INTELLIGENT MODELLING OF THE ENVIRONMENTAL BEHAVIOR OF CHEMICALS

A Dissertation
submitted to the Faculty of Mathematics and Natural Sciences
of the University Rostock

in fulfillment of the requirements
for the degree of
Doctor rerum naturalium (Dr. rer. nat.)

Shefali Kumar, born on July 19, 1978 in Jaipur (INDIA)
Rostock, October 2007

urn:nbn:de:gbv:28-diss2008-0059-3

Preface

In view of the new European Union chemical policy REACH (Registration, Evaluation, and Authorization of Chemicals), an interest in “non-animal” methods for assessing the risk potentials of chemicals towards human health and environment has increased. The incapability of classical modelling approaches in the complex and ill-defined modelling problems of chemicals’ environmental behavior, together with an availability of large computing power in modern times raise an interest in applying computational models inspired by the approaches coming from the area of artificial intelligence. This thesis is devoted to promote the applications of neuro/fuzzy techniques in assessing the environmental behavior of chemicals. Some of the bottlenecks lying in the neuro/fuzzy modelling of chemicals’ behavior towards environment have been identified and the solutions have been provided based on the techniques of computational intelligence.

The performance of modelling techniques is influenced by a number of factors regarding the choices of model inputs, model structure, model development criterion, and so on. These choices in many cases may not be suitable resulting into the development of a model with a low generalization capability (i.e. it doesn’t cover the whole range of considered chemicals to be assessed). We introduce a methodology to improve the generalization capability of a given modelling technique. This is done via incorporating an “intelligence” in the modelling technique. The effectiveness of the proposed methodology is demonstrated by studying the toxicity and bioconcentration factor modelling problems. As an application of the work to the field of Green Chemistry, a computer model was developed for predicting the toxicity of ionic liquids to *Vibrio fischeri*.

Acknowledgements

This study would not have been possible without the support, encouragement, and suggestions of my mentors, colleagues, friends, and family. Now, I would like to take this opportunity to acknowledge my debt to all these persons.

I express my deep gratitude to Prof. Udo Kragl for his constant support, guidance, and encouragement in completing this work. My thesis is greatly influenced by his useful comments and fruitful discussions. I am highly grateful to Dr. Mohit Kumar for helping me in understanding not only the mathematical theory of computational intelligent techniques but also the software "know-how" in a best scientific traditional manner. I am thankful to "Deutsche Bundesstiftung Umwelt" for providing the financial assistance during the study period of 3 years. I thank Dr. Emilio Benfenati (Instituto Mario Negri, Milan, Italy) for providing the toxicity data used in the European Community project IMAGETOX (Intelligent Modeling Algorithms for General Evaluation of TOXicities). I would also like to thank Prof. Ovanes Mekenyan (Laboratory of Mathematical Chemistry, "Prof. As. Zlatarov" University, 8010 Bourgas, Bulgaria) for providing the bioconcentration related data. I am thankful to all my colleagues for creating a friendly working atmosphere.

I acknowledge my deepest gratitude, love and affection to my parents and family. My gratitude for my husband for his care, inspiration, and support during the years is beyond words. Finally, this work is dedicated to "Pita Ji".

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The chemicals are known to have both beneficial as well as harmful effects on human and environment to which they are exposed. With the development of human civilization over the period of time, rapid industrialization took place. The modern industrial growth has led to the increase in the production of chemicals thereby an increase in their exposure to the human beings and environment. There is about 400 million tons global production of the chemicals. Due to their risk potentials toward human health and environment, there is a need of an assessment of the effects of these chemicals. Assessing chemical compound's risk to environment and health has become a part of the legislations and regulations world wide. European Commission has introduced the REACH (Registration, Evaluation, and Authorization of Chemicals) system [4].

1.1 REACH (Registration, Evaluation, and Authorization of Chemicals)

The aim of REACH is to provide an improvement in the protection of human health and environment while maintaining the competitiveness and innovation of the European Union chemicals industry. REACH came into force in June 2007 and requires all chemicals (which are manufactured in or imported into the European Union in quantities more than one tonne or more in volume each year) to be tested for health and safety and registered to a central European Chemicals Agency located in Helsinki, Finland. A free briefing which explains the key components of the REACH regulation has been provided by the "Lowell Center for Sustainable Production" [3]. The basic idea behind REACH is that the main responsibility for chemical safety is placed on the chemical producer or importer and not on public authorities or downstream users. The elements of REACH are [5]:

1. REACH is concerned with the all substances unless they are radioactive, subject to customs supervision, or are non-isolated intermediates. Waste is specifically exempted. Further details can be found out in [3].

2. Registration is the process that requires substance manufacturers and importers to send a registration dossier containing relevant information on the substance to a central European Chemicals Agency. This applies to substances manufactured or imported in quantities of one tonne per year or more as stated in [3]:

   > "*Save where this Regulation provides otherwise, any manufacturer of a substance in quantities of 1 tonne or more per year shall submit a registration to the Agency.*"
   >
   > "*Save where this Regulation provides otherwise, any importer of a substance, either on its own or in a preparation, in quantities of 1 tonne or more per year shall submit a registration to the Agency.*"

3. Data sharing is required for studies on vertebrate animals to reduce testing on vertebrate animals.

4. The communication requirement of REACH ensures that the information on hazards and risks and how to manage them will be passed down and up the supply chain.

5. Downstream users are required to consider the safety of their uses of substances. The downstream users will have to check that they use a substance within the

conditions described in the exposure scenarios in the Annex to the safety data sheet and apply these conditions.

6. The European Chemicals Agency will evaluate Registration dossiers and individual substances. This assessment may be used to prepare proposals for restrictions or authorisation.

7. The substance of very high concern will be subject to authorisation. Such high concern candidate substances will be listed and published by the agency. These substances are those that are

   (a) category 1 or 2 carcinogens, mutagens, and reproductive toxins

   (b) persistent, bio-accumulative and toxic

   (c) very persistent and very bio-accumulative

   (d) identified from scientific evidence as causing serious effects to humans or the environment giving rise to a concern equivalent to those mentioned above (e.g. endocrine disruptors) which will be identified on a case-by-case basis.

   The applicants have to demonstrate that the risk associated to the use of such substances is adequately controlled. If not, then socio-economic benefits of substance's uses must outweigh the risks.

8. The Restrictions procedure is designed to legally restrict the production or specific uses of certain dangerous chemicals whose use poses unacceptable risks to human health or the environment and need to be managed on an EU-wide basis.

9. A new European Chemicals Agency will be created in Helsinki, Finland to manage the technical, scientific, and administrative aspects of the REACH system.

10. A classification and labeling inventory of dangerous substances will help to promote an agreement within industry on the classification of a substance. Under

REACH this classification and labeling information will be entered into an inventory which will be published on the Agency's web site [3].

11. Some of the information generated by REACH will be publicly available via the internet. However, commercially sensitive information will be kept confidential.

1.2 Prediction of Environmental Behavior of Chemicals

1.2.1 The Motivation

The diseases caused by chemicals are assumed to account for some 1% of the overall burden of all types of disease in the European Union [39]. REACH is expected to reduce pollution of air, water, and soil. A study at the University of Leicester, UK has revealed that the implementation of REACH would need additional 12 million animals for testing of chemicals [39]. The ethical issues are involved in the animals testing. There is a public pressure to reduce animal testing [82]. Moreover, running of traditional bioassays for the testing are costly and time consuming. The alternative non animal test methods are required due to the cost and the very long time it would take to run animal tests for all chemicals to be assessed. It is expected that alternative methods would save the lives of at least 2 million animals [58]. A detailed explanation of the factors deriving the motivation for predicting toxicity and fate has been given in [82].

Quantitative Structure-Activity Relationship (QSAR) models have emerged as the promising "non-animal" alternative to predict the environmental behavior of the chemicals. The QSAR models describe and predict the effect of a given concentration or dose of a chemical on the health of population of certain biological species by the structure of the chemical. There is a co-relation between the chemical structure and biological activity, this has been recognized as structure activity relationship (SAR). The QSAR approach is based on the assumption that the activity of a chemical compound is determined by its molecular structure and the structure is represented by numerical descriptors which encapsulate the molecules properties relevant to the activity. In simple words by the term QSAR, we understand the process by which chemical structure is quantitatively correlated with a biological activity or chemical reactivity.

1.2.2 Some Historical Remarks

The modern science of predictive toxicology has grown and developed with the chain of historical events. A brief summary of the key historical events has been provided in [82]: It is since 5000 BC the knowledge of poisonous plants and animals venoms was known to human beings. Around 3000 BC the Egyptian had identified the toxic effects of some substances. Ebres Papyrus had described more than 800 recipes of poison around 1550 BC. During early 1500 AD Paracelsus had discovered that the toxicity of plant or animal poison are due to specific chemicals. It is in the modern era, that is early 1800's, Orfila had been credited as founder of toxicology. Cros, in 1863 had made an observation that the toxicity of the alcohol decreases with their water solubility. It was between 1860-1940 when several researchers had proposed the theories related to toxicology. In 1893 Richet had observed that toxicity was inversely related to the solubility. Meyer and Overton both had independently proposed in 1899-1901 that narcosis is related to partitioning between oil and water phases. Ferguson had proposed the solubility cutoff for the acute toxicity in 1939. In 1964, Hansch and Fujita developed the QSAR methods [50] which have been widely studied during last years in the light of new computational approaches coming from the area of artificial intelligence.

1.2.3 Intelligent Modelling Techniques

The modelling techniques could be roughly divided into following four approaches: "*white-box*", "*black-box*", "*gray-box*", and "*intelligent*" modelling.

The term "*white-box*" modelling refers to the mathematical treatment of process's nature and behavior with non-linear differential equations, based on the thorough understanding of the physical laws governing the behavior of the process. White-box models are fully derived by first principles. All equations and parameters can be determined by theoretical modelling. However, the environmental behavior modelling problem is characterized by complexity and uncertainty, and a complete understanding of the underlying mechanisms is virtually impossible. Therefore difficulties are encountered in conventional white-box modelling approaches, when complex and poorly understood systems are considered.

The black-box modelling approach consists of approximating the process by using some "*black-box*" structure. The modelling problem is simply the estimation of parameters describing the black-box structure using process data. Black-box models are based solely on measurement process data. The well known examples of this approach include regression and neural networks. The most severe limitation of this approach is the physical insignificance of structure and parameters of black-box model and therefore can not be used to analyze the process behavior.

A mixed approach combines the advantages of white-box and black-box approaches by modelling the known part of the process using physical laws and unknown part by black-box approximation using process data. This approach is termed as "*gray-box*" modelling or hybrid modelling. Typically, the determination of the model structure relies strongly on prior knowledge while the model parameters are mainly determined by process data.

The "intelligent" (also referred as artificial intelligent, computational intelligent, expert system) modelling employs techniques motivated by biological systems and human intelligence to model the process behavior. The motivation behind these methodologies is the human capability of handling complex tasks and making decisions under uncertainty. These techniques are based on the representation of process knowledge, using, for example, natural language, rules, etc. Fuzzy modelling and artificial neural networks are typical examples of intelligent modelling techniques. Artificial neural networks have the learning and adaptation capability by imitating the functioning of biological neurons on a simplified level. Fuzzy systems, on the other hand, are designed to handle uncertainty and vagueness by using fuzzy sets and if-then rules. Fuzzy systems model the complex input-output mappings based on statements that closely model the way people think, and these statements can be constructed in a heuristic fashion using application-specific knowledge. Therefore, fuzzy modelling, offer the beneficial feature of a way to incorporate heuristic information and to interpret the process behavior with linguistic terms.

The environmental behavior of chemicals is complex and ill-defined in nature and thus motivating the researchers to apply more powerful computational approaches (i.e. intelligent techniques) in modelling the toxicity, bioconcentration factor, etc. of chemicals. A large literature is available studying the applications of neural networks [36,46,59,79,112,118], expert systems [10,44], and hybrid systems e.g. neuro-fuzzy models [88].

1.3 The Methodological Problems in Building Predictive Models

The environmental behavior cann't be modelled using classical differential equations due to the complexity and lack of the complete knowledge of the system. The availability of large computing power in modern time raises an interest in applying computational models, inspired by the approaches coming from the area of artificial intelligence, in predicting the environmental behavior of chemicals. These methods are typically "data driven modelling" techniques. Mathematically, it is assumed that the activity of a chemical y (e.g. quantifiable toxicity end point for the fathead minnow) and the chosen molecular descriptors $(d_1, d_2, \cdots, d_n)$ of the structure are related through a functional relation:

$$y = f(d_1, d_2, \cdots, d_n). \tag{1.1}$$

Here, f is an unknown function that is identified using nonlinear models (e.g. neural networks, fuzzy models) called QSAR models. The existing experimentally measured activity data of chemicals and their molecular descriptor values are used to build a QSAR model. The fundamental concern is that the identified QSAR model achieves a good generalization of the model over the whole range of chemicals to be assessed.

Aiming at the good generalization capability of the QSAR model, following factors must be considered during the data-driven construction of the model:

P1: A few thousands of molecular descriptors can be calculated for a chemical structure. Which of the descriptors are most appropriate to serve as the inputs of the QSAR model?

P2: What model type and model structure should be chosen? For example, if a neural network is considered, then the modeling performance may be sensitive towards the choice of number of layers and neurons.

P3: The experimentally measured activity data may be noisy and thus the model identification method must be robust to the noise present in the data.

P4: The literature is flooded with the different neuro/fuzzy modelling techniques. The researchers, only interested in the applications, may find tough to understand the mathematics of the modelling techniques. How could a user keep track of the new proposed techniques based on advanced mathematics in his existing modelling software?

These problems (P1-P4) are though basic in nature but not solved till-now completely. The addressing of these problems is must for a wide acceptance of the QSAR approach in the research community.

1.4 The Central Problem of the Thesis

In view of the problems P1-P3, (1.1) should be modified as

$$y = f(d_1, d_2, \cdots, d_n) + n$$

where n is an uncertainty associated with the QSAR model, which takes into account any data noise and modeling errors (arising due to the non-optimal choice of chosen descriptors, model type, and model structure). In modelling literature, n is usually termed as disturbance or noise, however, we referred n to as uncertainty to emphasize

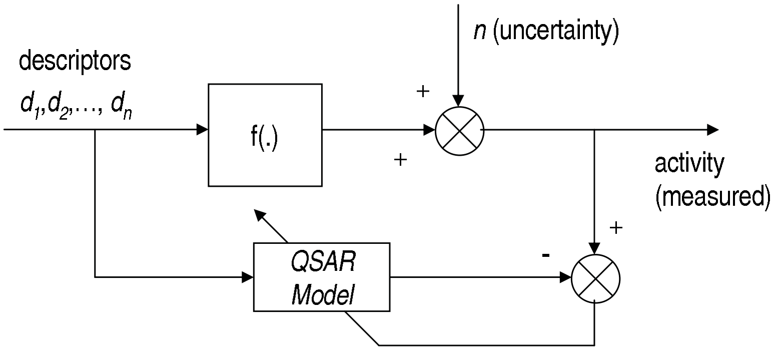


Figure 1.1: Developing a QSAR model using experimental data

that there is an uncertainty regarding the activity data, optimal choice of descriptors, model type, and model structure. Fig. 1.1 shows the identification of a QSAR model using experimentally measured activity data and descriptors values. The identification algorithm consists of tuning the adjustable parameters of the model so that the model output matches activity data in some "optimal manner". The optimal criterion is defined differently by the researchers resulting into the differences among the identification methods. The uncertainty value n, that affects the model identification procedure, is unknown.

The uncertainty n is the root cause of the poor generalization performance of the identified QSAR model. The central problem of this thesis is to propose a methodology for the development of QSAR models with an improved generalization via taking into account the underlying uncertainty in the modelling problem in a sensible way. The problem is stated formally as

> *Given the choice of molecular descriptors, model type, model structure, and model identification algorithm; How can the computational intelligence techniques be utilized in handling the underlying uncertainties? By handling of uncertainties it is meant that modelling performance of the given model identification algorithm is not affected adversely by the uncertainties. How can intelligence (i.e. capability of taking care of uncertainties) be incorporated in a given modelling problem (i.e. given molecular*

> *descriptors, model type, model structure, and model identification algorithm) to achieve robustness against uncertainties?*

1.5 Outline of the Work

The main aim of the thesis is to solve the central problem 1.4 with applications to the modelling of environmental behavior of chemicals. The work is organized into chapters as follows:

Neuro/fuzzy modelling of chemicals' behavior: The second chapter of the thesis reviews the effectiveness of the existing neural/fuzzy techniques in modelling the environmental behavior of chemicals. This is done by studying different techniques in modelling the toxicity and bioconcentration factor of chemicals. It will be demonstrated that due to the presence of uncertainties the existing neural/fuzzy techniques lead to the development of models with a low generalization performance. Thus, some research efforts are required to deal with the issue of uncertainties so that the QSAR models are general enough to cover the whole range of chemicals to be assessed.

Handling uncertainties using a fuzzy filter: The third chapter introduces a fuzzy filter based approach to handle the uncertainties. A fuzzy filter, designed on the basis of a mathematical criterion, is used to filter out the uncertainties from the modelling problem. The toxicity modelling problem is revisited and an improvement in the generalization performance of the models due to the handling of uncertainties is shown.

Incorporating intelligence in modelling: In the fourth chapter, a methodology that incorporates an intelligence (a capability of taking account of uncertainties in a sensible way during the development of the model) in a given modelling technique is introduced. The method improves the generalization capabilities of a given neuro/fuzzy modelling technique based on the information about uncertainties provided by the fuzzy filter. The approach is demonstrated by re-visiting the bioconcentration factor modelling problem.

A study on ionic liquids: Ionic liquids belong to a new class of chemicals which are not only of great industrial importance but also environmental friendly termed as "Green Chemicals". To facilitate their wide industrial acceptance in accordance to the new European chemical policy REACH, it is important to investigate their impact on the health and environment. Since the ionic liquids are known as "designer solvents" therefore the study of their environmental behavior will help in designing eco-friendly ionic liquids. Since there are theoretically over 10^6 ionic liquids, therefore it is necessary to develop a computer models for a fast and accurate prediction of their toxicity. In fifth chapter, some remarks about environmental behavior of ionic liquids have been provided and a neural network based model was constructed for the prediction of their toxicity (*Vibrio fischeri* EC_{50}).

Concluding remarks: Finally, thesis is concluded with an identification of main research findings and their contribution to the state of art. The limitations of the study and some details of future work are also provided.

1.6 Main Contributions

The overall aim of this study was to boost the QSAR modelling as promising non-animal alternative to predict the environmental behavior of the chemicals. The work contributes towards achieving the overall goal as follows:

- The identified bottlenecks of the QSAR approach to environmental behavior modelling, i.e. problems P1-P4 listed in section 1.3, have been reformulated to problem 1.4. Problems P1-P3 have been partially addressed by introducing an uncertainty n (which will be handled using computational intelligent techniques). And problem P4 will be completely addressed since the solution of the central-thesis-problem improves the generalization capability of a given modelling technique.

- A robustness of the modelling performance against uncertainties will be achieved.
- Unlike many studies, instead of proposing a new modelling techniques we incorporate intelligence in a given modelling technique. This is done via penalizing the data (used for identifying the QSAR model) on the basis of amount of uncertainties associated to the data. The generalization performance of even non-robust modelling techniques improves when penalized data instead of original data are used.
- The toxicity modelling problem has been studied with a data set built up by U.S. Environmental Protection Agency concerned with the acute toxicity 96-h LC_{50} to the fathead minnow fish (*Pimephales promelas*) [45, 88, 101, 105].
- The methodology has been demonstrated by considering a bioconcentration factor modelling problem for a data set of 511 chemicals taken from [31].
- A model to predict the *Vibrio fischeri* toxicity of ionic liquids has been developed.

Chapter 2

Neuro/Fuzzy Modelling of Chemicals' Behavior

This chapter evaluates the commonly used neuro/fuzzy techniques in modelling the environmental behavior of chemicals. In particular, toxicity modelling and bioconcentration factor (BCF) modelling problems are studied. It is shown that poor generalization performance is a typical characteristic of the modelling techniques in presence of uncertainties. The uncertainties associated to the toxicity modelling problem are visualized using self-organizing maps [64].

2.1 A Toxicity Modelling Problem

2.1.1 The data set

As a case study, we consider a data set built up by U.S. Environmental Protection Agency referring to acute toxicity 96-h LC_{50} to the fathead minnow fish (*Pimephales promelas*) [45, 88, 101, 105]. Fish are aquatic vertebrates which are most commonly used animal in toxicity testing for the environmental risk assessment. The fathead minnow (Fig. 2.1) is one of the different fish species used as model organism in ecotoxicology. The feathed minnow is a demersel cyprinid species which originate from the temperate water of North America and inhabits in small river, muddy pools of

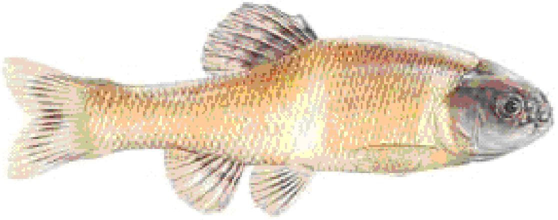


Figure 2.1: The fathead minnow (*Pimephales promelas*)

headwaters. Our toxicological data set contains 568 compounds representing several chemical classes and modes of action. This data set was used in the European Community project IMAGETOX (Intelligent Modeling Algorithms for General Evaluation of TOXicities). It was stated by authors in [88] that the heterogeneity of data set makes it difficult to model, and thus the QSAR models trained with this data set should be quite general.

A large number of descriptors are available in the literature for QSAR studies. Our concern here is not to make a comparison among them but to handle uncertainties for the given descriptors. We calculate for our analysis several constitutional, 2D autocorrelations, Burden eigenvalue, geometrical, 3D-MoRSE, WHIM, GETAWAY, and molecular properties based descriptors using E-DRAGON [114]. Further, topological structure descriptors (including molecular connectivity chi indices, kappa shape indices, E-State indices, molecular connectivity difference chi indices, atom-type E-State indices, group-type E-State indices, topological polarity, and counts of molecular features) have been obtained for the 568 organic compounds.

A pool of 20 descriptors, which showed highest absolute correlation with the target variable $-\log(LC_{50}(mmol/l))$, was created for consideration for possible QSAR model inputs. This was done simply by calculating the values of correlation coefficients among the variables. We use the method of "Principal Feature Analysis" [21] to choose 4 descriptors out of the 20, which retain most of the information, both in the sense of maximum variability of the descriptors in the lower dimensional space and in the sense of minimizing the reconstruction error. The various steps followed were [21]:

1. The data covariance matrix Σ were decomposed as

$$\Sigma = A \Lambda A^T$$

   where Λ is a diagonal matrix whose diagonal entries are the eigenvalues of Σ, and columns of matrix A are eigenvectors of Σ.

2. Let A_4 be a matrix containing first 4 columns of A and denote the rows of A_4 as $V_1, V_2, \cdots \in R^4$.

3. The vectors $|V_1|, |V_2|, \cdots$ were clustered into 4 clusters using $K-$means algorithm. For each cluster, the vector nearest to the mean of the cluster (say V_i) and the corresponding variable (i.e., i^{th} variable) was selected. This resulted in the choice of 4 descriptors.

The 4 descriptors are

- H3v (GETAWAY descriptor [23, 24]): H autocorrelation of lag 3 / weighted by atomic van der Waals volumes.

- ATS2p (2D autocorrelation descriptor [115]): Broto-Moreau autocorrelation of a topological structure - lag 2 / weighted by atomic polarizabilities.

- MLOGP (molecular property [115]): Moriguchi octanol-water partition coefficient.

- xv0 (topological structure descriptor [61]):- Valence chi 0 index.

2.1.2 Uncertainties associated to the data set

We are concerned with the mining of a 5-dimensional data set (4 descriptors and one target variable). Self-organizing maps (SOM) [64] provide a possibility of visualization of multi dimensional data onto a two dimensional map while preserving the topology of the data in the original space, i.e., the data points located near each other in the original space remains neighbors on the map. The map constitutes of neurons located on a regular 2-dimensional grid with hexagonal or rectangular lattices. Each neuron has an associated prototype vector of same dimension as the data. The map training procedure consists in adjusting the prototype vectors of the best matching neuron (i.e. the neuron most similar to the data vector in terms of Euclidean distance) and its neighbors so that the prototype vectors of the best matching neuron and its neighbors are more similar to the data vector. The common method to visualize a SOM is the U-matrix which shows the distances between prototype vectors of neighboring units. The location of a specific data sample on the map is determined by locating the best matching neuron of the data sample.

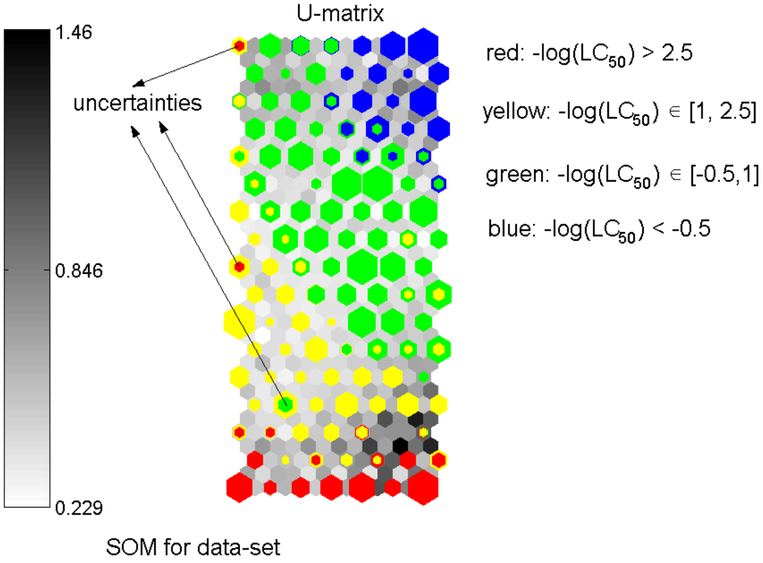


Figure 2.2: Distribution of 568 compounds on the map

In our study, molecular descriptors values and toxicity data were analyzed using a SOM that constitutes of 2-dimensional 17×7 grid with hexagonal lattices. Fig. 2.2 shows the distribution of 568 compounds on the map. Fig. 2.2 shows the visualization of U-matrix values and four hit histograms (red, yellow, green, and blue) corresponding to the four classes of compounds. A hit histogram corresponding to a specific compound class is calculated by aggregating the best matching neurons of all the data points of that class. Here, the data set of 568 compounds is divided into four classes according to the toxicity ranges: $-\log(LC_{50}(mmol/l)) > 2.5$ for red, $-\log(LC_{50}(mmol/l)) \in [1, 2.5]$ for yellow, $-\log(LC_{50}(mmol/l)) \in [-0.5, 1]$ for green, and $-\log(LC_{50}(mmol/l)) < -0.5$ for blue. As seen from Fig. 2.2, there are some compounds belonging to a specific class which are located on the map quite away from the other compounds belonging to the same class. This irregularity in the locations of the compounds of same class on the map may be a result of the uncertainties associated with the selected descriptors and toxicity data mapping problem.

2.1.3 Generation of training and testing data sets

The aim is to develop a QSAR model with H3v, ATS2p, MLOGP, xv0 as inputs and $-\log(LC_{50})$ as the output. The model will be trained with the data of 379 compounds and remaining 189 compounds will be used for the testing of model. The training and testing sets have been created as follows:

1. The point in the 5-dimensional space, whose coordinates correspond to the minimum values of H3v, ATS2p, MLOGP, xv0, $-\log(LC_{50})$, has been taken as the reference point.

2. The distance of each compound from the reference point is calculated and all the 568 compounds are arranged in the ascending order of their distances from the reference point.

3. Every third compound in the series of ascending order arranged compounds is taken as the testing compound and the remaining compounds as the training compounds.

This division of compounds into training and testing is meant for a sandwiching of testing compounds between training ones in the sense of Euclidean distance.

Fig. 2.3 shows the distribution of training and testing compounds by visualizing their hit histograms on the U-matrix. It can be seen from Fig. 2.3 that the training compounds (red in color) reflect the diversity of the data and the testing compounds (blue in color) are "sandwiched" between the training compounds.

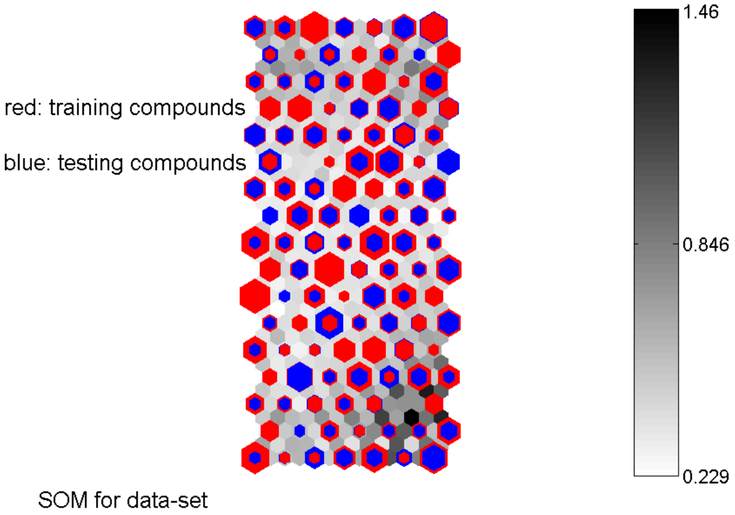


Figure 2.3: Distribution of training and testing compounds on the map

2.1.4 Performance of several neural network training algorithms

For the modelling of the toxicity data, the training of a 3-layer feed-forward neural network is considered. The first layer has 6 "tansig" (i.e. with hyperbolic tangent sigmoid transfer function) neurons, the second layer has 4 "tansig" neurons and the third layer one "purelin" (i.e. with linear transfer function) neuron. The network was initialized with random values of weights and biases. A number of standard training algorithms, available in MATLAB Neural Network Toolbox, have been used to train the network. The training stops if the number of epochs exceeds 5000 or the mean squared error drops below 0.01. For a comparison of the performance, coefficient of determination (R^2) and maximum absolute error (MAE) are calculated for each QSAR model.

Table 2.1 shows the performance of 8 different models ($N_1, \cdots, N_8$) which have been trained using different neural network training algorithms. We see that models $N_3, N_4, N_5, N_6,$ and N_7 undergo overtraining resulting in the loss of generalization

Table 2.1: Performances of different training algorithms

Model	Training algorithm	R^2 *training*	R^2 *testing*	MAE *training*	MAE *testing*
N_1	Batch Gradient Descent learning rate = 0.05	0.6876	0.6334	2.7567	2.5362
N_2	Batch Gradient Descent with Momentum learning rate = 0.05 momentum constant = 0.9	0.6615	0.6024	3.9133	4.1319
N_3	Resilient Backpropagation	0.7146	0.5121	2.5095	4.6666
N_4	Conjugate Gradient (Fletcher-Reeves) Charalambous search	0.7368	0.5097	2.6395	5.2900
N_5	Scaled Conjugate Gradient	0.7556	0.3884	2.3588	7.9039
N_6	One Step Secant algorithm Backtracking search	0.7624	0.5346	2.5777	4.8835
N_7	Levenberg-Marquardt	0.7548	0.2359	2.3186	17.8524
N_8	Bayesian regularization	0.7058	0.6250	2.7280	2.9219

performance as indicated by low R^2 and high MAE values on the testing data. However, the models N_1, N_2, N_8 showed a robustness towards uncertainties and have not been overtrained as indicated by their performance on the testing data. The poor performance of models N_3, N_4, N_5, N_6, and N_7 was a result of the fact that their training algorithms were not robust towards the underlying uncertainties in the modelling problem.

2.2 Bioconcentration Factor Modelling

Bioconcentration refers to the process of accumulation of chemicals in an aquatic organism as a result of exposure of the organism to a chemical concentration in the water via non-dietary routes. The extent of chemical bioconcentration is expressed in terms of bioconcentration factor (BCF) defined as the ratio of the chemical concentration in the organism to that in water [84]. The BCF is a measure of the tendency of a substance to bioconcentrate in aquatic organisms. For an assessment of the bioaccumulation potential of chemicals, BCF in marine or freshwater organisms is traditionally used as an indicator. A flow through method [38] is used for an experimental determination of BCF. The guidelines for characterizing potential bioconcentration in fish under flow-through conditions are provided in [97]. A method suitable for very hydrophobic chemicals has been outlined in [47]. There is a correlation found between BCF and $\log K_{OW}$ values. A good correlation has been established for a number of chemicals along with some limitations. An example of such limitations is that these correlations don't address the metabolic degradation of chemical compound within the organism and thus tends to over-predict.

The motivation for developing the computer models for predicting the BCF of chemicals is derived from the fact that the experimental measurements are time-consuming, expensive, and not feasible for many thousands of chemicals that are of potential regulatory interest. Another motivation of BCF modelling is due to the ethical issues involving animal testing. Many studies aiming at the prediction of BCF values, based on Quantitative Structure-Activity Relationship (QSAR) approach, have appeared in the literature [29]. Typically, the models that map the hydrophobicity ($\log K_{OW}$) of the chemicals to their ($\log BCF$) values are developed. Several modelling approaches including linear BCF models [83,116,117], bilinear BCF model [14], polynomial BCF model [22], fragment based additive BCF model [92], nonlinear empirical model [32] can be found in the literature. The researchers, in addition to the $\log K_{ow}$ based modelling, also examined the BCF models based on solubility in octanol [8], models based on aqueous solubility [28, 53, 60], models based on linear solvation energy relationships [54, 99], models based on connectivity indices [80], models based on fragment constants [113], models based on quantum chemical descriptors [121], models based on diverse theoretical descriptors [30, 48, 49].

2.2.1 The data set

We consider the modelling of a BCF data set of 511 chemicals taken from [31]. The data set includes following chemical classes: alkanes, alkenes, mono and di-aromatic hydrocarbons, polycyclic aromatic hydrocarbons (PAH), polychlorinated dibenzofuranes (PCDF), polychlorinated dibenzodioxines (PCDDO), polychlorinated biphenyles (PCB), cycloalkanes and cycloalkenes, chloraromatic chemicals, perfluorinated acids (PFA) with 6 to 13 difluoromethylene functions in the chain, chlorinated biphenyl esters, aliphatic esters, chlororganic chemicals, aliphatic and aromatic N-containing compounds, polycyclic aromatic N-containing compounds, organotin compounds, sulphur-containing heterocyclic compounds.

E-DRAGON [114] was used to calculate several molecular descriptors of the compounds. Out of the large number (several hundreds) of descriptors, a few descriptors, that serve as the inputs of the models for predicting the $\log BCF$ values, were chosen as follows:

1. descriptors with a standard deviation less than 10^{-6} were rejected.

2. a pool of 20 descriptors, which showed highest absolute correlation with the $\log BCF$ values, was created for consideration for possible QSAR model inputs. This was done simply by calculating the values of correlation coefficients among the variables.

3. the method of "Principal Feature Analysis" [21] was used to choose 5 descriptors out of the 20, which retain most of the information, both in the sense of maximum variability of the descriptors in the lower dimensional space and in the sense of minimizing the reconstruction error.

4. these 5 descriptors are

    - **H1v** (GETAWAY descriptor [23,24]): H autocorrelation of lag 1/weighted by atomic van der Waals volumes
    - **MATS4v** (2D autocorrelation descriptor [94]): Moran autocorrelation - lag 4/weighted by atomic van der Waals volumes
    - **BLTD48** (molecular property): Verhaar model of Daphnia base-line toxicity for Daphnia (48h) from MLOGP (mmol/l)
    - **R5p** (GETAWAY descriptor [23,24]): R autocorrelation of lag 5/weighted by atomic polarizabilities

- **TPSA(NO)** (molecular property [37]): topological polar surface area using N, O polar contributions

The aim is to develop a QSAR model with these 5 descriptors as inputs and $\log BCF$ value as the output. The model will be trained with the data of around 2/3 of the total compounds and the remaining 1/3 compounds will be used for the testing of the model. The training and testing sets, as in toxicity modelling problem, have been created as follows:

1. All descriptors and $\log BCF$ values are normalized to have zero mean and unit variance.

2. The point in the 6-dimensional space, whose coordinates correspond to the minimum values of 5 descriptors and $\log BCF$ value has been taken as the reference point.

3. The Euclidean distance of each compound from the reference point is calculated and all the compounds are arranged in the ascending order of their distances from the reference point.

4. Every third compound in the series of ascending order arranged compounds is taken as the testing compound and the remaining compounds as the training compounds.

2.2.2 The issue of uncertainties

The BCF modelling problem is studied using a neural network and a fuzzy model. Let us first consider the training of a 3-layer feed-forward neural network. The first layer has 6 “tansig” (i.e. with hyperbolic tangent sigmoid transfer function) neurons, the second layer has 4 “tansig” neurons and the third layer one “purelin” (i.e. with linear transfer function) neuron. The network was initialized with random values of weights and biases. The network was trained using two different training algorithms: “scaled conjugate gradient backpropagation” (MATLAB Neural Network Toolbox command “trainscg”) and “Levenberg-Marquardt backpropagation” (MATLAB Neural Network

Toolbox command "trainlm"). The training of the network stops if the number of epochs exceeds 10000.

Also, a Sugeno type fuzzy model was trained using an in-built training algorithm in MATLAB Fuzzy Logic Toolbox ("anfis" command). The "anfis" algorithm combines the least-squares and backpropagation gradient descent method to identify the parameters of the fuzzy model. The structure of the fuzzy model was generated from the training data using subtractive clustering (MATLAB Fuzzy Logic Toolbox command "genfis2"). The fuzzy model was trained till 1000 epochs.

Table 2.2: The performance of some neural/fuzzy modelling methods

method	R^2**-training**	**RMSE-training**	R^2**-testing**	**RMSE-testing**
"trainscg"	0.8924	0.4297	0.5596	0.9975
"trainlm"	0.9144	0.3831	0.5859	0.9298
"anfis"	0.8691	0.4739	0.4721	1.1631

The modelling performance is assessed by computing the coefficient of determination (R^2) and root mean squared error (RMSE) on training and testing data. Table 2.2 shows the performance of some of the standard neural/fuzzy modelling methods. We observe from table 2.2 that the modelling techniques show good performance on the training data, however, poor performance on the testing data. This indicates in the modelling problem the presence of uncertainties for the chosen molecular descriptors, chosen model type and structure, training algorithms related chosen parameters, and so on. These uncertainties resulted in the overtraining of the model and thus a poor generalization performance (as shown by a poor performance on the testing data).

2.3 Summary

We have seen that several neuro/fuzzy modelling techniques resulted in the overtraining and thus a poor generalization performance. To increase the generalization performance, one could argue for

1. a more appropriate selection of inputs based on some mathematical criterion,

2. a decrease in the number of training compounds,

3. a reduction in the number of adjustable parameters of the model,
4. the use of a robust training algorithm e.g. Bayesian regularized learning as did for model N_8 in table 2.1.

However, our aim here is to highlight the issue of overtraining (that results from the uncertainties regarding the optimal choices of aforementioned factors) and a method for dealing with the overtraining issue.

Chapter 3

Handling Uncertainties Using a Fuzzy Filter*

It was included in previous chapter that the development of a chemicals' behavior predicting model without considering uncertainties may produce a model with a low generalization performance. At the same time, the fundamental concern in QSAR approach is the good generalization capability of the model. To improve the generalization performance, Bayesian regularized neural networks have been suggested as robust QSAR models by authors in [16–18, 125]. The identification of a QSAR model using input-output data is an "ill-posed" problem. Regularization converts the identification problem into a "well-posed" problem. However, the choice of regularization parameters is usually not obvious. Bayesian regularization, under some stochastic assumptions on the nature of uncertainties, provides an optimal value of regularization parameters by applying Bayes' theorem [85].

Fuzzy inference systems based on fuzzy theory of Zadeh [127] are considered suitable for dealing with uncertainties. Fuzzy modeling framework provides a possibility of representing the knowledge in the environment of uncertainty and imprecision. The

*The method presented here will appear in
S. Kumar, M. Kumar, R. Stoll, and U. Kragl. Handling Uncertainties in Toxicity Modeling using a Fuzzy Filter. *SAR and QSAR in Environmental Research*, 18 (in press), 2007.

concern of this chapter is to handle the involved uncertainties in QSAR modeling using a fuzzy system. Our approach of handling uncertainties is based on following ideas:

1. Developing a fuzzy filter using experimental data that would filter out the uncertainty from experimentally measured activity data.

2. Developing QSAR models using filtered activity data. Since the uncertainties in data have been filtered out, therefore the training algorithm should result in the identification of a model with good generalization performance even if the training algorithm is non robust.

It is easy to realize that the development of the fuzzy filter is the bottleneck of the approach. The initial enthusiasm about fuzzy models was due to the fact that fuzzy models could be constructed from the knowledge of human experts. However, for the problems such as toxicity modeling, the human knowledge is not sufficient to establish the underlying relationships due to the complexity of the problem. Thus, different methods have been developed during the years for an automatic identification of fuzzy models with example input-output data [6, 7, 55]. Robustness against uncertainties becomes the main concern of any fuzzy model identification method in the modeling of complex real-world processes. The gradient-descent method is widely used for the recursive estimation of nonlinear fuzzy model parameters using input-output data. The non-robust nature of gradient-descent has motivated many researchers to develop the robust methods of fuzzy identification [19, 20, 51, 57, 62, 68–75, 119, 126].

Our aim is to identify a fuzzy model using example data that will filter out the uncertainties in toxicity modeling problem. We consider an “energy-gain bounding approach” [68] for the identification of such a fuzzy filter. The design criterion is to minimize the maximum possible value of energy-gain from uncertainties to the filtering errors. The maximum value of energy-gain (that will be minimized) is calculated over all possible finite uncertainties without making any statistical assumptions about the nature of uncertainties. It is easy to realize that any practical method of developing a fuzzy filter could not guarantee the 100% filtering of uncertainties. That is, some filtering errors are always involved. Thus, a robustness against filtering errors is rendered by defining fuzzy membership functions.

3.1 A Clustering based Fuzzy Filter and its Identification

This section outlines the mathematical theory of fuzzy filtering directly taken from [77]:

3.1.1 A clustering based fuzzy filter

Fuzzy filter, in our modeling approach, is essentially a mapping between descriptors values and corresponding filtered activity value. We want to create different clusters in descriptor input space and associate to each cluster the output activity value, i.e., filtered $-\log(LC_{50})$ or $\log BCF$ value. The mappings between input descriptors values (denoted by a vector $x = [d_1 \ d_2 \cdots d_n]^T \in R^n$) and output activity value (denoted by a scalar y) can be defined using different fuzzy rules:

$$\begin{aligned} R_1 : &\quad \textbf{If } x \textit{ belongs to a cluster having centre } c_1 \textbf{ then } y = \alpha^1, \\ &\qquad \vdots \\ R_K : &\quad \textbf{If } x \textit{ belongs to a cluster having centre } c_K \textbf{ then } y = \alpha^K, \end{aligned}$$

where $c_i \in R^n$ is the centre of i^{th} cluster, and the values $\alpha^1, \ldots, \alpha^K$ are real numbers. Such clustering based fuzzy mappings have been introduced by authors in [73].

The degree, by which a $n-$dimensional vector x belongs to the i^{th} cluster, can be defined by a fuzzy set, say A_i. Given a universe of discourse X, a fuzzy subset A_i of X is characterized by a mapping:

$$A_i : X \to [0, 1]$$

where for $x \in X$, $A_i(x)$ is a value in the closed interval [0,1] that represents the degree to which x belongs to A_i (i.e. i^{th} cluster). This mapping is called as membership function of the fuzzy set. For a given input vector x (i.e. for a given descriptor values

$d_1, \cdots, d_n$), the output of the filter (i.e. the corresponding filtered activity value) is calculated by aggregating the rules as

$$F(x) = \frac{\sum_{i=1}^{K} \alpha^i A_i(x)}{\sum_{i=1}^{K} A_i(x)}. \tag{3.1}$$

The membership function $A_i(x)$ should be chosen according to some fuzzy clustering criterion. By the method of fuzzy $c-$means (FCM), the membership function $A_i(x)$ must satisfy [13]:

$$\sum_{x \in X} \sum_{i=1}^{K} A_i^{\tilde{m}}(x) \|x - c_i\|^2 \rightarrow \mathit{Minimum}, \ \textstyle\sum_{i=1}^{K} A_i(x) = 1$$

where $\tilde{m} > 1$, is the *fuzzifier* and $\| \cdot \|$ denotes the Euclidean norm. The membership function that minimizes above objective function for a given choice of cluster centres $\{c_i\}_{i=1}^{K}$ follows as

$$FCM_i(x, c_1, \cdots, c_K) = \begin{cases} \dfrac{1}{\sum_{j=1}^{K} \left(\frac{\|x-c_i\|^2}{\|x-c_j\|^2} \right)^{\frac{1}{\tilde{m}-1}}} & \text{for } x \in X \setminus \{c_j\}_{j=1,\cdots,K}, \\ 1 & \text{for } x = c_i, \\ 0 & \text{for } x \in \{c_j\}_{j=1,\cdots,K} \setminus \{c_i\}. \end{cases} \tag{3.2}$$

However, a *possibilistic approach* for c-means clustering relaxes the unit sum constraint on the membership values so that $A_i(x)$ better reflects the typicality of x to the i^{th} cluster [67]. Another approach, called the noise clustering method has been introduced by Davé in [27] to deal with the noisy data. This approach considers noise a separate cluster such that membership of x to the noise cluster is defined as $1 - \sum_{i=1}^{K} A_i(x)$ and the noise prototype is always at the same distance from every point in the data-set. Another possible clustering criterion, assuming a noise cluster outside each data cluster, minimizes

$$J_c(A_i(x), c_1, \cdots, c_K) = \sum_{x \in X} \sum_{i=1}^{K} [A_i(x)\|x - c_i\|^2 + \{1 + A_i(x) \log A_i(x) - A_i(x)\}\delta_i]$$

where the second term in the objective function is intended to be a noise cluster. The term $\{1 + A_i(x)\log A_i(x) - A_i(x)\}$ may be interpreted as the degree to which x does not belong to the i^{th} cluster and thus the membership of x to the noise cluster. If the distance of x to the cluster centre c_i is greater than $\sqrt{\delta_i}$, then the minimization of $J_c(\cdot)$ forces a small value of $A_i(x)$ and a large value of membership of x_i to the noise cluster. Therefore, one of the strategies may be to set δ_i equal to the distance of nearest cluster centre from c_i, i.e., $\delta_i = \min\limits_j \|c_j - c_i\|^2$. Minimizing $J_c(A_i(x), c_1, \cdots, c_K)$ with respect to $A_i(x)$, leads to the following expression for the membership function:

$$RC_i(x, c_1, \cdots, c_K) = \exp(-\frac{\|x - c_i\|^2}{\delta_i}). \tag{3.3}$$

The membership functions of (3.2) and (3.3) can be combined by adopting a mixed clustering criterion [98, 128]. One way to do this is to assume that the membership function A_i has two components A_{1i} and A_{2i} such that

$$A_i = \frac{A_{1i}^{\tilde{m}}}{2} + \frac{A_{2i}}{2}$$

where A_{1i}, A_{2i} minimizes following constrained objective function:

$$\sum_{x \epsilon X} \sum_{i=1}^{K} [\left(A_{1i}^{\tilde{m}}(x) + A_{2i}(x)\right) \|x - c_i\|^2 + \{1 + A_{2i}(x)\log A_{2i}(x) - A_{2i}(x)\}\delta_i], \ \sum_{i=1}^{K} A_{1i}(x) = 1.$$

Now, A_{1i} will be given by (3.2) and A_{2i} by (3.3). Thus,

$$A_i(x, c_1, \cdots, c_K) = \frac{|FCM_i(x, c_1, \cdots, c_K)|^{\tilde{m}}}{2} + \frac{RC_i(x, c_1, \cdots, c_K)}{2}. \tag{3.4}$$

For any membership function $A_i(x)$, defined by (3.2), (3.3), or (3.4), if we define

$$G_i(x, c_1, \cdots, c_K) = \frac{A_i(x, c_1, \cdots, c_K)}{\sum_{i=1}^{K} A_i(x, c_1, \cdots, c_K)},$$

then the output of the fuzzy filter follows from (3.1) as

$$F(x) = \sum_{i=1}^{K} \alpha^i G_i(x, c_1, \cdots, c_K).$$

Introduce the notations: $\alpha = [\alpha^i]_{i=1,\ldots,K} \in R^K$, $\theta = [c_1^T \cdots c_K^T]^T \in R^{Kn}$, and $G(x,\theta) = [G_i(x,\theta)]_{i=1,\ldots,K} \in R^K$, so that output of the fuzzy filter for an input x can be expressed as

$$F(x) = G^T(x,\theta)\alpha.$$

Thus, a fuzzy filter is characterized by linear parameters α and non linear cluster centre parameters θ.

3.1.2 Robust identification of the fuzzy filter

The identification of the fuzzy filter using input-output data $\{x(j), y(j)\}_{j=0}^{k}$ involves the estimation of fuzzy filter parameters (α, θ). Here, $x(j)$ is the j^{th}-indexed input data (i.e. descriptor values) and $y(j)$ is the corresponding experimentally measured activity value. Assume that there exist some true fuzzy filter, characterized by parameters $(\alpha^*, \{\theta_j^*\}_{j=0}^{k})$, such that true fuzzy filter is functionally equivalent to the

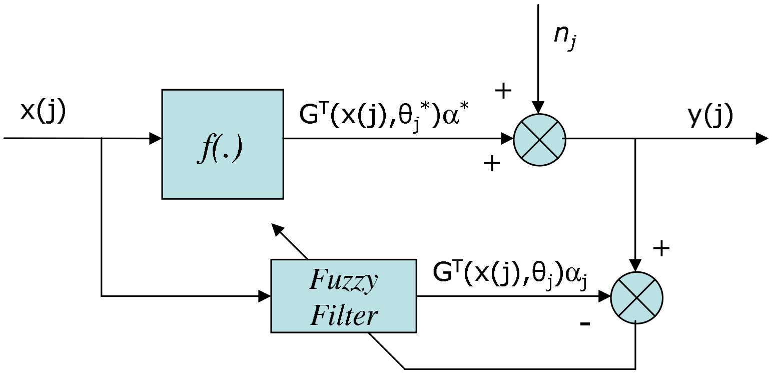


Figure 3.1: Identification of a fuzzy filter

unknown function $f(\cdot)$ of Fig. 1.1 in chapter 1, as illustrated in Fig. 3.1. That is,

$$y(j) = G^T(x(j), \theta_j^*)\alpha^* + n_j$$

where n_j is the uncertainty in j^{th}-indexed data. Let (α_j, θ_j) denote an estimate of (α^*, θ_j^*) using data $\{x(i), y(i)\}_{i=0}^{j}$ based on some recursive estimation strategy. The filtering error for j^{th}-indexed data is given as

$$e_j = G^T(x(j), \theta_j^*)\alpha^* - G^T(x(j), \theta_j)\alpha_j.$$

Any estimation strategy will be considered performing good if it results in a small energy of filtering errors, being measured as $\sum_{j=0}^{k} |e_j|^2$. The performance of any estimation strategy will be affected by three kind of unknown disturbances:

- the energy of uncertainties, $\sum_{j=0}^{k} |n_j|^2$,
- deviation of initial guess α_{-1} from true parameter α^*, assessed as $\|\alpha^* - \alpha_{-1}\|^2$,
- deviation of $\{\theta_j^*\}_{j=0}^{k}$ from their initial guess $\{\theta_{j-1}\}_{j=0}^{k}$, assessed as $\sum_{j=0}^{k} \|\theta_j^* - \theta_{j-1}\|^2$. Here, we follow the approach of [68], where the initial guess about θ_j^* is taken equal to the estimate of θ_{j-1}^*.

We are concerned with a robust identification method that is least sensitive to the disturbances. Our approach to the robust identification of fuzzy filter is based on energy-gain bounding criterion [68]:

$$\min_{\{\alpha_j, \theta_j\}_{j=0}^{k}} \max_{\alpha^*, \{\theta_j^*\}_{j=0}^{k}, \{n_j\}_{j=0}^{k}} \frac{\sum_{j=0}^{k} |G^T(x(j), \theta_j^*)\alpha^* - G^T(x(j), \theta_j)\alpha_j|^2}{\mu^{-1}\|\alpha^*\|^2 + \mu_\theta^{-1} \sum_{j=0}^{k} \|\theta_j^* - \theta_{j-1}\|^2 + \sum_{j=0}^{k} |n_j|^2}$$

where μ and μ_θ are positive constants. The identification method minimizes the maximum possible value of energy-gain from disturbances to the filtering errors. Such an identification method will guarantee that *small disturbances can not lead to large filtering errors.* The maximum value of energy-gain (that will be minimized) is calculated over all possible finite disturbances without making any statistical assumptions about the nature of signals. It follows from [68] that fuzzy filter parameters, based on energy-gain approach, are identified by performing for $j = 0, \cdots, k$, the recursions

$$\theta_j = \arg\min_{\theta} \left[\frac{|y(j) - G^T(x(j), \theta)\alpha_{j-1}|^2}{1 + \mu \|G(x(j), \theta)\|^2} + \mu_\theta^{-1} \|\theta - \theta_{j-1}\|^2 \right],$$

$$\alpha_j = \alpha_{j-1} + \frac{\mu G(x(j), \theta_j) \left[y(j) - G^T(x(j), \theta_j)\alpha_{j-1} \right]}{1 + \mu \|G(x(j), \theta_j)\|^2}, \ \alpha_{-1} = 0.$$

The optimal values of parameters (μ, μ_θ) in these recursions which result in the fast convergence and low steady-state error are given as

$$\mu(j) = \frac{\|\widehat{p}_j\|^2}{C\|G(x(j), \theta_j)\|^2}, \ \mu_\theta(j) = s_\theta \frac{\|\widehat{p}_j\|^2}{C\|G(x(j), \theta_j)\|^2}, \tag{3.5}$$

$$\widehat{p}_j = \omega \widehat{p}_{j-1} + (1 - \omega) \frac{y(j) - G^T(x(j), \theta_j)\alpha_{j-1}}{\|G(x(j), \theta_j)\|^2} G(x(j), \theta_j), \tag{3.6}$$

where s_θ is a predefined positive constant, ω $(0 < \omega < 1)$ is a smoothing factor, and C is a positive constant that should be chosen proportional to the magnitude of uncertainties.

3.2 Improving Modelling Performance via Fuzzy Filtering

The fuzzy filtering based approach to the chemicals' behavior modeling and prediction will be described in two parts. The first part involves the development of QSAR models and the second one deals with the implementation of developed QSAR models for prediction.

3.2.1 Development

The development procedure involves following three steps:

Identification of a fuzzy filter

Given the data of N training compounds i.e. descriptor values and experimentally measured activity data $\{x(j), y(j)\}_{j=0}^{N-1}$, identify a fuzzy filter using the method described in earlier section. The identification method can be implemented using a Gauss-Newton based algorithm suggested in Appendix A.

Filtering out the uncertainties from activity data

The output of the identified fuzzy filter represents the filtered activity value. If we denote the parameters of identified fuzzy filter by (α^I, θ^I), then the filtered activity value of $j^{th}-$indexed compound is given as

$$y_f(j) = G^T(x(j), \theta^I)\alpha^I. \tag{3.7}$$

Due to the filtering errors, it may be the case that uncertainties are not filtered by 100% i.e. $y_f(j) \neq f(x(j))$. Thus, for any further analysis of values $\{y_f(j)\}_{j=0}^{N-1}$, different fuzzy sets (such as *low*, *medium*, *high*) are defined for filtered activity data. The membership functions (i.e. fuzzy sets) would provide some tolerance, against

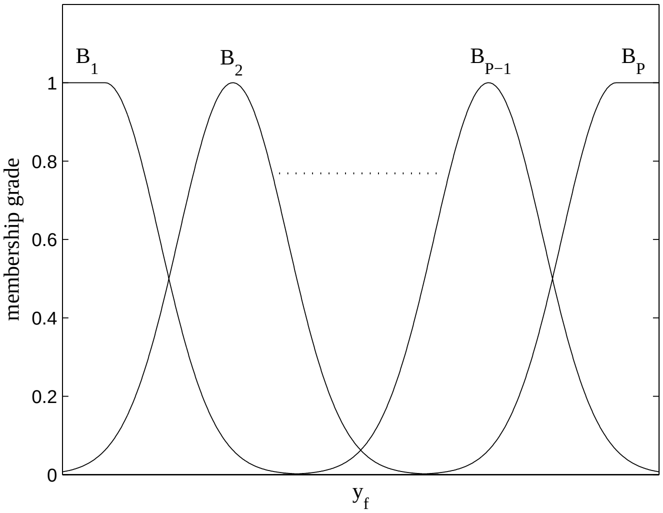


Figure 3.2: Defining membership functions for filtered activity y_f

the uncertainty lying in $y_f(j)$ due to the filtering errors, in any further analysis. As an illustration, P different fuzzy sets, represented as $B_1, B_2, \cdots, B_P$, are shown in Fig. 3.2.

Developing QSAR models using filtered activity data

To each of the fuzzy sets $B_1, B_2, \cdots, B_P$, some training data could be associated. This leads to the creation of P different data sets, $D_1, D_2, \cdots, D_P$, out of total data $\{x(j), y_f(j)\}_{j=0}^{N-1}$. Mathematically,

$$D_i = \{x(j), y_f(j) : \ 0 \leq j \leq N-1, \ B_i \left(y_f(j)\right) \geq \epsilon\}, \text{ where } 0 \leq \epsilon << 1.$$

Here, $B_i\left(y_f(j)\right)$ represents degree or grade to which $y_f(j)$ belongs to fuzzy set B_i. In other words, the data set D_i contains all those training compounds whose filtered activity value belongs to fuzzy set B_i at least by a degree of ϵ. Now, P different QSAR models $M_1, M_2, \cdots, M_P$, could be trained using data sets $D_1, D_2, \cdots, D_P$, respectively. The main point here is that the models are trained using filtered activity data, i.e., there are no disturbances (due to uncertainties) affecting aversely the training algorithm performance. Therefore, using even a non-robust training algorithm should not cause a loss in the generalization performance of the models.

> Remark: If models $M_1, \cdots, M_P$ are trained with a robust training algorithm (whose performance is not aversely affected by uncertainties), then data sets $(D_1, \cdots, D_P)$ are defined as
>
> $$D_i = \{x(j), y(j) : \ 0 \leq j \leq N-1, \ B_i \left(y_f(j)\right) \geq \epsilon\}, \text{ where } 0 \leq \epsilon << 1.$$

3.2.2 Implementation

Once the fuzzy filter, fuzzy sets $B_1, \cdots, B_P$, and QSAR models $M_1, \cdots, M_P$ are ready, the prediction of activity of any new compound follows by combining in a suitable manner the contributions of models $M_1, \cdots, M_P$, as shown in Fig. 3.3. Given a new compound's descriptors $x = [d_1, d_2, \cdots, d_n]^T$, the activity value can be predicted as

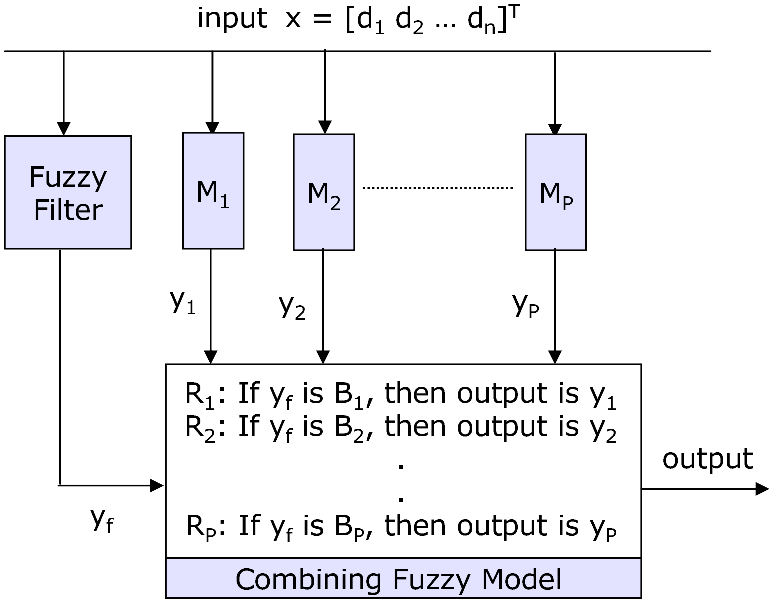

input x = [d1 d2 … dn]T
Fuzzy Filter
M1
M2
MP
y1
y2
yP
yf
R1: If yf is B1, then output is y1
R2: If yf is B2, then output is y2
.
.
RP: If yf is BP, then output is yP
Combining Fuzzy Model
output

Figure 3.3: Fuzzy filtering based approach to activity prediction

follows:

1. Compute the output of the fuzzy filter, $y_f = G^T(x, \theta^I)\alpha^I$.
2. Compute the output of models i.e. the values $y_1, \cdots, y_P$, where $y_i = M_i(x)$.
3. Combine the values $y_1, \cdots, y_P$ according to following fuzzy rule base:

$$R_i : \textbf{ If } y_f \text{ is } B_i \textbf{ then } \text{output} = y_i,\ i = 1, 2, \cdots, P$$

The predicted output value, $\hat{y}$, could be computed by taking the weighted average of the outputs provided by P rule:

$$\hat{y} = \frac{\displaystyle\sum_{i=1}^{K} y_i B_i(x)}{\displaystyle\sum_{i=1}^{K} B_i(x)}.$$

3.3 Toxicity Modelling Problem

The toxicity modelling problem of section 2.1 of chapter 2 is re-visited based on the ideas developed in this chapter. Now, it will be shown that the performance of the algorithms listed in table 2.1 of chapter 2 improves by the fuzzy handling of uncertainties. A fuzzy filter of 30 rules, with membership functions defined by (3.4) for $\tilde{m} = 2$, is identified based on the energy-gain bounding approach described in section 3.1. The identification method is implemented in MATLAB 6.5 using a Gauss-Newton based algorithm proposed in Appendix A. The identification parameters involved in (3.5-3.6) are chosen as $s_\theta = 0.05, \omega = 0.99$, and $C = 10$. The initial guess about cluster centres is taken by performing fuzzy c-means clustering on the training data. The identification algorithm was run till 100 epochs. The identified fuzzy filter is used to filter out the uncertainties and finding out the filtered toxicity data of the compounds, i.e., $\{y_f(j)\}_{j=0}^{567}$ values using (3.7).

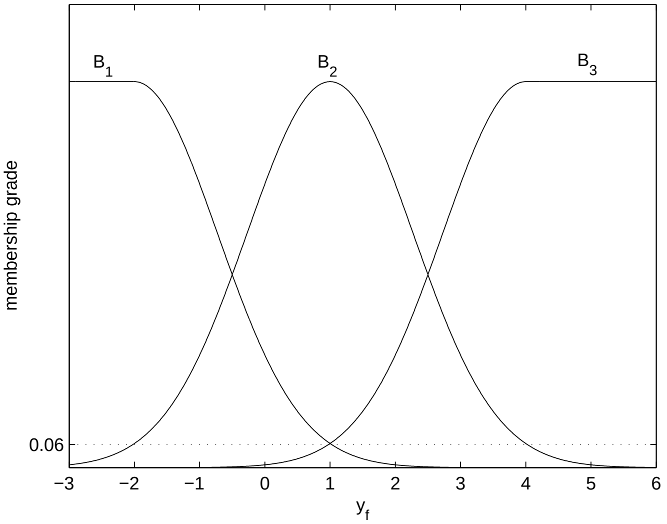


Figure 3.4: The membership functions for filtered toxicity data

We define three membership functions for the filtered toxicity values as shown in Fig. 3.4. Associated to the fuzzy sets of Fig. 3.4, three data sets (D_1, D_2, D_3) are created taking $\epsilon = 0.06$. Each of the training algorithms listed in table 2.1 of chapter 2 is used to train three different models M_1, M_2, M_3 using data sets D_1, D_2, D_3

respectively. Here, M_1, M_2, M_3 are neural networks with the same structure and initial conditions as of models of table 2.1. Finally, the three trained models are combined, as illustrated in Fig. 3.3.

Again, SOM is applied to the descriptors values and filtered toxicity data of 568 compounds. The distribution of the compounds on a map (with 17 × 7 grid with hexagonal lattices) is shown in Fig. 3.5. Unlike Fig. 2.2 of chapter 2, there are no uncertainties involved in Fig. 3.5. This is obviously due to the filtering action of the fuzzy system.

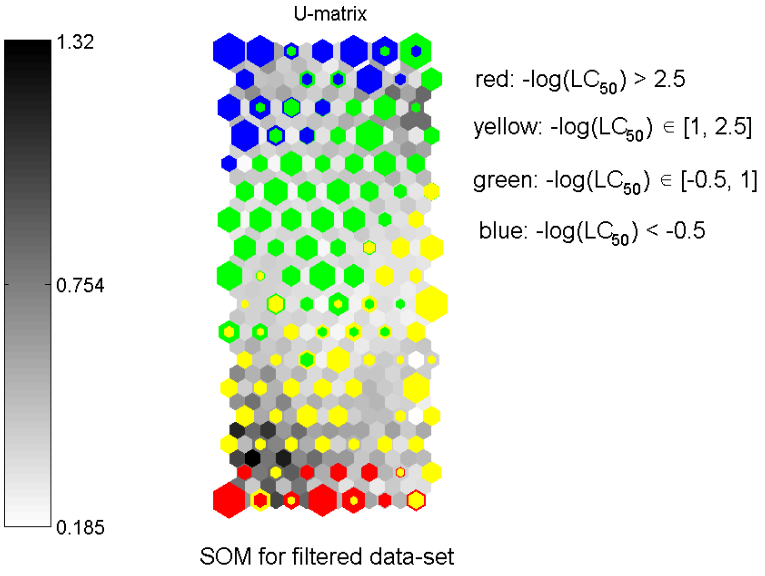


Figure 3.5: Distribution of 568 compounds on the map with uncertainties being filtered out

Our approach to toxicity modeling, illustrated in Fig. 3.3, is meant for rendering robustness to any algorithm used for training M_1, M_2, M_3 with data sets D_1, D_2, D_3 respectively. Table 3.1 shows the performance of different training algorithms in the proposed fuzzy filtering based method. A comparison of R^2 values for testing data between table 2.1 and 3.1 clearly shows an improvement in the generalization performance of all the algorithms. For example, "scaled conjugate gradient" algorithm shows a poor generalization performance in table 2.1 as indicated by a low R^2 and high MAE on the testing data. However, the generalization performance of the corresponding model in table 3.1 (i.e. N_5^f) is far better than N_5. As seen from table 3.1, Bayesian regularized learning of networks produced a model, N_8^f, that has best generalization performance (i.e. highest R^2 and lowest MAE value on testing data).

Table 3.1: Performances of different training algorithms in fuzzy filtering based toxicity modeling approach

Model	Training algorithm for M_1, M_2, M_3	R^2 *training*	R^2 *testing*	MAE *training*	MAE *testing*
N_1^f	Batch Gradient Descent learning rate = 0.05	0.7163	0.6743	2.4737	2.7327
N_2^f	Batch Gradient Descent with Momentum learning rate = 0.05 momentum constant = 0.9	0.7150	0.6437	2.6055	2.8382
N_3^f	Resilient Backpropagation	0.6425	0.6212	3.5268	2.9807
N_4^f	Conjugate Gradient (Fletcher-Reeves) Charalambous search method	0.6407	0.6036	3.5599	2.6762
N_5^f	Scaled Conjugate Gradient	0.6430	0.6275	3.4192	2.5194
N_6^f	One Step Secant algorithm Backtracking search method	0.6424	0.6001	3.6632	3.5620
N_7^f	Levenberg-Marquardt	0.6407	0.6098	3.7299	3.6575
N_8^f	Bayesian regularization	0.7324	0.6831	2.7526	2.4699

Fig. 3.6 shows the plots of predicted toxicity values using model N_8^f. Table 3.2 illustrates the performance of N_8^f on the first 30 testing compounds for which the prediction accuracy is highest. The maximum absolute prediction error (i.e. MAE) of N_8^f on testing data was 2.4699.

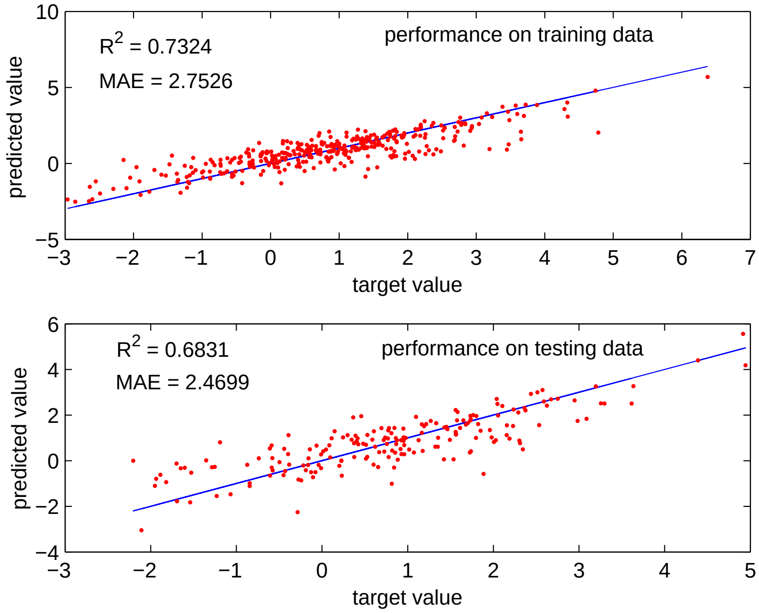


Figure 3.6: Toxicity prediction using model N_8^f

3.4 Summary

Achieving good generalization capability is the key concern in QSAR studies. Uncertainties are involved while establishing the model mappings between molecular descriptors values and activity data. These uncertainties arise due to noisy data and non-optimal (or sub-optimal) choice of descriptors, model type, and model structure. The non-robust training algorithms, due to the involved uncertainties, produce models with a low generalization performance, as seen the performance of models N_3, N_4, N_5, N_6, N_7 in table 2.1 of chapter 2. Thus, the robustness of the training algorithm against uncertainties is desired. The authors in [16–18,125] suggest Bayesian regularized neural networks as robust QSAR models. The robustness property of Bayesian regularized neural networks could be also observed in table 2.1.

This study handles the uncertainties using a fuzzy filter and thus the uncertainties are not allowed to affect the training algorithm performance. This improves the generalization performance of the training algorithms. The effectiveness of the fuzzy filtering approach in the toxicity modeling example can be seen by noting that

Table 3.2: The performance of model N_8^f on some of the testing compounds

	Compound	CAS	$-\log(LC_{50}(mmol/l))$	predicted value
1	2,4,6-Triiodophenol	609-23-4	2.59	2.59
2	4-Chloroaniline	106-47-8	0.62	0.62
3	Bis(*p*-fluorophenyl) ether	330-93-8	2.24	2.24
4	Permethrin	52645-53-1	4.39	4.40
5	6-Chloro-2-pyridinol	16879-02-0	-0.22	-0.21
6	Dibutyl terephthalate	1962-75-0	2.67	2.69
7	2,4-Dinitrotoluene	121-14-2	0.87	0.89
8	2-Ethoxyethyl methacrylate	2370-63-0	0.76	0.74
9	γ-Decanolactone	706-14-9	0.98	1.00
10	1-Chloro-3-nitrobenzene	121-73-3	0.92	0.90
11	2-Bromo-3-pyridinol	6602-32-0	-0.43	-0.46
12	Pentachloroethane	76-01-7	1.43	1.46
13	1,1,1,3,3,3-Hexafluoro-2-propanol	920-66-1	-0.16	-0.19
14	*n*-Octylcyanide	2243-27-8	1.45	1.48
15	Salicylanilide	87-17-2	1.73	1.77
16	2,3,4,5-Tetrachlorophenol	4901-51-3	2.75	2.71
17	1-Bromooctane	111-83-1	2.36	2.31
18	4-Nitrobenzamide	619-80-7	0.10	0.14
19	2,4,5-Trimethyloxazole	20662-84-4	-0.61	-0.65
20	4-Butylaniline	104-13-2	1.17	1.22
21	Pentyl ether	693-65-2	1.70	1.65
22	4-*n*-Nonyl phenol	104-40-5	3.2	3.26
23	(1S)-(-)-Camphor	464-48-2	0.95	1.02
24	2-Undecanone	112-12-9	2.06	1.98
25	*p*-*tert*-Butylphenol	98-54-4	1.46	1.38
26	3,5-Diiodo-4-hydroxybenzonitrile	1689-83-4	1.74	1.82
27	1-(2-Hydroxyethyl)piperazine	103-76-4	-1.69	-1.78
28	2-Allylphenol	1745-81-9	0.95	0.86
29	1-Bromohexane	111-25-1	1.68	1.58
30	Diphenylamine	122-39-4	1.65	1.77

- All models of table 3.1 (i.e. $N_1^f, \cdots, N_8^f$) have better generalization performance than their counterparts in table 2.1 (i.e. $N_1, \cdots, N_8$).

- The generalization performance of a QSAR model in table 3.1, unlike table 2.1, is not so sensitive towards the choice of training algorithm. For example, models N_7 and N_1 have a remarkable difference in their performances on the testing data, while this is not the case with N_7^f and N_1^f.

The difficulties in modeling the considered toxicity data have been already illustrated in [88], where different neural and fuzzy-neural networks were trained with the data set. Even the model with the best performance on testing data was found to have $R^2 = 0.5019$ (see table 5 in [88]). Thus, the chosen data set is a good example of illustrating the robustness of our approach. The QSAR models trained with this data set, because of the heterogeneity of data, should be quite general.

Chapter 4

Incorporating Intelligence in Modelling[*]

The previous chapter introduced a fuzzy filter for separating the uncertainties from the modelling problem. In this chapter, our goal is to incorporate an intelligence in a given modelling technique based on the fuzzy filter provided information about uncertainties. By intelligence we mean a capability of taking account of uncertainties in a sensible way during the development of the models for an improvement in the generalization performance of the modelling technique.

Given the descriptors-activity data of N training compounds $\{x(k), y(k)\}_{k=1}^{N}$, our approach to incorporate intelligence in a given modelling techniques is based on the following ideas:

1. A fuzzy filter is constructed using data $\{x(k), y(k)\}_{k=1}^{N}$ that would filter out any uncertainties arising due to the compounds behaving differently from the input-output data trend. For a compound, described by descriptors values $x(k)$, the fuzzy filter is used to obtain a filtered $y(k)$ value, denoted as $y_f(k)$. That is, N data-pairs $\{x(k), y_f(k)\}_{k=1}^{N}$ follow, without an exception, a trend of input-output mappings. The uncertainty associated to the compound is assessed as

[*] This work has been submitted as
S. Kumar, M. Kumar, K. Thurow, R. Stoll, and U. Kragl. Fuzzy Filtering for Robust Bioconcentration Factor Modelling. *Environmental Modelling and Software*, submitted for publication.

$\hat{n}_k = y(k) - y_f(k)$.

2. The uncertainties $\{\hat{n}_k\}_{k=1}^N$ and filtered output values $\{y_f(k)\}_{k=1}^N$ are assumed to have been produced by a set of random sources. We estimate the parameters of these random sources via modelling the N number of 2-dimensional data points $\{z_k = [\ y_f(k) \quad \hat{n}_k\ ]^T \in R^2\}_{k=1}^N$ using finite mixture models [89,91]. That is, we estimate the parameters of a set of probability density functions such that each data point z_k is modelled as having been generated by one of the probabilistic models in the set.

3. The finite mixture modelling leads to the clustering of the data via identifying which source (i.e. probabilistic model) produced each data point. Assume that C different sources, with the known probability density functions, have been identified producing the data $\{z_k\}_{k=1}^N$.

4. The data points associated to a source could be used to train (i.e. develop) a local model. A local model M_i (associated to the i^{th} source), if trained using a non-robust algorithm conventionally with data $\{x(k), y(k)\}$, may lead to a poor generalization performance. The reason being that in the training of model M_i, the data points associated to a higher magnitude of uncertainties might act as outliers and adversely affect the training of the model. Therefore, we want to train the models with some *penalized* data $\{x(k), y_p^i(k)\}$.

5. For any k^{th} data point used in the training of model M_i, the output value $y(k)$ is penalized (in a context of the i^{th} source) for the magnitude of the uncertainty associated to the k^{th} data point. This is done via defining a penalized output value $y_p^i(k)$ such that $y_p^i(k)$ is closer to $y(k)$ for the data points being treated as "regular" (typically characterized by a lower magnitude of estimated uncertainties), while $y_p^i(k)$ is closer to $y_f(k)$ for the data points being treated as outliers (typically characterized by a higher magnitude of estimated uncertainties). To define the penalized value y_p^i, we make use of the i^{th} probabilistic model provided information about the uncertainties.

6. A model M_i (associated to the i^{th} source) is not trained conventionally using data $\{x(k), y(k)\}$, however, trained using penalized data $\{x(k), y_p^i(k)\}$. Now, for the data points (might being acting as outliers), y_p^i is closer to y_f (i.e. closer to a point free from uncertainties) and thus training the model using y_p^i values should not adversely affect the training method.

7. Finally, the C different local models $M_1, \cdots, M_C$ are combined to estimate the final output.

Roughly speaking, our approach renders robustness in the identification of local models $M_1, \cdots, M_C$ via penalizing the data. The local models operate in the predefined regions. To penalize the data, as will be explained, we make use of the information about uncertainties provided by a fuzzy filter. The design of the fuzzy filter is based on the "energy-gain bounding approach" [68]. This approach improves the method of previous chapter in the followings:

1. In previous chapter, the local models are developed in the partitions of 1-dimensional real line of filtered values. In this study we partition the 2-dimensional space of filtered values and uncertainties, since the information about uncertainties will be used to penalize the data.

2. The method of previous chapter trains the models with the filtered data $\{x(k), y_f(k)\}$ and thus there are no uncertainties (in the training data) that could adversely affect the training procedure. However, here we use the penalized data $\{x(k), y_p^i(k)\}$ for the training of the models, offering the flexibility of "smooth switching" between $\{x(k), y(k)\}$ (for regular data points) and $\{x(k), y_f(k)\}$ (for outliers).

3. The previous method, unlike this one, penalizes all the data points (regular as well as outliers) and thus is over conservative.

4.1 The Methodology

The methodology for incorporating intelligence in a given modelling problem (i.e. given molecular descriptors, model type, model structure, and model identification

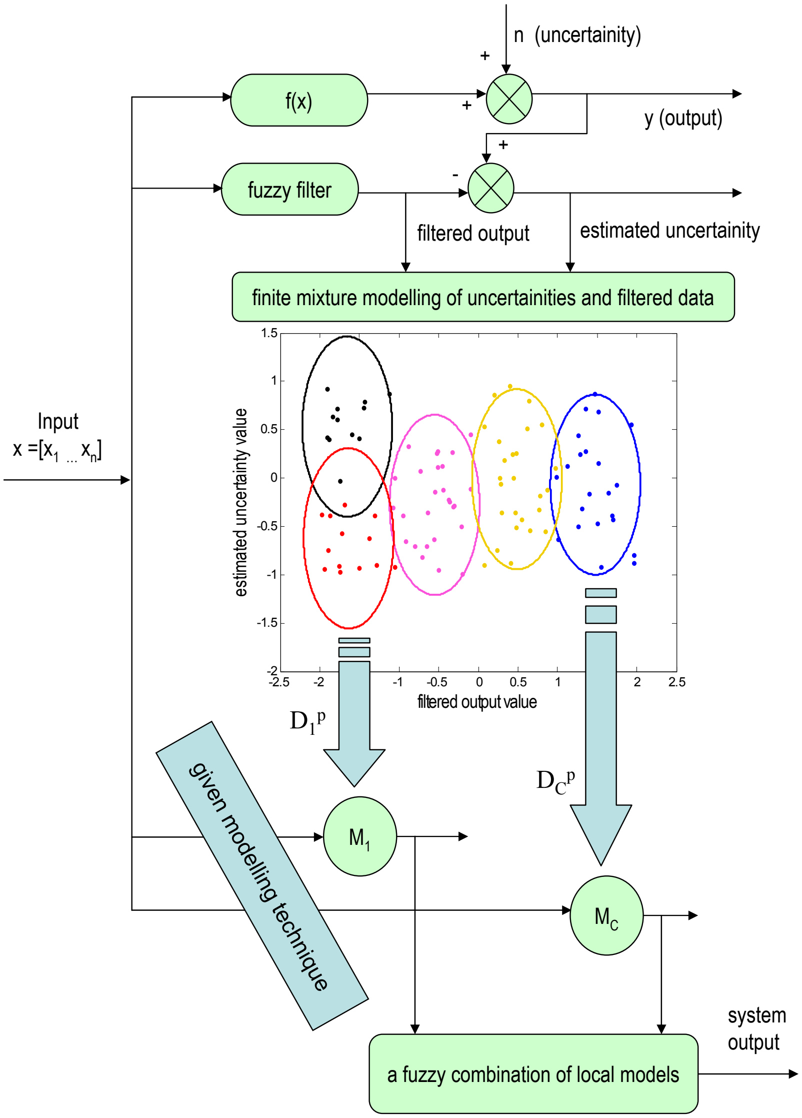


Figure 4.1: An intelligence is incorporated in a given modelling technique by using penalized data sets $D_1^p, \cdots, D_C^p$. The penalized data sets and a fuzzy rule base for combining the local models are carefully designed based on Gaussian mixture modelling of filtered data and uncertainties

algorithm) consists of following steps:

4.1.1 Identification of the parameters of a fuzzy filter

A fuzzy filter is identified based on the ideas outlined in section 3.1 using a Gauss-Newton based algorithm of Appendix A. For a choice of the number of rules in the fuzzy filter (i.e. number of clusters K) and initial guess about cluster centres θ_{-1}, a clustering on input data (e.g. using finite mixture models [40]) could be performed.

The output of the identified fuzzy filter represents the filtered output value. If we denote the parameters of identified fuzzy filter by (α^I, θ^I), then the filtered output value of $k^{th}-$indexed compound is given as

$$y_f(k) = G^T(x(k), \theta^I)\alpha^I. \tag{4.1}$$

The uncertainty associated to the $k^{th}-$indexed compound will be assessed as

$$\hat{n}_k = y(k) - y_f(k). \tag{4.2}$$

4.1.2 Gaussian mixture modelling of filtered data and uncertainties

Assume that the vector $z_k = [\ y_f(k) \quad \hat{n}_k\ ]^T$ represents one particular outcome of a 2-dimensional random variable $\mathbf{Z} \in R^2$ whose probability density function can be written as a mixture of the Gaussian distributions:

$$p(z) = \sum_{i=1}^{C} a_i p(z \mid m_i, \Sigma_i), \text{such that} \tag{4.3}$$

- the mixing probabilities $a_1, \cdots, a_C$ satisfy $a_i \geq 0$ and $\sum_{i=1}^{C} a_i = 1$,
- the parameters $m_i \in R^2$, Σ_i (a 2×2 positive definite matrix) characterize fully

the i^{th} Gaussian component:

$$p(z \mid m_i, \Sigma_i) = \frac{1}{\sqrt{(2\pi)^2|\Sigma_i|}} \exp\{-\frac{1}{2}(z - m_i)^T \Sigma_i^{-1}(z - m_i)\}. \tag{4.4}$$

An approach to the clustering of data $\{z_k\}_{k=1}^N$ is to fit finite mixture models (4.3) to the data, where a component distribution is used to model a specific cluster. That is, i^{th} cluster (with mean m_i and covariance Σ_i) is mathematically represented by Gaussian distribution $p(z \mid m_i, \Sigma_i)$. "Expectation-maximization" (EM) is the standard algorithm [90, 91] used to fit finite mixture models to data. In this study, however, we use the algorithm of [40] for estimating the parameters of the mixture (4.3). This algorithm is capable of automatically selecting the number of components C. The algorithm, unlike EM, is less sensitive to initialization and avoids the possibility of algorithm convergence to the boundary of the parameter space. As an illustration,

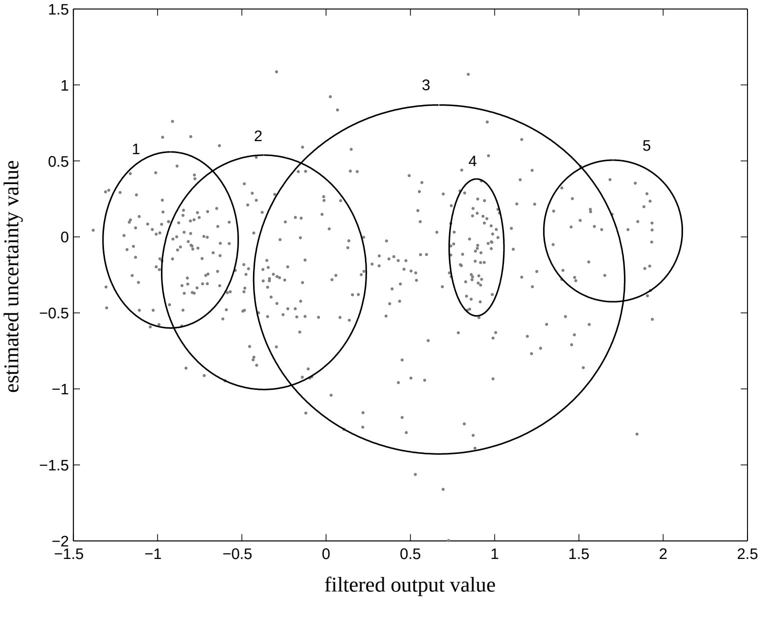


Figure 4.2: Gaussian mixture modelling of data: data points and level-curves (solid line) for the different components

Fig. 4.2 shows the Gaussian mixture modelling of an example data where the drawn ellipses are the level-curves of component distributions. The data points in Fig. 4.2

could be clustered via associating each point to one of the 5 components. The matrix Σ_i in (4.4) could be chosen to be a diagonal matrix (i.e. the two random variables are independent). If $m_i = \begin{bmatrix} m_i^1 \\ m_i^2 \end{bmatrix}$ and $\Sigma_i = \begin{bmatrix} \Sigma_i^1 & 0 \\ 0 & \Sigma_i^2 \end{bmatrix}$, then

$$p(z \mid m_i, \Sigma_i) = p(y_f \mid m_i^1, \Sigma_i^1) p(\hat{n} \mid m_i^2, \Sigma_i^2), \text{ where} \tag{4.5}$$

$$p(y_f \mid m_i^1, \Sigma_i^1) = \frac{\exp\{-\frac{(y_f - m_i^1)^2}{2\Sigma_i^1}\}}{\sqrt{2\pi\Sigma_i^1}}, \; p(\hat{n} \mid m_i^2, \Sigma_i^2) = \frac{\exp\{-\frac{(\hat{n} - m_i^2)^2}{2\Sigma_i^2}\}}{\sqrt{2\pi\Sigma_i^2}}. \tag{4.6}$$

The data points in Fig. 4.2 are taken from a case study to be discussed in the latter part of the chapter.

4.1.3 A combination of local models

Given the knowledge of component distributions $p(z \mid m_i, \Sigma_i), \; i = 1, \cdots, C$, we want to utilize this information in the development of neural, fuzzy, or of any other type local models $(M_1, \cdots, M_C)$ valid in the predefined operating regions. The operating regions can be represented by fuzzy sets and the local models can be combined using a fuzzy rule base:

$$R_1 : \begin{array}{c} \text{For input } x \text{, if the fltered value } y_f = G^T(x, \theta^I)\alpha^I \text{ is } A_1(y_f), \\ \text{then output } = M_1(x) \end{array} \quad , [w_1]$$

$$\vdots$$

$$R_C : \begin{array}{c} \text{For input } x \text{, if the fltered value } y_f = G^T(x, \theta^I)\alpha^I \text{ is } A_C(y_f), \\ \text{then output } = M_C(x) \end{array} \quad , [w_C]$$

Here, $(A_1(y_f), \cdots, A_C(y_f))$ are the membership functions, $M_i(x)$ denotes the i^{th} model output for the input x, and $w_i \in [0, 1]$ is the weight of the rule that represents the belief in the accuracy of the i^{th} rule R_i. The *degree of fulfillment* of the i^{th} rule is given by $\beta_i(y_f) = w_i A_i(y_f)$. The overall output y, for input x, is estimated by taking the weighted average of the output provided by each rule:

$$y = \frac{\sum_{i=1}^{C} w_i A_i(y_f) M_i(x)}{\sum_{i=1}^{C} w_i A_i(y_f)}.$$

We want to define the membership function $A_i(y_f)$ in such a way that the data points, belonging to the region covered by $A_i(y_f)$, are most likely to be generated by the i^{th} probabilistic model $p(y_f \mid m_i^1, \Sigma_i^1)$. This is done by defining $A_i(y_f)$ as follows

$$A_i(y_f) = k_n p(y_f \mid m_i^1, \Sigma_i^1),\ i = 1, \cdots, C \tag{4.7}$$

where k_n is a normalizing constant that ensures that $A_i(y_f) \in [0, 1]$. In view of this choice of the membership functions, the natural choice of the rule weight w_i is the prior probability of observing a data point from i^{th} source i.e. $w_i = a_i$. Thus, the overall output by combining the local models is given as

$$y = \frac{\sum_{i=1}^{C} a_i p(y_f \mid m_i^1, \Sigma_i^1) M_i(x)}{\sum_{i=1}^{C} a_i p(y_f \mid m_i^1, \Sigma_i^1)}. \tag{4.8}$$

4.1.4 The development of local models

One would normally expect to train a local model M_i (associated to fuzzy set $A_i(y_f(k))$) with input-output data set D_i defined as

$$D_i = \{x(k), y(k),\ 1 \leq k \leq N,\ A_i(y_f(k)) \geq \epsilon\},\ 0 \leq \epsilon << 1. \tag{4.9}$$

The data set D_i contains all those training compounds whose filtered output value belongs to fuzzy set A_i at least by a degree of ϵ. As an illustration, the output values of set D_i have been displayed (marked as "·") in Fig. 4.3. However, as stated earlier,

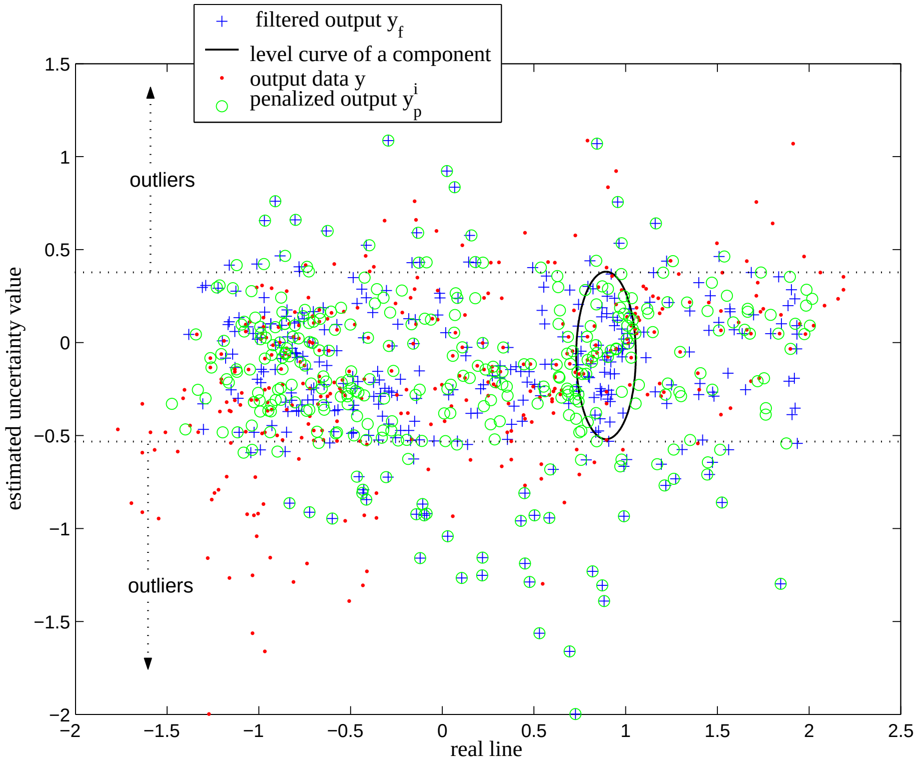


Figure 4.3: Display of data output y, filtered output y_f, and penalized output y_p^i

the training of M_i with data set D_i using a non robust algorithm may lead to a poor generalization performance of the model. The reason being that in the training of model M_i, the data points lying far away from i^{th} cluster centre along the estimated-uncertainty-axis might act as outliers and adversely affect the training of the model. Therefore, we want to train the models with some *penalized* data $\{x(k), y_p^i(k)\}$. A penalized value $y_p^i(k)$ is defined such that $y_p^i(k)$ is closer to $y(k)$ for the data points being treated as "regular" (lying closer to the i^{th} cluster centre), while $y_p^i(k)$ is closer to $y_f(k)$ for the data points being treated as outliers (far away from i^{th} cluster centre along the estimated-uncertainty-axis). Now, for the data points (might being acting as outliers), y_p^i is closer to y_f (i.e. closer to a point free from uncertainties) and thus training the model M_i using $\{x(k), y_p^i(k)\}$ values should not adversely affect the training method.

Fig. 4.3 displays an example of the penalized values (marked as "o"), shifting from

$\{y(k)\}$ (marked as "$\cdot$") to the $\{y_f(k)\}$ (marked as "+"), as we move away from the cluster centre along the estimated-uncertainty-axis. To define the penalized values, we make use of the information (provided by i^{th} probabilistic model) on uncertainties. One of the possible methods for defining the penalized values is as follows:

$$y_p^i(k) = \bar{\omega}_k^i y_f(k) + (1 - \bar{\omega}_k^i) y(k), \text{ where} \tag{4.10}$$

$$\bar{\omega}_k^i = \left(1 - \frac{p(\hat{n}_k \mid m_i^2, \Sigma_i^2)}{p_{max}^i}\right)^{s_p}, \; p_{max}^i = \max_k \; p(\hat{n}_k \mid m_i^2, \Sigma_i^2), \; s_p > 0. \tag{4.11}$$

Here, s_p is a "switching parameter" that controls the rate at which the switching of y_p^i from y to y_f, with a decrease in $p(\hat{n}_k \mid m_i^2, \Sigma_i^2)$ (i.e. while moving away from i^{th} cluster centre along the estimated-uncertainty-axis), takes place. A lower value of s_p results in a faster switching and vice-versa. Let D_i^p denotes the penalized training data set for M_i:

$$D_i^p = \{x(k), y_p^i(k), \; 1 \le k \le N, \; A_i(y_f(k)) \ge \epsilon\}, \; 0 \le \epsilon << 1. \tag{4.12}$$

Finally, the data sets $D_1^p, \cdots, D_C^p$ could be used to train the local models $M_1 \cdots, M_C$ respectively. Fig. 4.1 summarizes our methodology for incorporating intelligence in a given modelling technique.

4.1.5 Implementation of the methodology for prediction

The given training data $\{x(k), y(k)\}_{k=1}^N$ is used to estimate the parameters (α^I, θ^I), $\{(m_i^1, \Sigma_i^1), (m_i^2, \Sigma_i^2), a_i, \; i = 1, \cdots, C\}$ and thus the training of local models $M_1, \cdots, M_C$ is accomplished. Now, the prediction of the output value for a given input (i.e. prediction of activity of a compound that may or may not be included in training set) follows as

- for an input x, compute the filtered output $y_f = G^T(x, \theta^I)\alpha^I$,

- the outputs of the local models could be combined to predict the output according to (4.8):

$$y = \frac{\sum_{i=1}^{C} a_i p(y_f \mid m_i^1, \Sigma_i^1) M_i(x)}{\sum_{i=1}^{C} a_i p(y_f \mid m_i^1, \Sigma_i^1)}, \quad p(y_f \mid m_i^1, \Sigma_i^1) = \frac{\exp\{-\frac{(y_f - m_i^1)^2}{2\Sigma_i^1}\}}{\sqrt{2\pi\Sigma_i^1}}. \quad (4.13)$$

4.2 The Bioconcentration Factor Modelling Problem

We demonstrate, by re-visiting the bioconcentration factor modelling problem of section 2.2 of chapter 2, that the proposed fuzzy filtering based methodology could be used for incorporating intelligence in the modelling methods and thus achieves a robustness against uncertainties. It will be seen that the training algorithms of table 2.2 of chapter 2, if used to train the local models with penalized data (as suggested by our methodology), would result in an improvement in the generalization performance.

We employed a fuzzy filter, with membership functions defined by (3.4) for $\tilde{m} = 2$, for filtering out the uncertainties from the modelling problem. The number of rules in the fuzzy filter (i.e. K) and initial guess about cluster centres (θ_{-1}) were chosen via performing clustering on the 5-dimensional input training data using finite mixture models [40]. The identification algorithm was run till 100 epochs taking $\mu = \mu_\theta = 0.1$. The identified fuzzy filter was used to obtain for the training compounds the filtered values (4.1) and the underlying uncertainties (4.2).

The Gaussian mixture modelling of the 2-dimensional data (filtered and uncertainties values) identified 5 different component distributions describing the behavior of the data. These 5 component distributions have been displayed in Fig. 4.2. Associated to these components, the penalized data sets $D_1^p, \cdots, D_5^p$ (obtained using (4.12)) could be used to train the local models $M_1, \cdots, M_5$ respectively. The local models are finally combined using (4.13) to predict the overall output. Table 4.1 shows the system performance when the local models $M_1, \cdots, M_5$ are neural networks trained

Table 4.1: The performance of "trainscg" network training algorithm via proposed technique

s_p	R^2**-training**	**RMSE-training**	R^2**-testing**	**RMSE-testing**
0.01	0.7916	0.6355	0.6725	0.8542
0.03	0.7933	0.6309	0.6835	0.8405
0.05	0.7970	0.6230	0.6682	0.8598
0.1	0.8036	0.6084	0.6734	0.8433
0.2	0.8072	0.5962	0.6915	0.8188
0.4	0.8139	0.5793	0.7017	0.8009
0.5	0.8196	0.5681	0.7329	0.7485
0.75	0.8269	0.5541	0.6918	0.8022
1	0.8277	0.5516	0.7109	0.7790
2	0.8420	0.5253	0.7180	0.7627

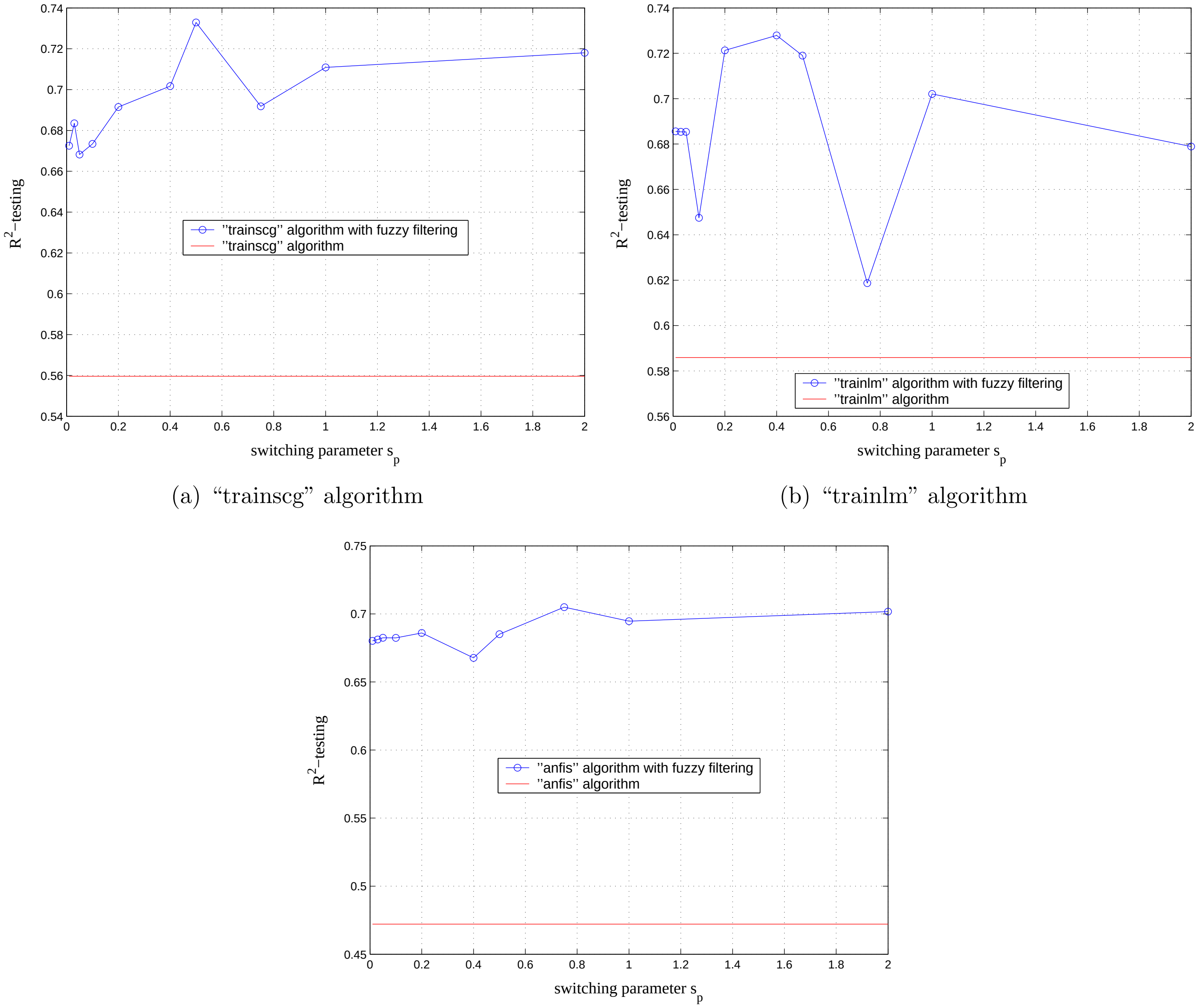
with the "trainscg" algorithm. Here, $M_1, \cdots, M_5$ have the same structure, initial conditions, and training parameters (e.g. number of epochs) as of the network trained with "trainscg" in table 2.2. The parameter ϵ in (4.12), to define the penalized data sets for different values of switching parameter s_p, was chosen as $\epsilon = 0$. In this text, we made no comment on the choice of switching parameter s_p, thus we consider the different values of switching parameter s_p ranging from 0.01 to 2.

Similarly, the tables 4.2 and 4.3 show the performance of the "trainlm" and "anfis" algorithms respectively via proposed fuzzy filtering based technique.

A comparison of tables 4.1, 4.2, and 4.3 with table 2.2, shown in Fig. 4.4, verifies that the generalization performance (i.e. testing data performance) of the modelling methods improved considerably via proposed approach. The type, structure, and training conditions of the local models in the studies are the same as of the models in table 2.2 of chapter 2. However, none of the modelling method resulted in the overtraining of the model via proposed fuzzy filtering based technique. This indicates that the robustness offered to the modelling methods is obviously a result of

1. penalizing the data,

2. combining the local models using a fuzzy rule base that has been carefully designed,



(a) “trainscg” algorithm

(b) “trainlm” algorithm

(c) “anfis” algorithm

Figure 4.4: An improvement in the generalization performance of the modelling methods via proposed approach

Table 4.2: The performance of "trainlm" network training algorithm via proposed technique

s_p	R^2-training	RMSE-training	R^2-testing	RMSE-testing
0.01	0.7905	0.6366	0.6856	0.8371
0.03	0.7928	0.6306	0.6854	0.8346
0.05	0.7935	0.6282	0.6854	0.8314
0.1	0.8032	0.6077	0.6475	0.8721
0.2	0.8129	0.5879	0.7213	0.7752
0.4	0.8158	0.5747	0.7279	0.7484
0.5	0.8202	0.5672	0.7190	0.7676
0.75	0.8333	0.5444	0.6187	0.8920
1	0.8405	0.5315	0.7021	0.7821
2	0.8457	0.5198	0.6789	0.8127

based on Gaussian mixture modelling of filtered data and uncertainties.

If the chosen training algorithm is robust towards uncertainties, then the local model could be trained with data sets $D_1, \cdots, D_C$ defined by (4.9). Since the training algorithm is robust, there is no need of penalizing the training data. In this case, an improvement in the modelling performance could be still expected as a result of the fuzzy combination of local models. As an illustration, we consider the Bayesian regularized neural networks that have been accepted as a robust methods of QSAR modelling [16–18,125]. The local models are trained with data sets $D_1, \cdots, D_5$ defined by (4.9) for $\epsilon = 0.01$ using Bayesian regularized neural network training algorithm (MATLAB Neural Network Toolbox command "trainbr"). Table 4.4 illustrates the performance of a Bayesian regularized neural network on BCF modelling problem and an improvement (although slightly) to this as a result of the fuzzy combination of local models.

4.3 Summary

Several modelling methods have been proposed in the literature aiming at the good generalization performance of the models. This work, unlike many studies, doesn't

Table 4.3: The performance of "anfis" training algorithm via proposed technique

s_p	R^2-training	RMSE-training	R^2-testing	RMSE-testing
0.01	0.7966	0.6293	0.6802	0.8421
0.03	0.7976	0.6248	0.6812	0.8383
0.05	0.7987	0.6208	0.6825	0.8343
0.1	0.7958	0.6197	0.6824	0.8257
0.2	0.8068	0.5961	0.6860	0.8152
0.4	0.8115	0.5830	0.6677	0.8352
0.5	0.8130	0.5789	0.6851	0.8067
0.75	0.8155	0.5715	0.7050	0.7805
1	0.8196	0.5636	0.6947	0.7938
2	0.8294	0.5459	0.7017	0.7841

Table 4.4: The performance of Bayesian regularized neural network training algorithm

method	R^2-training	RMSE-training	R^2-testing	RMSE-testing
"trainbr"	0.8731	0.4666	0.7112	0.7604
"trainbr" via proposed method	0.8787	0.4579	0.7466	0.7129

propose a new modelling method but provides a tool for rendering robustness in any modelling method. A case study dealing with the bioconcentration factor modelling of chemicals was provided to illustrate the effectiveness of our technique.

The uncertainties, affecting adversely the generalization capabilities of the modelling methods, are filtered using a fuzzy filter. Based on the available information about uncertainties, the local models are developed in a manner that uncertainties are not allowed to affect the training of the local models. This improves the generalization performance of a modelling technique. The combination of the local models using a fuzzy rule base (that has been carefully designed based on Gaussian mixture modelling of filtered data and uncertainties) provides additional tolerance towards uncertainties.

The aim of this study is to provide to the researchers a piece of software that would improve the robustness performance of their favourite modelling methods. One could observe in Fig. 4.4 a considerable improvement in the performance of the different modelling methods via proposed technique. However, there are some issues which remain to be addressed in our future work. The automatic selection of the value of switching parameter s_p is a part of our future work. Fortunately, the effectiveness of our approach has been observed at all considered values of s_p ranging from 0.01 to 2. For a choice $s_p = 0$ (i.e. training of local models with filtered data), the technique becomes close to the method of previous chapter.

Chapter 5

A Study on Ionic Liquids

This chapter presents a study on the ionic liquids (which has gained importance in the field of green chemistry), their environmental behavior, and a computer model for their toxicity prediction.

5.1 Ionic Liquids

The out comes of the constant efforts and hardwork of scientists revealed the new class of chemicals which are not only of great industrial importance but are also environmental friendly. Such chemicals are termed as "Green Chemicals". One of such class of green chemicals is of ionic liquids which are considered to possess environmental benign properties. The ionic liquids are known as green solvents due to the fact that they exert immeasurable low vapour pressure at standard condition and do not contribute to air pollution at all. The ionic liquids have enormous potential for wide industrial application and are considered as potential replacement for environmentally harmful, volatile organic solvents. The ranges of applications of ionic liquids are wider than fluorous solvents and supercritical carbon dioxide. There is a steady increase in the number of ionic liquids related publications in recent years. The ionic liquids have drawn considerable attention as an alternative to conventional organic solvents in a variety of significant synthetic, catalytic, electrochemical applications, separation and extraction process, biotransformation etc [9, 15, 25, 26, 35, 65, 66, 86, 93, 95, 96, 104, 106, 110, 120, 123, 124].

Molten organic salts like imidazolium and the quaternary ammonium salts exhibits room temperature liquid-like behavior and interesting solvent properties for both chemical reaction and extraction. The first room temperature molten salt systems were reported in 1951 by Hurly and Wier [52]. The liquids that are comprised entirely of ions could be called as ionic liquids, in this regard they are anhydrous aprotic solvents [63]. Ionic liquids are salts with a melting point below 100°C. One of the salient features of the ionic liquids is strong ion-ion interaction that are not often seen in the higher temperature molten salts.

The typical ionic liquids have an organic cation and an inorganic, polyatomic anion. The general chemical composition of ionic liquid is consistent despite the chemical and physical properties and specific composition vary tremendously. The potential number of ionic liquids is large due to many known potential cation and anions. Some examples of cations and anions commonly used for the formation of the ionic liquids are shown in Fig. 5.1. The anion (X^-) can be any of a variety of species including nitrate [NO_3^-], acetate [CH_3COO^-], terafluoroborate [BF_4^-] etc.

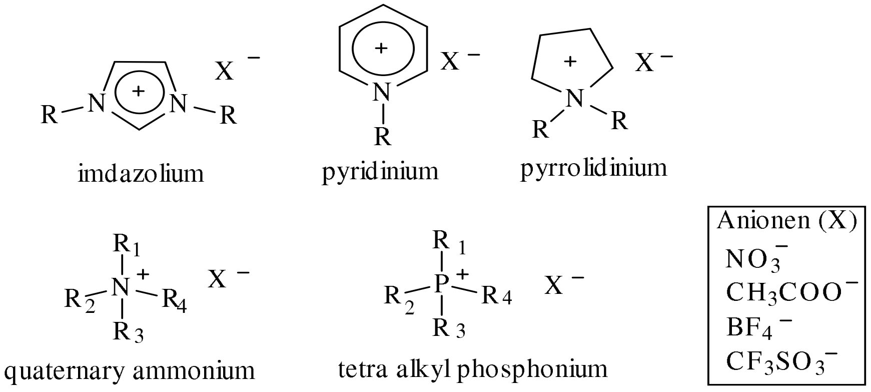


Figure 5.1: Some examples of cations and anions of ionic liquids

5.2 Environmental Behavior of Ionic Liquids

The ionic liquids are labeled as green solvent as they exert negligible vapour pressure and prevent air pollution but they are at least to some degree water soluble and may escape in water bodies which may cause water pollution. To facilitate their wide industrial acceptance in accordance to the new European chemical policy REACH, it is important to investigate their impact on the health and environment. Since the ionic liquids are known as “designer solvents” [107], therefore the study of their environmental behavior will help in designing eco-friendly ionic liquids.

Some of the ionic liquids show strong antimicrobial activity. The C-1 alkyl chains substituents in imidazolium, pyridinium and quaternary ammonium based cation, plays an important role in influencing the toxicity of the ionic liquid. The ionic liquids with a longer alkyl chain are more toxic to the microbes [100]. Similar observation was made in the detailed biological studies of dialkylimidazolium ionic liquids in luminescent bacteria as well as in the the IPC-81 (leukemia cells) and C6 (glioma cells) rat cell lines [34,103]. The toxicity of ionic liquids was found to be lower than the conventional solvents such as acetone, acetonitrile, methanol and methyl *tert*-butyl ether. The acetylcholinesterase can be inhibited by ionic liquids containing a cation with a positively charged nitrogen and with a certain lipophilicity [108]. The effects of ionic liquids on the Daphnia magna, algae (*Secnedesmus spp*), fresh water snail (*Physa acuta*), plant (*Lemna minor*), nematodes(*Caenorhabditis elegans*), fish (*Danio rerio*) have been also investigated [11, 12, 102, 111, 122]. The influence of the anion moiety and side chain of ionic liquids have been also investigated in eco-toxicological test battery. The side chain length effect was distinct and consistent than the anion effect [87, 109]. The toxicity of the ionic liquids may also have a negative impact on their biodegradation. The authors in [56] have reported a theoretical environmental risk analysis on a set of dialkylimidazolium ionic liquids and illustrated the theoretical metabolism scheme for [BMIM] cation. The suggested breakdown products can be identified by a combination of strong ion-exchange SPE and GC-MS [78].

The biodegradation of the ethyl ester and amide imidazolium ionic liquids using standard OECD sturm and closed bottle test has been examined in [42]. The ester imidazolium ionic liquids were found out to be more biodegradable than amide imidazolium ionic liquids. The tested imidazolium based ionic liquids were turned out to be “not readily biodegradable”. However the introduction of the group that is susceptible to enzymatic hydrolysis improves the extent of biodegradation [41]. It was found that octylsulfate anion conferred higher levels of biodegradability [43]. The introduction of an ester group in the side chain of the 1,3-dialkylimidazolium cation improves the biodegradation to a large extent [43]. This ionic liquid has been obtained by combining cation 3-methyl-1-(propyloxycarbonyl) imidazolium and anion octylsulphate. It shows a biodegradation of 49% in the closed bottle test.

We have performed a closed bottle test [1] on a set of five commonly used ionic liquids (see Appendix C.3.1). The investigated ionic liquids were [BMIM][BF_4], [BMPy][BF_4], [EMIM][OTos], [EMIM][$EtSO_4$], and ECOENG2122P. The reference substance taken was Sodium *n*-dodecyl sulphate (SDS). The results have been summarized in Fig. 5.2. As seen from Fig. 5.2, none of the tested ionic liquids could be classified as readily biodegradable ionic liquid. According to OECD standards a compound that achieves a biodegradation level higher than 60% (in 28 days) is referred to as readily biodegradable compound.

A recent study [33] has also indicated that none of the considered imidazolium based ionic liquids could be classified as readily biodegradable. This suggest that imidazolium ring does not get mineralized easily, however, imidazolium ring with longer alkyl chain such as *hexyl* and *octyl* substituents are partially mineralized. In contrary the pyridinum based ionic liquids with *hexyl* and *octyl* substituents are fully mineralized. This implies that biodegradation rate increases with an increase in the length of alkyl chain [33].

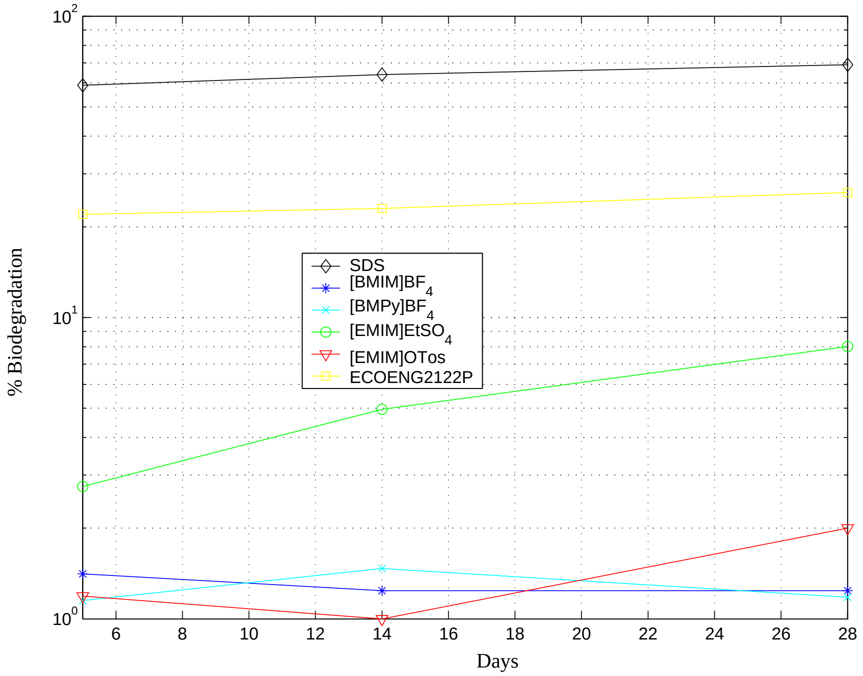


Figure 5.2: Biodegradation curves of studied ionic liquids

5.3 A Computer Model for Predicting the Toxicity of Ionic Liquids

The eco-toxicological assays are expensive and time consuming. Further, there are theoretically over 10^6 ionic liquids which cann't be tested for their toxicity. Thus, it is necessary to develop QSAR models for a fast and accurate prediction of toxicity. The QSAR method is based on the assumption that the toxicity of a chemical compound is determined by its molecular structure and the structure is represented using molecular descriptors.

In the earlier chapters, we calculated several constitutional, 2D autocorrelations, Burden eigenvalue, geometrical, 3D-MoRSE, WHIM, GETAWAY, and molecular properties based descriptors using E-DRAGON [114]. However, in this study we consider "group contributing molecular descriptors" [81] of ionic liquids. According to this approach, the toxicity of a ionic liquid is assumed to depend upon the contribution

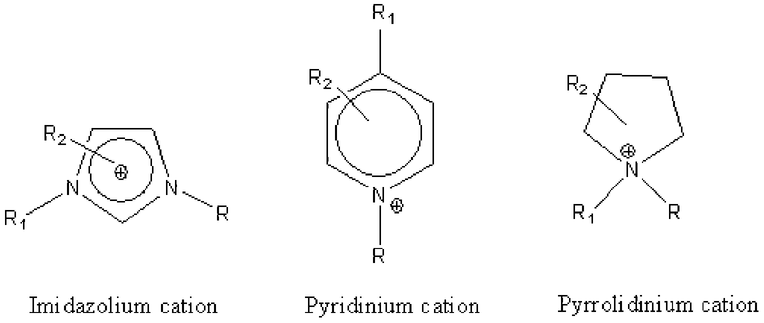


Figure 5.3: Structures of cations of ionic liquids

of anions, cations, and alkyl-chain substitutions [81]. The typical ionic liquid cations are imidazolium, pyridinium, and pyrrolidium (see Fig. 5.3). Here R is the long n-alkane chain, R_1 is short chain substitution, and R_2 is an additional substitution on cation apart from R and R_1 as shown in Fig. 5.3. The anions examples are BF_4^-, PF_6^-, Cl^-, Br^-, $N(CN_2)_2^-$, $CH_3SO_4^-$, etc. The different anions have been divided into three groups: the A_1 group includes BF_4^-, Cl^-, tosylate, diethylphosphate; the A_2 group includes PF_6^-, Br^-, $N(CN_2)_2^-$, $CH_3SO_4^-$, $C_2H_5SO_4^-$; and A_3 group includes octylsulphate, $(CF_3SO_2)_2N^-$, $[(O\text{-}OPhO)_2B]^-$.

The structure of a ionic liquid, in our analysis, will be represented by a set of 9 descriptors $(a_1, a_2, a_3, c_1, c_2, c_3, r, r_1, r_2)$ defined in table 5.1. Table 5.2 shows some

Table 5.1: Definitions of group contributing descriptors (similar to [81])

descriptor	value
a_1	equal to 1, if A_1 group is present in the molecule, otherwise, equal to 0
a_2	equal to 1, if A_2 group is present in the molecule, otherwise, equal to 0
a_3	equal to 1, if A_3 group is present in the molecule, otherwise, equal to 0
c_1	equal to 1, if imidazolium cation is present in the molecule, otherwise, equal to 0
c_2	equal to 1, if pyridinium cation is present in the molecule, otherwise, equal to 0
c_3	equal to 1, if pyrrolidium cation is present in the molecule, otherwise, equal to 0
r	equal to number of carbons in long chain R
r_1	is equal to number of carbons in chain R_1
r_2	is equal to number of carbons in chain R_2

of the ionic liquid compounds, their group contributing molecular descriptors, and toxicity (*Vibrio fischeri* EC_{50}) data either collected from literature or generated experimentally by performing bioluminescence inhibition assay (see Appendix C.3.2).

Our aim is to develop a model that takes the values of these 9 descriptors as inputs and outputs the $\log EC_{50}$ value. The data set has been divided into training and testing sets, as in sections 2.1.3 and 2.2.1 of chapter 2, such that testing compounds were sandwiched between the training ones in the sense of Euclidean distance. This was done by including every second compound in the series of ascending order arranged compounds in the testing set and the remaining compounds in the training set. Thus each set contains about half of the total compounds.

Bayesian regularized neural networks are considered as a robust method of QSAR modelling in literature [16–18, 125]. This fact has been observed in our studies too in previous chapters. Thus for the modelling of ionic liquid toxicity data, we train, using MATLAB Neural Network Toolbox software, a 3-layer feed-forward Bayesian regularized network that has 2 "tansig" (i.e. with hyperbolic tangent sigmoid transfer function) neurons in first layer, 1 "tansig" neuron in second layer, and one "purelin" (i.e. with linear transfer function) neuron in third layer. The performance (i.e. coefficient of determination R^2 and root mean squared error RMSE) of the trained network on training and testing data is listed in the first row of table 5.3 and in Fig. 5.4.

Although the performance of Bayesian regularized neural model on testing data is acceptable, but a large difference exists between the training performance and testing performance. Therefore, uncertainties exist in the formulated modelling problem. Chapter 4 has outlined a methodology, summarized in Fig. 4.1, for incorporating an intelligence in a given modelling technique to achieve a robustness of the modelling performance against uncertainties. Now, we try to improve the performance of the Bayesian regularized neural model based on the approach of Fig. 4.1. However, descriptors data in this case is binary (i.e. either 0 or 1), see table 5.2, and thus we don't want to define fuzzy membership functions for the binary data (although theoretically it is possible to do so). In this particular case (when a fuzzy filter is not preferred), we replace the fuzzy filter in Fig. 4.1 by the Bayesian regularized trained

Table 5.2: Ionic liquids and their descriptors

compound	a_1	a_2	a_3	c_1	c_2	c_3	r	r_1	r_2	$\log EC_{50}(\mu mol/L)$	Ref.
[MIM]	0	0	0	1	0	0	0	1	0	4.17	[34]
[C$_1$MIM][CH$_3$SO$_4$]	0	1	0	1	0	0	1	1	0	4.76	[81]
[C$_2$MIM][C$_2$H$_5$SO$_4$]	0	1	0	1	0	0	2	1	0	4.02	[81]
[C$_2$MIM][Cl]	1	0	0	1	0	0	2	1	0	4.55	[81]
[C$_3$MIM][BF$_4$]	1	0	0	1	0	0	3	1	0	3.94	[103]
[C$_4$MIM][PF$_6$]	0	1	0	1	0	0	4	1	0	3.07	[41]
[C$_4$MIM][BF$_4$]	1	0	0	1	0	0	4	1	0	3.55	[103]
[C$_4$MIM][BF$_4$]	1	0	0	1	0	0	4	1	0	3.10	[41]
[C$_4$MIM][Br]	0	1	0	1	0	0	4	1	0	3.07	[103]
[C$_4$MIM][Br]	0	1	0	1	0	0	4	1	0	4.01	[34]
[C$_4$MIM][Br]	0	1	0	1	0	0	4	1	0	3.27	[41]
[C$_4$MIM][Cl]	1	0	0	1	0	0	4	1	0	3.71	[34]
[C$_4$MIM][Cl]	1	0	0	1	0	0	4	1	0	3.34	[41]
[C$_4$MIM][N(CN$_2$)$_2$]	0	1	0	1	0	0	4	1	0	3.67	[41]
[C$_4$EIM][BF$_4$]	1	0	0	1	0	0	4	2	0	2.80	[103]
[C$_5$MIM][BF$_4$]	1	0	0	1	0	0	5	1	0	3.14	[103]
[C$_6$MIM][Br]	0	1	0	1	0	0	6	1	0	1.42	[34]
[C$_6$MIM][Cl]	1	0	0	1	0	0	6	1	0	1.94	[81]
[C$_6$MIM][Cl]	1	0	0	1	0	0	6	1	0	2.32	[41]
[C$_6$MMIM][Cl]	1	0	0	1	0	0	6	1	1	1.74	[81]
[C$_6$MIM][PF$_6$]	0	1	0	1	0	0	6	1	0	2.17	[41]
[C$_6$MIM][BF$_4$]	1	0	0	1	0	0	6	1	0	3.18	[103]
[C$_6$EIM][BF$_4$]	1	0	0	1	0	0	6	2	0	2.15	[103]
[C$_7$MIM][BF$_4$]	1	0	0	1	0	0	7	1	0	2.44	[103]
[C$_8$MIM][Br]	0	1	0	1	0	0	8	1	0	0.63	[34]
[C$_8$MIM][Cl]	1	0	0	1	0	0	8	1	0	1.19	[41]
[C$_8$MIM][PF$_6$]	0	1	0	1	0	0	8	1	0	0.95	[41]
[C$_8$MIM][BF$_4$]	1	0	0	1	0	0	8	1	0	1.41	[103]
[C$_9$MIM][BF$_4$]	1	0	0	1	0	0	9	1	0	0.72	[103]
[C$_{10}$MIM][Cl]	1	0	0	1	0	0	10	1	0	0.50	[103]
[C$_{10}$MIM][BF$_4$]	1	0	0	1	0	0	10	1	0	-0.18	[103]
[MPy]	0	0	0	0	1	0	1	0	0	3.06	[34]
[C$_4$Py][Br]	0	1	0	0	1	0	4	0	0	3.39	[34]
[C$_4$MPy][Br]	0	1	0	0	1	0	4	1	0	2.75	[34]
[C$_4$MMPy][Br]	0	1	0	0	1	0	4	1	1	2.69	[34]
[C$_4$Py][Cl]	1	0	0	0	1	0	4	0	0	3.40	[34]
[C$_4$Py][N(CN$_2$)$_2$]	0	1	0	0	1	0	4	0	0	3.30	[34]
[C$_4$MPy][N(CN$_2$)$_2$]	0	1	0	0	1	0	4	1	0	2.65	[34]
[C$_4$MMPy][N(CN$_2$)$_2$]	0	1	0	0	1	0	4	1	1	2.38	[34]
[C$_6$MPy][Br]	0	1	0	0	1	0	6	1	0	2.06	[34]
[C$_6$MPy][Cl]	1	0	0	0	1	0	6	1	0	1.44	[81]
[C$_8$MPy][Br]	0	1	0	0	1	0	8	1	0	0.79	[34]
[C$_6$MPyRR][Cl]	1	0	0	0	0	1	6	1	0	2.99	[81]
[C$_4$MIM][BF$_4$]	1	0	0	1	0	0	4	1	0	3.60	experiment
[C$_2$MIM][C$_2$H$_5$SO$_4$]	0	1	0	1	0	0	2	1	0	4.05	experiment
[C$_2$MIM][(C$_2$H$_5$)$_2$PO$_4$]	1	0	0	1	0	0	2	1	0	4.63	experiment
[C$_2$MIM][C$_7$H$_7$SO$_3$]	1	0	0	1	0	0	2	1	0	4.59	experiment
[C$_2$MIM][(2-OPhO)B]	0	0	1	1	0	0	2	1	0	2.96	[87]
[C$_4$MIM][(CF$_3$)$_2$N]	0	1	0	1	0	0	4	1	0	3.46	[87]
[C$_4$MIM][(CF$_3$SO$_2$)$_2$N]	0	0	1	1	0	0	4	1	0	2.47	[87]
[C$_4$MIM][octylOSO$_3$]	0	0	1	1	0	0	4	1	0	1.82	[87]
[C$_4$MIM][Cl]	1	0	0	1	0	0	4	1	0	3.47	[87]
[C$_1$OIM][BF$_4$]	1	0	0	1	0	0	1	8	0	1.40	[87]

Table 5.3: Ionic liquids toxicity modelling performance

method	R^2**-training**	**RMSE-training**	R^2**-testing**	**RMSE-testing**
Bayesian regularized	0.9837	0.1440	0.7395	0.6509
Bayesian regularized with intelligence	0.9796	0.1644	0.8047	0.5374

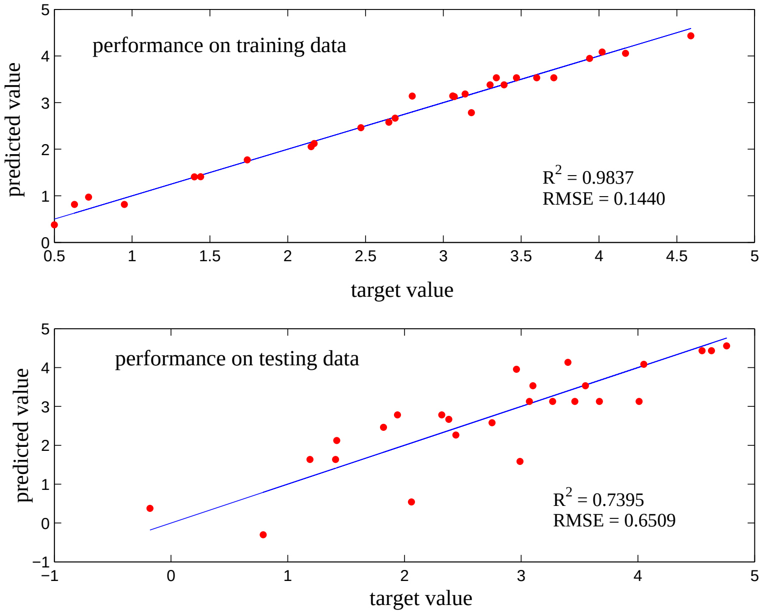


Figure 5.4: Plots for the Bayesian regularized neural modelling of ionic liquid toxicity

neural model.

Gaussian mixture modelling of data, by using the algorithm of [40], produced three different clusters shown in Fig 5.5, i.e., 3 local models M_1, M_2, M_3 will be trained. As stated earlier in section 4.2 of chapter 4, for Bayesian regularized modelling technique (being robust), the local models in Fig. 4.1 could be trained with data sets $D_1, \cdots, D_3$ defined by (4.9) without any penalization. The parameter ϵ in (4.9) was taken in such a way that all data lying at $\pm 2\sqrt{\Sigma_i^1}$ from mean m_i^1 are included in i^{th} data set D_i which was then used for the training of i^{th} local model M_i with Bayesian regularized training algorithm. The overall output is computed by combining the local models using a fuzzy rule base constructed from the Gaussian mixture model as explained in the previous chapter. The performance of intelligence incorporated

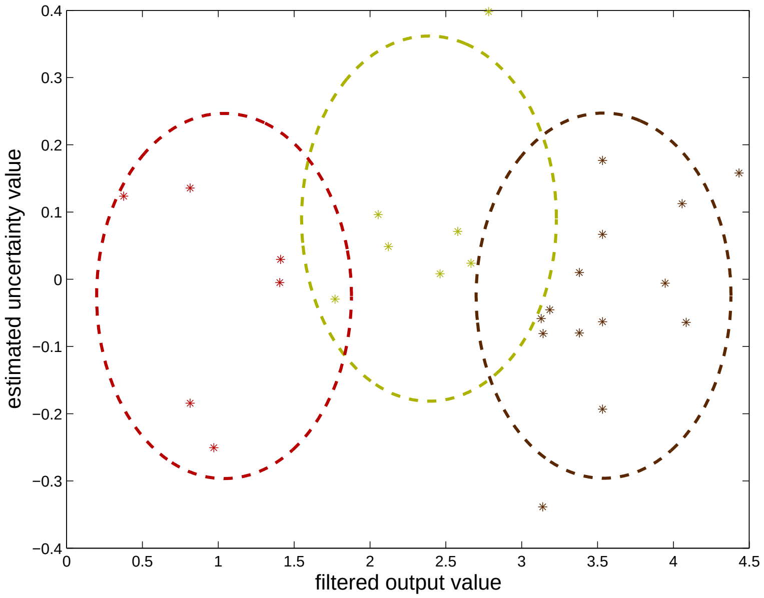


Figure 5.5: Data points and level curves for the different Gaussian components

Bayesian regularized neural modelling is listed in the second row of table 5.3 and also in Fig. 5.6. We observe an improvement in the performance as a result of incorporating intelligence. Finally, the experimental and model predicted $\log EC_{50}(\mu mol/L)$ values for the testing compounds are listed in table 5.4. Table 5.4 also lists the absolute difference between the two.

5.4 Summary

This chapter has presented a computer model to predict the *Vibrio fischeri* toxicity of ionic liquids. We achieved a prediction accuracy of $R^2 = 0.9796$ on training compounds and of $R^2 = 0.8047$ on testing compounds. We observe clearly the presence of a outlier in the testing compounds. This compound, marked in table 5.4, has the maximum prediction error (equal to 1.5367) among the testing compounds. The reason for behaving this compound (i.e. [C_6MPyRR][Cl]) as outlier is that it is the only compound with pyrrolidium cation (i.e. descriptor c_3 in table 5.2 takes value equal to 1 only for this compound). In simple words, our training data didn't include

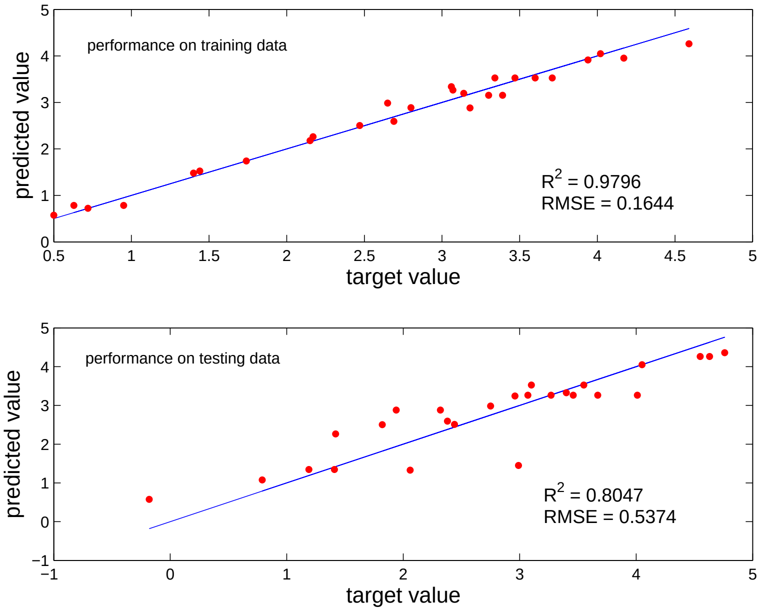


Figure 5.6: Plots for the Bayesian regularized neural modelling (with intelligence) of ionic liquid toxicity

such types of compound. Thus, for a fair assessment of prediction performance of our model, this compound should be excluded from the testing set. After excluding the outlier, the prediction accuracy of the model on remaining 25 testing compounds increased to $R^2 = 0.8838$. The testing compounds (i.e. unseen compounds not used in the training of models) were in number nearly equal to the training compounds. These results (i.e. a prediction accuracy of $R^2 = 0.8838$ on testing compounds) are encouraging and verify the effectiveness of our approach when it comes to the generalization capability of the model.

Table 5.4: Prediction of toxicity of testing ionic liquids

compound	**experimental**	**predicted**	**absolute difference**
	$\log EC_{50}(\mu mol/L)$	$\log EC_{50}(\mu mol/L)$	
[C$_2$MIM][(2-OPhO)B]	2.96	3.2424	0.2824
[C$_4$MIM][octylOSO$_3$]	1.82	2.5021	0.6821
[C$_2$MIM][C$_2$H$_5$SO$_4$]	4.05	4.0489	0.0011
[C$_4$MMPy][N(CN$_2$)$_2$]	2.38	2.5937	0.2137
[C$_4$MPy][Br]	2.75	2.9845	0.2345
[C$_1$MIM][CH$_3$SO$_4$]	4.76	4.3588	0.4012
[C$_2$MIM][Cl]	4.55	4.2614	0.2886
[C$_4$MIM][Br]	3.07	3.2644	0.1944
[C$_4$MIM][BF$_4$]	3.10	3.5269	0.4269
[C$_2$MIM][(C$_2$H$_5$)$_2$PO$_4$]	4.63	4.2614	0.3686
[C$_4$Py][Cl]	3.40	3.3253	0.0747
[C$_4$MIM][Br]	3.27	3.2644	0.0056
[C$_4$MIM][(CF$_3$)$_2$N]	3.46	3.2644	0.1956
[C$_4$MIM][BF$_4$]	3.55	3.5269	0.0231
[C$_4$MIM][N(CN$_2$)$_2$]	3.67	3.2644	0.4056
[C$_4$MIM][Br]	4.01	3.2644	0.7456
[C$_6$MIM][Br]	1.42	2.2626	0.8426
[C$_6$MIM][Cl]	1.94	2.8812	0.9412
[C$_6$MPy][Br]	2.06	1.3287	0.7313
[C$_6$MIM][Cl]	2.32	2.8812	0.5612
[C$_6$MPyRR][Cl]	2.99	1.4533	1.5367 (outlier)
[C$_7$MIM][BF$_4$]	2.44	2.5112	0.0712
[C$_8$MPy][Br]	0.79	1.0773	0.2873
[C$_8$MIM][Cl]	1.19	1.3471	0.1571
[C$_8$MIM][BF$_4$]	1.41	1.3471	0.0629
[C$_1$OMIM][BF$_4$]	-0.18	0.5731	0.7531

Chapter 6

Concluding Remarks

When it comes to the data-driven modelling of environmental behavior of chemicals, neural/fuzzy techniques have a lot to offer. These techniques are potentially considered suitable for dealing with the complex and ill-defined problems where classical modelling approaches either fail or don't perform up to the acceptable level. Thus we tried to apply neuro/fuzzy techniques in chemicals's environmental behavior modelling problems. However, the success of such modelling techniques depends upon number of factors including choice of descriptors, choice of model structure, and the robustness of training algorithm. If the uncertainties regarding the made choices is high (i.e. the made choices are far away from the optimal ones), then a non-robust model construction algorithm would typically overtrain the model resulting into a low generalization performance. However, generalization is the key concern of such studies.

The low generalization capability of the model is probably the most commonly faced problem by the QSAR research community. In second chapter, we have highlighted this issue in the modelling of fathead minnow toxicity and bioconcentration factor data of chemicals. Our aim in this thesis was to remove the generalization-capability related bottleneck of the neural/fuzzy techniques in modelling of chemicals' environmental behavior. In third and fourth chapter of the thesis, we have introduced a methodology to remove this bottleneck. Many examples have been provided to demonstrate that generalization performance of different neuro/fuzzy modelling techniques improved using the proposed methodology of incorporating intelligence.

Bayesian regularized neural networks due to their robustness properties perform, in general, better than any other neuro/fuzzy technique for modelling in presence of uncertainties. In chapter 4 and 5, we demonstrate that performance of Bayesian regularized neural networks can be further improved using the proposed methodology. We feel that this work, due to its basic nature, may be useful in several real-world modelling problems. Chapter 5 outlines an application of the work in the emerging field of green chemistry. An interesting feature of our work is that we don't ask someone to replace his favourite modelling technique by our technique, however, our methodology improves the performance of a given modelling technique.

In this thesis, our concern was to address the problem:

> *How to improve the generalization performance of chemicals' environmental behavior predicting models for a given choice of descriptors, model type, and model structure*?

Our future research work is concerned with the development of algorithms for an automatic selection of descriptors and model structure.

Bibliography

[1] ISO guideline 10707, Water quality, Closed Bottle Test, 1994.

[2] EN-ISO 11348-1 guideline, Determination of inhibitory effect of water samples on the light emission of vibrio fischeri. Luminescent bacteria test, 1998.

[3] A brief introduction to the European Commissions regulatory proposal on Registration, Authorisation and Evaluation of Chemicals (REACH), April 2006.

[4] REGULATION (EC) No 1907/2006 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL. Official Journal of the European Union, OJ L396, December 2006.

[5] REACH in brief, February 2007.

[6] R. Babuška. *Fuzzy Modeling for Control*. Kluwer Academic Publishers, Boston, 1998.

[7] R. Babuška and H.B. Verbruggen. *Fuzzy Model Identification: Selected Approaches*. Springer, Berlin, Germany, 1997.

[8] S. Banerjee and G. L. Baughman. Bioconcentration Factors and Lipid Solubility. *Environmental Science Technology*, 25:536–539, 1991.

[9] E. D. Bates, R. D. Mayton, I. Ntai, and J. H. Davis. CO_2 capture by a task-specific ionic liquid. *J. Am. Chem. Soc.*, 124:926, 2002.

[10] E. Benfenati and G. Gini. Computational predictive programs (expert systems) in toxicology. *Toxicology*, 119:213–225, 1997.

[11] R. J. Bernot, M. A. Brueseke, M. A. Evans-White, and G. A. Lamberti. Acute and chronic toxicity of imidazoluim - based ionic liquids on Daphnia magna. *Environmental Toxicology and Chemistry*, 24(1):87–92, 2005.

[12] R. J. Bernot, E. E. Kennedy, and G. A. Lamberti. Effects of Ionic liquids on the survival, movement, and feeding behaviour of the freshwater snail, Physca acuta. *Environmental Toxicology and Chemistry*, 24(7):1759–1765, 2005.

[13] J. C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms.* New York:Plenum, 1981.

[14] S. Bintein, J. Devillers, and W. Karcher. Nonlinear dependence of fish bioconcentration on n-Octanol/water partition coefficients. *SAR and QSAR in Environmental Research*, 1:29–39, 1993.

[15] J. A. Boon et al. Friedel Crafts reactions in ambient-temperature molten-salts. *J. Organic Chemistry*, 51:480–483, 1986.

[16] Frank R. Burden, Martyn G. Ford, David C. Whitley, and David A. Winkler. Use of Automatic Relevance Determination in QSAR Studies Using Bayesian Neural Networks. *J. Chem. Inf. Comput. Sci.*, 40:1423–1430, 2000.

[17] Frank R. Burden and David A. Winkler. New QSAR Methods Applied to Structure-Activity Mapping and Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci.*, 39:236–242, 1999.

[18] Frank R. Burden and David A. Winkler. Robust QSAR Models Using Bayesian Regularized Neural Networks. *J. Med. Chem.*, 42:3183–3187, 1999.

[19] M. Burger, H.W. Engl, J.Haslinger, and U.Bodenhofer. Regularized data-driven construction of fuzzy controllers. *J. Inverse and Ill-posed Problems*, 10:319–344, 2002.

[20] D. S. Chen and R. C. Jain. A robust back propagation learning algorithm for function approximation. *IEEE Trans. Neural Networks*, 5:467–479, May 1994.

[21] I. Cohen. Automatic facial expression recognition from video sequences using temporal information. Masters thesis, Univ. of Illinois at Urbana Champaign, 2000.

[22] D. W. Connell and D. W. Hawker. Use of polynomial expressions to describe the bioconcentration of hydrophobic chemicals by fish. *Ecotoxicology and Environmental Safety*, 16:242–257, 1988.

[23] V. Consonni, R. Todeschini, and M. Pavan. Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 1. Theory of the novel 3D molecular descriptors. *J. Chem. Inf. Comp. Sci.*, 42:682–692, 2002.

[24] V. Consonni, R. Todeschini, M. Pavan, and P. Gramatica. Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 2. Application of the novel 3D molecular descriptors to QSAR/QSPR studies. *J. Chem. Inf. Comp. Sci.*, 42:693–705, 2002.

[25] S. G. Cull et al. Room-temperature ionic liquids as replacements for organic solvents in multiphase bioprocess operations. *Biotechnol. Bioeng.*, 69(2), Jul 2000.

[26] S. Dai, Y. H. Ju, and C. E. Barnes. Solvent extraction of strontium nitrate by a crown ether using room-temperature ionic liquids. *J. Chem. Soc., Dalton Trans.*, 8:1201–1202, 1999.

[27] R. N. Davé. Characterization and detection of noise in clustering. *Pattern Recognition Lett.*, 12(11):657–664, 1991.

[28] R. P. Davies and A. Dobbs. The prediction of bioconcentration in fish. *Water Research*, 18:1253–1262, 1984.

[29] J. C. Dearden. QSAR modelling of bioaccumulation. In M. T. D. Cronin and D. J. Livingstone, editors, *Predicting Chemical Toxicity and Fate*. CRC Press LLC, Boca Raton, Florida, 2004.

[30] J. C. Dearden and N. M. Shinnawei. Improved prediction of fish bioconcentration factor of hydrophobic chemicals. *SAR and QSAR in Environmental Research*, 15:449–455, 2004.

[31] S. Dimitrov, N. Dimitrova, T. Parkerton, M. Comber, M. Bonnell, and O. Mekenyan. Base-line model for identifying the bioaccumulation potential of chemicals. *SAR and QSAR in Environmental Research*, 16(6):531–554, 2005.

[32] S. D. Dimitrov, O. G. Mekenyan, and J. D. Walker. Non-linear modeling of bioconcentration using partition coefficients for narcotic chemicals. *SAR and QSAR in Environmental Research*, 13:177–188, 2002.

[33] K. M Docherty, J. K. Dixon, and C. F. Kulpa. Biodegradability of imidazolium and pyridinium ionic liquids by an activated sludge microbial community. *Biodegradation.*, 18(4):481–493, 2007.

[34] K. M. Docherty and C. F. Kulpa. Toxicity and antimicrobial activity of imidazolium and pyridinium ionic liquids. *Green Chem.*, 7:185–189, 2005.

[35] M. Eckstein, P. Wasserscheid, and U. Kragl. Enhanced enantioselectivity of lipase from Pseudomona sp. at high temperatures and fixed water activity in the ionic liquid 1-butyl-3-methyl bis((trifluoromethyl)sulfonyl)amide. *Biotechnology Letters*, 24:763–767, 2002.

[36] D. V. Eldred and P. C. Jurs. Prediction of Acute Mammalian Toxicity of Organophosphorus Pesticide Compounds from Molecular Structure. *SAR and QSAR in environmental research*, 10:75–99, 1999.

[37] P. Ertl, B. Rohde, and P. Selzer. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *J. Med. Chem.*, 43:3714–3717, 2000.

[38] European Centre for Ecotoxicology and Toxicology of Chemicals. The Role of Bioaccumulation in Environmental Risk Assessment: The Aquatic Environment and Related Food Webs. Technical report 67, Brussel, Belgium, 1995.

[39] F. Lemke and E. Benfenati and J. A. Müller. Data-drivern Modeling and Prediction of Acute Toxicity of pesticide residue. *SIGKDD Newsletter*, 8(1):71–79, 2006.

[40] M. A. T. Figueiredo and A. K. Jain. Unsupervised Learning of Finite Mixture Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):381–396, 2002.

[41] M. T. Garcia, Gathergood N., and P. J. Scammells. Biodegradable ionic liquids: Part II. Effect of the anion and toxicology. *Green Chem.*, 7:9–14, 2005.

[42] N. Gathergood, M. T. Garcia, and P. J. Scammells. Biodegradable ionic liquids: Part I. Concept, preliminary targets and evaluation. *Green Chem.*, 6:166–175, 2004.

[43] N. Gathergood, M. T. Garcia, and P. J. Scammells. Biodegradable ionic liquids: Part III. The first readily biodegradable ionic liquids. *Green Chem.*, 8:156–160, 2006.

[44] G. Gini. Predictive Toxicology of Chemicals: Experience and Impact of AI tools. *AI Magazine*, 21:81–84, 2000.

[45] G. Gini, M. V. Cracium, and C. König. Combining Unsupervised and Supervised Artificial Neural Networks to Predict Aquatic Toxicity. *J. Chem. Inf. Comput. Sci.*, 44:1897–1902, 2004.

[46] G. Gini, M. Lorenzini, E. Benfenati, P. Grasso, and M. Bruschi. Predictive Carcinogenicity: A Model for Aromatic Compounds, with Nitrogen-Containing Substituents, Based on Molecular Descriptors Using an Artificial Neural Network. *J. Chem. Inf. Comput. Sci.*, 39:1076–1080, 1999.

[47] F. A. P. C. Gobas and X. Zhang. Measuring bioconcentration factors and rate constants of chemicals in aquatic organism under condition of variable water concentration and short exposure time. *Chemosphere*, 25:1961–1971, 1995.

[48] P. Gramatica and E. Papa. QSAR modelling of bioconcentration factor by theoretical molecular descriptors. *QSAR & Combinatorial Science*, 22:374–385, 2003.

[49] P. Gramatica and E. Papa. An Update of the BCF QSAR Model Based on Theoretical Molecular Descriptors. *QSAR & Combinatorial Science*, 24:953–960, 2005.

[50] C. Hansch and T. Fujita. $\rho - \sigma - \pi$ Analysis. A method for the correlation of biological activity and chemical structure. *J. Am. Chem. Soc.*, 86:1616–1626, 1964.

[51] X. Hong, C. J. Harris, and S. Chen. Robust neurofuzzy rule base knowledge extraction and estimation using subspace decomposition combined with regularization and D-optimality. *IEEE Trans. Syst., Man., Cybern. B*, 34(1):598–608, 2004.

[52] F. H. Hurly and T. P. Weir. Electrodeposition of metals from fused quaternary ammonium salts. *J. Electrochem. Soc.*, 98:203, 1951.

[53] P. Isnard and S. Lambert. Estimating Bioconcentration factors from octanol-water partition coefficient and aqueous solubility. *Chemosphere*, 17:21–34, 1988.

[54] O. Ivanciuc. Artificial neural networks applications. Part 7. Estimation of bioconcentration factors in fish using solvatochromic parameters. *Revue Roumaine Chimie*, 43:347–354, 1998.

[55] J. S. R. Jang, C. T. Sun, and E. Mizutani. *Neuro-Fuzzy and Soft Computing; a Computational Approach to Learning and Machine Intelligence.* Prentice-Hall, Upper Saddle River, 1997.

[56] B. Jastorff et al. How hazadous are ionic liquids? Structure-activity relationships and biological testing as important elements for sustainability evaluation. *Green Chemistry*, 5:136–142, 2003.

[57] T. Johansen. Robust identification of Takagi-Sugeno-Kang fuzzy models using regularization. In *Proc. IEEE conf. Fuzzy Systems*, pages 180–186, New Orleans, USA, 1996.

[58] K. van der Jagt and S. Munn and J. Tørsløv and J. de Bruijn. Alternative approaches can reduce the use of test animals under REACH. Joint Research Centre, Report EUR 21405, 2004.

[59] K. L. E. Kaiser and S. P. Niculescu. Using Probabilistic Neural Networks to Model the Toxicity of Chemicals to the Fathead Minnow (*Pimephales promelas*): A Study Based on 865 Compounds. *Chemosphere*, 38:3237–3245, 1999.

[60] E. E. Kenaga and C. A. I. Goring. Relationship between water solubility and soil sorption, octanol-water partitioning and bioconcentration of chemicals in biota. In *Aquatic Toxicology, Special Technical Publication 707*, pages 78–115. American Society for Testing and Materials, Philadelphia, PA, 1980.

[61] L. B. Kier and L. H. Hall. *Molecular Connectivity in Structure-Activity Analysis.* John Wiley and Sons, New York, 1986.

[62] J. Kim, Y. Suga, and S. Won. A New Approach to Fuzzy Modeling of Nonlinear Dynamic Systems With Noise: Relevance Vector Learning Mechanism. *IEEE Trans. on Fuzzy Systems*, 14(2):222–231, April 2006.

[63] M. Koel. Physical and chemical properties of ionic liquids based on the dialkylimidazolium cation. In *Proc. Estonian acad. Sci. Chem.*, volume 49, pages 145–155, May 2000.

[64] T. Kohonen. *Self-Organizing Maps.* Springer-Verlag, Berlin, Germany, 2001.

[65] U. Kragl, M. Eckstein, and N. Kaftzik. Enzyme catalysis in ionic liquids. *Current Opinion in Biotechnology*, 13:565–571, 2002.

[66] U. Kragl, M. Eckstein, and N. Kaftzik. *Ionic Liquids in Synthesis.* VCH-Wiley, Weinheim, Germany, 2002.

[67] R. Krishnapuram and J. M. Keller. A possibilistic approach to clustering. *IEEE Trans. on Fuzzy Systems*, 1:98–110, May 1993.

[68] M. Kumar, N. Stoll, and R. Stoll. An energy-gain bounding approach to robust fuzzy identification. *Automatica*, 42(5):711–721, May 2006.

[69] M. Kumar, R. Stoll, and N. Stoll. Robust Adaptive Fuzzy Identification of Time-Varying Processes with Uncertain Data. Handling Uncertainties in the Physical Fitness Fuzzy Approximation with Real World Medical Data: An Application. *Fuzzy Optimization and Decision Making*, 2:243–259, September 2003.

[70] M. Kumar, R. Stoll, and N. Stoll. SDP and SOCP for outer and robust fuzzy approximation. In *Proc. 7^{th} IASTED International Conference on Artificial Intelligence and Soft Computing*, Banff, Canada, July 2003.

[71] M. Kumar, R. Stoll, and N. Stoll. Robust Adaptive Identification of Fuzzy Systems with Uncertain Data. *Fuzzy Optimization and Decision Making*, 3(3):195–216, September 2004.

[72] M. Kumar, R. Stoll, and N. Stoll. Robust Solution to Fuzzy Identification Problem with Uncertain Data by Regularization. Fuzzy Approximation to Physical Fitness with Real World Medical Data: An Application. *Fuzzy Optimization and Decision Making*, 3(1):63–82, March 2004.

[73] M. Kumar, R. Stoll, and N. Stoll. A Min-Max Approach to Fuzzy Clustering, Estimation, and Identification. *IEEE Trans. on Fuzzy Systems*, 14(2):248–262, April 2006.

[74] M. Kumar, R. Stoll, and N. Stoll. A Robust Design Criterion for Interpretable Fuzzy Models with Uncertain Data. *IEEE Trans. on Fuzzy Systems*, 14(2):314–328, April 2006.

[75] M. Kumar, R. stoll, and N. stoll. Deterministic Approach to Robust Adaptive Learning of Fuzzy Models. *IEEE Trans. Syst., Man., Cybern. B*, 36(4):767–780, August 2006.

[76] M. Kumar, K. Thurow, N. Stoll, and R. Stoll. Robust fuzzy mappings for QSAR studies. *European Journal of Medicinal Chemistry*, 42(5):675–685, 2007.

[77] S. Kumar, M. Kumar, R. Stoll, and U. Kragl. Handling uncertainties in toxicity modelling using a fuzzy filter. *SAR and QSAR in Environmental Research*, 18 (in press), 2007.

[78] S. Kumar, W. Ruth, and U. Sprenger, B. Kragl. On the biodegradation of ionic liquid 1-Butyl-3-methylimidazolium tetrafluoroborate. *Chimica oggi, Chemistry Today*, 24(2):24–26, March/April 2006.

[79] L. Sztandera and M. Trachtman and C. Bock and J. Velga and A. Garg. Soft computing and density functional theory in the design of safe textile chemicals. In L. M. Sztandera and C. Pastore, editor, *Soft computing in textile sciences*. Physica-Verlag GmbH, Heidelberg, Germany, 2003.

[80] X. Lu, S. Tao, H. Hu, and R. W. Dawson. Estimation of bioconcentration factors of nonionic organic compounds in fish by molecular connectivity indices and polarity correction factors. *Chemosphere*, 41:1675–1688, 2000.

[81] P. Luis, I. Ortiz, R. Aldaco, and A. Irabien. A novel group contribution method in the development of a QSAR for predicting the toxicity (*Vibrio fischeri* EC_{50}) of ionic liquids. *Ecotoxicology and Environmental Safety*, 67(3):423–429, Jul 2006.

[82] M. T. D. Cronin. Predicting Chemical Toxicity and Fate in Humans and the Environment - An Introduction. In M. T. D. Cronin and D. J. Livingstone, editor, *Predicting Chemical Toxicity and Fate*. CRC Press LLC, Boca Raton, Florida, 2004.

[83] D. Mackay. Correlation of bioconcentration factors. *Environmental Science & Technology*, 16:274–278, 1982.

[84] D. Mackay and A. Fraser. Bioaccumulation of persistent Organic Chemicals: mechanisms and Models. *Environmental Pollution*, 110:375–391, 2000.

[85] D. J. C. MacKay. Bayesian interpolation. *Neural Computation*, 4:415–447, 1992.

[86] C. J. Mathews et al. In situe formation of mixed phosphine-imidazolyidene palladium complexes in room-temperature ionic liquids. *Organometallics*, 20(18):3848–3850, 2001.

[87] M. Matzke et al. The influence of anion species on the toxicity of 1-alkyl-3-methylimidazolium ionic liquids observed in an (eco)toxicological test battery. *Green Chemistry*, page in press, 2007.

[88] P. Mazzatorta, E. Benfenati, C.-D. Neagu, and G. Gini. Tuning Neural and Fuzzy-Neural Networks for Toxicity Modeling. *J. Chem. inf. Comput. Sci.*, 43:513–518, 2003.

[89] G. McLachlan and K. Basford. *Mixture models: Inference and Applications to Clustering*. New York: Marcel Dekker, 1988.

[90] G. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. New York: John Wiley & Sons, 1997.

[91] G. McLachlan and D. Peel. *Finite Mixture Models*. New York: John Wiley & Sons, 2000.

[92] W. M. Meylan, P. H. Howard, R. S. Boethling, D. Aronson, H. Printup, and S. Gouchie. Improved method for estimating Bioconcentration/Bioaccumulation Factor from Octanol/Water Partition Coefficient. *Environmental Toxicology and Chemistry*, 18:664–672, 1999.

[93] A. L. Monteiro, F. K. Zinn, and R. F. de Souza. Asymmetric hydrogenation of 2-arylacrylic acids catalyzed by immobilized Ru-BINAP complex in 1-n-butyl-3-methylimidazolium tetrafluoroborate molten salt.

[94] P. A. P. Moran. Notes on continuous stochastic phenomena. *Biometrika*, 37:17–23, 1950.

[95] A. Noda and M. Watanabe. Highly conductive polymer electrolytes prepared by in situ polymerization of vinyl monomers in room temperature molten salts. *salts, Electrochim. Acta*, 45:1265, 2000.

[96] H. Olivier-Bourbigou and L. Magna. Ionic liquids: Perspectives for organic and catalytic reactions. *J. Molec. Catal. A.*, 182-183:419–437, 2002.

[97] Organization for Economic Cooperation and Development. Bioconcentration: Flow-through fish test. OECD Guide-Line for Testing of Chemicals: Draft Guideline 305, Paris, France, 1994.

[98] N. R. Pal, K. Pal, J. M. Keller, and J. C. Bezdek. A possibilistic fuzzy c-means clustering algorithm. *IEEE Trans. on Fuzzy Systems*, 13(4):517–530, August 2005.

[99] J. H. Park and H. J. Lee. Estimation of bioconcentration factor in fish, adsorption coefficient of soils and sediments and interfacial tension with water for organic nonelectrolytes based on the linear solvation energy relationships. *Chemosphere*, 26:1905–1916, 1993.

[100] J. Pernak, K. Sobaszkiewicz, and I. Mirska. Anti-microbial activites of ionic liquids. *Green Chemistry*, 5(1):52–56, 2003.

[101] M. Pintore, N. Piclin, E. Benfenati, G. Gini, and J. R. Chrétien. Predicting Toxicity against the fathead Minnow by Adaptive Fuzzy Partition. *QSAR Comb. Sci.*, 22:210–219, 2003.

[102] C. Pretti et al. Acute toxicity of ionic liquids to the zebrafish (*Danio rerio*). *Green Chem.*, 8(3):238–240, 2006.

[103] J. Ranke et al. Biological effects of imidazolium ionic liquids with varying chain lenghts in acute Vibrio fischeri and WST-1 cell viability assays. *Ecotoxicology and Environmental Safety*, 28(3):396–404, 2004.

[104] A. Ranwell and M. A. Tshamano. Potential application of ionic liquids for olefin oligomerization. *ACS Symposium Series 818*, pages 147–160, 2002.

[105] C. L. Russom, S. P. Bradbury, and S. J. Broderius. Predicting modes of toxic action from chemical structure: Acute toxicity in the fathead minnow (Pimephales promelas). *Environmental Toxicology and Chemistry*, 16(5):948–967, 1997.

[106] S. H. Schofer, N. Kaftzik, P. Wasserscheid, and U. Kragl. Enzyme catalysis in ionic liquids: lipase catalysed kinetic resolution of 1-phenylethanol with improved enantioselectivity. *Chem. Commun.*, pages 425–426, 2001.

[107] R. A. Sheldon. Green solvents for sustainable organic synthesis: State of the art. *Green Chem.*, 7:267–278, 2005.

[108] Stock, F. and Hoffmann, J. and Ranke, J. and Störmann, R. and Ondruschka, B. and Jastorff, B. Effects of ionic liquids on the acetylcholinesterase a structureactivity relationship consideration. *Green Chem.*, 6:286–290, 2004.

[109] S. Stolte et al. Anion effects on the cytotoxicity of ionic liquids. *Green Chem.*, 8:621–629, 2006.

[110] P. A. Z. Suarez et al. Two-phase catalytic hydrogenation of olefins by Ru(II) and Co(II) complexes dissolved in 1-n-butyl-3-methylimidazolium cation. *Inorganica Chimica Acta*, 255:207, 1997.

[111] R. P. Swatloski, J. D. Holbery, S. B. Memon, G. A. Caldwell, K. A. Caldwell, and R. D. Roger. Using Caenorhabditis elegans to probe toxicity of 1-alkyl-3-methylimidazolium chloride based ionic liquids. *Chem. Commun.*, 24(7):668–669, 2004.

[112] L. Sztandera, M. Trachtman, C. Bock, J. Velga, and A. Garg. Soft Computing in the Design of Nontoxic Chemicals. *J. Chem. Inf. Comput. Sci.*, 43:189–198, 2003.

[113] S. Tao, H. Hu, F. Xu, R. Dawson, B. Li, and J. Cao. Fragment constant method for prediction of fish bioconcentration factors of non-polar chemicals. *Chemosphere*, 41:1563–1568, 2000.

[114] I. V. Tetko, J. Gasteiger, R. Todeschini, A. Mauri, D. Livingstone, P. Ertl, V. A. Palyulin, E. V. Radchenko, N. S. Zefirov, A. S. Makarenko, V. Y. Tanchuk, and V. V. Prokopenko. Virtual computational chemistry laboratory - design and description. *J. Comput. Aid. Mol. Des.*, 19:453–463, 2005.

[115] R. Todeschini and V. Consonni. Handbook of molecular descriptors. In R. Mannhold, H. Kubinyi, and H. Timmerman, editors, *Methods and Principles in Medicinal Chemistry.* Wiley-VCH, Weinheim, 2000.

[116] G. D. Veith, D. L. DeFoe, and B. V. Bergstedt. Measuring and estimating the bioconcentration factor of chemicals on fish. *Journal of Fisheries Research Board of Canada*, 36:1040–1048, 1979.

[117] G. D. Veith and P. Kosian. Estimating bioconcentration potential from octanol/water partition coefficients. In D. Mackay, S. Paterson, S. J. Eisenreich, and M. S. Simons, editors, *Physical behavior of PCBs in the Great Lakes*, pages 269–282. Ann Arbor Sciences Publishers, Ann Arbor, 1983.

[118] M. Vracko. A Study of Structure-Carcinogenic Potency Relationship with Artificial Neural Networks. The Using of Descriptors Related to Geometrical and Electronic Structures. *J. Chem. Inf. Comput. Sci.*, 37:1037–1043, 1997.

[119] W. Y. Wang, T. T. Lee, C. L. Liu, and C. H. Wang. Function approximation using fuzzy neural networks with robust learning algorithm. *IEEE Trans. Syst., Man., Cybern. B*, 27:740–747, September 1997.

[120] P. Wasserscheid and W. Keim. New Solutions for Transition Metal Catalysis. *Angew. chem. Int. Ed.*, 39:3772–3789, 2000.

[121] D. Wei, A. Zhang, C. Wu, S. Han, and L. Wang. Progressive study and robustness test of QSAR model based on quantum chemical parameters for predicting BCF of selected polychlorinated organic compounds (PCOCs). *Chemosphere*, 44:1421–1428, 2001.

[122] A. S. Wells and V. T. Coombe. On the freshwater ecotoxicity and biodegradation properties of some common ionic liquids. *Organic Process Research and Development*, 10(4):794–798, 2006.

[123] J. S. Wilkes and M. J. Zaworodtko. Air and water stable 1-ethyl-3-methylimidazolium based ionic liquids. *J. Chem. Soc., Chem. Commun.*, 13:965–967, 1992.

[124] J.S. Wilkes, J.A. Levisky, R.A. Wilson, and C.L. Hussey. Dialkylimidazolium chloroaluminate melts: A new class of room temperatureionic liquids for electro chemistry, spectroscopy, and synthesis. *Inorg. Chem.*, 21:1263–1264, 1982.

[125] David A. Winkler. Neural Networks as Robust Tools in Drug Lead Discovery and Development. *Molecular Biotechnology*, 27:139–167, 2004.

[126] Wen Yu and Xiaoou Li. Fuzzy identification using fuzzy neural networks with stable learning algorithms. *IEEE Trans. on Fuzzy Systems*, 12(3):411–420, June 2004.

[127] L. A. Zadeh. Outline of a new approach to the analysis of complex systems and decision processes. *IEEE Trans. Syst., Man, Cybern.*, 3:28–44, January 1973.

[128] J. S. Zhang and Y. W. Leung. Improved possibilistic c-means clustering algorithms. *IEEE Trans. on Fuzzy Systems*, 12(2):209–217, April 2004.

Appendix A

A Gauss-Newton based Algorithm

Given N input-output data pairs $\{x(j), y(j)\}_{j=0}^{N-1}$, to compute the parameters

$$\begin{aligned}\theta_j &= \arg\min_\theta \left[\frac{[y(j) - G^T(x(j),\theta)\alpha_{j-1}]^2}{1+\mu\|G(x(j),\theta)\|^2} + \mu_\theta^{-1}\|\theta - \theta_{j-1}\|^2\right] \\ &= \arg\min_\theta \|r(\theta)\|^2, \text{ where } r(\theta) = \begin{bmatrix} \left[\frac{[y(j)-G^T(x(j),\theta)\alpha_{j-1}]^2}{1+\mu\|G(x(j),\theta)\|^2}\right]^{1/2} \\ \left(\mu_\theta^{-1}\right)^{1/2}(\theta - \theta_{j-1}) \end{bmatrix},\end{aligned}$$

$$\alpha_j = \alpha_{j-1} + \frac{\mu G(x(j),\theta_j)\left[y(j) - G^T(x(j),\theta_j)\alpha_{j-1}\right]}{1+\mu\|G(x(j),\theta_j)\|^2},$$

we use a Gauss-Newton based algorithm taken from [76]. The algorithm consists of following steps:

1. Choose initial guess about cluster centres θ_{-1}, number of maximum epochs E_{max}, $\alpha_{-1} = 0$, epoch count $EC = 0$, and data index $j = 0$.

2. If $EC < E_{max}$,

   (a) if $j \leq (N-1)$,

      i. define $r(\theta) = \begin{bmatrix} \left[\frac{[y(j)-G^T(x(j),\theta)\alpha_{j-1}]^2}{1+\mu\|G(x(j),\theta)\|^2}\right]^{1/2} \\ \left(\mu_\theta^{-1}\right)^{1/2}(\theta - \theta_{j-1}) \end{bmatrix}$ and let $s^*(\theta)$ be the unique solution of following linear least-squares problem:

$$s^*(\theta) = \arg \min_{s} \, [\|r(\theta) + r'(\theta)s\|^2],$$

where $r'(\theta)$ is the Jacobian matrix of vector r with respect to θ, determined by the method of finite-differences. The Jacobian $r'(\theta)$ is a full rank matrix, as a result of using regularization.

ii. compute $\theta_j = \theta_{j-1} + s^*(\theta_{j-1})$.

iii. compute

$$\alpha_j = \alpha_{j-1} + \frac{\mu G(x(j), \theta_j) \left[y(j) - G^T(x(j), \theta_j)\alpha_{j-1} \right]}{1 + \mu \|G(x(j), \theta_j)\|^2}.$$

iv. $j := j + 1$ and go to step 2(a).

(b) $EC := EC + 1$, $\alpha_{-1} := \alpha_{N-1}$, $\theta_{-1} := \theta_{N-1}$, $j = 0$, and go to step 2.

A MATLAB (a product of Mathworks, MA, USA) code was developed to implement the above algorithm. MATLAB is a high-level language and interactive environment to perform computationally intensive tasks.

Appendix B

List of Abbreviation

BCF	Bioconcentration factor
BOD	Biological oxygen demand
$[BF_4]$	tetrafluoroborate
[BMIM]	1-Butyl-3-methyl-imidazolium
[BMPy]	1-butyl-4-methylpyridinium
BTA	bis[(trifluoromethyl)sulfonyl]amid
COD	Chemical oxygen demand
$[C_1MIM][CH_3SO_4]$	1-*n*-Methyl-3-methyl-imidazolium methyl sulfate
$[C_2MIM][C_2H_5SO_4]$	1-*n*-Ethyl-3-methyl-imidazolium ethylsulfate
$[C_2MIM][Cl]$	1-*n*-Ethyl-3-methyl-imidazolium chloride
$[C_3MIM][BF_4]$	1-*n*-Propyl-3-methyl-imidazolium chloride
$[C_4MIM][PF_6]$	1-*n*-Butyl-3-methyl-imidazolium hexafluorophosphate
$[C_4MIM][BF_4]$	1-*n*-Butyl-3-methyl-imidazolium tetrafluoroborate
$[C_4MIM][Br]$	1-*n*-Butyl-3-methylmidazolium bromide
$[C_4MIM][Cl]$	1-*n*-Butyl-3-methylimidazolium chloride
$[C_4MIM][N(CN_2)_2]$	1-*n*-Butyl-3-methylimidazolium dicynamide
$[C_4EIM][BF_4]$	1-*n*-Butyl-3-ethylimidazolium tetrafluoroborate
$[C_5MIM][BF_4]$	1-*n*-Pentyl-3-methyl-imidazolium tetrafluoroborate
$[C_6MIM][Br]$	1-*n*-Hexyl-3-methyl-imidazolium bromide
$[C_6MIM][Cl]$	1-*n*-Hexyl-3-methyl-imidazolium chloride
$[C_6MMIM][Cl]$	1-*n*-Hexyl-2,3-dimethylimidazolium chloride
$[C_6MIM][PF_6]$	1-*n*-Hexyl-3-methyl-imidazolium hexafluorophosphate
$[C_6MIM][BF_4]$	1-*n*-Hexyl-3-methyl-imidazolium tetrafluoroborate
$[C_6EIM][BF_4]$	1-*n*-Hexyl-3-ethyl-imidazolium tetrafluoroborate
$[C_7MIM][BF_4]$	1-*n*-Heptayl-3-methyl-imidazolium tetrafluoroborate
$[C_8MIM][Br]$	1-*n*-Octyl-3-methyl-imidazolium bromide
$[C_8MIM][Cl]$	1-*n*-Octyl-3-methyl-imidazolium chloride

[C_8MIM][PF_6]	1-*n*-Octyl-3-methyl-imidazolium hexafluorophosphate
[C_8MIM][BF_4]	1-*n*-Octyl-3-methyl-imidazolium tetrafluoroborate
[C_9MIM][BF_4]	1-*n*-Nonecyl-3-methyl-imidazolium tetrafluoroborate
[C_{10}MIM][Cl]	1-*n*-Decacyl-3-methyl-imidazolium chloride
[C_{10}MIM][BF_4]	1-*n*-Decacyl-3-methyl-imidazolium tetrafluoroborate
[MPy]	3-Methyl pyridine
[C_4Py][Br]	1-*n*-Butyl pyridinium bromide
[C_4MPy][Br]	1-*n*-Butyl-3-methyl pyridinium bromide
[C_4MMPy][Br]	1-*n*-.Butyl-3,5-dimethyl pyridinium bromide
[C_4Py][Cl]	1-*n*-Butyl pyridinium chloride
[C_4Py][$N(CN_2)_2$]	1-*n*-Butyl pyridinium dicynamide
[C_4MPy][$N(CN_2)_2$]	1-*n*-Butyl-3-methyl pyridinium dicynamide
[C_4MMPy][$N(CN_2)_2$]	1-*n*-Butyl-3,5-dimethyl pyridinium dicynamide
[C_6MPy][Br]	1-*n*-Hexyl-3-methyl pyridinium bromide
[C_6MPy][Cl]	1-*n*-Hexyl-3-methyl pyridinium chloride
[C_8MPy][Br]	1-*n*-Octyl-3-methyl pyridinium bromide
[C_6MPyRR][Cl]	1-*n*-Hexyl-1-methyl pyrrolidinium chloride
[C_2MIM][$(C_2H_5)_2PO_4$]	1-*n*-Ethyl-3-methyl-imidazolium diethylphosphate
[C_2MIM][$C_7H_7SO_3$]	1-*n*-Ethyl-3-methyl-imidazolium tosylate
[C_2MIM][(2-OPhO)B]	1-*n*-Ethyl-3-methyl-imidazolium bis(1,2-benzenediolate)borate
[C_4MIM][$(CF_3)_2N$]	1-*n*-Butyl-3-methyl-imidazolium bis(trifluoromethyl)imide
[C_4MIM][$(CF_3SO_2)_2N$]	1-*n*-Butyl-3-methyl-imidazolium bis(trifluoromethylsulfonyl)imide
[C_4MIM][octylOSO_3]	1-*n*-Butyl-3-methyl-imidazolium octylsulfate
[C_1OMIM][BF_4]	1-*n*-methyl-3-octyl-imidazolium tetrafluoroborate
EC_{50}	Half maximal effective concentration
ECOENG212	1-Ethyl-3-methyl-imidazolium ethylsulfate
ECOENG2122P	1-Ethyl-3-methyl-imidazolium diethylphosphate
e.g	for example
[$EtSO_4$]	ethylsulfate
EU	European Union
GC-MS	Gaschromatography Massenspectrometery
IL	Ionic liquid
LC_{50}	Half maximum Lethal Concentration
[MIM]	3-Methylimidazolium
PFA	Principal Feature Analysis
[PF_6]	hexafluorophosphate
QSAR	quantitative structure-activity relationship
REACH	Registration, Evaluation, and Authorization of Chemicals
SDS	Sodium n-dodecyl sulfate
SOM	Self organizing maps
SPE	Solid phase extraction
[TOS]	Tosylate

Appendix C

Materials and Methods

C.1 List of Chemicals

chemical	CAS-Nr.	address
1-*n*-Butyl-3-methyl-imidazolium tetrafluoroborate	17451-65-6	Solvent innovation GmbH, Köln
1-*n*-Butyl-4-methyl pyridinium bromide	343952-33-0	Fluka (Sigma-Aldrich Laborchemikalien GmbH)
1-*n*-Ethyl-3-methyl-imidazolium diethylphosphate	not known	Solvent innovation GmbH, Köln
1-*n*-Ethyl-3-methyl-imidazolium tosylate	328090-25-1	Solvent innovation GmbH, Köln
1-*n*-Ethyl-3-methyl-imidazolium ethylsulfate	342573-75-5	Solvent innovation GmbH, Köln
1-*n*-Ethyl-3-methyl-imidazolium bis[(trifluoromethyl)sulfonyl]amid	not known	Solvent innovation GmbH, Köln
Sodium *n*-dodecyl sulfate	8012-56-4	Carl Roth GmbH & Co. Kg, Karlsruhe

C.2 List of Apparatus

	type
Microscope	Zeiss, sterikroscope stemi 2000-C KL 750, HWS, Germany
Luminometer	Bio-Orbit 1250, Labsystems, Turku, Finland
Oximeter	WTW Oxi 330

C.3 Methods

C.3.1 Closed bottle test

The biodegradability of a set of ionic liquids was investigated using standard "closed bottle test". The reference substance taken was Sodium *n*-dodecyl sulfate. The test and reference substances were prepared in an aerated mineral medium with a concentration of 2 mg/L. The solutions were inoculated with the secondary effluent (collected form activated sludge treatment plant). After well-mixing, the solutions were filled into the BOD bottles. For each ionic liquid as well as for blank and reference, triplicate bottles for each of three series (i.e. of 7 days, 14 days, and 28 days) were analyzed immediately for dissolved oxygen and closed tightly. The BOD bottles were incubated at 20°C in dark. These bottles were withdrawn in triplicate for an analysis of the dissolved oxygen over the period of 7 days (first series), 14 days (second series), and 28 days (final series). The chemical oxygen demand for each ionic liquid and reference was determined. And the biodegradation was expressed as the ratio of BOD (mg O_2) to COD (mg O_2).

C.3.2 Bioluminescence inhibition assay with marine bacteria *Vibrio fischeri*

A standard bioluminescence inhibition assay [2] was carried out for a set of ionic liquids. The test bacteria *Vibrio fischeri* DSM 7151/ NRRL B-11177 was purchased from DSMZ (Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH, Braunschweig, Germany). The ionic liquids were diluted to 10000 ppm in sterile 2% NaCl solution. The entire procedure is followed according to the standard guidelines of test protocol. The 500μL aliquots containing the bacterial suspension were pre-incubated at 15°C for 15 minutes before measuring the initial luminescence and before adding the diluted ionic liquids. The control was also run in parallel. The EC_{50} (the effective concentration resulting in the 50% reduction of the light produced by bacteria) values were determined at 15 and 30 minutes thrice for each ionic liquids. The luminescence was measured with the luminometer. The marine bacteria *Vibrio fischeri* exhibits bioluminescence as a result of a series of metabolic reaction. The decrease in the bioluminescence is proportional to the toxicity of the test substance.

Appendix D

Declaration of Originality

I, Shefali Kumar, hereby certify that this material, which I now submit for assessment leading to the award of Doctor rerum naturalium (Dr. rer. nat.) is entirely my own work and to the best of my knowledge, it contains no material previously published, or substantially overlapping with material submitted for the award of any other degree at any institution, except where due acknowledgment is made in the text.

. . . .

Rostock, October 26, 2007 Shefali Kumar

Appendix E

Curriculum Vitae

Fritz Reuter Str. 56
D-18057 Rostock
fon: +49(381)498-6451 (O)
fon: +49(381)461-3851 (H)

SHEFALI KUMAR

Education

PhD	University of Rostock. 2007 (expected).
Dilpoma	Chemistry, University of Rostock. 2004.
B.Sc.	Chemistry, University of Rajasthan Jaipur, India. 2000.
10+2	T.P.S senior Secondary School, Jaipur. 1997.

Projects and Training

Three Months training at MDC BERLIN (Germany), BUCH, Liposome Chemistry sponsored by IAESTE-DAAD, 2000.

“Isolation and Enrichment of Dye Degrading Micro organisms”, Birla Institute of Scientific Research, Jaipur (India), 1999.

Certified Work on “AIDS a Social Pollution” at National Children Science Congress, Assam (India), 1996.

"Effect of Gulf War on Environment", Project at Nehru Science Center Bombay (India), 1995.

Honors

Scholarship awarded by "Deutsche Bundesstiftung Umwelt" to pursue PhD.

Received "DAAD Student Award" at University of Rostock in 2004.

Received "Gold Medal" for securing top position in the University of Rajasthan, 2000.

Received "Best Student Award" from University of Rajasthan, 2000.

Second position in university level environmental quiz, department of microbiology, University Maharani College, Jaipur, 1999.

Awarded as "Best Child Scientist" at National Children Science Congress, Department of Science and Technology, Govt. of India. 1995-1996.

Field of interest:- Environmental Chemistry, Application of Artificial intelligence in Environmental Sciences.

Personal Details

Date of Birth: July 19, 1978.

Place of Birth: Jaipur (INDIA).

Marital Status: Married.

Appendix F

List of Publications

- S. Kumar, W. Ruth, B. Sprenger, and U. Kragl. On the biodegradation of ionic liquid 1-Butyl-3-methylimidazolium tetrafluoroborate. *Chemistry Today*, 24(2):24-26, March/April 2006.

- S. Kumar, M. Kumar, R. Stoll, and U. Kragl. Handling Uncertainties in Toxicity Modeling using a Fuzzy Filter. *SAR and QSAR in Environmental Research*, 18 (in press), 2007.

- S. Kumar, M. Kumar, K. Thurow, R. Stoll, and U. Kragl. Fuzzy Filtering for Robust Bioconcentration Factor Modelling. *Environmental Modelling and Software*, submitted for publication.

- S. Kumar, W. Ruth, B. Sprenger, and U. Kragl. Environmental Behaviour of Ionic Liquids. In *Proc. COIL*, Salzburg, Austria, June 2005.

- M. Kumar, S. Kumar, U. Kragl, and R. Stoll. Modelling Structural Behaviour of LD_{50} Toxicity. *Dokumentation Deutsche Gesellschaft für Arbeitsmedizin und Umweltmedizin e.V. 47, Jahrestagung*, Mainz, March 2007.

- S. Kumar. Neue Anwendungen für ionische Flüssigkeiten und ihr Verhalten in der Umwelt. *DBU-Workshop "Ionische Flssigkeiten und alternative Lsungsmittel"*, Osnabrück, April 2005.