

Artificial Intelligence and Data Mining for Toxicity

Prediction

* C. Helma and J. Kazius

December 13, 2005

Abstract

Tools for artificial intelligence and data mining can derive (Quantitative) Structure–Activity Relationships ((Q)SARs) for toxicity in an objective and reproducible manner. This review provides a conceptual description of the most important data mining algorithms for the identification of chemical features and the extraction of relationships between these descriptors and toxic activities. We will discuss the compliance of these techniques with the OECD guidelines for (Q)SAR requirements as well as performance implications. Special emphasis will be given to validation procedures for (Q)SAR models.

Keywords Predictive Toxicology, QSAR, Artificial Intelligence, Data Mining, Machine Learning, Pattern Recognition, Data-Driven Learning, Chemoinformatics

Abbreviations

1D/2D/3D One/Two/Three-Dimensional

AD Applicability Domain

AI Artificial Intelligence

ANN Artificial Neural Network

CoMFA Comparative Molecular Field Analysis

CV Cross-Validation

kNN k-Nearest Neighbour

MLR Multiple Linear Regression

OECD Organisation for Economic Cooperation and Development

PCA Principal Component Analysis

PCR Principal Component Regression

PLS Partial Least Squares / Projection to Latent Structures

(Q)SAR (Quantitative) Structure–Activity Relationship

RP Recursive Partitioning

SVM Support Vector Machine

1 Introduction

The goal of predictive toxicology is to accurately predict adverse effects of chemicals that lack experimental data. *In silico* predictive toxicology techniques are fast and cheap alternatives to *in vivo* and *in vitro* toxicity bioassays as they require neither experimental materials nor a physically available compound. For this reason, they are applicable in all cases where potential toxicity has to be evaluated rapidly and/or with limited resources (e.g. early toxicity screening of drug candidates, screening of all untested compounds in production). In addition, they can be useful for the prioritisation of compounds for toxicity bioassays and for the generation and evaluation of hypotheses about toxicological mechanisms.

Artificial Intelligence (AI) techniques are gaining increasing popularity in this area because they can deal with very complex relationships between chemical structures and toxic activities (*Quantitative*) *Structure–Activity Relationships ((Q)SARs)*. The important research topics of *Expert Systems* [1] and *Data Mining* [2] are relevant for predictive toxicology and are strongly related to artificial intelligence.

An expert system is a program that mimics human reasoning. An expert system for toxicity endpoint(s) can be developed by constructing a knowledge base, e.g. from interviews with human experts. The implementation of a knowledge base requires a close collaboration between toxicological, chemical and computer science experts and can be limited by improperly formalised or incomplete knowledge. This process is resource-intensive and the resulting knowledge base requires continuous updates to account for corrections, new scientific evidence and user experience.

Due to space constraints we will not cover expert systems for the rest of the article, readers who are interested in this topic can refer to a recent article by Parsons and McBurney [1].

This review will focus on methods that derive predictions from a dataset with experimental toxicity data. Most people use *Machine Learning*, *Pattern Recognition*, artificial intelligence and data mining to describe related concepts. These areas overlap heavily with statistics since all fields study the analysis of data. Here, we will mostly handle the term data mining for the computational extraction of useful information from data.

It is beyond the scope of a single paper to give an exhaustive review of current applications of data mining techniques for toxicity prediction. Instead, we intend to provide a conceptual survey of present techniques, discuss their capabilities and limitations and stress the important subject of model validation.

Readers who are interested in a more detailed description of the presented algorithms can consult a recent book [3], its companion website (www.predictive-toxicology.org) or refer to introductory data mining and machine learning literature [2, 4].

In this article we will define the most important steps for the construction and validation of predictive models in Section 2. Section 3 will briefly cover the data preprocessing step and Section 4 will discuss the available repertoire of features (*descriptors*) for representing the chemicals in the dataset. The fifth section will deal with the construction of predictive models from these features and Section 6 will highlight the requirements and pitfalls of evaluating the predictive performance of trained models.

2 Ingredients and requirements for a predictive toxicology system

Data mining methods, including classification and regression methods, aim to extract relationships between chemical structures (or their properties) and toxic activities ((Q)SARs). They objectively derive (Q)SARs from *training data* with chemical structures and experimentally determined toxicity data (Section 5). Predictions can be obtained from the application of these (Q)SARs to structures with unknown activities. Moreover, the resulting relationships can provide an initial knowledge base for expert systems and help to detect hypotheses about toxicological mechanisms. Despite these capabilities, (Q)SARs do not intend to model all biochemical and physiological processes that lead to a particular toxic effect.

Currently, most predictive toxicology techniques use the following procedure to generate (Q)SAR models from experimental data:

1. Selection and preprocessing of a dataset with a well-defined toxicity endpoint (Section 3: *Dataset preprocessing*)
2. Identification and calculation of features (*descriptors*), such as substructures or properties, that are relevant for toxicity (Section 4: *Chemical representation*).
3. Detection of relationships between these features and toxic activities, i.e. (Q)SAR models (Section 5: *Construction of classification and regression models*).
4. Evaluation of the predictive performance of the model and of its learned rela-

tionships (Section 6: *Validation of (Q)SAR models*).

5. Interpretation of these relationships in terms of toxic mechanisms (Sections 4 and 5).

Data mining plays an important role in this procedure. Recent data mining techniques can e.g. identify features that are relevant for toxicity in a comprehensive and unbiased manner (Section 4) and use them for the identification of complex Structure–Activity Relationships (Section 5). Many (Q)SAR investigations use a hybrid approach by determining relevant features via expert knowledge before applying computational regression methods.

Data mining algorithms are independent of the investigated endpoint. It is therefore possible to model every conceivable toxicity endpoint as long as sufficient experimental data is available. The use of human data (e.g. from epidemiological or clinical studies and adverse effect registries) offers exciting ways of predicting human health effects directly without surrogate endpoints.

The requirements for a specific predictive toxicology system will depend to a large extent on its application area. Regulatory applications for example, will need accurate predictions that can be interpreted in terms of current toxicological knowledge, but they can compromise on speed because relatively few compounds have to be processed. Interpretable predictions are on the contrary not very important for high throughput screening, but predictions have to be calculated with high speed for a large number of structures.

As most predictive toxicology applications will benefit from their consideration, we will present a brief overview of the *OECD Principles for Validating (Quantitative)*

Structure-Activity Relationship ((Q)SAR) Models [5].

The OECD requires the following information to facilitate the consideration of a (Q)SAR model for regulatory purposes:

- a defined endpoint
- an unambiguous algorithm
- a defined domain of applicability
- appropriate measures of goodness-of-fit, robustness and predictivity
- a mechanistic interpretation

A personal view of the authors on these topics will be presented in the following subsections. The implications of these OECD requirements on the individual steps of data mining will be discussed in Sections 3 – 6.

2.1 Defined endpoint

This principle shall ensure clarity about the endpoint being predicted and identify the experimental system and protocol that is being modelled by the (Q)SAR. As experimental protocols influence the outcome of toxicity assays, it is vital to specify and document them exactly (e.g. in accordance with good laboratory practice). If data from multiple assays are aggregated into a single endpoint (e.g. “*Salmonella* Mutagenicity” from Ames tests with different strains and activation schemes) it is essential to clarify the exact procedures for data aggregation. Section 3 will give an overview of preprocessing procedures that can be applied to transform real world data into a well-defined dataset of compounds.

2.2 Unambiguous algorithm

This principle shall guarantee the transparency of the modelling algorithm, which ensures that predictions can be recalculated independently from a specific program implementation. Although most data mining algorithms are well documented in the literature, it is important to specify the exact algorithm and parameter settings for each step (feature identification, model fitting (random seed, if any), calculation of predictions) as well as all problem-specific modifications of the original algorithm. If a program uses modifications that cannot be reimplemented with standard data mining tools it should be possible to obtain the source code for a closer inspection. This requirement can be problematic for companies that do not want to disclose the internal working of their software releases. Nondisclosure will likely represent a barrier for regulatory acceptance.

2.3 Defined domain of applicability

As (Q)SAR models are based on training datasets that cover only a small fraction of the chemical universe, they can provide reliable predictions only for a limited set of similar structures. The *applicability domain (AD)* is determined by the size and composition of the training set as well as by the capabilities of the descriptors and the modelling algorithm. It is crucial to provide formal definitions for applicability domains in order to minimise unreliable predictions for compounds that fall beyond the scope of a (Q)SAR model.

However, up to now, few (Q)SAR models have defined their applicability domain. Recent studies have shown that similarity to molecules in the training set can be a

good general purpose AD estimator [6, 7, 8], which can even be applied in retrospect to existing (Q)SAR models. AMBIT, a descriptor-based tool for the AD assessment of (Q)SAR models can be accessed at <http://ambit.acad.bg/run.php>. Limited applicability domains have important consequences for the validation of (Q)SAR models. This topic will be covered in Section 6.

2.4 Appropriate measures of goodness-of-fit, robustness and predictivity

Appropriate procedures for model validation are still a topic of hot discussions in the (Q)SAR community. For this reason, the OECD guidelines provide relatively little specific advice, apart from distinguishing between internal and external model performance. As there is presently no consensus about appropriate validation procedures, Section 6 will present the authors' view on this subject.

2.5 Mechanistic interpretation

Although it is not the intention of data mining techniques to model biochemical processes that lead to toxic events, many (Q)SAR models can provide information that is interpretable in terms of toxic mechanisms. In most cases this will be an association between chemical features and the endpoint being predicted. The accessible information depends on the features that are used to describe chemical characteristics, the data mining algorithm itself and the representation of its results. The possibilities for mechanistic interpretations for various types of descriptors and models will be discussed in Sections 4 and 5.

3 Data preprocessing

Real world toxicity datasets are usually not created especially for (Q)SAR purposes. For this reason, most datasets require preprocessing to obtain a dataset that is consistent and valid for the investigated endpoint. A recent review summarises quality requirements for toxicity data and gives an overview of high quality toxicity data sources [9]. The DSSTox database network (<http://www.epa.gov/nheerl/dsstox/> [10]) is a source for public datasets that have been preprocessed especially for (Q)SAR studies.

Large datasets increase the overall chance of finding statistically significant structure–toxicity relationships and broaden the applicability domain of the resulting descriptive or predictive model. Some (Q)SAR studies on the other hand focus on a specific toxic interaction of one chemical class and thus have a very limited applicability domain.

Estimates of experimental reproducibility can ensure the quality of the training set and provide a benchmark for performance evaluations. If it is economically or ethically unfeasible to perform interlaboratory reproducibility studies, such estimates can be obtained from reproducibility studies or replicate datasets in the literature, as long as the protocols are comparable. Additionally, concordance estimations are needed if data from different protocols or bioassays are combined into a single endpoint for modelling.

Compounds in most real world datasets have been tested because of existing toxicological concerns, high potential exposures or regulatory requirements. As a result, they contain atypical compounds as well as chemical and toxicological

classes that are overrepresented.

3.1 Removal of atypical compounds and duplicates

Real world datasets contain a surprisingly large fraction of incorrect structures. This can result from a wrong representation of the structure or from the confusion with another compound. Most computational chemistry packages can be used to check the correct syntax of structures, but the other cases are harder to spot. An inspection by a chemist is usually required because the appropriate structure can be identified only from the original experimental data (e.g. from nomenclature, synonyms, trade names or CAS numbers [11]). Depending on the aim of the study, peptides, inorganic and/or organometallic compounds can be removed from the dataset. And, for scientific or technical reasons, mixtures can be removed as well as compounds without completely defined structures (e.g. polymers).

Duplicate compounds can unbalance the distribution of chemicals in the dataset, which is problematic for some predictive toxicology techniques. Unfortunately, the computers perception of duplicates does not necessarily reflect the chemists perception. The perception of duplicates can furthermore depend on the applied (Q)SAR technology. It may be necessary for some techniques, to remove counterions if they do not affect the experimental outcome. This may lead to duplicates that did not exist in the original database and it is necessary to define how to deal with them. A related problem is that a single compound can be described by multiple virtual states. Prior to comparison, all molecular structures should therefore be standardised in terms of protonation state, tautomerism, formal charges and resonance state [11, 12]. If isotope atoms are expected not to influence the ex-

perimental result, they should be converted into normal atoms prior to the detection of duplicates. If only 2D features are used during the chemical representation step, then stereoisomers will have to be considered as duplicates.

Only after these standardisation steps all duplicates can be removed, which is required for most (Q)SAR techniques. If conflicting experimental results are encountered for a single compound, it must be justified and documented how a single experimental outcome is selected.

A different problem is that some compounds can differ from the bulk of other retrieved compounds in terms of molecular weight, reactivity, functional groups, solubility or other characteristics. Although such atypical compounds can be excluded with justification, their exclusion will reduce the applicability domain of the resulting model. For instance, if only organic, drug-like compounds are considered during model construction, then the model cannot be expected to make reliable predictions for inorganic or non-drug-like compounds.

4 Chemical Representation

Although some data mining methods can be applied directly to chemical structures [13] or to relational representations of compounds [14], most techniques require that each compound is represented as a row of features. Note that it is also possible to use measurements of biological short-term assays as chemical features for the low-throughput assessment of a more complex toxicity endpoint. It is important to consider that no statistical or data mining method is able to extract relationships from information that was not provided to it. Therefore, the most im-

portant bottleneck of relating chemical structures to experimental activities lies in the representation of compounds in the form of features.

4.1 Experimentally determined properties

Experimental data of physicochemical properties such as acid-base dissociation constant (pK_a) octanol/water partition coefficient ($\log P$ or $\log K_{ow}$) or water solubility ($\log S$) are often relevant for toxicity endpoints and should thus be taken into account. Although physicochemical properties are much faster and cheaper to determine than toxicity endpoints, the availability of experimental measurements is still a limiting factor for the creation of (Q)SAR models. The search for structure-toxicity relationships therefore typically relies on properties that can be calculated directly from molecular structures.

4.2 Computed properties

(Physico)chemical characteristics of a compound can be computed from its atoms (1D), structure (2D) or conformation (3D). The “Handbook of Molecular Descriptors” by Todeschini and Consonni [15] provides an encyclopaedic reference to molecular descriptors that are suitable for (Q)SAR studies. These features can estimate local characteristics, such as LUMO energy, or whole-molecule characteristics, such as flexibility indices [15, 16]. While some features are easy to interpret in terms of biochemical mechanisms (e.g. molecular weight, $\log P$, number of hydrogen acceptors), others present more difficulties for the average toxicologist (e.g. topological indices [15, 16, 17]).

Given the abundance of molecular descriptors, it is difficult to make an *a pri-*

ori selection that is optimal for a particular endpoint (Sections 5 and 6). Although feature selection techniques can help with the identification of relevant descriptors (Section 5.2), the danger is to miss important properties that have not been calculated in the first place.

To overcome such limitations, several techniques have been developed that allow an exhaustive search of training structures for classes of interpretable substructures, which can then be related to toxicity.

4.3 Substructures in standard chemical notation

The 2D structure of a compound consists of atoms, identified by their atom type (C, N, O, etc.), that are connected through bonds. In principle, this chemical structure, or chemical graph, can be used to determine all occurring atoms and substructures. Substructure search techniques have gained popularity during the last years as they can objectively derive all features of a certain type from the training compounds, such as linear chains of atoms, pharmacophore points and substructure graphs. Considering their 2D structure, substructures can have different degrees of complexity, namely linear (chains of atoms), tree-shaped (with branches, but without cycles) and substructures of any shape. Various programs exist that can mine moderately [18, 19] and considerably [8] sized toxicity datasets for all linear substructures within acceptable time limits. The number of distinct non-linear substructures (with branches and cycles) in the same dataset, on the other hand, is several orders of magnitude higher. Calculating them has been a performance problem, but recent algorithmic advances allow us to rapidly mine datasets of thousands of compounds for linear, branched and cyclic substructures of any

size [12].

Substructure mining algorithms are designed to extract all substructures from a dataset of chemical graphs that satisfy one or more predefined constraints. As the most frequently applied constraints concern frequencies (e.g. in toxic vs. non-toxic compounds), this step is also called frequent graph mining. Substructure mining starts with single atoms that are extended atom by atom until the extended substructure is not frequent enough on the given dataset. For efficiency reasons, it is essential to prevent duplicate searches for the same substructure as well as searching for infrequent or non-existent substructures. This can be achieved with search algorithms that prevent the extension of infrequent substructures, use canonical forms to prevent repeated searches and use caching techniques to reduce the search space. In addition, the search process can be split into dedicated search phases for linear, branched and cyclic substructures [12, 20].

4.4 Substructures in elaborate chemical representation

Characteristics of atoms depend on more than only their atom types. For example, the environment of a nitrogen atom determines whether or not it is aromatic and whether it is predominantly cationic, uncharged or anionic in water of neutral pH . As a result, it is advantageous to detect similarities between atoms or substructures as these similarities are not considered in the standard chemical notation. The SMARTS language (www.daylight.com/dayhtml/doc/theory/theory.smarts.html) was developed to precisely specify characteristics of atoms and larger substructures, thus enabling detailed substructure searches. The SMARTS language is particularly useful for identifying atoms or substructures that share similar

characteristics in terms of aromaticity, electronegativity, hydrogen bonding ability and more. A recent publication describes a preprocessing method that enables the incorporation of such characteristics directly in the substructure mining step [12]. Apart from its completeness and objectivity, an advantage of this method is that assumptions regarding the protonation state of individual heteroatoms can be circumvented by using multiple descriptions of a single atom. In theory, additional types of substructure-specific data can be considered, such as atom- and substructure-based estimates of various types of physicochemical data [21, 22].

4.5 3D-based features

Reaction mechanisms and interactions of compounds with biological target molecules (e.g. proteins or DNA) are at least three-dimensional in nature. This could be a reason to prefer 3D-based methods over simple 1D- and 2D-based representations. Nevertheless, 3D methods do not outperform 2D methods for most toxicological endpoints. This can be the result of high biological variability, but it might also be the consequence of problems that are inherently associated with 3D feature calculations.

The values of 3D-based features depend on the exact 3D conformation. This dependency is probably their most important drawback as one compound can adopt significantly different conformations. Computational chemistry tools can estimate one or more energy-minimised conformations for each compound [23], but the relevance of each conformation for toxicity remains unknown and the computations depend on several assumptions (e.g. about protonation states of heteroatoms).

3D features can be global properties of structures (e.g. hydrophobic surface

area), but ideally they should point to the part of the surface that is relevant for toxicity (e.g. metabolic activation). 3D-(Q)SAR techniques like CoMFA [24] align all structures to a reference compound and derive descriptors from a grid of measurement points, but this only works for compounds with a large common substructure. 3D substructure mining would be a logical extension of future investigations into subgraph mining.

In general, features that require fewer assumptions should be preferred over features that assume, for instance, a 3D conformation or a protonation state. Features that express chemically relevant information and that can locate critical parts of the chemical structure can provide interpretable structure–toxicity hypotheses. Such hypotheses can facilitate the design of follow-up experiments. Moreover, they are better accepted by experts than SARs with very abstract descriptors.

5 Construction of classification and regression models

This section will present data-driven methods that aim to derive predictive models from training sets with chemical features and toxic activities. We will briefly review the methodologies of the most important algorithms and discuss their suitability in terms of runtime implications and interpretability. The OECD guidelines for a defined applicability domain and appropriate measures of goodness-of-fit will be discussed in Section 6. For the sake of simplicity, we will focus on algorithms in terms of classification (e.g. distinction between toxic/nontoxic substances), but most of the presented techniques can be adopted for regression purposes (e.g. prediction of LD_{50} values). Algorithmic details for the pre-

sented techniques can be found in a recent book [13, 25, 26] as well as in general introductions to machine learning and data mining [2, 4]. Most of the described techniques have been implemented in many popular statistical, data mining and chemoinformatics packages as well as in specific predictive toxicology tools. A non-comprehensive list of links to some of these tools can be found at <http://www.predictive-toxicology.org/programs.html>.

5.1 Basic ideas of learning theory

The purpose of a predictive toxicology model is to predict the toxicity of untested compounds on the basis of existing experimental data of other compounds (*training data*). The previous section discussed different features, or descriptors, for the representation of chemical characteristics. Now we will present techniques for the identification of prediction models that relate chemical descriptors to toxicity values.

More formally speaking, this involves seeking a function for predicting new (unseen) cases. A learning algorithm identifies this function by searching in a set of suitable functions (the *hypothesis space*) in order to identify a function that minimises the *empirical error* (i.e. the difference between predictions and real values).

Intuitively, models that are complex (in terms of function complexity and/or the amount of descriptors considered) can fit almost any set of training data with high accuracy. Such models, however, behave poorly on future structures as they are unable to extract general relationships from the training data (high *generalisation error*). This phenomenon is commonly referred as *overfitting*, as opposed to *underfitting* where models fail to represent a good solution because their model fitting

procedure is not complex enough or because their descriptors are inadequate.

5.2 Selection of relevant features

Theoretical considerations and practical experience show that a large number of irrelevant and/or highly correlated features deteriorate the performance of data mining algorithms. Without detailed knowledge of all involved biological mechanisms it is, however, hard to identify which features are relevant by expert knowledge alone. It is even harder to ascertain *a priori* that no relevant features are missed.

Automated techniques exist for the removal of correlated and irrelevant features. As opposed to supervised techniques, unsupervised techniques do not consider toxic activities as selection criteria. Simple tests can, for instance, be used to remove features that have identical values for all training examples and features that are highly inter-correlated. A more elaborate unsupervised method, called *Principal Component Analysis (PCA)*, transforms the initial set of features into a smaller, uncorrelated set of descriptor-based functions. However, these functions, or principal components, are more difficult to interpret than the individual features.

With supervised techniques it is not only possible to produce a smaller set of features, but also to determine the relevance of descriptors for a particular toxicity endpoint. This can be achieved with simple statistical filters that decide, for example, whether a feature occurs more frequently in toxic than in nontoxic structure (*Filter Methods*). A more time-consuming strategy is the application of various optimisation techniques to select the most useful features for a particular classification or regression algorithm (e.g. genetic and evolutionary algorithms [27, 28, 29]). Overall, they work by eliminating features that are redundant with respect to other

features or by randomly selecting a small feature set, which is then grown or mutated until a termination criteria has been reached. These techniques only use statistical criteria to determine which of all correlated features are selected. They can therefore converge towards suboptimal selections of descriptors, especially if the fitness function is not handled properly. As the results of most optimisation techniques depend on stochastic processes, several runs will yield slightly different results.

Backward elimination is a very general and efficient way to select relevant descriptors for various learning algorithms [30]. In each iteration, a certain fraction of the most irrelevant features are removed from the training set until the predictive performance of the derived (Q)SAR model drops significantly.

Data mining techniques can identify mathematically independent and relevant features in an unbiased manner, but they can also contribute to overfitting (especially optimisation and backward elimination techniques). As a result, it is crucial to validate the robustness of these techniques properly. It is essential to perform feature selection only on the training data and to exclude all test set information from this process (Section 6).

5.3 Generalised linear models

Generalised linear models use statistical regression techniques to minimise the difference between predictions and real values. *Multiple Linear Regression (MLR)* [25] has been the workhorse for (Q)SAR model development during the last decades. It attempts to identify a linear function that relates descriptors to toxicity values. MLR assumes a normal distribution, an independence of descriptors and a linear

relation between descriptors and the toxicity endpoint.

As MLR uses a relatively simple error minimisation algorithm models that can be generated in short time. The final model is a portable, linear (Q)SAR equation that can make predictions for new instances with high speed, even with a pocket calculator.

MLR cannot capture nonlinear relationships and does not work well with correlated descriptors or with many (irrelevant) descriptors. To overcome the latter drawbacks, feature selection can be applied (e.g. with an optimisation technique [27, 28, 29]). Alternatively, a PCA transformation of features (and the endpoint) can be performed prior to MLR (*Principal Component Regression (PCR)* (or *Partial Least Squares (PLS)*)).

Several extensions have also been developed that account for different distributions (e.g. logistic regression) or introduce nonlinear relationships (e.g. polynomial regression). Given a small set of relevant descriptors, overfitting is rarely a problem as MLR cannot generate very complex functions. Underfitting can occur if the structure–toxicity relationships are not linear.

The interpretation of an MLR model is relatively straightforward because many researchers are already familiar with the resulting linear (Q)SAR equation with descriptor-specific weights.

5.4 Bayesian techniques

Naive Bayes (NB) is a conceptually and computationally simple learning scheme that estimates the probability for toxicity based on the presence or absence of structural features [4, 26]. The contributions of individual features to toxicity are

estimated from experimental data by dividing the number of toxic compounds with a particular feature with the number of all compounds (i.e. toxic and nontoxic) with this feature.

For classification, NB assumes an independence of features and calculates the overall probability for toxicity by multiplying the contributions of all fragments with the default probability for toxicity and does the same for non-toxicity. A compound is classified as toxic if the probability for toxicity exceeds the probability for non-toxicity.

Naive Bayes can rapidly generate models because it requires only a single scan through the database to count the occurrences of features in each class. Predictions are also fast because of the simplicity of the classification model.

The assumption of feature independence is rarely fulfilled in a (Q)SAR setting, but it can be obtained with feature selection procedures like PCA. The crucial problem for practical applications is however the determination of *a priori* probabilities for toxicity, which have a high impact on future predictions. If the distribution of toxic/nontoxic structures deviates in future predictions from the training set, a large number of false predictions are the consequence. It is, however, impossible to detect this effect with test sets that have a similar composition as the training set (Section 6).

It is relatively easy to interpret NB models and individual Bayesian predictions because the probabilities of features indicate their importance and contribution to toxicity.

5.5 Recursive partitioning

Recursive partitioning (RP) algorithms use a “divide-and-conquer” approach to split the training data into subsets with (almost) identical classes [2, 4, 26]. They start e.g. with a search for the substructure that provides the best separation between toxic and nontoxic compounds. As it is usually impossible to provide a clear cut separation with a single feature, the search proceeds with the identification of other substructures that provide a better separation of the subsets. Although the same procedure can be repeated until the complete training set has been fitted, a predefined stopping criterion is normally used to prevent overfitting. The concept can be easily extended to account for numerical descriptors and regression problems.

For the classification of a new compound, the descriptor-based splitting criteria of the (Q)SAR model are repeatedly applied to determine its toxicity class.

Compared to the previously mentioned techniques, the model fitting process is somewhat more time consuming, because RP needs to search recursively over the descriptor space. The prediction of new structures is however still relatively fast, because only simple splitting criteria have to be applied.

Because a relaxed stopping criterion promotes overfitting and hinders interpretability (by making the model larger and thus more complex), the stopping criterion should be quite strict. The results of RP algorithms can be presented visually in various ways, most frequently as *Decision Trees* and *Rules*. Although it is relatively easy to interpret small decision trees and rules, practical experience shows that most toxicological experts have problems to interpret larger models in terms of toxicological mechanisms [31].

5.6 Artificial Neural Networks

Artificial Neural Networks (ANNs) try to mimic the working of biological neural networks by feeding input (descriptor) data through interconnected computational units (neurons) to an output unit (toxicity values) [13]¹. Typically, there are at least three layers in an ANN - an input layer, a hidden layer, and an output layer. The input layer does no processing - it simply feeds data into the hidden layer. The hidden layer, in turn, feeds into the output layer. The actual processing in the network occurs in the nodes of the hidden layer and the output layer.

The most popular approach to train an ANN is the *backpropagation* algorithm. During the training phase, the algorithm adjusts the weights of each connection in order to reduce the difference between predicted and observed output values. After repeating this process for a sufficiently large number of training cycles, the network will usually converge to a state with a small prediction error. In contrast to the approaches discussed so far, this process is not deterministic (i.e. several runs of the algorithm lead to slightly different results) because it depends on the starting condition and stochastic processes. A typical problem of the back-propagation algorithm is the possibility to end up in a local minimum of the error function.

Predictions are obtained by feeding the descriptors of the new structure through the optimised network. They are relatively fast with making predictions, but the speed of convergence of the back-propagation algorithm can pose efficiency problems.

¹The `comp.ai.neural-nets` FAQ (<http://www.faqs.org/faqs/ai-faq/neural-nets/part1/index.html>) is a good starting point for finding more information about ANN theory and implementations.

Neural networks are *universal approximators*, i.e. the associated hypothesis space can be made arbitrarily large by adding hidden units. This means that underfitting will not occur if enough hidden units are used and the complete set of descriptors is sufficiently relevant. If the capacity of the network is too large, overfitting will be a problem because the number of examples is insufficient to restrict the hypothesis space. For this reason, it is necessary to provide strict techniques to avoid overfitting (e.g. weight decay, early stopping).

Due to the complexity of the models, the interpretation of ANNs is not straightforward. It is, however, possible to use extraction techniques to generate lists of important features and similar compounds [32] or to convert an ANN to a better understandable representation like rules [33].

5.7 Support Vector Machines

Support Vector Machines (SVMs) analyse the high-dimensional descriptor space for a hyperplane that separates toxic from nontoxic structures [34]². They choose the *maximum margin hyperplane* that maximises the distance from the closest training examples (*support vectors*). Unlike ANN, backpropagation and optimisation methods (simulated annealing and genetic and evolutionary algorithms), a given SVM will always deterministically converge to the same solution for a given data set.

For the classification of a new structure, SVM models determine if the structure falls on the toxic or on the nontoxic side of the decision surface.

SVMs are considered to be slower at run-time than other techniques with similar

²A detailed description of SVM learning theory can be found in several books and tutorials, most of them are listed on <http://www.kernel-machines.org/> together with links to SVM implementations.

generalisation performance (e.g. ANNs). Once a model has been trained predictions can be obtained relatively fast.

The original SVM algorithm was developed for the identification of linear hyperplanes. Without changing this rest of the algorithm, SVMs were adopted for nonlinear cases through a simple mathematical trick that substitutes dot products with nonlinear kernel functions. This increases the expressive power of the SVMs but it also leads to a greater risk of overfitting training data and a poorer generalisation performance.

Linear SVM models can be transformed into equations that assign weights to individual features. Although the support vectors can provide some information about structurally similar compounds, other types of SVMs are essentially black-box classifiers because of their mathematical complexity.

5.8 k-Nearest Neighbour methods

The k-Nearest Neighbour (kNN) technique is different from the previously described algorithms as it does not attempt to find a model that is valid for all cases, but derives predictions directly from the examples in the training set [2, 4]. For this purpose kNN searches in the training set for compounds with structures that are similar to the test structure (*neighbours*) and predicts according to the weighted activity of the closest k compounds.

kNN skips the model fitting step, but requires more efforts for predictions than other methods because the complete training set has to be consulted for each prediction instance. For this reason, prediction runtimes increase linearly with the size of the training set.

As kNN methods do not train a model that relates descriptors to activities, there is no risk of over- or underfitting the training set. kNN methods make no assumptions about the relation between descriptors and toxic activities, but base their prediction on the hypothesis that similar chemical structures have similar toxic properties.

As a result, an adequate definition of chemical similarity is crucial for the success of kNN techniques. Similarity can be measured in terms of model-specific molecular descriptors [7, 35] or in terms of *chemical similarity*, which is based on model-independent features [36, 37, 6]. As these similarity measures consider the complete chemical structure independent of their relevance for toxicity, a recent publication [8] has proposed a procedure that considers only features that are relevant for toxicity. Apart from classification and regression tasks, the determination of similarity can play an important role for the definition of applicability domains (Section 2).

kNN presents the rationales for predictions in terms of structurally similar compounds, which is close to the arguing of toxicological experts in the absence of experimental evidence. Clues for toxicological mechanisms can be obtained by a literature search on these neighbours and predictions can be easily rejected if they are implausible from a toxicology standpoint.

5.9 Further learning techniques

The previous sections gave only a brief overview of the most important data mining techniques. Extensions and special purpose modifications exist for most of these algorithms as well as hybrid approaches, which combine different approaches. Fur-

thermore, unsupervised data mining methods (e.g. clustering, self organising maps [38]) do not consider toxic activities, but can be used to place a new structure into a group of compounds with similar chemical properties.

6 Validation of (Q)SAR models

The proper validation of (Q)SAR models is still a controversial topic. For this reason, the following statements represent predominantly the authors' view on this subject. Many of the arguments are supported by two recent papers by Tropsha [35] and Hawkins [39] as well as by discussions within the data mining and (Q)SAR communities.

6.1 Impact of the test set

Validation procedures aim to estimate the accuracy that can be expected for *future* predictions of a (Q)SAR model. In principle, the concordance between predictions and measured activities can be evaluated with a *test set* that is a *representative sample* of future predictions (e.g. potential drugs, high production volume chemicals, the "chemical universe", ...). As it is in most cases impossible to determine in advance which types of novel compounds will be predicted, real world test sets are in most cases *not representative* for future predictions. For this reason, the validation results for a single model can vary significantly from test set to test set [40].

The consequences can be illustrated by assuming a model that is capable to provide 100% accurate predictions for structures within its applicability domain (AD),

but that can only perform random guessing for compounds with new substructures and/or properties (i.e. for compounds beyond the AD). In this case, the validation accuracy will be determined predominantly by the fraction of test set compounds within the applicability domain, which will vary from test set to test set.

6.2 Separation of training data from test data

This extreme example illustrates that the composition of test sets is an important factor that influences the resulting predictivity estimates. But there are more factors that have an impact on validation results - fortunately they are easier to control.

By now there is a consensus that the recreation of training set results (sometimes called goodness-of-fit or internal validation) is no indication of the model's capability to predict unseen instances [35]³. Instead, validation has to be performed on a test set that is clearly separated from training examples. This means that all test set information has to be excluded from the training phase, not only the test set instances but also all derived information (e.g. relevant features, feature-specific weights and parameter settings). However, it is still frequently observed in (Q)SAR validation studies that test set information leaks into the training phase through e.g. feature-specific weights or model parameters.

If, prior to validation, the selection of features is based on the whole dataset (either by an automated procedure or by expert inspection), the distinction between training and test sets is blurred because the selected features contain test set infor-

³It is even possible to develop algorithms that per definition guarantees 100% accuracy on test set instances (e.g. 1-nearest neighbour, ANN's with a sufficient number of hidden neurons).

mation. A related problem occurs if parameter settings that were optimised for the full dataset are repeatedly applied in validation runs, that is, without recalculation for the training dataset. Also in this case, the selected “optimal” model parameters contain implicit information from the test set.

An incomplete separation between training and test sets leads to models that are overfitted for the validation test set [30]. The validation results will be overly optimistic, but the performance for really unknown prediction instances will degrade substantially. With data mining procedures these problems can be easily avoided, namely by using exclusively training set information for feature selection and parameter optimisations.

6.3 Validation with an existing test set

It is often unfeasible to test a representative sample of compounds solely for the purpose of (Q)SAR validation. For this reason, most (Q)SAR validation studies have to rely on existing data.

If two separate datasets exist for the same endpoint, it is possible to train the (Q)SAR model with one dataset and predict the activities of the other dataset as a validation exercise. Many experts consider this procedure as a gold standard for (Q)SAR validation and it has been frequently used for the comparison of (Q)SAR models [5, 41, 42]. However, there are also some pitfalls.

First of all, it is necessary to ensure that both datasets use comparable protocols (see “Defined Endpoint” in the OECD Guidelines) and that all instances from the training set are excluded from the test set.

Like any other dataset, the external test set will be a real world toxicity dataset,

which has been built for a specific purpose (Section 3). As a result, its chemical content differs from the set of structures that require predictions (e.g. potential drugs). The test set is then *no representative* sample of the future compound collection and the validation results can vary significantly with different test sets [40]. This can lead to over- or underestimations of the future performance. Chance correlations can be an additional confounder, especially for small test set sizes.

Nevertheless, when experts have been involved in the construction of the model (e.g. through feature selection or with an expert system), the use of a confidential external test set is the only method of validation because experts are unable to forget test set-related information and this is a requirement for cross-validation (see below). Even if experts select features from the general literature without examining the structures or their activities, an external test set can help to evaluate the features in an objective manner.

In the absence of two distinct datasets, it is necessary to split the available experimental data artificially into a training set for computational model fitting and a test set for the evaluation of predictions.

6.4 Validation with a single artificial test set

Splitting the available data into a single training and a single testing set (e.g. 2/3 of the examples are used as training set and 1/3 are used as a test set) is a very straightforward and frequently used approach. If the split is performed randomly, the composition and relevance of the test set is purely up to chance. Therefore several methods have been developed to ensure that the distribution of compounds in the test set is comparable to that of the training set in terms of toxic activities

and/or chemical features [35]. Such procedures, however, make the assumption that the chemical distribution within the training set is *representative* for future predictions. As a result, the overall predictive performance are likely overestimated.

A major problem with splitting the dataset is the trade-off between reliable validation results (i.e. large test sets) and complete models (i.e. large training sets). This can be more problematic for diverse datasets with few structures and for algorithms that, internally, require additional evaluation sets for parameter optimisations (e.g. ANNs).

6.5 Validation with multiple artificial test sets (Cross-Validation)

To obtain a more robust estimate of predictivity, the separation between training and test set can be performed repeatedly. The combination of several validation runs has the advantage that it is possible to report the performance for a larger number of structures, which is much more reliable than the results of a single validation run. As an additional benefit, the variability of predictions can be estimated from the different predictive performances.

In cross-validation (CV), all experimental data are separated into n (usually 5 or 10) parts (*folds*). For each validation run, one fold is removed as a test set from the dataset. The model is trained with the remaining examples of the dataset and then evaluated on the test set. The whole procedure is repeated n times until all examples from the training set have served as test set instance.

Like all other validation procedures, cross-validation is limited by the assumption that the tested instances (in this case the complete training set) are representative for future prediction instances. If test set information has leaked into

(Q)SAR model, then cross-validation gives no guarantee against an overestimated predictivity. To ensure proper cross-validation, feature selection and parameter optimisation have to be repeated separately for each fold. If this is the case, the n -fold cross-validation procedure provides a realistic assessment of both overall predictivity and test set dependent fluctuations.

6.6 Validation and applicability domain

Up to now, most of the discussions about validation techniques did not consider the consequences of the model's applicability domain (AD) on the capability to predict future instances. Although no reliable predictions can be expected for compounds beyond the applicability domain, most validation schemes treat the misclassification of these compounds equally to that of compounds within the applicability domain. This can distort the estimated overall predictivity because the proportion of compounds within and outside a model's applicability domain may vary largely between test sets.

Many disagreements about validation procedures possibly result from the neglect of applicability domains. For example, it is likely that an external test set (of a different origin than the training set) has a lower fraction of compounds within the applicability domain than a test set that has been drawn randomly from the training set. This may be an explanation for the frequent observation that the predictive accuracy is lower for an external test set than for cross-validation experiments. This hypothesis is substantiated by an experimental comparison of leave-one-out cross-validation with results for an external test set with several thousand structures, that gave very similar results for compounds within the AD [8].

Applicability domains do not have clear cut borders between predictable and unpredictable structures. Most methods provide a numerical value, such as a similarity measure of the test compound to the training set. This number can be used to weight the importance of individual predictions in validation study, or to set a threshold for the AD. An adjustment of the threshold for acceptable predictions provides an elegant solution to balance the scope of a model against the accuracy of its predictions for different application areas. For practical purposes, we advise to report the overall predictive performance, the predictive performance for compounds within the AD and the fraction of the test set(s) that fall within the AD. In general, the consideration of applicability domains in validation studies should lead to more realistic and less variable performance estimates than techniques that treat all predictions equally.

7 Conclusion

The procedure of deriving a predictive (Q)SAR model is a multi-disciplinary effort that, despite the possible pitfalls, holds great promises. Chemical and toxicological expertise aid the selection of descriptor classes that represent relevant chemical information as well as the interpretation and acceptance of the resulting models. Tools for artificial intelligence and data mining can derive (Q)SAR models from experimental toxicity data in an objective and reproducible manner. First of all, data mining can be used to calculate descriptors (e.g. substructures) for chemical structures. Secondly, various data mining techniques can be applied for the identification of relationships between these features and toxic activities. (Q)SAR models can

then make predictions for new compounds on the basis of the extracted relationships.

Irrespective of the applied techniques, it is necessary to validate models properly. A clear cut separation between training and test sets is crucial for the realistic estimation of predictive performance. The applicability domain has important consequences for the validation of a (Q)SAR model and the authors advocate a better consideration of applicability domains in future validation studies.

Acknowledgements

C. Helma was supported by a grant from the Centre for Documentation and Evaluation of Alternatives to Animal Experiments (ZEBET). We thank Prof. J. Kok, Prof. T. Bäck and Prof. A.P. IJzerman for proofreading.

References

- [1] Parsons, S.; McBurney, P. The Use of Expert Systems for Toxicology Risk Prediction. In *Predictive Toxicology*; Helma, C., Ed.; Taylor & Francis: Boca Raton, 2005.
- [2] Witten, I.; Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques*; Morgan Kaufmann Publishers: San Francisco, California, 2000.
- [3] Helma, C., Ed.; *Predictive Toxicology*; Taylor & Francis: Boca Raton, 2005.
- [4] Mitchell, T. M. *Machine Learning*; The McGraw-Hill Companies, Inc.: Columbus, OH, 1997.

- [5] Organisation for Economic Cooperation and Development (OECD), "Report from the Expert group on (Quantitative) Structure-Activity Relationships [(Q)SARs] on the Principles for the Validation of (Q)SARs", Technical Report OECD Environment Health and Safety Publications, Series on Testing and Assessment No. 49, OECD, 2004.
- [6] Sheridan, R.; Feuston, B.; Maiorov, V.; Kearsley, S. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1912-1928.
- [7] Jaworska, J.; Nikolova-Jeliazkova, N.; Aldenberg, T. *Altern. Lab. Anim.* **2005**, *33*, 445-459.
- [8] Helma, C. *Molecular Diversity* **2006**, *in press*.
- [9] Cronin, M. Toxicological Information for Use in Predictive Modeling: Quality, Sources, and Databases. In *Predictive Toxicology*; Helma, C., Ed.; Taylor & Francis: Boca Raton, 2005.
- [10] Richard, A. *Preclinica* **2004**, *2*, 103-108.
- [11] Helma, C.; Kramer, S.; Pfahringer, B.; Gottmann, E. *Environ. Health Perspect.* **2000**, *108*, 1029-1033.
- [12] Kazius, J.; Nijssen, S.; Kok, J.; Bäck, T.; IJzerman, A. P. *J. Chem. Inf. Model.* **2006**, *in press*.
- [13] Frasconi, P. Neural Networks and Kernel Machines for Vector and Structured Data. In *Predictive Toxicology*; Helma, C., Ed.; Taylor & Francis: Boca Raton, 2005.

- [14] King, R.; Muggleton, S.; Srinivasan, A.; Sternberg, M. *Proc. Natl. Acad. Sci. USA* **1996**, *93*, 438–42.
- [15] Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; VCH: Weinheim, 2000.
- [16] Kier, L. Indexes of molecular shape from chemical graphs. In *Computational Chemical Graph Theory*; Rouvray, D. H., Ed.; Nova Science: New York, 1990.
- [17] Balaban, A. T. *Chem. Phys. Lett.* **1982**, *89*, 399-404.
- [18] Klopman, G. *J. Am. Chem. Soc.* **1984**, *106*, 7315-7321.
- [19] Malacarne, D.; Pesenti, R.; Paolucci, M.; Parodi, S. *Environ. Health Perspect.* **1993**, *101*, 332-42.
- [20] Nijssen, S.; Kok, J. N. A Quickstart in Frequent Structure Mining can make a Difference. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*; ACM: New York, 2004.
- [21] Nys, G.; Rekker, R. *Chim. Therap.* **1973**, *8*, 521-537.
- [22] Kuhne, R.; Ebert, R.; Kleint, F.; Schmidt, G.; G., S. *Chemosphere* **1995**, *30*, 2061-2077.
- [23] Gasteiger, J.; Rudolph, C.; Sadowski, J. *Tetrahedron Comp. Method.* **1990**, *3*, 537-547.
- [24] Podlogar, B.; Ferguson, D. *Drug Des. Discov.* **2000**, *17*, 4-12.

- [25] Eriksson, L.; Johansson, E.; Lundstedt, T. Regression- and Projection-Based Approaches in Predictive Toxicology. In *Predictive Toxicology*; Helma, C., Ed.; Taylor & Francis: Boca Raton, 2005.
- [26] Kramer, S.; Helma, C. Machine Learning and Data Mining. In *Predictive Toxicology*; Helma, C., Ed.; Taylor & Francis: Boca Raton, 2005.
- [27] Rogers, D.; Hopfinger, A. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854-866.
- [28] Kubinyi, H. *Quant. Struct.–Act. Relat.* **1994**, *13*, 285-294.
- [29] So, S. S.; Karplus, M. *J. Med. Chem.* **1996**, *39*, 1521-1530.
- [30] Guyon, I.; Elisseeff, A. *J. Machine Learning Res.* **2004**, *3*, 1157-1182.
- [31] Helma, C.; Kramer, S. *Bioinformatics* **2003**, *19*, 1179-1182.
- [32] Guha, R.; Jurs, P. *J. Chem. Inf. Model.* **2005**, *45*, 800-806.
- [33] Zhou, Z.; Jiang, Y.; Chen, S. *AI Communications, 2003, 16(1): 3-15* **2003**, *16*, 3-15.
- [34] Vapnik, V. *The Nature of Statistical Learning Theory*; Springer: Berlin Heidelberg New York, 1995.
- [35] Tropsha, A.; Gramatica, P.; Gombar, V. K. *QSAR Comb. Sci.* **2003**, *22*, 69-77.
- [36] Willett, P.; Barnard, J.; Downs, G. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983-996.
- [37] Raymond, J.; Willett, P. *J. Comput.–Aided Mol. Des.* **2002**, *16*, 59-71.
- [38] Kohonen, T. *Self-Organizing Maps*; Springer-Verlag: Berlin, 1997.

- [39] Hawkins, D. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1-12.
- [40] Benigni, R. *Chem. Rev.* **2005**, *105*, 1767-1800.
- [41] Toivonen, H.; Srinivasan, A.; King, R. D.; Kramer, S.; Helma, C. *Bioinformatics* **2003**, *19*, 1183-1193.
- [42] Kazius, J.; McGuire, R.; Bursi, R. *J. Med. Chem.* **2005**, *48*, 312-320.