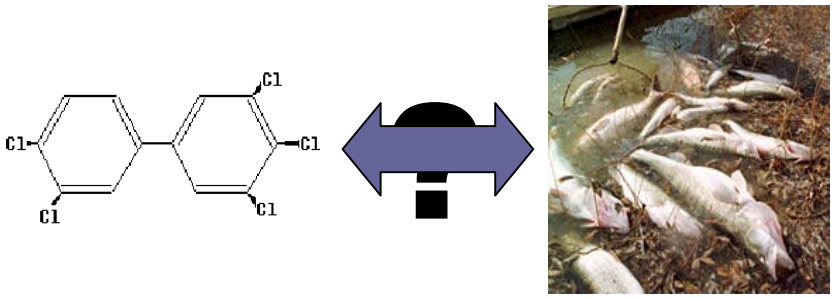
IVL report
IVL Swedish Environmental Research Institute

Estimating environmentally important properties of chemicals from the chemical structure



Erik Furusjö, Magnus Andersson, Magnus Rahmberg, Anders Svenson

B1517
Mars 2003

IVL

Organisation/Organization

IVL Svenska Miljöinstitutet AB
IVL Swedish Environmental Research Institute Ltd.

Adress/address

Box 21060
100 31 Stockholm

Telefonnr/Telephone

08-598 563 00

RAPPORTSAMMANFATTNING
Report Summary

Projekttitel/Project title

Verktyg för prognosticering av kemikaliers bioackumulerbarhet, toxicitet samt persistens

Anslagsgivare för projektet/
Project sponsor

CFs Miljöfond, SIVL

Rapportförfattare/author

Erik Furusjö, Magnus Andersson, Magnus Rahmberg, Anders Svenson

Rapportens titel och undertitel/Title and subtitle of the report

Estimating environmentally important properties of chemicals from the chemical structure

Sammanfattning/Summary

The development of models to predict important environmental properties is easily recognised in the light of the great number of existing chemicals that still need to be characterised. To meet the needs for testing new chemicals such models may also be useful. Here new quantitative structure-activity relationship (QSAR) models are presented to predict acute and subacute aquatic toxicity to a green alga (*Pseudokirschneriella subcapitata*), a crustacea (*Daphnia magna*), two fish species (*Lepomis macrochirus*, *Leuciscus idus*) and a bacterial bioluminescence inhibition test (Microtox3). The toxicity is predicted from more than 1400 molecular descriptors using the multivariate statistical method partial least squares (PLS) regression. The models are based on descriptors calculated from the chemical structure only and can be applied to substances that have not yet been isolated or synthesised.

QSAR models were obtained for which the standard prediction errors in logarithmic units correspond to the following concentration factors: Microtox 15 min bioluminescence inhibition EC_{50} – a factor 3.4; green alga 96 h growth rate inhibition EC_{50} – a factor 2.8; *Daphnia magna* 48 h immobilisation EC_{50} – a factor 2.3; *Lepomis macrochirus* 96 h toxicity LC_{50} – a factor 2.4; *Leuciscus idus* 96 h toxicity LC_{50} – a factor 3.5

In addition to development of prognosis models, the aim of this project was to develop methodology to obtain more reliable QSAR model predictions of toxicity. Two methodologies that are very important in this respect are systematic selection of the training set by statistical molecular design (SMD) and outlier detection. Partial least squares (PLS) modelling provides unique diagnostic tools when the model is used to predict the toxicity of new substances. Using these, it can be detected if the model does not cover the substance that the model is applied to, *i.e.* if the substance is a model outlier and the prediction is likely to be inaccurate. It is shown that reliable automatic outlier detection with a high efficiency can be obtained. This is a huge advantage for routine use of QSAR models and a leap forward towards reliable QSAR estimates of substance properties without requiring expert knowledge by the user.

Nyckelord samt ev. anknytning till geografiskt område eller näringsgren /Keywords

Quantitative structure activity relationships, QSAR, SAR, chemical structure, aquatic toxicity, screening new chemicals, multivariate data analysis, MVA, partial least squares, PLS, statistical molecular design, SMD, outlier detection, robust models

Kvantitativa struktur-aktivitetssamband, QSAR, SAR, kemisk struktur, akvatisk toxicitet, kemikalier, multivariat dataanalys, MVA, partial least squares, PLS, statistical molecular design, SMD, avvikare, robusta modeller

Bibliografiska uppgifter/Bibliographic data

IVL Rapport/report B1517

Beställningsadress för rapporten/Ordering address

IVL, Publikationsservice, Box 21060, S-100 31 Stockholm
fax: 08-598 563 90, e-mail: publicationservice@ivl.se eller via www.ivl.se/rapporter

Table of Contents

Abstract

The development of models to predict important environmental properties is easily recognised in the light of the great number of existing chemicals that still need to be characterised. To meet the needs for testing new chemicals such models may also be useful. Here new quantitative structure-activity relationship (QSAR) models are presented to predict acute and subacute aquatic toxicity to a green alga (*Pseudokirschneriella subcapitata*), a crustacea (*Daphnia magna*), two fish species (*Lepomis macrochirus*, *Leuciscus idus*) and a bacterial bioluminescence inhibition test (Microtox3). The toxicity is predicted from more than 1400 molecular descriptors using the multivariate statistical method partial least squares (PLS) regression. The models are based on descriptors calculated from the chemical structure only and can be applied to substances that have not yet been isolated or synthesised.

QSAR models were obtained for which the standard prediction errors in logarithmic units correspond to the following concentration factors:

- Microtox 15 min bioluminescence inhibition EC_{50} – a factor 3.4
- green alga 96 h growth rate inhibition EC_{50} – a factor 2.8
- *Daphnia magna* 48 h immobilisation EC_{50} – a factor 2.3
- *Lepomis macrochirus* 96 h toxicity LC_{50} – a factor 2.4
- *Leuciscus idus* 96 h toxicity LC_{50} – a factor 3.5

In addition to development of prognosis models, the aim of this project was to develop methodology to obtain more reliable QSAR model predictions of toxicity. Two methodologies that are very important in this respect are systematic selection of the training set by statistical molecular design (SMD) and outlier detection. Partial least squares (PLS) modelling provides unique diagnostic tools when the model is used to predict the toxicity of new substances. Using these, it can be detected if the model does not cover the substance that the model is applied to, *i.e.* if the substance is a model outlier and the prediction is likely to be inaccurate. It is shown that reliable automatic outlier detection with a high efficiency can be obtained. This is a huge advantage for routine use of QSAR models and a leap forward towards reliable QSAR estimates of substance properties without requiring expert knowledge by the user.

1 Introduction

The work described in this report is related to methods for estimating environmentally important properties, such as aquatic toxicity, from the structure of the chemical substance. The use of quantitative structure activity relationships (QSAR) is becoming established and accepted for estimating the ecotoxicity of many chemicals in the absence of results from actual toxicity tests[1]. However, there are some limitations associated with most QSAR models used today:

- Application relies on the availability of measured physicochemical parameters such as octanol/water partition coefficient, density, refractive index, boiling- and melting point, *etc.*
- A prediction from the model does not give any diagnostic information on whether or not the model is valid for this compound and, thus, what the quality of the prediction can be expected to be.
- The applicability of the models is in general limited to narrow classes of compounds

The aim of the research presented here is to investigate to what degree these limitations can be relaxed and what modelling methods and molecular descriptor are best in this respect.

The first limitation is serious, since it is often of great interest to assess the environmental properties of a substance that has not been isolated in a laboratory. Further, laboratory tests, even simple tests like solubility and partition coefficients are time consuming and expensive even when the substance has been isolated. Thus, our work has focused completely on models that are based on the structure of the substance alone, *i.e.* without requiring access to the actual substance and any physicochemical measurements.

The lack of prediction diagnostics that indicates what the quality of the prediction can be expected can easily over-confidence in the value produced by the model. A value without uncertainty measure is often perceived as exact, although the opposite is usually more adequate. In our opinion, a value should be treated with scepticism in the absence of an uncertainty estimate of some kind.

[1] See *e.g.* the homepage of the ECOSAR software under the USEPA New chemicals program, http://www.epa.gov/oppt/newchems/21ecosar.htm. Please note the discussion about these models in chapter 6 of this report.

Multivariate modelling methods based on latent variables, such as principal component analysis (PCA) and partial least squares (PLS) can provide prediction diagnostics with each prediction due to the fact that the covariance structure in the descriptor set is modelled. Although PLS has been used in several QSAR studies [Giraud *et al.* 2000, Shi *et al.* 2001, Tong *et al.* 1998, Eriksson *et al.* 2000], this important feature of the algorithm is often neglected.

The third limitation listed above, that the models are valid only for a narrow class of compounds, is probably the most difficult to solve. The reason for this is that different groups of substances can act with different mechanisms, which may be difficult to capture in a single model, especially for more specific and complex responses. The research presented here deals primarily with more general responses like aquatic toxicity. In order to reach the goal of general applicability, the work has been focused on models covering a wide range of chemicals.

2 Toxicity

The toxicity is an important property used in risk assessment and classification of substances. Acute toxicity, usually synonymous to lethality, is characterised by short-term exposure in relation to the life cycle of the organism. Long-term exposure, usually to lower doses may cause chronic effects. Effects on reproduction and exposure to more than one life cycle represent such effects.

In this study the bioluminescence inhibition of a marine bacterium kept at non-reproducing conditions (Microtox), the *Daphnia magna* 48 h immobilisation test and the 96 h fish lethality tests all represent acute toxic effects. The alga growth rate inhibition test, however, could be considered at least as a sub-acute or sub-chronic test although the duration was only 96 h. Several life cycles pass within this time and effects on the reproduction may be tested.

3 Theory

3.1 Molecular descriptors

In the scope of the investigation presented in this report, the purpose of molecular descriptors is to be the basis of models describing some aspect or aspects of the behaviour of chemical substances. Some general requirements that need to be fulfilled in order to make this possible are:

- The descriptors should contain relevant information for the purpose of the modelling, *i.e.* the aspect of the behaviour modelled. This means that the descriptor should allow for, and take into account, flexibility in the chemical structure if this is necessary to capture the behaviour of the substance.

- Most modelling methods require that the size of the descriptor set is independent of the size of the molecule.

Molecular descriptors can be classified by origin into measured and calculated descriptors. The major difference from an application point-of-view is that the chemical substance in question is required in order to obtain a measured descriptor while the calculated descriptors can be obtained for substances that cannot be isolated or have not yet been synthesised.

Andersson *et al*. [2000] have compared the information content in measured physicochemical and some calculated descriptors. Their results show that the descriptor sets contain similar information for the data sets investigated. The aim of the work presented in this report is to forecast environmental properties of large sets of new chemical substances. Measured properties are frequently not available for such sets and the aim is often to prioritise the substances for tests. Thus, the work presented is focused on the use of calculated descriptors and measured descriptors are discussed only very briefly below.

The distinction between measured and calculated descriptors is only one of many distinctions that can be made. Other possible classifications are global and local (depending on if the descriptor describes a property of the whole or a part of the molecule), static and dynamic (depending on whether dynamics of *e.g.* conformational changes are considered) as well as relative and absolute [Wehrens *et al*. 1999].

3.1.1 Measured descriptors

Undoubtedly, the single most important descriptor used in QSAR is hydrophobicity, which is usually measured as the logarithm of the octanol/water partition coefficient, log K_{OW}.

Other examples of useful measured descriptors include [Andersson *et al*. 2000, Livingstone 2000]:

- solubilities in different solvents
- boiling, melting and flash points
- spectroscopic properties such as NMR shifts or IR/Raman stretching frequencies
- molecular volume and density
- specific refraction and molecular refractivity

It is not difficult to understand why properties like log K_{OW} and solubility are important since they reflect the way the substance is distributed within an organism, which is of course important for its biological activity.

There are numerous methods for estimating log K_{OW} from the chemical structure based on different algorithms. Frequently such estimates are used as a descriptor for further QSAR modelling if experimental log K_{OW} values are not available. In such cases, the descriptor is not measured but often it is still denoted measured since both measured and calculated values of log K_{OW} are used as basis for the same model and no distinction is made between them.

3.1.2 Calculated descriptors

In order to relate chemical structure to biological activity or other molecular properties, it is necessary to describe the chemical structure numerically in some manner. A *calculated molecular descriptor* is a number extracted by a well-defined algorithm from a structural representation of the molecule. The descriptors are *defined by the algorithm* used for the calculation. Often, the chemical/physical interpretation of this number is not straightforward. However, this does not mean that the descriptor does not contain useful information about the properties of the molecule. We quote professor Roberto Todeschini of the Chemometrics and QSAR research group, Dept. of Environmental Sciences, University of Milano-Bicocca, Italy: "There is good reason to believe that often our difficulties in attributing a meaning to this number lie ultimately in the lack of deeper chemical theories and higher level languages and not from esoteric approaches to the descriptor definition." [web site http://www.disat.unimib.it/chm].

Numerous types of descriptors have been developed to numerically describe chemical structures. They can be coarsely classified into the groups 0D, 1D, 2D, 3D and other. These groups are briefly reviewed below.

3.1.2.1 0D descriptors

0D descriptors are constitutional in character and independent of molecular connectivity and conformations. Typical examples are atom and bond type counts, molecular weight and sum of atomic van der Waals volumes.

This type of descriptors cannot distinguish most molecular isomers and similar molecules, *e.g.* m-nitrophenol from p-nitrophenol.

3.1.2.2 1D descriptors

Counts of **functional groups** and **atom-centred fragments**, *i.e.* fractions of a molecule involving a few atoms, are often termed 1D molecular descriptors.

Molecular holograms. Holographic QSAR (HQSAR) is a recently developed technique that uses molecular holograms as descriptors [Burden, Winkler 1999]. We have chosen to classify this type of descriptors as 1D since it is based on structure fragments similar to the other 1D descriptors. The calculation procedure is roughly as follows: the molecule is divided into fragments of a number of atoms. Typically, a range like 3 to 8 atoms per fragment excluding hydrogen is used. Each fragment is mapped to an integer number. The integers are arranged in a number of bins (similar to a histogram) and the descriptors are the number of fragments in each bin. Typically, the number of bins used is in the range 20-400. HQSAR descriptors have not been used to obtain the results in this report.

3.1.2.3 2D descriptors

2D descriptors are dependent on the constitution and connectivity of the molecule but independent of conformation. Thus, 2D descriptors can be calculated from a 2D-structure representation of the molecule, *e.g.* a structure formula of an organic molecule.

2D autocorrelations. An autocorrelation function of the form $A(d) = sum_{ij}(p_i p_j)$ can be used to encode the topology of a molecular graph. p_i and p_j represent the values of an atomic property at atoms *i* and *j* , respectively, and d is the topological distance between the two atoms measured in bonds along the shortest path. The function has the useful property that no matter how large and complex the molecule, it can be encoded in a fixed length vector of small rank. Typically, only path lengths of 2 to 8 are considered. Atomic properties include *e.g.* atomic mass, volume, polarisability and electronegativity.

BCUT descriptors are calculated as the eigenvalues of the so-called adjacency matrix with the diagonal elements weighted by atomic masses. The adjacency matrix is a square matrix with each row/column corresponding to one atom. The *ij* element is 0 if atoms *i* and *j* are not connected, 1 if they are connected by a single bond, √2 if they are connected by a double bond *etc.*

Galvez topological charge indices are similar to the BCUT descriptors but the diagonal elements of the adjacency matrix are weighted by atomic charges instead of atomic weights.

Molecular walk counts. Counts walks and self-returning walks in the molecule of different length.

Various topological descriptors. A diverse set of descriptors, *e.g.* Wiener type indices and connectivity indices can be calculated from the 2D molecular structure.

3.1.2.4 3D descriptors

The 0D, 1D and 2D descriptors discussed above are independent of the 3D geometry of the molecule. It is reasonable to believe that the 3D structure of a molecule has a large influence of the biological activity of the molecule. Thus, descriptors that contain information on 3D structure should be valuable for QSAR studies.

3D descriptors are calculated from the 3D structure of the molecule, *i.e.* they are dependent on the conformation, including bond angles, interatomic distances *etc.* Since these properties are not available for any given chemical substance, some type of geometry optimisation must be included in the modelling (and prediction) process if a generally applicable method is sought.

A relatively simple and fast method that is applicable for small as well as large molecules is to optimise the geometry of the molecule by molecular mechanics. All structures used for modelling in this work were optimised by this method. Various force fields can be used but the MM+ force field is a versatile force field that suits the aim of general applicability of the results.

The geometry optimisation is performed by a local optimisation algorithm, which means that it may converge to different local minima depending on the initial geometry. One approach is to perform optimisations starting with a number of different conformations and choosing the 3D structure with the lowest energy. Chemical knowledge can be used to start with reasonable conformations, which makes the probability of reaching the global energy minimum relatively large, but there is no guarantee that the global optimum is reached, especially for very large molecules.

It should be noted that it is by no means certain that it is the conformation lowest in energy that is active in a biological system. This is a drawback of using the geometry dependent 3D descriptors as is done in this work. There are some methods that take the possibility of different active conformations into account, *e.g.* the CoRePa method discussed below. The conformation problems do not apply to 0D, 1D and 2D descriptors since these are conformation independent.

Other possible geometry optimisation methods include, semi-empirical (*e.g.* AM1) geometry optimisation and quantum mechanical methods, but these require substantially more computing power and are less suited for large molecules for this reason. In addition, they require more advanced and expensive software that may not be as widely available.

Randic molecular profiles characterise molecular shape in the form of a shape profile (a series of numbers) [Randic 1995, Randic, Razinger 1995].

Radial distribution functions (RDF) contains information about the interatomic distances in a molecule, unweighted or weighted by different atomic properties such as atomic mass, electronegativity, van der Waals volume and atomic polarisability [Hemmer *et al*. 1999].

3D-MoRSE descriptors reflect the three-dimensional distribution of different properties in the molecule. The transformation is derived from calculations used when determining molecular structure from electron diffraction measurements. The descriptors are obtained by summing products of atomic properties (mass, electronegativity, polarisability) weighted by different angular scattering functions and have been shown to preserve information about *e.g.* branching [Schur *et al*. 1996].

WHIM (weighted holistic invariant molecular) descriptors are based on principal component analysis of atomic co-ordinates with different weighting schemes. Weighting by atomic mass, electronegatvity, atomic polarisability, van der Waals volume, electrotoplogical state as well as unweighted analysis gives a total of 99 descriptors. The descriptors are of two types: directional (shape related) and non-directional (size related) [Livingstone 2000].

GETAWAY (Geometry Topology and Atom Weights Assembly) descriptors are calculated from a leverage matrix based on atomic co-ordinates called the molecular influence matrix. Weighted by different atomic properties such as atomic mass, electronegativity, van der Waals volume and atomic polarisability.

Various 3D geometrical descriptors based on molecular geometry, *e.g.* sums of interatomic geometrical distances.

Quantum mechanical/semi-empirical descriptors. As discussed above, geometry optimisation can be performed by quantum mechanical or semi-empirical methods. When such methods are applied to a molecule a description of the molecule is obtained that potentially contains large amounts of information about the properties of the molecule. A large number of descriptors can be extracted, *e.g.* energies of molecular orbitals (HOMO and LUMO), molecular polarisability, charge distribution, heat of formation, ionisation potential *etc.* We have not used quantum mechanical/semi-empirical descriptors to obtain the results presented below.

EVA (Eigenvalue) descriptors are vectors based on eigenvalues corresponding to a molecule's vibrational modes.

3.1.2.5 Other calculated descriptors

Estimated physical properties. As noted above, QSAR estimates of physical properties, most commonly the octanol/water partition coefficient log K_{OW} are used as descriptors for further QSAR modelling.

3.1.2.6 CoMFA and GRID

There exist specific QSAR descriptors that are based on a more physical model or understanding of the molecular interactions behind the biological response measured. Two methods that are closely related and based on superposition and alignment of molecular structures are Comparative molecular field analysis (CoMFA) and GRID [Livingstone 2000]. Both involve the use of a molecular probe and calculation of the interaction between the probe and the molecule that is being analysed. Interactions are measured at a (usually large) number of points in space defined by a grid placed around the molecular structure. PLS (see below) is usually used as the regression method in CoMFA.

CoMFA and GRID require that molecules be aligned relative to some common reference, *e.g.* the centre of mass. Aligning molecules with a similar structure is usually not that difficult, but a more diverse data set poses problems for all methods requiring alignment [Buydens *et al*. 1999]. CoMFA and GRID descriptors have not been used in the present work.

3.1.3 Software

A survey of available software for calculation of molecular descriptors was performed during the first stages of the project. The survey showed a variety of software packages of which most are strongly focused on drug discovery and drug design. Examples of software packages that compute molecular descriptors are Tsar, Dragon, AMPAC, MolconnZ and MOPAC. The software packages are more or less advanced; some of them only allow descriptor calculation but a few of them are also capable of QSAR modelling. An important aspect in the choice of a tool for calculation of descriptors is the licence fee for the software. Almost all of the software packages are licensed for a substantial fee, which means that they are not generally available to potential users of the QSAR models. This would limit the possible use of the models.

Thus, the criteria for our choice of software are calculation of a wide variety of relevant molecular descriptors at a reasonable price on a computer running Windows. Evaluation of these criteria led to a choice of the Dragon software. Dragon is a free software package developed by the Chemometrics and QSAR research group at Milan University, Italy. Dragon can be used to calculate a large number (1481) of molecular

descriptors from molecular structures saved in several different file formats, *e.g.* the standard format .mol and HyperChem .hin. The descriptors calculated by Dragon are discussed in Appendix A.

3.2 Modelling methods

This section describes some modelling methods that can be used to relate the chemical structure to environmental properties. The emphasis is on multivariate regression methods based on latent variables, since it is one of these methods, partial least squares (PLS), that has been used to obtain the results presented in this report. Other methods are discussed but less in-depth. Molecular docking algorithms are not considered at all.

Clustering of substances prior to regression modelling is often beneficial as reported by several authors, see *e.g.* [Suzuki *et al*. 2001]. Classification of substances prior to modelling has not been performed in this study, since the aim was to obtain models covering a broad range of chemicals in order to facilitate forecasting of environmental properties of large sets of new chemicals. This means that the predictive performance, measured as prediction errors for the environmental properties predicted by the models, is probably larger than what would be the case if clustering was used prior to regression modelling. On the other hand, the models are more generally applicable which is considered to be of greater importance.

3.2.1 Linear regression

The simplest forms of QSAR models are simple univariate linear regression models of the form

$$response = k_1 \times descriptor + k_0$$

These very simple models are of limited use since such a simple relationship is usually inadequate. An extension of this equation is

$$response = k_0 + \sum_{i=1}^{p} k_i \times descriptor^i$$

p is usually chosen as $p = 2$ or $p = 3$. The extension allows non-linear relations between the response and the single descriptor. However, a single descriptor is usually not sufficient to capture the behaviour of a substance, although successful applications have been reported for narrow groups of substances, usually with log K_{OW} as the descriptor.

Multiple linear regression (MLR) can be used to model the dependence of several descriptors according to the equation

$$response = k_0 + \sum_{i=1}^{p} k_i \times x_i$$

x_i is the ith descriptor. The number of descriptors, p, can vary widely from $p = 2$ to relatively large numbers. However, if many descriptors are used that contain similar information, *i.e.* are co-linear, problems with so-called variance inflation occurs, which means that the models become very sensitive to small variations in the descriptors and that their predictive performance becomes poor. To solve this problem, different variable selection algorithms can be used to select a small set of variables with high information content. Another approach is to use multivariate projection methods, described in the next section, that handle, and even utilise, the co-linearity in the descriptor set.

3.2.2 Multivariate projection methods

Typical examples of multivariate projection methods are principal component analysis (PCA) and partial least squares (PLS). Sometimes this type of methods is denoted multivariate data analysis (MVA) methods, which is a rather non-descriptive name but nevertheless adopted here due to convention. More informative names are multivariate projection methods or latent variable methods.

The fundamental MVA method is PCA. Only a very brief description of PCA is given here. More detailed introductory descriptions are references [Wold *et al*. 1987], [Martens, Naes 1989] and [Esbensen *et al*. 1996]. PCA decomposes a data matrix **X** (a table, in the current context the rows correspond to the substances while the columns correspond to descriptors) according to:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E}$$

PCA can be considered a co-ordinate transformation from the original variable space to a model hyper-plane of much lower dimensionality that captures the variance in the data in the most efficient way. The scores, denoted **t** or **T**, are the co-ordinates in the new co-ordinate system and thus describe the objects (here: chemical substances). The loadings, denoted **p** or **P**, describe the relation between the latent variables (principal components) that span the model space and original variables.

The matrix **E** in the equation above contains the residuals, *i.e.* the part of the data not captured by the model hyper-plane. Substances that do not conform to the "pattern" found among the other substances will be badly described by the model and thus have

large residuals. This can be caused by corrupted data or that the substance in question is different from the others, which may indicate that a QSAR model based on the rest of the compounds will not be valid.

The substantial dimensionality reduction achieved by applying PCA to molecular descriptor data sets leads to enhanced interpretation abilities which facilitate classification and clustering of substances. This is utilised in a methodology known as statistical molecular design (SMD), see the separate discussion below.

PCA is not a regression method and cannot be used for finding quantitative relationships between descriptors and responses. The most common multivariate regression method is PLS.

3.2.2.1 Partial least squares

PLS is a latent variable based regression method described in several references [Martens, Naes 1989, Esbensen *et al*. 1996, Geladi, Kowalski 1986]. PLS has several benefits compared to ordinary multiple linear regression:

- Co-linearity is handled in a natural way and even utilised to find a robust estimate of the data structure. This means that variable selection methods are of less importance than in MLR.

- The latent variable approach means that outlier diagnostics can be obtained both for training and prediction substances.

The prediction outlier diagnostics obtained has no counterpart in MLR or the non-linear regression methods, such as artificial neural networks (ANN) discussed below, and are the greatest advantage of latent variable regression methods according to us. For a new sample it is possible to calculate a probability that the sample belongs to the sample population the model was estimated from and thus that the model is likely to yield a valid prediction. It should be noted that, as shown below, it is quite possible for a model to yield good predictions although the sample is classified as not belonging to the model. The opposite, that the sample is classified as belonging to the model and poorly predicted is uncommon. This is the behaviour required for risk assessment of substances, since a false prediction that is not detected may lead to a substance being erroneously classified as likely to be non-toxic and thus that further testing of the substance is given low priority.

3.2.2.2 Hierarchical modelling

During recent years, hierarchical multivariate modelling methods has undergone rapid development and several successful applications within process modelling have been

published [Westerhuis *et al*. 1998, Qin *et al*. 2001, Westerhuis, Coenegracht, 1997]. A hierarchical model structure can be beneficial when several distinct and separate blocks of data are used for modelling. In process modelling this usually corresponds to process data from different process sections (reactors, coolers, distillation columns *etc.*) that influence product properties in different ways. A separate model (*e.g.* a PLS model) is built for each block. The scores calculated from each of these block models are then used as an input to a top-level hierarchical model. It is out of the scope of this report to go into any detail regarding the theory for hierarchical multivariate modelling. More details are given in Westerhuis *et al*. [1998]. The benefit of using a hierarchical model structure is that the complexity of the individual models is decreased. Still, the interaction between different blocks can be modelled and combination of information between blocks can still be achieved in the top-level model.

In QSAR modelling the different groups of descriptors reflect different aspects of the substance and can be treated as blocks in a hierarchical model structure. Interpretation of the top-level model gives insights into which descriptor groups contain most information about the biological response and how the information is combined.

3.2.3 Non-linear methods

Non-linear methods are not applied in the work presented in this report but several investigations presented in the literature indicate that they give superior performance to linear methods in some cases. The most common group of methods is artificial neural networks (ANN) that exist in a variety of different forms.

It should be noted that ANNs have some drawbacks that sometimes are neglected: the large number of parameters means that a large amount of training data is needed and that validation must be performed rigorously in order to avoid over-fitting that leads to poor model performance. Further, prediction diagnostics are not obtained from ANN models. One needs to ensure in some other way, independently from the ANN model, that the model is valid for the substance in question or, which is common practice, predict and pray.

3.2.4 Common Reactivity Pattern

The modelling methods discussed above are general empirical regression methods that can in principle be applied to any regression problem and that can be used for QSAR modelling when applied to molecular descriptors and molecular properties of substances.

Another approach is Common Reactivity Pattern (CoRePa) [Mekenyan *et al*. 1997], which accounts for conformer flexibility in the structures. A brief description of

CoRePA is as follows. A set of chemicals that are most (or sometimes least) active, *i.e.* that exceed (fall short of) a threshold for the biological activity in question, is selected. Then, a set of parameters that are hypothesised to be potentially important for the biological activity are identified. These are evaluated for a distribution of conformers for each compound to give a distribution of the parameter per substance. All distributions for a certain parameter are superimposed and common regions are identified. The common regions identified (*i.e.* for different parameters) constitute the common reactivity pattern.

3.2.5 Model validation and model accuracy measures

It is important to be able to measure model performance for different reasons, including ranking of models and estimating the reliability of predictions, when the model is used on new substances. An accuracy measure is essential in order to be able to trust and use a model prediction.

The data used to estimate the model, the training set, cannot be used to reliably estimate model performance. Two validation methods are commonly used:

- **Cross-validation**. In cross-validation the model is estimated a number of times. In each round, a part of the training substances are kept out. The toxicities of these substances are then predicted by the model and compared to the known (reference) values. The procedure is repeated until all samples have been kept out exactly once and cross-validation prediction errors have been obtained for all substances.

- **Test set validation**. Test set validation is used when there are enough data available to exclude some of it, called the test set, from the model estimation and use it solely for validation. The model is estimated from the remaining data, the training set.

Test set validation is the most reliable method to estimate the true model performance, since if the test set is adequately selected, it is exactly equal to future model use; substances that are completely unknown to the model are predicted. Cross-validation is a reasonable substitute method if the amount of data is limited but the reliability is lower; slightly over-optimistic results are usually obtained.

For multivariate modelling methods and some other modelling methods there is a further complication. Validation is usually used both for model complexity selection (*e.g.* the number of PLS components in PLS regression) and for estimation of model performance. Since the model complexity selection is usually based on a prediction error criterion this can lead to so-called selection bias, which means that over-optimistic estimates of model performance are obtained. One way to deal with this problem that has been used in this work is to use cross-validation to select model complexity and test

set validation to estimate model performance. This means that selection bias is avoided and that very reliable estimates of model performance can be obtained.

Model performance can be measured by different metrics:

- **R^2** (or R^2Y) is the part of the variance explained in the training data, *i.e.* without validation. Thus, it does not give information about model performance for new substances. If R^2 is 1 the model explains the data perfectly, if R^2 is zero it is as good to guess a random number as to use the model.

- **Q^2** is the validation counterpart to R^2. It measures the part of the variance explained in the validation data. Q^2 can be calculated both for cross-validation, in which case it is sometimes denoted Q^2_{CV}, and for test set validation.

- **RMSEP** (root mean square error of prediction) is a measure of the prediction error and has the same unit as the response predicted by the model. It is calculated similarly to a standard deviation and can be used roughly as a standard deviation of predictions. In the formula, y is the reference value and $\hat{y}$ is the predicted value.

$$RMSEP = \sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{n}}$$

- **RMSECV** (root mean square error of cross-validation) the cross-validation version of RMSEP, *i.e.* corresponding to Q^2_{CV}.

- **RMSEE** (root mean square error of estimation) the non-validated version of RMSEP, *i.e.* corresponding to R^2.

3.2.6 Outliers in QSAR models

An outlier in a QSAR model is a substance that is in some way different from the rest (majority) of the substances used to estimate the QSAR model and for which the model is not valid. The difference can be caused by different features in the chemical structure, which is closely related to the discussion above on classification of substances prior to modelling.

The common explanation of a model outlier is that it is badly predicted (has a large y residual) but this is a somewhat limited definition since a good prediction may be purely due to chance, although the substance class in question is not at all present in the training data. In multivariate statistics, it is common to define three types of outliers:

- X/Y outliers are outliers in the normal meaning, *i.e.* substances for which the relationship between the descriptors (X variables) and the environmental property (Y variable) is not valid, *e.g.* due to different toxicity mechanisms.

- X outliers. In short, a substance is an X outlier if the molecular descriptors for this substance do not conform to the "pattern" (covariance structure) in the (rest of the) training data. A different pattern in the descriptors indicates that the substance is different from the training data and thus that the prediction is likely to be inaccurate, *i.e.* a substance that is an X outlier is likely to be an X/Y outlier as well.

- Y outliers are only defined for training or test samples. They are substances for which the reference value of the response is bad for some reason.

It is important to note that outliers can be present both during training (model estimation) and model use (prediction). Naturally, since no Y value is normally available during prediction (this is why the model is used to estimate the property in question), Y cannot be present and X/Y outliers cannot be detected directly.

However, if multivariate prediction methods are used X outliers can be detected during prediction from the X residuals of the projection (also known as: distance to model in X space). This is a significant advantage of multivariate projection methods, like PLS, that facilitates automatic detection of outliers during the use of a QSAR model. This possibility is a property of the PLS method and not of the descriptors used. Thus, the advantage is present regardless of the molecular descriptors used although the success is of course dependent on the information content in the descriptors.

Lipnick [1991] discussed possible reasons for outliers (X/Y only) in QSAR models and related them to different mechanisms of action.

3.2.7 Statistical Molecular Design (SMD)

Statistical molecular design (SMD) is a method introduced by researchers in Umeå, Sweden [Eriksson, Johansson 1996, Andersson *et al*. 2000, Eriksson *et al*. 2000]. The purpose of SMD is to apply experimental design methodology in QSAR modelling. The goal of experimental design is to select a training set for modelling that contains maximal information given the number of experiments that can be performed. In QSAR, the experiments correspond to substances but their properties (molecular descriptors) cannot be designed since they are impossible to control independently in practically all cases.

SMD uses a large number of candidate structures for which the response (y) variable does not need to be measured or known. Molecular descriptors are calculated or measured for all candidate substances and PCA is performed on the data set. The

principal components that are combinations of the molecular properties are referred to as the principal properties (PP) of the data set, since they are the combinations that explain the variation among the molecules in an optimal way.

The design is then performed with respect to the principal properties by selecting a subset of substances that are most efficient in spanning the substance (or PCA model) space and thus are the best selection of training set for a QSAR model. The selection can be done manually from the score plots if the number of principal components (properties) is three to four or less. An algorithm based on D-optimality is necessary when a higher number of PCs are used. Such an algorithm can be used for low-dimensional models as well, but it is often sufficient to select samples manually.

An excellent illustration of the usefulness of SMD and multivariate techniques for exploration of principal properties can be found in a recent publication [Giraud *et al*. 2000].

4 Methods

4.1 Toxicity data

4.1.1 Microtox toxicity

The toxicity of various substances to the marine bacterium *Vibrio fischerii* was taken from literature [Kaiser, Palabrica 1991]. The EC_{50} (in mmoles/L) of the bacterial luminescence inhibition at 15 min exposure was selected as the toxic endpoint and transformed to the log of the inverse of the millimolar concentration to yield pEC_{50} values.

The Microtox toxicity of ethylene diamine was also tested experimentally following essentially the procedure of the manufacturer's manual [Azur Environmental, Carlsbad, USA (www.azurenv.com), Svenson 1993]. Before testing, the solution of the toxicant was adjusted to pH 7.3 ± 0.05. The procedure involving a combined duplicate of tests was repeated three times to generate a log-normal average value of the EC_{50}. Ethylene diamine obtained from Merck, freshly distilled prior to use, was a kind gift from Fredrik Rahm at the Department of Organic Chemistry, Royal Institute of Technology in Stockholm.

4.1.2 Alga toxicity

The unicellular green alga *Pseudokirschneriella subcapitata* was chosen as the organism for prediction of alga toxicity. The species, also known by its synonyms

Selenastrum capricornutum and *Raphidocelis subcapitata*, is the most widely used freshwater organism for test of alga toxicity. The inhibition in growth rate was selected as the toxic endpoint [Nyholm, Källqvist 1989]. The 96 h EC_{50} values were collected for a set of substances from published sources [Alexander *et al*. 1988, Blaylock *et al*. 1985, Calamari *et al*. 1979, 1980, 1983, Draper, Brewer 1979, Eloranta 1982, Galassi, Vighi 1981, Galassi *et al*. 1988, IUCLID database, 2000 Kuivasniemi *et al*. 1985, Macri, Sbardella 1984, Shigeoka *et al*. 1988]. Before use, the data was transformed as the logarithms of the inverse EC_{50} in mmoles/L. EC_{50} values for some substances were calculated by non-linear regression of the logarithmic rate from growth data given in literature [Adams, Dobbs 1984].

4.1.3 *Daphnia* toxicity

The toxicity of various substances to *Daphnia magna* exposed for 48 h at specified conditions was used to derive a prognosis model for a crustacean. Toxicity data were collected from a published source [Devillers *et al*. 1987] and used as logarithms of the inverse in mmoles/L.

4.1.4 Fish toxicity

Data from two fish species were selected to model fish toxicity. The lethal toxicity to *Leuciscus idus* was taken from Juncke, Lüdemann [1978], *i.e.* data dertermined in one of the two laboratories reported in the literature source, and *Lepomis macrochirus* from Buccafusco *et al.* [1981]. Data represents LC_{50} at 96 h exposure and the values were transformed as the logarithms of the inversed millimolar concentrations.

4.2 Descriptor calculation and QSAR modelling

The following procedure was used to obtain the results presented in the following section.

For each substance:

- The structure of the substance was obtained from Internet databases, *e.g.* ChemIDplus[2], or, if not available, drawn manually.
- The structure was imported into a molecular modelling software, HyperChem[3], and the minimum energy conformation was determined by molecular mechanics with the MM+ force field. Different initial conformations were used in order to decrease the risk of finding local energy minima. The optimised structure was saved.

[2] http://chem.sis.nlm.nih.gov/chemidplus/cmplxqry.html
[3] HyperCube Inc., http://www.hyper.com

For all substances belonging to a data set:

- From the saved structures, 1481 molecular descriptors were calculated for each substance by the Dragon software[4] and the results were saved.
- The descriptor file was extended with the biological response variable.
- The data set was imported into the multivariate modelling software SIMCA P-10[5] for modelling.

Auto-scaling (also known as unit variance scaling) have been used throughout this work, since the variables are on different scales and no a priori information about variable importance was available that could motivate other scaling schemes.

PCA was used to detect trends and groupings in the data. In the cases SMD was used, samples were selected manually from the score plots as discussed in chapter 5 below for each model.

Regression was performed by the PLS method. All regression models were validated by both cross-validation and a separate test set. Cross-validation was used to select model complexity, and on some occasions to perform variable selection, while the test set was used solely for estimating prediction error and judging the ability of the model to detect outliers in the prediction stage. This procedure gives a reliable and objective estimate of model performance.

An unusually large proportion, usually about 50 % of the available data has been used for model testing. This is motivated by the fact that the focus of the work is to develop methodology to obtain reliable QSAR models. The only way to estimate reliability is by the test set and the estimate is better the more samples are used for this purpose.

5 Results

The research presented in this report was performed with multiple aims. The discussion in this chapter and chapter 6 is meant to reflect all of these.

- To develop accurate and useful QSAR models for toxicity of chemical substances.
- To develop and evaluate methodology for increasing the reliability of QSAR predictions
- To investigate the information content and usefulness of different groups and types of descriptors

[4] http://www.disat.unimib.it/chm/Dragon.htm
[5] Umetrics Inc., http://www.umetrics.com

Before viewing the modelling results it can be interesting to note the distribution and span of the toxicity values in the data sets used for modelling. These are shown in Figure 1 below and values are also given in Appendix B. It can be noted that the span for the fish species and for green alga toxicity was significantly shorter than the span for the other two.

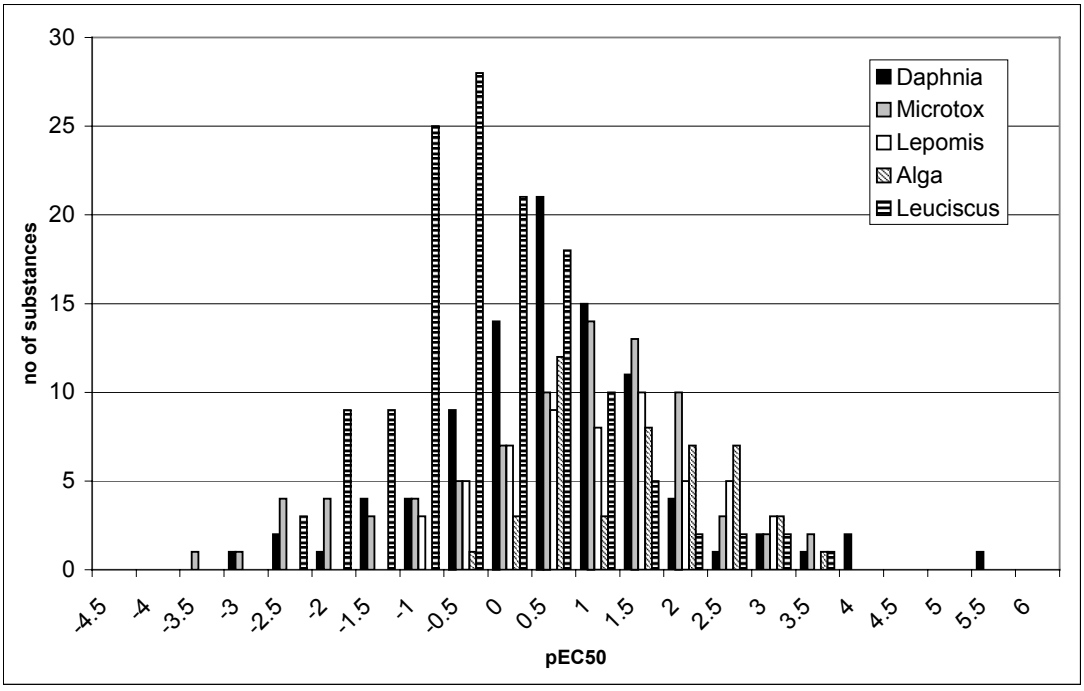


Figure 1. Distribution of reference values used for QSAR modelling.

5.1 Microtox prognosis model

83 substances with reference values between pEC_{50}=-3.3 and pEC_{50}=3.8 were available for modelling, which means that the substances span 7 orders of magnitude in the concentration domain.

A preliminary evaluation of Microtox QSAR models showed that ethylene diamine appeared as a Y outlier; it had a very low predicted pEC_{50} (about –2 to –3) in all models while the reference value obtained from the literature was 0.47. Therefore a new test of the compound was conducted. The average EC_{50} at 15 min exposure of ethylene diamine was 17.9 g/L (limits of one standard deviation 17.6-18.2 g/L). This corresponds to pEC_{50} = -2.47, which agrees very well with the value predicted by the preliminary models but is considerably lower than the published value. The higher toxic value probably depended on insufficient pH control. Now, after testing, pH 7.3 was recorded, indicating that pH was maintained at a constant and optimal level throughout the 15 min exposure period. Dissolving the compound in an aqueous solution with low buffering

capacity changes pH, which in turn affects the toxic behaviour. Therefore a careful pH control is extremely important. Alkaline condition itself will inhibit the luminescence, and the amine will probably have a higher toxicity due to a higher proportion as the dissociated, uncharged species, as recently was shown in alga toxicity for ammonia [Källqvist, Svenson 2003].

The ability to detect erroneous literature values, which was confirmed by new experiments, shows the power of QSAR modelling. The new reference value Microtox $pEC_{50} = -2.47$ was used in all further modelling. A list of all Microtox toxicity values used for modelling and validation is shown in Appendix B.

5.1.1 Prediction outlier detection

A number of QSAR models for Microtox toxicity were developed based on different sets of descriptors. These were used to both evaluate the performance of the descriptors and different prediction outlier detection methods. Evaluation of criteria for detection of outliers during model application (*i.e.* prediction) is discussed in this section. The actual modelling results are discussed in the next section.

Outlier detection during prediction, *i.e.* to detect substances that do not fit in the model and thus have a high risk of being poorly predicted, is very important. Since the aim of the prognosis model is typically screening of new substances and prioritising further testing, it is serious if substances are classified as false negatives. On the other hand, to predict false positives is less serious, since this will be revealed by the testing performed as a result of the QSAR prediction. However, also such malpredictions decrease the efficiency of the screening and prioritisation and should, naturally, be avoided if possible.

Outlier detection during prediction aims at completely avoiding grossly erroneous predictions. If the substance in question risks being badly predicted, this should be detected and the prediction should be considered unreliable and not used. Other methods for screening and prioritisation should then be used, *e.g.* other QSAR models or testing toxicity. In outlier detection, it is inefficient but not serious if reliable prediction is classified as unreliable, *i.e.* if a substance that is well predicted is classified as an outlier, since this will lead to prediction by other methods or toxicity testing of the substance. The opposite mistake, on the other hand, *i.e.* to classify a bad prediction as reliable, is serious. This should be kept in mind while reading the discussion in this section.

When PLS regression is used as the modelling method, as in this work, two measures can be considered when judging whether or not a new sample belongs to the model. The first is the distance to the model plane (also called residual magnitude) and the second is the distance between the model centre and the projection in the model plane. In the

SIMCA software, the distance to the model plane of a prediction is known as DModXPS (Distance to Model in X space for the Prediction Set), while also considering the distance in the model plane leads to the statistic DModXPS+. From theses distances and the corresponding distances in the training set, it is possible to calculate a probability that a (new) substance belongs to the model. These probabilities are known as PModXPS and PModXPS+, respectively, in the software.

In order to classify substances as falling within or outside the domain of the model, one must choose a significance level. Initial investigations with significance levels corresponding to 5 % or 1 % theoretical risk of erroneously classifying a valid prediction as an outlier showed that these levels gave a very large number of erroneous outlier indications. The reason is probably that the theoretical assumptions, *e.g.* normally distributed data, are not fulfilled. In such cases, it is common to use empirical significance levels in statistical tests.

Results from further investigations with both PModXPS and PModXPS+ at 0.5 % and 0.1 % risk levels are shown in Table 1 for 9 different PLS models based on different sets of descriptors. The RMSEE and RMSEP values in the second and third column are the root mean squared error of estimation for the training data and the root mean squared error of prediction for the full test set. In addition, the table shows the number of outlying substances in the test set according to each method and the RMSEP after these were removed from the test set.

RMSEE is expected to be significantly lower than RMSEP since it is calculated from predicting the same substance from which the model parameters were estimated. The RMSEP values are aggregated values for the whole prediction set but it is clear from plots of predicted versus measured toxicities that the high RMSEP values for some models are caused by one or a few substances being poorly predicted by the model, *i.e.* they are outliers. This is visualised in Figure 2 and even more clearly in Figure 3.

No differences were encountered between the PModXPS and PModXPS+ methods as shown in the table. Therefore, this is not further discussed but the following discussion is devoted to the choice of significance level and the reliability of the outlier detection method.

Table 1. Outlier detection in the Microtox toxicity data set.

Model[a]	RMSEE	RMSEP	PModXPS+ 0.5 % outliers	PModXPS+ 0.5 % RMSEP[b]	PModXPS 0.5 % outliers	PModXPS 0.5 % RMSEP[b]	PModXPS+ 0.1 % outliers	PModXPS+ 0.1 % RMSEP[b]	PModXPS 0.1 % outliers	PModXPS 0.1 % RMSEP[b]
PLS2	0.34	0.64	7	0.53	7	0.53	7	0.53	7	0.53
PLS3	0.46	0.73	7	0.65	7	0.65	5	0.64	5	0.64
PLS4	0.56	0.77	7	0.66	7	0.66	6	0.70	6	0.70
PLS5	0.42	1.06	11	0.62	11	0.62	7	0.58	7	0.58
PLS6	0.65	3.72	5	0.82	5	0.82	4	0.82	4	0.82
PLS7	0.52	0.98	7	0.66	7	0.66	7	0.66	7	0.66
PLS8	1.03	0.95	9	0.89	9	0.89	8	0.90	8	0.90
PLS9	0.50	1.47	6	0.74	6	0.74	5	0.74	5	0.74
PLS10	0.38	0.86	3	0.54	3	0.54	2	0.55	2	0.55

[a] For explanation of model notation see 5.1.2 below.

[b] RMSEP for the test set after removal of the outliers indicated by this method.

In Figure 2 and 3 the substances have different symbols according to their probability of belonging to the model according to PModXPS+. Substance marked with squares are outliers according at both the 0.1 % and 0.5 % levels, while the substances marked with diamonds in Figure 3 are outlier only at the 0.5 % level.

It is clear from Figure 2 that at least p,p-DDT (ppDDT) and carbon tetrachloride (ccl4) are poorly predicted by the model PLS2. However, 5 more substances are classified as outliers although they are quite well predicted: tetrachloroethene (teke), nitrilotriacetic acid (nta), methanol (meoh), dichloromethane (dkm), and dioxane (dioxan). There are two possible reasons:

1. They are outliers in the model that are reasonably well predicted 'by chance'. This should be the case for at least methanol and nitrilotriacetic acid. Methanol is an extreme sample with a toxicity value lower than any compound in the training set. Thus the model is extrapolated which is uncertain and the outlier classification is correct although the prediction happens to be correct in this case. Nitrilotriacetic acid with its large number of polar bonds in such a small molecule is different from substances in the training set.

2. They are falsely classified as outliers although they are similar to the compounds in the training set. This can be the case for dioxane that is not structurally dissimilar to substances in the training set.

For dichloromethane and tetrachloroethene it is questionable if they are outliers or not. There were relatively few smaller chlorinated compounds in the training data (chloroform, trichloroethene).

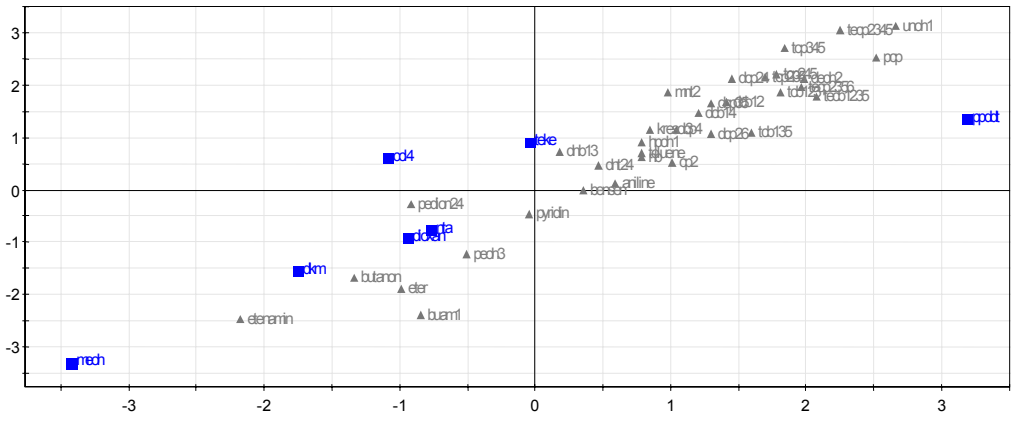


Figure 2. Measured versus predicted Microtox pEC50 values for the PLS2 model based on all descriptors. Substances marked with squares have a probability of less than 0.1 % of belonging to the model according to PModXPS+.

For the model PLS5 visualised in Figure 3 similar results were obtained at the 0.1 % level (blue substances):

- p,p-DDT, methanol, CCl_4 and tetrachloroethene were inaccurately predicted and this is detected by the outlier detection method.

- Diethylether, nitrilotriacetic acid and 1,3,5-trichlorobenzene were correctly predicted but nevertheless classified as outliers. For nitrilotriacetic acid, this should be considered a coincidence as discussed for the PLS2 model above. For the other two substances the classification is more questionable and they are probably falsely detected as outliers. Nevertheless, the RMSEP of the test set is decreased from 1.06 to 0.58 when the substances indicated as outliers were removed, which according to our experience was in reasonable relation to the RMSEE of 0.42.

The substances classified as outliers only at the 0.5 % level in model PLS5 (diamonds in Figure 3), dichloromethane, 4-chlorophenol, 3,5-dichlorophenol and 1,2,3-trichlorobenzene, were all predicted correctly by the model. At least the three aromatic compounds should not be outliers considering their structural similarity to the training set. Similar results were obtained for other models (not shown). It can be noted in Table 1 that although more outliers were detected at the 0.5 % level for several models, the RMSEP of the test set has not decreased significantly.

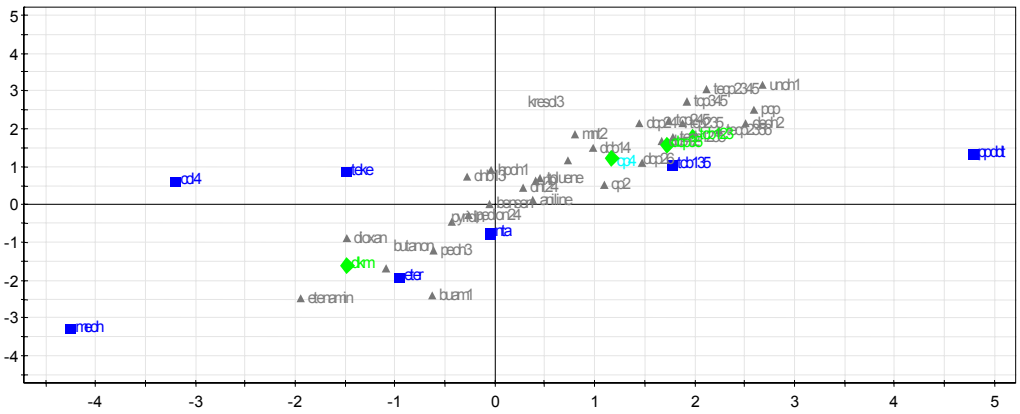


Figure 3. Measured versus predicted Microtox pEC50 values for the PLS5 model based on 2D autocorrelation descriptors. Substances marked with squares have a probability of less than 0.1 % of belonging to the model according to PModXPS+ and are classified as outliers. Substances marked with diamonds have probabilities between 0.1 %-0.5 %.

From inspection of predicted versus measured plots for all the other models investigated it was observed that not a single clear outlier was missed at the 0.1 % level. Similar results as those discussed above were obtained when evaluating outlier detection method for the models based on systematic selection of training set discussed in 5.1.3 (not shown).

To summarise, it can be concluded that outlier detection at the 0.1 % level is sufficiently safe and more efficient than at the 0.5 % level. In the rest of the report, outlier detection at the 0.1 % level with the PModXPS+ statistic has been used.

5.1.2 Random training set selection

A number of QSAR models for Microtox toxicity were developed based on different sets of descriptors. The complete data set collected consisted of 83 substances, see Appendix B. Initial modelling showed that 1-pentadecanol, which was the longest carbon chain in the data set, was difficult to fit into a model with the rest of the substances and, hence, it was removed. The remaining 82 substances were split non-systematically into a training set and a test set each comprising 41 substances. To leave 50 % of the substances in a test set is unusual but was motivated by the intention to develop methodology for reliable and robust QSAR predictions. The only way to objectively test the reliability of the models is to use a test set and the larger the test set the better the estimate of the degree of reliability. Information about the models developed is shown in Table 2. No outliers were removed from the training set for any of the models.

Table 2. Model information for PLS models developed for Microtox pEC_{50} with random selection of training set.

Model	Descriptors[a]	k[b]	R^2Y	Q^2Y[c]	RMSEE	RMSEP	Outliers (test set)	RMSEP[d]
PLS2	all	4	0.95	0.87	0.34	0.84	7	0.53
PLS3	Constitutional + topological + mol. walk counts	2	0.91	0.83	0.46	0.73	5	0.64
PLS4	BCUT	4	0.87	0.79	0.56	0.77	6	0.70
PLS5	2D autocorr.	4	0.93	0.71	0.42	1.06	7	0.58
PLS6	RDF	4	0.83	0.62	0.65	3.72	4	0.82
PLS7	3D-MoRSE	4	0.89	0.77	0.52	0.98	7	0.66
PLS8	WHIM	2	0.56	0.30	1.03	0.95	8	0.90
PLS9	GETAWAY	4	0.90	0.65	0.50	1.47	5	0.74
PLS10	PLS3-PLS9 scores[f]	1	0.94	0.92[e]	0.38	0.86	2	0.55

[a] See 3.1

[b] Number of PLS components

[c] Relative explained variance of the response determined by cross-validation. Seven cross-validation segments were used

[d] RMSEP of the test set after removal of the detected outliers

[e] Should not be compared with the other Q2Y values since the cross-validation is based on non-validated scores. RMSEP values from test set prediction are comparable, however.

[f] Top level hierarchical model based on the scores from PLS3-PLS9, see 3.2.2.2.

From the table, it can be noted that the number of outliers detected during prediction was between 2 and 7. It is not surprising that outliers were detected, since the training and test data were selected non-systematically. It can be expected that some of the test substances do not fall within the domain of the model, *cf.* the discussion about models based on systematic training set selection in 5.1.3 below. In addition, it was not surprising that the performance and the number of outliers detected are different in different models since they were based on different descriptors that reflect different molecular properties.

None of the individual descriptor groups (PLS3-PLS9) yielded models that were as good as models based on all data (PLS2 and PLS10), which indicates that a combination of information from different descriptor groups were needed. The best single descriptor group was 2D autocorrelations that include Broto-Moreau, Moran and Geary autocorrelations of lag 1-8.

The model based on all descriptors directly (PLS2) and the model based on all models but in a hierarchical model structure had the same performance within statistical uncertainty. The RMSEP values for test data were about 0.5 log-units, *i.e.* predictions were correct within a factor 3 on the concentration scale, which was considered to meet the needs of screening models. These results should be compared to those based on systematic training set selection in 5.1.3 below.

It is worth noting that the PLS2 model has 7 samples classified as outliers. As discussed above and shown in Figure 2, several of these were in reality predicted well by the model. In Figure 4, the predictions of the PLS10 model are shown. As seen in the figure and in Table 2, only two substances were classified as outliers. It is clear from the figure that these two substances were the two that were inaccurately predicted. Hence, the outlier detection performed well for this model. Predicted versus measured values after removing the two outliers, which illustrate the good model performance, are shown in Figure 5. Please note the discussion about outlier detection in hierarchical models in the *Lepomis* data in 5.4 below.

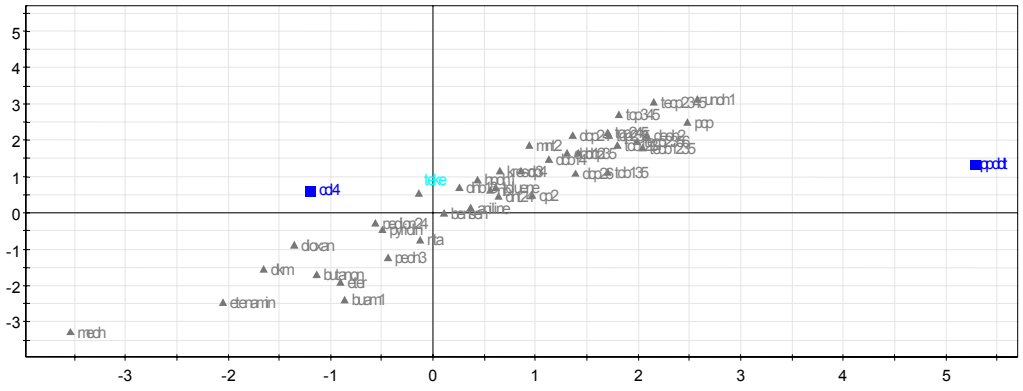


Figure 4. Measured versus predicted Microtox pEC50 values for the test set of the hierarchical top-level model (PLS10). Substances marked with squares are outliers.

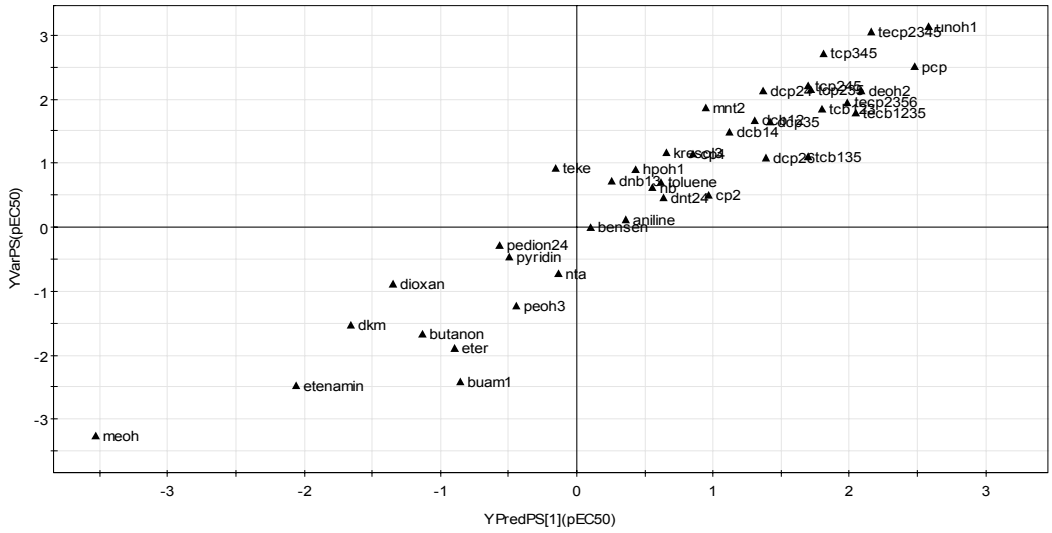


Figure 5. Measured versus predicted Microtox pEC50 values for the test set of the hierarchical top-level model (PLS10) after removal of two outliers (p,p-DDT and CCl_4). RMSEP = 0.55 pEC_{50} units.

5.1.3 Systematic training set selection

To further illustrate the importance of a representative training set, Microtox pEC_{50} prognosis models were developed based on a systematically selected training set. The approach used for training set selection was similar to statistical molecular design (SMD, see 3.2.7) but the selection was done graphically from the score plots rather than mathematically based on a D-optimality criterion, which is common in SMD.

A PCA model based on all descriptors was used as basis for the selection. To facilitate comparison to the models based on random training set selection, the same sizes of training and test set, 41 substances each, was used. The score plots used and the selected substances are shown in Figure 6

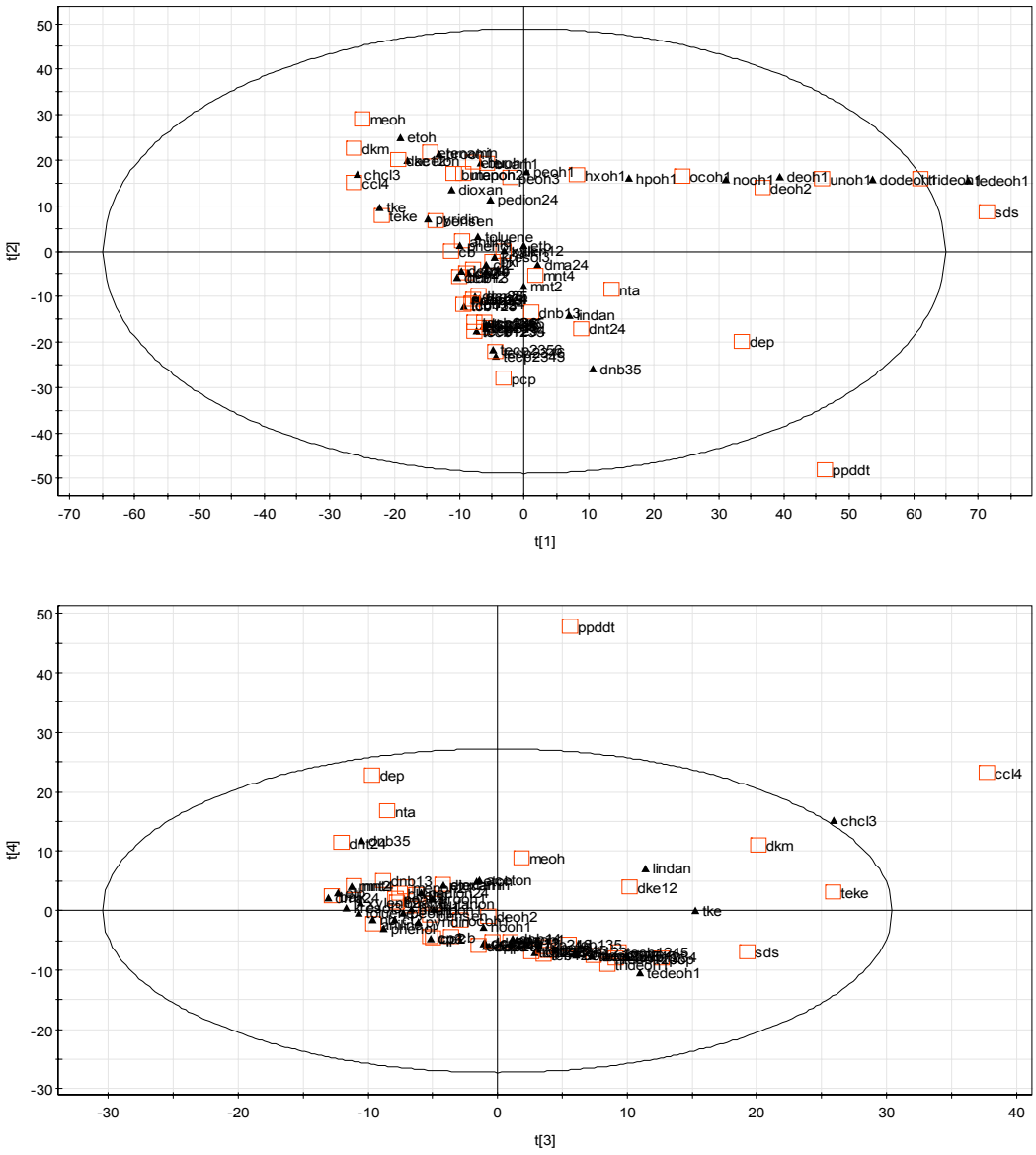


Figure 6. PCA score plots and systematic training set selection for Microtox pEC_{50} models. The training set (41 of 82 substances) is marked by squares.

Modelling results are shown in Table 3 and can be compared to those in Table 2. It is clear that, as expected, the number of detected outliers was much lower when systematic training set selection was used, since the selection was done to avoid extrapolation. The models based on all descriptors, PLS2 and PLS12, had approximately the same RMSEP for test data: 0.53 and 0.54 respectively, but the number of outliers detected for PLS2 was 7 and only 3 for PLS12. The same trend is visible in the results from the top-level hierarchical models, PLS10 and PLS20. The RMSEPs were 0.55 and 0.64 and the number of outliers 2 and 0, respectively.

Table 3. Model information for PLS models developed for Microtox pEC_{50} with systematic selection of training set.

Model	Descriptors[a]	*k*[b]	R^2Y	Q^2Y[c]	RMSEE	RMSEP	outliers	RMSEP[d]
PLS12	all	4	0.96	0.83	0.31	0.54	3	0.54
PLS13	Constitutional + topological + mol. walk counts	4	0.94	0.83	0.38	0.58	2	0.56
PLS14	BCUT	3	0.80	0.72	0.70	0.81	0	0.81
PLS15	2D autocorr.	1	0.64	0.53	0.92	0.96	0	0.96
PLS16	RDF	2	0.73	0.58	0.82	0.94	4	0.92
PLS17	3D-MoRSE	5	0.90	0.70	0.52	0.85	2	0.71
PLS18	WHIM	2	0.63	0.48	0.94	1.05	1	1.03
PLS19	GETAWAY	3	0.82	0.63	0.66	0.89	2	0.88
PLS20	PLS13-PLS19 scores[f]	2	0.92	0.87[e]	0.44	0.64	0	0.64

[a] See 3.1

[b] Number of PLS components

[c] Relative explained variance of the response determined by cross-validation. Seven cross-validation segments were used

[d] RMSEP of the test set after removal of the detected outliers

[e] Should not be compared with the other Q2Y values since the cross-validation is based on non-validated scores. RMSEP values from test set prediction are comparable, however.

[f] Top level hierarchical model based on the scores from PLS13-PLS19, see 3.2.2.2.

The training set selection was performed from a PCA model based on all descriptors. Hence, it is not necessarily so that the substances selected span the X space optimally, when only one of the descriptor groups are used for modelling, since different groups reflect different molecular properties. Thus, ideally a PCA model based on exactly the same descriptors as the quantitative PLS model should have been used. However, selection based on a single PCA model was considered to be enough for the purposes of this work. In addition, if different training and test sets had been used, it would have precluded thorough validation of the hierarchical model, since the different block models contributing would have been based on different training data.

In conclusion, about the same prediction error, slightly above 0.5 pEC_{50} units, was obtained from the models based on systematic training set selection as for models based on random training set selection. However, the number of outliers decreased when systematically selected training data was used, since the risk of model extrapolation decreased.

5.2 Green alga toxicity prognosis model

45 substances with reference values between pEC_{50} = -0.2 and pEC_{50} = 4.0 were used for modelling. For both random and systematic selection of the training set, 24 substance were used for training and the remaining 21 substances for testing the model.

5.2.1 Random training set selection

The models developed from random training set of 24 substances are shown in Table 4. The PLS10 model based on all descriptors gave an RMSEP of the test set of 0.50 pEC_{50} units after removal of four outliers that were found by the outlier detection method discussed in 5.1.1. The model based on only BCUT descriptors (PLS12) had a lower RMSEP but a higher number of substances did not fit into the model. Thus, we conclude that as discussed above for the Microtox models, a combination of information from different descriptor groups was needed.

Table 4. Model information for PLS models developed for alga toxicity pEC_{50} with random selection of training set. No prediction errors were calculated for models with $Q^2_{cv}<0.5$.

Model	Descriptors[a]	*k*[b]	R^2Y	Q^2Y[c]	RMSEE	RMSEP	outliers	RMSEP[d]
PLS10	all	3	0.87	0.59	0.39	0.81	4	0.50
PLS11	Constitutional + topological + mol. walk counts	3	0.82	0.65	0.44	0.75	2	0.56
PLS12	BCUT	7	0.95	0.74	0.27	1.05	7	0.35
PLS13	2D autocorr.	1	0.67	0.56	0.58	0.78	5	0.61
PLS14	RDF	4	0.85	0.72	0.42	0.80	4	0.51
PLS15	3D-MoRSE	1	0.70	0.38				
PLS16	WHIM	3	0.77	0.49				
PLS17	GETAWAY	1	0.67	0.44				
PLS18	PLS11-PLS17 scores[f]	2	0.92	0.84[e]	0.30	0.82	1	0.70

[a] See 3.1

[b] Number of PLS components

[c] Relative explained variance of the response determined by cross-validation. Seven cross-validation segments were used

[d] RMSEP of the test set after removal of the detected outliers

[e] Should not be compared with the other Q2Y values since the cross-validation is based on non-validated scores. RMSEP values from test set prediction are comparable, however.

[f] Top level hierarchical model based on the scores from PLS11-PLS17, see 3.2.2.2.

5.2.2 Systematic training set selection

Of the models based on a systematically selected training set, shown in Table 5, the best predictive ability was obtained when all descriptors were used (model PLS1). Despite the systematic selection of training set, outliers were detected in the test set but it was noted that they were fewer than in the models based on a randomly selected training set, as expected. A further example of how well the outlier detection method worked, can be found in Figure 7, which shows the test set predictions for the model PLS1. Of the three substances detected as outliers, two had grossly erroneous predictions.

Table 5. Model information for PLS models developed for alga toxicity pEC_{50} with systematic selection of training set. No prediction errors were calculated for models with $Q^2Y<0.5$.

Model	**Descriptors**[a]	***k***[b]	**R^2Y**	**Q^2Y**[c]	**RMSEE**	**RMSEP**	**outliers**	**RMSEP**[d]
PLS1	all	4	0.88	0.69	0.42	0.56	3	0.39
PLS2	Constitutional + topological + mol. walk counts	3	0.8	0.53	0.52	0.87	2	0.43
PLS3	BCUT	5	0.9	0.73	0.39	0.61	3	0.55
PLS4	2D autocorr.	2	0.73	0.55	0.59	0.67	1	0.68
PLS5	RDF	2	0.77	0.63	0.54	0.69	2	0.48
PLS6	3D-MoRSE	5	0.9	0.56	0.4	0.5	4	0.51
PLS7	WHIM	3	0.72	0.21				
PLS8	GETAWAY	2	0.71	0.6	0.61	0.61	0	0.61
PLS9	PLS2-PLS8 scores[f]	1	0.85	0.78[e]	0.43	0.55	1	0.52

[a] See 3.1

[b] Number of PLS components

[c] Relative explained variance of the response determined by cross-validation. Seven cross-validation segments were used

[d] RMSEP of the test set after removal of the detected outliers

[e] Should not be compared with the other Q2Y values since the cross-validation is based on non-validated scores. RMSEP values from test set prediction are comparable, however.

[f] Top level hierarchical model based on the scores from PLS2-PLS8, see 3.2.2.2.

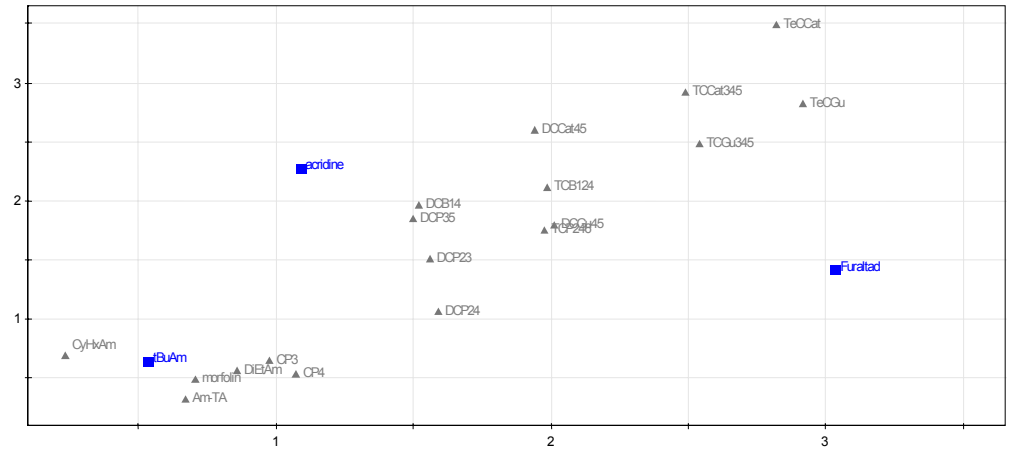


Figure 7. Measured versus predicted green alga pEC50 values for the test set of the model PLS1 based on all descriptors. Substances marked with squares are those detected as outliers by the outlier detection method used.

5.3 Daphnia toxicity prognosis model

The data set available for *Daphnia* pEC_{50} prognosis model development consisted of 93 substances with a wide range of pEC_{50} values spanning 7 orders of magnitude on the concentration scale, from –2.6 to 5.5.

Models based on all descriptors were built with both random and systematic training set selection as for Microtox pEC_{50}. In both cases 46 substances were selected as the test set, leaving 47 substances in the training set. The results are shown in Table 6.

Table 6. Model information for PLS models developed for *Daphnia* pEC_{50}.

Model	Training set	k[a]	R^2Y[b]	Q^2Y[c]	RMSEE	RMSEP	Outliers (test set)	RMSEP[d]
PLS1	Random	3	0.93	0.84	0.37	1.92	13	0.34
PLS10	Systematic	3	0.93	0.84	0.44	0.36	2	0.36

[a] Number of PLS components

[b] Relative explained variance of the response in the training set

[c] Relative explained variance of the response determined by cross-validation. Seven cross-validation segments were used

[d] RMSEP of the test set after removal of the detected outliers.

The accuracy of the models was very good with RMSEP values of approximately 0.35 pEC_{50} units, which corresponded to a factor 2.3 in concentration units. This is more than enough for a screening model. The difference between random and systematic selection

of training set was the same as observed for the Microtox toxicity models. The model performance as measured by prediction of the test set was approximately equal for both models but the model based on a randomly selected training set gives a higher number of prediction outliers (13 of 46 substances). As discussed above, this occurs because the training set does not span the space of the test set, *i.e.* because the model is extrapolated. Again, this stresses the need for outlier detection capabilities.

5.3.1 Simple variable selection

In order to investigate the possibility of decreasing the number of descriptors used in the modelling, a simple variable selection algorithm was used. The variables were ranked according to their importance as calculated by:

$$Z_j = \sum_{i=1}^{a} \left(w_{ij}^2 SS_i \right)$$

where Z_j is the importance of variable j, w_{ij} is the PLS loading weight of variable j for component i, SS_i is the explained sum of squares for the ith PLS component and a is the number of PLS components in the model. This algorithm is simple and naïve since it does not take into account that the most important variables found may very well contain the same information. Algorithms for stepwise variable selection that can be used to find combinations of variables that contain complementary information exist but has not been used in this work, since variable selection was not the primary objective. The ten most important variables in the PLS10 model based on systematic selection of training data were: BEHm1, XMOD, MW, X2sol, BEHm2, X0v, X1sol, RDF045p, X0sol and RDF045v. The ten most important variables in the PLS10 model based on random selection of training data were: X1sol, XMOD, VEA1, VED1, VEv1, VEp1, VEe1, VEZ1, VEm1 and X0sol[6].

It is interesting to note the different descriptors selected depending on the training set. The descriptor set selected for the model based on the more diverse, systematically selected training set, was more diverse. It contained BCUT descriptors, connectivity indices, molecular weight and radial distribution functions, which reflected that a combination of information from different descriptor groups improved prediction as discussed for the Microtox toxicity models in 5.1. The 10 descriptors selected from the model based on random selection of training data were of only two types: connectivity

[6] MW is molecular weight. Descriptors starting with (*cf.* 3.1): BEHm are BCUT descriptors weighted by atomic mass, X are different types of connectivity indices, RDF are radial distribution function descriptors, VE are topological descriptors based on eigenvalues of the adjacency matrix or different weightings of the distance matrix.

indices and topological descriptors based on eigenvalues of the adjacency matrix or the distance matrix.

Information for the models based on variable selection from the models in Table 6 is shown in Table 7. The PLS11 models based on variable selection from the PLS10 model has about the same performance as the PLS10 model while the performance of the PLS2 model based on variable selection from the PLS1 model has lower predictive ability. This was interpreted as a consequence of the lower diversity of the selected variables that originate in the lower diversity in the training data, which had two effects:

1. The variables selected do not contain enough information about the biological response, which is reflected in the higher RMSEE of model PLS2 compared to model PLS1.

2. The reduced variable set of model PLS2 does not allow efficient outlier detection as indicated that only one outlier is detected although the predicted versus measured plot indicates that several others exist, as can be expected since this is the case for model PLS1.

Table 7. Model information for PLS models developed for *Daphnia* pEC_{50} based on a reduced descriptor set.

Model	**Training set**	k[a]	R^2Y[b]	Q^2Y[c]	**RMSEE**	**RMSEP**	**Outliers (test set)**	**RMSEP**[d]
PLS2	Random	1	0.97	0.84	0.56	0.83	1	0.82
PLS11	Systematic	1	0.90	0.87	0.53	0.41	1	0.42

[a] Number of PLS components

[b] Relative explained variance of the response in the training set

[c] Relative explained variance of the response determined by cross-validation. Seven cross-validation segments were used

[d] RMSEP of the test set after removal of the detected outliers

In conclusion, the studied data set indicated that it is dangerous to perform variable selection on a data set that is not completely representative for future application of the model. Although not all descriptors included in the PLS model contain significant information about the biological response modelled, they can add information that is useful when detecting substances that are not covered by the model. In real application of QSAR models for screening, the substances for which the model will be used are not known beforehand. Hence, variable selection should be done with caution if outlier detection is a consideration. This is a modelling aspect that has been neglected in most previous QSAR work.

5.4 Fish toxicity prognosis model

The data set available for modelling of toxicity to *Lepomis macrochirus* is smaller than for *e.g.* Microtox toxicity. A total of 55 substances with reference values were available. The pLC_{50} values of these substances were between –0.6 and 3.2.

Initial modelling showed that one of the substances, 1,3-dichloropropene, was badly predicted by all models although several similar substances were available in the data set and no indication of the substance as an X outlier was obtained. This led to the suspicion that the substance is an Y outlier, *i.e.* that the published toxicity used as reference value, $pLC_{50} = 1.26$, was inaccurate. No other published value of toxicity to *Lepomis macrochirus* was found in the literature but a 96 h LC_{50} of 0.239 g/l for another species, Fathead Minnow (*Pimephales promelas*), was found [Geiger *et al*. 1990]. This value corresponds to $p[LC_{50}$ mmol/l] = - 0.3. The difference is much larger than what can be expected from the fact that the tests are performed on two different fish species. Further, a comparison of *Lepomis macrochirus* toxicity values with structurally similar compounds, *e.g.* 1,1-dichloropropane, 1,1-dichloroethylene (Appendix B), showed that these had significantly lower pLC_{50} values than 1,3-dichloropropene. This is no evidence of a bad value for 1,3-dichloropropene, but in combination with the toxicity for Fathead Minnow, it is a strong indication that the value is not reliable. Thus, 1,3-dichloropropene was excluded from all further modelling and was not included in the models presented in this section.

5.4.1 *Lepomis macrochirus* toxicity model with random training set selection

A number of models were developed based on different descriptors as shown in Table 8. The randomly selected training set consisted of 32 substances, leaving 22 substances in the test set. It is clear that the models developed based on single descriptor groups had very poor performance, which was not always the case for the Microtox and *Daphnia* responses. There were two likely explanations for this and both probably contributed to the poor performance:

1. The response was supposedly more complex since fish are higher organisms than bacteria, algae and *Daphnia*.

2. The data set was of lower quality for modelling since the spanned effect range was smaller than for the other responses. This is further discussed in 6 below.

The models PLS10 and PLS18 that were both based on all descriptors but with different model structures had better performance in terms of Q^2 and RMSEP values. For the PLS10 model, the RMSEP was significantly lower than the RMSEE, which is

unexpected and should be attributed to chance, since the training and test sets were chosen randomly. Model performance was considered satisfying and fulfilling the need for screening of substances. To illustrate the results, predicted vs. measured pLC_{50} values for the PLS10 model training and test sets are shown in Figure 8.

It should be noted that PLS10 and PLS18 had very similar predictive properties, although outlier detection properties differed. If the 4 outliers from PLS10 are removed from the PLS18 test set (one of them was already detected in PLS18), the resulting RMSEP = 0.38, which was practically the same as the RMSEP = 0.34 obtained for PLS10. Hence, prediction was adequate but outlier detection seemed to be inadequate in the hierarchical model. We would again like to emphasise that outlier detection is as important as predictive ability, since no model will cover every substance that people try to use it for.

Table 8. Model information for PLS models developed for Lepomis pLC_{50} with random selection of training set. No calculation of prediction errors were made for models with Q^2Y^c <0.5.

Model	**Descriptors**[a]	***k***[b]	**R^2Y**	**Q^2Y^c**	**RMSEE**	**RMSEP**	**Outliers (test set)**	**RMSEP**[c]
PLS10	all	2	0.79	0.70	0.48	4.02	4	0.34
PLS11	Constitutional + topological + mol. walk counts	2	0.73	0.61	0.54	0.92	3	0.57
PLS12	BCUT	3	0.67	0.54	0.60	0.75	1	0.75
PLS13	2D autocorr.	2	0.69	0.43				
PLS14	RDF	2	0.45	0.16				
PLS15	3D-MoRSE	1	0.47	0.30				
PLS16	WHIM	2	0.67	0.40				
PLS17	GETAWAY	2	0.62	0.39				
PLS18	PLS11-PLS18 scores[e]	1	0.76	0.73[f]	0.50	3.79	1	0.75

[a] See 3.1

[b] Number of PLS components

[c] Relative explained variance of the response determined by cross-validation. Seven cross-validation segments were used

[d] RMSEP of the test set after removal of the detected outliers

[e] Top level hierarchical model based on the scores from PLS3-PLS9, see 3.2.2.2.

[f] Should not be compared with the other Q2Y values since the cross-validation is based on non-validated scores. RMSEP values from test set prediction are comparable, however.

It is likely that all descriptors were not needed to obtain results similar to those from PLS10, although Table 8 clearly shows that a single descriptor group was not enough.

Investigations regarding the descriptor groups actually needed have not been performed as a part of this work. As noted in the discussion on outlier detection in 5.1.1, it was definitely not only predictive ability that should be considered when selecting variables, since descriptors that did not contribute significantly to predictive ability may contain information that can help in detecting substances that were not covered by the model.

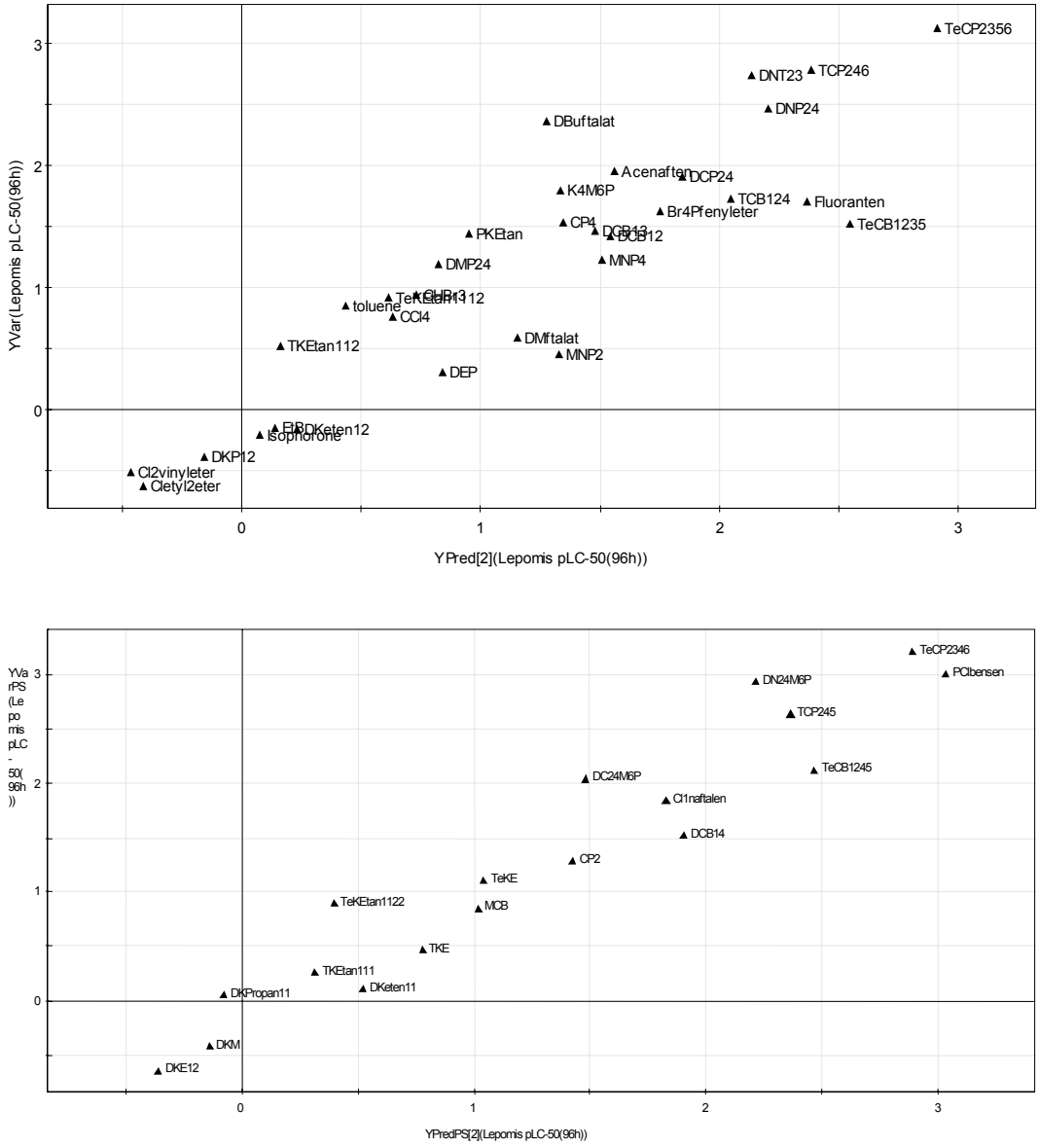


Figure 8. Predicted vs. measured toxicities for the PLS10 model: training data (top) and test set after removal of four outliers (bottom).

5.4.2 *Lepomis macrochirus* toxicity model with systematic training set

Systematic selection of training data was performed based on a PCA model of all substances and all descriptors as discussed in 3.2.7 and for Microtox models in 5.1.3. The same sizes of the training and test sets were used as for the random selection, *i.e.* 32 substances for training and 22 substances for testing.

When modelling was attempted it was clear from the diagnostics available in PLS that two of the substances selected for training could not be used for model estimation, since they were outliers that distorted the model structure. These two substances, 2,4,6-trinitrophenol and dibutylphthalate, were not discarded but moved to the test set. Note that the fact that they did not fit in the model during estimation should mean that they were detected as outliers during prediction. Moving them to the test set thus gave further testing of the outlier detection possibilities of the different models.

The modelling results, shown in Table 9, were similar to those obtained for random training set selection in 5.4.1 above. No single descriptor group contained enough information to give a good model, but the model based on all descriptors, PLS1, gave a good accuracy with an RMSEP of 0.40 for the test set after removal of outliers. It is surprising that the number of outliers was in general higher for the models presented in this section, based on systematic selection of training data, compared to the models based on a randomly selected training set. The opposite would have been expected and this was also the result when the two training set selection methods were compared for other responses (Microtox in 5.1.2 and 5.1.3, Alga in 5.2 and *Daphnia* in 5.3).

Predictions for the test set, including outliers, for model PLS1 are shown in Figure 9. It is clear that the two substances removed from the training set, 2,4,6-trinitrophenol and dibutylftalate, were badly predicted and that this was detected by the outlier detection method used. In addition, 1,2-diphenylhydrazine and hexachloroethane, were poorly predicted, which was detected. The remaining four substances that were indicated as outliers were not inaccurately predicted, but as discussed before, a false positive prediction is better than a false negative.

Table 9. Model information for PLS models developed for Lepomis pLC_{50} with systematic selection of training set. No calculation of prediction errors were made for models with $Q^2<0.5$.

Model	**Descriptors**[a]	***k***[b]	**R^2Y**	**Q^2Y**[c]	**RMSEE**	**RMSEP**	**Outliers (test set)**	**RMSEP**[d]
PLS1	all	2	0.88	0.78	0.37	0.92	8	0.40
PLS2	Constitutional + topological + mol. walk counts	2	0.87	0.74	0.39	1.10	6	0.69
PLS3	BCUT	4	0.87	0.67	0.40	0.96	2	0.89
PLS4	2D autocorr.	2	0.82	0.65	0.45	0.98	4	0.96
PLS5	RDF	2	0.52	0.31				
PLS6	3D-MoRSE	2	0.73	0.60	0.55	1.03	4	0.67
PLS7	WHIM	2	0.62	0.31				
PLS8	GETAWAY	2	0.68	0.38				
PLS9	PLS2-PLS8 scores[e]	1	0.85	0.82[f]	0.40	1.21	3	0.70

[a] See 3.1

[b] Number of PLS components

[c] Relative explained variance of the response determined by cross-validation. Seven cross-validation segments were used

[d] RMSEP of the test set after removal of the detected outliers

[e] Top level hierarchical model based on the scores from PLS2-PLS8, see 3.2.2.2.

[f] Should not be compared with the other Q2Y values since the cross-validation is based on non-validated scores. RMSEP values from test set prediction are comparable, however.

Similar to the results for random training data selection, the two models based on all data, PLS1 and PLS9, had similar predictive ability but differed in outlier detection properties. If the outliers detected in PLS1 were removed from the PLS9 test set, one obtains RMSEP = 0.37, *i.e.* very close to the value for PLS1.

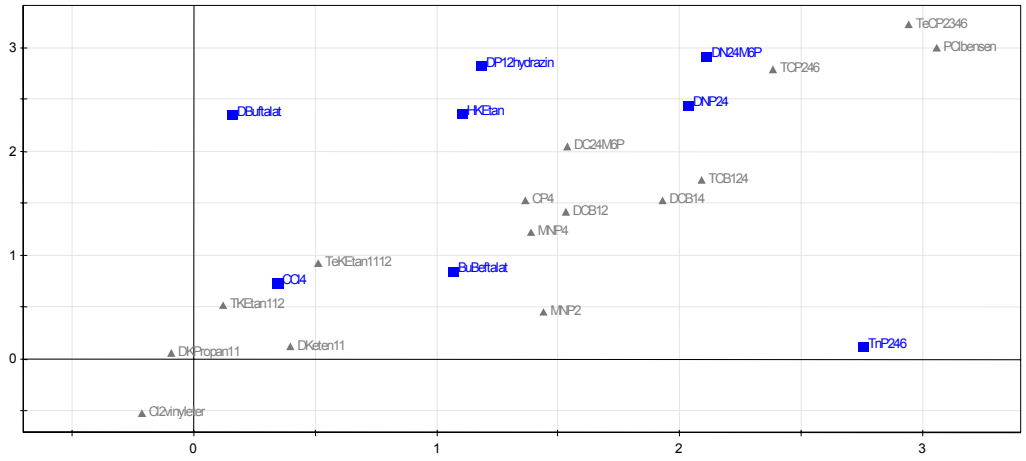


Figure 9. Predicted vs. measured pLC_{50} values for the test set for model PLS1

5.4.3 *Leicuscus idus* toxicity models

The following toxicity data was not used due to large differences (more than 0.5 log units) in LC_{50} values between two sets of data reported in Juncke and Lüdemann [1978]. The differences indicated uncertainties in fish toxicity determinations, that would obscure QSAR modelling: acrolein, amylacetate, benzaldehyde, carbontetrachloride, chlorobenzene, cycloheptene, cyclohexane, cyclohexene, cyclohexylamine, isobutyronitrile, isopropylbenzene and lindane. Dodecylbenzene was omitted since its LC_{50} value was higher than its water solubility. In addition, ethyleneimine, acetone cyanohydrine, propargyl alcohol and hydroquinone were not used due to strange behaviour during initial modelling. The reason for this was not clear, but a possible explanation could be erroneous reference values or very specific mechanisms of toxic action that were not represented among other substances in the data set. The remaining data set consisted of 117 substances with pLC_{50} values between –2.1 and 3.2. Of these 47 were used as test set and the remaining 60 as training data.

The models had poorer predictive ability as measured by cross-validation Q^2 (*cf.* Table 10) compared to most other models presented in this report. Q^2 values of less than 0.5 were usually considered to indicate a poor fit. Nevertheless, the RMSEP values were not much larger than those in the Microtox toxicity models. The reason is that Q^2 measures relative explained variance while RMSEP are absolute measures. The reference value range was smaller for *Leuciscus idus* toxicity than for *e.g.* Microtox toxicity.

However, it can be noted that the RMSEP values of the test sets were not different from the RMSEE values. The latter values were calculated from training data and corresponded to the R^2 values in the table. We have not calculated Q^2 values for the test

set but it was clear that they were close to the R^2 values. Thus cross-validation is probably slightly pessimistic in this case due to the low number (7) of segments used. It should be noted however, that we used cross-validation only to select the model complexity as discussed in 3.2.5 and not to measure model performance.

Table 10. Model information for PLS models developed for *Leuciscus idus* pLC_{50}.

Model	Training set	k[a]	R^2Y[b]	Q^2Y[c]	RMSEE	RMSEP	Outliers (test set)	RMSEP[d]
PLS1	Systematic	3	0.74	0.47	0.56	0.62	2	0.58
PLS10	Random	3	0.77	0.43	0.54	-[e]	5	0.51

[a] Number of PLS components

[b] Relative explained variance of the response in the training set

[c] Relative explained variance of the response determined by cross-validation. Seven cross-validation segments were used

[d] RMSEP of the test set after removal of the detected outliers

[e] Not calculated because some of the outlying substances have very poor predictions.

This data set was the most diverse regarding the chemical structure of the substances and better model performance would probably be obtained if substances were classified into groups with similar behaviour and properties before modelling. This has not been done in the present work, however, but is an interesting topic for continued work.

6 Discussion

In order to discuss the quality of models, one must choose a method to measure model performance. Frequently, R^2 and Q^2 values are reported in QSAR studies as measures of model performance. However, although these certainly tell us something about the significance of the fit, they do not help us in judging whether or not the model is accurate enough for application purposes, since they say nothing about the expected uncertainty of a prediction when the model is used. Therefore, we have chosen to use RMSEP as the primary model quality criterion in this study. From an application point-of-view this is the relevant performance measure. RMSEP values can be compared for models based on different training data, which is not the case for R^2 and Q^2.

Table 11shows a summary of the models developed for the toxic properties considered in this work based on all descriptors. For the two properties with a large test set (Microtox pEC_{50} and *Daphnia* pEC_{50}) the RMSEPs for systematic and random selection was practically equal (although the number of outliers was different as discussed above).

For the two properties with small test sets (alga pEC_{50} and *Lepomis macrochirus* pLC_{50}) the difference between the two RMSEP estimates was larger. A possible explanation for this fact is that expectation values for prediction errors are known to be difficult to estimate (this is the reason we have used as large test sets as considered possible in all cases). The lower number of substances used in the estimation, as shown in the second column of the table, means that RMSEP estimation was less accurate than in the other two cases.

The last column of the table shows pooled RMSEP values for the five properties. These are believed to be the best available RMSEP estimates for a model based on all substances available in each case, *i.e.* the model that will be used for prediction of toxicity of new substances. Thus, these were the prediction error estimates used in the discussion of model performance below.

Table 11. Prediction errors for models based on all descriptors with systematic and random selection of the training set.

Organism and property	**No. of test substances**	**RMSEP systematic**	**RMSEP random**	**RMSEP pooled**
Microtox pEC_{50}	41	0.54	0.53	0.54
Green alga pEC_{50}	21	0.39	0.50	0.45
Daphnia magna pEC_{50}	47	0.36	0.34	0.35
Lepomis macrochirus pLC_{50}	22	0.40	0.34	0.37
Leuciscus idus pLC_{50}	47	0.58	0.51	0.55

The EPA QSAR evaluation study defined successful prediction of properties as being within a factor 10 of the experimental value [EPA 1994]. Because RMSEP values can be treated approximately as standard deviations of model predictions, we can transfer this criterion to the models in this work. The Microtox, alga toxicity and *Leuciscus idus* models have an RMSEP (in log units) of approximately 0.5, which means that ± one order of magnitude corresponded to ± 2 RMSEP values, *i.e.* two standard deviations, which is approximately a 95 % confidence interval. For the *Daphnia magna* and *Lepomis macrochirus* toxicity models, the RMSEP was approximately 0.35 log units, which means that one order of magnitude corresponded to 3 standard deviations, which corresponded roughly to a 99.5 % confidence interval. Thus, of the substances not indicated as outliers by the outlier detection diagnostics, roughly 99.5 % of the *Daphnia magna* and *Lepomis macrochirus* toxicity predictions and roughly 95 % of the Microtox, alga and *Leuciscus idus* toxicity predictions will be correct according to the EPA criterion.

This can be compared to the results obtained in a performance study of the ECOSAR software provided by the US Environmental Protection Agency for estimation of

aquatic toxicity of different types of chemicals. The models are based solely on the descriptor log K_{OW} and are of highly questionable quality [Kaiser *et al*. 1999, http://www.disat.unimib.it/chm/QSARnews3.htm 2003-02-19]. In some cases they are based on only 2 or 3 substances. The performance study report [EPA 1994] discussed some of the EPA QSAR models routinely used to estimate environmental effects of new chemicals. For physicochemical properties like log K_{OW}, water solubility and vapour pressure, the models were of limited accuracy with between 50% and 70% of the predictions considered correct (within one order of magnitude from the experimental value). For aquatic toxicity, the hit rates were higher with 71% and 82% correct predictions for *Daphnia* and fish toxicity, respectively. It should be noted that the set of about 140 chemicals used in the study were diverse, containing organics, metallo-organics and polymers.

However, we stress that in our opinion the most alarming issue is not the quite low hit rates but the lack of diagnostic information that could indicate that a prediction is unreliable. Using these models there is no way of telling whether or not the prediction is among the correct 50-70 %, except from experience and expert knowledge of QSAR modelling, which is unsatisfactory. The models developed in this work gives a much higher rate of correct predictions for a limited set of substances but are not applicable for the diverse set of chemicals used in the test of the EPA models. However, as we have shown, if a substance was not correctly predicted by the PLS models this was detected by the outlier detection method. If so, the user knows that the prediction is not correct, and that he has to use other tools to predict or to measure toxicity for this particular substance.

To finalise this discussion, it should be mentioned that the prediction being within one order of magnitude from the literature value is not equal to the prediction being within one order of magnitude from the actual value. Errors in experimental determination and other sources of uncertainty, such as impurities, are implicitly included in the estimated model error. The so-called apparent root mean square error of prediction ($RMSEP_{apparent}$), which is the one obtained during test set validation of models, can be written as:

$$RMSEP_{apparent} = \sqrt{RMSEP^2 + s_{ref}^2}$$

here, RMSEP is the true prediction error of the model, *i.e.* the estimated difference between model predictions and the actual ("true") value of the property predicted and s_{ref} is the error in the reference values expressed as a standard deviation. Thus, the $RMSEP_{apparent}$ is expected to be higher than the true RMSEP. Therefore, the number of correct model predictions, measured in comparison with the true toxicity value is higher than indicated above.

7 Conclusions

QSAR prognosis models for five different toxicity end-points have been developed: Microtox, green alga, *Daphnia magna*, *Lepomis macrochirus* and *Leuciscus idus* toxicity. The resulting models have good predictive ability and are well suited for screening of new and existing chemicals and priority setting for further testing. They are valid for organic molecules that are 'similar' to the substances used for model training (Appendix B).

It may be difficult for a model user to objectively assess if a substance is 'similar' to training substances. It has been shown that outlier diagnostics based on the distance to the PLS model space can be used to reliably detect prediction outliers. This reduces dramatically the risk of false negative predictions, which is the most alarming risk when using QSAR models in prediction of environmental effects.

The difference in the number of outliers between models with a randomly and systematically selected training set illustrates well the need of representative training data, since models of the type used do in general not extrapolate well. Prediction outliers can never be completely avoided for a QSAR model that is used in reality. Thus, it is of great importance to be able to detect when the model is not valid and predictions should not be trusted.

8 Acknowledgements

This research project was financed by the Environmental Foundation of the Swedish Association of Graduate Engineers (CFs Miljöfond) and the Foundation for the Swedish Environmental Research Institute (SIVL).

9 References

Adams, N. Dobbs, A. J. 1984. A comparison of results from two test methods for assessing the toxicity of aminotriazole to Selenastrum capricornutum, Chemosphere **13**, 965-971.

Alexander, H. C., Dill, D. C., Smith, L. W., Guiney, P. D., Dorn, P. 1988. Bisphenol A: Acute aquatic toxicity, Environ. Toxicol. Chem. **7**, 19-26.

Andersson, P. M., Sjöström, M., Wold, S., Lundstedt, T. 2000. Comparison between physicochemical and calculated molecular descriptors, J. Chemom. **14,** 629-642.

Blaylock, B. G., Frank, M. L., McCarthy, J. F. 1985. Comparative toxicity of copper and acridine to fish, Daphnia and algae, Environ. Toxicol. Chem. **4,** 63-71.

Buccafusco, R. J., Ellis, S. J., LeBlanc. G. A.1981. A*cute toxicity of priority pollutants to bluegill (*Lepomis macrochirus*)*. Bull. Environ. Contam. Toxicol. **26,** 446-452.

Burden, F. R., Winkler, D. A. 1999. *New QSAR methods applied to structure-activity mapping in combinatorial chemistry*, J. Chem. Inf. Comput. Sci. **39,** 236-242.

Buydens, L. M. C., Reijmers, T. H., Beckers, M. L. M., Weherens, R. 1999. *Molecular data-mining: a challenge for chemometrics*, Chemom. Intell. Lab. Syst. **49,** 121-133.

Calamari, D., Da Gasso, R., Galassi, S., Provini, A. Vighi, M. 1980. *Biodegradation and toxicity of selected amines on aquatic organisms*, Chemosphere **9,** 753-762.

Calamari, D., Galassi, S., Da Gasso, R. 1979. *A system of tests for the assessment of toxic effects on aquatic life: An experimental preliminary approach*, Ecotoxicol. Environ. Safe. **3,** 75-89.

Calamari, D., Galassi, S., Setti, F., Vighi, M. 1983. *Toxicity of selected chlorobenzenes to aquatic organisms*, Chemosphere **12,** 253-262.

Devillers, J.,Chambon, P., Zakarya, D., Chastrette, M.,Chambon, R. 1987. *A predictive structure-toxicity model with* Daphnia magna. Chemosphere **16**, 1149-1163.

Draper III, A. C., Brewer, W. S. 1979. *Measurement of the aquatic toxicity of volatile nitrosamines*, J. Toxicol. Environ. Health **5,** 985-993.

Eloranta, V. 1982*Effect of the slimecide Fennosan F 50 on algal growth in different test media*. Pap. Puu **64,** 129-135.

EPA, US Environmental Protection Agency, 1994. *US EPA/EC joint project on the evaluation of (quantitative) structure activity relationships*, EPA report 743-R-94-001.

EPA, US Environmental Protection Agency, 2000. *Summary of the endocrine disruptor priority-setting workshop*, U.S. Environmental Protection Agency, EPA web site

Eriksson, L., Johansson, E. 1996. *Multivariate design and modelling in QSAR*, Chemom. Intell. Lab Syst. **34,** 1-19.

Eriksson, L., Johansson, E., Müller, M., Wold, S. 2000. *On the selection of the training set in environmental QSAR analysis when compounds are clustered,* J. Chemom. **14,** 599–616.

Esbensen, K, Schönkopf, S., Midtgaard, T. 1996. *Multivariate analysis in practice*, Camo A/S, Trondheim.

Flower, D. R., 1998. *On the properties of bit string-based measures of chemical similarity*, J. Chem. Inf. Comput. Sci. **38,** 379-386.

Galassi, S., Mingazzini, M., Vignanò, L., Cessareo, D., Tosato, M. L. 1988. *Approaches to modelling toxic responses of aquatic organisms to aromatic hydrocarbons*, Ecotoxicol. Environ. Safe. **16,** 158-169.

Galassi, S., Vighi, M. 1981. *Testing toxicity of volatile substances with algae*, Chemosphere **10,** 1123-1126.

Geiger, D.L., Brooke, L.T., Call, D.J., 1990, *Acute Toxicities of Organic Chemicals to Fathead Minnows (Pimephales promelas)*, Vol. 5, Center for Lake Superior Environmental Studies, University of Wisconsin, Superior, WI

Geladi, P., Kowalski, B. R. 1986. *Partial least squares regression: a tutorial*, Anal. Chim. Acta **185,** 1-17.

Giraud, E., Lutmann, C., Lavelle, F., Riou, J-F., Mailliet, P., Laoui, A. 2000. *Multivariate data analysis using D-optimal designs, partial least squares and response surface modelling: a directional approach for the analysis of farnesyltransferase inhibitors*, J. Med. Chem. **43,** 1807-1816.

Hemmer, M.C., Steinhauer, V., Gasteiger, J., 1999. *Deriving the 3D structure of organic molecules from their infrared spectra*, Vib. Spectrosc. **19,** 151-164.

IUCLID database 2000, European Chemicals Bureau, EU JRC, Ispra, Italy

Juhnke, I., Lüdemann, D. 1978. *Ergebnisse der Untersuchung von 200 chemischen Verbindungen auf akute Fischtoxizität mit dem Goldorfentest*. Z. Wasser Abw. Forsch. **11**, 161-164.

Kaiser, K. L. E., Dearden, J. C., Klein, W., Schultz, T. W. 1999. *A note of caution to users of ECOSAR*, Water Qual. Res. J. Canada **34,** 179-182.

Kaiser, K. L. E., Palabrica, V. S. 1991. Photobacterium phosphoreum *toxicity data index*. Water Poll. Res. J. Canada **26,** 361-431.

Kuivasniemi, K., Eloranta, V., Knuutinen, J. 1985. *Acute toxicity of some chlorinated phenolic compounds to* Selenastrum capricornutum *and phytoplankton*, Arch. Environ. Contam. Toxicol. **149,** 43-49.

Källqvist, T., Svenson, A. 2003. *Assessment of ammonia toxicity in tests with the microalga,* Nephroselmis pyriformis*, Chlorophyta*. Wat. Res. **37**, 377-384.

Lipnick, R. L. 1991. *Outliers: their origin and use in the classification of molecular mechanisms of toxicity*, Sci. Tot. Environ. **109/110**, 131-153.

Livingstone, D. J. 2000. *The characterisation of chemical structure using molecular properties. A survey*, J. Chem. Inf. Comput. Sci. **40,** 195-209.

Macri, A, Sbardella, E. 1984. *Toxicological evaluation of nitrofurazone and furaltadone on* Selenastrum capricornutum, Daphnia magna *and* Musca domestica, Ecotoxicol. Environ. Saf. **8,** 101-105.

Martens, H., Naes, T., 1989. *Multivariate calibration*, John Wiley & Sons, Chichester.

Mekenyan, O., Ivanov, J., Karabunarliev, S., Bradbury, S. P., Ankley, G. T., Archer, W. 1997. *A computationally-based hazard identification algorithm that incorporates ligand flexibility. 1 Identification of potential androgen receptor ligands*, Env. Sci. Technol. **31,** 3702-3711.

Nyholm, N., Källqvist, T., 1989. methods for growth inhibition toxicity tests with freshwater algae, Environ. Toxicol. Chem. 8, 689-703.

Qin, S J., Valle. S., Piovoso, M J., 2001, On unifying multiblock analysis with apllication to decentralized process monitoring, J. Chemometrics 15, 715-742

Randic, M., 1995. *Molecular shape profiles,* Chem. Inf. Comput. Sci. **35**, 373-382

Randic, M., Razinger, M. 1995. *On characterisation of molecular shapes*, J. Chem. Inf. Comput. Sci. **35**, 594-606.

Schur, J. H., Selzer, P., Gasteiger, J. 1996. The coding of three-dimensional structure of molecules by molecular transforms and its application to structure-spectra correlations and studies of biological activity, J. Chem. Inf. Comput. Sci. **36**, 334-344.

Shi, L. M., Fang, H., Tong, W., Wu, J. Perkins, R., Blair, R. M., Branham, W. S. Dial, S. L., Moland, C. L., Sheehan, D. M. 2001, QSAR models using a large diverse set of estrogens, J. Chem. Inf. Sci. 41, 186-195

Shigeoka, T., Sato, Y., Takeda, Y., Yoshida, K., Yamauchi, F. 1988. Acute toxicity of chlorophenols to green algae, Selenastrum capricornutum and Chlorella vulgaris, and quantitative structure-activity relationships, Environ. Toxicol. Chem., **7,** 847-854.

Suzuki, T., Ide, K., Ishida, M., Shapiro, S., 2001. *Classification of environmental estrogens by physicochemical properties using principal component analysis and hierarchical cluster analysis*, J. Chem. Inf. Comput. Sci. **41,** 718-726.

Svenson, A. 1993. *Microtox-test, en metodbeskrivning*, IVL-publ. B 1100.

Tong, W., Lowis, D. R., Perkins, R. Chen, Y., Welsh, W. J., Godette, D. W., Heritage, T. W. and Sheehan D. M. 1998, *Evaluation of QSAR methods for large-scale prediction of chemicals binding to the estrogen receptor*, J. Chem. Inf. Comput. Sci. 38, 669-677

Wehrens, R., de Gelder, R., Kemperman, G. J., Zwanenburg, B.,Buydens, L. M. C., 1999. *Molecular challenges in modern chemometrics*, Anal. Chim. Acta 400**,** 413-424.

Westerhuis, J A., Kourti, T., MacGregor, J F., 1998, *Analysis of multiblock and hierarchical PCA and PLS models*, J. Chemometrics 12, 301-321

Westerhuis, J A., Coenegracht, P M J., 1997, *Multivariate modelling of the pharmaceutical two-step process of wet granulation and tableting with multiblock partial least squares*, J. Chemometrics 11, 379-392

Wold, S., Esbensen, K., Geladi, P. 1987. *Principal component analysis*, Chemom. Intell. Lab Syst. **2,** 37-52.

Appendix A: Descriptors calculated by the Dragon software

The descriptors calculated by the Dragon software can be divided into 18 logical blocks:

1. constitutional descriptors (47)
2. topological descriptors (266)
3. molecular walk counts (21)
4. BCUT descriptors (64)
5. Galvez topological charge indices (21)
6. 2D autocorrelations (96)
7. charge descriptors (14)
8. aromaticity indices (4)
9. Randic molecular profiles (41)
10. geometrical descriptors (70)
11. RDF descriptors (150)
12. 3D-MoRSE descriptors (160)
13. WHIM descriptors (99)
14. GETAWAY descriptors (197)
15. functional groups (121)
16. atom-centred fragments (120)
17. empirical descriptors (3)
18. properties (3)

The full list of all 1481 descriptors is too extensive to include in this report but can be exported from the Dragon software or found on the web site of the Milano Chemometrics and QSAR Research Group (http://www.disat.unimib.it/chm/Varfile.pdf).

Appendix B: Substances and reference data used

The short name given in the table corresponds to the labels in the plots that show modelling results.

Name	Short name	Microtox pEC-50	Daphnia pEC-50	Leuciscus pLC50	Lepomis pLC50	Alga pEC-50
Acenaphthene	Acenaften				1.96	
Acetaldehyde	Acetaldehyd			-0.50		
Acetoacetate	Acetoacetic ester			-0.60		
Acetone	aceton	-2.23	-2.17	-2.29		
Acetone cyanohydrin [7]	Acetoncyanhydrin			1.70		
Acetonitrile	Acetonitril			-2.15		
Acridine [8]	acridine					2.30
Acrolein	Acrolein			1.35		
Acrylic acid	AcryCOOH		-0.21	-0.64		
Acrylic acid, butylester	AcrylBuester			0.75		
Acrylic acid, 2-ethylhexylester	AcrylEt2Hxester			0.90		
Acrylonitrile	Acrylnitril			0.28		
Allylamine	Allylamin			-0.12		
Allyl-N-thiourea	Allyltiourea			-1.50		
Aminotriazole	Am-TA		-1.31			0.32
Amylacetate	Amylacetat			0.00		
Amylethylketone	Amyletylketon			0.24		
2-Methyl-2-butanol[9]	Amyl-t-OH			-1.44		
Aniline	aniline	0.13	0.31	0.18		0.69
Anisole	anisol			-0.05		
Atrazine	atrazine			0.82		
Benzylalcohol	Balk	0.23	0.29			
Benzene	bensen	0.00	0.45	0.37		
Benzoic acid	Bensoicacid			-0.58		
Benzaldehyde	Benzaldehyd			0.23		

[7] (2-Hydroxyisobutyronitrile)
[8] (2,3-benzoquinoline)
[9] (tert. Amylalcohol)

Name	Short name	Microtox pEC-50	Daphnia pEC-50	Leuciscus pLC50	Lepomis pLC50	Alga pEC-50
Benzonitrile	Benzonitril			0.03		
Benzotrichloride	Benzotriklorid			-1.33		
Benzylchloride	BKL	1.75		1.63		
Bromo-4-phenyl-phenylether	Br4Pfenyleter				1.63	
Butylamine	BuAm1	-2.40		-0.51		
Butylbenzylphthalate	BuBeftalat				0.86	
Butylglycol	Buglycol			-1.13		
Butylether	Bu-n-eter			0.26		
n-Butylacetate	BuOAc		0.69	-0.08		
Butanol-1	BuOH1	-1.61	-1.40	-1.13		
Butanol-2	BuOH2			-1.60		
t-Butylacetate	Bu-t-acetat			-0.49		
Butanone	Butanon	-1.67		-1.81		
t-Butylbenzene	Bu-t-bensen			0.31		
Butyldiglycol	Butyldiglycol			-1.05		
n-Butyraldehyde	Butyraldehyd			-0.20		
Butyric acid	Butyric acid			-0.74		
Carbaryl	Carbaryl			1.00		
Chlorobenzene	CB	0.99				
Carbon tetrachloride	CCl4	0.66	0.20	0.21	0.76	
Bromoform	CHBr3				0.94	
Chloroform	CHCl3	-0.75	0.25	-0.13		
Chlordane	Chlordan		3.41			
Citric acid	Citronsyra			-0.60		
1-Chloronaphthalene	Cl1naftalen				1.85	
2-Chloroethyl-vinyl ether	Cl2vinyleter				-0.52	
Bis-2-chloroethylether	Cletyl2eter				-0.62	
2-Chlorophenol	CP2	0.51	0.86		1.29	0.26
3-Chlorophenol	CP3	0.99	0.91			0.65
4-Chlorophenol	CP4	1.15	1.20		1.53	0.53
2-chlorotoluene	CT2	1.32		0.21		
Cycloheptane	CyHeptan			-0.32		
Cycloheptene	Cyhepten		-0.07	0.23		

Name	Short name	Microtox pEC-50	Daphnia pEC-50	Leuciscus pLC50	Lepomis pLC50	Alga pEC-50
Cyclohexane	CyHexan			0.18		
Cyclohexanone	CyHexanon			-0.74		
Cyclohexene	CyHexen			0.32		
Cyclohexylacetate	CyHexylOAc			0.23		
Cyclohexanamine	CyHxAm		0.31	-0.29		0.70
Cyclopentanol	CyPentanOH			-1.24		
Cyclopentanone	Cypentanon			-1.58		
Di-n-butylphtalate	DBuftalat				2.37	
2,4-Dichloro-6-methylphenol	DC24M6P				2.04	
1,2-Dichlorobenzene	DCB12	1.67		0.70	1.42	1.82
1,3-Dichlorobenzene	DCB13	1.57	1.41		1.47	
1,4-Dichlorobenzene	DCB14	1.48			1.53	1.96
3,4-Dichlorocatechol	DCCat34					2.85
4,5-Dichlorocatechol	DCCat45					2.60
4,5-Dichloroguaiacol	DCGu45					1.80
2,3-Dichlorophenol	DCP23	1.53	1.50			1.51
2,4-Dichlorophenol	DCP24	2.14	1.78	1.51	1.91	1.07
2,5-Dichlorophenol	DCP25	1.23	1.52			
2,6-Dichlorophenol	DCP26	1.08	1.24			0.75
3,4-Dichlorophenol	DCP34	1.99	1.77			1.71
3,5-Dichlorophenol	DCP35	1.65	1.89			1.85
1-Decanol	DeOH1	2.38	1.16	2.42		
2-Decanol	DeOH2	2.13				
Diethylphthalate	DEP	0.29		0.62	0.31	
1,2-Diethylbenzene	Detbensen			0.62		
Diacetone alcohol	Diacetonalkoh			-1.89		
Diallylphthalate	Diallylftalat			2.79		
Dibutylamine	diBuAm					0.83
Diethylamine	DiEtAm		-0.49	1.48		0.56
Diethanolamine	Dietanolamin			-1.25		
Diethylnitrosoamine	diEt-NA					0.96
Dietyloxalate	Dietyloxalat			-0.32		
Dimethylamine	diMeAm					0.86

Name	Short name	Microtox pEC-50	Daphnia pEC-50	Leuciscus pLC50	Lepomis pLC50	Alga pEC-50
Dimethylnitrosamine	diMe-NA					1.27
2,4-Dinitro-5-sec-butylphenol	Dinoseb			2.08		
Dioxane	Dioxan	-0.88		-1.98		
Diphenylether	DiPheeter			1.75		
Diphenylmethane	Diphemetan			1.35		
Diisopropylamine	diPrAm					0.70
Disulfoton	Disulfoton		4.14			
1,2-Dichloroethane	DKE12	-1.05	-0.74		-0.64	
1,1-Dichloroethylene	DKeten11				0.12	
1,2-Dichloroethylene	DKeten12		-0.25		-0.16	
Dichloromethane	DKM	-1.53	-1.39	-0.79	-0.41	
1,2-Dichloropropane	DKP12		0.11		-0.39	
1,1-Dichloropropane	DKPropan11				0.06	
1,3-Dichloropropane	DKPropen13				1.26	
2,4-Dimethylaniline	DMA24	0.87		-0.21		
Dimethylphthalate	DMftalat				0.59	
2,3-Dimethylphenol	DMP23		0.96			
2,4-Dimethylphenol	DMP24		1.02		1.20	
2,5-Dimethylphenol	DMP25		1.03			
2,6-Dimethylphenol	DMP26		0.93			
3,4-Dimethylphenol	DMP34		0.81			
3,5-Dimethylphenol	DMP35		0.74			
2,4-Dinitro-6-methylphenol	DN24M6P				2.94	
1,3-Dinitrobenzene	DNB13	0.72		1.23		
3,5-Dinitrobenzoic acid	DNB35	-0.28				
2,4-Dinitrophenol	DNP24		1.41		2.47	
2,5-Dinitrophenol	DNP25		1.34			
2,3-Dinitrotoluene	DNT23				2.74	
2,4-Dinitrotoluene	DNT24	0.45				
Dodecylbenzene	Dodebensen			-0.51		
Dodecanol	DodeOH1	3.79				
1,2-Diphenylhydrazine	DP12hydrazin				2.83	
2-Ethyl-1-hexanol	E2HxOH1		0.70			

Name	Short name	Microtox pEC-50	Daphnia pEC-50	Leuciscus pLC50	Lepomis pLC50	Alga pEC-50
EDTA	EDTA			-0.74		
Endrin	Endrin		3.58			
Ethanolamine	Etanolamin			-0.93		
Ethylbenzene	EtB	1.14		0.38	-0.15	
Ethylenediamine	Etylendiamin	-2.47		-0.83		
Diethylether	eter	-1.88	-1.16	-1.58		
Ethylacetate	EtOAc			-0.58		
Ethanol	EtOH	-2.91	-2.34			
2-Ethylhexylamine	Etyl2hexylami			0.91		
Ethylbutyrate	Etylbutyrat			0.17		
Ethylglycolacetate	Etylglykolacet			-0.03		
Ethylpropionate	Etylpropionat			-0.09		
Acetic acid, bromo-, 2-butene-1,4-diyl ester (Fennosan F50)	FeF50					3.99
Fenitrothione	Fenitrotion		4.49			
Floroglucinol	FloGlu		0.00			
Fluoranthene	Fluoranten				1.70	
Furaltadone[10]	Furaltad					1.44
Furfurylalcohol	Furfurylalkohol			-1.14		
1-Heptene	Hepten-n-1			-0.31		
Hexachlorobutadiene	Hexaklorbutadien			-0.26		
Hexachloroethane	HKEtan				2.38	
Acetic acid	HOAc		-0.40	-0.83		
1-Heptanol	HpOH1	0.92	0.15	0.45		
1-Hexanol	HxOH1	0.47	-0.29	-0.11		
2-Hexanol	HxOH2			-0.45		
3-Hexanol	HxOH3			-0.54		
Hydroquinone	Hydrokinon			2.87		
Hydroquinone monomethylether	HydrokinonOMe			0.58		
Diethanolamine	Iminodieta		-0.44			
Isoamylalcohol	IsAmOH		-0.60			

[10] (5-morpholinomethyl-3-(5-nitrofurfurylideneamino)-2-oxazolidinone)

Name	Short name	Microtox pEC-50	Daphnia pEC-50	Leuciscus pLC50	Lepomis pLC50	Alga pEC-50
Isobutylacetate	IsobutylOAc			-0.03		
Isobutyronitrile	Isobutyronitril			-0.55		
Iso-octanol	IsoocOH			0.73		
Isophorone	Isophorone				-0.20	
Isopropylbenzene	Isoprobensen			0.41		
Isopropylacetate	IsoproOAc			-0.55		
Isopropylacetone	Isopropylacet			-0.87		
4-Chloro-6-methylphenol	K4M6P				1.79	
2-Methylphenol	Kresol2		0.78	0.76		
3-Methylphenol	Kresol3	1.16	0.75			
4-Methylphenol	Kresol4		0.94			
Dipentene (dl-Limonene)	Limonen-dl			0.60		
Lindane	Lindan	1.42	2.46	3.02		
Malathione	Malation		5.52			
4-Chloroaniline	MCA4		1.59			
Chlorobenzene	MCB		0.97	3.75	0.85	0.95
2-Methylcyclohexanone	Me2Cyhexano			-0.62		
2-Methylfuran	Me2Furan			-0.46		
3-Methyl-4-chlorophenol	Me3Cl4P		1.50			
5-Methyl-2-hexanone	Me5Hexanon2			-0.16		
Methylacrylate	MeAcrylat			1.06		
Methylmethacrylate	MeMetakrylat			-0.54		
Methanol	MeOH	-3.26	-2.63			
2-Methyl-1-propanol	MePOH21	-1.22		-1.31		
Methylpropionate	MePropionat			-0.34		
4-Nitroaniline	MNA4		1.15	0.60		
2-Nitrophenol	MNP2		0.50			
3-Nitrophenol	MNP3		0.80			
4-Nitrophenol	MNP4		1.09		1.22	
2-Nitrotoluene	MNT2	1.87	0.93	0.67		
4-Nitrotoluene	MNT4	1.10	1.18			
Monofluoroacetic acid	MonofluorAc			-0.55		
Monolinuron	Monolinuron			0.46		
Morpholine	Morfolin		-0.06	-0.51		0.49

Name	Short name	Microtox pEC-50	Daphnia pEC-50	Leuciscus pLC50	Lepomis pLC50	Alga pEC-50
Nitrobenzene	NB	0.62		0.31	0.46	
Nitrofurazone	NO2Fuzo					2.14
Nonanol	NoOH1	2.01				
Nitrilotriacetic acid	NTA	-0.72		-0.40		
1-Octanol	OcOH1	1.58	0.81	0.81		
Octafonium chloride[11]	octafon					2.96
Oxalic acid	Oxalsyra			-0.56		
Pentachlorobenzene	PClbensen				3.00	
Pentachlorophenol	PCP	2.51	2.55			2.80
Pentadecanol	PedeOH1	0.25				
2,4-Pentanedione	Pedion24	-0.27		-0.06		
1-Pentanol	PeOH1	-0.66	-0.68	-0.74		
3-Pentanol	PeOH3	-1.23	-0.60			
Phenol	Phenol	0.44	0.40	0.58		-0.20
Phenylacetate	PheOAc			1.02		
Pentachloroethane	PKEtan				1.45	
p,p'-DDT	ppDDT	1.39		3.25		
Propylacetate	ProOAc			-0.28		
1-Propanol	ProOH1	-2.21	-1.87	-1.88		
2-Propanol	ProOH2			-2.17		
Propargylalcohol[12]	PropynOH			1.47		
Pyridine	Pyridin	-0.47		-0.48		
Pyrogallol	PyroGa		0.74			
Resorcinol	Resorc		0.01			
Salicylaldehyde	Salicylaldehyd			1.47		
Dodecylsulfat	SDS	2.28		1.12		
Styrene	Styren			0.79		
Tartaric acid	TartCOOH		0.05			
t-Butylamine	tBuAm					0.66
1,2,3-Trichlorobenzene	TCB123					2.30
1,2,4-Trichlorobenzene	TCB124		1.96		1.73	2.11

[11] (Octaphen) (N,N-diethyl-N-(2-(4-(1,1,3,3-tetramethylbutyl)phenoxy)ethyl)benzene methanamine chloride)
[12] (2-Propyn-1-ol)

Name	Short name	Microtox pEC-50	Daphnia pEC-50	Leuciscus pLC50	Lepomis pLC50	Alga pEC-50
3,4,5-Tricholorcatechol	TCCat345					2.92
3,4,6-Tricholorcatechol	TCCat346					3.04
3,4,5-Trichloro-2,6-dimethoxyphenol	TCDMP26					2.48
3,4,5-Trichloroguaiacol	TCGu345					2.48
4,5,6-Trichloroguaiacol	TCGu456					2.70
1,2,3-Trichlorobenzene	TCP123	1.86				
1,2,4-Trichlorobenzene	TCP124	1.69				
1,3,5-Trichlorobenzene	TCP135	1.11				
2,3,4-Trichlorophenol	TCP234	2.09	1.95			1.99
2,3,5-Trichlorophenol	TCP235	2.16	1.94			
2,3,6-Trichlorophenol	TCP236	1.17	1.43			
2,4,5-Trichlorophenol	TCP245	2.21	1.98		2.64	
2,4,6-Trichlorophenol	TCP246	1.38	1.56		2.79	1.75
3,4,5-Trichlorophenol	TCP345	2.71	2.35			
Tebuthiuron[13]	Tebuthon					3.27
1,2,3,4-Tetrachlorobenzene	TeCB1234	2.06				
1,2,3,5-Tetrachlorobenzene	TeCB1235	1.79			1.53	
1,2,4,5-Tetrachlorobenzene	TeCB1245	1.52			2.13	
Tetrachlorocatechol	TeCCat					3.49
Tetrachloroguaiacol	TeCGu					2.82
2,3,4,5-Tetrachlorophenol	TeCP2345	3.05	2.12			
2,3,4,6-Tetrachlorophenol	TeCP2346	2.20			3.22	2.25
2,3,5,6-Tetrachlorophenol	TeCP2356	1.96	2.01		3.13	
Tetradecanol	TedeOH1	2.58				
Tetrahydrofuran	TeHyFuran			-1.59		
Tetrachloroethene	TeKE	0.93	1.71		1.11	
1,1,1,2-Tetrachloroethane	TeKEtan1112				0.92	
1,1,2,2- Tetrachloroethane	TeKEtan1122				0.90	
1,2,4,5-Tetramethylbenzene	TeMebensen1245			0.65		

[13] (N-(5-(1,1-dimethylethyl)-1,3,4-thiadiazol-2-yl)-N,N'-dimethylurea)

Name	Short name	Microtox pEC-50	Daphnia pEC-50	Leuciscus pLC50	Lepomis pLC50	Alga pEC-50
Thiram	Thiram		3.11			
Trichloroethene	TKE	-0.16	0.24	-0.01	0.47	
1,1,1-Trichloroethane	TKEtan111				0.27	
1,1,2-Trichloroethane	TKEtan112				0.52	
2,3,5-Trimethylphenol	TMP235		0.71			
2,3,6-Trimethylphenol	TMP236		0.85			
2,4,6-Trimethylphenol	TMP246		0.68			
2,4,6-Trinitrophenol	TnP246		0.37		0.13	
Toluene	Toluene	0.71	0.57	0.12	0.85	
2-Toluidine	Toluidin2			-0.04		
Tridecanoic acid	TrideCOOH		1.25			
Tridecanol	TrideOH1	3.57				
2,4,5- Trichloro-phenoxyacetic acid	Triklor-2,4,5 -fenoxisyra			-0.31		
Tri-n-butylphosphate	Tri-n-BuP			1.54		
Undecanol	UnOH1	3.14				
Vinyl acetate	VinylOAc		0.22	0.52		
1,2-Xylene	Xylen12	1.06				

IVL Svenska Miljöinstitutet AB

P.O.Box 210 60, SE-100 31 Stockholm
Hälsingegatan 43, Stockholm
Tel: +46 8 598 563 00
Fax: +46 8 598 563 90

IVL Swedish Environmental Research Institute Ltd

P.O.Box 470 86, SE-402 58 Göteborg
Dagjämningsgatan 1, Göteborg
Tel: +46 31 725 62 00
Fax: +46 31 725 62 90

Aneboda, SE-360 30 Lammhult
Aneboda, Lammhult
Tel: +46 472 26 77 80
Fax: +46 472 26 77 90

www.ivl.se