Data-driven Modeling and Prediction of Acute Toxicity of Pesticide Residues

Frank Lemke
KnowledgeMiner Software
Dürerstr. 40
16341 Panketal, Germany
frank@knowledgeminer.net

Emilio Benfenati
Istituto di Ricerche Farmacologiche
"Mario Negri"
Via Eritrea 62, 20157 Milan, Italy
benfenati@marionegri.it

Johann-Adolf Müller
KnowledgeMiner Software
Dürerstr. 40
16341 Panketal, Germany
jamueller@knowledgeminer.net

ABSTRACT

This paper outlines and implements a concept for developing alternative tools for toxicity modeling and prediction of chemical compounds to be used for evaluation and authorization purposes of public regulatory bodies to help minimizing animal tests, costs, and time associated with registration and risk assessment processes. Starting from a general problem description we address and introduce concepts of multileveled self-organization for high-dimensional modeling, model validation, model combining, and decision support within the frame of a knowledge discovery from noisy data.

Keywords

knowledge discovery workflow, self-organizing modeling, model validation, DEMETRA, pesticide toxicity, predictive QSAR models, European chemicals policy.

1. THE PROBLEM OF ECOTOXICITY

Besides the economical importance of the chemical industry as Europe's third largest manufacturing industry, it is also true that certain chemicals have caused serious damage to human health resulting in suffering and premature death and to the environment. The incidence of some diseases, e.g. testicular cancer in young men and allergies, has increased significantly over the last decades. While the underlying reasons for this have not yet been identified, there is justified concern that certain chemicals play a causative role for allergies.

The global production of chemicals has increased from 1 million tons in 1930 to 400 million tons today. There are about 100.000 different substances registered in the European market of which 10.000 are marketed in volumes of more than 10 tons, and a further 20.000 are marketed at 1-10 tons per year. The present system for general industrial chemicals distinguishes between "existing substances" i.e. all chemicals declared to be on the market in September 1981, and "new substances" i.e. those placed on the market since that date. There are some 3000 new substances. Testing and assessing their risks to human health and the environment according to the European Commission Directive 67/548 are required before marketing in volumes above 10 kg per year. For higher volumes more in-depth testing, focusing on long-term and chronic effects, has to be provided [1]. In contrast, existing substances amount to more than 99% of the total volume of all substances on the market, but they are not subject to the same testing requirements. Some of them have never been tested at all. The number of existing substances reported in 1981 was 100.106, the current number of existing substances marketed in volumes above 1 ton is estimated at 30.000. In result, there is a general lack of knowledge about the properties and the uses of existing substances. The risk assessment process is slow and resource-intensive and does not allow the system to work efficiently and effectively [1].

To address these problems and to achieve the overriding goal of sustainable development one political objective formulated by the European Commission in its "White Paper on the Strategy for a future Chemicals Policy" [1] is the implementation of the so-called REACH system (*R*egistration, *E*valuation and *A*uthorization of *Ch*emicals). Some more important objectives of the REACH framework are the protection of human health and the environment, an increased overall registration transparency, integration with international efforts, and the promotion of non-animal testing methods.

A consequence of this new chemicals policy, which passed European and national parliaments in 2005, is that every existing single substance on the market for the last 15 years will have to subsequently pass an official risk assessment and registration procedure as defined by the REACH framework, starting from high volume substances. But also for substances in articles (e.g., manufactured goods such as cars, textiles, electronic chips) a special regime applies.

Based on World Bank estimates and a number of prudent assumptions, diseases caused by chemicals are assumed to account for some 1% of the overall burden of all types of disease in the European Union (EU). Assuming a 10% reduction in these diseases as a result of REACH would result in a 0.1% reduction in the overall burden of disease in the EU. This would be equivalent to around 4.500 deaths being avoided every year [2]. Due to lack of data it is not possible to get a quantitative idea of the impacts on the environment. All in all, however, it is expected that REACH will contribute to reduced pollution of air, water, and soil as well as to reduced pressure on biodiversity and to reduced effects from endocrine disrupting chemicals [2].

According to a study of the University of Leicester, UK, one cost for implementing REACH would be an additional need of about 12 million animals for testing purposes. Because of this costs and the very long time it would take to run animal tests for all chemicals to be assessed (> 30.000), alternative, standardized, validated and accepted, by both industry and regulatory bodies, non-animal test methods are required. Current estimates expect that such alternative methods would save the lives of at least 2 million animals [3].

A current and promising way in that direction is building mathematical models – QSARs, Quantitative Structure-Activity Relationship models - based on already existing animal test data that describe and predict the impact of a given dose or concentration of a chemical compound (pollutant) on the health of a population of a certain biological species by the chemical's molecular properties. Typical parameters that are used in QSAR for expressing the chemical's impact on the population's health are the lethal dose LD_{50} or the lethal concentration LC_{50}. LC_{50}, for example, specifies the experienced concentration of a chemical compound where 50% of the population died within a given time, for example within a period of 96 hours ($LC_{50/96h}$), after introduction of the chemical into the system.

The issue of modeling and prediction of ecotoxicity of a very specific type of chemicals - pesticides – was considered in the international project DEMETRA [4] funded by the European Commission. To satisfy the multi-disciplinary nature of modeling ecotoxicity this project is composed of chemists, toxicologists, information scientists, and engineers from science and industry:

- Istituto di Ricerche Farmacologiche "Mario Negri", Italy,
- Central Science Laboratory, UK,
- Biochemics Consulting, France,
- University of Galati, Romania,
- Politecnico di Milano, Italy,
- University of Patras, Greece,
- Syngenta Crop Protection AG, Swizerland,
- BASF AG, Germany,
- KnowledgeMiner Software, Germany.

The major objective of this project was to develop a public piece of software for toxicity prediction of pesticides and related compounds (such as metabolites), directly and immediately useful for evaluation of pesticides and related compounds within the dossier preparation for pesticide registration. This software aims specifically at users such as national and EU regulatory bodies, industries, non-governmental organizations, and researchers who are involved in official registration and authorization procedures. It will allow processing of chemicals, one by one, for prediction of toxicity for pesticides and related compounds. It will also support regulatory evaluators to assess data submitted in approvals applications.

Compared to the general target of the REACH system of assessing and predicting toxicity of industrial chemicals, as outlined above, the target of DEMETRA was more focused on pesticides, and thus is somehow more difficult, because pesticides are typically very active compounds, complex on a chemical point of view (many functional groups are present, often several of them within the same compound) and on a toxicological point of view (for the occurrence of many toxic mode of action caused by the compounds). Furthermore, pesticides are limited, and the number of data available is small.

2. THE PROBLEM OF MODELING ECOTOXICITY

Besides the ethical, cost, and time considerations of running traditional bioassays to evaluate the ecotoxic effects of a chemical, there are also methodological problems of building predictive QSAR models. Ecotoxicological systems are complex, ill-defined systems, which are characterized by [5]:

- Inadequate a priori information about the system. Creating models for predicting toxic or other negative effects on the environment and human health is a highly interdisciplinary challenge. Scientists from chemistry, toxicology, biology, systems theory, information technology and machine learning, but also, not to forget, users from industry and public, regulatory bodies have to work together for finding a real working solution. There is no domain knowledge available, from any single domain, that would suffice to solve the problem by theory-driven approaches.
- Large number of potential, often immeasurable or simply unknown variables. A few hundred to a few thousand input variables are not uncommon in toxicity QSAR modeling.
- Noisy and few data samples. Reliable experimental toxicity data derived from past bioassays are rather rarely available and to obtain. Some tens to a few hundred data samples are common in toxicity QSAR modeling, though.
- Fuzzy objects. Experimental toxicity data are result of animal tests. Depending on the species used in an assay its inherent bio-variability can be quite high and can vary very much from species to species.

The economical, ethical, and methodological problems resulting from applying traditional bioassay and theory based methods but also dedicated expert systems [6] suggest and demand using a data–driven approach for finding an alternative tool for the evaluation and authorization of the huge amount of chemicals on the market.

Concluding from a systems theoretical analysis of the toxicity QSAR modeling problem [7, 8] the final, simplified nonlinear static model used in QSAR modeling to describe acute toxicity is shown in figure 1:

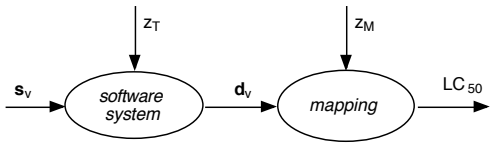

Figure 1. Simplified model for describing acute toxicity

with

$LC_{50} = f_2(f_1(s_v, z_T), z_M) = f(s_v, z_T, z_M)$, and

LC_{50} – experienced lethal concentration for a certain species and chemical compound (taken from past animal tests),

$\mathbf{s}_v$ – the (graphical) structure of the tested chemical compound in the chemical domain,

z_T – noise of the chemical structure to molecular descriptor transformation process,

z_M – noise transformed from the ecotoxicological test system,
$\mathbf{d}_v$ – vector of numerical molecular descriptors of the test compound to be used as input information for QSAR modeling.

The external disturbance z_T which adds noise to descriptor input space used for modeling can be reduced by fixing bugs and manual failures and by finding a most consistent chemical structure-to-descriptor transformation – although it is not clear a priori which transformation or optimization will add and which will reduce noise. The disturbance z_M, which finally results from the experimental animal tests, in contrast, adds noise to the output LC_{50} and is a given fact that cannot be changed afterwards. The overall noise dispersion in the data used for building toxicity QSAR models is expected of being up to 400%.

Apparently, inductive modeling of ecotoxicological systems implies dealing with very noisy data. Sets of data, generally, are not perfect reflections of the world. The measuring process necessarily captures uncertainty, distortion and noise. Noise is not errors that can infect data but is part of the world. Therefore, a modeling tool, but also results and decisions, must deal with the noise in the data. For a small level of noise dispersion, all regression-based methods using some internal criterion can be applied: Self-organizing Statistical Learning Networks (also known as Group Method of Data Handling; GMDH [5, 9, 10]) with internal selection criteria, statistical methods, or Neural Networks. For considerably noisy data – which always includes small data samples – GMDH or other algorithms based on external criteria are preferable. For a high level of noise dispersion, i.e., processes that show a highly random or chaotic behavior, finally, nonparametric algorithms of clustering, Analog Complexing pattern recognition, graphic-based methods, or fuzzy modeling should be applied [5, 11] to satisfy Stafford Beer's adequateness law [12]. This implies, of course, that with increasing noise in the data the model results and their descriptive language become fuzzier and more qualitative too.

In practice, inductive modeling means handling mountains of data, i.e. tables with high dimension. Besides the known theoretical dimensionality problem there is also a dimension limit of all known tools regarding computing time and physical memory. Therefore, a step of high priority is the objective choice of essential variables - state space reduction. In many fields, such as toxicology, there are only a small number of observations but many observed or calculated variables, which is the reason for uncertain results.

Furthermore, there is only very limited domain knowledge that could be used for modeling purposes so it calls for tools that perform a highly automated knowledge extraction from data.

3. KNOWLEDGE EXTRACTION FROM DATA

Deriving knowledge from data is an interactive and iterative workflow process of various subtasks and decisions and is called Knowledge Discovery from Databases (KDD) [13]. Usually, the single data mining process, only, has been automated in form of algorithms (Neural Networks, Decision Trees, fuzzy modeling, Genetic Algorithms, classical statistical methods, for instance) and software. The remaining parts require user interaction, manual work, and they are overall most time-consuming. This means, the result of knowledge discovery is very much dependent from knowledge, skills, ideas of the person who is running the analysis, and it is barely transparent and reproducible by other persons. Seen from a user perspective, however, in many cases these are key features for generating acceptance, trust and reliability in the results. So it is in the case of toxicity QSAR modeling.

We have been developing an integrating algorithm based on multileveled self-organization. This technology has been used intensively in the DEMETRA project for the first time. Our approach to a multileveled self-organization was motivated by the initial idea of KDD by making the overall workflow process more automated and more objective and to limit the user involvement to the inclusion of well-known a priori knowledge and to some pre- and post-processing tasks that are hardly to automate. Figure 2 shows the KDD workflow process when implementing an automated multileveled self-organization.

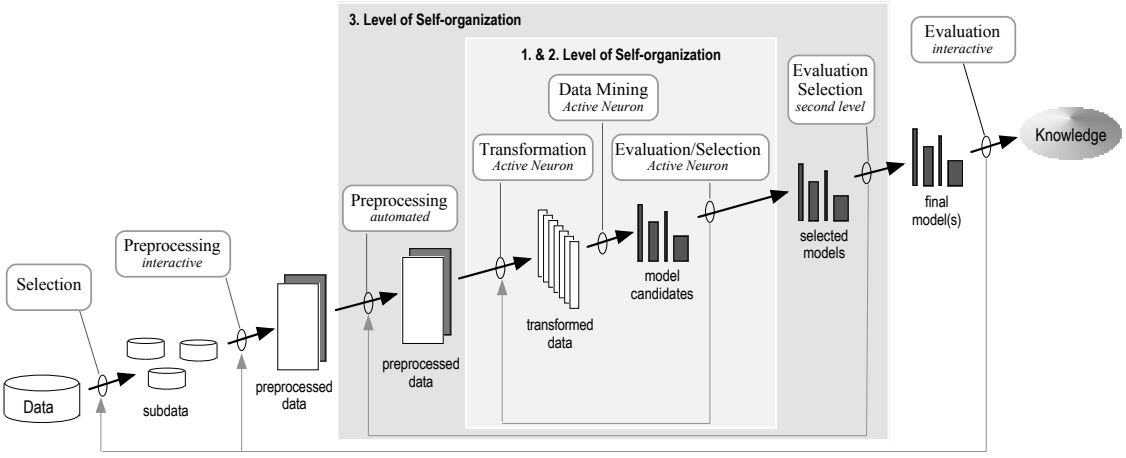

Figure 2. Multileveled self-organization displayed in gray boxes as a tool for KDD workflow processing

The concept of a multileveled self-organization starts with data preprocessing tasks that can be automated and may include:

- Missing values detection and handling,
- Further pre-selection of input variables according to some a priori given or intended constraints like a variable's diversity, type (continuous or discrete), or origin,
- Generation of additional, derived potential input variables,
- Deterministic or stochastic subdivision of data sets,
- Dimension reduction in state and/or sample space. In our work we introduced and used a wrapper approach, where the selection of relevant variables is evaluated by the implemented data mining algorithm directly, i.e., by the quality of results or the appropriateness of a variable to contribute to solving the given modeling task. In this case, variable selection is based on the so far reached model quality in the data mining process, i.e., we have an iterative procedure. The basic idea here is dividing high-dimensional modeling problems into smaller, more manageable problems by creating a new self-organizing network level composed of active neurons, where an active neuron is represented by an inductive learning algorithm in turn (lower levels of self-organization) applied to disjunctive data sets. The objective of this approach is based on the principle of regularization of ill-posed tasks, especially on the requirement of defining the actual task of modeling a priori to allow the algorithm selecting a set of correspondingly best models. In the context of a knowledge discovery from databases, however, this idea consequently requires using this principle in every stage of the knowledge extraction process – data pre-selection, pre-processing including dimension reduction, modeling (data mining), and model evaluation – consistently.

The proposed approach of multileveled self-organization integrates pre-processing, modeling, and model evaluation into a single, automatically running process and it therefore allows for directly building reliable models from high-dimensional data sets (up to 30.000 variables, currently), objectively. The external information necessary to run the new level of self-organization is provided by the implemented algorithm's noise sensitivity characteristic as explained in [14, 15] (fig. 3).

The first two levels of self-organization have been the basic idea of Self-organizing Statistical Learning Networks for more than 20 years [5, 9, 17]. They are built on three main concepts:

- The black-box method as a basic approach to analyze systems from input-output data,
- The concept of connectionism as a description of complex functions by networks of elementary functions, and
- The principle of model induction [5].

These two levels of self-organization incorporate these essential tasks:

- Self-organization of neuron transfer functions,
- Self-organization of the network's structure (topology) by generating alternative model candidates of different input variables and of growing complexity, and
- The first level of model evaluation and model selection.

The last step in multileveled self-organization is further evaluation and selection of models that passed the lower self-organization levels by calculating the models' Descriptive Power as described in more detail in [7, 14, 15, 16]. A key problem in data mining and knowledge discovery from data is final evaluation of generated models. This evaluation process is an important condition for application of models obtained by data mining. From data mining, only, it is impossible to decide whether the estimated model can reflect the causal relationship between input and output, adequately, or if it's just a stochastic model with non-causal correlations. Model evaluation needs - in addition to a properly working noise filtering for avoiding overfitting the learning data (first level of validation) - some new, external information to justify a model's quality, i.e., both its predictive and descriptive power. Again, the algorithm's noise sensitivity characteristic provides key information here.

The objective of a second level of model validation is:

- Noise filtering implemented in first level of validation is very likely to not being an ideal noise filter and thus not working properly in any case (see fig. 3) and
- To get a new model quality measure, Descriptive Power, that is adjusted by the noise filtering power of the algorithm.

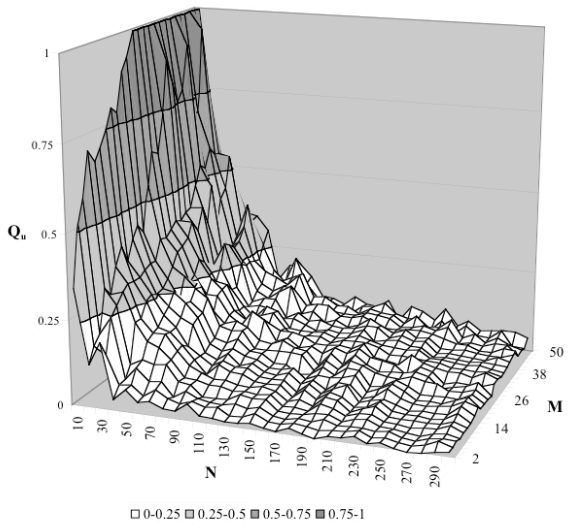

Figure 3. Noise sensitivity characteristic of a Self-organizing Statistical Learning algorithm
M: number of potential inputs
N: number of samples
Q_u: virtual quality of a model
Q_u=1: noise filtering does not work at all
Q_u=0: ideal filtering

The noise sensitivity characteristic (fig. 3) expresses a virtual model quality Q_u that can be obtained when using a data set of M potential inputs of N random samples. It is virtual model quality, because, by definition, there is not any causal relationship between stochastic variables (true model quality $Q = 0$, by definition [14]), but there are actually models of quality $Q > 0$, which, when using random samples, just reflect stochastic correlations. By implementing an algorithm's noise sensitivity characteristic into a data mining tool it is possible for any given number of potential inputs M and number of samples N to calculate a threshold quality $Q_u=f(N, M)$ that any model's quality Q must exceed to be stated valid in that it describes some relevant relationship between input and output. Otherwise, a model of quality $Q \leq Q_u$ is assumed invalid, since its quality Q can also be reached when simply using independent variables, which means that this model does not differ from a model of just stochastic correlations. The implemented two-stage model validation approach now allows, for the first time, to get on the fly an active decision support in model evaluation based on the model's descriptive power calculated on the learning data, only, for minimizing the risk of false interpreting models and using invalid models that just reflect some non-causal correlation [14, 15, 16, 18].

The overall process of knowledge extraction based on multileveled self-organization is highly computationally intensive – the self-organization of a nonlinear regression model of about 10 self-selected relevant input variables from 1000 potential inputs and 200 samples, for example, may take up to 2 days of computing time – however, since it doesn't require any user interaction it can run in the background while saving the user's attention and time for other work. Increased transparency and reproducibility are other features of this approach.

4. RESULTS

The results shown here were obtained within the DEMETRA project and they can be seen as milestones towards QSAR models that can be applied within REACH system implementation.

Based on five data sets - D_1 (Trout), D_2 (Daphnia), D_3 (Oral Quail), D_4 (Dietary Quail), D_5 (Bee), - we first created many individual regression and classification models (> 500) using different modeling and data mining algorithms like Partial Least Squares, different types of Neural Networks, fuzzy modeling, and multileveled self-organization as described above.

From this pool of individual models we then created a hybrid model for each data set by combining corresponding individual models.

Since the focus of public regulatory bodies is on regression models, we report results from these models here, only.

4.1 The Data

Biological data are affected by factors relative to the biological system itself and by factors dependent on the investigation technique used. While natural variability cannot be eliminated, and is part of the real world, many attempts have been done to reduce the influence of the technique used to study the biological system, through the introduction of standardized procedures. Commonly, the term variability is used in relation to the natural factors, while uncertainty is used in the case of factors related to the technique to study the biological phenomenon. In our case, we used only data on pesticide ecotoxicity originating from experiments, which have been conducted according to official guidelines. In particular, Dr. Brian Montague from the US Environmental Protection Agency, Washington, DC, provided the data for this work. In many cases several different values for the same compound was reported, resulting from different experiments conducted all according to the official guidelines. We defined some criteria for the selection of appropriate values, in order to use experiments with a higher quality and a lower variability [19]. Furthermore, we checked the values with other databases, in order to increase their reliability. We studied five different toxicological endpoints, and the number of compounds was less than 300 in the most favorable case (toxicity towards rainbow trout) to about 100 in the case of bee toxicity. The limited number of examples is, indeed, a common problem for this type of study, mainly – like in our case on pesticides - when a heterogeneous set of compounds is used, referring to many different kinds of bio-mechanisms responsible for the observed toxicity phenomenon.

To describe the chemical nature of the compounds we used several software tools, such as DRAGON, CODESSA, PALLAS, CACHE (see also fig. 1). As a result, thousands of molecular descriptors are available for each chemical compound.

4.2 Individual Models

Based on the five data sets a large set of individual QSAR models were created by different project partners using different data mining algorithms. To allow comparison and combination of these models three strict preconditions were defined:

- The official data sets produced within the DEMETRA project have to be used for modeling, only.
- Although some of the data sets have rather few compounds N, only (N~100), each data set D_i was randomly subdivided by a 6:1 split into a learning subset $D_{i,A}$ (or $D_{i,A}$ and $D_{i,B}$) and an out-of-sample test data subset $D_{i,C}$, with $N_{A,B} + N_C = N$. The data in the test subset was never to be used for modeling at all, but was hold out for validating all created individual models on this new data.
- For comparison purposes, for every model the Coefficient of Determination R^2 calculated on both learning and test data subsets had to be provided:

$$R^2 = 1 - \delta^2, \delta^2 = \frac{\sum_{i \in N} (y_i - \hat{y}_i)^2}{\sum_{i \in N} (y_i - \bar{y}_i)^2} \leq 1,$$

  where y_i, $\hat{y}_i$, and $\bar{y}$ are the true, estimated, and mean values of the output variable, respectively.

The results of the five best individual QSAR models for the trout data set are listed in table 1 exemplarily. Some QSAR

models were created using 2-dimensional (2D) molecular descriptors (inputs), only, others were built on 3-dimensional (3D) or 2D and 3D descriptors.

$R^2_{A,B,C}$	$Q^2_{A,B}$	R^2_C	*m*	*model type*	*DM-method*
0.67	0.69	0.59	10	explicit linear model	multileveled self-organization
0.66	0.66	0.64	15	explicit linear model	multileveled self-organization
0.65	0.66	0.63	6	implicit nonlinear model	Neural Network (GA-MLP)
0.63	0.63	0.65	8	implicit nonlinear model	Neural Network (GA-MLP)
0.63	0.71	0.64	11	explicit nonlinear model	multileveled self-organization

N = 275 $N_{A,B}$ = 229 N_C = 46 M: up to 1800

with

$R^2_{A,B,C}$ - R^2 calculated on the entire data set *D*

$Q^2_{A,B}$ - leave-one-out cross-validation on the data subset $D_{A,B}$

R^2_C - R^2 calculated on the test data subset D_C

m - number of variables used in the model

M - number of potential input variables; state space dimension

multileveled self-organization: High-dimensional modeling algorithm using multileveled self-organization with GMDH Networks as Active Neurons

Neural Network (GA-MLP): Genetic Algorithm for dimension reduction; Multilayer Percepton Neural Network for modeling

Table 1. Five best models for the data set D_1 – Trout - with respect to $R^2_{A,B,C}$

The *model type* column of table 1 distinguishes between implicit and explicit regression models. While Neural Networks typically distribute and hide the created model in the network the result of multileveled self-organization are explicit analytical models. Figure 4 shows, for example, the regression equation of the first model of table 1. Neither the formal model structure nor the input variables composition was given a priori; the model is completely self-organized. This true knowledge extraction from data has proven very useful and advantageous for model interpretation, evaluation, and implementation issues. So it is possible to implement these types of models in a MS Excel sheet, automatically, for immediate use for further analysis, evaluation, or just application purposes [18].

$$LC_{50}\ (trout)\ [mmol/l] = -1.6023\ (\text{C-031})^{-1} - 1.53\ \text{MATS3e} - 1.3148\ (\text{nOH})^{-1} - 27.1340\ \text{GATS3m} - 0.8957\ \text{nxch3} + 2.1469\ (\text{SEigZ})^{-1} - 0.2699\ \text{LogDpH7} + 0.7736\ (\text{D/Dr09})^{-1} - 0.0313\ \text{D/Dr03} + 5.8706\ (\text{Mp})^{-1} + 28.220$$

Figure 4. Self-organized linear regression model in chemical notation

Similar results of individual models were obtained for the other four data sets.

4.3 Combined Models

All methods of automatic model selection lead to a single "best" model while the accuracy of model result depends on the variance of the data. A common way for variance reduction is aggregation of similar model results following the idea: Generate many versions of the same predictor/classifier and combine them in a second step. If modeling aims at prediction, it is helpful to use alternative models that estimate alternative forecasts. These forecasts can be combined using several methods to yield a composite forecast of a smaller error variance than any of the models have individually. The desire to get a composite forecast is motivated by the pragmatic reason of improving decision-making rather than by the scientific one of seeking better explanatory models. Composite forecasts can provide more informative inputs for a decision analysis, and therefore, they make sense within decision theory, although they are often unacceptable as scientific models in their own right, because they frequently represent an agglomeration of often conflict theories.

Based on the five sets of individual models that now served as input information, we generated a combined model for each data set by a Self-organizing Statistical Learning Network algorithm. The result is five self-selected, optimally composed linear or nonlinear regression models, including their regression equation. It is shown from table 2 that the overall model performance for all 5 data sets increases sufficiently. Figures 5 plots the combined model again for the trout data set exemplarily.

It should be noted that the combined models are not just a composition, or the mean, of the five or seven best individual models of a data set but are an a priori unknown, optimal mix of models that – combined – decrease the error variance of the combined model most.

However, every individual or combined model is not able to also reflect the uncertainty given by the initial experimental toxicity data. Here the idea of a prediction interval is useful.

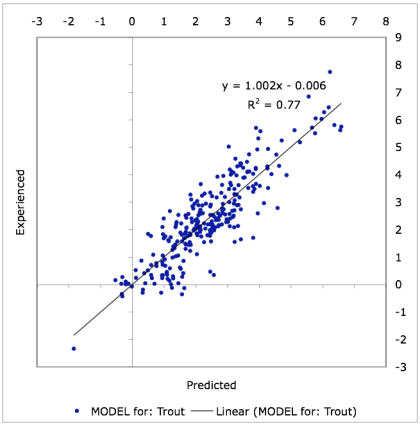

Figure 5. Scatter plot of the combined model for data set D_1 – Trout

data set	$R^2_{A,B,C}$	$Q^2_{A,B}$	R^2_C	*m*	*models*
D_1 - Trout	0.74	0.77	0.56	7	NN(1), F-NN(1), MSO(5)
D_2 - Daphnia	0.81	0.84	0.62	7	NN(2), F-NN(2), PLS(1), MSO(2)
D_3 - Oral Quail	0.84	0.9	0.53	4	NN(3), PLS(1)
D_4 - Dietary Quail	0.85	0.88	0.71	7	PLS(1), MSO(6)
D_5 - Bee	0.8	0.8	0.78	5	NN(2), MSO(3)

with
$R^2_{A,B,C}$ - R^2 calculated on the entire data set D_i
$Q^2_{A,B}$ - leave-one-out cross-validation on the data subset $D_{i,A,B}$
R^2_C - R^2 calculated on the test data subset $D_{i,C}$
m - number of models implemented in the combined model

models column: The modeling method a model was generated with followed by the number of models of this type used in the combined model

NN - Neural Network
F-NN - Fuzzy Neural Network
PLS - Partial Least Squares
MSO - Multileveled Self-organization

Table 2. Model performance summary of five combined models

4.4 Model Uncertainty and Prediction Interval

As pointed out in this paper, toxicity data are highly noisy and therefore require adequate modeling, results interpretation, and decision support methods. Additionally, all methods of automatic model selection lead to a single "best" model. On this base are made conclusions and decisions as if the model was the true model. However, this ignores the major component of uncertainty, namely uncertainty about the model itself. In toxicity modeling it is not possible that a single crisp prediction value can cover and reflect the uncertainty given by the initial object's data. If models can be obtained in a comparing short time it is useful to create and apply several alternative reliable models on different data subsets or using different modeling methods and then to span a prediction interval from the models' various predictions for describing the object's uncertainty more appropriately. In this way a most likely, a most pessimistic (or most safe prediction from a toxicity point of view), and a most optimistic (or least safe) prediction is obtained, naturally, based on the already given models only, i.e., no additional (statistical) model has to be introduced for confidence interval estimation, for example, which would had to make some new assumptions about the predicted data, and therefore, would include the confidence about that assumptions, which, however, is not known a priori.

A prediction interval has two implications:

- The decision maker is provided a set or range of predicted values that are possible and likely representations of a virtual experimental animal test including the uncertainty once observed in corresponding past real-world animal tests. The decision maker can base its decision on any value of this interval according to importance, reliability, safety, impact or effect or other properties of the actual decision to make. This keeps the principle of freedom of choice for the decision process.
- Depending on which value is actually used, a prediction interval also results in different overall model quality values like R^2, starting from the highest accuracy for most likely predictions.

Figure 6 displays the prediction intervals for selected test set compounds (D_C) obtained from the predictions of the individual models contained in the combined model for the data set D_1 as reported in table 2.

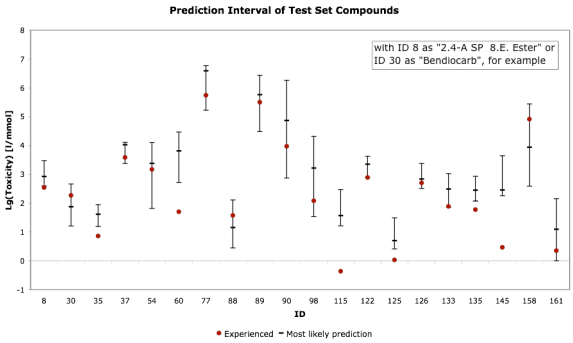

Figure 6. Prediction interval for test set compounds of data set $D_{1,C}$

5. DECISION SUPPORT MODEL IMPLEMENTATION

The data-driven concept for developing adequate toxicity prediction and decision models outlined in this paper to be used as alternative, substituting tools for animal tests during the projected extended evaluation of existing chemical compounds is implemented, exemplarily, for the trout data set in Microsoft Excel. This prototype is a fully working toxicity prediction tool that works on both any single compound of the given data set D_1 and any new compound when the required descriptor values for this compound are provided. The result is a most likely toxicity value in two common toxicity data spaces – mmol/l and mg/l – along with the prediction uncertainty expressed by the compound's predicted highest and lowest toxicity, displayed numerically and graphically.

Figures 7 and 8 show the interface of this tool.

Some features, which are relevant for the specific purposes of this tool, should be noted here. Our approach was driven by the overall goal of providing a tool for regulatory use of QSAR models. A major problem with currently published QSAR models, from a regulatory point of view, is that they are much closer to a research tool than to a practical tool.

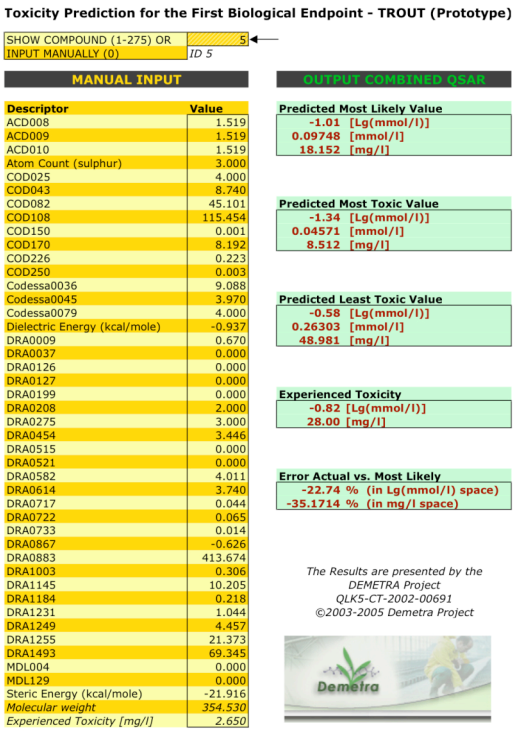

Figure 7. Interface of the implemented decision support model for predicting a chemical compound's toxicity on the biological endpoint trout – page 1.

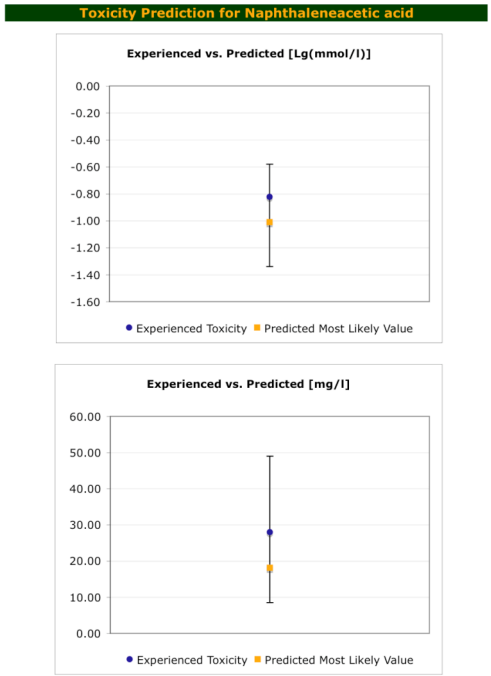

Figure 8. Interface of the implemented decision support model for predicting a chemical compound's toxicity on the biological endpoint trout – page 2.

In particular, they are sensitive to the human experience of the individual researcher. Typically, a researcher with a specific skill is using those more research-oriented models. And, it is expected that on the basis of her/his experience different results may be obtained. Such a situation, which is common in the research field, is neither a most favorable nor an acceptable case for regulatory uses during the authorization process of chemicals, if different results are expected depending on the person who is using the QSAR model.

The tool we present here, vice versa, is tailored for regulatory uses, because it calculates a unique output value from the model, along with its uncertainty. The user does not require any particular experience in the QSAR model itself. However, for a given chemical compound, she/he has to calculate the values for the chemical descriptors indicated in the tool using certain publicly available software, but no further experience in QSAR modeling is needed.

6. SUMMARY

In this paper we outlined a concept for developing alternative tools for toxicity prediction of chemical compounds to be used for evaluation and authorization purposes of public regulatory bodies to help minimizing animal tests, costs, and time associated with registration and risk assessment processes.

Toxicity QSAR modeling is described by these major preconditions and requirements:

- Animal tests as the source of toxicity data for QSAR modeling are described by a complex, nonlinear dynamic ecotoxicological system. However, the toxicity QSAR modeling problem, finally, transforms to building static, linear or nonlinear models. This, all together, is a strong simplification of the ecotoxicological system and adds high uncertainty to results.
- Toxicity data is very noisy due to a biological species' natural variability and due to the uncertainty of the animal test procedure. Also, there is not a single valid toxicity value but a certain range of experienced toxicities for a given chemical compound that can be seen all as true, reliable values.
- Toxicity QSAR modeling is an ill-defined and high-dimensional modeling problem that requires adequate modeling tools.
- Decision support has to take into account the uncertainty of the underlying system and the models.

Within the DEMETRA project, we generated five data sets for five biological endpoints that show very high quality. This quality feature refers to the reliability of the experimental toxicity data derived from past animal tests as well as to the calculation of molecular descriptors for the pesticides under study.

We addressed the problem of high-dimensional modeling of an ill-defined system by introducing multileveled self-organization, which incorporates state space dimension reduction, variables selection, data mining, and model evaluation into a single, autonomously running algorithm. We paid special attention to model validation and we

suggested and implemented a two-stage model validation idea, which is composed of applying cross-validation and an algorithm's identified noise sensitivity, subsequently.

We combined several individual QSAR models to model ensembles that all show significantly increased model accuracy and, in addition, we assigned to every single prediction of a given compound a prediction interval to describe uncertainty.

Finally, this concept is implemented exemplarily in Microsoft Excel for real-world application and demonstration purposes. All five final models are currently developed in Java for public web-based access [4].

A future work on the way to reliable toxicity prediction models is the definition of standards for toxicity data, toxicity QSAR modeling, and model validation for improving reproducibility, transparency and acceptability of data-driven toxicity prediction tools to be established as a real alternative and supplement to animal tests.

7. ACKNOWLEDGMENTS

The authors acknowledge financial support from the European Commission for the project DEMETRA, QLK5-CT-2002-00691.

8. REFERENCES

[1] European Commission: White Paper. Strategy for a future Chemicals Policy, 27.02.2001

[2] European Commission: REACH in brief, 15.09.2004

[3] van der Jagt, K., Munn, S., Tørsløv, J., de Bruijn, J. (editors): Alternative approaches can reduce the use of test animals under REACH. Institute for Health and Consumer Protection, European Commission, *Joint Research Centre, Report EUR 21405 EN*, Ispra, 2004

[4] DEMETRA, EC project, http://www.demetra-tox.net, 2005

[5] Müller, J.-A., Lemke, F.: *Self-Organising Data Mining. Extracting Knowledge From Data*, BoD, Hamburg, 2000

[6] Gini, G., Lorenzini, M., Benfenati, E.: Predictive Carcinogenicity: A model for Aromatic Compounds with Nitrogen-Containing Substituents Based on Molecular Descriptors Using Artificial Neural Network, *Journal of Chem. Inform. And Comp. Sci.*, (39)1999(6), pp. 1076-1080

[7] Lemke, F., Müller, J.-A.: Benfenati, E.: Modelling and Prediction of Toxicity of Environmental Pollutants. *LNAI 3303* (Eds. J. A. Lopez et al.), Springer, Berlin, Heidelberg 2004, pp. 221-234

[8] Lemke, F., Benfenati, E., Müller, J.-A.: Data-driven Modeling of Acute Toxicity of Pesticide Residues as Alternative Tool within Official Registration, Evaluation and Authorization Procedures, *Data Mining Case Study Workshop, 5th IEEE ICDM*, Houston, Texas, 2005

[9] Farlow, S. J. (ed.): *Self-Organizing Methods in Modeling: GMDH-Type Algorithm*, Marcel Dekker, New York, 1984

[10] Barron, A. R., Barron, R. L.: Statistical learning networks: a unifying view, *Proceedings of the 20th Symposium Computer Science and Statistics*, 1988, pp. 192-203

[11] Ivakhnenko, A.G., Müller, J.-A.: Parametric and nonparametric procedures in experimental systems analysis, *SAMS*, 9(1992), pp. 157-175

[12] Beer, S. T.: *Cybernetics and Management*, English University Press, London, 1959

[13] Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P.: From Data Mining to Knowledge Discovery: An Overview. In: Fayyad, U.M. et al.: *Advances in Knowledge Discovery and Data Mining*. AAAI Press/The MIT Press, Menlo Park, California, 1996, pp. 107-128

[14] Lemke, F., Müller, J.-A.: Validation in self-organising data mining, *Proceedings 1st International Conference on Inductive Modelling*, 2002

[15] http://www.knowledgeminer.com/pdf/validation.pdf

[16] Lemke, F.: Does my model reflect a causal relationship? http://www.knowledgeminer.com/isvalid.htm, 2002

[17] Ivakhnenko, A.G., Müller, J.-A.: *Selbstorganisation von Vorhersagemodellen*, Verlag Technik, Berlin, 1984, (in German)

[18] KnowledgeMiner: Self-organizing data mining and prediction tool, http://www.knowledgeminer.com

[19] Roncaglioni, A., Benfenati, E., Boriani, E., Clook, M.: A Protocol to Select High Quality Datasets of Ecotoxicity Values for Pesticides. *Journal of Environmental Science and Health*, Part B, B39, 641-652, 2004.

About the authors:

Frank Lemke is an independent software developer and consultant specialized in self-organizing modeling for 15 years. He authored several papers on theory and application of inductive modeling and a book on self-organizing data mining. He contributed to a number of international research projects in information and life sciences. He has been developing the "KnowledgeMiner" data mining software package.

Emilio Benfenati, head of the Laboratory of Environmental Chemistry and Toxicology at the "Mario Negri" Institute, Milan, is working in predictive toxicology, environmental toxicology and environmental analysis. He coordinated more than 10 international projects. He is author or co-author of about 200 articles in international journals and books.

Johann-Adolf Müller has been working in the area of inductive modeling and simulation of complex systems for more than 30 years. He earned a Ph.D. in Electrical Engineering and a Ph.D. from the Institute of Economics, Berlin, in 1978. He is author of many papers and books on model self-organization, and he is visiting professor at the Chengdu University of Science and Technology, China