A Strategy for Ranking Environmentally Occurring Chemicals

by

Lennart Eriksson

Akademisk avhandling

Som med tillstånd av rektorsämbetet vid Umeå Universitet för erhållande av Filosofie Doktorsexamen vid Matematisk-Naturvetenskapliga fakulteten, framlägges till offentlig granskning vid Kemiska Institutionen, hörsal B, Lu 0, fredagen den 7:e juni 1991, kl. 10.00.
Title: A Strategy for Ranking Environmentally Occurring Chemicals

Author: Lennart Eriksson, Research Group for Chemometrics, Department of Organic Chemistry, University of Umeå, S-901 87 Umeå, Sweden.

Abstract: A systematic methodology for quantitative structure-activity relationship (QSAR) development in environmental toxicology is provided. The methodology is summarized in a strategy with six sequential steps.

The strategy rests on two cornerstones, namely (1) the use of statistical design to select a series of representative compounds (the so-called training set) on which to base a QSAR, and (2) the multivariate modelling of the relationship between physico-chemical and biological properties of the training set compounds.

The first step of the strategy is the division of chemicals into classes of structurally similar compounds. Briefly, steps 2 to 6 are: (2) physico-chemical and structural characterization of the compounds in a class, (3) selection of a training set of representative compounds, (4) biological testing of the selected training set, (5) QSAR model development, and (6) experimental validation of the QSAR and predictions for non-tested compounds.

The thesis summarizes the results obtained from the application of the strategy to the class of halogenated aliphatic compounds. Biological measurements were made in four biological test systems, reflecting acute toxicity, mutagenicity, relative cytotoxicity and genotoxicity. QSARs were developed relating each biological endpoint to the structural descriptors of the compounds. Multivariate PLS modelling was used in the data analysis. The developed QSARs were used for predicting the biological activity pattern of the non-tested compounds in the class. These predictions may be used as a starting point for a priority ranking for further biological testing of these compounds.

The strategy has not been developed solely for establishing QSARs for the halogenated aliphatics class. On the contrary, this work is intended to demonstrate a generally applicable QSAR methodology.

Keywords: QSAR, hazard ranking, statistical design, multivariate modelling, halogenated aliphatics, PCA, PLS.
A Strategy for Ranking Environmentally Occurring Chemicals

by

Lennart Eriksson

Akademisk avhandling

Som med tillstånd av rektorsämbetet vid Umeå Universitet för erhållande av Filosofie Doktorsexamen vid Matematisk-Naturvetenskapliga fakulteten, framlägges till offentlig granskning vid Kemiska Institutionen, hörsal B, Lu 0, fredagen den 7:e juni 1991, kl. 10.00.
Title: A Strategy for Ranking Environmentally Occurring Chemicals

Author: Lennart Eriksson, Research Group for Chemometrics, Department of Organic Chemistry, University of Umeå, S-901 87 Umeå, Sweden.

Abstract: A systematic methodology for quantitative structure-activity relationship (QSAR) development in environmental toxicology is provided. The methodology is summarized in a strategy with six sequential steps.

The strategy rests on two cornerstones, namely (1) the use of statistical design to select a series of representative compounds (the so-called training set) on which to base a QSAR, and (2) the multivariate modelling of the relationship between physico-chemical and biological properties of the training set compounds.

The first step of the strategy is the division of chemicals into classes of structurally similar compounds. Briefly, steps 2 to 6 are: (2) physico-chemical and structural characterization of the compounds in a class, (3) selection of a training set of representative compounds, (4) biological testing of the selected training set, (5) QSAR model development, and (6) experimental validation of the QSAR and predictions for non-tested compounds.

The thesis summarizes the results obtained from the application of the strategy to the class of halogenated aliphatic compounds. Biological measurements were made in four biological test systems, reflecting acute toxicity, mutagenicity, relative cytotoxicity and genotoxicity. QSARs were developed relating each biological endpoint to the structural descriptors of the compounds. Multivariate PLS modelling was used in the data analysis. The developed QSARs were used for predicting the biological activity pattern of the non-tested compounds in the class. These predictions may be used as a starting point for a priority ranking for further biological testing of these compounds.

The strategy has not been developed solely for establishing QSARs for the halogenated aliphatics class. On the contrary, this work is intended to demonstrate a generally applicable QSAR methodology.

Keywords: QSAR, hazard ranking, statistical design, multivariate modelling, halogenated aliphatics, PCA, PLS.

ISBN 91 - 7174 - 577 - 7
Contents

1 List of papers ..................................................................................................................7

2 Introduction ........................................................................................................................9
   2.1 An environmental perspective .................................................................................9
   2.2 A historical perspective .........................................................................................11
   2.3 A multivariate analogy perspective .......................................................................12
   2.4 An experimental design perspective ......................................................................13
   2.5 Aims and scope ......................................................................................................14

3 The strategy and a summary of the results ...............................................................17
   3.1 Division of chemicals into classes (step 1) .........................................................17
      3.1.1 Structurally similar compounds .....................................................................17
      3.1.2 High volume chemicals ...............................................................................17
      3.1.3 The halogenated aliphatics class ..................................................................18

   3.2 Structural description and definition of design variables (step 2) ..................20
      3.2.1 Structural description ....................................................................................20
      3.2.2 The need for statistical design .......................................................................22
      3.2.3 Principal component analysis .......................................................................23
      3.2.4 Definition of design variables .......................................................................26

   3.3 Selection of a training set of compounds (step 3) .............................................28
      3.3.1 Some remarks on statistical design ...............................................................28
      3.3.2 Using statistical design to select a training set .............................................30

   3.4 Biological testing (step 4) ......................................................................................32
      3.4.1 Multivariate data ...........................................................................................32
      3.4.2 The measured biological data .......................................................................32
      3.4.3 Dose-response curves ...................................................................................33
3.5 Additional structural description and QSAR model development (step 5) ........................................................................................................... 35
  3.5.1 The two-fold objective ..................................................................... 35
  3.5.2 Chemical model systems for the AX-class ..................................... 35
  3.5.3 Modelling the relation between X and Y ..................................... 37
  3.5.4 The PLS method ........................................................................... 37
  3.5.5 QSAR model development ........................................................... 40

3.6 Validations and predictions (step 6) .................................................. 44
  3.6.1 Experimental validation ............................................................... 44
  3.6.2 Using statistical design to select a validation set ........................ 44
  3.6.3 Predictions for non-tested compounds .................................... 45

4 An overview of the twelve endpoints .................................................... 47
  4.1 A principal component analysis ......................................................... 47
  4.2 Interpretation of the loadings ............................................................ 47
  4.3 Interpretation of the scores ................................................................. 49
  4.4 An overall ranking of the halogenated aliphatics ............................... 51
  4.5 A PLS analysis ................................................................................. 52
  4.6 Some possible causes to the biological observations ........................ 53

5 Summary and conclusions ...................................................................... 55
  5.1 Some comments on the strategy ....................................................... 55
  5.2 The two cornerstones of the strategy ............................................... 56
  5.3 An additional benefit of using design ............................................. 57
  5.4 The use of linear and quadratic terms of the descriptors .................. 57
  5.5 A future outlook .............................................................................. 58

6 Acknowledgements ................................................................................ 59

7 References ............................................................................................ 61

8 Appendices ............................................................................................ 67
LIST OF PAPERS

This thesis is based on the following papers and will be referred to by the Roman numerals I-VI.


2 INTRODUCTION

2.1 An environmental perspective

In 1945, the Chemical Abstract Services listed almost 1 million individual chemical compounds, but by 1988 this figure had risen to more than 8 million substances that have actually been made or isolated [1]. This rapid increase is still on-going and shows no signs of tapering off in the foreseeable future. Although the vast majority of known chemicals never reach the market nor are released to the environment, there is still a substantial number of compounds in common use. Attempts made to estimate the number of these everyday chemicals have produced figures in the range 20,000 - 70,000 [1,2], and it is believed that between 500 and 1000 new ones are added each year. In addition, there are also many products and mixtures of varying composition routinely being handled. Some of these mixtures are well-defined blends of chemicals, whereas others, like most plant effluents, are poorly known or characterized.

Being surrounded by chemicals, it is easy to get the impression that we are trapped in a chemical jungle, with neither overview nor sense of direction in which to move for control and regulation. At the same time, our increasing knowledge of the acute and long-term biological effects of chemicals both to ourselves and to the environment, tells us that the need for biological testing of environmentally occurring compounds is urgent [2]. However, a direct biological testing of all existing chemicals is not feasible. It would lock test resources for decades and also be contrary to the interests of animal welfare. Furthermore, "new" biological effects are likely to be discovered for many compounds, which in turn would require that previously tested compounds be reexamined in new biological test systems to study these effects.

Thus, to save time, money and research animals, it is of great interest to develop efficient means of predicting biological effects of non-tested compounds. One way to accomplish the estimation of the potential hazards of environmental chemicals is to construct mathematical models, so-called
quantitative structure-activity relationships (QSARs). These relate the variation in biological activity within a series of compounds to their chemical and structural properties [3].

Typically, the development of a QSAR can be formalized as follows: For a certain set of compounds appropriate biological activity variables are measured. Together these data form the \((N \times M)\) biological data matrix \(Y\), with \(N\) being the number of compounds in the training set (the compounds used to calculate the QSAR) and \(M\) being the number of biological activity variables. Moreover, the chemical and structural properties of the compounds are translated to numbers. These data form the \((N \times K)\) chemical descriptor matrix \(X\), where \(N\) is the same as above and \(K\) the number of descriptor variables. The biological data \((Y)\) are then "modelled" by the chemical data \((X)\) in terms of

\[
Y = F(X; \beta) + E
\]  

where \(F(X; \beta)\) represents the systematic part of the data, \(E\) the residuals (measurement errors, model imperfections, etc.) and \(\beta\) the regression coefficients (see also Figure 1). If a model works, the chemical data for new compounds may be inserted and the biological activities of these compounds predicted.

![Figure 1. A schematic representation of QSAR.](image-url)
A QSAR thus informs how a certain biological activity variable varies as a function of the variation in chemical properties, and particularly which chemical properties correlate with a certain change in a biological response. Provided that it is properly developed and validated, a QSAR may be a useful screening facility in an environmental application, especially for setting priorities of where to concentrate further research and test efforts.

2.2 A historical perspective

The origin of structure-activity modelling dates back to the 1860s [4,5], when Crum Brown and Fraser [6] first suggested that a biological activity could depend on a physico-chemical property. Just before the turn of the century, Overton [7] and Meyer [8] extended this work and found relations between the narcotic activity of a series of compounds and their hydrophobicity. The general breakthrough, however, for QSAR as a discipline came with the pioneering work of Hansch and coworkers in the 1960s [9]. They were able to demonstrate how QSARs could be formulated for series of similar compounds, using model systems reflecting hydrophobicity and electronic properties. At that time, the QSAR methodology was mainly developed and employed in the areas of drug design and pesticide research. It was not until the 1970s that the QSAR concept started to be applied in the environmental sciences [2]. Spurred by the sudden awareness of all the chemicals being released to the environment, it became popular to use QSARs for trying to predict biological activity, fate, exposure and other outcomes of environmental interest. During the 1980s the use of QSARs in environmental toxicology steadily increased and broadened, and today is a well-established branch of the QSAR research field.

Two recent review articles, one with the emphasis on giving a critical assessment of QSARs in toxicology and ecotoxicology [10], and one focussed on QSARs for environmental pollutants [11], provide good insight into the present status of QSAR for risk assessment of environmental chemicals. Over the years many different types of physico-chemical descriptors have been used, the most popular ones being Hansch’s log P [12], Hammet’s $\sigma$ [13] and Taft’s $E_\alpha$ [14]. These and many others have been used for correla-
ting chemical structure changes to various biological activities (acute toxicity, mutagenicity, carcinogenicity, teratogenicity, etc.) and fate processes (uptake, bioconcentration, biodegradation, etc.). However, surprisingly often only one or a few of these chemical property descriptors are used in constructing a QSAR, which is unfortunate in terms of the informational inefficiency that may result from the limited data used in such cases.

2.3 A multivariate analogy perspective

The goal in any structure-activity modelling is to find the mathematical expression that best describes the relationship between physico-chemical descriptors and biological responses. To adequately describe the often complex nature of many biological systems, it is necessary to utilize a battery of as many relevant physico-chemical descriptors as possible. This view, the multivariate analogy approach to QSAR modelling, introduced by Wold, Dunn, Sjöström and Hellberg [15-17], assumes that the factors governing the events in a biological test system are represented in the multitude of physico-chemical descriptors characterizing the compounds. In other words, within a series of compounds it is assumed that a small change in chemical structure will be accompanied by an analogous small change in biological activity, and that the multivariate physico-chemical description will reveal these analogies.

Analogy models can be regarded as linearizations of the "real", complicated relationships between chemical properties and biological responses. Wold and Dunn [17] have shown that such analogy models typically have local validity only, that is, can encompass only compounds fairly similar in structure, and showing commonality in biological mechanisms. Thus, a QSAR should be based on a series of chemically and biologically similar compounds. It is noted, however, that the compounds must be dissimilar enough to cause some systematic change in biological activity.

Besides being multivariate, QSAR data often are crude, imprecise and strongly collinear [18]. This implies that traditional statistical techniques, like multiple linear regression (MLR), that assume the physico-chemical descriptors to be exact, 100% relevant, and independent, will not always work well [19]. Since the majority of the QSARs for environmental chemicals
so far published are based on MLR, some caution in interpreting some of these results may be warranted. However, it should be stressed that examples of well founded MLR QSARs are found in the literature [20].

In concluding this paragraph, it is emphasized that in situations where many and strongly collinear physico-chemical descriptors and/or biological responses operate together, data analytical methods other than the traditional MLR techniques must be utilized. All data analysis in this thesis has been conducted with multivariate projection methods, namely principal component analysis (PCA) [21] and partial least squares projections to latent structures (PLS) [21,22].

2.4 An experimental design perspective

An often overlooked step in the QSAR research field is the selection of the compounds on which to base the model, i.e. how to design the so-called training set. This ignorance may result in unbalanced test series, which in turn give rise to QSARs of poor quality [17,23]. A QSAR critically depends on the selection of the training set. The training set compounds must be representative for the class of compounds from which they originate [24], that is, they must be chosen in such a way that they efficiently span the physico-chemical and structural domain of the class. One efficient way to accomplish this is to utilize some kind of experimental plan, such as factorial or fractional factorial designs [25,26], which are informationally efficient schemes facilitating the selection of a pertinent training set.
2.5 Aims and scope

To provide a statistically sound framework for environmental QSARs, a new strategy has been developed. This six-step strategy (see Figure 2) rests on two cornerstones, (1) the statistically designed selection of a training set of compounds, and (2) the multivariate modelling of the relationships between physico-chemical descriptors and biological responses. This thesis outlines the strategy and presents the application to the class of halogenated aliphatic hydrocarbons. The strategy was presented in paper I. However, during the course of the project the strategy has been subjected to minor adjustments. The modified strategy was presented in paper III and adhered to in the following papers. The summary below follows the same outline.

Figure 2. The six steps of the strategy.
The thesis introduces a systematized methodology for QSAR development in environmental toxicology. The methodology is based on some important principles and concepts, which are summarized below (cf. Figure 2):

1) **Division of chemicals into classes.** Since the mechanism of biological action differs between different types (classes) of compounds, one cannot construct models relating structure to activity that apply to structurally diverse compounds [17]. Thus, the present approach is linked to the division of chemicals into classes of structurally similar compounds.

2) **Multivariate structural description.** Traditionally, a QSAR is based on rather few (say, two or three) chemical descriptors, either arbitrary chosen or selected by a step-wise procedure from a larger set of descriptors. However, the intrinsic properties of chemical compounds is best captured by a multitude of chemical descriptors. Accordingly, the use of a multivariate chemical and structural description is advocated.

3) **The use of statistical design to select a balanced test series (training set).** Commonly, chemical and biological data are collected directly from the literature or drawn from data bases. Such data, however, often have an unbalanced distribution. They may come from various sources and thereby vary in precision and reliability. Moreover, there may be many measurements for a few compounds of central interest, and only scanty observations for the remaining species in a series of compounds. Thus, there is a need for creating balanced and representative training sets. This can be accomplished by using statistical design.

4) **A broad biological testing on a small series of representative compounds.** The common way of laying out the biological testing is to conduct one or a few biological tests on a large series of chemicals, or, the reverse, lots of tests on a few, arbitrarily chosen compounds. However, instead of testing randomly selected compounds in arbitrary biological test systems, it is informationally optimal to concentrate the test resources on a small series of representative compounds. This will give a broad and stable picture of their biological properties. Besides money, this also saves time and research animals.
5) **Multivariate modelling of the relationship between structure and activity.** In general, QSARs in environmental toxicology are developed by means of multiple linear regression (MLR). MLR suffers from the disadvantages that it cannot cope with missing data and can only treat one biological response at a time. These limitations can be overcome by the use of the multivariate PLS method.

6) **Experimental validation of a QSAR.** It is sometimes argued that a QSAR is constructed only for understanding the nature of a certain biological activity, and not for predictive purposes. However, QSARs may well be used for predictive purposes, provided that the predictive capability of a QSAR has been experimentally assessed on a set of validation compounds.
3 THE STRATEGY AND A SUMMARY OF THE RESULTS

3.1 Division of chemicals into classes (step 1)

3.1.1 Structurally similar compounds

The first step of the strategy is the division of chemicals into classes on the basis of their chemical structure. This step must be carried out carefully, so that one does not violate the condition of local validity when attempting to formulate a QSAR. Thus, the ideal situation corresponds to classes where within each class all the chemicals are structurally similar and act according to the same biological mechanism. In reality, this is difficult to achieve and some deviations from ideality are to be expected.

Another consideration is the question of which chemical data source to use. The potential hazard of a chemical is a combination of its intrinsic properties and the extent to which it is released to the environment (exposure). Thus, the task is to find a relevant data base providing this information on the chemicals of concern.

3.1.2 High volume chemicals

The United States Environmental Protection Agency (USEPA) has ranked some 55,000 chemicals and simple mixtures in their 1977 inventory according to production volume. The top 1,000 chemicals accounted for almost 99% of the total production volume of all compounds, i.e. the remaining 54,000 chemicals only accounted for about 1% of this volume [2]. Thus, though the production volume is not the best estimate of exposure, it is still a reasonable first approximation of which chemicals may be encountered in our surroundings.

In a preliminary study [27], the practicality of using high production volume lists as starting points for dividing chemicals into classes, was tested. The USEPA top 1,000 list, supplemented with the corresponding top 229 list of Sweden, was chosen. Paper I briefly reports the results from this initial study. After removal of mixtures (such as petroleum blends), inorganic salts
and structurally diverse insecticides, a total of 458 compounds remained. On the basis of their structural features, backbone and functional groups, these compounds were sorted into appropriate classes. In the end, 408 (or 89%) of these substances had been divided into 56 classes, each maximally containing about 40 compounds.

3.1.3 The halogenated aliphatics class

In paper I, one of the 56 classes, the one containing halogenated aliphatic hydrocarbons with 1 to 4 carbon atoms, was selected for further studies. This class was chosen for several reasons. It comprises many compounds of environmental concern, such as CFCs (chlorinated fluorocarbons) and chlorinated organic solvents, and was therefore judged to be of significance from a regulatory point of view. Moreover, the description of the physico-chemical properties of these compounds was rather straightforward, since many well established QSAR descriptors could be extracted from the literature.

Originally, in the theoretical example investigated in paper I, the halogenated aliphatics class included both saturated and unsaturated compounds. However, before the strategy was ever tested in reality, it was decided to further subdivide this class by separating the saturated and unsaturated chemicals from each other. This was done because it was anticipated (see e.g. [28]) that unsaturated compounds might exhibit mechanisms of biological action different from their saturated analogues. Thus, when put into practise (paper III and onwards), the strategy was applied to a class consisting solely of saturated compounds. This class, the AX-class [27], comprised 58 chemicals (see Table 1) of which 23 originated from the two high volume lists initially scrutinized. The remaining 35 compounds, selected on the basis of the availability of physico-chemical descriptors, were added to broaden and stabilize the class.
Step 1 of the strategy was performed in a slightly different manner in paper II. The aim of this paper was to illustrate the whole sequence of steps involved in the strategy and gain experience prior to our own measurements on the AX-class. To fulfil this objective, compounds were selected for which biological data were found in the literature [29]. To achieve stable classes, the two sets of compounds, chlorinated aliphatics and aliphatic alcohols, also contained some additional compounds not biologically tested.
3.2 Structural description and definition of design variables (step 2)

3.2.1 Structural description

After dividing chemicals into classes and choosing which class to work with, the next question is how to appropriately describe the structural and physico-chemical variation of the selected chemicals. Obviously, the demands on the physico-chemical description depend not only on the considered chemicals, but also on the nature of the biological system under investigation. In general, the more complicated an observed system is, the more unlikely it is that a single descriptor variable will contain sufficient information about a given biological problem. Thus, the structural and physico-chemical description is multivariate, but to what extent varies from case to case. Depending on the application, the type and number of descriptor variables included is different and is chosen as deemed appropriate.

Often the number of chemicals in a class is rather large, above 50 or 60, which makes the compilation of some types of descriptor variables difficult. Boiling points, melting points and similar tabulated data, are usually always available, but descriptors derived from chemical model reactions or spectroscopic measurements are less abundant. Thus, in reality, the choice of the training set compounds must be based on a somewhat restricted (though multivariate) set of descriptors. However, at later stages of the strategy (step 5), when the training set has been selected and the considered compounds are fewer, additional description of the few training set compounds may enrich and supplement the structural data.

According to Dunn [30], structural and physico-chemical descriptor variables can be categorized into two groups: (1) global types and (2) substituent types. Global variables, like log P, are based on the whole molecule, whereas substituent descriptors are connected to a certain part or moiety of a molecule. Depending on the application, the two groups of variables can be used independently or in conjunction with each other. Regardless of which type of variable, global or substituent, is chosen, it is
often difficult to predict in advance exactly which descriptor variables will be useful. If no prior knowledge or information exists about the importance of certain factors, it is usually recommended that at least the hydrophobic, steric and electronic properties of the compounds are described [16].

To parametrize the structural variation among the chemicals considered in this thesis, both global and substituent descriptors have been used. In paper III, for example, out of a total of thirteen descriptors used to characterize the 58 saturated halogenated aliphatics, eight variables were global and five were indicator (substituent) variables (see Table 2). Among the variables in Table 2, molecular weight and van der Waals volume are linked to the size (steric effect), whereas log P accounts for hydrophobicity and ionization potential reflects electronic properties. The remaining structural descriptors also contribute to delineating these factors, but may account for other phenomena as well.

### Table 2. The descriptors used to characterize the 58 halogenated aliphatics.

| 1) Molecular weight | 8) Ionization potential |
| 2) Boiling point     | 9) Number of carbons   |
| 3) Melting point     | 10) * bromines         |
| 4) Density           | 11) * chlorines        |
| 5) Refractive index  | 12) * fluorines        |
| 6) van der Waals volume | 13) * iodines      |
| 7) log P             |                      |

Essentially, the descriptor variables listed in Table 2 are the same throughout the whole series of papers (I-VI), with minor deviations in paper II. In this second part, three of the indicator variables were excluded because only chlorinated compounds were considered. The 15 aliphatic alcohols treated in the same paper were characterized with the same basic set of descriptors. However, some additional descriptors, specific for the alcohols, were also included (see paper II).
3.2.2 The need for statistical design

The structural description provides the basis for the third step of the strategy, the selection of the training set. This training set must not be selected by starting with a primary chemical and then changing one structural factor (descriptor) at a time to yield the other compounds. Such a COST-design is informationally inefficient [31] in that it does not allow interactions between the structural factors to be studied, that is, the joint influence of descriptor variables on biological activity (see Figure 3). Instead of using the COST-approach, all relevant structural factors should be varied simultaneously according to some experimental protocol (see Figure 3). This introduces systematic variation of several factors concurrently and allows the detection of interactions.

Figure 3. The COST approach (left): The vertical line corresponds to holding factor A constant while varying B. One finds the "optimal" value of B. Next, factor B is kept constant while varying A (the horizontal line). The factorial design approach (right): The two factors A and B are changed simultaneously. This provides information about the direction in which to move for finding the real optimum, as indicated by the arrow.

With the factorial design approach, one obtains a better mapping of the variation in the biological response(s), than using the COST-approach. This is accomplished because the structural factor domain of the chemicals is spanned in a balanced and more efficient way.
Firstly, the relevant structural factors, the so-called design variables, are specified. These may, for instance, be derived from a multivariate analysis of the battery of structural descriptors, which reveals the dominant, latent patterns ("factors") in the data (see section 3.2.3). Alternatively, in simple and well-understood situations, independent principal variables known to exert an influence on the biological activity under study, may be used as design variables. Such variables may, for example, be hydrophobicity (log P) or ionization potential.

In this thesis, multivariate analysis of the structural descriptors in step 2 of the strategy has been carried out with principal components analysis (PCA) to give few underlying dimensions, the so-called principal components (PCs). The PCs are ideal as design variables because they are few and orthogonal to (independent of) each other. Before continuing, we shall therefore thoroughly discuss PCA.

### 3.2.3 Principal component analysis

Principal component analysis (PCA) is a data analytical method designed to (A) extract and (B) highlight the systematic variation inherent in a multivariate data matrix X [32,33], like the multiproperty matrix compiled for the halogenated aliphatics. In the present case X is a table of physico-chemical descriptors with N rows (N compounds) and K columns (K descriptor variables). This means that the primary objectives of PCA are (A) to evaluate the underlying dimensionality (complexity) of the data and (B) to get an overview of the dominant patterns or major trends in the data. This is accomplished by forming linear combinations of the original variables to give a few descriptive principal components. The principal components (PCs) approximate the multivariate data matrix and provide means of interpreting its general features. PCA focusses not only on relationships between variables, but also makes it possible to look at relationships between objects (here: compounds).

Operationally, PCA is a projection method, that is, it projects the multivariate data X down onto a lower dimensioned subspace. Such projections may be calculated by starting from any number of variables, and
in general they get more stable and precise the more variables are involved. The projection procedure leads to a dimensionality reduction which simplifies the interpretation of the data.

Mathematically, PCA corresponds to a decomposition of the multivariate data matrix $X$ into the product of two smaller matrices, $T$ and $P'$. In matrix notation, this is:

$$X = 1 \cdot \overline{x} + T \cdot P' + E$$  \hspace{1cm} (2)

Here, $1$ is a column vector with the element one in all positions, $\overline{x}$ a row vector with the means of all variables and $E$ the residual matrix comprising the part of the data not accounted for by the PC model. The matrices $T$ and $P'$ together capture the main features of $X$ and provide an approximation of this matrix. The matrix $T$ contains object related information (here: object = compound). Plotting its columns against each other (score plots) reveals relationships between objects. Analogously, the matrix $P'$ holds variable specific information (the so-called loadings) rendering it possible to analyze relationships between variables (here: physico-chemical descriptors or biological activities). An overview of PCA in its matrix form is given in Figure 4.
PCA may also be interpreted geometrically. Consider the fictive situation sketched in Figure 5, where a number of objects are plotted in a three-dimensional space. This point swarm might represent a series of compounds being characterized by three physico-chemical descriptor variables. When PCA is applied to this data set, the first principal component (PC 1) will orient in a manner making it account for as much of the variation among the objects as possible. Next, the second principal component (PC 2), orthogonal to the first one, will describe the second largest variation, and so on. Together, the first and second PCs define a plane. The compounds (i.e. the point swarm) can be projected onto this plane, making it constitute a two-dimensional window into the three-dimensional physico-chemical descriptor space. The variable averages ($\bar{x}$) and the loadings $p_1$ and $p_2$ give the location and direction of the plane, respectively, whereas the scores $t_1$ and $t_2$ give the position (the coordinates) of each object when projected down onto the plane. The angle between a variable and a PC can be interpreted as the arc cosine of the loading for that variable (provided that $\Sigma p^2 = 1$). Moreover, with $K$ descriptor variables an analogous situation applies, only then the compound data constitute a swarm of points in a space with $K$ dimensions.

![Figure 5. A geometrical representation of PCA. Each compound, represented by a point in the three-dimensional descriptor space, is projected onto a two-dimensional subspace (the PC-plane).](image-url)
In PCA, the scaling of the descriptor variables is important. If no prior information is available about the variables, they are scaled to unit variance (autoscaling). This scaling gives the variables the same chance to influence the model building. However, if prior knowledge exist, this may be used to influence the scaling. Several descriptor variables may for instance originate from the same data analytical method (IR, GC, etc.), or a number of variables may be of the same type (like indicator variables). In such situations a blockwise (or battery) scaling may be warranted, which gives each category of variables appropriate scaling weights such that their sum of variances is one. In this way one prevents a certain type of variables from dominating the PC model. Besides autoscaling, the data is also preprocessed by subtracting the variable averages (mean-centering).

In PCA, the significance testing is carried out by a technique called crossvalidation [34], which estimates the predictive ability of each PC. One can then select the model dimensionality that gives the best predictions.

3.2.4 Definition of design variables

Recall that the 58 halogenated aliphatic compounds (see Table 1) were characterized structurally and physico-chemically with 13 descriptor variables (see Table 2). These descriptors are more or less strongly correlated to one another. Hence, PCA is used to reduce the dimensionality by defining a few, orthogonal principal components that are utilized as design variables.

The PCA of the 58x13 data matrix is reported in paper III. A four-component model accounting for 87% of the variance in the matrix was attained. Representing the "principal properties" [18] of the halogenated aliphatics, these four PCs can be used in a statistical design to select the training set. How this is accomplished is discussed below (section 3.3). However, the PCs can also be used to form plots visualizing the systematic information in the physico-chemical description. This is done in Figure 6, where the two first and dominating PCs, explaining nearly 70% of the variation, are plotted against each other. Figure 6 thus summarizes the essential features of the descriptor matrix and provides a two-dimensional window into the 13-dimensional descriptor space. The corresponding loading
plot (Figure 7) is useful for the interpretation of these components. It is seen that PC 1 mainly is related to the size/bulk of the compounds, whereas PC 2 describes a combination of the hydrophobicity and size/bulk. The meanings of the third and fourth PCs are not as evident.

![Figure 6](image1.png) **Figure 6.** Plot of the first two PCs of the AX-class. The training set compounds are numbered and marked by filled triangles. See paper III for details and Table 1 for numbers.

![Figure 7](image2.png) **Figure 7.** Loading plot corresponding to Figure 6. For the numbering of the variables, see Table 2.

In paper II, PCA was also utilized to establish design variables. Here, however, the results were two-component PC models, indicating less complexity in the multivariate characterization. This is rather natural, since the classes contain comparatively few compounds selected on a narrow physico-chemical and structural basis.

The design variables (principal components) for the class of interest are the basis for the third step of the strategy, the selection of the training set. In this step, a statistical design is utilized to choose an appropriate combination of test compounds.
3.3 Selection of a training set of compounds (step 3)

3.3.1 Some remarks on statistical designs

Statistical designs are useful for selecting series of compounds for biological testing [35,36]. This stems from the fact that they generate the training set by introducing systematic variations in all design variables simultaneously, and not in only one variable at a time. One family of statistical designs of great practical importance is the factorial and fractional factorial designs [25,26]. These are easy to construct and modify, and the results are interpretable by means of rather simple arithmetic.

In a factorial design, each design variable (factor) is given a certain number of fixed levels, usually 2-3 depending on the purpose of the investigation, the nature (complexity) of the application and the costs of the experimentation involved.

The factorials usually are used with two levels per design variable. An example aimed at illustrating this is given in Figure 8. Suppose there is a need for studying a chemical reaction where the outcome (yield) may be expected to depend on three external factors. To evaluate the importance of each such factor, a factorial design in three design variables may be constructed. In doing this, the three design variables (designated DV₁, DV₂ and DV₃) are varied between two fixed levels, one high (+) and one low (-), giving a so-called $2^3$ factorial design. This leads to eight factor combinations in total, which geometrically can be interpreted as a cube where each experiment is situated at a corner of the cube.

With more than three design variables, full factorials give too many experiments (compounds). Then, the fractional factorial designs are of great utility [25,26]. They reduce the number of required experiments, by dividing the original factorial design into balanced fractions. A fractional factorial variant of the above example is also presented in Figure 8. In this case, the used combinations are four diagonally opposite corners of the cube. This is called a $2^{3-1}$ fractional factorial design.
Factorial and fractional factorial designs only allow linear (DV$_i$) and interaction (DV$_i$DV$_j$) terms to be estimated. However, if appropriately augmented with interior center-points, they also permit rough estimation of quadratic terms (DV$_i^2$) reflecting curvature. These designs may also be augmented to form composite designs [25,37], which allow a more rigorous quantification of curvature.
3.3.2 Using statistical design to select a training set

In the chemical reaction example just outlined, each combination of plus and minus signs in the $2^3$ factorial design represented a certain chemical experiment. When used for QSARs, a design point encodes a compound having certain structural and physico-chemical properties, rather than experimental conditions. The design is formed analogously anyway, and is executed in the design variables defined for the class of compounds.

For the class of 58 halogenated aliphatics, four design variables were derived (paper III), which corresponds to the representation of the 58 compounds in a subspace with four dimensions. The four design variables were used to construct a $2^{4-1}$ fractional factorial design encoding eight compounds. This was done by selecting compounds with values of their design variables (i.e. the score values of the principal components) matching the points specified by the design as well as possible, see Table 3. However, to avoid compounds with extreme chemical properties, species positioned at about two-thirds of the maximum or minimum values of the design variables were selected.

Table 3. The ten training set compounds of the AX-class.

<table>
<thead>
<tr>
<th>Factorial design</th>
<th>Settings in PC-Scores</th>
<th>Compound no.</th>
</tr>
</thead>
<tbody>
<tr>
<td>- - - -</td>
<td>-0.72 -1.26 -1.29 -0.51</td>
<td>(52) CH$_3$-CH$_2$-CH$_2$-CH$_2$Br</td>
</tr>
<tr>
<td>+ - - +</td>
<td>1.96 -0.86 -0.81 0.15</td>
<td>(48) CH$_3$-CHCl-CH$_3$</td>
</tr>
<tr>
<td>- + - +</td>
<td>-1.77 1.22 -0.14 -0.08</td>
<td>(33) CH$_3$-CHBr$_2$</td>
</tr>
<tr>
<td>+ + - -</td>
<td>1.20 0.89 -0.90 -0.12</td>
<td>(30) CH$_3$-HBr</td>
</tr>
<tr>
<td>- - + +</td>
<td>-1.69 -0.83 0.92 0.70</td>
<td>(15) CHCl$_2$-CHCl$_2$</td>
</tr>
<tr>
<td>+ - + -</td>
<td>1.14 -0.40 0.95 -1.10</td>
<td>(7) CCl$_3$F</td>
</tr>
<tr>
<td>- + + -</td>
<td>-3.20 1.68 1.07 -1.90</td>
<td>(39) CBr$_3$F</td>
</tr>
<tr>
<td>+ + + +</td>
<td>1.92 0.79 0.13 0.70</td>
<td>(2) CH$_2$Cl$_2$</td>
</tr>
<tr>
<td>0 0 0 0</td>
<td>0.52 0.22 0.70 0.48</td>
<td>(3) CHCl$_3$</td>
</tr>
<tr>
<td>0 0 0 0</td>
<td>0.56 0.03 0.28 1.54</td>
<td>(11) CH$_2$Cl-CH$_2$Cl</td>
</tr>
</tbody>
</table>
Ideally, with three design variables, the selected compounds should form a regular eight-corner half hypercube. However, in reality such a hypercube will never be perfectly symmetric, but rather somewhat distorted. This arises from the discrete nature of chemical compounds, rendering it difficult to find compounds exactly fitting to the design. Moreover, practical constraints (especially volatility) among the halogenated aliphatics also indirectly affected the shape of the hypercube, by reducing the candidate list of possible training set members. In view of these facts, emphasis primarily was placed on selecting compounds having PC1 and PC2 scores in agreement with the design, because these PCs explained most of the variance. However, to the greatest possible extent, acceptable matching was also sought in terms of the last two PCs.

The eight selected corner compounds were also supplemented with two center-points with near zero values in the design variables. These compounds, located in the interior of the design, were added to provide information about the curvature, and to make it possible to roughly estimate quadratic terms when needed. Thus, the whole training set for the halogenated aliphatics class comprised 10 compounds (see Table 3).

In paper II, slightly different factorial designs were used to select training sets of compounds. In both applications, a (full) factorial design in two variables \(2^2\) was utilized. However, due to a triangular distribution of the compounds in the PC score plot, two designs points were collapsed into one, making the selected set of compounds only include three corner compounds. A fourth compound was also selected to represent the variation in the middle of the design.
3.4 Biological testing (step 4)

3.4.1 Multivariate data

Above it was noted that the structural and physico-chemical description of chemicals is, by necessity, multivariate, which also holds true for the biological properties. This implies that a large number of biological measurements should be conducted, preferably at various dose levels, on several animal species, etc., so that the response matrix contains variables that span as many aspects as possible of the biological effects. Thus, in the fourth step of the strategy (biological testing), the basic idea is to subject the training set compounds to an extensive biological and toxicological evaluation, to get a broad, stable picture with ample information about their biological properties. The concentration of test resources to a few compounds also has the benefit of minimizing the biological testing.

3.4.2 The measured biological data

The training set of the AX-class (10 compounds, see Table 3) was submitted to the following biological tests: (1) acute oral toxicity to albino rat [38], (2) the micronucleus test [39], (3) the Ames test [40], (4) the DNA precipitation assay [41] and (5) the 28 day chronic toxicity to rat [42]. All the biological testing has been carried out at the Defence Research Establishment in Umeå. So far, all tests except the last one have been completed. This means that biological test data originating from systems (1) to (4) constitute the basis for the QSARs dealt with in this thesis.

The first two tests gave measures of the acute toxicities of the AX-compounds. Test (1) rendered $LD_{50}$ to rat and the micronucleus test the highest non-lethal dose, henceforth abbreviated HNLD, to mouse. Paper III in particular, but to some extent also paper IV, addresses the QSAR analyses of these two endpoints. Besides the HNLD, the micronucleus test also provided information on the mutagenicity of the compounds. However, most compounds were found to be inactive, making it impossible to establish a mutagenicity QSAR. The remaining two tests provided data relating to other
biological phenomena. The DNA precipitation assay yielded estimates of the
genotoxicity of the compounds, which was modelled in papers IV and V.
Paper VI, finally, reports the analysis of the mutagenicity data originating
from the Ames test. A summary of the endpoints is given in section 3.5.5.

In paper II, the fourth step never was carried out, because the biological
data were taken from the literature.

3.4.3 Dose-response curves

Biological measurements are commonly recorded as dose-response
curves, showing the relationship between the administered doses and the
responses they cause. Often such curves are summarized in a single value,
like LD$_{50}$, EC$_{50}$, etc. This need not be a problem if the curves are congruent
and exhibit the same general features. Then, a single value will adequately
reflect the existing information. However, if the curves are incongruent, i.e.
influenced by more than one factor, summarizing the curves as single values
may lead to information being overlooked [43,44].

Some questions of how to evaluate dose-response data from the Ames
test are discussed in paper VI. Four dose-response curves were registered for
each compound. When examining these curves, it was evident that they could
not be condensed into single values adequately portraying their attributes.
Two values were needed to sufficiently characterize the interesting parts of
the curves. One value was needed to describe the shape (steepness) of the
curve and the other was required to reflect the location along the dose axis.
This is illustrated in Figure 9, where nine dose-response curves from the
same tester strain are plotted.

Figure 9 illustrates the importance of occasionally considering more than
one parameter when evaluating biological data recorded as dose-response
curves. The use of two characteristics to describe the properties of such
curves is not unique to this application. On the contrary, this probably holds
true for many series of dose-response curves, regardless of their origin.
Moreover, when confronted with more complicated phenomena, additional
parameters beyond the two used here may be needed to supplement the
shape and location measures.
Figure 9. The distribution of the nine dose-response curves from the TA 1535 tester strain with metabolic activation. One value, the slope, is needed to describe the shape of a curve and one value - the mutagen dose giving 50% increase in the number of revertants (MD<sub>50</sub>) - the location along the dose axis. For further details, see paper VI.
3.5 Additional structural description and QSAR model development (step 5)

3.5.1 The two-fold objective

In the fifth step of the strategy the main objective is to find a suitable mathematical expression linking the chemical descriptors to the biological responses. During this procedure, information regarding the essential features in the chemical and biological data structure is obtained. There may, for instance, be a need for transforming some of the variables. Moreover, cognizance about the adequacy of the physico-chemical and structural description is gained. The QSAR analysis also gives information on whether a descriptor variable is relevant or inappropriate for a certain application.

A secondary objective of the fifth step is, if needed, to further develop the characterization of the physico-chemical and structural properties of the compounds in the training set. The purpose of the initial characterization is to found a basis for the statistical design and not necessarily to provide a complete framework for constructing QSARs. Thus, for the final QSAR development, the initial description of the chemicals may be insufficient. Hence, the goal of this additional characterization is to enrich the structural description of the training set compounds. However, it may happen that the initial characterization will, indeed, suffice in accounting for a particular biological response, such as when an endpoint of non-specific mode of action is considered. An example of the latter is found in paper II, where no additional structural description was required.

3.5.2 Chemical model systems for the AX-class

When developing chemical model systems, the idea is to find systems that mirror the factors governing the biological activity. Attention must be paid to both the type of compounds and the nature of the biological system under investigation. Paper IV reports on the extended characterization of the halogenated aliphatics. In total, eight model systems were developed, which reflect chemical reactivity and hydrophobicity (see Table 4).
Table 4. The eight chemical model systems for the AX-class.

* log (retention time) on a GC Supelco SPB-01 column (9,23)*
* log (retention time) on a GC Supelco Vocol column (10,24)*
* log (retention time) on an LC Nucleosil C-18 column (11,25)*
* log (retention time) on an LC Supelcosil C-8 column (12,26)*
* log (rate constant) of the Finkelstein reaction (13,27)*
* the relative response using a flame ionization detector (14,28)*
* log (rate constant) of the reaction with the hydroxyl radicalb
* the UV wavelength at which the absorbance exceeded 1.0 abs.unitb

a) The numbers in parentheses refer to the variable numbers the linear and quadratic terms of the model systems were given in the QSAR calculations in papers III, V and VI (see text).
b) Variables considered in paper IV, but not in papers III, V, VI and this summary.

The halogenated aliphatics may alkylate hydroxyl-, amino- and thiol-moieties present in amino acids, DNA bases and other biomolecules. Hence, it was of interest to describe the alkylating potency of these compounds. The rate constant of the Finkelstein substitution reaction was used as a model. This system resembles the reaction with 4-nitrobenzylpyridine (4-NBP), used by Hermens et al. [45] to describe the alkylating potency of 15 reactive aliphatic and aromatic halides. The second reactivity parameter included was the rate constant for the gas-phase reaction between the hydroxyl radical and an AX-compound. This variable was either taken from an extensive compilation [46] or measured [47] according to a standardized procedure. The two other reactivity parameters were the relative response using a flame ionization detector (FID) and the UV wavelength at which the specific absorbance exceeded 1.0 abs.unit. Both these latter descriptor variables relate to the electronic properties of the compounds, but the FID variable relates to the number of reducible carbons as well.

The compounds were also investigated in two gas chromatographic and two liquid chromatographic model systems. This was done in order to strengthen and stabilize the description of the hydrophobic character of the compounds.
3.5.3 Modelling the relation between X and Y

Having compiled a multitude of chemical descriptor data (matrix X) and biological activity data (matrix Y) for the training set compounds, the next phase is to develop a mathematical link between X and Y. Since the matrices are multivariate and collinear, the data analytical method must be based on projections. Partial least squares projections to latent structures, PLS (see the following section), has been used for developing all QSAR models in this thesis. PLS is not hampered by the restriction of MLR that the number of compounds must exceed the number of physico-chemical and structural descriptors.

3.5.4 The PLS method

The PLS method [21,22] correlates the systematic variation in the biological response data (matrix Y or a variable vector y) to the systematic variation in the physico-chemical descriptor data (matrix X), with the purpose of predicting Y from X. PLS operates in a way analogous to PCA, but simultaneously calculates latent variables for the two matrices plus a relation between them. PLS makes a projection of the physico-chemical descriptor data onto a lower dimensioned subspace T, and simultaneously projects the biological response data onto the same subspace. As in PCA, the data is preprocessed by means of autoscaling and mean-centering.

In PLS, each model dimension consists of the X score vector \( t \), the Y score vector \( u \), the X weight vector \( w \), the X loading vector \( p \) and the Y loading vector \( c \). The weight vector \( w \) is computed to achieve maximal correlation between \( t \) and \( u \). The matrix representation of PLS is visualized in Figure 10, together with the underlying algorithm, and the corresponding geometrical representation in Figure 11.
A Strategy for Ranking Environmentally Occurring Chemicals

$$w' = u'X/u'u$$

$$t = Xw/w'w$$

$$c' = t'Y/t't$$

$$u = Yc/c'c$$

$$p' = t'X/t't$$

$$E = X-tp'$$

$$F = Y-tc'$$

Figure 10. The matrix representation of PLS. The arrows show the order of calculations within one round of the iteration procedure. Solid parts of the arrows indicate data "participating" in the calculations and dashed parts "inactive" data. For clarity, $T$, $U$, $P'$, $C'$ and $W'$ are given as matrices, though in each iteration round only the last vector is updated.
As seen, PLS decomposes the X-matrix in a similar manner to PCA (cf. Eq. 2). The Y-matrix is modelled analogously:

\[ Y = 1 \cdot \bar{y} + U \cdot C' + F \]  

(3)

Similar to the previous expression (section 3.2.3), 1 is a column vector with ones as all elements, \( \bar{y} \) a row vector of averages and F a residual matrix. The
matrices \( U \) and \( C' \) model the systematic variation in the biological activity data matrix \( Y \).

As in PCA, the statistical significance of each model dimension is assessed by crossvalidation [34]. Once the PLS model has been developed, predictions of biological activities for other compounds (besides the training set compounds) can be made. This is accomplished by inserting their chemical descriptor data into the PLS model according to the sequence (see also Figs 10 and 11):

\[
X \Rightarrow t \Rightarrow u \Rightarrow Y
\]  

The residual matrix \( E \) of the physico-chemical descriptor data can be used to compute a tolerance interval around the model. This interval can subsequently be used for classifying new compounds as being comparable to the training set compounds or not.

For a given model dimensionality, biased regression coefficients can be calculated for each \( y \)-variable. However, when these coefficients \( b_k \) \((k=1,2,\ldots,K)\) are calculated from a PLS model with \( A \) dimensions where \( A < k \), these coefficients are not independent. Hence, individual contributions to \( y \) cannot be assessed in contrast to MLR where variables \( x_k \) and thereby regression coefficients are assumed to be independent. (This assumption is often far from fulfilled, however).

### 3.5.5 QSAR model development

Paper III presents the development of the physico-chemical and structural descriptor matrix \( (X) \) for the halogenated aliphatics. Originally, this matrix contained 13 variables (cf. Table 2). However, the QSAR analyses showed that the five indicator variables did not significantly contribute to the models, and accordingly they were omitted. To further enrich the structural description, the information from six chemical model systems was added to the \( X \)-matrix. Paper IV lists eight chemical model systems in total, but at the time when the first QSAR calculations were performed on the halogenated aliphatics class, only six were completed. Moreover, it was found that
quadratic terms of the variables were necessary for describing non-linear relations between chemical and biological data. The interaction (cross) terms, however, had no significant effect on the modelling. Thus, in summary, the X-matrix consisted of 28 (14 linear and 14 quadratic) descriptor variables. This battery remained largely unaltered in papers III, V and VI, the minor exception being paper V where only linear terms were used.

In paper III, QSARs for LD₅₀ to rat and HNLD to mouse were presented. Both models were two-dimensional and accounted for nearly 90% of the variation in biological responses. Figure 12 shows the results for one of these QSARs, the LD₅₀ one. It is seen that the training set compounds differ widely in toxicity, ranging from the practically non-toxic fluoro-trichloromethane to the severely toxic fluorotribromomethane. The corresponding PLS loading plot (strictly, this should be called the PLS weight plot since it is the columns in W that are considered) in Figure 13, gives an interpretation of the model and shows the influence of the physico-chemical and structural descriptors. In this case, it is primarily the size and hydrophobicity describing variables that are most influential.

Figure 12 (left). Results for the LD₅₀ QSAR with observed data plotted versus the calculated. For the numbering, see Table 3.
Figure 13 (right). The PLS loading (weight) plot for the LD₅₀ QSAR. The descriptors are: Mw (1), Bp (2), Mp (3), D (4), n₀ (5), vdW (6), log P (7), Ip (8), GC1 (9), GC2 (10), LC1 (11), LC2 (12), kₑ (13), R_PID (14), Mw² (15), Bp² (16), Mp² (17), D² (18), n₀² (19), vdW² (20), [log P]¹ (21), Ip² (22), GC1² (23), GC2² (24), LC1² (25), LC2² (26), kₑ² (27) and Ip² (28), see also paper III.
Following this, paper V presented the results of the analysis of the genotoxicity data from the DNA precipitation assay. Here, the two QSAR models needed only the linear descriptor variables, indicating a less non-linear relation than in the previous cases. The two QSARs, dealing with the slope (shape) and the $EC_{20}$ values (location) of the dose-response curves, were two-dimensional and explained 90% or more of the biological variation. Lastly, paper VI presented the QSAR modelling of the Ames test data. In this instance, the first multireponse QSARs were calculated. Two models concerned the mutagenic potency of the compounds as evidenced by the slope of the dose-response curves, whereas one quadruple-response model handled the relative cytotoxicity (as indicated by the location of the dose-response curves) of the compounds. All three QSARs described high amounts of the biological variation (80-93%). In contrast to the two-dimensional slope-QSARs, the latter QSAR was three-dimensional, which is explicable by it treating four endpoints at the same time.

An overview of the biological responses and a summary of the results of the QSAR models, are given in Tables 5 and 6, respectively.

Table 5. The twelve biological responses.

<table>
<thead>
<tr>
<th>no.</th>
<th>Description</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>29</td>
<td>Acute toxicity, $LD_{50}$ to rat</td>
<td>III</td>
</tr>
<tr>
<td>30</td>
<td>Acute toxicity, $HN LD$ to mouse</td>
<td>III</td>
</tr>
<tr>
<td>31</td>
<td>Genotoxicity (slope), the DNA precipitation assay</td>
<td>V</td>
</tr>
<tr>
<td>32</td>
<td>Genotoxicity ($EC_{20}$), the DNA precipitation assay</td>
<td>V</td>
</tr>
<tr>
<td>33</td>
<td>Mutagenic potency (slope, TA 100 +S9), Ames test</td>
<td>VI</td>
</tr>
<tr>
<td>34</td>
<td>Mutagenic potency (slope, Ta 100 -S9), Ames test</td>
<td>VI</td>
</tr>
<tr>
<td>35</td>
<td>Mutagenic potency (slope, TA 1535 +S9), Ames test</td>
<td>VI</td>
</tr>
<tr>
<td>36</td>
<td>Mutagenic potency (slope, TA 1535 -S9), Ames test</td>
<td>VI</td>
</tr>
<tr>
<td>37</td>
<td>Relative cytotoxicity ($MD_{50}$, TA 100 +S9), Ames test</td>
<td>VI</td>
</tr>
<tr>
<td>38</td>
<td>Relative cytotoxicity ($MD_{50}$, TA 100 -S9), Ames test</td>
<td>VI</td>
</tr>
<tr>
<td>39</td>
<td>Relative cytotoxicity ($MD_{50}$, TA 1535 +S9), Ames test</td>
<td>VI</td>
</tr>
<tr>
<td>40</td>
<td>Relative cytotoxicity ($MD_{50}$, TA 1535 -S9), Ames test</td>
<td>VI</td>
</tr>
</tbody>
</table>
Table 6. A summary of the results of the QSAR models.

<table>
<thead>
<tr>
<th>Var.</th>
<th>EV</th>
<th>CV</th>
<th>EMV</th>
<th>EAPE</th>
<th>Model dim.</th>
</tr>
</thead>
<tbody>
<tr>
<td>29</td>
<td>89</td>
<td>38</td>
<td>63.5</td>
<td>0.60</td>
<td>2</td>
</tr>
<tr>
<td>30</td>
<td>89</td>
<td>54</td>
<td>71.5</td>
<td>0.53</td>
<td>2</td>
</tr>
<tr>
<td>31</td>
<td>90</td>
<td>75</td>
<td>82.5</td>
<td>0.42</td>
<td>2</td>
</tr>
<tr>
<td>32</td>
<td>95</td>
<td>70</td>
<td>82.5</td>
<td>0.42</td>
<td>2</td>
</tr>
<tr>
<td>33</td>
<td>86</td>
<td>50</td>
<td>68.0</td>
<td>0.56</td>
<td>2</td>
</tr>
<tr>
<td>34</td>
<td>72</td>
<td>37</td>
<td>54.5</td>
<td>0.67</td>
<td>2</td>
</tr>
<tr>
<td>35</td>
<td>92</td>
<td>35</td>
<td>63.5</td>
<td>0.60</td>
<td>2</td>
</tr>
<tr>
<td>36</td>
<td>86</td>
<td>22</td>
<td>54.0</td>
<td>0.68</td>
<td>2</td>
</tr>
<tr>
<td>37</td>
<td>95</td>
<td>52</td>
<td>73.5</td>
<td>0.51</td>
<td>3</td>
</tr>
<tr>
<td>38</td>
<td>97</td>
<td>55</td>
<td>76.0</td>
<td>0.49</td>
<td>3</td>
</tr>
<tr>
<td>39</td>
<td>90</td>
<td>56</td>
<td>73.0</td>
<td>0.52</td>
<td>3</td>
</tr>
<tr>
<td>40</td>
<td>89</td>
<td>46</td>
<td>67.5</td>
<td>0.57</td>
<td>3</td>
</tr>
</tbody>
</table>

a) For the numbering of the variables, see Table 5.
b) Explained variance (%).
c) Crossvalidated predictive variance (%). The crossvalidated variance gives a measure of the predictive capability of a QSAR. However, with designed data, the X-space becomes spherical, which causes the crossvalidation procedure to seriously underestimate the true predictive capability. The omission of one object at a time causes shifts in the projections of the X-space, as there is no preferred projection-direction.
d) Estimated modelled variance. Calculated as (EV+CV)/2.
e) Estimated average prediction error. Calculated as SQR[(100-EMV)/100]. The values are given in the standard deviation unit.
f) Dimensionality of the QSAR model (i.e. the number of PLS-components).

The two QSARs in paper II were also two-dimensional. They were derived using linear terms only. Both models described around 85% of the biological activity variance, but it should be remembered they were based on four compounds only.
3.6 Validation and predictions (step 6)

3.6.1 Experimental validation

The final purpose of a QSAR model is to utilize it for predicting the biological activities of non-tested compounds. Before doing this, it is important that the predictive capability the model is verified experimentally. This is accomplished by actually testing some of the additional compounds and then comparing the experimental results with the values predicted by the QSAR. If the QSAR predicts within acceptable limits, it may be used for more extensive predictive tasks. The adequacy of the predictions should be compared with the precision and the range of the biological measurements.

3.6.2 Using statistical design to select a validation set

Thus, the sixth and last step of the strategy corresponds to (1) the validation of the predictive capability of a QSAR model and (2) the prediction of the biological activities of other substances belonging to the class. For this purpose, a validation set comprising six compounds of the AX-class, was selected according to a $2^{3-1}$ fractional factorial design. Table 7 lists the chemicals in this set and Figure 14 shows their distribution in a score plot analogous to Figure 6. In paper IV, five of the six chemicals were subjected to the same extended chemical characterization as the training set compounds. The sixth compound was included in paper V, where the selection of the validation set was discussed as well. The validation set was selected according to a $2^{3-1}$ fractional factorial design to keep the number of compounds low, but still representative (see paper V).
3.6.3 Predictions for non-tested compounds

The validation set was tested in the DNA precipitation assay (see paper V for details), enabling the two genotoxicity QSARs to be evaluated. The validation of one of these models is shown in Figure 15. Considering the often complex nature of biological systems and the uncertainty associated with such data, the predictions compare quite well with the experimentally found values. Thus, this particular QSAR may be used for further predictive tasks.
So far, validation sets have not been used to test any of the other QSARs. However, the QSARs established for the Ames test results and the LD$_{50}$ to rat will be validated in the near future. Nevertheless, preliminary predictions have been made for non-tested compounds, but these must be viewed with caution until such time as the models have been rigorously validated.

No particular validation sets of compounds were chosen in paper II. Rather, the compounds remaining after the training sets were selected functioned as validation compounds. The efficacy of both these QSARs was verified in a satisfactory fashion by means of these remaining compounds. Thus, the QSARs subsequently were used for predictions.
4 AN OVERVIEW OF THE TWELVE ENDPOINTS

4.1 A principal component analysis

An interesting question is how the monitored biological responses relate to each other. To investigate this, a principal component analysis was carried out on the 10 (compounds) by 12 (responses) biological data matrix (see also Table 5 for an explanation of the responses). Note, however, that eight of the response variables lack observations for fluorotrichloromethane (for a discussion, see papers V and VI). All response variables were scaled to unit variance (autoscaling).

The PCA yielded a three-component model explaining 73% of the variance, the individual PCs describing 30, 34, and 9% of the variance, respectively. The fourth component modelled 5% of the variance, but was barely significant. Figures 16-19 show the loading- and score plots of the analysis.

The most important result of this PCA is that all compounds in the training set fit well to the same model. This indicates that they are indeed biologically similar in their action and that the present modelling is warranted. This ability of multivariate response data, to evaluate similarities in biological action, is most important in application of semi-empirical QSAR modelling.

4.2 Interpretation of the loadings

The second loading vector is plotted against the first in Figure 16 and the third versus the first in Figure 17. It is seen in Figure 16 that the four MD$_{50}$ variables from the Ames test (no.s 37-40) are grouped together. Clearly, they provide very similar information. Moreover, three pairs of responses are discernable. Not surprisingly, the two acute toxicities (LD$_{50}$ and HNLD, no.s 29 and 30) lie together. This applies also to the two types of slope variables for the Ames test (no.s 33-36), reflecting the mutagenic potencies of the training set compounds. It appears that the information inherent in the acute toxicity variables to some extent parallels the information given by the TA...
100 slope variables. The two genotoxicity responses of the DNA precipitation assay (no.s 31 and 32) are distributed differently from the other variables. However, the intrinsic information of variable 31 is also related to the variation of the four clustered MD$_{50}$ responses. Figure 17, which displays the behaviour of the response variables in the third dimension, corroborates the apparent phenomena seen in Figure 16. The DNA precipitation assay obviously provides information not given by the Ames and acute toxicity tests. The four MD$_{50}$ variables are still located near each other, as are the acute toxicity variables and the TA 100 slope variables. The third dimension also indicates that the TA 1535 slope variables contribute specific information.

Figure 16 (left). Loading plot with the second loading vector plotted against the first. For the numbering of the biological responses, see Table 5.

Figure 17 (right). Loading plot with the third loading vector plotted against the first. Notation as in Figure 16.
4.3 Interpretation of the scores

The score plot in Figure 18, showing PC2 versus PC1, represents 64% of the systematic variation in the twelve biological responses. This graph gives a summarizing picture of the biological properties of the training set compounds in the current test systems. However, note that the position of fluorotrichloromethane (no. 7) is tentative, because it is only based on the four variables 29-32. Figure 19, shows PC3 versus PC1.

The compounds are triangularly distributed in Figure 18. One "corner" is fluorotribromomethane (no. 39), positioned in the lower left part of the plot. This compound is by far the most potent of all the ten compounds. It has the lowest LD<sub>50</sub> and HNLD values and was found most cytotoxic in the Ames test (lowest MD<sub>50</sub> values). In the four mutagenic potency variables, however, CBr<sub>3</sub>F holds an intermediate position.
In the lower right part of the score plot, compounds of low mutagenicity and low relative cytotoxicity form the second "corner". This applies to trichloromethane (no. 3), 1,2-dichloroethane (no. 11), 1,1,2,2-tetrachloroethane (no. 15) and 1,1-dibromoethane (no. 33). These compounds have moderate to high (but not extremely high) acute toxicities. Hence, in this area of the plot one finds compounds of relatively low cytotoxicity and mutagenicity, but intermediate acute toxicity.

Lastly, the third "corner". It is composed of the two compounds dichloromethane (no. 2) and bromoethane (no. 30), both of which induce low acute toxicity, but high mutagenic activity. Bromoethane, in fact, was found most mutagenic using the Ames test. The compound 2-chloropropane (no. 48) is situated near bromoethane. Overall, it has the same biological properties as bromoethane, though the magnitudes differ. The substance 1-bromobutane (no. 52) lies closest to the aggressive CBr₃F. This compound is among the most mutagenic, genotoxic and acute toxic.

The concluding interpretation of Figure 18 is that the two PCs jointly describe several biological factors. When moving along the diagonal going from the lower right to the upper left part of the plot, one first encounters compounds of low mutagenicity and finally approaches the area where the highly mutagenic bromoethane is positioned. This diagonal thus seems to represent the change in mutagenic potency of the compounds. The other diagonal appears to separate the compounds depending on their genotoxicity. Fluorotribromomethane is the most and dichloromethane is among the least genotoxic compounds. The first PC in itself accounts for differences in the relative cytotoxicity, by placing fluorotribromomethane to the left and the fairly non-cytotoxic ones to the right (compounds 3, 11, 15 and 33). In an analogous way, the second PC is responsible for discriminating the compounds according to their acute toxicities. The most acute toxic ones lie in the lower region of the plot, whereas the least potent ones show up in the upper portion.

The third principal component, plotted against the first in Figure 19, merely described 9% of the biological activity variance. As is evident from the loading plot in Figure 17, this component mainly is associated with
features of the genotoxicity test not already explained by the two preceding components. However, the separation of the compounds in this dimension does not follow the general changes in genotoxicity.

### 4.4 An overall ranking of the halogenated aliphatics

The calculated PC model can be used for obtaining an overview of the predictions of the remaining compounds in the AX-class. Here, the predictions considered are those listed in papers III, V and VI. By inserting these predictions into the PC model, it is possible to project them down onto the PC planes of Figures 18 and 19. During this classification one also obtains information about whether or not a compound fits into the model.

The results of the projection of the predictions onto the existing PC model is visualized in Figure 20. For clarity, only the PC1-PC2 plane is presented. These two PCs also accounted for almost two-thirds of the biological variation among the training set compounds.

![Figure 20. Scatter plot visualizing the projection of the predictions onto the PC model of the biological data of the training set compounds. For the numbering of the compounds, see Table 1.](image)

Out of the 48 biologically non-tested compounds, 34 are fitted to the PC-model. 14 compounds that previously (papers III, V, and VI) have shown a poor fit to any of the developed QSAR models, are not included in this overall projection.
The PC model was interpreted in the previous paragraphs (sections 4.2 and 4.3). Bearing this interpretation in mind when looking at Figure 20, it is possible to get an overall appraisal of the predicted biological activity pattern of the non-tested compounds.

For example, the location of the chlorinated ethanes (no.s 12-14) in the lower right corner, suggests that these compounds are predicted to have low mutagenicity and relative cytotoxicity, intermediate genotoxicity and rather high acute toxicity. Some monohalogenated compounds, such as bromomethane (no. 9), iodoethane (no. 42) and 1-iodopropane (no. 49), lie in the upper part of the plot, and are likely to be highly mutagenic. Their positioning in the plot also indicates that they are predicted to be of low to intermediate genotoxicity and relative cytotoxicity.

Another interesting grouping is the chlorinated compounds 1-chloropropane (no. 47), 1-chlorobutane (no. 54) and chloroethane (no. 10), together with the CFCs 1,1,2-trichloro-1,2,2-trifluoroethane (no. 18), 1-chloro-2,2,3-trifluoropropene (no. 28) and 1-chloro-2,2-difluoropropene (no. 29). These compounds are predicted to have fairly high genotoxicity and mutagenicity, intermediate relative cytotoxicity and low acute toxicity. It is also interesting to note that these CFCs fall near the training set compound fluorotrichloromethane (no. 7).

Similar interpretations can be made for all the other compounds. Simply by looking at the location of a compound in Figure 20, one can overview its predicted biological activity pattern. This plot can serve as a starting point for further risk assessment of these chemicals. However, the predictions should not be used in isolation, but rather in combination with other data on exposure and use pattern information.

4.5 A PLS analysis

Subsequent to the PCA, a PLS analysis was carried out on the battery of twelve biological responses, using the 28 physico-chemical and structural descriptors. This resulted in a two-component model, significant according to cross-validation. The model described 58% of the variance in the biological data, which is an acceptable result. Recall that the corresponding
PCA (section 4.1) indicated 73% of the variance as an upper limit.

The multivariate PLS model indicates no particular advantage of treating the twelve endpoints together. It is probably more favourable to treat the responses in smaller groups of related endpoints. However, at present it is not possible to decide which approach is the best. This conclusion has to await the experimental validation of some of the QSARs.

4.6 Some possible causes to the biological observations

It is tempting to try to explain the biological behaviour of the compounds in terms of chemical and toxicological reasonings. However, one must be aware of the fact that the halogenated aliphatic hydrocarbons can undergo a variety of chemical processes and can be bioactivated (metabolized) by several different mechanisms [28,48], some of which are well understood but others not. Here, the intention is merely to give a brief view of this topic and not to present an exhaustive discussion.

Commencing with the LD$_{50}$ and HNLD variables, it is noted that they both measure acute toxicity. There are several physico-chemical descriptor variables that, in an equivalent manner, correlate to these endpoints, see for instance Figure 7. Apparently, it is the molecular bulk and polarizability properties of the compounds that have a dominating influence on the biological activity. Descriptor variables such as the molecular weight, the boiling point and the refractive index therefore make significant contributions to modelling the biological variation.

Another interesting observation concerns the results of the Ames test, which shows that the mutagenic potency of the compounds tends to weakly correlate with the degree of halogenation. The three monohalogenated compounds bromoethane, 2-chloropropane and 1-bromobutane, are much more mutagenic (high values in variables 33-36) than the other polyhalogenated compounds. The better alkylating potency of the monohalogenated alkanes appears to be a plausible explanation for this observed tendency. Such compounds may well alkylate for instance the bases in DNA or DNA repair enzymes. Bromoethane is more harmful than the others, probably due to steric factors, its α-carbon being the least sterically hindered and thereby
the most accessible for attack by a nucleophile. The polyhalogenated aliphatics have a lower propensity for acting as alkylating agents, which may be one reason for their generally lower mutagenic potency.

Concerning the genotoxicity test, the results suggest that the hydrophobicity of the compounds is an important property. This was also evident in the corresponding QSARs (paper V), where log P had a significant and dominant influence. Thus, it seems reasonable to assume that the ability of a compound to enter the membrane of the V79 cells affects the outcome of this test. However, the results also tell that log P is not the sole determinant of the genotoxicity of the compounds; there are other regulating factors as well.
5 SUMMARY AND CONCLUSIONS

5.1 Some comments on the strategy

The strategy for ranking the halogenated aliphatic hydrocarbons is the main theme of this thesis. However, is should be emphasized that the strategy has not been developed solely for establishing QSARs for this class. On the contrary, this work is intended to demonstrate a generally applicable QSAR methodology, suitable for many kinds of chemical classes. Thus, the strategy should be viewed as summarizing and representing a philosophy for risk assessment of environmental chemicals. The strategy is based on two cornerstones: (1) the use of statistical experimental design to select a representative set of chemicals on which to base a QSAR and (2) the use of multivariate data analytical techniques to find the relationships between physico-chemical descriptor data and biological responses of the training set compounds. The two cornerstones are discussed further in section 5.2.

The multivariate approach is not restricted to a certain set of chemicals, physico-chemical and structural descriptors, or biological test systems. The selection of chemicals may be carried out starting from any list, inventory or data base containing the chemicals of interest. Likewise, the selection of physico-chemical and structural descriptors should include descriptors that have proven beneficial in other investigations. The biological tests may also be chosen freely, however, it is recommended to use well established tests.

The strategy consists of six sequential steps, some of which are intimately linked to each other. For example, there exists no fundamental theory for how to divide chemicals into classes. This makes steps 1 and 2 dependent on one another. Usually, the chemicals of interest are categorized according to their structural similarities (step 1). Structural similarity, however, is not easily defined in advance. Thus, the subsequent structural description of the compounds (step 2) and the results of the multivariate analysis of this description, may identify a need for revision of the compounds included in a class [49]. For instance, some of the compounds may be outliers or there may exist strong groupings indicating further subdivision to be necessary, etc.
Generally, when processing a class of chemicals through the steps of the strategy, feedback from one step to another is valuable. For instance, results from the fourth step can indicate a compound to behave biologically differently from the others. This can be viewed in connection with the results of the previous steps, to see if there are other indications pointing in the same direction. Likewise, the quality of the predictions in the last step can be compared with the precision of the biological measurements, to ascertain that the predictions are within acceptable limits.

In the end, the strategy makes it possible to obtain predictions of biological effects for many compounds based on measurements on just a few. However, such predictions should neither be used as a substitute for biological testing, nor directly as a basis for regulation and decision making. Rather, the strategy should be regarded as a screening facility, indicating which chemicals can be ranked as the first to be tested and which can be given a lower priority. Predictions of toxicity, however, should not be the sole basis for a priority list for further biological testing of chemicals. Such data should also be integrated with exposure and other effect data, to finally arrive at ranking chemicals according to their potential hazards.

5.2 The two cornerstones of the strategy

The first cornerstone, statistical design, is used to select a small set of training set compounds in a balanced and representative way. Statistical design should be applied to a few and orthogonal design variables and not to a block of collinear structural descriptors, otherwise the number of compounds cannot be kept low and at the same time be information-rich. However, a multitude of structural descriptors is beneficial for a successful QSAR modelling. Due to the discrete nature of the structural variation among chemical compounds, it may sometimes be difficult to find compounds exactly matching the design. This can lead to somewhat distorted (hyper)-cubes of chemicals. However, as pointed out by Skagerberg [18], the most important aspect is that systematic and representative variation is introduced into the training set.
The second cornerstone is the multivariate modelling of the relations between chemical and biological data by means of PLS. Traditionally, MLR is used for this purpose. However, PLS offers three distinct advantages: (1) it can handle many and collinear descriptor variables even exceeding the number of objects, (2) it can cope with missing data and (3) it allows the treatment of several biological responses simultaneously.

5.3 An additional benefit of using design

Even though the halogenated aliphatics class is selected on the basis of the structural similarity concept, it contains a divergent body of chemicals representing various halogenation patterns, sizes, lipophilicities, etc. Thus, the training set compounds, which should represent this variation, can be expected to exhibit different biological properties. In fact, the training set compounds differ substantially in biological activities. Moreover, in all but one aspect - the mutagenic potency - the compound fluorotribromomethane is the most active. In the majority of cases its potency is at least two orders of magnitude larger than the least potent one. This contrasting difference in biological activities must be credited to the use of statistical design. By this, it was possible to select training set compounds spanning the dominant structural features in a balanced fashion.

5.4 The use of linear and quadratic terms of the descriptors

One of the most important observations that can be made based on this study, is the low modelling capability of the interaction \((x_i \cdot x_j)\) terms and the high influence of the quadratic \((x_i^2)\) terms. Since fractional factorial designs supplemented with center-points only allow rough estimations of quadratic terms, in future studies one should perhaps choose another type of design more amenable to quantifying quadratic relationships. A composite design [37] in the design variables defined for a class, or a D-optimal design [18,37] with constraints, are rational alternatives.
5.5 A future outlook

Concerning the AX-class, besides the completion of the 28 day chronic test, no additional screening studies in other biological test systems are planned. However, there are still options open to investigation. One question of particular interest is to investigate whether some of the biological responses can serve as biological model systems for improving the QSAR modelling of other biological responses. Furthermore, the chemical and structural characterization of the AX-chemicals may continue. Especially, it is planned to expand the spectroscopic part of the characterization (NIR spectroscopy).

The LD$_{50}$ QSAR and the QSARs of the Ames test will be validated in the near future, by the testing of validation compounds. Simultaneously to this testing, three of the training set compounds will be reinvestigated to constitute a calibration set. This will facilitate checking of whether systematic differences occur in the biological results, and also to correct for such discrepancies.

The results from the application of the strategy to the halogenated aliphatics class are encouraging. The strategy will hopefully be applied to other classes of chemicals in the future. It has, in fact, already been used in conjunction with a class of monosubstituted benzenes [50].

It is, of course, rather difficult to foresee what will happen within the field of QSAR research during the next few decades. The development of structural descriptors will continue, perhaps with more emphasis on theoretical and calculable descriptors than experimentally determined. Hopefully, however, the use of chemical model systems reflecting the biological system of interest will increase. Another type of descriptors, which hitherto have received only scarce attention but may become increasingly popular, is the biological model systems. As chemical model systems are useful for the modelling of biological activities, it is logical to presume that biological model systems will be valuable as well.
I wish to thank all my friends and colleagues at the Department of Organic Chemistry for creating a stimulating and humorous atmosphere, which made the everyday work both endurable and enjoyable. I also wish to dedicate special thoughts to:

My advisors, professor Svante Wold and professor Michael Sjöström, without whose wisdom and encouragement this thesis would surely not have been printed. I am greatly indebted to them for introducing me to the research field of chemometrics, and especially to the fascinating sub-field of quantitative structure-activity relationships. I specially thank Svante for always providing new inputs and alternative solutions to all types of problems, and Michael for always being there and for listening to my questions and thoughts with a seemingly never failing patience.

Jörgen Jonsson, my dear friend and colleague during all these years, with whom I have shared many unforgettable moments at the department and during travels. I thank Jörgen for critically assessing my work and formulating alternative theories and conclusions.

Dr. Sven Hellberg for always having some spare time for valuable scientific discussions and support.

Fredrik Lindgren for carefully planning and carrying out essential parts of the chemical characterization of the halogenated aliphatic compounds.

Dr. Rune Berglind and his staff at the Defence Research Establishment for taking care of all the biological measurements, and for valuable assistance concerning the interpretation of the results from a biological and toxicological perspective.
Dr. Maria Livia Tosato, head of the structure-activity research group in Rome, for her enthusiasm and optimism, which made our collaboration both fruitful and enjoyable.

All other close coworkers in the chemometrics research group, including our secretary Ann-Helén Waara, for valuable support and assistance.

Börje Nenner, my teacher in chemistry at Östra Gymnasiet, who really awoke my interest for chemistry and environmental sciences and encouraged me to follow the path I did.

Dr. Douglas Baxter for correcting erroneous English in this thesis.

This work was made possible through financial support from the Centre for Environmental Research (CMF) in Umeå and the National Chemicals Inspectorate in Stockholm, which I am grateful for having received.
7 REFERENCES


8 Appendices


