition of the entire dataset, in which almost all compounds are inactive, so the modeling and validation will be biased toward correct prediction of the larger class. Thus, reducing the number of compounds included in the larger class is necessary. This can be achieved easily by calculating the distance (or similarity) matrix between compounds belonging to different classes followed by excluding compounds of the larger class that are dissimilar beyond certain threshold from those of the smaller class. Ideally, after excluding dissimilar compounds of the larger class, the number of remaining compounds of this class should be more or less equal to the number of compounds of the smaller class. Classification QSAR models are developed then only for compounds that remain in the balanced dataset. In other words, the modeling subset will not include compounds of the (initially) larger class that were excluded by the procedure as more dissimilar to the smaller class than the re-

maining molecules of the (initially) bigger class. This approach makes it more challenging to achieve a successful QSAR model that discriminates, say, active compounds from most chemically similar inactive compounds; therefore we consider it inherently more robust over alternative approaches reported in the literature when random samples of the bigger class are used to create balanced datasets for classification modeling.

Alternative approaches that could help balancing datasets include undersampling of the bigger class[32,33] or oversampling of the smaller class.[34] The extended discussion of these approaches is beyond the scope of this review.

2.3 Detection and Removal of Outliers Prior to QSAR Studies

Success of QSAR modeling depends on the appropriate selection of a dataset for QSAR studies. In a recent editorial of the Journal of Chemical Information and Modeling, Maggiora[18] noticed that one of the main deficiencies of many chemical datasets is that they do not fully satisfy the main hypothesis underlying all QSAR studies: similar compounds have similar biological activities or properties. Maggiora defines the "cliffs" in the descriptor space where the properties change so rapidly, that, in fact adding or deleting one small chemical group can lead to a dramatic change in the compound's property. In other words, small changes of descriptor values can lead to large changes in molecular properties. Generally, in this case there could be not just one outlier, but a subset of compounds whose properties are different from those on the other "side" of the cliff. In other words, cliffs are areas where the main QSAR hypothesis (similar compounds have similar properties) does not hold. So cliff detection is a major QSAR problem. In QSAR area, many people were aware of these and other problems related to outlier detection, but have not yet paid a sufficient attention to addressing them in automated QSAR procedures.

There are two types of outliers we must be aware of: *leverage* (or structural) outliers and *activity* outliers. Structural outliers can be defined as singletons in a dataset clustered using any of available techniques described in standard statistical literature and activity outliers are essentially defined as activity "cliffs" (see above). One should keep in mind that both types of outliers can be real or due to errors in structure representation or biological activity annotation but in any case, preserving outliers in a modeling dataset will likely lead to model instability; the latter can be manifested in significant differences in external predictive power of models built with n-fold external validation strategy. Thus, outliers should be removed before proceeding with model development and analyzed separately for possible errors; however, no current QSAR modeling techniques provides a reliable approach to build models taking outliers into account. Finding such approaches is one of the challenges facing the field.

2.4 Critical Importance of Model Validation

In our important paper titled "Beware of q^2!",[35] we have demonstrated the insufficiency of the training set statistics for developing externally predictive QSAR models and formulated the main principles of model validation. Despite earlier observations and warnings of several authors[36–38] that high cross-validated correlation coefficient R^2 (q^2) is the necessary, but insufficient condition for the model to have high predictive power, many studies continue to consider q^2 as the only parameter characterizing the predictive power of QSAR models. In this respect, we have shown[35] that the predictive power of QSAR models can be claimed only if the model was successfully applied for predicting the external test set compounds, which were not used in the model development.

Indeed, it is important to emphasize that the true predictive power of a QSAR model can be established only through model validation procedure which consists of prediction of activities of compounds which were not included in model building, i.e., compounds in the *test set*. In contrast to the test set, compounds used for model building constitute the *training set*. In many QSAR studies multiple models are built and from them "best" models are selected, which are defined as those based on the prediction statistics for the test set. Thus, the test set is actually used to select models. This use of the test set for model selection practically negates the consideration of such routine as an adequate external model validation. In fact, it does not guarantee at all that models selected in this way will make accurate predictions if used for chemical *database mining* (i.e. predicting activities of compounds in truly external database). In our workflow, to simulate the use of QSAR models for database mining, a so called *external evaluation set* is employed. It should consist of compounds with known activities that are not included in either training or test sets. External evaluation set can be selected randomly from the entire initial dataset. In general, the size of the external evaluation set should be about 15%–20% of the entire dataset. The remaining part of the dataset is called *modeling set* that can be divided into training and test sets. Algorithms for dividing a modeling set into diverse and representative training and test sets were developed in our group previously and are reported and discussed in detail elsewhere.[39]

We have demonstrated earlier[35] that the majority of models with high q^2 values have poor predictive power when applied for prediction of compounds in the external test set. In the subsequent publication[40] the importance of rigorous validation was again emphasized as a crucial, integral component of model development. Several examples of published QSPR models with high fitted accuracy for the training sets, which failed rigorous validation tests, have been considered. We presented a set of simple guidelines for developing validated and predictive QSPR models and discussed several validation strategies such as the randomi-

zation of the response variable (*Y*-randomization), external validation using rational division of a dataset into training and test sets, and others. We highlighted the need to establish the domain of model applicability in the chemical space to flag molecules for which predictions may be unreliable, and discussed some algorithms that can be used for this purpose. We advocated the broad use of these guidelines in the development of predictive QSPR models.[40–42]

The importance of model validation could now be regarded as collective wisdom within the community of molecular modelers. At the 37-th Joint Meeting of Chemicals Committee and Working Party on Chemicals, Pesticides & Biotechnology, held in Paris on 17–19 November 2004, the OECD (Organization for Economic Co-operation and Development) member countries adopted the following five principles that valid (Q)SAR models should follow to allow their use in regulatory assessment of chemical safety: (i) a defined endpoint; (ii) an unambiguous algorithm; (iii) a defined domain of applicability; (iv) appropriate measures of goodness-of-fit, robustness and predictivity; (v) a mechanistic interpretation, if possible. Since then, most of the European authors publishing in QSAR area include a statement that their models fully comply with OECD principles (e.g., see References[43–46]).

Validation of QSAR models is one of the most critical problems of QSAR. Recently, we have extended our requirements for the validation of multiple QSAR models selected by acceptable statistics criteria of prediction for the test set.[47] Additional studies on this critical component of QSAR modeling should establish reliable and commonly accepted applicability domain criteria, which should make models increasingly useful for virtual screening.

2.5 Applicability Domains and Model Acceptability Criteria

One of the most important problems in QSAR analysis is establishing the domain of applicability for each model. In the absence of the applicability domain restriction, each model can formally predict the activity of any compound, even with a completely different structure from those included in the training set. Thus, the absence of the model applicability domain as a mandatory component of any QSAR model would lead to the unjustified extrapolation of the model in the chemistry space and, as a result, a high likelihood of inaccurate predictions. In our research we have always paid particular attention to this issue.[40,48–55] A good overview of commonly used applicability domain definitions can be found elsewhere.[56,57]

In our earlier publications[35,40] we have recommended a set of statistical criteria which must be satisfied by a predictive model. For continuous QSAR, criteria that we will follow in developing activity/property predictors are as follows: (i) correlation coefficient R between the predicted and observed activities; (ii) coefficients of determination[58] (predicted versus observed activities ${R_0}^2$, and observed versus predicted activities ${R'_0}^2$ for regressions through the origin); (iii) slopes k and k' of regression lines through the origin. We consider a QSAR model predictive, if the following conditions are satisfied (i) $q^2 > 0.5$; (ii) $R^2 > 0.6$; (iii) $({R_0}^2 - {R_0}^2)/R^2 < 0.1$ and $0.85 \le k \le 1.15$ or $({R_0}^2 - {R'_0}^2)/R^2 < 0.1$ and $0.85 \le k' \le 1.15$; (iv) $|{R_0}^2 - {R'_0}^2| < 0.3$ where q^2 is the cross-validated correlation coefficient calculated for the training set, but all other criteria are calculated for the test set (for additional discussion, see[8]).

3 Predictive QSAR Models as Virtual Screening Tools

In our recent studies we were fortunate to recruit experimental collaborators who have validated computational hits identified by virtual screening of commercially available compound libraries using rigorously validated QSAR models. Examples include anticonvulsants,[53] HIV-1 reverse transcriptase inhibitors,[59] D1 antagonists,[60] antitumor compounds,[61] beta-lactamase inhibitors,[62] Human Histone Deacetylase (HDAC) inhibitors,[63] and geranylgeranyltransferase-I inhibitors.[64] Thus, models resulting from predictive QSAR workflow could be used to prioritize the selection of chemicals for the experimental validation. To illustrate the power of validated QSAR models as virtual screening tools we shall discuss the examples of studies that resulted in experimentally confirmed hits. We note that such studies could only be done if there is sufficient data available for a series of tested compounds such that robust validated models could be developed. The following examples illustrate the use of QSAR models developed with predictive QSAR modeling and validation workflow (Figure 2) for virtual screening of commercial libraries to identify experimentally confirmed hits.

3.1 Discovery of Novel Anticancer Agents

A combined approach of validated QSAR modeling and virtual screening was successfully applied to the discovery of novel tylophorine derivatives as anticancer agents.[61] QSAR models have been initially developed for 52 chemically diverse phenanthrine-based tylophorine derivatives (PBTs) with known experimental EC_{50} using chemical topological descriptors (calculated with the MolConnZ program) and variable selection k nearest neighbor (kNN) method. Several validation protocols have been applied to achieve robust QSAR models. The original dataset was divided into multiple training and test sets, and the models were considered acceptable only if the leave-one-out cross-validated R^2 (q^2) values were greater than 0.5 for the training sets and the correlation coefficient R^2 values were greater than 0.6 for the test sets. Furthermore, the q^2 values for the actual dataset were shown to be significantly higher than those obtained for the same dataset with randomized target properties (*Y*-randomization test), indicating that models were statistically significant. Ten best models were then em-

ployed to mine a commercially available ChemDiv Database (ca. 500 K compounds) resulting in 34 consensus hits with moderate to high predicted activities. Ten structurally diverse hits were experimentally tested and eight were confirmed active with the highest experimental EC_{50} of 1.8 μM implying an exceptionally high hit rate (80%). The same ten models were further applied to predict EC_{50} for four new PBTs, and the correlation coefficient (R^2) between the experimental and predicted EC_{50} for these compounds plus eight active consensus hits was shown to be as high as 0.57.

3.2 Discovery of Novel Histone Deacetylase (HDAC) Inhibitors

Histone deacetylases (HDAC) play a critical role in transcription regulation. Small molecule HDAC inhibitors have become an emerging target for the treatment of cancer and other cell proliferation diseases. We have employed variable selection k Nearest Neighbor approach (*k*NN) and Support Vector Machines approach (SVM) to generate QSAR models for 59 chemically diverse compounds with inhibition activity on class I HDAC. MOE[22] and MolConnZ[65] based 2D descriptors were combined with *k* nearest neighbor (*k*NN) and support vector machines (SVM) approaches independently to improve the predictive power of models. Rigorous model validation approaches were employed including randomization of target activity (*Y*-randomization test) and assessment of model predictability by consensus prediction on two external datasets. Highly predictive QSAR models were generated with leave-one-out cross validation R^2 (q^2) values for the training set and R^2 values for the test set as high as 0.81 and 0.80, respectively with MolconnZ/*k*NN approach and 0.94 and 0.81, respectiveley with MolconnZ/SVM approach. Validated QSAR models were then used to mine four chemical databases: National Cancer Institute (NCI) database, Maybridge database, ChemDiv database and ZINC database, including a total of over 3 million compounds. The searches resulted in 48 consensus hits, including two reported HDAC inhibitors that were not included in the original data set. Four database hits with novel structural features were purchased and tested using the same biological assay that was employed to assess the inhibition activity of the training set compounds. Three of these four compounds were confirmed active with the best inhibitory activity (IC_{50}) of 1 μM.

3.3 Discovery of Novel Geranylgeranyltransferase Type I (GGTase-I) Inhibitors

In another recent study,[64] we employed our standard QSAR modeling workflow (Figure 2) to discover novel Geranylgeranyltransferase type I (GGTase-I) inhibitors. Geranylgeranylation is critical to the function of several proteins including Rho, Rap1, Rac, Cdc42, and G-protein gamma subunits. GGTase-I inhibitors (GGTIs) have therapeutic potential to treat inflammation, multiple sclerosis, atherosclerosis, and many other diseases. Following our standard QSAR modeling workflow, we have developed and rigorously validated models for 48 GGTIs using variable selection *k* nearest neighbor[66] and automated lazy learning,[54] and genetic algorithm-partial least square[67] QSAR methods. The QSAR models were employed for virtual screening of 9.5 million commercially available chemicals yielding 47 diverse computational hits. Seven of these compounds with novel scaffolds and high predicted GGTase-I inhibitory activities were tested in vitro, and all were found to be bona fide and selective micromolar inhibitors.

Figure 4 shows the structures of both representative training set compounds as well as confirmed computational hits. We should emphasize that QSAR models have been traditionally viewed as lead optimization tools capable of predicting compounds with chemical structure similar to the structure of molecules used for the training set. However, this study clearly indicates (Figure 4) that with enough attention given to the model development process and using chemical descriptors characterizing whole molecules (as opposed to, e.g., chemical fragments) it is indeed possible to discover compounds with novel chemical scaffolds. Furthermore, in our study we have additionally demonstrated that these novel hits could not be identified using traditional chemical similarity search,[64] which highlights the power of robust QSAR models as the drug discovery tool.

In summary, several examples above demonstrate that QSAR models could be used successfully as virtual screening tools to discover compounds with the desired biological activity in chemical databases or virtual libraries.[53,60,61,68,69] It should be stressed that the total number of compounds selected for virtual screening based on QSAR model predictions is typically relatively small, only a few dozen. Obviously, the total number of computational hits is controlled by the value of applicability domain. In

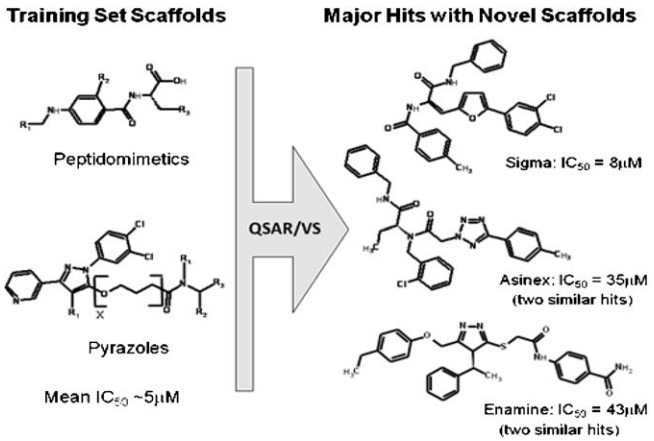


Figure 4. The use of QSAR modeling, virtual screening of commercial libraries, and experimental validation of computational hits afforded the discovery of geranylgeranyltransferase-I inhibitors with novel scaffolds.

most published cases, because we were limited in both time and resources, we chose a very conservative applicability domain leading to the selection of a small library of computational hits with an expectation that a large fraction of these would be confirmed as active compounds. In the industrial size projects it may be more reasonable to loosen the applicability domain requirement and increase the size of virtual hit library. One may expect that the increase in the library size will result in lower relative accuracy of prediction but the absolute number of confirmed hits may actually increase. Thus, scientists using QSAR models that incorporate the applicability domain should always be aware of the interplay between the size of the domain, the coverage of the virtual screening library, and the prediction accuracy so they should use the applicability domain as a tunable parameter to control this interplay. The discovery of novel bioactive chemical entities is the primary goal of computational drug discovery, and the development of validated and predictive QSAR models is critical to achieve this goal.

4 Best Practices for Contests in QSAR Modeling: Competitive Collaboration and Consensus Modeling

The title of this section may appear contradictory and perhaps controversial because competition is perhaps one of the major (and for the most part, healthy) attributes of scientific research. Nevertheless, we believe that QSAR modeling may provide unique environment to advance the field by a mechanism that we may regard as "competitive collaboration". The following example may help illustrate our point.

4.1 Study Design

In a recent study,[15] the combinational QSAR modeling approach was applied to a diverse series of organic compounds tested for aquatic toxicity in *Tetrahymena pyriformis* in the same laboratory of Prof. T. Schultz over nearly a decade.[70–76] The unique aspect of this research was that it was conducted in collaboration between six academic groups specializing in cheminformatics and computational toxicology. The common goals for our virtual collaboratory were to explore the relative strengths of various QSAR approaches in their ability to develop robust and externally predictive models of this particular toxicity end point. The members of our collaboratory included scientists from the University of North Carolina at Chapel Hill in the United States (UNC); University of Louis Pasteur (ULP) in France; University of Insubria (UI) in Italy; University of Kalmar (UK) in Sweden; Virtual Computational Chemistry Laboratory (VCCLAB) in Germany; and the University of British Columbia (UBC) in Canada. Each group relied on its own QSAR modeling approaches to develop toxicity models using the same modeling set, and we agreed to evaluate the realistic model performance using the same external validation set(s).

The *T. pyriformis* toxicity dataset used in this study was compiled from several publications of the Schultz group as well as from data available at the Tetratox database website of (http://www.vet.utk.edu/TETRATOX/). After deleting duplicates as well as several compounds with conflicting test results and correcting several chemical structures in the original data sources, our final dataset included 983 unique compounds. The dataset was randomly divided into two parts: 1) the modeling set of 644 compounds; 2) the validation set including 339 compounds. The former set was used for model development by each participating group and the latter set was used to estimate the external prediction power of each model as a universal metric of model performance. In addition, when this project was already well underway, a new dataset had become available from the most recent publication by the Schultz group.[77] It provided us with an additional external set to evaluate the predictive power and reliability of all QSAR models. Among compounds reported in[77] 110 were unique, i.e., not present among the original set of 983 compounds; thus, these 110 compounds formed the second independent validation set for our study.

Naturally, different groups employed different techniques and (sometimes) different statistical parameters to evaluate the performance of models developed independently for the modeling set. To harmonize the results of this study the same standard parameters were chosen to describe each model's performance as applied to the modeling and external test set predictions. Thus, we have employed Q^2_{abs} (squared leave-one-out cross-validation correlation coefficient) for the modeling set, R^2_{abs} (frequently described as coefficient of determination) for the external validations sets, and *MAE* (mean absolute error) for the linear correlation between predicted (Y_{pred}) and experimental (Y_{exp}) data (here, $Y = pIGC_{50}$); these parameters are defined as follows:

$$Q^2_{abs} = 1 - \sum_Y (Y_{exp} - Y_{LOO})^2 / \sum_Y (Y_{exp} - <Y>_{exp})^2 \quad (1)$$

$$R^2_{abs} = 1 - \sum_Y (Y_{exp} - Y_{pred})^2 / \sum_Y (Y_{exp} - <Y>_{exp})^2 \quad (2)$$

$$MAE = \sum_Y |Y - Y_{pred}| / n \quad (3)$$

Many other statistical characteristics can be used to evaluate model performance; however, we restricted ourselves to these three parameters that provide minimal but sufficient information concerning any model's ability to reproduce both the trends in experimental data for the test sets as well as mean accuracy of predicting all experimental values. The models were considered acceptable if $R_{abs}{}^2$ exceeded 0.5.

4.2 QSAR Models of Aquatic Toxicity; Comparison Between Methods and Models

The objective of this study from methodological prospective was to explore the suitability of different QSAR modeling tools for the analysis of a dataset with an important toxicological endpoint. Typically, such datasets are analyzed with one (or several) modeling techniques, with a great emphasis on the (high value of) statistical parameters of the training set models. In this study, we went well beyond the modeling studies reported in the original publications from the Schultz group in several respects. First, we have compiled all reported data on chemical toxicity against *T. pyriformis* in a single large dataset and attempted to develop global QSAR models for the entire set. Second, we have employed multiple QSAR modeling techniques thanks to the engagement of six collaborating groups. Third, we have focused on defining model performance criteria not only using training set data but most importantly using external validation sets that were not used in model development in any way (unlike any common cross-validation procedure).[78] This focus afforded us the opportunity to evaluate and compare all models using simple and objective universal criteria of external predictive accuracy, which in our opinion is the most important single figure of merit for a QSAR model that is of practical significance for experimental toxicologists. Fourth, we have explored the significance of applicability domains and the power of consensus modeling in maximizing the accuracy of external predictivity of our models.

The results of this exercise demonstrated that all models performed quite well for the training set with even the lowest Q_{abs}^2 among them as high as 0.72. However, there was much greater variation between these models when looking at their (universal and objective) performance criteria as applied to the external validation sets. Thus, of 15 QSAR approaches used in this study, nine implemented method-specific applicability domains. Models that did not define the AD showed a reduced predictive accuracy for the validation set II even though they yielded reasonable results for the validation set I.

4.3 The Power of Consensus

For the most part all models succeeded in achieving reasonable accuracy of external prediction especially when using the AD. It then appeared natural to bring all models together to explore the power of consensus prediction. Thus, the consensus model was constructed by averaging all available predicted values taking into account the applicability domain of each individual model. In this case we could use only nine of 15 models that had the AD defined. Since each model had its unique way of defining the AD, each external compound could be found within the AD of anywhere between one and nine models so for averaging we only used models covering the compound. The advantage of this data treatment is that the overall coverage of the prediction is still high because it was rare to have an external compound outside of the ADs of all available models. The results showed that the prediction accuracy for both the modeling set ($MAE = 0.22$) and the validation sets I and II (0.27 and 0.34, respectively) was the best compared to any individual model. The same observation could be made for the correlation coefficient R^2_{abs}. The coverage of this consensus model was actually 100% for all three data sets. This observation suggests that consensus models afford both high space coverage and high accuracy of prediction.

In summary, this study presents an example of a fruitful international collaboration between researchers that use different techniques and approaches but share general principles of QSAR model development and validation. Significantly, we did not make any assumptions about the purported mechanisms of aquatic toxicity yet were able to develop statistically significant models for all experimentally tested compounds. However, the most significant single result of our studies is the demonstrated superior performance of the consensus modeling approach when all models are used concurrently and predictions from individual models are averaged. We have shown that both the predictive accuracy and coverage of the final consensus QSAR models were superior as compared to these parameters for individual models. The consensus models appeared robust in terms of being insensitive to both incorporating individual models with low prediction accuracy and the inclusion or exclusion of the AD. Another important result of this study is the power of addressing complex problems in QSAR modeling by forming a virtual collaboratory of independent research groups leading to the formulation and empirical testing of *best modeling practices*. This study confirms the power of the "competitive collaboration" principle that we proposed in the beginning of this section.

5 Summary and Conclusions

As is true perhaps for any computational field, QSAR modeling has been both blessed and sometimes, cursed in the literature. Our group was among the first emphasizing the importance of statistical validation of QSAR models.[35] As we pointed out and demonstrated with examples in this review (cf. Section 4.2), the high accuracy of the training set model characterized with leave-one-out cross validated R^2 (q^2), i.e., model fitness, is not indicative of the high external predictive power of the model. Thus, the exclusive reliance on training set modeling without any external validation is one of the reasons why many models cannot be considered reliable. Another important paper examined the reasons behind the failure of in silico ADME/Tox models[6] linking the frequent failures to the inappropriate use of models, false expectations, or procedures used to develop models. In a brief but very important editorial note G. Mag-

giora[18] outlined limitations and some reasons for failures of QSAR modeling that relate to the so called "activity cliffs", which are known cases when a small change in chemical structure leads to dramatic changes in the target activity. Such cases are indeed difficult to foresee and hard to capture and explain using QSAR models since the models work best in reflecting relatively smooth trends in structure-activity correlations. Addressing the activity cliffs problem is indeed a hard problem in QSAR modeling and in some cases it is a source of poor predictions. A summary of various reasons leading to erroneous QSAR models was given in a recent critical overview of the field.[79] Another recent important paper listed as many as 21 possible sources of error when developing QSAR models[7] and provided some recipes as to how to avoid at least some common errors in QSAR model development.

In most cases the authors concerned with the quality and practical utility of QSAR models looked deeply into possible sources of errors or offered approaches to improve the robustness of models. On the other hand, the author of the negative opinion letter published in early 2008[19] made an unfortunate attempt to equate the fraction of papers not paying enough attention to the statistical quality of models with the entire field. As we discuss in this review, it is critically important to avoid the oversimplification of the QSAR modeling process and employ statistically robust approaches for both model development and validation. Authors ignoring the complexity of the problem or those paying insufficient attention to model validation do end up developing and in some cases, publishing models that could not be regarded as reliable. Conversely, the criticism of the field should be balanced and based on the thorough analysis of possible sources of error rather than equating the entire field to one large error as the aforementioned opinion letter[19] did. Thus, in this review we have made a fair attempt to outline the objective challenges facing (but not dooming!) the field (such as activity cliffs) and emphasize the importance of developing and practicing rigorous approaches to both model development and validation.

In conclusion, we have discussed current best practices for developing robust and externally predictive QSAR models. As with any computational molecular modeling approach, it is imperative that QSAR method is used expertly. Therefore, this review has focused on the discussion of critical components of QSAR modeling procedures that should be studied and executed rigorously to enable their successful application. We have shown that with enough attention paid to critical issues of model validation and applicability domain definition, the models could be indeed used successfully to mine external virtual libraries, especially of commercially available chemicals, to generate reliable computational hits. The methods and applications discussed in this review should be of help to both computational and synthetic chemists as well as experimental biologists working in the areas of biological screening of chemical libraries.

Acknowledgements

The author would like to acknowledge the support from *National Cancer Institute NIH* (Grant R01GM066940). The author is also thankful to several colleagues in his laboratory, especially Profs. *Golbraikh*, *Fourches*, *Zhu* and *Wang*, who have conducted many studies cited in this review as well as were engaged in many scientific discussions that helped the author formulate the best current practices for QSAR modeling discussed herein.

References

[1] *PubChem*, http://pubchem.ncbi.nlm.nih.gov/, **2008**.
[2] T. Oprea, A. Tropsha, *Drug Discov. Today* **2006**, *3*, 357 – 365.
[3] A. Williams, V. Tkachenko, C. Lipinski, A. Tropsha, S. Ekins, *Drug Discovery World* **2010**, *10*, 33 – 39.
[4] A. Tropsha, in *Comprehensive Medicinal Chemistry II*, Vol. 4 (Ed: Y. C. Martin), Elsevier, Amsterdam **2006**, pp. 149 – 165.
[5] R. Todeschini, V. Consonni, *Handbook of Molecular Descriptors*, Wiley-VCH, Weinheim, Germany, **2000**.
[6] T. R. Stouch, J. R. Kenyon, S. R. Johnson, X. Q. Chen, A. Doweyko, Y. Li, *J. Comput. Aided Mol. Des.* **2003**, *17*, 83 – 92.
[7] J. C. Dearden, M. T. Cronin, K. L. Kaiser, *SAR QSAR. Environ. Res.* **2009**, *20*, 241 – 266.
[8] A. Tropsha, A. Golbraikh, *Curr. Pharm. Des.* **2007**, *13*, 3494 – 3504.
[9] A. Golbraikh, A. Tropsha, *Mol. Divers.* **2002**, *5*, 231 – 243.
[10] A. Golbraikh, M. Shen, Z. Xiao, Y. D. Xiao, K. H. Lee, A. Tropsha, *J. Comput. Aided Mol. Des.* **2003**, *17*, 241 – 253.
[11] J. J. Irwin, B. K. Shoichet, *J. Chem. Inf. Model.* **2005**, *45*, 177 – 182.
[12] M. Olah, M. Mracec, L. Ostopovici, R. Rad, A. Bora, N. Hadaruga, I. Olah, M. Banda, Z. Simon, M. Mracec, T. I. Oprea, in *Chemoinformatics in Drug Discovery* (Ed: T. I. Oprea), Wiley-VCH, New York, **2005**, pp. 223 – 239.
[13] M. Olah, R. Rad, L. Ostopovici, A. Bora, N. Hadaruga, D. Hadaruga, R. Moldovan, A. Fulias, M. Mracec, T. I. Oprea, in *Chemical Biology: From Small Molecules to Systems Biology and Drug Design* (Eds: S. L. Schreiber, T. M. Kapoor, G. Weiss), Wiley-VCH, Weinheim, **2007**, pp. 760 – 786.
[14] D. Young, T. Martin, R. Venkatapathy, P. Harten, *QSAR Comb. Sci.* **2008**, *27*, 1337 – 1345.
[15] H. Zhu, A. Tropsha, D. Fourches, A. Varnek, E. Papa, P. Gramatica, T. Oberg, P. Dao, A. Cherkasov, I. V. Tetko, *J. Chem. Inf. Model.* **2008**, *48*, 766 – 784.
[16] I. V. Tetko, I. Sushko, A. K. Pandey, H. Zhu, A. Tropsha, E. Papa, T. Oberg, R. Todeschini, D. Fourches, A. Varnek, *J. Chem. Inf. Model.* **2008**, *48*, 1733 – 1746.
[17] W. L. Jorgensen, *J. Chem. Inf. Model.* **2006**, *46*, 937.
[18] G. M. Maggiora, *J. Chem. Inf. Model.* **2006**, *46*, 1535.
[19] S. R. Johnson, *J. Chem. Inf. Model.* **2008**, *48*, 25 – 26.
[20] D. Fourches, E. Muratov, A. Tropsha, *J. Chem. Inf. Model.* **2010**, DOI: 10.1021/ci100176x in press.
[21] L. B. Kier, L. H. Hall, *Molecular Connectivity in Chemistry and Research*, Academic Press, New York, **1976**.
[22] *MOE*, Chemical Computing Group. http://www.chemcomp.com/index.htm, **2010**.
[23] *ChemAxon*, ChemAxon JChem (http://www.chemaxon.com), **2010**.

[24] *OpenBabel, the OpenSource Chemistry Toolbox*, Openbabel.org, **2010**, 2 – 1–2010.
[25] *OpenEye Scientific Software*, http://www.eyesopen.com/products/applications/filter.html, **2010**.
[26] *ISIDA software*, Laboratoire d'Infochimie, Louis Pasteur University, Strasbourg, France (infochim.u-strasbg.fr), **2010**.
[27] V. E. Kuz'min, A. G. Artemenko, E. N. Muratov, *J. Comp. Aid. Mol. Des.* **2008**, *22*, 403 – 421.
[28] *Hyleos*, http://www.hyleos.net/, **2010**.
[29] Molecular Networks GmbH, (http://www.molecular-networks.com/products), **2010**.
[30] CambridgeSoft, http://www.cambridgesoft.com/, **2009**.
[31] D. Fourches, J. C. Barnes, N. C. Day, P. Bradley, J. Z. Reed, A. Tropsha, *Chem. Res. Toxicol.* **2010**, *23*, 171 – 183.
[32] S.-J. Yen, Y.-S. Lee, *Lecture Notes in Control and Information Sciences* **2006**, *344*, 731 – 740.
[33] M. Kubat, S. Matwin, *Proc. 14th Conf. on Machine Learning*, **1977**, pp. 179 – 186.
[34] N. Japkowicz, *Proc. Learning from Imbalanced Data Sets, Papers from the AAAI Workshop*, Technical Report WS-00-05 (Ed: N. Japkowicz), pp. 10 – 15.
[35] A. Golbraikh, A. Tropsha, *J. Mol. Graph. Model.* **2002**, *20*, 269 – 276.
[36] E. Novellino, C. Fattorusso, G. Greco, *Pharm. Acta Helv.* **1995**, *70*, 149 – 154.
[37] U. Norinder, *J. Chemomet.* **1996**, *10*, 95 – 105.
[38] A. Tropsha, S. J. Cho, in *3D QSAR in Drug Design* (Eds: H. Kubinyi, G. Folkers, Y. C. Martin), Kluwer Academic, Dordrecht, The Netherlands, **1998**, pp. 57 – 69.
[39] A. Golbraikh, M. Shen, Z. Xiao, Y. D. Xiao, K. H. Lee, A. Tropsha, *J. Comput. Aided Mol. Des.* **2003**, *17*, 241 – 253.
[40] A. Tropsha, P. Gramatica, V. K. Gombar, *QSAR Comb. Sci.* **2003**, *22*, 69 – 77.
[41] A. Golbraikh, A. Tropsha, *J. Comput. Aided Mol. Des.* **2002**, *16*, 357 – 369.
[42] A. Golbraikh, M. Shen, Z. Xiao, Y. D. Xiao, K. H. Lee, A. Tropsha, *J. Comput. Aided Mol. Des.* **2003**, *17*, 241 – 253.
[43] M. Pavan, T. I. Netzeva, A. P. Worth, *SAR QSAR Environ. Res.* **2006**, *17*, 147 – 171.
[44] M. Vracko, V. Bandelj, P. Barbieri, E. Benfenati, Q. Chaudhry, M. Cronin, J. Devillers, A. Gallegos, G. Gini, P. Gramatica, C. Helma, P. Mazzatorta, D. Neagu, T. Netzeva, M. Pavan, G. Patlewicz, M. Randic, I. Tsakovska, A. Worth, *SAR QSAR Environ. Res.* **2006**, *17*, 265 – 284.
[45] A. G. Saliner, T. I. Netzeva, A. P. Worth, *SAR QSAR Environ. Res.* **2006**, *17*, 195 – 223.
[46] D. W. Roberts, A. O. Aptula, G. Patlewicz, *Chem. Res. Toxicol.* **2006**, *19*, 1228 – 1233.
[47] S. Zhang, A. Golbraikh, A. Tropsha, *J. Med. Chem.* **2006**, *49*, 2713 – 2724.
[48] A. Golbraikh, D. Bonchev, A. Tropsha, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 147 – 158.
[49] A. Kovatcheva, G. Buchbauer, A. Golbraikh, P. Wolschann, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 259 – 266.
[50] A. Kovatcheva, A. Golbraikh, S. Oloff, Y. D. Xiao, W. Zheng, P. Wolschann, G. Buchbauer, A. Tropsha, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 582 – 595.
[51] M. Shen, Y. Xiao, A. Golbraikh, V. K. Gombar, A. Tropsha, *J. Med. Chem.* **2003**, *46*, 3013 – 3020.
[52] M. Shen, A. LeTiran, Y. Xiao, A. Golbraikh, H. Kohn, A. Tropsha, *J. Med. Chem.* **2002**, *45*, 2811 – 2823.
[53] M. Shen, C. Beguin, A. Golbraikh, J. P. Stables, H. Kohn, A. Tropsha, *J. Med. Chem.* **2004**, *47*, 2356 – 2364.
[54] S. Zhang, A. Golbraikh, S. Oloff, H. Kohn, A. Tropsha, *J. Chem. Inf. Model.* **2006**, *46*, 1984 – 1995.
[55] A. Golbraikh, M. Shen, Z. Xiao, Y. D. Xiao, K. H. Lee, A. Tropsha, *J. Comput. Aided Mol. Des.* **2003**, *17*, 241 – 253.
[56] L. Eriksson, J. Jaworska, A. P. Worth, M. T. Cronin, R. M. McDowell, P. Gramatica, *Environ. Health Perspect.* **2003**, *111*, 1361 – 1375.
[57] T. I. Netzeva, S. A. Gallegos, A. P. Worth, *Environ. Toxicol. Chem.* **2006**, *25*, 1223 – 1230.
[58] L. Sachs, *Handbook of Statistics*, Springer, Heidelberg, **1984**.
[59] J. L. Medina-Franco, A. Golbraikh, S. Oloff, R. Castillo, A. Tropsha, *J. Comput. Aided Mol. Des.* **2005**, *19*, 229 – 242.
[60] S. Oloff, R. B. Mailman, A. Tropsha, *J. Med. Chem.* **2005**, *48*, 7322 – 7332.
[61] S. Zhang, L. Wei, K. Bastow, W. Zheng, A. Brossi, K. H. Lee, A. Tropsha, *J. Comput. Aided Mol. Des.* **2007**, *21*, 97 – 112.
[62] J. H. Hsieh, X. S. Wang, D. Teotico, A. Golbraikh, A. Tropsha, *J. Comput. Aided Mol. Des.* **2008**, *22*, 593 – 609.
[63] H. Tang, X. S. Wang, X. P. Huang, B. L. Roth, K. V. Butler, A. P. Kozikowski, M. Jung, A. Tropsha, *J. Chem. Inf. Model.* **2009**, *49*, 461 – 476.
[64] Y. K. Peterson, X. S. Wang, P. J. Casey, A. Tropsha, *J. Med. Chem.* **2009**, *52*, 4210 – 4220.
[65] MolconnZ. http://www.edusoft-lc.com/molconn/, **2010**.
[66] W. Zheng, A. Tropsha, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 185 – 194.
[67] S. J. Cho, W. Zheng, A. Tropsha, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 259 – 268.
[68] A. Tropsha, W. Zheng, *Curr. Pharm. Des.* **2001**, *7*, 599 – 612.
[69] A. Tropsha, in *Cheminformatics in Drug Discovery* (Ed: T. Oprea), Wiley-VCH, Weinheim, **2005**, pp. 437 – 455.
[70] A. O. Aptula, D. W. Roberts, M. T. D. Cronin, T. W. Schultz, *Chem. Res. Toxicol.* **2005**, *18*, 844 – 854.
[71] T. W. Schultz, G. D. Sinks, L. A. Miller, *Environ. Toxicol.* **2001**, *16*, 543 – 549.
[72] T. W. Schultz, M. T. Cronin, T. I. Netzeva, A. O. Aptula, *Chem. Res. Toxicol.* **2002**, *15*, 1602 – 1609.
[73] T. W. Schultz, T. I. Netzeva, M. T. Cronin, *SAR QSAR Environ. Res.* **2003**, *14*, 59 – 81.
[74] T. W. Schultz, T. I. Netzeva, M. T. Cronin, *SAR QSAR Environ. Res.* **2004**, *15*, 385 – 397.
[75] T. W. Schultz, T. I. Netzeva, D. W. Roberts, M. T. Cronin, *Chem. Res. Toxicol.* **2005**, *18*, 330 – 341.
[76] T. W. Schultz, *Chem. Res. Toxicol.* **1999**, *12*, 1262 – 1267.
[77] T. W. Schultz, M. Hewitt, T. I. Netzeva, M. T. D. Cronin, *QSAR Comb. Sci.* **2007**, *26*, 238 – 254.
[78] P. Gramatica, *QSAR Comb. Sci.* **2007**, *26*, 694 – 701.
[79] A. M. Doweyko, *J. Comput. Aided Mol. Des.* **2008**, *22*, 81 – 89.

Received: May 28, 2010
Accepted: June 8, 2010
Published online: July 6, 2010

