DISSERTATIONES CHIMICAE UNIVERSITATIS TARTUENSIS

69

DISSERTATIONES CHIMICAE UNIVERSITATIS TARTUENSIS
69

QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIPS OF ENVIRONMENTALLY RELEVANT PROPERTIES

IIRIS KAHN

TARTU UNIVERSITY
PRESS

Department of Chemistry, University of Tartu, Estonia
Institute of Chemical Physics, Molecular Engineering

Dissertation is accepted for the commencement of the degree of Doctor of Philosophy in Molecular Engineering on June 29, 2007, by the Doctoral Committee of the Department of Chemistry, University of Tartu.

Supervisors: Prof. Mati Karelson, University of Tartu
Dr. Uko Maran, University of Tartu

Opponents: Dr. Aynur O. Aptula, Safety & Environmental Assurance Centre, Unilever Inc., UK
Prof. Peeter Burk, University of Tartu

Commencement: August 29, 2007 at 18 Ülikooli Str., Scientific Council room

ISSN 1406–0299
ISBN 978–9949–11–672–0 (trükis)
ISBN 978–9949–11–673–7 (PDF)

Tartu Ülikooli Kirjastus
www.tyk.ee
Tellimuse nr 296

CONTENTS

LIST OF ORIGINAL PUBLICATIONS

This hesis consists of four articles that are denoted in the text by Roman numerals I–IV, respectively.

I. Iiris Kahn, Dan Fara, Mati Karelson, Uko Maran and Patrik L. Andersson, **QSPR Treatment of the Soil Sorption Coefficients of Organic Pollutants.** *J. Chem. Inf. Model.* **2005**, *45*, 94–105.

II. Uko Maran, Sulev Sild, Iiris Kahn and Kalev Takkis, **Mining of the Chemical Information in GRID Environment.** *Future Generat. Comput. Syst.* **2007**, *23*, 76–83.

III. Iiris Kahn, Uko Maran, Emilio Benfenati, Tatiana I. Netzeva, T. Wayne Schultz and Mark T. D. Cronin, **Comparative Quantitative Structure-Activity-Activity Relationships for Toxicity to *Tetrahymena pyriformis* and *Pimephales promelas*.** *ATLA-Altern. Laborat. Anim.* **2007**, *35*, 15–24.

IV. Iiris Kahn, Sulev Sild and Uko Maran, **Modeling the Toxicity of Chemicals to *Tetrahymena pyriformis* Using Heuristic Multi-Linear Regression and Heuristic Back-Propagation Neural Networks.** *J. Chem. Inf. Model.* ***submitted.***

Author's Contribution

Publication I: the main contributor performing the calculations and writing the manuscript

Publication II: contributed to the QSAR model formation, namely data slitting with PCA

Publication III: the main contributor performing the calculations and writing the manuscript

Publication IV: one of the main contributors to preparing data, performing calculations, analyzng results and writing the manuscript

LIST OF ABBREVIATIONS

3D	three-dimensional
ADME/Tox	absorption, distribution, metabolism, excretion and toxicity
AM1	Austin Model 1
ANN	Artificial Neural Network
B3LYP	Becke 3-parameter hybrid functional using Lee-Yang-Parr correlation functional
BMLR	Best MultiLinear Regression
BNN	Back-propagation Neural Network
BOD	biological oxygen demand
CODESSA	Comprehensive Descriptors for Structural and Statistical Analyses
HIV-1	human immunodeficiency virus of type 1
IGC_{50}	median growth inhibition concentration
K_{OC}	soil sorption coefficient, normalized to organic carbon
LC_{50}	median lethal concentration
MCMM	Monte Carlo Multiple Minimum
MLR	Multilinear Regression
MMFF	Merck Molecular Force Field
MOPAC	Molecular Orbital PACkage
OpenMolGRID	Open computing GRID for Molecular science and engineering
PBT	persistent, bioaccumulative and toxic
PCA	Principal Component Analysis
PC	Principal Component
PLS	Partial Least Squares
QAAR	Quantitative Activity-Acitvity Relationship
QSAAR	Quantitative Structure-Activity-Acitvity Relationship
QSAR	Quantitative Structure-Acitvity Relationship
QSPR	Quantitative Structure-Property Relationship
REACH	Registration, Evaluation and Authorisation of Chemicals
SCF	Self Consistent Field
SIMCA	Soft Independent Modeling of Class Analogy
SOM	Self Organizing Map
US EPA	United States Environmental Protection Agency

INTRODUCTION

Quantitative Structure-Activity Relationships (QSARs) are empirical models that relate experimental properties/activities of compounds with their molecular structures. The rapid development of quantum theory and *ab initio* computational methods have made it possible to predict molecular properties of small isolated molecules within experimental error. However, the majority of industrially and environmentally important chemical processes, and biochemical transformations in living organisms take place in heterogeneous condensed media and hence the use of QSARs that proceed directly from the endpoint of interest is an attractive and fast alternative to predict the molecular properties in complex environments.

Documented evidence of a structure-activity relationship dates back to the year of 1863, to a PhD thesis by A. F. A. Cros, University of Strasbourg[1] who observed an increase in the toxicity of alcohols to mammals, with decreasing water solubility. The potential of QSAR as of the present was revealed about a hundred years later. After the use of various empirical quantities in correlations with the studied properties of compounds, theoretical molecular descriptors, derived using only the information encoded in the chemical structure, began to emerge. QSAR became more attractive due to the development of new software tools[2] and expanded rapidly to various areas of industrial and environmental chemistry. QSARs became extremely productive in pharmaceutical chemistry and in computer-assisted drug design[3] including the development of ADME/Tox profiles.

In recent years, QSARs have become essential tools for the estimation of the physicochemical, biological and environmental properties of chemicals for regulatory purposes. In 2003, the European Commission created a new chemical management system called REACH (Registration, Evaluation and Authorisation of Chemicals),[4] which requires environmental and toxicology data for new and existing industrial chemicals. The use of valid QSARs was proposed as a source for primary data in the interests of time- and cost-effectiveness as well as animal welfare.[5] Four years later, in June 1st 2007, the REACH legislation entered into force by opening the new European Chemicals Agency in Helsinki.[6] Over a period of 11 years, about 30 000 chemical substances in use today (in volumes starting at 1 tonne) are to be registered with the required data for the risk evaluation.

The main aim of the work presented in this thesis was the assessment of the environmental impact of organic pollutants by means of QSAR, namely, the soil sorption of organic pollutants and acute aquatic toxicity to the freshwater organisms, and the development of QSAR models of high predictive ability on the studied endpoints. In the course of the work, aspects of the QSAR methodology were assessed and improved. In Chapter 1, an overview is given about the methodology and environmentally relevant properties, and in Chapter 2, our original results of QSAR/QSPR modeling on selected activities/ properties are summarized.

1. LITERATURE OVERVIEW

1.1. Quantitative Structure-Activity Relationships

Quantitative Structure-Property Relationships (QSPRs) are mathematical models where structural descriptors are related to the chemical activity under study:

$$Property = f\,(\text{structure})$$

In biological contexts, these are also called quantitative structure-activity relationships (QSARs). Once the QSAR model is established it can be used to predict properties of untested compounds based on their molecular structure. A number of steps and conditions apply for successful development of valid QSARs:

1. Evaluation of the experimental data
2. Optimization of the molecular geometries and generation of the molecular descriptors
3. Formation of the representative training and validation sets
4. QSAR model development
5. Selection of the best model according to the statistical criteria and relevance of the model descriptors to the studied property; Consensus modeling
6. Validation and interpretation of the model

1.1.1. Selection of the data set

Evaluation of the quality of the experimental data and the design of the training set are crucial steps in the model development. Reliable data are required to build reliable predictive models. Such data should ideally be measured from well-standardized assays and even in the same laboratory.[7] Excellent examples of such procedures include the 96-h LC_{50} *Pimephales promelas* (fathead minnow) database of the United States Environmental Protection Agency Duluth Laboratory[8]; and the 40-h growth inhibition of the ciliated protozoan *Tetrahymena pyriformis* database[9]. High confidence and lower experimental error may be assigned to these data in the development of predictive QSARs.

1.1.2. Molecular descriptors

Molecular descriptors are numerical representations of the molecular structure used as the independent variables in QSAR modeling. In order to encode structural information relevant to the specific modeling tasks, thousands of descriptors have been formulated.[10] Empirical and theoretical descriptors can be distinguished.[11] The empirical descriptors are usually experimentally measured physico-chemical properties, such as water solubility, or hydrophobicity. For the modeling of environmental and biological properties, the hydrophobicity, measured as n-octanol/water partition coefficient, is widely used because of the importance of partitioning phenomena between water and other phases in biological systems. Although a few reliable methods[12,13,14] based on atom/ fragment contribution have been developed for the estimation of hydrophobicity, the experimental values are generally preferred if available. A shortcoming of empirical descriptors is that they can be difficult and costly to obtain, especially for unknown or hypothetical compounds.

Theoretical descriptors are attractive due to the available computational methods and tools that require only the chemical structure of the compound as input. The simplest descriptors involve counts of certain structural features, such as atoms, bonds or functional groups. Simple algorithms can also be used to generate molecular connectivity indices[15], that are often used to characterize atomic constitution, size and steric effects. The number of descriptors that can be calculated expands greatly when the three-dimensional (3D) structure and charge distribution of molecules are considered. The molecular geometries can be optimized using molecular mechanics or quantum mechanical methods. The charge distributions can be obtained either empirically or from quantum mechanical calculations; the latter approach is usually preferred. The semiempirical AM1 Hamiltonian[16] has become a method of choice over more rigorous *ab initio* methods in calculation of molecular descriptors. In a comparative study of the effectiveness of the AM1 Hamiltonian and density functional (B3LYP/6-31G**) method, it was concluded that for large-scale predictions the use of precise but time-consuming *ab initio* methods did not offer considerable advantage compared to the semiempirical calculations.[17] Among the so-called 3D descriptors mention could be made of various shape and surface area descriptors[18], partial atomic charges and reactivity indices, including frontier orbital energies. New descriptors are continually formulated for different modeling tasks. For example, the intermolecular terms of the modified scoring function obtained from the docking results were used as descriptors in prediction of the binding affinity of ligands to protein.[19]

1.1.3. Formation of training set

QSARs are based on structural similarity and therefore, prediction of the properties is possible only if the training set adequately represents the domain of the studied chemicals. In publication II of the present thesis, the use of PCA as a multivariate clustering method was investigated for the selection of a representative training set. PCA is an unsupervised learning method that takes molecular descriptors as input and transforms them into uncorrelated latent variables or principal components.[20] Another useful method for clustering the chemicals by structural similarities is Kohonen artificial neural networks or self organizing maps (SOM).[21] PCA and SOM, both being unsupervised learning methods cluster the compounds according to their structures represented by the molecular descriptors and then from each cluster a specified number of representatives can be selected in the training set. D-optimal design, a method of statistical experimental design, has also proven to provide a well-balanced structural representativity of the data space considering both the descriptors and the response.[22,23] The simplest approaches are random selection and activity sampling both of which do not proceed from the molecular structure. Use of these methods is recommended only for homogeneous or large data sets of hundreds of compounds.

Presently, intense discussions are being held on the subject of the applicability domain of the QSAR models for regulatory purposes.[24] The applicability domain of a QSAR model is considered to be the response and chemical structure space in which the model makes predictions with a given reliability.[25] According to the international recommendations, the application domain of the model must be clearly defined and carefully considered when the model is used for predictions.[26] It is due to the reductionist nature of QSARs that does not allow extrapolations. In the formation of the training set the applicability limitations of the model are consequently determined.

1.1.4. Modeling methods

Linear methods. Multilinear regression[27] is the simplest of the linear methods for relating the descriptors to the property producing the equation:

$$Property = \sum_{i}^{n} \alpha_i D_i + const$$

where i is the serial number and n is the total number of descriptors involved in the model, D_i denote the descriptors (with low multicollinearity), α_i and *const* are adjustable coefficients found by the least-squares method. This approach is characterized with the statistical parameters: mean square errors and *t*-values of the regression coefficients, the *F*-value, the standard deviation (*s*), and the square of the correlation coefficient (R^2).

Another widely used method in this category is Partial Least Squares (PLS) regression. PLS uses variables transformed into orthogonal principal components in the model instead of original variables. The latent variables have been obtained similarly to the PCA except that they have been related to the property values, the dependent variable, through a weight vector.[28,29]

Nonlinear methods. In the case of an intrinsically non-linear dependence between the experimental property of compounds and molecular descriptors, non-linear regression methods can be applied for the development of QSAR/QSPR equations. The intrinsic non-linear dependence may also be encoded in the respective artificial neural networks[30] (ANNs) that are capable of modeling extremely complex functional relationships. Many different types of neural network architectures have been developed over the decades with various "training" algorithms.[31] It has been argued, that the obtained ANN models lack "transparency", i.e. it is not possible to determine the amount or direction of the influence of the model parameters controlling the property. Nevertheless, ANNs have continuously been shown to be able to model diverse data sets with higher accuracy than the linear methods.[32]

Selection of descriptors to the models is a critical part of the QSAR development. It can be accomplished by logical reasoning according to the molecular forces governing the studied phenomenon. Or, in case of availability of a large pool of theoretical descriptors, a fully automated statistical procedure can be used for mining of the chemical information; the selected descriptors may reveal new knowledge and broaden understanding about the mechanism of action. An optimal set of descriptors for description of the studied property is the minimal set needed to reveal only the main similarities and differences present in the data in order to avoid over-fitting and loss of generalization. As a rule, the choice of the descriptor selection algorithm is made according to the nature of the modeling method. In publication IV of this thesis we have followed this direction by using a specific heuristic computational module to facilitate non-linear parameter selection for the neural network model development. For linear relationships PCA, simulated annealing or forward selection are frequently used; for non-linear modeling genetic algorithms can be utilized.[33]

Usually, the modeling algorithm provides a battery of QSARs with similar statistical characteristics. The selection of the appropriate model is made by the analysis of the relevance of the molecular descriptors to the studied property and by the contribution of the individual descriptors to the improvement of the R^2 and the cross-validated R^2 of the model. However, recently, the concept of consensus modeling was introduced to the predictive QSARs. In this approach, ten of the best QSAR models with nearly equivalent statistical characteristics but consisting of different descriptors are used to predict the property and the received values are averaged. As a rule, the consensus model appears to have higher fitting and prediction ability than any of the individual models.[34]

For molecular science and engineering an open computing grid, OpenMolGRID, system has been recently created.[35] It provides grid enabled components, such as a data warehouse for chemical data and software for building QSPR/QSAR models, including conformational search and geometry optimization, descriptor generation and the statistical tools for descriptor selection and model building. In addition, molecular engineering tools for generating compounds with predefined chemical properties or biological activities, i.e. for solving the reverse problem of QSAR, are provided. The effectiveness of this system in generating QSAR models is shown in publication II.

1.1.5. Validation of the models

Once the model is established, its reliability and predictive ability must be determined. It is argued that unless a QSPR model is validated neither predictions nor mechanistic interpretations based on the model descriptors should be made.[36] Validation of the stability of the model can be done by a suitable internal validation method which employs regrouping of the data that was used in model development. The most common internal validation methods for MLR models are the leave-one-out and leave-many-out cross-validation, expressed by the respective squared correlation coefficients, R^2_{cv}. In these methods, one data point or a certain part of the data, respectively, is predicted using a model developed on the rest of the data. The obtained values are then collected and correlated with the experimental values to provide the R^2_{cv}. A low cross-validated coefficient compared to the R^2 of the model, is an indication of instability of the model.

Other internal validation strategies include randomization of the modeled property, also called Y-scrambling, and bootstrap resampling. In bootstrap resampling[37] regrouping of the model data is made randomly. As a result, some data points occur in the same random sample more than once, while others may never be selected. Similarly to the cross-validation methods, high R^2_{cv} value is expected for a reliable model. In the Y-randomization test[38], the Y-vector of dependent variables, is randomly shuffled and a new QSPR model is developed using the original independent-variable matrix. This process is repeated several times and the R^2 and R^2_{cv} of the obtained models are expected to have low values.

The real criterion for the validity of a QSAR model is the predictivity of an external validation set that was not involved in the model development. The validation set should span the range of the property of the training set and cover the structural application domain of the model. If the prediction result, i.e. the squared correlation coefficient R^2 of the relationship of the observed versus predicted values of the validation set, is considerably lower than that of the model, it is concluded that the model has either been over-fitted or poorly represented by the training set. A valid model with high generalization ability has the prediction R^2 and standard deviation, s, similar to those of the model. Although better statistical parameters are desirable, the appropriate level is determined by the standard deviation of the experimental measurements.

Observation of the residuals of the relationships between the observed versus predicted values may reveal compounds with highly over- or underestimated properties that are addressed as outliers to the model. Outliers are usually chemicals with an exceptional mechanism of action, or compounds outside the structural domain of the model. The latter can be determined by calculating the leverage[39] for both training and new compounds. A training set compound with high leverage has great influence on the model parameters making the model unstable.[40] High leverage of the new compound means that the predicted value is extrapolated from the model and therefore, is not reliable.

1.2. Environmentally Relevant Properties

Global development of chemical industry has brought up serious issues of human and environmental safety. According to the REACH legislation, in relation to the environment, the avoidance of chemical contamination of air, water, soil and buildings, as well as preventing damage to biodiversity are the major goals. Improved control of persistent, bioaccumulative and toxic (PBT) substances is of particular importance in this respect.[41] To this end, the following properties need to be assessed: ecotoxicology, mobility, persistence and degradability, and bioaccumulative potential. The evaluation of the variety of risks posed by industrial chemicals is a step toward the use of safer alternatives.

The properties named above are closely interrelated. For example, the fate of a chemical in soil is affected by biodegradation and abiotic transformation as well as its mobility in the soil. In real ecosystems, environmental variables, such as temperature, rainfall or sunlight, generally cannot be controlled and it may be difficult to distinguish whether an observed effect is a result of one fate process versus another. Model systems, known as microcosms, can be used to replicate the processes affecting the fate of a chemical in complex ecosystems. However, QSAR methodology needs experimental values from reproducible standardized laboratory tests for modeling individual environmentally important phenomena.

Atmospheric degradation of volatile organic compounds that cause damage to the ozone layer by reaction with photochemically generated oxidants, such as OH and NO_3 radicals and ozone is characterized via oxidation rate constants. Hydroxyl radical reactions are the predominant pathway and they are typically grouped into four main types: (1) hydrogen atom abstraction, (2) addition to double and triple bonds, (3) addition to aromatic rings, and (4) reactions with nitrogen, sulfur, and phosphorus. Models have been developed both for the individual reaction types and including all four of the reaction classes.[42,43] Atkinson's estimation method using group/fragment methodology combined with the knowledge of reaction mechanisms has been implemented in the US EPA's AOPWIN estimation software.[44,45]

Advances in the QSAR study of atmospheric degradation of chemicals are quite recognized,[46,47] while modeling of biodegradability in water and soil has produced very modest results ($R^2 < 0.5$)[48]. Biodegradability has been expressed in a diversity of parameters including half-lives, various biodegradation rates and constants, theoretical oxygen demand, biological oxygen demand (BOD), etc., and mainly models for homological compounds have been made. The group contribution technique based on BIODEG evaluated biodegradation database presented by Howard *et al.*[49] has been highlighted as the most advantageous for use in broad screening for tendency to biodegrade. The assessment of CO_2 and BOD allows an accurate determination of biodegradation processes, and continuous methods can be used for analysis. Especially the end product, carbon dioxide, is an important parameter in the estimation of the mineralization of a test compound. (CO_2-evolution test is used for determination of the ultimate biodegradability of organic compounds by aerobic microorganisms.)

The mobility and distribution of the pollutants or their degradation products in soil (as well as sewage sludge) can be assessed by the soil sorption coefficient, the partitioning capacity of the compound between soil and water, normalized to the content of organic carbon, K_{OC}.[50,51] With EPA's Office of Pollution Prevention and Toxics, Meylan *et al.* developed a log K_{OC} prediction method based on the first-order molecular connectivity index that is available as Pckoc program.[52] The nonpolar compounds were correlated with the connectivity index. The second regression was developed by using the deviations between the measured log K_{OC} and estimated log K_{OC} values with the nonpolar equation and the number of certain structural fragments in the polar compounds.

Considerable efforts have been devoted to the QSAR studies of acute toxicity to the aquatic species.[53,54,55] Toxicity is viewed as one of the biggest challenges for QSAR modeling due to the biochemical complexity of the living organisms. Quantitative as well as classification models of the toxic mechanisms have been developed to the various aquatic species, such as fish, water flea and algae, using median lethal or growth inhibitory concentrations as endpoints. All of these species have different susceptibilities to the toxicants that have to be taken into consideration. Inert organics reveal non-specific effects (baseline toxicity) on most aquatic organisms, e.g., algae, daphnids and fish, whereas inhibitors of photosynthesis that are specific toxicants towards algae, are often baseline toxicants towards daphnids and fish. A flow-through fish test is used in assessment of bioconcentration.[56,23] For environmental toxicity measurements also such endpoints as earthworm toxicity, and honeybee oral or contact acute toxicity are suggested for testing.

2. SUMMARY OF ORIGINAL PUBLICATIONS

2.1. Outline of Methods for QSAR Model Development in Current Thesis

In the current thesis, diverse and high quality data sets are used for QSAR model building and the main flow of the participating methods and software is outlined in Fig.1. The first three steps, including descriptor generation, are common in all articles. In article III, no dataset partitioning is used and in article IV, due to the very large size of the data set, one third of the data is extracted into the validation set by a simple activity sampling procedure: the data is sorted by the increasing activity values and every third compound is set aside for the validation set.

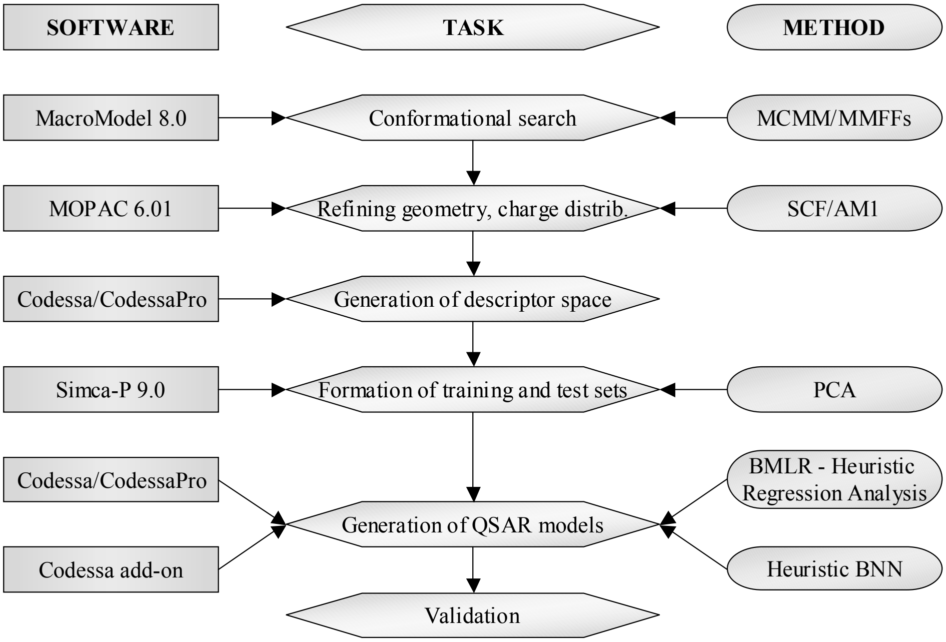


Figure 1. Software and methods used in QSAR model building in the current thesis. (Literature references are provided in the original publications of thesis, I–IV)

In all articles I–IV, BMLR is used to derive QSARs; ANN is used in article IV. For internal validation, leave-one-out and leave-many-out cross-validation are used. External validation is part of all publications with the exception of publication III of the comparative QSAAR of two aquatic species, ciliate and fish.

2.2. Soil Sorption Coefficients

Evaluation of the soil mobility of chemicals is a primary task in estimating their environmental distribution. Soil sorption coefficient is the ratio between the chemical concentration in soil and in water, normalized to organic carbon (K_{OC}). The typical model molecules for humic and fulvic acids, the constituents of the organic carbon layer of the soil, are shown in Fig. 2.

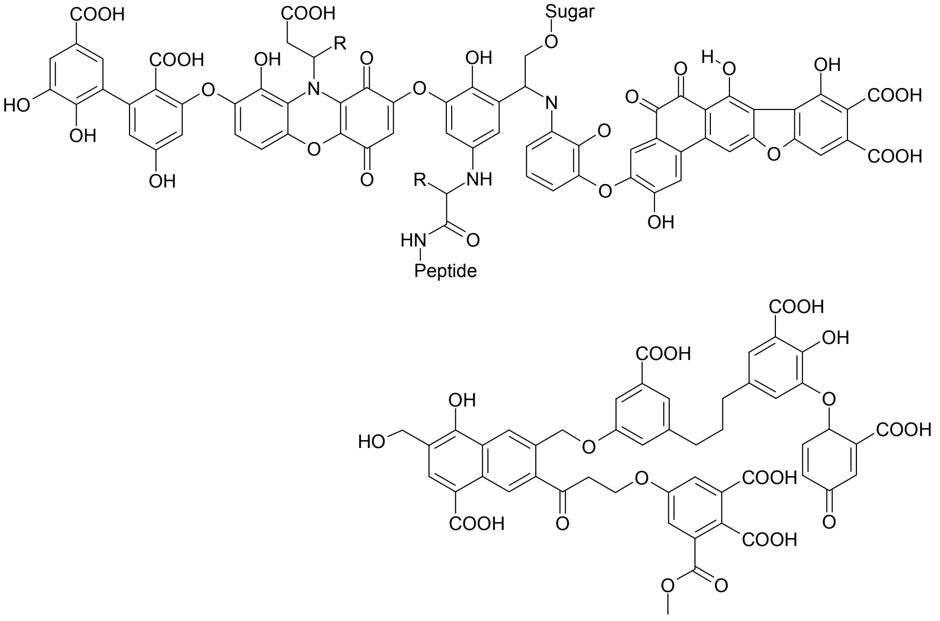


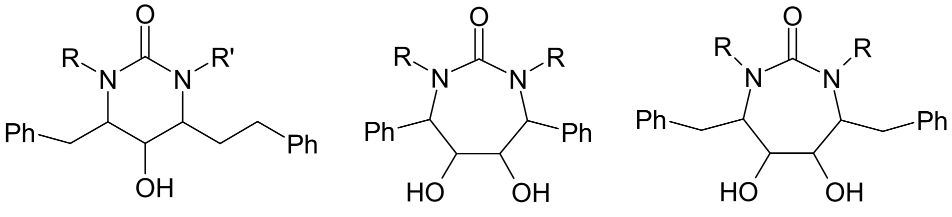
Figure 2. Typical model molecules for humic (above) and fulvic acids (below) (Stevenson, 1982, Langford *et al*., 1983)

In publication I, global and class-specific multilinear QSAR models were developed for the prediction of $\log K_{OC}$ for a structurally diverse set of 344 organic pollutants. A two-parameter model was obtained for a representative set of 68 compounds ($R^2 = 0.76$, $s = 0.44$) and a four-parameter model ($R^2 = 0.76$, $s = 0.41$) for the full set of 344 compounds. According to the extensive analysis of the model performance by internal and external validation, and by the chemical classes, the 4-parameter model demonstrated higher stability and was proposed as the model for use in prediction of the soil sorption coefficients. The applicability domain of this QSAR was determined by the 14 chemical classes involved in the development of the model and by the range of the n-octanol/water partition coefficient, $\log K_{OW}$, from –1 to 7. Amides and triazines were poorly described by this model; for these classes use of the class-specific models was suggested.

Modeling of the individual chemical classes resulted in one- to four-parameter QSARs, logK_{OW} being present only in four of them. Analysis of the descriptors and the model coefficients indicated that larger size and bulkier shape favor nonspecific interactions with the soil constituents and the humic matrix. The charge distribution describes nonspecific polar and specific interactions either with water while floating through the soil or with the soil reducing the mobility of the contaminants. The presence of reactivity indices in the QSAR models indicates that chemical reactivity affects soil sorption for some chemical classes.

2.3. Data Splitting with Principal Component Analysis

In publication II, the OpenMolGRID system was characterized in detail and used for the modeling of 80 non-peptide aspartyl protease inhibitors based on cyclic ureas (Fig. 3) as an example. Efficient inhibition of this enzyme can combat HIV-1 via the production of non-infectious viral particles. The inhibitory activity was expressed as log(1/K_i).



1 2 3

Figure 3. Structures of the substituted six- and seven-membered cyclic ureas used in modeling.

One objective of this study was to investigate the potential of Principal Component Analysis (PCA) as a clustering method for designing a structurally representative training set for the QSAR modeling. Five clusters of structurally different HIV-inhibitors were identified from the score plots of the first two principal components (PCs) based on 394 calculated molecular descriptors. Further, for each cluster, a separate PCA was performed and five compounds (four from the corners and one from the center) from the score plots of the two first PCs were selected for the training set. In the end, the H-substituted compound of type 1, R=R'=H, was also included, to provide a set of 26 compounds. Proceeding from this training set, the selected three-parameter

QSAR model had $R^2 = 0.86$, $R^2_{cv} = 0.81$, $s = 0.71$, and $F = 44.88$. The external validation of the model with 54 compounds gave $R^2 = 0.61$ and $s = 0.82$.

The cluster-based factor analysis resulted in a small representative training set and good statistical parameters for the developed QSAR model. The validation result was also satisfactory exceeding that of predictions obtained from other models that were trained on twice as many compounds but used less consistent methods, such as random selection and activity sampling, for training set formation. The relatively large difference between the R^2 of the fit and prediction shows that there is still room for improvement of the structural representation of the training set which could also lead to a different descriptor content of the QSAR model.

2.4. Comparative QSAAR for Toxicity to *Tetrahymena pyriformis* and *Pimephales promelas*

The emphasis of the European REACH legislation in generating toxicological information is on methods and models that provide an alternative to the use of animals in toxicity testing. Using surrogate assays of lower species to predict the toxicity of higher species by Quantitative Activity-Activity Relationship (QAAR) methods has been recognized as a primary source to fill the existing data gaps in toxicological databases. Compared to *in vivo* testing, *in vitro* surrogates are more economical and rapid, allow extrapolations to other species and can broaden knowledge about mechanisms of toxic action.

In publication III, relative toxic effects of 364 compounds common to both species were assessed to the freshwater fish *Pimephales promelas* log ($1/LC_{50}$) and ciliate *Tetrahymena pyriformis* log ($1/IGC_{50}$) endpoints. Good agreement, with the slope of unity, between toxic potencies ($R^2 = 0.75$) was found using ciliate toxicity as a surrogate for the fish toxicity. The intercept of about half a log unit indicates the higher sensitivity of the fish test to the toxicants that is due to the flow-through design of the fish test as opposed to the static design of the ciliate assay and the longer duration of the fish test. With the addition of three theoretical molecular descriptors, the model was significantly improved ($R^2 = 0.82$). The received Quantitative Structure-Activity-Activity Relationship (QSAAR) also showed high reliability according to the internal validation methods, the leave-one-out ($R^2CV = 0.818$) and leave-50%-out cross-validation ($R^250 = 0.815$). The structural features describing the difference of the two organisms were related to the average bond order of the carbon skeleton of the toxicant, its hydrogen-bonding ability and relative nitrogen content. Compounds considerably more toxic to the fish than the ciliate were small and intrinsically reactive via various electrophilic mechanisms. Among them were allyl and propargyl alcohols that are pro-electrophiles that undergo oxidation via alcohol dehydrogenase to an *α,β*-unsaturated aldehyde or ketone which can act as strong electrophiles.

2.5. Toxicity of Chemicals to *Tetrahymena pyriformis*

For the assessment of the environmental impact of toxicants, the unicellular ciliated protozoan, *Tetrahymena pyriformis*, is attractive for its fast growth rates and inexpensive assays. In publication IV, median population growth impairment concentration data ($\log(1/IGC_{50})$) to *Tetrahymena* from a 40-h assay of 1371 compounds spanning a variety of mechanisms of toxic action, including narcoses and electrophilic mechanisms, was used in QSAR development. The ability of a back-propagation ANN coupled to the heuristic feature selection algorithm (hBNN) to model compounds with a variety of toxic mechanisms in one global model was investigated.

The BMLR model (R^2 = 0.726, s = 0.551) showed very high stability according to the internal validation (R^2_{CV} = 0.721) and external validation on 457 compounds (R^2 = 0.720, s = 0.561). The compounds with reactive mechanisms that appear more toxic to *Tetrahymena*, were modeled with moderate accuracy revealing several series of chemicals with strongly underestimated toxicity. The largest residuals belonged to the group of carbonyl-containing *α,β*-unsaturated compounds that act via irreversible covalent mode of action as direct acting electrophiles. The second significant group of aliphatic chemicals with highly underestimated toxicity values was the *α*-haloactivated compounds that react preferably by the S_N2 displacement mechanism with the halogen atom as a strong leaving group. From among the aromatic compounds, hydroquinones and *p*-substituted phenols had high residuals. These compounds are susceptible to oxidation to the respective quinones that react by free radical formation initiating a number of competing processes within the cell. All of these chemicals are typical outliers in linear QSAR models.

The heuristic feature selection module incorporated to the ANN algorithm was able to relate different descriptors to the studied response compared to the multilinear procedure. Only the n-octanol/water partition coefficient, logP, was common in both models. The R^2 of the hBNN model improved considerably in comparison with the multi-linear regression, from 0.726 to 0.826, respectively. Considering the very high diversity of the data set, the hBNN model provided excellent prediction with R^2=0.794 and s=0.484 on the set of 457 compounds not used in model development.

2.6. Concluding Remarks

In current thesis, the emphasis of QSAR model development was focused on the environmental properties and ecotoxicology. Models with good predictive power were obtained for the soil sorption coefficient and the acute toxicity to the aquatic unicellular organism, *Tetrahymena pyriformis*. Both data sets are recognized as of high quality for the study of environmental fate of organic industrial pollutants.

Potential of the QSAR methodology was also explored for the possibility to use a nonvertebrate species as a surrogate in modeling the toxicity to the vertebrate as one way to reduce expensive animal testing. In this case, toxic potency of the unicellular ciliate, *Tetrahymena pyriformis*, was used to model the toxicity to the freshwater fish, *Pimephales promelas*. The ability of the molecular descriptors to take into account the interspecies differences was addressed in particular. Significant improvement in the correlation coefficient was obtained with including three easily interpretable molecular descriptors into the model.

In the course of the work, several aspects of the QSAR methodology were evaluated or improved. HIV-1 aspartyl protease inhibitors were used as a suitable data set for investigating the ability of PCA as a multivariate clustering method to form a structurally representative training set for QSAR model development. Considerable improvement of the prediction was obtained compared to the models derived on the training sets obtained by more arbitrary selection methods carried out previously on the same data.

In the publication about soil sorption coefficients it was shown that modeling with a small size of the training set benefits highly from the use of a validation set during the selection of the final model with the highest generalization ability and hence the highest potential prediction capability. This statement was proved by using an additional independent external validation set for making predictions compiled from a different source in the literature.

And finally, a heuristic feature selection module incorporated to the ANN algorithm was tested with modeling the median population growth inhibition concentration to the ciliate *Tetrahymena pyriformis*. This method enabled to consider the nonlinear relationship of each descriptor with the studied property in course of the model development. The resulting model contained different descriptors and had significantly higher statistical parameters of prediction compared to the corresponding linear model.

REFERENCES

1. Cros, A. F. A. PhD Thesis, University of Strasbourg 1863; cited from: S. Borman, New QSAR Techniques Eyed for Environmental Assessments. *Chem. Eng. News* **1990**, February 19, 20–23.
2. Katritzky, A. R.; Fara, D. C.; Petrukhin. R.; Tatham, D. B.; Maran, U.; Lomaka, A.; Karelson, M. The Present Utility and Future Potential for Medicinal Chemistry of QSAR/QSPR with Whole Molecule Descriptors. *Curr. Top. Med. Chem.* **2002**, *2*, 1333–1356.
3. Jorgensen, W. L. The Many Roles of Computation in Drug Discovery. *Science* **2004**, *303*, 1813–1818.
4. Anon. *Proposal Concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH)*. COM(2003)644 final. Brussels, Belgium: European Commission, 2003. Website http://europa.eu.int (Accessed 16.06.07).
5. Van der Jagt, K.; Munn, S.; Tørsløv, J.; de Bruijn, J. Alternative Approaches can Reduce the Use of Test Animals under REACH. Addendum to the Report "Assessment of Additional Testing Needs under REACH. Effects of (Q)SARs, Risk Based Testing and Voluntary Industry Initiatives". JRC Report EUR 21405 EN, 25 pp., 2004. Ispra, Italy: European Commission Joint Research Centre. Website http://ecb.jrc.it (Accessed 16.06.07).
6. Anon. (Brussels, 1st June 2007) New European Chemicals Agency Starts Operations as REACH Enters into Force. IP/07/745
7. Cronin, M. T. D.; Schultz, T. W. Pitfalls in QSAR. *J. Mol. Struct.–THEOCHEM* **2003**, *622*, 39–51.
8. Russom, C. L.; Bradbury, S. P.; Broderius, S. J.; Hammermeister, D. E.; Drummond, R. A. Predicting Modes of Toxic Action from Chemical Structure: Acute Toxicity in the Fathead Minnow (*Pimephales promelas*). *Environ. Toxicol. Chem.* **1997**, *16*, 948–967.
9. Schultz, T. W. Tetratox: *Tetrahymena pyriformis* Population Growth Impairment Endpoint – A Surrogate for Fish Lethality. *Toxicol. Meth*. **1997**, *7*, 289–309.
10. Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*, Wiley- VCH: Weinheim, Germany, 2000.
11. Karelson, M. *Molecular Descriptors in QSAR/QSPR*; John Wiley & Sons: New York, U.S.A., 2000.
12. Meylan, W. M.; Howard, P. H. Atom/Fragment Contribution Method for Estimating Octanol-Water Partition Coefficients. *J. Pharm. Sci.* **1995**, *84*, 83–92.
13. ACD/LogP, prediction of the octanol-water partition coefficient for neutral compounds. Website: http://www.acdlabs.com
14. ClogP 4.0, Biobyte. Website: http://www.biobyte.com
15. Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure-Activity Analysis;* John Wiley and Sons: New York, 1986.
16. Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902–3909.

17. Netzeva, T. I.; Aptula, A. O.; Benfenati, E.; Cronin, M. T. D.; Gini, G.; Lessigiarska, I.; Maran, U.; Vracko, M.; Schüürmann, G. Description of the Electronic Structure of Organic Chemicals Using Semiempirical and *ab initio* Methods for Development of Toxicological QSARs. *J. Chem. Inf. Model.* **2005**, *45*, 106–114.
18. Stanton, D. T.; Jurs, P. Development and Use of Charged Partial Surface Area Structural Descriptors in Computer-Assisted Quantitative Structure-Property Relationship Studies. *Anal. Chem.* **1990**, *62*, 2323–2329.
19. Hetenyi, C.; Paragi, G.; Maran, U.; Timar, Z.; Karelson, M.; Penke, B. Combination of a Modified Scoring Function with Two-Dimensional Descriptors for Calculation of Binding Affinities of Bulky, Flexible Ligands to Proteins. *J. Am. Chem. Soc.* **2006**, *128*, 1233–1239.
20. Jolliffe, I. T. *Principal Component Analysis.* Springer-Verlag, New York, 1986.
21. Zupan, J.; Novic, M.; Ruisa´nchez, I. Kohonen and Counter Propagation Artificial Neural Networks in Analytical Chemistry. *Chemom. Int. Lab. Syst.* **1997**, *38*, 1–23.
22. Wu, W.; Walczak, B.; Massart, D. L.; Erni, F.; Last, I. R.; Prebble. K, A. Artificial Neural Networks in Classification of NIR Spectral Data: Design of the Training Set. *Chemometr. Intell. Lab. Syst.* **1996**, *33*, 35–46.
23. Gramatica, P.; Papa, E. QSAR Modeling of Bioconcentration Factor by Theoretical Molecular Descriptors. *QSAR Combin. Sci.* **2003**, *22*, 374–385.
24. Jaworska, J.; Nikolova-Jeliazkova, N.; Aldenberg, T. QSAR Applicability Domain Estimation by Projection of the Training Set in Descriptor Space: A Review. *ATLA-Altern. Laborat. Anim.* **2005**, *33*, 445–459.
25. Netzeva, T. I.; Worth, A. P.; Aldenberg, T.; Benigni, R.; Cronin, M. T. D.; Gramatica, P.; Jaworska, J. S.; Kahn, S.; Klopman, G.; Marchant, C. A.; Myatt, G.; Nikolova-Jeliazkova, N.; Patlewicz, G. Y.; Perkins, R.; Roberts, D. W.; Schultz, T. W.; Stanton, D. T.; van de Sandt, J. J. M.; Tong, W.; Veith, G.; Yang, C. Current Status of Methods for Defining the Applicability Domain of (Quantitative) Structure-Activity Relationships. – The Report and Recommendations of ECVAM Workshop 52. *ATLA-Altern. Laborat. Anim.* **2005**, *33*, 155–173.
26. De Roode, D.; Hoekzema, C.; de Vries-Buitenweg, S.; van de Waart, B.; Van der Hoeven, J. QSARs in Ecotoxicological Risk Assessment. *Regul. Toxicol. Pharmacol.* **2006**, *45*, 24–35.
27. Draper, N. R.; Smith, H. *Applied Regression Analysis*, Wiley, New York, 1966.
28. Dunn, W. J.; Wold, S.; Edlund, U.; Hellberg, S. Multivariate Structure-Activity Relationships Between Data from a Battery of Biological Tests and an Ensemble of Structure Descriptors: The PLS Method. *QSAR*, **1984**, *3*, 131–137.
29. Hoskuldsson, A. PLS Regression Methods. *J. Chemom.* **1988**, *2*, 211–228.
30. Zupan, J.; Gasteiger, J. Neural Networks for Chemists: An Introduction. VCH, 1993.
31. Schneider, G.; Wrede, P. Artificial Neural Networks for Computer-Based Molecular Design. *Progr. Biophys. Mol. Biol.* **1998**, *70*, 175–222.
32. Kaiser, K. L. E. The Use of Neural Networks in QSARs for Acute Aquatic Toxicological Endpoints. *J. Mol. Struct.–THEOCHEM* **2003**, *622*, 85–95.
33. Serra, J. R.; Jurs, P. C.; Kaiser, K. L. E. Linear Regression and Computational Neural Network Prediction of *Tetrahymena* Acute Toxicity for Aromatic Compounds from Molecular Structure. *Chem. Res. Toxicol.* **2001**, *14*, 1535–1545.

34. Gramatica, P.; Giani, E.; Papa, E. Statistical External Validation and Consensus Modeling: A QSPR Case Study for K_{OC} Prediction. *J. Mol. Graphics Modell.* **2007**, *25*, 755–766.
35. Sild, S.; Maran, U.; Lomaka, A.; Karelson, M. Open Computing Grid for Molecular Science and Engineering. *J. Chem. Inf. Model.* **2006**, *46*, 953–959.
36. Tropsha, A.; Gramatica, P.; Gombar, V. K. The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. *QSAR Combin. Sci.* **2003**, *22*, 69–77.
37. Efron, B.; Tibshirani, R. J. *An Introduction to the Bootstrap*, Chapman & Hall: New York, 1993.
38. Wold, S.; Eriksson, L. Statistical Validation of QSAR Results. In *Chemometrics methods in Molecular Design*, van de Waterbeemd, H., Ed.; VCH: Veinheim Germany, 2001; pp. 309–318.
39. Atkinson, A. C. *Plots, Transformations and Regression*. Oxford: Clarendon Press, 1985.
40. Eriksson, L.; Jaworska, J.; Worth, A. P.; Cronin, M. T. D.; McDowell, R. M.; Gramatica, P. Methods for Reliability and Uncertainty Assessment and for Applicability Evaluations of Classification- and Regression-Based QSARs. *Environ. Health Perspect.* **2003**, *111*, 1361–1375.
41. REACH COM (2003) 644, 29 October 2003 – CBI position on European Parliament and Council regulation.
42. Gramatica P, Pilutti P, Papa E. Validated QSAR Prediction of OH Tropospheric Degradation of VOCs: Splitting into Training-Test Sets and Consensus Modeling. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1794–1802.
43. Öberg T. A QSAR for the Hydroxyl Radical Reaction Rate Constant: Validation, Domain of Application, and Prediction. *Atmos. Environ.* **2005**, *39*, 2189–2200.
44. *AOPWIN,* Ver.1.90. Environmental Protection Agency, U.S.A., 2000.
45. Atkinson, R. A. Structure-Activity Relationship for the Estimation of Rate Constants for the Gas-Phase Reactions of OH Radicals with Organic Compounds. *Int. J. Chem. Kinet.* **1987**, *19*, 799–828.
46. Güsten, H. Predicting the Abiotic Degradability of Organic Pollutants in the Troposphere. *Chemosphere* **1999**, *38*, 1361–1370.
47. Meylan, W. M.; Howard, P. H. A Review of Quantitative Structure-Activity Relationship Methods for the Prediction of Atmospheric Oxidation of Organic Chemicals. *Environ. Toxicol. Chem.* **2003**, *22*, 1724–1732.
48. Raymond, J. W.; Rogers, T. N.; Shonnard, D. R.; Kline, A. A. A Review of Structure-Based Biodegradation Estimation Methods. *J. Hazard. Mater.* **2001**, *84*, 189–215.
49. Howard, P. H.; Boethling, R. S.; Stiteler, W. M.; Meylan, W. M.; Hueber, A. E.; Beauman, J. A.; Larosche, M. E. Predictive Model for Aerobic Biodegradability Developed from a File of Evaluated Biodegradation Data. *Environ. Toxicol. Chem.* **1992**, *11*, 593–603.
50. Wauchope, R. D.; Yeh, S.; Linders, J. B. H. J.; Kloskowski, R.; Tanaka, K.; Rubin, B.; Katayama, A.; Kördel, W.; Gerstl, Z.; Lana, M.; Unsworth, J. B. Review. Pesticide Soil Sorption Parameters: Theory, Measurement, Uses, Limitations and Reliability. *Pest Management Sci.* **2002**, *58*, 419–445.

51. Doucette, W. J. Quantitative Structure-Activity Relationships for Predicting Soil-Sediment Sorption Coefficients for Organic Chemicals. *Environ. Toxicol. Chem.* **2003**, *22*, 1771–1788.
52. Meylan, W.; Howard, P. H.; Boethling, R. S. Molecular Topology Fragment Contribution Method for Predicting Soil Sorption Coefficients. *Environ. Sci. Tech*. **1992**, *26*, 1560–1567.
53. Bradbury, S. P.; Russom, C. L.; Ankley, G. T.; Schultz, T. W.; Walker, J. D. Overview of Data and Conceptual Approaches for Derivation of Quantitative Structure-Activity Relationships for Ecotoxicological Effects of Organic Chemicals. *Environ. Toxicol. Chem.* **2003**, *22*, 1789–1798.
54. Lessigiarska, I.; Worth, A. P.; Sokull-Kluttgen, B.; Jeram, S.; Dearden, J. C.; Netzeva, T. I.; Cronin, M. T. D. QSAR Investigation of a Large Data Set for Fish, Algae and Daphnia Toxicity. *SAR QSAR Environ. Res.* **2004**, *15*, 413–431.
55. Moore, D. R. J.; Breton, R. L.; MacDonald, D. B. A Comparison of Model Performance for Six Quantitative Structure-Activity Relationship Packages That Predict Acute Toxicity to Fish. *Environ. Toxicol. Chem.* **2003**, *22*, 1799–1809.
56. Barron, M. G. Bioconcentration. *Environ. Sci. Technol.* **1990**, *24*, 1612–1618.

SUMMARY IN ESTONIAN

KESKKONNATÄHTSATE OMADUSTE KVANTITATIIVSED STRUKTUUR-AKTIIVSUS SÕLTUVUSED

Kvantitatiivsed struktuur-aktiivsus sõltuvused (QSAR'id) on empiirilised mudelid, mis seovad ainete eksperimentaalseid omadusi nende molekulaarstruktuuridega. Kvantteooria ning *ab initio* arvutusmeetodid võimaldavad eksperimentaalse täpsusega arvutada ainete omadusi väikestele isoleeritud molekulidele. Reaalsed tööstuslikud või biokeemilised protsessid elusorganismides toimuvad aga erinevates kondenseeritud keskkondades ning seetõttu omavad QSAR'id, mis lähtuvad otseselt vaadeldavast aine omadusest, teatud eeliseid aineomaduste ennustamisel mistahes keskkonnas.

Struktuur-aktiivsus sõltuvusi on dokumenteeritud juba alates 19. sajandi teisest poolest. Oma täieliku potentsiaali saavutas meetod umbes sada aastat hiljem, kui algas teoreetiliste molekulaardeskriptorite formuleerimine, mis ei vaja mingeid eksperimentaalseid andmeid ja on seetõttu arvutatavad veel sünteesimata või hüpoteetilistele ainetele. Laialdasematest reakendusaladest võib nimetada meditsiini ja ravimidisaini. Hiljuti omandas QSAR olulise positsiooni rahvusvahelises seadusandluses, milles nähakse valideeritud QSAR mudeleid kui primaarsete andmete saamise allikat keskkonda ja toksilisust puudutavate andmete genereerimisel nii olemasolevatele kui ka tulevastele laialdaselt kasutatavatele kemikaalidele. Käesoleva väitekirja peamiseks eesmärgiks oli hinnata saasteainete mõju keskkonnale, täpsemalt, ainete absorbeerumiskonstanti mullas ja toksilisust magevee organismidele, QSAR meetodite abil. Töö käigus hinnati ka mitmeid QSAR metodoloogilisi aspekte.

Esimene peatükk annab ülevaate QSAR metodoloogiast ja keskkonnatähtsatest omadustest. Tähtsaimad nõuded QSAR mudelite edukaks moodustamiseks algavad kõrge usaldatavusega ja ühtse määramistäpsusega eksperimentaalse andmebaasi valikust. Seejärel teostatakse vajalikul teooria tasemel molekulide geomeetria optimiseerimine ning arvutatakse molekulaardeskriptorid lähtudes saadud geomeetriast. Olulise tähtsusega on treeningkomplekti koostamine mudeli moodustamiseks, kuna sellega defineeritakse ühtlasi ka mudeli kasutuspiirid. Järgneb sobiva statistilise meetodi valik deskriptorite selekteerimiseks ja regressiooni tuletamiseks, mille ennustusvõimet kasutuspiiride ulatuses hinnatakse sõltumatu valideerimiskomplekti abil.

Rahvusvaheliste keskkonnakaitse organite poolt on esitatud nõudmised ennustusvõimeliste QSAR mudelite moodustamiseks järgmistes valdkondades: õhu, vee, mulla ja ehitiste keemilise saaste ja loodusliku mitmekesisuse hävitamise vältimiseks. Selleks on vaja hinnata ainete transporti, keemilist stabiilsust või lagunemist keskkonnas, ning akumulatsiooni ja toksilisust elusorganismides. Nõuetekohased QSAR mudelid on loodud vaid üksikutele neid omadusi kirjeldavatele suurustele.

Väitekirja teises osas on kokkuvõte tehtud uurimistöö raames saadud originaalsetest tulemustest. Publikatsioonis I moodustati multilineaarse meetodi abil globaalseid ja aineklassi-spetsiifilisi QSPR mudeleid mulla absorptsiooni koefitsiendile, log K_{OC}, mis iseloomustab ainete mobiilsust mulla orgaanilist süsinikku sisaldavas kihis. Parima ennustusvõimega globaalne mudel saadi nelja deskriptori abil, kasutades kogu andmekomplekti, 344 ainet. Polaarsed ained käitusid erinevalt mittepolaarsetest, mistõttu mõnede suuremate kõrvalekaldujate puhul, nagu amiidid, soovitati kasutada klassispetsiifilist mudelit. Artiklis II uuriti põhikomponentanalüüsi (PCA) klastrimoodustamise võime kasutamist struktuuriliselt esindusliku treenimiskomplekti valikul QSAR mudeli formuleerimiseks HIV-1 aspartüülproteaasi inhibiitorite inhibeerimisvõime ennustamiseks. Võrreldes eelnevalt saadud ennustustega sama andmekomplekti jaoks, paranes tulemus uue treenimiskomplekti valikuga märgatavalt. Artiklis III kasutati magevee ainurakse organismi toksilist aktiivsust, $\log(1/IGC_{50})$, surrogaadina ehk ühena deskriptoritest toksilise aktiivsuse, $\log(1/LC_{50})$, ennustamiseks kalale. Saadud kahe liigi vahelisele korrelatsioonile lisati veel kolm teoreetilist molekulaardeskriptorit, mis tunduvalt parandasid korrelatsiooni ja üldjoontes kirjeldasid liikidevahelisi erinevusi tundlikkuse suhtes uuritud ainetele. Publikatsioonis IV kasutati eriti mitmekesise struktuurse koostise ja mehhanistliku käitumisega toksilisuste andmebaasi (1371 ainet) ainuraksele magevee organismile *Tetrahymena pyriformisele*. Andmebaasi abil testiti heuristilist parameetrite selekteerimise meetodit ühenduses kunstlike närvivõrkudega (ANN), mis võimaldab deskriptorite valiku käigus arvesse võtta nende mittelineaarset iseloomu uuritud aktiivsuse suhtes. Võrreldes samadel tingimustel saadud multilineaarse mudeliga, sisaldas ANN mudel erinevaid deskriptoreid ning andis oluliselt parema ennustustulemuse andmekomplektile, mis oli eraldatud selleks otstarbeks enne mudeli ehitamise algust.

ACKNOWLEDGEMENTS

I wish to express my greatest gratitude to my doctoral advisors Professor Mati Karelson and Dr. Uko Maran for their expertise and guidance throughout the research. Many thanks are due to my colleges Dr. Sulev Sild for excellent collaboration and Dr. Andre Lomaka for valuable discussions.

Sincere thanks will go to Kenan Professor Alan R. Katritzky for the opportunity to work in his group of computational chemistry in the University of Florida, Gainesville, FL, USA, and to Professor Emilio Benfenati to join his group in the Mario Negri Pharmacological Institute in Milan, Italy.

Sincere thanks are also in store for my father, Aavo (d. 2004), for his shared inspiration for biochemistry, my mother, Elme, and my daughter, Karoli, for their support and loving presence.

PUBLICATIONS

I

Iiris Kahn, Dan Fara, Mati Karelson, Uko Maran and Patrik L. Andersson,
QSPR Treatment of the Soil Sorption Coefficients of Organic Pollutants.
Journal of Chemical Information and Modeling **2005**, 45, 94–105.

II

Uko Maran, Sulev Sild, Iiris Kahn and Kalev Takkis,
Mining of the Chemical Information in GRID Environment.
Future Generation Computer Systems **2007**, 23, 76–83.

III

Iiris Kahn, Uko Maran, Emilio Benfenati, Tatiana I. Netzeva,
T. Wayne Schultz and Mark T. D. Cronin,

Comparative Quantitative Structure-Activity-Activity Relationships for Toxicity to *Tetrahymena pyriformis* and *Pimephales promelas*.

ATLA-Alternatives to Laboratory Animals **2007**, 35, 15–24.

IV

Iiris Kahn, Sulev Sild and Uko Maran,
Modeling the Toxicity of Chemicals to *Tetrahymena pyriformis* Using Heuristic Multi-Linear Regression and Heuristic Back-Propagation Neural Networks.
Journal of Chemical Information and Modeling **2007**, submitted.

CURRICULUM VITAE

IIRIS KAHN

Born: June 7th, 1970, Tartu, Estonia
Citizenship: Estonian
Address: Tallinn Technical University, Institute of Chemistry,
Akadeemia tee 15, Tallinn 12618, Estonia
Tel: +372 620 2819, 55602081
E-mail: ikahn@ttu.ee

Education

2003–present *Ph.D.* student of Molecular Engineering, doctoral advisors prof. Mati Karelson and Dr. Uko Maran, University of Tartu, Estonia

2003 *M.Sc.* in Molecular Engineering, University of Tartu, Estonia

1994 *B.Sc.* in Chemistry (specialty biochemistry), University of Tartu, Estonia

Professional Experience

2006–present Researcher, Tallinn Technical University, Institute of Chemistry, Estonia

2005–2006 Extraordinary Researcher, Tallinn Technical University, Institute of Chemistry, Estonia

2000–2005 Chemist, Department of Chemistry, University of Tartu, Estonia

1997–1999 Laboratory Assistant, University of Missouri-Columbia, Department of Biochemistry, USA

Publications

1. Kahn, I.; Maran, U.; Karelson, M. Quantum-Chemical Modelling of the Solvatochromic Shift in Electronic Spectra. *Acta et Comment. Univ. Tartuensis* **1994**, *975*, 15–29.
2. Kahn, I.; Fara, D.; Karelson, M.; Maran, U.; Andersson, P.L. QSPR Treatment of the Soil Sorption Coefficients of Organic Pollutants. *J. Chem. Inf. Model.* **2005**, *45*, 94–105.
3. Maran, U.; Sild, S.; Kahn, I.; Takkis, K. Mining of the Chemical Information in GRID Environment. *Future Generat. Comput. Syst.* **2007**, *23*, 76–83.
4. Kahn, I.; Maran, U.; Benfenati, E.; Netzeva, T.I.; Schultz, T.W.; Cronin, M.T.D. Comparative Quantitative Structure-Activity-Activity Relationships for Toxicity to *Tetrahymena pyriformis* and *Pimephales promelas*. *ATLA – Altern. Laborat. Anim.* **2007**, *35*, 15–24.
5. Kahn, K.; Kahn, I. Improved Efficiency of Focal Point Conformational Analysis with Truncated Correlation Consistent Basis Sets. *J. Comput. Chem.* **2007**, accepted.
6. Kahn, I.; Sild, S.; Maran, U. Modelling the Toxicity of Chemicals to *Tetrahymena pyriformis* Using Heuristic Multi-Linear Regression and Heuristic Back-Propagation Neural Networks. *J. Chem. Inf. Model.* **2007**, submitted.

ELULOOKIRJELDUS

IIRIS KAHN

Sündinud:	7. juuni 1970, Tartu, Eesti
Kodakondsus:	eesti
Aadress:	Tallinna Tehnikaülikool, keemiainstituut, Akadeemia tee 15, Tallinn 12618, Eesti Tel.: +372 620 2819, 55602081 E-post: ikahn@ttu.ee

Haridus

2003–praegu	molekulaartehnoloogia doktorant, juhendajad prof. Mati Karelson ja Dr. Uko Maran, Tartu Ülikool
2003	*M.Sc.* molekulaartehnoloogias, Tartu Ülikool
1994	*B.Sc.* keemias biokeemia erialal, Tartu Ülikool

Erialane kogemus

2006 – praegu	Teadur, Tallinna Tehnikaülikool, Keemiainstituut
2005 – 2006	Erakorraline teadur, Tallinna Tehnikaülikool, Keemiainstituut
2000 – 2005	Keemik, Tartu Ülikool, Keemilise Füüsika Instituut
1997 – 1999	Laborant, Missouri-Columbia Ülikool, biokeemia osakond, USA

Publikatsioonid

1. Kahn, I.; Maran, U.; Karelson, M. Quantum-Chemical Modelling of the Solvatochromic Shift in Electronic Spectra. *Acta et Comment. Univ. Tartuensis* **1994**, *975*, 15–29.
2. Kahn, I.; Fara, D.; Karelson, M.; Maran, U.; Andersson, P.L. QSPR Treatment of the Soil Sorption Coefficients of Organic Pollutants. *J. Chem. Inf. Model.* **2005**, *45*, 94–105.
3. Maran, U.; Sild, S.; Kahn, I.; Takkis, K. Mining of the Chemical Information in GRID Environment. *Future Generat. Comput. Syst.* **2007**, *23*, 76–83.
4. Kahn, I.; Maran, U.; Benfenati, E.; Netzeva, T.I.; Schultz, T.W.; Cronin, M.T.D. Comparative Quantitative Structure-Activity-Activity Relationships for Toxicity to *Tetrahymena pyriformis* and *Pimephales promelas*. *ATLA-Altern. Laborat. Anim.* **2007**, *35*, 15–24.
5. Kahn, K.; Kahn, I. Improved Efficiency of Focal Point Conformational Analysis with Truncated Correlation Consistent Basis Sets. *J. Comput. Chem.* **2007**, accepted.
6. Kahn, I.; Sild, S.; Maran, U. Modelling the Toxicity of Chemicals to *Tetrahymena pyriformis* Using Heuristic Multi-Linear Regression and Heuristic Back-Propagation Neural Networks. *J. Chem. Inf. Model.* **2007**, submitted.

DISSERTATIONES CHIMICAE UNIVERSITATIS TARTUENSIS

1. **Toomas Tamm.** Quantum-chemical simulation of solvent effects. Tartu, 1993, 110 p.
2. **Peeter Burk.** Theoretical study of gas-phase acid-base equilibria. Tartu, 1994, 96 p.
3. **Victor Lobanov.** Quantitative structure-property relationships in large descriptor spaces. Tartu, 1995, 135 p.
4. **Vahur Mäemets.** The ^{17}O and ^{1}H nuclear magnetic resonance study of H_2O in individual solvents and its charged clusters in aqueous solutions of electrolytes. Tartu, 1997, 140 p.
5. **Andrus Metsala.** Microcanonical rate constant in nonequilibrium distribution of vibrational energy and in restricted intramolecular vibrational energy redistribution on the basis of slater's theory of unimolecular reactions. Tartu, 1997, 150 p.
6. **Uko Maran.** Quantum-mechanical study of potential energy surfaces in different environments. Tartu, 1997, 137 p.
7. **Alar Jänes.** Adsorption of organic compounds on antimony, bismuth and cadmium electrodes. Tartu, 1998, 219 p.
8. **Kaido Tammeveski.** Oxygen electroreduction on thin platinum films and the electrochemical detection of superoxide anion. Tartu, 1998, 139 p.
9. **Ivo Leito.** Studies of Brønsted acid-base equilibria in water and non-aqueous media. Tartu, 1998, 101 p.
10. **Jaan Leis.** Conformational dynamics and equilibria in amides. Tartu, 1998, 131 p.
11. **Toonika Rinken.** The modelling of amperometric biosensors based on oxidoreductases. Tartu, 2000, 108 p.
12. **Dmitri Panov.** Partially solvated Grignard reagents. Tartu, 2000, 64 p.
13. **Kaja Orupõld.** Treatment and analysis of phenolic wastewater with microorganisms. Tartu, 2000, 123 p.
14. **Jüri Ivask.** Ion Chromatographic determination of major anions and cations in polar ice core. Tartu, 2000, 85 p.
15. **Lauri Vares.** Stereoselective Synthesis of Tetrahydrofuran and Tetrahydropyran Derivatives by Use of Asymmetric Horner-Wadsworth-Emmons and Ring Closure Reactions. Tartu, 2000, 184 p.
16. **Martin Lepiku.** Kinetic aspects of dopamine D_2 receptor interactions with specific ligands. Tartu, 2000, 81 p.
17. **Katrin Sak.** Some aspects of ligand specificity of P2Y receptors. Tartu, 2000, 106 p.
18. **Vello Pällin.** The role of solvation in the formation of iotsitch complexes. Tartu, 2001, 95 p.

19. **Katrin Kollist.** Interactions between polycyclic aromatic compounds and humic substances. Tartu, 2001, 93 p.
20. **Ivar Koppel.** Quantum chemical study of acidity of strong and superstrong Brønsted acids. Tartu, 2001, 104 p.
21. **Viljar Pihl.** The study of the substituent and solvent effects on the acidity of OH and CH acids. Tartu, 2001, 132 p.
22. **Natalia Palm.** Specification of the minimum, sufficient and significant set of descriptors for general description of solvent effects. Tartu, 2001, 134 p.
23. **Sulev Sild.** QSPR/QSAR approaches for complex molecular systems. Tartu, 2001, 134 p.
24. **Ruslan Petrukhin.** Industrial applications of the quantitative structure-property relationships. Tartu, 2001, 162 p.
25. **Boris V. Rogovoy.** Synthesis of (benzotriazolyl)carboximidamides and their application in relations with *N*- and *S*-nucleophyles. Tartu, 2002, 84 p.
26. **Koit Herodes.** Solvent effects on UV-vis absorption spectra of some solvatochromic substances in binary solvent mixtures: the preferential solvation model. Tartu, 2002, 102 p.
27. **Anti Perkson.** Synthesis and characterisation of nanostructured carbon. Tartu, 2002, 152 p.
28. **Ivari Kaljurand.** Self-consistent acidity scales of neutral and cationic Brønsted acids in acetonitrile and tetrahydrofuran. Tartu, 2003, 108 p.
29. **Karmen Lust.** Adsorption of anions on bismuth single crystal electrodes. Tartu, 2003, 128 p.
30. **Mare Piirsalu.** Substituent, temperature and solvent effects on the alkaline hydrolysis of substituted phenyl and alkyl esters of benzoic acid. Tartu, 2003, 156 p.
31. **Meeri Sassian.** Reactions of partially solvated Grignard reagents. Tartu, 2003, 78 p.
32. **Tarmo Tamm.** Quantum chemical modelling of polypyrrole. Tartu, 2003. 100 p.
33. **Erik Teinemaa.** The environmental fate of the particulate matter and organic pollutants from an oil shale power plant. Tartu, 2003. 102 p.
34. **Jaana Tammiku-Taul.** Quantum chemical study of the properties of Grignard reagents. Tartu, 2003. 120 p.
35. **Andre Lomaka.** Biomedical applications of predictive computational chemistry. Tartu, 2003. 132 p.
36. **Kostyantyn Kirichenko.** Benzotriazole — Mediated Carbon–Carbon Bond Formation. Tartu, 2003. 132 p.
37. **Gunnar Nurk.** Adsorption kinetics of some organic compounds on bismuth single crystal electrodes. Tartu, 2003, 170 p.
38. **Mati Arulepp.** Electrochemical characteristics of porous carbon materials and electrical double layer capacitors. Tartu, 2003, 196 p.

39. **Dan Cornel Fara.** QSPR modeling of complexation and distribution of organic compounds. Tartu, 2004, 126 p.
40. **Riina Mahlapuu.** Signalling of galanin and amyloid precursor protein through adenylate cyclase. Tartu, 2004, 124 p.
41. **Mihkel Kerikmäe.** Some luminescent materials for dosimetric applications and physical research. Tartu, 2004, 143 p.
42. **Jaanus Kruusma.** Determination of some important trace metal ions in human blood. Tartu, 2004, 115 p.
43. **Urmas Johanson.** Investigations of the electrochemical properties of polypyrrole modified electrodes. Tartu, 2004, 91 p.
44. **Kaido Sillar.** Computational study of the acid sites in zeolite ZSM-5. Tartu, 2004, 80 p.
45. **Aldo Oras.** Kinetic aspects of dATPαS interaction with $P2Y_1$ receptor. Tartu, 2004, 75 p.
46. **Erik Mölder.** Measurement of the oxygen mass transfer through the air-water interface. Tartu, 2005, 73 p.
47. **Thomas Thomberg.** The kinetics of electroreduction of peroxodisulfate anion on cadmium (0001) single crystal electrode. Tartu, 2005, 95 p.
48. **Olavi Loog.** Aspects of condensations of carbonyl compounds and their imine analogues. Tartu, 2005, 83 p.
49. **Siim Salmar.** Effect of ultrasound on ester hydrolysis in aqueous ethanol. Tartu, 2006, 73 p.
50. **Ain Uustare.** Modulation of signal transduction of heptahelical receptors by other receptors and G proteins. Tartu, 2006, 121 p.
51. **Sergei Yurchenko.** Determination of some carcinogenic contaminants in food. Tartu, 2006, 143 p.
52. **Kaido Tämm.** QSPR modeling of some properties of organic compounds. Tartu, 2006, 67 p.
53. **Olga Tšubrik.** New methods in the synthesis of multisubstituted hydrazines. Tartu. 2006, 183 p.
54. **Lilli Sooväli.** Spectrophotometric measurements and their uncertainty in chemical analysis and dissociation constant measurements. Tartu, 2006, 125 p.
55. **Eve Koort.** Uncertainty estimation of potentiometrically measured ph and pK_a values. Tartu, 2006, 139 p.
56. **Sergei Kopanchuk.** Regulation of ligand binding to melanocortin receptor subtypes. Tartu, 2006, 119 p.
57. **Silvar Kallip.** Surface structure of some bismuth and antimony single crystal electrodes. Tartu, 2006, 107 p.
58. **Kristjan Saal.** Surface silanization and its application in biomolecule coupling. Tartu, 2006, 77 p.
59. **Tanel Tätte.** High viscosity $Sn(OBu)_4$ oligomeric concentrates and their applications in technology. Tartu, 2006, 91 p.

60. **Dimitar Atanasov Dobchev**. Robust QSAR methods for the prediction of properties from molecular structure. Tartu, 2006, 118 p.
61. **Hannes Hagu**. Impact of ultrasound on hydrophobic interactions in solutions. Tartu, 2007, 81 p.
62. **Rutha Jäger.** Electroreduction of peroxodisulfate anion on bismuth electrodes. Tartu, 2007, 142 p.
63. **Kaido Viht.** Immobilizable bisubstrate-analogue inhibitors of basophilic protein kinases: development and application in biosensors. Tartu, 2007, 88 p.
64. **Eva-Ingrid Rõõm.** Acid-base equilibria in nonpolar media. Tartu, 2007, 156 p.
65. **Sven Tamp.** DFT study of the cesium cation containing complexes relevant to the cesium cation binding by the humic acids. Tartu, 2007, 102 p.
66. **Jaak Nerut.** Electroreduction of hexacyanoferrate(III) anion on Cadmium (0001) single crystal electrode. Tartu, 2007, 180 p.
67. **Lauri Jalukse.** Measurement uncertainty estimation in amperometric dissolved oxygen concentration measurement. Tartu, 2007, 112 p.
68. **Aime Lust.** Charge state of dopants and ordered clusters formation in CaF_2:Mn and CaF_2:Eu luminophors. Tartu, 2007, 100 p.