Realia et naturalia

DISSERTATIONES
CHIMICAE
UNIVERSITATIS
TARTUENSIS
143

GEVEN PIIR

Environmental risk assessment of chemicals using QSAR methods

DISSERTATIONES CHIMICAE UNIVERSITATIS TARTUENSIS
143

DISSERTATIONES CHIMICAE UNIVERSITATIS TARTUENSIS
143

GEVEN PIIR

Environmental risk assessment of chemicals using QSAR methods

UNIVERSITY OF TARTU
PRESS
1632

Institute of Chemistry, Faculty of Science and Technology, University of Tartu, Estonia

This Dissertation is accepted for the commencement of the degree of Doctor of Philosophy in Molecular Engineering in 29 April 2015, by the Doctoral Committee of the Institute of Chemistry, University of Tartu

Supervisor: Dr. Sulev Sild, University of Tartu, Tartu, Estonia

Opponent: Prof. Alexandre Varnek, University of Strasbourg, Strasbourg, France

Commencement: August 28, 2015 at 10:00, Ravila 14A, room 1021

This work has been partially supported by Graduate School „Functional materials and technologies" receiving funding from the European Social Fund under project 1.2.0401.09-0079 in University of Tartu, Estonia.

ISSN 1406-0299
ISBN 978-9949-32-843-7 (print)
ISBN 978-9949-32-844-4 (pdf)

University of Tartu Press
www.tyk.ee

Öökapile

CONTENTS

LIST OF ORIGINAL PUBLICATIONS

This thesis is based on three publications, listed below. All these papers are denoted in the text by Roman numerals I–III.

I. "QSAR model for the prediction of bio-concentration factor using aqueous solubility and descriptors considering various electronic effects"
   Piir, G.; Sild, S.; Roncaglioni, A.; Benfenati, E.; Maran, U. *SAR QSAR Environ. Res.* **2010**, 21, 711–729.
II. "Comparative analysis of local and consensus quantitative structure activity relationship approaches for the prediction of bioconcentration factor"
   Piir, G.; Sild, S.; Maran, U. *SAR QSAR Environ. Res.* **2013**, 24, 175–199.
III. "Classifying bio-concentration factor with random forest algorithm, influence of the bio-accumulative vs. non-bio-accumulative compound ratio to modelling result, and applicability domain for random forest model"
   Piir, G.; Sild, S.; Maran, U. *SAR QSAR Environ. Res.* **2014**, 25, 967–981.

Author's contribution

Publication I: The author is responsible for preparation of data sets, calculations, analysis, and manuscript preparation.

Publication II: The author is responsible for preparation of data sets, calculations, analysis, and manuscript preparation.

Publication III: The author is responsible for preparation of data sets, calculations, analysis, and manuscript preparation.

LIST OF ABBREVIATIONS

ERA	Environmental risk assessment
BCF	Bioconcentration factor
QSAR	Quantitative structure-activity relationship
QSPR	Quantitative structure-property relationship
log P	Logarithm of n-octanol/water partition coefficient
LOO	Leave-One-Out cross-validation
LMO	Leave-Many-Out cross-validation
PPV	Positive predictive value
NPV	Negative predictive value
TP	True positive prediction
TN	True negative prediction
FP	False positive prediction
FN	False negative prediction
MLR	Multi-linear regression
PLS	Partial least squares regression
ANN	Artificial neural network
SVM	Support vector machine
RF	Random Forest algorithm
OOB	Out-of-bag sample
EM	Expectation-maximization algorithm
AD	Applicability domain

INTRODUCTION

Risk assessment in general deals with estimating magnitudes and probabilities of adverse effects associated with an event[1]. Earliest risk assessments were related to gambling, but in seventeenth-century they were implemented to determine the likelihood that a trading ship would return successfully[2]. Since then, risk assessments have been conducted in many fields from finance to human health. Risk assessment in different fields considers series of factors that may cause an adverse effect. These factors vary from field to field, but the main goal is to systematically analyse those factors and estimate the probability of adverse effects.

Every year, thousands of chemical compounds are released into the environment. Logically, one would assume that we know how those chemicals behave in the environment. However, for many chemicals used today there is no information about their environmental properties. In addition, chemicals not produced commercially are also present in the environment. Usually, such chemicals are degradation products or impurities of industrial substances[3]. Over the years, awareness of the importance of environmental risk assessment has increased. Unfortunately, boosted increase in awareness follows major ecological catastrophes. For example, major chemical (Sandoz chemical spill, 1986) and oil (Deepwater Horizon oil spill, 2010) spills[4]. However, chemical behaviour in the environment is a complex process. Therefore, estimating the impact on the environment involves large uncertainty[3]. Many experimental tests were (and are being) developed to assess the risk chemicals pose to the environment. Yet, many experimental tests last days or even months, the cost of one test can amount to 100,000 euros, and hundreds of animals are used during the test[5]. Considering the factor of time, money, and animal testing, it is not feasible to measure the properties of all the chemicals experimentally.

To fill the gaps in the data and reduce animal testing, quantitative structure-activity relationships (QSAR) are proposed as an alternative solution. With the approval of the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH)[6] legislation, the usage of QSARs in regulatory purposes got legal encouragement. Even though legal support for using QSARs in regulatory purposes exists, one must understand that not all QSARs are applicable. Valid QSAR must meet certain requirements (OECD principles[7]). According to Technical Guidance Document on Risk Assessment[8], validated model can be used to:

- assist in data evaluation
- make a decision for further testing strategies
- establish specific parameters needed in the risk assessment
- identify data gaps on effects of potential concern.

Another point worth mentioning is that QSAR is a prediction tool and therefore there is a possibility that the prediction can be inaccurate. For that reason, instead of using one model, it is encouraged to use multiple models to build confidence in decision-making.

The full process of environmental risk assessment consists of four steps: hazard identification, exposure assessment, effects assessment, and risk characterisation[4]. Those steps consider effects on terrestrial, aquatic and air compartments. The hazard of chemicals is usually assessed by persistence, bioaccumulation potential, toxicity, and potential for long-range transport[9]. This thesis focuses on exposure assessment in the aquatic environment, more precisely on the assessment of bioaccumulation potential of chemicals using different QSAR methods. It provides a general overview about ERA and endpoints used in ERA. A more thorough overview is given about the bioaccumulation potential and the methodologies used in this work. Finally, it summarises the results of the original research. It includes a global model proposed for predicting BCF and methods to derive subsets for local models to enhance prediction accuracy for BCF. In addition, classification models with new applicability domain approach for BCF are proposed. Classification models explore how different chemical distribution schemas on training phase influence the results. The proposed applicability domain approach is specific for Random Forest based QSAR models.

1. ENVIRONMENTAL RISK ASSESSMENT

Environmental risk assessment (ERA) tries to eliminate or reduce risks that can be posed by chemicals to the environment. The complexity of ecological systems makes the estimation of risks highly uncertain. Despite that, there is a great interest in environmental risk assessment to scientifically prioritise problems with the greatest environmental risk and to find ways to reduce uncertainty[1].

ERA is part of the risk management process. The whole process consists of eight steps[4]:

- Hazard identification
- Exposure assessment
- Effects assessment
- Risk characterisation
- Risk benefit analysis
- Risk reduction
- Monitoring and review

From these steps, the first four belong to the risk assessment phase. This thesis focuses only on the risk assessment of chemicals in the environment. Therefore, a short overview is given only for the risk assessment steps.

Hazard identification or problem formulation is the phase where potential hazards are acknowledged and further studies planned. This means that the available data is analysed to identify the source and exposure characteristics, to determine which ecosystems are potentially at risk and what kind of ecological effects are present[1]. In this step, assessment endpoints are selected. Assessment endpoint is defined as "an explicit expression of the environmental value that is to be protected, operationally defined by an ecological entity and its attributes."[10] Endpoints are used to give a mechanistic interpretation to the environmental value that needs protection[1]. It is important that selected endpoints are ecologically relevant, susceptible, and relevant to management goals[11]. Next step is to develop a conceptual model, which describes hypothesised relationships between sources and the assessment endpoint response[2]. The last step in this phase is to formulate an analysis plan. This plan specifies the assessment design, data needs, assumptions, extrapolations, and methods used for the analysis[1].

Exposure and effects assessments can be grouped together as analysis phase. Exposure assessment deals with measurements or predictions of the emission, transport, fate and behaviour of the chemicals in the environment[12]. This assessment consists of three different tasks: measuring exposure, analysing exposure and profiling exposure[2]. Exposure measurements determine concentrations of chemicals in different environmental compartments. Concentrations

can be measured experimentally or predicted by using some model[4]. The next step is to analyse samples taken from different spaces at different times and finally build an exposure profile, which summarises the results[2].

Effects assessment (dose-response assessment) is used to estimate the relationship between chemical concentration and observable effect[1]. Exposure assessment is also done in three steps. The first step deals with data gathering. Data is often obtained from QSAR models, read-across analysis, *in vitro* studies or from experimental laboratory studies[4]. The second step analyses gathered data to identify the relationships between chemicals and ecological effects. The final step determines the relationship between the scale and length of exposure and the endpoint effects[2].

The final part of the chemical risk assessment is risk characterisation where results from the analysis phase are put together to estimate potential risks. Risks can be characterised as effect/no effect ratios or probabilities[12]. Therefore, in this phase the assessment strengths, limitations, assumptions, and main uncertainties are outlined[1]. The main goal of this step is to provide the risk manager with a detailed overview of the probable risks.

2. ENDPOINTS IN ENVIRONMENTAL RISK ASSESSMENT

The environmental risk of chemicals can be assessed with many different endpoints. Each of them gives some specific piece of information about the environmental risk. Obviously, not all of them have equal impact for risk assessment. Therefore, different agencies have developed so-called priority lists, which describe endpoints that are relevant to them. According to REACH legislation[6], the endpoints needed for the registration of chemicals are categorized into three groups: physico-chemical information, toxicological information, and eco-toxicological information. This chapter gives a brief overview of eco-toxicological endpoints. As the sole endpoint used in this work, bioaccumulation process is described more thoroughly.

Eco-toxicological endpoints in REACH are divided into six general groups[6]:

1. Aquatic toxicity
   - Short-term toxicity testing on invertebrates
   - Growth inhibition study aquatic plants
   - Short-term toxicity testing on fish
   - Activated sludge respiration inhibition testing
   - Long-term toxicity testing on invertebrates
   - Long-term toxicity testing on fish
   - Fish early-life stage (FELS) toxicity test
   - Fish short-term toxicity test on embryo and sac-fry stages
   - Fish juvenile growth test
2. Effect on terrestrial organisms
   - Short-term toxicity to invertebrates
   - Effects on soil micro-organisms
   - Short-term toxicity to plants
   - Long-term toxicity testing on invertebrates
   - Long-term toxicity testing on plants
3. Effects on sediment organisms
   - Long-term toxicity to sediment organisms
4. Toxicity to birds
   - Long-term or reproductive toxicity to birds
5. Degradation
   - Biotic
     - Ready biodegradability
     - Simulation testing on ultimate degradation in surface water
     - Soil simulation testing
     - Sediment simulation testing

- Abiotic
  - Hydrolysis as a function of pH
  - Identification of degradation products

6. Fate and behaviour in the environment
   - Adsorption/desorption screening
   - Bioaccumulation in aquatic species
   - Further information on adsorption/desorption
   - Further information on the environmental fate and behaviour of the substance and/or degradation products

In European Union alone (in 2008), more than 12 million vertebrates were used for different kinds of tests[13]. The number of animals that were subjects for tests above was around two million. These numbers encourage seeking alternative approaches where possible. One of the alternatives is to use predictive models. Not all of the endpoints above are covered equally well. For example, first four groups of endpoints contain mostly different acute and chronic toxicity tests for various organisms. However, most of the predictive models are developed to cover acute aquatic toxicity. Different approaches for modelling acute aquatic toxicity has been reviewed by Netzeva and coworkers[14]. Other toxicity endpoints are rarely used for modelling. Degradation endpoints cover decomposition of chemicals through biotic or abiotic processes. Here are most models developed for biodegradability. Available models and known problems about modelling biodegradability were recently summarised by Rücker and Kümmerer[15]. If the fate and behaviour of the chemicals in soils are assessed by adsorption/desorption screening, then the fate and behaviour of the chemicals in organisms are tested with bioaccumulation tests. Some soil sorption models have been reviewed by Doucette[16] and bioaccumulation models by Pavan and coworkers[17]. Bioaccumulation is covered in more detail in the following sections.

2.1. Bioconcentration factor

Bioaccumulation is a process where the concentration of chemicals in an organism can reach a higher level than in the surrounding environment. In the bioaccumulation process, all possible routes are considered as an entry route to the organism[18]. In a laboratory, bioaccumulation is usually measured as a bioconcentration. The latter is a special case of bioaccumulation where absorption of chemicals from environment takes place only through non-dietary routes. Therefore bioconcentration factor (BCF) is defined as a proportion of chemical concentration in the organism (C_o) and concentration in the water (C_w) at steady state (equation 1)[19]. In essence, bioconcentration is the total of uptake, distribution and elimination processes of a chemical due to aqueous exposure[18]. According to REACH legislation[6], the preferable organism for BCF measure-

ments is fish, but other aquatic organisms, like mussels, are also used. Guidelines to measure bioconcentration experimentally are described in OECD test no. 305[5]. Since the evaluation of chemical compound's bioaccumulation potential is based mainly on the bioconcentration factors in fish, this overview concentrates on the effects that influence the bioconcentration in fish.

$$BCF = \frac{C_o}{C_w} \quad (1)$$

2.1.1. Uptake

Uptake of the chemicals starts by transporting the chemicals to the potential sites of absorption. By definition, BCF reckons only non-dietary routes of uptake; therefore, skin and gills are the most probable sites of absorption in fish. The most significant site of chemical uptake is believed to be gills. The reasons for that are high countercurrent flows of blood and water, large absorbing surface, and a small diffusion distance between blood and water[20]. Skin is considered as less important chemical uptake site. There is an exception though. For very small fish, absorption through skin has higher importance because of the large surface-to-volume ratio[20]. Once the chemicals are transported to the absorbing epithelium, they must pass the cell membrane. In principle, there are four methods how chemicals can penetrate the cell membrane. The most important process for uptake is passive diffusion followed by carrier molecule mediated transport. Less significant are filtration through aqueous pores and endocytosis[21].

Chemical's passive diffusion through the cell membrane is influenced by the concentration differences on both sides of the membrane and by the ease or difficulty with which the molecules can transfer through the lipophilic inner part of the membrane[22]. The first part of the passive diffusion is explained with the concentration gradient between water and organism. This can result only concentration equilibrium in both mediums. Therefore, passive diffusion for the bioconcentration phenomena can be described as a fugacity difference between water and organism. Compared to water, higher concentration in fish can be explained by the higher capacity of fish to store chemicals per unit of volume[18]. In equations, concentrations are usually used instead of fugacities. Secondly, not all chemicals move through the cell membrane with the same ease. Lipophilicity, molecular size and degree of ionisation significantly influence the passage through the cell membrane. The most important is lipophilicity (lipid solubility). The lipophilicity of a chemical is often characterised by n-octanol/water partition coefficient (log P). Chemicals with higher log P value are more soluble in lipids. Usually, more lipophilic chemicals move across the cell membrane more easily. Molecular size and shape can significantly reduce the cell penetration ability. From molecules with similar lipophilicity, larger and

branched molecules may penetrate cell membrane more slowly because of the frictional resistance and steric hindrance[22]. Reduced permeability can also be explained by the degree of ionisation. Due to the lipid nature of the cell membrane, non-ionised form of chemicals is more permeable than ionised form[21].

Filtration uses the same mechanism as passive diffusion. It is used to transport small water-soluble molecules through small pores or water channels in the membrane. Carrier molecules (proteins bonded with membrane) are used to transport molecules that do not have sufficient lipophilicity to pass membrane using passive diffusion. This transportation mode can be divided into two: facilitated diffusion and active transport. The first one uses concentration gradient and does not need additional energy. The second one requires additional energy input and can move chemicals through the membrane even when equilibrium on both sides is achieved[21]. With endocytosis, large and non-soluble chemicals can be transported into the cell. In this case, the cell surrounds the chemical and forms a vesicle, which is engulfed into the cell interior[22].

2.1.2. Distribution

Once the chemical is absorbed from the water, it ends up in the vascular system. Chemicals are then distributed around the body into different tissues by the blood flow. Transport of the chemicals into the tissues is influenced by the rate of blood flow and the ability to bind to blood plasma protein[21]. At first, chemicals are delivered from blood to high-perfusion tissues, such as liver. Then to low-perfusion tissues, such as muscle, and finally to lipoidal tissues[20]. It is a reversible movement of chemical from blood to tissue or from extracellular to intracellular compartments[21].

Processes familiar from uptake are also responsible for distribution. Most chemicals move from the blood to the tissues by passive diffusion. Therefore, it is influenced by the same physico-chemical properties (lipophilicity etc.) as discussed before. At first, chemicals are distributed to better-perfused tissues. Nevertheless, over time many chemicals are redistributed to other tissues. Chemicals are relocated in accordance with their relative affinity for tissue constituents[23]. For example, water-soluble chemicals will spread throughout the water present in the body, while lipophilic chemicals are stored in fat containing tissues.

2.1.3. Elimination

The elimination of chemicals from organism occurs simultaneously with the distribution. Elimination is a combination of different processes. For example, neutral, water-soluble and low-molecular-weight chemicals are mostly eliminated through the gills by passive diffusion. In addition, many organic

chemicals are eliminated by urine or bile[23]. The latter two are controlled by the kidneys and liver, respectively. In fish 80% of the cardiac output goes through kidneys[23] where blood is being filtrated. Filtered chemicals are then either excreted with urine or reabsorbed using passive diffusion[22]. Excreting chemicals is one of the liver's key roles. The excretion rate depends on the chemical uptake by the hepatocytes, the formation and secretion of bile, and the transport of chemicals into bile[23]. Liver also supports elimination by metabolising chemicals. Chemicals are converted to more polar metabolites[22] which usually leads to faster elimination because of the decreased reabsorbing ability. Growth and reproductive transfer can also lead to lower concentration of the chemicals in the organism[18]. The same amount of moles of a chemical in a bigger organism means smaller concentration and the mother can transfer large amounts of chemical to the egg. The last two, however, do not contribute significantly to the result of the experiment.

3. METHODS IN BCF MODELLING

Quantitative structure-activity (property) relationships seek relations between molecular structure and some biological activity or physico-chemical property. The theory of QSAR relies on the assumption that similar chemicals have similar properties. Because of this and since experimental measurements of different endpoints are expensive and time-consuming, it makes sense to use QSAR models in environmental risk assessment and in other areas of research. The current chapter discusses the methods and information needed for QSAR analysis. The focus is on methods that are present in the models developed for BCF. The methods used in the current work are covered in more detail.

QSAR models can be used for different purposes. Regression (quantitative) models are used to predict numerical values, and classification (qualitative) models are used to predict belonging to a certain class. A short overview of the methods used in both of these models is given by Yee and Wei[24]. The complexity of the models is determined by the nature of the used mathematical algorithms. Simple relationships can be predicted by linear models, but relationships that are more complex need non-linear models. Algorithms used in linear and non-linear modelling are reviewed by Dudek and others[25]. Models also have different applicability boundaries. The molecular composition of the data set is used to distinguish global and local models. Global models are built using data sets with wide variety of chemicals, while local models are built using a distinct class of chemicals. At the same time, all models can be part of the consensus model. Consensus model uses multiple models and the prediction is made by averaging the predictions of all of the used models.

Evidently, models have different development and usage strategies. A reliable model, however, must meet certain requirements. According to the Organisation for Economic Co-operation and Development (OECD)[7] a reliable QSAR model must have a defined endpoint, an unambiguous algorithm, a defined domain of applicability, appropriate measures of goodness-of-fit, robustness and predictivity, and if possible, a mechanistic interpretation. All of those points will be discussed throughout this thesis.

3.1. Quantitative and qualitative models

3.1.1. Quantitative (regression) models

Regression analysis is a widely used technique in QSAR analysis. As a result, it has been used many times for the estimation of BCF[26–71] and it was also used in Article I and Article II of this thesis. This technique seeks relationships between activity (dependent variable) and molecular descriptors (independent variables). The latter are numerical representations of chemical structures. A general mathematical representation is shown in equation 2, where A is activity under investigation and D stands for a set of molecular descriptors. How relationship

between dependent and independent variables in the regression function is determined defines whether the model is linear or non-linear (see chapter 3.2).

$$A = f(D) \tag{2}$$

For model building, two kinds of data sets are needed. Experimental data with continuous values is the first item needed for the modelling. Articles accompanying this thesis use experimental data for BCF (see chapter 2.1). One must bear in mind though that the quality of the model depends on the quality of the data set. A set of molecular descriptors is the second item necessary for the modelling. The molecular descriptors are either empirical or theoretical. Empirical descriptors use experimental data, but theoretical descriptors are calculated directly from molecular structure[72]. The most relevant descriptors used to model BCF are more thoroughly discussed in chapter 3.4.

Goodness-of-fit of the regression models is assessed with different statistical approaches[72]. The most frequent are the coefficient of determination (R^2, equation 3) and standard deviation (σ, equation 4).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(A_i - \hat{A}_i\right)^2}{\sum_{i=1}^{n}(A_i - \bar{A})^2} \tag{3}$$

In equation 3, $\hat{A}_i$ is a calculated activity, $\bar{A}$ is a mean value of experimental activity and A_i is an experimental activity. R^2 is used to determine how well the calculated values correspond to the experimental values. The standard deviation is defined as:

$$\sigma = \sqrt{\sum_{i=1}^{n}\frac{\left(\hat{A}_i - A_i\right)^2}{n}} \tag{4}$$

where n is the number of experimental values. Both, coefficient of determination and standard deviation, are used to select the final model. Usually the model with the best balance between those two is selected.

Goodness-of-fit measures only show how well the model works with the data that was used to derive the model. Models with good statistical parameters might be useless for predictions, because of overtraining or chance correlations. Therefore, models must be validated to ensure their predictive power. Several methods exist for validation. In general, they can be divided into two: internal and external validation methods. The most common internal validation is cross-validation, where a fraction of chemicals is left out and the selected model is rebuilt. Either one (leave-one-out, LOO) or many (leave-many-out, LMO) chemicals can be excluded. Each time one or more chemicals are left out, a

model is rebuilt with the same descriptors, and a prediction for left-out chemical(s) is made. This is repeated until all chemicals have predictions. Those predictions are used to calculate the cross-validated coefficient of determination (equation 3). Internally predictive model's R^2 and $R^2_{(LOO, LMO)}$ should have more or less the same value. It is proposed[73] that $R^2_{(LOO, LMO)}$ should be higher than 0.7, but a superior indicator of the prediction power is an external validation. External validation uses only chemicals that did not take part of the model building. The external validation set is compiled by excluding a certain amount of chemicals from the original data set. Another option is to use a new data set from a different source. It is important that the designed external validation set use chemicals that are representative of the chemicals used in the modelling process. To find out the "real" predictive power of the model, predictions for all the chemicals in the external validation set are used to calculate R^2_{ext}[73] (equation 3).

3.1.2. Qualitative (classification) models

Similarly, classification models seek relationships between activity and molecular descriptors. If regression models use experimental data with continuous values, then classification models use categorical values. Compared to the regression models, there are only a few classification models for BCF[74–77]. We can only speculate about the reasons. One of the reasons might be that it is logical to build a regression model when continuous values are available. The second reason might be that in the classification modelling, equal distribution of the classes is preferred. The real world data sets, however, are usually imbalanced. For classification, regression models can also be applied by using cut-off values to classify chemicals[33,39,51,78]. This prompts the question, why build a classification model when continuous values are available? In the case of BCF modelling, pure classification models give better results. It is because regression models have larger uncertainties in the areas that are near the cut-off value. Classification models for BCF were also proposed in Article III.

As for any model, the performance of the model should be assessed. The performance of the classification model is usually measured with five metrics. These are sensitivity (equation 5), specificity (equation 6), accuracy (equation 7), positive predictive value (PPV, equation 8) and negative predictive value (NPV, equation 9). These terms are calculated using the number of true positive (TP), true negative (TN), false positive (FP) and false negative (FN) predictions.

$$Sensitivity = \frac{TP}{TP + FN} \tag{5}$$

$$Specificity = \frac{TN}{TN + FP} \tag{6}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{7}$$

$$PPV = \frac{TP}{TP + FP} \tag{8}$$

$$NPV = \frac{TN}{TN + FN} \tag{9}$$

A confusion matrix (Figure 1) shows relations of those terms to one another more clearly. Each row in the confusion matrix displays how many chemicals are there in the actual class and each column shows chemicals in the predicted class.

		Predicted class		
		Class A	**Class B**	
Actual class	**Class A**	True Positive	False Negative	Sensitivity (**Eq. 5**)
	Class B	False Positive	True Negative	Specificity (**Eq. 6**)
		PPV (**Eq. 8**)	NPV (**Eq. 9**)	Accuracy (**Eq. 7**)

Figure 1. Confusion matrix

The sensitivity shows the proportion of correctly classified chemicals in the class A. The specificity measures the proportion of correctly classified chemicals in the class B. The accuracy gives the ratio between correctly classified chemicals (TP and TN) in the data set and all chemicals in the data set. PPV and NPV express the proportion of correctly classified chemicals within all chemicals classified as a certain class. To summarise, the higher the values of performance metrics, the better the performance of the model. These metrics are applicable to the training set and to the validation sets. Of course, it is best to evaluate the quality of the classification models with an external validation set.

3.2. Linear and non-linear methods

QSARs can be divided by the algorithm to linear and non-linear models. Most of the QSARs derived for BCF are based on linear algorithms. The simplest form of regression model is a one-parameter model (equation 10), where b_0 is the intercept of the model and b_1 is the slope of the regression line.

$$A = b_0 + b_1 x_1 \tag{10}$$

Traditionally, *log P* is used in the one-parameter BCF models[54]. However, other experimental descriptors, such as aqueous solubility[42], a chromatographic retention factor[26] and an artificial membrane accumulation index[37] are also proposed. Highly hydrophobic chemicals have rather weak correlation between *log BCF* and *log P*. Several methods were proposed to improve the correlation. Bintein and others[27] showed that compared to linear models, parabolic and bilinear models describe hydrophobic chemicals much better. Similarly, Dimitrov and others[32] built a non-linear model with *log P* as a single descriptor. Meylan and co-workers[57], however, used an approach where the applicable linear equation depends on the chemical's *log P* value. Those equations use *log P* together with other structural correction factors. Recently, Garg and others[38] proposed a model for highly hydrophobic chemicals. They used *log P* based parabolic model and complemented it with molecular volume.

Most of the linear models for BCF are multi-linear regression (MLR) models [29,40,41,45,49,50,52,53,55,58–69]. They all follow equation 11, but use different methods to select relevant descriptors (x_k) into the models.

$$A = b_0 + \sum_k b_k x_k \tag{11}$$

Short review of the methods used in QSAR/QSPR analysis is compiled by Liu and Long[79]. Most of the MLR models for BCF use either genetic algorithm[40,41,55,58,60] or step-wise regression[29,49,52,59,63,65,69] to select descriptors into the models. Article I and Article II use the best multi-linear regression (BMLR) method to select the descriptors. BMLR is a simple and straightforward step-wise forward selection approach and it is described in more detail in chapter 3.5.1. Another linear method that is used to model BCF in several works is partial least squares regression (PLS)[36,56,63,69]. PLS uses latent variables (linear combinations) from original descriptors to predict the activity[80].

Several models for BCF use non-linear algorithms[28,30–33,35,39,47,49,70,71,74–77]. The simplest form of non-linear model uses non-linear regression function[30–33,53]. Using artificial neural networks (ANN) is another way to develop a non-linear QSAR model[28,35,39,47,70,71]. Generally, ANN is constructed as a series of layers. Each layer contains one or more interconnected neurons. Neurons use non-linear activation functions with optimized weights to estimate the connections

between input (descriptors) and output (activity) variables. Another technique that is being used is support vector machine (SVM)[49,59,77], which uses previously constructed hyperplane to separate chemicals into different compartments[81]. Decision trees are by nature also non-linear methods. They are mostly used to solve classification problems[74–77]. Article III uses Random Forest (RF) algorithm[82], which is based on an ensemble of decision trees. The working principles of the RF-algorithm are explained in chapter 3.5.2.

3.3. Global, local and consensus models

Depending on the molecular composition of training sets, QSAR models are either global or local. Global models are built using data sets that cover a broad range of chemical space. Usually those data sets contain hundreds or even thousands of chemicals from different chemical classes. Models in Articles I and III are built using such data sets. Most of the developed models for BCF can be identified as global models[26,28–35,39–41,45,47,49–54,58–61,63–69,71,74–78]. For example, the largest data set (1,036 chemicals) used for the modelling of BCF was used by Toropova and co-workers[68]. As the name suggests, global models are all-purpose models and ideally could predict activities for all kinds of chemicals. Local models, in contrast, are built for a certain subset of chemicals, which share some structural or chemical similarity. These models can be used to predict activities only for chemicals that are analogous to corresponding subsets. Subsets are often designed by a human expert and in the case of modelling BCF are mostly based on structural similarity. Most of the local models for BCF are built for predicting polychlorinated biphenyls[43,56,62,70]. Chlorobenzenes[44], polybrominated biphenyl ethers[55] and polycyclic aromatic hydrocarbons[36] are also used as structurally similar subsets. Where subsets contain only highly hydrophobic chemicals, property based local models exist as well[29,38,58]. Another way of creating subsets for local models is creating them automatically using structural rules or some kind of machine-learning algorithm. The latter is used in Article II and a short overview of used algorithms is presented in chapter 3.5.3.

Another class of models is so-called consensus models. The basic idea behind consensus models is that multiple models are used to make a final prediction. Using predictions from different models has usually positive effect on the final prediction. One model might fail to predict some specific chemicals, but average prediction from multiple models usually cancels out those prediction errors. It is because one erroneous prediction has less weight than multiple correct predictions. Consensus of multiple models is often used in regulatory purposes, because for decision-making, one needs to use as much relevant information as possible. In the BCF modelling some consensus models were proposed[71,76–78] and Article II proposes several consensus models as well.

3.4. Most relevant descriptors in BCF modelling

To describe some features of a chemical compound, thousands of different molecular descriptors are defined and can be calculated using various software packages. Details about many of them are gathered by Karelson[72] and Todeschini et al[83,84]. This thesis focuses only on descriptors that have mechanistic relations to BCF.

The most commonly used descriptor in BCF modelling is the n-octanol/ water partition coefficient (*log P*) and it describes chemical's lipophilicity. Calculation of the experimental *log P* follows equation 12, but lack of the experimental *log P* data conduces to use predictive *log P* models. Models for BCF contain *log P* as a single descriptor[32,54] or with complementary descriptors[28–31,38,45,60,71,77]. It is logical to see *log P* in many models, because—as previously stated—lipophilicity plays an important role in penetrating the cell membrane. The models in Article II as well as those in Article III use *log P*. *Log P* is considered as an approximation of a cell membrane but it does not entirely describe the effects of bioaccumulation. Therefore, it is often complemented (or modelled altogether) with other descriptors that describe other properties that influence the bioaccumulation process. Usually, properties such as metabolism, molecular size, branching, flexibility, but also hydrogen bonding and ionisation are considered.

$$\mathrm{Log}\,P = log\frac{[C]_{n-octanol}}{[C]_{water}} \tag{12}$$

Molecular size with molecular branching and flexibility usually lessen bioaccumulation. Longer, more branched and flexible molecules are usually less bioaccumulative because of the frictional resistance and steric hindrance. Descriptors that consider the size of the molecule are present in many models. Molecular weight[39,64,75], molecular surface area[35], molecular volume[38,59] or maximal molecular diameter[30–33,75] are describing molecular size. Different connectivity indices add valuable information about molecular branching and flexibility[28,29,35,48,52,53,56,63,65,69]. Connectivity indices are calculated from the vertex degree of atoms in the H-depleted molecular graph[84]. Vertex degree counts atom's σ bonds, dismissing bonds with hydrogen.

Molecule's ability to form hydrogen bonds increases the probability that the molecule stays in the aqueous phase, at the same time, decreasing the molecule's potential to pass the cell membrane. The effects of hydrogen bonding are mostly described by counting either hydrogen bond donating/ acceptor groups[40,41] or using descriptors indirectly related to hydrogen bonding ability[28,29,58]. All these descriptors consider functional groups that have either strong electron-withdrawing ability (-OH, -NH) or lone pairs (-O, -N).

Cell permeability is also determined by chemicals' electrical nature. Ionised and polar chemicals have usually lower permeability. Different electronic

effects are accounted by using electro-topological state indices[45,56,69] and molecular electronegativity distance vectors (MEDV)[50,61,62]. From quantum-chemical descriptors, the most frequent descriptors are the highest occupied molecular orbital energy (HOMO) and the lowest unoccupied molecular orbital energy (LUMO)[35,59,64,65]. Electro-topological state indices account both sigma and valence electrons in the H-depleted molecular graph. They are not purely electronic descriptors, but describe atom polarity and steric accessibility[83]. Therefore, they show the possibility of interactions with other molecules. MEDV descriptors are modified electro-topological state indices and represent relative electronegativities of the molecule. HOMO energy is related to the ionisation potential and describes nucleophilicity of a molecule[72]. LUMO energy is related to the electron affinity and describes electrophilicity of a molecule[72]. Both of these descriptors give information about the molecule's reactivity or stability.

Mechanistically, metabolism is the most important cause of the reduced bioaccumulation potential of chemicals. Computationally, it is difficult to describe the mitigating effects of metabolism. Unfortunately, there is little experimental data for fish metabolism. Dimitrov and others[30,31] used an approximation for fish metabolism. They accounted effects of the metabolism by using information from the tissue metabolism simulator. The latter creates metabolic maps based on a library of experimental data of rat liver.

3.5. Methods used in this work

3.5.1. The best multi-linear regression

Searching for the best MLR QSAR model is not an easy task. The large initial descriptor pool makes it even harder. For example, let us say we have a descriptor pool of 1,000 descriptors and we want to find the best five-parameter model. Different combinations to choose from in this case are about $8*10^{12}$. Calculation of all the combinations is time-consuming and therefore, not practical[72]. Using BMLR algorithm, one can significantly reduce the search space. BMLR algorithm rejects all the descriptors with missing values or insignificant variance (less than $1*10^{-6}$). From highly correlated ($R^2 > 0.6$) descriptors, one is rejected. In the next step, all orthogonal pairs of descriptors from the remaining descriptor pool are found. With all orthogonal pairs of descriptors, two-parameter regression models are built and 400 models with the highest R^2 are selected. For each previously selected model, new descriptors are added and three-parameter regression models are built. If the R^2 of three-parameter models is lower than in the best two-parameter model, the latter is chosen as the final model. Otherwise, the best 400 models with the highest R^2 are selected and another descriptor is added. This step is repeated until the R^2 value is starting to decrease. A final model with the highest R^2 value is considered as the best representation of the activity in the given descriptor pool[72].

3.5.2. Random Forest

The Random Forest algorithm[82] builds on an ensemble of decision trees (Figure 2) and can be used in regression and classification analysis. The basic idea of RF is simple and can be summarised as follows.

In the training phase, the RF algorithm grows *N* decision trees that form a forest. Each tree is grown using a random sample from the initial data set. By default, RF draws from the training set of *n* chemicals a bootstrap sample, which typically contains 2/3 of the chemicals from the initial data set. The left out sample (1/3 of the chemicals) is called out-of-bag (OOB) sample. The best descriptor for the sample splitting is selected from a randomly selected subset of descriptors (m_{try}). The default value of m_{try} for the regression analysis is *p/3* of the descriptors. For the classification analysis, it is equal to the square root of *p*, where *p* is the total number of descriptors. Each tree is grown until no more splits are possible. Predictions of the regression models are calculated by averaging the predictions from all trees. The majority of the votes of all trees is used to make the final decision for the classification models. OOB sample is used to get the estimated error of the model. At first, all OOB chemicals are predicted using the grown tree. All the predictions are kept and the majority or average of those predictions is compared with real values to get the estimated error.

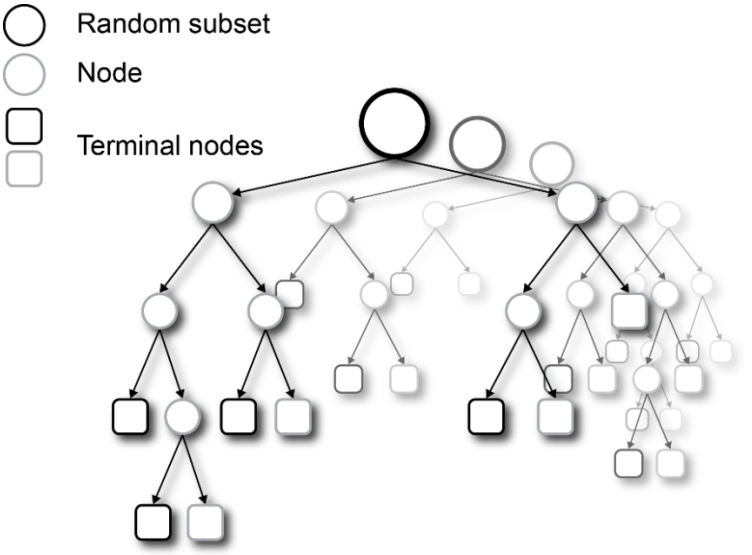

Figure 2. General structure of random forest

3.5.3. Clustering algorithms

Automatic construction of structurally similar subsets of chemicals requires the selection of a clustering algorithm and a decision how to represent molecular structures. One can use either molecular descriptors or chemical fingerprints[85]. The main idea behind clustering is to find common patterns from the given data and group similar chemicals together. Although the idea behind various clustering algorithms is the same, the way of achieving it differs. Therefore, clusters produced with different algorithms are different, even though they use the same data as a starting point. There are many algorithms out there and reviewing all of them would be an enormous task. Many of them are reviewed elsewhere[86–88], but here we focus on three of them (K-means algorithm, EM-algorithm and hierarchical clustering algorithm). These were used in Article II.

K-means algorithm[89] starts by selecting a predefined number (K) of random centres for subsets. Then each chemical is assigned to the nearest subset centre by the Euclidean distance. This is followed by the recalculation of the new centres for the created subsets, and the regrouping of chemicals based on the new subset centres. This is repeated until chemicals do not change subset memberships.

Expectation-maximisation (EM) clustering algorithm was first described by Dempster and others[90] and an excellent outline of it is given by Mitchell[91]. The EM-algorithm uses a similar iterative process like k-means algorithm. Instead of random subset centres, it uses random normal distributions. At each step, subset membership probabilities are estimated for each chemical. This is followed by the recalculation of distribution parameters in order to maximize the likelihood of the model to fit the given data. The process is repeated until the convergence criterion is fulfilled.

Hierarchical clustering is either agglomerative or divisive. The latter takes the whole data set as a single cluster and splits the data set into smaller clusters. The agglomerative hierarchical clustering algorithm takes a whole data set and presents each chemical as an individual subset. The next step is to group the most similar subsets together based on a distance function and a linkage criterion. This is repeated until the desired number of subsets is obtained. From those two methods, divisive hierarchical clustering is more time-consuming and is rarely used. In Article II, agglomerative hierarchical clustering was used.

3.5.4. Different applicability domain approaches

Building a QSAR model is just a first step to get a good prediction for the modelled property. In an ideal world, model should be able to predict activities for any chemicals. In reality, models have limitations. Therefore, it is utmost important to define an applicability domain (AD) for a QSAR model to minimise the risks of wrong predictions. AD needs some predefined boundaries and violation of those boundaries is a sign that the model might fail to predict

such chemical. A review of different AD methods used in QSAR modelling is given by Sahigara and co-workers[92]. Approaches relevant for this work receive individual attention henceforward.

The first approach uses descriptor ranges. It is based on the maximum and minimum values of the descriptors from the training set. The acceptable difference for the descriptors is set as 15% of the descriptor range. Meaning that descriptor value used in the prediction can exceed the maximum and minimum value of the descriptor by a predefined amount. If a chemical under investigation has at least one descriptor out of bounds it is considered as out of AD and the model might give unreliable prediction.

The second approach is called leverage and it is suggested to be used for identifying chemicals out of AD[93]. The leverage (h) is calculated according to equation 13, where x is the vector of descriptors of a predicted chemical and X is the matrix of the descriptors in the training set. If a predicted chemical has the leverage value higher than the critical leverage value (h', equation 14) it is considered out of AD and the prediction might be unreliable. In equation 14, K is the number of descriptors in the model and N is the size of the training set.

$$h = x^T(X^TX)^{-1}x \tag{13}$$

$$h' > \frac{3(K+1)}{N} \tag{14}$$

The third approach is RF specific. The concept of this approach was developed during the preparation of Article III. It uses proximity matrix that is provided by RF algorithm. Proximity matrix is a square matrix with proximity values for each pair of chemicals. Proximity is calculated as follows. At the beginning, each pair of chemicals has proximity value equal to zero. Every time two chemicals end up at the same terminal node, one is added to the proximity value. For the proximity value, the final sum is divided by the number of trees in the model. The proximity value illustrates the pairwise chemical similarity based on the descriptor values. The closer the proximity value is to one, the more similar two chemicals are. For the definition of AD, two user-defined rules are used. The first rule is a proximity cut-off value (for example 0.7), which finds all the chemicals from the training set that exceed this value. The second rule defines how many chemicals in the training set must have proximity value equal to or greater than the cut-off value. Applying these rules helps to identify the composition of the training set compared to the chemical under investigation. If there are not enough similar chemicals in the training set, the prediction might be unreliable. In addition, if most of the chemicals in the training set are predicted incorrectly, there is a high probability that the prediction is also wrong.

4. SUMMARY OF THE ORIGINAL ARTICLES

4.1. QSAR model for the prediction of bio-concentration factor using aqueous solubility and descriptors considering various electronic effects

Article I describes the development of a global QSAR model for predicting BCF. For the modelling, two data sets were compiled. The first data set consists of 473 chemicals and was divided into a training and a validation set. The second data set was compiled from different sources and consists of 161 chemicals. It was used as a second external validation set. Using the best multi-linear regression approach, the following relationship was obtained:

$$\text{Log BCF} = -0.420 \times \log S - 0.005 \times \text{WPSA-1} + 0.252 \times \text{LUMO} + 0.004 \times \text{THSA} - 0.096 \times \min(\#\text{HA},\#\text{HD}) - 0.405$$

Set	N_{comp}	R^2	R^{2*}	$R^2_{(LOO)}$	$R^2_{(LMO)}$	s	s^*
Training	310	0.75	0.78	0.74	0.74	0.67	0.63
Validation 1	156	0.62	0.73	–	–	0.83	0.68
Validation 2	161	0.46	0.67	–	–	1.03	0.78
All together	627	0.64	–	–	–	0.81	–

*value when outliers with error greater than 2.5 deviations excluded

The model above contains five descriptors, which all have a mechanistic interpretation regarding BCF. *Log S* is aqueous solubility and it shows that chemicals with lower solubility in water have higher log BCF value. It is logical that if chemicals have problems dissolving in water, they are more likely adsorbed by various surfaces, like cell membranes and soil. Surface-weighted charged partial positive surface area (WPSA-1) considers the molecule's ability to penetrate or interact with the cell membrane. The energy of the lowest unoccupied molecular orbital (LUMO) describes molecules' stability and identifies strong and weak electrophiles. It also describes the chemical's ability to interact with the cell membrane. Total hydrophobic surface area (THSA) considers hydrophobic interactions that are relevant for membrane penetration. Finally, min(#HA,#HD) counts the minimum number of hydrogen bond acceptor or hydrogen bond donor sites. It describes the chemical's ability to stay in the aqueous phase, which reduces chemical's cell penetration capabilities and therefore lowers BCF.

A thorough analysis of the outliers pointed out certain chemical features that have higher uncertainty in predictions. Complementary to the outliers analysis, the applicability domain of the model was analysed by descriptor range and leverage approach. AD analysis alone could not identify most of the outliers.

Therefore, it is important to use outliers' analysis information together with applicability domain analysis. Altogether, the proposed model is in accordance with OECD principles for regulatory use and considering all the information available it gives the user a higher confidence to use the predictions in risk assessment.

4.2. Comparative analysis of local and consensus quantitative structure activity relationship approaches for the prediction of bioconcentration factor

Article II examines the possibilities for improving BCF predictions with the help of local and consensus models. Three clustering algorithms and seven different sets of descriptors helped to construct homogeneous training sets for modelling. In addition, two rule-based approaches were used for training set designs. Consensus models were created by averaging predictions from multiple models. All results were compared to the predictions from the global model developed in Article I.

The clustering algorithms used were EM algorithm, k-means algorithm and hierarchical clustering algorithm. A comparison between algorithms showed that models developed with subsets produced by EM and k-means algorithms have slightly better statistical parameters than models based on the subsets from hierarchical clustering. Nevertheless, in most of the cases, statistical parameters improved compared to the global model. The same trend was there in the case of rule-based subset creation. To be exact, the best set of local models was achieved with EM algorithm. Compared to the global model, R^2 for all the chemicals improved from 0.64 to 0.77. The best subset of rule-based models was based on the atomic constitution of chemicals. In this case, R^2 improved to 0.79. Some of the models in this subset had overlapping content and, therefore, some predictions were average predictions from multiple models. The usage of multiple predictions makes this a consensus model. Superior results compared to other local models indulged us to create other consensus models. Using consensus (super-consensus model) from all the available predictions gave the best statistical performance compared to all the other models. R^2 in super-consensus model improved up to 0.81.

111 local QSAR models were produced to predict BCF and altogether 164 different descriptors were used in those models. Comprehensive analysis of those descriptors was not feasible, but certain trends emerged. The most frequent descriptors were *log S* and *log P*, which play an important role in cell penetration. Other descriptors mostly describe mitigating factors such as molecular size, branching, flexibility, hydrogen bonding and ionisation.

Detailed analysis of local models showed that some models had rather poor statistical parameters. This information can be used to reveal subsets of chemicals that are difficult to model. In addition, compared to the global model, most of the outliers were the same, but local models introduced many new

outliers. Using consensus models, however, a single wrong prediction usually gets a smaller weight than multiple correct predictions. Therefore, most of the wrong predictions were evened out. The super-consensus model could not correctly predict ten chemicals. The same chemicals had wrong predictions also with the global model. Therefore, using local models together with consensus model can identify inaccurate data or uniquely behaving chemicals by recording chemicals that have wrong predictions in most of the models.

4.3. Classifying bio-concentration factor with random forest algorithm, influence of the bio-accumulative *vs.* non-bio-accumulative compound ratio to modelling result, and applicability domain for Random Forest model

Article I and Article II sought ways to improve BCF predictions with regression modelling. However, we do not always need an exact value of BCF. Sometimes we only need to know whether the chemical is bio-accumulative or not. Most of the BCF data sets have a normal distribution, but dividing those chemicals into bio-accumulative (B) and non-bio-accumulative (nB) chemicals results in data sets with more nB-chemicals. Such imbalanced data sets can influence classification results towards the majority class, making the building of classification models more difficult.

Article III focuses on Random Forest based classification models for classification of the BCF. It studies the influence of the data set class balance on the models. Classification models were built on a large data set of 1,007 chemicals. The balance of the data set had a ratio of four to one towards nB-chemicals. Three class distribution schemas were studied and each of them chose only a certain number of chemicals from both classes as an input sample. As a result, the training sets for modelling had three configurations. The first was imbalanced towards nB-chemicals, the second was balanced and the third was imbalanced towards the B-chemicals.

The imbalanced model towards nB-chemicals correctly classified around 85% of the chemicals. However, the classification of nB-chemicals was more precise than the classification of B-chemicals. For example, the misclassification of B-chemicals was 33% compared to the misclassification of only 9% in the case of nB-chemicals. Considering the environmental risk of inaccurate prediction of both classes, misclassifying B-chemicals is obviously more severe. The balanced model showed similar overall accuracy than the previous model, but it could predict B-chemicals more precisely than nB-chemicals. Ideally, the model should predict both classes equally well, but considering the environmental risk, it is more important to classify B-chemicals correctly. Tilting the balance towards the B-chemicals resulted in a model that could predict B-chemicals nearly perfectly (only one chemical was misclassified). However, 33% of the nB-chemicals were misclassified. Despite the high misclassification rate for nB-chemicals, this model could securely identify 67%

of the nB-compounds, because all of the chemicals predicted as nB-chemicals were nB-chemicals.

All the models are easy to use and consist only of three descriptors. All used descriptors are based on 2D-structures, making the reproducibility of descriptor values trivial. At the same time, all the descriptors were calculated using freely distributed software and were consistent with mechanistic interpretation of BCF. The most important descriptor in all the models was *log P*, which was complemented with descriptors that describe the size and branching of the chemical, polarisability/electrostatic effects and hydrogen-bonding ability of the chemicals.

Additionally, the new applicability domain approach for Random Forest based QSAR models was proposed. This proximity-based AD approach helps to assess the similarity of the target chemical compared to the chemicals in the training set. In this way, one can easily extract similar compounds from the training set and by analysing classification results for those chemicals, make it easier to estimate how trustworthy the prediction for the target chemical is. The more information about the behaviour of the chemical one has, the more confident one can be in decision-making. Therefore, by using different models together with AD approach, one can more easily judge the correctness of the decision.

5. SUMMARY

Bioconcentration is an important endpoint for the determination of the fate and behaviour of chemicals in the environment. One area where BCF is extensively used is environmental risk assessment. However, experimental measurement of BCF for one chemical can take up to six months, cost around 100,000 euros, and need about one hundred animals. Therefore, thousands of chemicals are not being experimentally measured. This creates the need for the development of faster and more economical QSAR models to predict BCF for chemicals with no experimental data. To fill the gaps, many theoretical models have been developed. Wide chemical space makes it hard to use one universal model for all the chemicals. Therefore, at risk assessment, applicability of the chosen model is assessed for each chemical. On top of that, for more reliable results, multiple models are used.

The goal of this thesis is to provide an outline for risk assessment procedures, bioconcentration factor and different QSAR methodologies. The modelling part of the thesis is divided into two. The first part focuses on the regression analysis and the second part on the classification problems. At first, a global regression model was proposed for predicting BCF. The global model could predict a wide variety of chemicals and provide information about the model's applicability domain. The creation of the global model laid the foundation for the exploration of the possibilities to improve prediction quality using smaller, more focused data sets. Most of the subsets of focused models showed better predictive power compared to the global model. Additionally, consensus model was compared against the global model and local models. Proposed consensus model outperformed all of them. To separate bio-accumulative and non-bio-accumulative chemicals three classification models with different training set compositions were proposed. All three developed models had their strengths in different classification scenarios, but the most all-purpose model was the model where classes were distributed evenly. To identify whether a chemical fits into the boundaries of the model, a new approach was proposed for assigning applicability domain for Random Forest based models. Applying AD shows how many similar chemicals were used to develop the model and how well they were predicted. The information provided by the AD schema allows making a more confident final decision about the correctness of the prediction.

Building a QSAR model is not a trivial task. The purpose of the model declares which aspects should receive special attention. For risk assessment, it is important to use relevant endpoints and unambiguous algorithms. All the models built during this work use well-defined algorithms and BCF as an endpoint. Attention was paid to the requirement of model validation and defined applicability domain. In addition, all the used descriptors have a sound mechanistic interpretation in relation to BCF. Therefore, all of these models can be used in environmental risk assessment to get additional information about the bioaccumulation potential of chemicals.

6. SUMMARY IN ESTONIAN

Kemikaalide keskkonnariskide hindamine QSAR meetoditega

Kemikaalide keskkonnariskide hindamisel uuritakse mitmeid omadusi, mida kemikaal võib mõjutada. Erilise tähelepanu all on kemikaali käitumine keskkonnas. Üheks uuritavaks omaduseks selles valdkonnas on biokontsentratsioon (BCF) ehk kui suures ulatuses kemikaal võib ladestuda organismi. Olemasolevate eksperimentaalsete mõõtmismeetodite rakendamist kõikidele kemikaalidele takistavad aeg, raha ja loomkatsed. Näiteks ühe kemikaali eksperimentaalne määramisel kasutatakse üle saja kala, see võtab aega kuni kuus kuud ning maksab umbes 100 000 eurot. Seetõttu puudub eksperimentaalselt määratud väärtus tuhandetel kemikaalidel. Viimane innustab looma kiiremad ning odavamad QSAR mudeleid, et täita tühimikud andmekogudes, kus kemikaalidel puuduvad eksperimentaalselt mõõdetud väärtused. On loodud mitmeid mudeleid, kuid ükski neist pole universaalne, ega rakendatav kõikidele kemikaalidele. Seetõttu kasutatakse riskianalüüsis erinevatele kemikaalidele omaduse väärtuse ennustamisel erinevaid mudeleid. Üldiselt kasutatakse ennustamisel mitmeid mudeleid, kus lõplik otsus tehakse mitme mudeli keskmisest ennustusest.

Käesolev dissertatsioon on ülesehitatud põhimõttel anda ülevaade riskianalüüsi etappidest, biokontsentratsiooni tegurist ning erinevatest QSAR analüüsi meetoditest. Eksperimentaalne osa keskendub kirjeldatud meetodite rakendamisele BCF-i modelleerimisel. Modelleerimisel kasutati kahte lähenemist. Esimeses keskenduti regressiooni mudelite loomisele ning teises tegeleti klassifitseerimisprobleemidega. Alustuseks loodi laia kemikaalide spektrit kattev globaalne mudel BCF-i ennustamiseks. Mudelile määratud rakenduspiirkond võimaldas otsustada, millistele kemikaalidele tehtud ennustused olid usaldusväärsed ning millised mitte. Ennustustäpsuse parandamiseks, tükeldati andmekomplekt rohkem fokusseeritud alamkomplektideks ning loodi lokaalsed mudelid. Võrreldes globaalse mudeliga suutsid enamus lokaalsete mudelite komplekte ennustada BCF-i täpsemalt. Veel kõrgem ennustustäpsus saadi kui kõik loodud lokaalsed mudelid ühendati üheks suureks konsensusmudeliks. Bioakumuleeruvate kemikaalide eraldamiseks mitte bioakumuleeruvatest koostati kolm erineva klassijaotusega klassifitseerimismudelit. Loodud kolm mudelit on rakendatavad erinevatel juhtudel, kuid kõige üldisemaks võib pidada mudelit, kus treenimise faasis oli klasside jaotus võrdne. Lisaks pakuti välja uus rakenduspiirkonna määramise meetod Random Forest'i mudelite jaoks. Viimane näitab kui palju sarnaseid kemikaale kasutati mudeli loomisel ning kui hästi langesid ennustused kokku eksperimentaalsete andmetega. Kasutades seda informatsiooni on võimalik hinnata kui täpne on ennustus tundmatule kemikaalile.

QSAR mudeli loomine ei ole lihtne ülesanne, sest sõltuvalt mudeli eesmärgist tuleb arvestada mitmete nõuetega. Riskianalüüsis on oluline kasutada asjakohaseid omadusi ja ühemõttelisi algoritme. Kõik käesolevas töös loodud mudelid kasutavad omadusena BCF-i ja täpselt defineeritud algoritme. Samuti pöörati tähelepanu mudelite valideerimisele ning rakenduspiirkonna määramisele, mis on tähtsad nõuded riskianalüüsis. Lisaks omasid kõik kasutatud deskriptorid modelleeritud omaduse suhtes mehhanistlikku seletust. Seetõttu annavad kõik loodud mudelid oma panuse riskianalüüsi, seletamaks kemikaalide võimet bioakumuleeruda.

REFERENCES

(1) Shea, D. In *A Textbook of Modern Toxicology*; John Wiley & Sons, Inc., 2004; pp. 501–517.
(2) Suter II, G. W. *Ecological risk assessment, Second Edition*; CRC press, 2006.
(3) Benfenati, E.; Azimonti, G.; Auteri, D.; Lodis, M. In *Computational Toxicology*; John Wiley & Sons, Inc., 2006; pp. 625–650.
(4) Leeuwen, C. J. V. In *Risk Assessment of Chemicals*; Leeuwen, C. J. va.; Vermeire, T. G., Eds.; Springer Netherlands, 2007; pp. 1–36.
(5) *Test No. 305: Bioaccumulation in Fish: Aqueous and Dietary Exposure*; lbraceOECDrbrace Publishing, 2012.
(6) Parliament, E. Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) establishing a European Chemicals Agency, amending Directive 1999/45/EC and repealing Council Regulation (EEC) No 793/93 and Commission Regulation (EC) No 1488/94 as well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC, 2006.
(7) OECD. Guidance document on the validation of (quantitative)structure-activity relationships [(q)sar] models, 2007.
(8) Health, I. for; Production, C.; Centre, E. C. J. R.; Bureau, E. C. *Technical Guidance Document on Risk Assessment*; Commission Directive; Office for Official Publications of the European Communities, 2003.
(9) In *Chemicals in the Environment: Assessing and Managing Risk*; Hester, R. E.; Harrison, R. M., Eds.; The Royal Society of Chemistry, 2006; Vol. 22, pp. 132–153.
(10) Forum, U. S. E. P. A. R. A. *Guidelines for Ecological Risk Assessment*; The Agency, 1998.
(11) Forum, U. S. E. P. A. R. A. *Generic Ecological Assessment Endpoints (GEAEs) for Ecological Risk Assessment*; Risk Assessment Forum, U.S. Environmental Protection Agency, 2003.
(12) Maltby, L. In *Chemicals in the Environment: Assessing and Managing Risk*; Hester, R. E.; Harrison, R. M., Eds.; The Royal Society of Chemistry, 2006; Vol. 22, pp. 84–101.
(13) Scholz, S.; Sela, E.; Blaha, L.; Braunbeck, T.; Galay-Burgos, M.; Garcıa-Franco, M.; Guinea, J.; Klüver, N.; Schirmer, K.; Tanneberger, K.; Tobor-Kaplon, M.; Witters, H.; Belanger, S.; Benfenati, E.; Creton, S.; Cronin, M. T. D.; Eggen, R. I. L.; Embry, M.; Ekman, D.; Gourmelon, A.; Halder, M.; Hardy, B.; Hartung, T.; Hubesch, B.; Jungmann, D.; Lampi, M. A.; Lee, L.; Léonard, M.; Küster, E.; Lillicrap, A.; Luckenbach, T.; Murk, A. J.; Navas, J. M.; Peijnenburg, W.; Repetto, G.; Salinas, E.; Schüürmann, G.; Spielmann, H.; Tollefsen, K. E.; Walter-Rohde, S.; Whale, G.; Wheeler, J. R.; Winter, M. J. *Regulatory Toxicology and Pharmacology* **2013**, *67*, 506–530.
(14) Netzeva, T. I.; Pavan, M.; Worth, A. P. *QSAR & Combinatorial Science* **2008**, *27*, 77–90.
(15) Rücker, C.; Kümmerer, K. *Green Chem.* **2012**, *14*, 875.
(16) Doucette, W. J. *Environmental Toxicology and Chemistry* **2003**, *22*, 1771–1788.

(17) Pavan, M.; Worth, A. P.; Netzeva, T. I. *European Commission Joint Research centre (EUR 22327 EN)* **2006**.
(18) Sijm, D.; Rikken, M.; Rorije, E.; Traas, T.; McLachlan, M.; Peijnenburg, W. In *Risk Assessment of Chemicals*; Springer, 2007; pp. 73–158.
(19) Barron, M. G. *Environmental science & technology* **1990**, *24*, 1612–1618.
(20) Barron, M. In *Handbook of Ecotoxicology, Second Edition*; CRC Press, 2002.
(21) Baynes, R. E.; Hodgson, E. In *A Textbook of Modern Toxicology*; John Wiley & Sons, Inc., 2004; pp. 75–110.
(22) O'Flaherty, E. J. In *Principles of Toxicology*; John Wiley & Sons, Inc., 2000; pp. 35–55.
(23) Kleinow, K.; Nichols, J.; Hayton, W.; McKim, J.; Barron, M. In *The Toxicology of Fishes*; CRC Press, 2008; pp. 55–152.
(24) Yee, L. C.; Wei, Y. C. In *Statistical Modelling of Molecular Descriptors in QSAR/QSPR*; Wiley-VCH Verlag GmbH & Co. KGaA, 2012; pp. 1–31.
(25) Dudek, A.; Arodz, T.; Galvez, J. *Combinatorial Chemistry & High Throughput Screening* **2006**, *9*, 213–228.
(26) Bermúdez-Saldaña, J. M.; Escuder-Gilabert, L.; Medina-Hernández, M. J.; Villanueva-Camañas, R. M.; Sagrado, S. *Journal of Chromatography A* **2005**, *1063*, 153–160.
(27) Bintein, S.; Devillers, J.; Karcher, W. *SAR and QSAR in Environmental Research* **1993**, *1*, 29–39.
(28) Dearden, J.; Hewitt, M. *SAR and QSAR in Environmental Research* **2010**, *21*, 671–680.
(29) Dearden, J.; Shinnawei, N. *SAR and QSAR in Environmental Research* **2004**, *15*, 449–455.
(30) Dimitrov, S.; Dimitrova, N.; Georgieva, D.; Vasilev, K.; Hatfield, T.; Straka, J.; Mekenyan, O. *SAR and QSAR in Environmental Research* **2012**, *23*, 17–36.
(31) Dimitrov, S.; Dimitrova, N.; Parkerton, T.; Comber, M.; Bonnell, M.; Mekenyan, O. *SAR and QSAR in Environmental Research* **2005**, *16*, 531–554.
(32) Dimitrov, S.; Mekenyan, O.; Walker, J. *SAR and QSAR in Environmental Research* **2002**, *13*, 177–184.
(33) Dimitrov, S. D.; Dimitrova, N. C.; Walker, J. D.; Veith, G. D.; Mekenyan, O. G. *QSAR & Combinatorial Science* **2003**, *22*, 58–68.
(34) Dimitrov, S. D.; Dimitrova, N. C.; Walker, J. D.; Veith, G. D.; Mekenyan, O. G. *Pure and Applied Chemistry* **2002**, *74*.
(35) Fatemi, M.; Jalali-Heravi, M.; Konuze, E. *Analytica chimica acta* **2003**, *486*, 101–108.
(36) Ferreira, M. *Chemosphere* **2001**, *44*, 125–146.
(37) Fujikawa, M.; Nakao, K.; Shimizu, R.; Akamatsu, M. *Chemosphere* **2009**, *74*, 751–757.
(38) Garg, R.; Smith, C. J. *Food and Chemical Toxicology* **2014**, *69*, 252–259.
(39) Gissi, A.; Gadaleta, D.; Floris, M.; Olla, S.; Carotti, A.; Novellino, E.; Benfenati, E.; Nicolotti, O. *Altex* **2014**, *31*, 23–36.
(40) Gramatica, P.; Papa, E. *QSAR & Combinatorial Science* **2005**, *24*, 953–960.
(41) Gramatica, P.; Papa, E. *QSAR & Combinatorial Science* **2003**, *22*, 374–385.
(42) Isnard, P.; Lambert, S. *Chemosphere* **1988**, *17*, 21–34.
(43) Ivanciuc, T.; Ivanciuc, O.; Klein, D. J. *Molecular diversity* **2006**, *10*, 133–145.
(44) Ivanciuc, T.; Ivanciuc, O.; Klein, D. J. *Journal of chemical information and modeling* **2005**, *45*, 870–879.

(45) Jackson, S. H.; Cowan-Ellsberry, C. E.; Thomas, G. *Journal of agricultural and food chemistry* **2009**, *57*, 958–967.
(46) Khadikar, P. V.; Singh, S.; Mandloi, D.; Joshi, S.; Bajaj, A. V. *Bioorganic & medicinal chemistry* **2003**, *11*, 5045–5050.
(47) Kumar, S.; Kumar, M.; Thurow, K.; Stoll, R.; Kragl, U. *Environmental Modelling & Software* **2009**, *24*, 44–53.
(48) Lin, K.-H.; Jaw, C.-G.; Yen, J.-H.; Wang, Y.-S. *Ecotoxicology and environmental safety* **2009**, *72*, 1942–1949.
(49) Liu, H.; Yao, X.; Zhang, R.; Liu, M.; Hu, Z.; Fan, B. *Chemosphere* **2006**, *63*, 722–733.
(50) Liu, S.-S.; Qin, L.-T.; Liu, H.-L.; Yin, D.-Q. *Journal of molecular modeling* **2008**, *14*, 83–92.
(51) Lombardo, A.; Roncaglioni, A.; Boriani, E.; Milan, C.; Benfenati, E. *Chemistry Central Journal* **2010**, *4*, S1.
(52) Lu, X.; Tao, S.; Cao, J.; Dawson, R. *Chemosphere* **1999**, *39*, 987–999.
(53) Lu, X.; Tao, S.; Hu, H.; Dawson, R. *Chemosphere* **2000**, *41*, 1675–1688.
(54) Mackay, D. *Environmental Science & Technology* **1982**, *16*, 274–278.
(55) Mansouri, K.; Consonni, V.; Durjava, M. K.; Kolar, B.; Öberg, T.; Todeschini, R. *Chemosphere* **2012**, *89*, 433–444.
(56) Melo, E. B. de. *Ecotoxicology and environmental safety* **2012**, *75*, 213–222.
(57) Meylan, W. M.; Howard, P. H.; Boethling, R. S.; Aronson, D.; Printup, H.; Gouchie, S. *Environmental Toxicology and Chemistry* **1999**, *18*, 664–672.
(58) Papa, E.; Dearden, J.; Gramatica, P. *Chemosphere* **2007**, *67*, 351–358.
(59) Peng, S.; Jian-Wei, Z.; Peng, Z.; Lin, X. *Chemosphere* **2011**, *83*, 1045–1052.
(60) Pramanik, S.; Roy, K. *Environmental Science and Pollution Research* **2014**, *21*, 2955–2965.
(61) Qin, L.-T.; Liu, S.-S.; Liu, H.-L. *Molecular diversity* **2010**, *14*, 67–80.
(62) Qin, L.-T.; Liu, S.-S.; Liu, H.-L.; Ge, H.-L. *Chemosphere* **2008**, *70*, 1577–1587.
(63) Roy, K.; Sanyal, I.; Roy, P. *SAR and QSAR in Environmental Research* **2006**, *17*, 563–582.
(64) Sahu, V. K.; Singh, R. K. *CLEAN–Soil, Air, Water* **2009**, *37*, 850–857.
(65) Saçan, M. T.; Erdem, S. S.; Özpinar, G. A.; Balcioglu, I. A. *Journal of chemical information and computer sciences* **2004**, *44*, 985–992.
(66) Tao, S.; Hu, H.; Xu, F.; Dawson, R.; Li, B.; Cao, J. *Journal of Environmental Science and Health, Part B* **2001**, *36*, 631–649.
(67) Toropov, A.; Toropova, A.; Benfenati, E. *European journal of medicinal chemistry* **2009**, *44*, 2544–2551.
(68) Toropova, A.; Toropov, A.; Lombardo, A.; Roncaglioni, A.; Benfenati, E.; Gini, G. *European journal of medicinal chemistry* **2010**, *45*, 4399–4402.
(69) Wang, Y.; Li, Y.; Ding, J.; Jiang, Z.; Chang, Y. *SAR and QSAR in Environmental Research* **2008**, *19*, 375–395.
(70) Zarei, K.; Salehabadi, Z. *Structural Chemistry* **2012**, *23*, 1801–1807.
(71) Zhao, C.; Boriani, E.; Chana, A.; Roncaglioni, A.; Benfenati, E. *Chemosphere* **2008**, *73*, 1701–1707.
(72) Karelson, M. *Molecular descriptors in QSAR/QSPR*; Wiley-Interscience; Wiley-Interscience, 2000.
(73) Gramatica, P. *QSAR & combinatorial science* **2007**, *26*, 694–701.
(74) Nendza, M.; Herbst, T. *SAR and QSAR in Environmental Research* **2011**, *22*, 351–364.

(75) Nendza, M.; Müller, M. *SAR and QSAR in Environmental Research* **2010**, *21*, 495–512.
(76) Strempel, S.; Nendza, M.; Scheringer, M.; Hungerbühler, K. *Environmental Toxicology and Chemistry* **2013**, *32*, 1187–1195.
(77) Sun, X.; Li, Y.; Liu, X.; Ding, J.; Wang, Y.; Shen, H.; Chang, Y. *Molecular diversity* **2008**, *12*, 157–169.
(78) Fernández, A.; Lombardo, A.; Rallo, R.; Roncaglioni, A.; Giralt, F.; Benfenati, E. *Environment international* **2012**, *45*, 51–58.
(79) Liu, P.; Long, W. *International journal of molecular sciences* **2009**, *10*, 1978–1998.
(80) Wold, S.; Sjöström, M.; Eriksson, L. *Chemometrics and intelligent laboratory systems* **2001**, *58*, 109–130.
(81) Vapnik, V. N. *Statistical learning theory*; Wiley New York, 1998; Vol. 2.
(82) Breiman, L. *Machine learning* **2001**, *45*, 5–32.
(83) Todeschini, R.; Consonni, V.; Mannhold, R.; Kubinyi, H.; Folkers, G. *Molecular Descriptors for Chemoinformatics*; Methods and Principles in Medicinal Chemistry; Wiley, 2009.
(84) *Handbook of Molecular Descriptors*; Todeschini, R.; Consonni, V., Eds.; Wiley-lbraceVCHrbrace Verlag lbraceGmbHrbrace, 2000.
(85) Daylight Fingerprints.
(86) Jain, A. K. *Pattern Recognition Letters* **2010**, *31*, 651–666.
(87) Jain, A. K.; Murty, M. N.; Flynn, P. J. *ACM computing surveys (CSUR)* **1999**, *31*, 264–323.
(88) Xu, R.; Wunsch, D.; others. *Neural Networks, IEEE Transactions on* **2005**, *16*, 645–678.
(89) Lloyd, S. *Information Theory, IEEE Transactions on* **1982**, *28*, 129–137.
(90) Dempster, A. P.; Laird, N. M.; Rubin, D. B. *Journal of the Royal Statistical Society. Series B (Methodological)* **1977**, 1–38.
(91) Mitchell, T. M. *Machine Learning*; 1st ed.; McGraw-Hill, Inc.: New York, NY, USA, 1997.
(92) Sahigara, F.; Mansouri, K.; Ballabio, D.; Mauri, A.; Consonni, V.; Todeschini, R. *Molecules* **2012**, *17*, 4791–4810.
(93) Tropsha, A.; Gramatica, P.; Gombar, V. K. *QSAR & Combinatorial Science* **2003**, *22*, 69–77.

ACKNOWLEDGEMENTS

First, I would like to thank my supervisor Dr. Sulev Sild who provided helpful recommendations and guidance during my studies and research. I would also like to thank Dr. Uko Maran for sharing his knowledge, which often made me look at things from different angle. I am grateful to all my colleagues and each and every one whose contribution helped me on this journey. For bring back to the memory the long lost word "*the*", I would like to thank my good friend Marit. Without her linguistic advice, this thesis had good chance to be one sentence long.

My biggest gratitude goes to my parents whose support made my studies possible. My studies would not been the same without friendly mocking between my brother and sister. I would also thank my dear Marit for her understanding and support during all these years. Her assurance that renovating the kitchen and writing the thesis at the same time is a good idea makes sense now. Immense thank you goes to my son Lennar, who made it clear that he needs a new bedtime story and I should write it. Thank you again, here it is. I also want to thank all my friends for rewarding discussions about everything. Unfortunately, we have not figured out the theory of everything, but we still have time. Few friends I would like to mention are Martti, Erkki, Veiko, Argo, Kert, Alar, Markus, Kalev, Maikki, Mare, Kadriann, Anneli, Anna and Karl. Special place in my heart is reserved for Guido, Illar, Aivar and Alexander. Without their impossible and sometimes hilarious ideas, many roads have been unexplored.

This work was supported by the Estonian Science Foundation (grants 7153 and 7709), Estonian Ministry for Education and Research (SF0140031Bs09), EU 6th FP project Chemomentum (IST-033437) and European Union Regional Development Fund (grant number 3.2.1201.13-0021). This work has been partially supported by graduate school „Functional materials and technologies" receiving funding from the European Social Fund under project 1.2.0401.09-0079 in Estonia.

PUBLICATIONS

CURRICULUM VITAE

Name: Geven Piir
Born: 17.04.1982, Tartu, Estonia
E-mail geven.piir@ut.ee
Citizenship: Estonian

Education:

2009–present	Ph.D. student of Molecular Technology, University of Tartu
2007–2009	M.Sc. in Molecular Technology, University of Tartu
2000–2006	B.Sc. in Chemistry, University of Tartu
1988–2000	Ülenurme Gymnasium

Professional Employment:

2014–present	University of Tartu, Junior Researcher
2013–2014	University of Tartu, Chemist
2013–2013	National Institute of Chemistry Slovenia, Visiting Scientist
2007–2009	University of Tartu, Chemist-Specialist

Scientific publications:

1. Piir, G.; Sild, S.; Roncaglioni, A.; Benfenati, E.; Maran, U. QSAR model for the prediction of bio-concentration factor using aqueous solubility and descriptors considering various electronic effects. *SAR QSAR Environ. Res.*, **2010**, 21, 711–729.
2. Piir, G.; Sild, S.; Maran, U. Comparative analysis of local and consensus quantitative structure activity relationship approaches for the prediction of bioconcentration factor. *SAR QSAR Environ. Res.*, **2013**, 24, 175–199.
3. Piir, G.; Sild, S.; Maran, U. Classifying bio-concentration factor with random forest algorithm, influence of the bio-accumulative vs. non-bio-accumulative compound ratio to modelling result, and applicability domain for random forest model. *SAR QSAR Environ. Res.*, **2014**, 25, 967–981.

Oral presentations:

1. CMTPI 2011 (6th International Symposium on Computational Methods in Toxicology and Pharmacology Integrating Internet Resources), 3–7 September 2011, Maribor, Slovenia – „Local and Global QSAR Modelling Approaches for the Prediction of Bio-Concentration Factor"
2. QSAR 2014 (16th International Workshop on QSAR in Environmental and Health Sciences), 16–20 June 2014, Milan, Italy – "Classification models for the bio-concentration factor"

Poster presentations:

1. QSAR 2010 (14th International Workshop on Quantitative structure-activity relationships in Environmental and Health Sciences), 24–28. May 2010, Montréal, Canada – "QSAR model for the prediction of bio-concentration factor by water solubility as the major descriptor"
2. EURO QSAR 2010 (18th European Symposium on Quantitative Structure Activity Relationships), 19–24 September 2010, Rhodes, Greece – "The Effect of the 3D-geometry on Molecular Descriptors and QSAR Models"
3. Eesti XXXII Keemiapäevad, 14–15 April 2011, Tartu, Estonia – "Molekulide konformatsiooni mõju QSAR-mudeli reprodutseeritavusele"
4. QSAR 2014 (16th International Workshop on QSAR in Environmental and Health Sciences), 16–20 June 2014, Milan, Italy – "Reproducibility of QSAR/QSPR models in the scientific literature – perspective for regulatory use"

ELULOOKIRJELDUS

Nimi: Geven Piir
Sünniaeg: 17.04.1982, Tartu, Eesti
E-post geven.piir@ut.ee
Kodakondsus: Eesti

Haridus:

2009–tänaseni	Molekulaartehnoloogia doktorant, Tartu Ülikool
2007–2009	M.Sc. molekulaartehnoloogias, Tartu Ülikool
2000–2006	B.Sc. keemias, Tartu Ülikool
1988–2000	Ülenurme Gümnaasium

Töökogemus:

2014–tänaseni	Tartu Ülikool, nooremteadur
2013–2014	Tartu Ülikool, keemik
2013–2013	Sloveenia Rahvuslik Keemia Instituut, külalisteadur
2007–2009	Tartu Ülikool, keemik-Spetsialist

Teaduspublikatsioonid:

1. Piir, G.; Sild, S.; Roncaglioni, A.; Benfenati, E.; Maran, U. QSAR model for the prediction of bio-concentration factor using aqueous solubility and descriptors considering various electronic effects. *SAR QSAR Environ. Res.*, **2010**, 21, 711–729.
2. Piir, G.; Sild, S.; Maran, U. Comparative analysis of local and consensus quantitative structure activity relationship approaches for the prediction of bioconcentration factor. *SAR QSAR Environ. Res.*, **2013**, 24, 175–199.
3. Piir, G.; Sild, S.; Maran, U. Classifying bio-concentration factor with random forest algorithm, influence of the bio-accumulative vs. non-bio-accumulative compound ratio to modelling result, and applicability domain for random forest model. *SAR QSAR Environ. Res.*, **2014**, 25, 967–981.

Suulised ettekanded:

1. CMTPI 2011 (6th International Symposium on Computational Methods in Toxicology and Pharmacology Integrating Internet Resources), 3.–7. september 2011, Maribor, Sloveenia – "Local and Global QSAR Modelling Approaches for the Prediction of Bio-Concentration Factor"
2. QSAR 2014 (16th International Workshop on QSAR in Environmental and Health Sciences), 16.–20. juuni 2014, Milano, Itaalia – "Classification models for the bio-concentration factor"

Posterettekanded:

1. QSAR 2010 (14th International Workshop on Quantitative structure-activity relationships in Environmental and Health Sciences), 24.–28. mai 2010, Montréal, Kanada – "QSAR model for the prediction of bio-concentration factor by water solubility as the major descriptor"
2. EURO QSAR 2010 (18th European Symposium on Quantitative Structure Activity Relationships), 19.–24. september 2010, Rhodos, Kreeka – "The Effect of the 3D-geometry on Molecular Descriptors and QSAR Models"
3. Eesti XXXII Keemiapäevad, 14.–15. aprill 2011, Tartu, Eesti – "Molekulide konformatsiooni mõju QSAR-mudeli reprodutseeritavusele"
4. QSAR 2014 (16th International Workshop on QSAR in Environmental and Health Sciences), 16.–20. juuni 2014, Milano, Itaalia – "Reproducibility of QSAR/QSPR models in the scientific literature – perspective for regulatory use"

DISSERTATIONES CHIMICAE UNIVERSITATIS TARTUENSIS

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

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

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

60. **Dimitar Atanasov Dobchev**. Robust QSAR methods for the prediction of properties from molecular structure. Tartu, 2006, 118 p.
61. **Hannes Hagu**. Impact of ultrasound on hydrophobic interactions in solutions. Tartu, 2007, 81 p.
62. **Rutha Jäger.** Electroreduction of peroxodisulfate anion on bismuth electrodes. Tartu, 2007, 142 p.
63. **Kaido Viht.** Immobilizable bisubstrate-analogue inhibitors of basophilic protein kinases: development and application in biosensors. Tartu, 2007, 88 p.
64. **Eva-Ingrid Rõõm.** Acid-base equilibria in nonpolar media. Tartu, 2007, 156 p.
65. **Sven Tamp.** DFT study of the cesium cation containing complexes relevant to the cesium cation binding by the humic acids. Tartu, 2007, 102 p.
66. **Jaak Nerut.** Electroreduction of hexacyanoferrate(III) anion on Cadmium (0001) single crystal electrode. Tartu, 2007, 180 p.
67. **Lauri Jalukse.** Measurement uncertainty estimation in amperometric dissolved oxygen concentration measurement. Tartu, 2007, 112 p.
68. **Aime Lust.** Charge state of dopants and ordered clusters formation in CaF_2:Mn and CaF_2:Eu luminophors. Tartu, 2007, 100 p.
69. **Iiris Kahn**. Quantitative Structure-Activity Relationships of environmentally relevant properties. Tartu, 2007, 98 p.
70. **Mari Reinik.** Nitrates, nitrites, N-nitrosamines and polycyclic aromatic hydrocarbons in food: analytical methods, occurrence and dietary intake. Tartu, 2007, 172 p.
71. **Heili Kasuk.** Thermodynamic parameters and adsorption kinetics of organic compounds forming the compact adsorption layer at Bi single crystal electrodes. Tartu, 2007, 212 p.
72. **Erki Enkvist.** Synthesis of adenosine-peptide conjugates for biological applications. Tartu, 2007, 114 p.
73. **Svetoslav Hristov Slavov**. Biomedical applications of the QSAR approach. Tartu, 2007, 146 p.
74. **Eneli Härk.** Electroreduction of complex cations on electrochemically polished Bi(*hkl*) single crystal electrodes. Tartu, 2008, 158 p.
75. **Priit Möller.** Electrochemical characteristics of some cathodes for medium temperature solid oxide fuel cells, synthesized by solid state reaction technique. Tartu, 2008, 90 p.
76. **Signe Viggor.** Impact of biochemical parameters of genetically different pseudomonads at the degradation of phenolic compounds. Tartu, 2008, 122 p.
77. **Ave Sarapuu.** Electrochemical reduction of oxygen on quinone-modified carbon electrodes and on thin films of platinum and gold. Tartu, 2008, 134 p.
78. **Agnes Kütt.** Studies of acid-base equilibria in non-aqueous media. Tartu, 2008, 198 p.

79. **Rouvim Kadis.** Evaluation of measurement uncertainty in analytical chemistry: related concepts and some points of misinterpretation. Tartu, 2008, 118 p.
80. **Valter Reedo.** Elaboration of IVB group metal oxide structures and their possible applications. Tartu, 2008, 98 p.
81. **Aleksei Kuznetsov.** Allosteric effects in reactions catalyzed by the cAMP-dependent protein kinase catalytic subunit. Tartu, 2009, 133 p.
82. **Aleksei Bredihhin.** Use of mono- and polyanions in the synthesis of multisubstituted hydrazine derivatives. Tartu, 2009, 105 p.
83. **Anu Ploom.** Quantitative structure-reactivity analysis in organosilicon chemistry. Tartu, 2009, 99 p.
84. **Argo Vonk.** Determination of adenosine A_{2A}- and dopamine D_1 receptor-specific modulation of adenylate cyclase activity in rat striatum. Tartu, 2009, 129 p.
85. **Indrek Kivi.** Synthesis and electrochemical characterization of porous cathode materials for intermediate temperature solid oxide fuel cells. Tartu, 2009, 177 p.
86. **Jaanus Eskusson.** Synthesis and characterisation of diamond-like carbon thin films prepared by pulsed laser deposition method. Tartu, 2009, 117 p.
87. **Marko Lätt.** Carbide derived microporous carbon and electrical double layer capacitors. Tartu, 2009, 107 p.
88. **Vladimir Stepanov.** Slow conformational changes in dopamine transporter interaction with its ligands. Tartu, 2009, 103 p.
89. **Aleksander Trummal.** Computational Study of Structural and Solvent Effects on Acidities of Some Brønsted Acids. Tartu, 2009, 103 p.
90. **Eerold Vellemäe.** Applications of mischmetal in organic synthesis. Tartu, 2009, 93 p.
91. **Sven Parkel.** Ligand binding to 5-HT_{1A} receptors and its regulation by Mg^{2+} and Mn^{2+}. Tartu, 2010, 99 p.
92. **Signe Vahur.** Expanding the possibilities of ATR-FT-IR spectroscopy in determination of inorganic pigments. Tartu, 2010, 184 p.
93. **Tavo Romann**. Preparation and surface modification of bismuth thin film, porous, and microelectrodes. Tartu, 2010, 155 p.
94. **Nadežda Aleksejeva.** Electrocatalytic reduction of oxygen on carbon nanotube-based nanocomposite materials. Tartu, 2010, 147 p.
95. **Marko Kullapere.** Electrochemical properties of glassy carbon, nickel and gold electrodes modified with aryl groups. Tartu, 2010, 233 p.
96. **Liis Siinor.** Adsorption kinetics of ions at Bi single crystal planes from aqueous electrolyte solutions and room-temperature ionic liquids. Tartu, 2010, 101 p.
97. **Angela Vaasa.** Development of fluorescence-based kinetic and binding assays for characterization of protein kinases and their inhibitors. Tartu 2010, 101 p.

98. **Indrek Tulp.** Multivariate analysis of chemical and biological properties. Tartu 2010, 105 p.
99. **Aare Selberg.** Evaluation of environmental quality in Northern Estonia by the analysis of leachate. Tartu 2010, 117 p.
100. **Darja Lavõgina.** Development of protein kinase inhibitors based on adenosine analogue-oligoarginine conjugates. Tartu 2010, 248 p.
101. **Laura Herm.** Biochemistry of dopamine D_2 receptors and its association with motivated behaviour. Tartu 2010, 156 p.
102. **Terje Raudsepp.** Influence of dopant anions on the electrochemical properties of polypyrrole films. Tartu 2010, 112 p.
103. **Margus Marandi.** Electroformation of Polypyrrole Films: *In-situ* AFM and STM Study. Tartu 2011, 116 p.
104. **Kairi Kivirand.** Diamine oxidase-based biosensors: construction and working principles. Tartu, 2011, 140 p.
105. **Anneli Kruve.** Matrix effects in liquid-chromatography electrospray mass-spectrometry. Tartu, 2011, 156 p.
106. **Gary Urb.** Assessment of environmental impact of oil shale fly ash from PF and CFB combustion. Tartu, 2011, 108 p.
107. **Nikita Oskolkov.** A novel strategy for peptide-mediated cellular delivery and induction of endosomal escape. Tartu, 2011, 106 p.
108. **Dana Martin.** The QSPR/QSAR approach for the prediction of properties of fullerene derivatives. Tartu, 2011, 98 p.
109. **Säde Viirlaid.** Novel glutathione analogues and their antioxidant activity. Tartu, 2011, 106 p.
110. **Ülis Sõukand.** Simultaneous adsorption of Cd^{2+}, Ni^{2+}, and Pb^{2+} on peat. Tartu, 2011, 124 p.
111. **Lauri Lipping.** The acidity of strong and superstrong Brønsted acids, an outreach for the "limits of growth": a quantum chemical study. Tartu, 2011, 124 p.
112. **Heisi Kurig.** Electrical double-layer capacitors based on ionic liquids as electrolytes. Tartu, 2011, 146 p.
113. **Marje Kasari.** Bisubstrate luminescent probes, optical sensors and affinity adsorbents for measurement of active protein kinases in biological samples. Tartu, 2012, 126 p.
114. **Kalev Takkis.** Virtual screening of chemical databases for bioactive molecules. Tartu, 2012, 122 p.
115. **Ksenija Kisseljova**. Synthesis of aza-β^3-amino acid containing peptides and kinetic study of their phosphorylation by protein kinase A. Tartu, 2012, 104 p.
116. **Riin Rebane.** Advanced method development strategy for derivatization LC/ESI/MS. Tartu, 2012, 184 p.

117. **Vladislav Ivaništšev.** Double layer structure and adsorption kinetics of ions at metal electrodes in room temperature ionic liquids. Tartu, 2012, 128 p.
118. **Irja Helm.** High accuracy gravimetric Winkler method for determination of dissolved oxygen. Tartu, 2012, 139 p.
119. **Karin Kipper.** Fluoroalcohols as Components of LC-ESI-MS Eluents: Usage and Applications. Tartu, 2012, 164 p.
120. **Arno Ratas.** Energy storage and transfer in dosimetric luminescent materials. Tartu, 2012, 163 p.
121. **Reet Reinart-Okugbeni**. Assay systems for characterisation of subtype-selective binding and functional activity of ligands on dopamine receptors. Tartu, 2012, 159 p.
122. **Lauri Sikk.** Computational study of the Sonogashira cross-coupling reaction. Tartu, 2012, 81 p.
123. **Karita Raudkivi.** Neurochemical studies on inter-individual differences in affect-related behaviour of the laboratory rat. Tartu, 2012, 161 p.
124. **Indrek Saar.** Design of GalR2 subtype specific ligands: their role in depression-like behavior and feeding regulation. Tartu, 2013, 126 p.
125. **Ann Laheäär.** Electrochemical characterization of alkali metal salt based non-aqueous electrolytes for supercapacitors. Tartu, 2013, 127 p.
126. **Kerli Tõnurist.** Influence of electrospun separator materials properties on electrochemical performance of electrical double-layer capacitors. Tartu, 2013, 147 p.
127. **Kaija Põhako-Esko.** Novel organic and inorganic ionogels: preparation and characterization. Tartu, 2013, 124 p.
128. **Ivar Kruusenberg.** Electroreduction of oxygen on carbon nanomaterial-based catalysts. Tartu, 2013, 191 p.
129. **Sander Piiskop.** Kinetic effects of ultrasound in aqueous acetonitrile solutions. Tartu, 2013, 95 p.
130. **Ilona Faustova**. Regulatory role of L-type pyruvate kinase N-terminal domain. Tartu, 2013, 109 p.
131. **Kadi Tamm.** Synthesis and characterization of the micro-mesoporous anode materials and testing of the medium temperature solid oxide fuel cell single cells. Tartu, 2013, 138 p.
132. **Iva Bozhidarova Stoyanova-Slavova.** Validation of QSAR/QSPR for regulatory purposes. Tartu, 2013, 109 p.
133. **Vitali Grozovski**. Adsorption of organic molecules at single crystal electrodes studied by *in situ* STM method. Tartu, 2014, 146 p.
134. **Santa Veikšina**. Development of assay systems for characterisation of ligand binding properties to melanocortin 4 receptors. Tartu, 2014, 151 p.
135. **Jüri Liiv.** PVDF (polyvinylidene difluoride) as material for active element of twisting-ball displays. Tartu, 2014, 111 p.

136. **Kersti Vaarmets.** Electrochemical and physical characterization of pristine and activated molybdenum carbide-derived carbon electrodes for the oxygen electroreduction reaction. Tartu, 2014, 131 p.
137. **Lauri Tõntson**. Regulation of G-protein subtypes by receptors, guanine nucleotides and Mn^{2+}. Tartu, 2014, 105 p.
138. **Aiko Adamson.** Properties of amine-boranes and phosphorus analogues in the gas phase. Tartu, 2014, 78 p.
139. **Elo Kibena**. Electrochemical grafting of glassy carbon, gold, highly oriented pyrolytic graphite and chemical vapour deposition-grown graphene electrodes by diazonium reduction method. Tartu, 2014, 184 p.
140. **Teemu Näykki.** Novel Tools for Water Quality Monitoring – From Field to Laboratory. Tartu, 2014, 202 p.
141. **Karl Kaupmees.** Acidity and basicity in non-aqueous media: importance of solvent properties and purity. Tartu, 2014, 128 p.
142. **Oleg Lebedev**. Hydrazine polyanions: different strategies in the synthesis of heterocycles. Tartu, 2015, 118 p.