

A step forward in using QSARs for hazard and exposure assessment of chemicals

Aleksandra Rybacka



Department of Chemistry
Umeå University, 2016

© Aleksandra Rybacka

ISBN: 978-91-7601-504-9

Cover picture: WordleTM, DeviantArt

Electronic version available at <http://umu.diva-portal.org/>

Printed at the KBC Service Centre, Umeå University

Umeå, Sweden, 2016

To my beloved parents

– tradycyjnie, pracę dedykuję Wam, kochani rodzice, więc teraz musicie chociaż spróbować ją przeczytać...

Table of Contents

Abstract	1
Sammanfattning (Summary in Swedish)	3
List of publications	5
List of abbreviations	7
1. Introduction	8
2. Aims and scope	9
3. Background	10
3.1. Chemical safety assessment	10
3.1.1. Hazard assessment	12
3.1.2. Exposure assessment	15
3.1.3. Risk characterisation	16
4. Chemometrics	18
4.1. PCA	18
4.2. QSAR/QSPR	19
4.2.1. A defined endpoint	21
4.2.2. An unambiguous algorithm	22
4.2.3. Applicability domain	24
4.2.4. Validation	25
4.2.5. Interpretation	26
5. Hazard assessment of industrial chemicals – identification of SVHCs	28
5.1. CMR properties	28
5.2. Potential EDCs	34
5.3. WoE approach and consensus modelling	40
5.4. Metabolism simulator	43
5.5. AD assessment	47
6. Exposure assessment of pharmaceuticals – environmental fate at a WWTP	52
7. Conclusions and future prospects	57
Acknowledgements	61
References	63

Abstract

According to the REACH regulation chemicals produced or imported to the European Union need to be assessed to manage the risk of potential hazard to human health and the environment. An increasing number of chemicals in commerce prompts the need for utilizing faster and cheaper alternative methods for this assessment, such as quantitative structure-activity or property relationships (QSARs or QSPRs). QSARs and QSPRs are models that seek correlation between data on chemicals molecular structure and a specific activity or property, such as environmental fate characteristics and (eco)toxicological effects.

The aim of this thesis was to evaluate and develop models for the hazard assessment of industrial chemicals and the exposure assessment of pharmaceuticals. In focus were the identification of chemicals potentially demonstrating carcinogenic (C), mutagenic (M), or reprotoxic (R) effects, and endocrine disruption, the importance of metabolism in hazard identification, and the understanding of adsorption of ionisable chemicals to sludge with implications to the fate of pharmaceuticals in waste water treatment plants (WWTPs). Also, issues related to QSARs including consensus modelling, applicability domain, and ionisation of input structures were addressed.

The main findings presented herein are as follows:

- QSARs were successful in identifying almost all carcinogens and most mutagens but worse in predicting chemicals toxic to reproduction.
- Metabolic activation is a key event in the identification of potentially hazardous chemicals, particularly for chemicals demonstrating estrogen (E) and transthyretin (T) related alterations of the endocrine system, but also for mutagens. The accuracy of currently available metabolism simulators is rather low for industrial chemicals. However, when combined with QSARs, the tool was found useful in identifying chemicals that demonstrated E- and T-related effects *in vivo*.
- We recommend using a consensus approach in final judgement about a compound's toxicity that is to combine QSAR derived data to reach a consensus prediction. That is particularly useful for models based on data of slightly different molecular events or species.
- QSAR models need to have well-defined applicability domains (AD) to ensure their reliability, which can be reached by e.g. the

conformal prediction (CP) method. By providing confidence metrics CP allows a better control over predictive boundaries of QSAR models than other distance-based AD methods.

- Pharmaceuticals can interact with sewage sludge by different intermolecular forces for which also the ionisation state has an impact. Developed models showed that sorption of neutral and positively-charged pharmaceuticals was mainly hydrophobicity-driven but also impacted by Pi-Pi and dipole-dipole forces. In contrast, negatively-charged molecules predominantly interacted via covalent bonding and ion-ion, ion-dipole, and dipole-dipole forces.
- Using ionised structures in multivariate modelling of sorption to sludge did not improve the model performance for positively- and negatively charged species but we noted an improvement for neutral chemicals that may be due to a more correct description of zwitterions.

Overall, the results provided insights on the current weaknesses and strengths of QSAR approaches in hazard and exposure assessment of chemicals. QSARs have a great potential to serve as commonly used tools in hazard identification to predict various responses demanded in chemical safety assessment. In combination with other tools they can provide fundamentals for integrated testing strategies that gather and generate information about compound's toxicity and provide insights of its potential hazard. The obtained results also show that QSARs can be utilized for pattern recognition that facilitates a better understanding of phenomena related to fate of chemicals in WWTP.

Sammanfattning (Summary in Swedish)

Enligt kemikalielagstiftningen REACH måste kemikalier som produceras i eller importeras till Europeiska unionen riskbedömas avseende hälso- och miljöfara. Den ökande mängden kemikalier som används i samhället kräver snabbare och billigare alternativa riskbedömningsmetoder, såsom kvantitativa struktur-aktivitets- eller egenskapssamband (QSARs eller QSPRs). QSARs och QSPRs är datamodeller där samband söks korrelationer mellan data för kemikaliers struktur-relaterade egenskaper och t.ex. kemikaliers persistens eller (eko)toxiska effekter.

Målet med den här avhandlingen var att utvärdera och utveckla modeller för riskbedömning av industri kemikalier och läkemedel för att studera hur QSARs/QSPRs kan förbättra riskbedömningsprocessen. Fokus i avhandlingen var utveckling av metoder för identifiering av potentiellt cancerframkallande (C), mutagena (M), eller reproduktionstoxiska (R) kemikalier, och endokrint aktiva kemikalier, att studera betydelsen av metabolism vid riskbedömning och att öka vår förståelse för joniserbara kemikaliers adsorption till avloppsslam. Avhandlingen behandlar även konsensusmodellering, beskrivning av modellens giltighet och betydelsen av jonisering för kemiska deskriptorer.

De huvudsakliga resultaten som presenteras i avhandlingen är:

- QSAR-modeller identifierade nästan alla cancerframkallande ämnen och de flesta mutagener men var sämre på att identifiera reproduktionstoxiska kemikalier.
- Metabolisk aktivering är av stor betydelse vid identifikationen av potentiellt toxiska kemikalier, speciellt för kemikalier som påvisar östrogen- (E) och sköldkörtel-relaterade (T) förändringar av det endokrina systemet men även för mutagener. Träffsäkerheten för de tillgängliga metabolisimulatorerna är ganska låg för industriella kemikalier men i kombination med QSARs så var verktyget användbart för identifikation av kemikalier som påvisade E- och T-relaterade effekter *in vivo*.
- Vi rekommenderar att använda konsensusmodellering vid *in silico* baserad bedömning av kemikaliers toxicitet, d.v.s. att skapa en sammanvägd förutsägelse baserat på flera QSAR-modeller. Det är speciellt användbart för modeller som baseras på data från delvis olika mekanismer eller arter.
- QSAR-modeller måste ha ett väldefinierat giltighetsområde (AD) för att garantera dess pålitlighet vilket kan uppnås med t.ex. conformal

prediction (CP)-metoden. CP-metoden ger en bättre kontroll över prediktiva gränser hos QSAR-modeller än andra distansbaserade AD-metoder.

- Läkemedel kan interagera med avloppsslam genom olika intermolekylära krafter som även påverkas av joniseringsstillståndet. Modellerna visade att adsorptionen av neutrala och positivt laddade läkemedel var huvudsakligen hydrofobicitetsdrivna men också påverkade av Pi-Pi- och dipol-dipol-krafter. Negativt laddade molekyler interagerade huvudsakligen med slam via kovalent bindning och jon-jon-, jon-dipol-, och dipol-dipol-krafter.
- Kemiska deskriptorer baserade på joniserade molekyler förbättrade inte prestandan för adsorptionsmodeller för positiva och negativa joner men vi noterade en förbättring av modeller för neutrala substanser som kan bero på en mer korrekt beskrivning av zwitterjoner.

Sammanfattningsvis visade resultaten på QSAR-modellers styrkor och svagheter för användning som verktyg vid risk- och exponeringsbedömning av kemikalier. QSARs har stor potential för bred användning vid riskidentifiering och för att förutsäga en mängd olika responser som krävs vid riskbedömning av kemikalier. I kombination med andra verktyg kan QSARs förse oss med data för användning vid integrerade bedömningar där data sammanvägs från olika metoder. De erhållna resultaten visar också att QSARs kan användas för att bedöma och ge en bättre förståelse för kemikaliers öde i vattenreningsverk.

List of publications

The thesis is based on the following papers, which are referred to in the text by the corresponding Roman numerals.

- I. **Aleksandra Rybacka**, Christina Rudén, and Patrik L Andersson. 2014. “On the Use of In Silico Tools for Prioritising Toxicity Testing of the Low-Volume Industrial Chemicals in REACH.” *Basic & Clinical Pharmacology & Toxicology* 115 (1): 77–87.
- II. **Aleksandra Rybacka**, Christina Rudén, Igor V Tetko, and Patrik L Andersson. 2015. “Identifying Potential Endocrine Disruptors among Industrial Chemicals and Their Metabolites--Development and Evaluation of in Silico Tools.” *Chemosphere* 139 (November): 372–78.
- III. Kamel Mansouri, Ahmed Abdelaziz, **Aleksandra Rybacka**, Alessandra Roncaglioni, Alexander Tropsha, Alexandre Varnek, Alexey Zakharov, Andrew Worth, Ann M. Richard, Christopher M. Grulke, Daniela Trisciuzzi, Denis Fourches, Dragos Horvath, Emilio Benfenati, Eugene Muratov, Eva Bay Wedebye, Francesca Grisoni, Giuseppe F. Mangiatordi, Giuseppina M. Incisivo, Huixiao Hong, Hui W. Ng, Igor V. Tetko, Ilya Balabin, Jayaram Kancherla, Jie Shen, Julien Burton, Marc Nicklaus, Matteo Cassotti, Nikolai G. Nikolov, Orazio Nicolotti, Patrik L. Andersson, Qingda Zang, Regina Politi, Richard D. Beger, Roberto Todeschini, Ruili Huang, Sherif Farag, Sine A. Rosenberg, Svetoslav Slavov, Xin Hu and Richard S. Judson. 2016. “CERAPP: Collaborative Estrogen Receptor Activity Prediction Project.” Accepted in *Environmental Health Perspectives*.
- IV. Ulf Norinder, **Aleksandra Rybacka**, and Patrik L Andersson. 2016. “Conformal Prediction to define applicability domain – a case study on predicting ER and AR binding.” *SAR and QSAR in Environmental research* 27 (4): 301-316.
- V. **Aleksandra Rybacka** and Patrik L Andersson. “Considering ionic state in modelling sorption of pharmaceuticals to sewage sludge.” Manuscript.

Published papers are reproduced with permission from publisher (John Wiley & Sons, Elsevier, Environmental Health Perspectives, and Taylor & Francis)

Author's contributions

Paper I

I was involved in planning of the study. I was responsible for data collection, modelling, analysis and interpretation, and writing the paper.

Paper II

I was involved in planning of the study. I was responsible for data collection, modelling, analysis and interpretation, and writing the paper.

Paper III

I was involved in planning of the study. I was responsible for a part of data modelling, as well as analysis and interpretation and corrections of the paper.

Paper IV

I was involved in planning of the study. I was responsible for a part of data modelling, analysis and interpretation, and writing a respective part of the paper.

Paper V

I was involved in planning of the study. I was responsible for data collection, modelling, analysis and interpretation, and writing the paper.

List of abbreviations

AD	applicability domain
ADME	absorption, distribution, metabolism and excretion
API	active pharmaceutical ingredient
ASNNs	associative neural networks
CLP	classification, labelling and packaging
CMR	carcinogenic, mutagenic, or toxic for reproduction
CP	conformal prediction
CSA	chemical safety assessment
DNEL	derived no effect level
ECHA	European Chemicals Agency
FN	false negative
FP	false positive
HPVCs	high production volume chemicals
HTS	high-throughput screening
ITS	integrated testing strategies
k-NN	k-nearest neighbour
K_{ow}	octanol-water partition coefficient
LD ₅₀	median lethal dose
LPVCs	low production volume chemicals
LV	latent variable
MLR	multiple linear regression
OECD	Organisation for Economic Co-operation and Development
PBPK	physiologically based pharmacokinetic
PCA	principal component analysis
PEC	predicted environmental concentration
pKa	logarithmic acid dissociation constant
PLS	partial least squares
PLS-DA	partial least squares discriminant analysis
PNEC	predicted no effect concentration
QSAR	quantitative structure-activity relationships
QSPR	quantitative structure-property relationships
RF	random forest
SVHC	substances of very high concern
SVM	supporting vector machine
TN	true negative
TP	true positive
WWTP	waste water treatment plant

1. Introduction

Once upon a time risk identification was based on human misfortune rather than animal experimentation. As strange as it may sound, a few generations of people all over the world had to unwillingly become “mad as a hatter” before mercury’s neurotoxic effect was discovered and appropriate laws were passed to protect hat makers from occupational exposure to mercuric nitrate (Waldron 1983). The increase in radiation- and cancer-related adverse effects during the 19th and 20th centuries (McClellan 1999) inspired the development of methods that could assess the potential toxic effects of ubiquitous chemicals. This led to a long series of animal tests that focused on the most hazardous substances in use. However, animal tests are expensive, slow, and result in an immense number of killed animals, so it is not surprising that cheaper and faster alternative methods were investigated to assess the constantly growing list of emerging chemicals. A fear of the unknown also led to new style of thinking – a precautionary approach with the axiom “better safe than sorry”.

When Hansch and his group correlated biological activity of pesticides with their structure (Hansch et al. 1962), they may not have realized the full significance of their pioneering work. Quantitative Structure-Activity Relationship (QSAR) models that correlate chemical structure with activity or property (QSPRs) through mathematical equations started emerging in the next years. These models brought more attention to the use of computational tools in risk assessment, which eventually led to a co-operation among OECD member countries in the mid-1990s that aimed to assess QSAR methodology for the prediction of various endpoints related to chemical hazards and exposure. These endpoints included aquatic toxicity (OECD 1992a), biodegradation (OECD 1993b), and octanol-water coefficient (OECD 1993a). Furthermore, during the co-operation recommendations for the future use of QSARs (OECD 1995a; OECD 2007a) were given. The recently introduced legislation on cosmetics (EC 2009b) that bans animal tests in hazard assessments demonstrate that regulatory makers and the scientific community are giving QSARs and other computational tools more trust. The use and development of QSAR tools is a step towards a utopian world, where the fear of wrong prediction does not exist, chemical risk is characterized and fully controlled, and all people, producers and importers of chemical products, workers, governmental agencies, scientists, consumers and other stakeholders, live in peace and harmony.

2. Aims and scope

The research underlying this thesis focuses on evaluating and developing QSARs for the hazard assessment of industrial chemicals and the exposure assessment of pharmaceuticals. The research, identified with roman numerals, was particularly focused on:

- 1) Studying the use of QSARs as 1st tier prioritization tools to identify substances of very high concern (SVHC), that are carcinogenic, mutagenic or toxic for reproduction (CMR) (paper I), and to demonstrate toxic effects related to endocrine disruption (papers II-III).
- 2) Developing new models that can be commonly used to identify SVHCs that can impair the endocrine system (papers II-III).
- 3) Investigating how computational tools that predict the metabolism of chemicals can improve the identification of SVHCs (papers I-II).
- 4) Studying how the use of consensus and weight-of-evidence (WoE) approaches in QSAR modelling can aid the identification of SVHCs (papers I and III).
- 5) Studying the chemical variation of compounds that are defined as outside of the applicability domain of QSAR models developed for identifying potential endocrine disruptors (papers II and IV) and how the use of the conformal prediction method can aid the applicability domain assessment (paper IV).
- 6) Studying factors that influence the environmental fate of pharmaceuticals, notably, their elimination through sorption to sludge in a waste water treatment plant (WWTP) (paper V). The work focused on using QSAR modelling to understand the mechanisms governing this process.

3. Background

Chemicals are a part of our everyday life. We are exposed to a vast number of substances during any stage of their life cycle (extraction, production, use, or waste treatment, as shown in Figure 1). The number of compounds that are industrially produced in high tonnage (over 100 tonnes per year) already exceeds 8000 chemicals (ECHA 2014b), and more than 4000 pharmaceuticals and personal care products are currently in use (Boxall et al. 2012). Some industrial chemicals can be chemical hazards, potentially causing acute or long-term detrimental health effects. Certain compounds, although present at low concentrations, can also be hazardous if they can accumulate in organisms or persist in any of environmental compartments. Moreover, they can have effects that are of high concern for organisms, such as impairing reproduction in vertebrates. Constant exposure to pharmaceuticals that have been designed to interact with specific pathways and processes in humans and animals can also have a negative impact on the environment and human health (Boxall 2004; Jones et al. 2003).

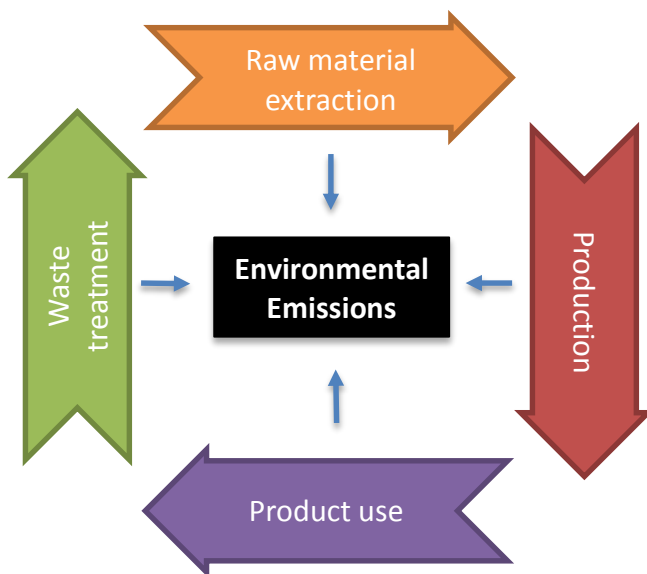


Figure 1: Possible stages in the life cycle of anthropogenic substances.

3.1. Chemical safety assessment

All of the chemicals released during a product life cycle needs to be assessed to manage the risk of potential hazard to human health and the environment. A number of European regulations have been developed to

mandate risk assessment and risk management of chemical substances. For example, industrial chemicals that are produced or imported in quantities of more than 1 ton per year are under REACH regulation (REACH 2006), while EudraLex (EC 2016c) regulates medicinal and veterinary products in the European Union. Additional guidelines for risk assessment have been issued by the European Chemical Agency (ECHA) (ECHA 2011a; ECHA 2012a; ECHA 2012b; ECHA 2014a) and by the European Medicines Agency (EMA) (CVMP 2004; CHMP 2006). Active pharmaceutical ingredients (APIs) and excipients are exempt from registration, evaluation, and authorization within REACH (REACH 2006) (article 2 in REACH) if they are already registered in EMA, but can be subject to REACH if they have other uses that are not medicinal, such as roles of processing aids or growth media. Furthermore, plant protection products, biocides, and food additives can be exempt from REACH since their marketing authorization is controlled by other regulations (EC 2008b; EC 2009a; EC 2012).

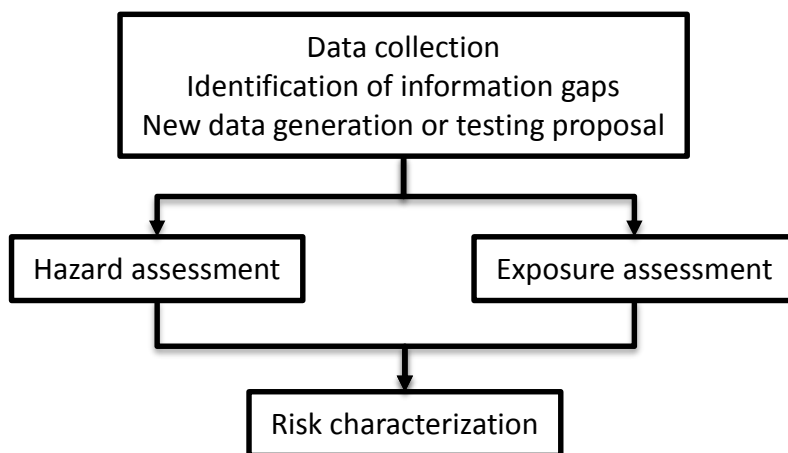


Figure 2: Steps in the chemical safety assessment (CSA) process (based on (ECHA 2011a))

The introduction of REACH led to a paradigm shift, as risk management went from being a separate step to becoming integrated in the risk assessment process (named in the REACH chemical safety assessment, as shown in Figure 2). This meant that chemical safety assessment (CSA) became an iterative process; it includes the improvement of the data and/or methods that are used in the assessment as well as adjustments to the appropriate risk management measures to ensure that risks are controlled (Christensen et al. 2011). While the risk assessment process focuses on the scientific part of identifying the risk, the risk management process focuses on finding mutually accepted risk reduction measures and is often

influenced by economic factors as well as the available alternatives (Leeuwen 2007). Since REACH also aims to reduce animal testing, the risk assessment should include, and combine, various data (of sufficient quality and relevance) before proposing further *in vivo* testing. For this reason, data sharing between the stakeholders of a CSA plays an important role in avoiding unnecessary animal testing. The use of alternative non-animal testing methods is considered to be a major change in the way toxicity testing will be conducted in the future (Krewski et al. 2010) (NRC 2007).

The steps of CSA (Figure 2) will be described next, followed by information on how non-testing tools (in particular QSARs) can be used at every stage. The requirements are given according to REACH.

3.1.1. Hazard assessment

During hazard assessment risk assessors identify a substance's adverse effects. This step includes three elements: evaluation and integration of available information, classification and labelling based on the classification, labelling, and packaging (CLP) regulation (CLP 2008), and derivation of the hazard threshold levels for human health and the environment (ECHA 2011a). This process results in data concerning toxicology, ecotoxicology, and the physicochemical properties of the substances, all of which are used to determine the relationship between dose, or level of exposure, and the severity of the adverse effect. Risk assessors estimate the derived no effect level (DNEL) in the case of human health endpoints and the predicted no effect concentration (PNEC) in the case of environmental endpoints. Data are generally obtained from animal (*in vivo*) and cell (*in vitro*) studies, along with computational (*in silico*) tools, but can also be obtained from epidemiological studies.

CSA is required for substances that are manufactured or imported in quantities of 10 tonnes or more per year. However, for substances that are produced or imported at a lower tonnage (above 1 ton) some information about hazard and exposure is also required, such as hazard classification under CLP, details regarding the chemical use (industrial, professional, and/or consumer), and significant routes of exposure for humans and the environment (ECHA 2011a; REACH 2006). Chemicals that demonstrate a toxic effect within certain endpoints should also be subject to careful attention aimed at limiting exposure to the substance. As a part of the precautionary principle included in REACH to ensure a high level of protection for human health and the environment, substances of very high concern (SVHC) should be identified and their use should be either eventually phased out or become subject to an authorization process

granting allowance for specific use. For a substance to be identified as a SVHC it must meet any of three criteria: 1) the substance is a carcinogen, mutagen, or toxic to reproduction (CMR) according to CLP regulation (CLP 2008); 2) the substance is either persistent, bioaccumulative, and toxic (PBT) or very persistent and very bioaccumulative (vPvB) according to criteria specified in Annex XIII in REACH; 3) the substance shows an equivalent level of concern, for example, as an endocrine disruptor (article 57 in REACH). The identification of endocrine disruptors is currently hampered because, as of yet, no science-based criteria have been proposed. Nevertheless, a roadmap for defining these criteria is currently being developed by the European Commission (EC 2016b). Additionally, the European Commission compiled a priority list of 564 chemicals that have been suggested to be endocrine disruptors by various organisations and the scientific community (EC 2016a). Some of the potential endocrine disruptors are also present on the SIN List (“Substitute It Now!”), which was developed by the International Chemical Secretariat (ChemSec) as an indicator of substances that may be SVHCs (Chemsec 2015). As of 23 September 2015, the SIN list includes 844 chemicals, whereas the Candidate List of SVHCs for Authorisation comprises 168 hits (ECHA 2016a).

3.1.1.1. ITS and computational tools

REACH demands that all of the available information is utilised before any additional tests are conducted. This triggered the development of various integrated testing strategies (ITS) that aim to gather and assess available information. ITS serve to gather and generate information through a weight-of-evidence (WoE) approach that uses a decision tree to address a specific regulatory question, for example, the toxic effect of a substance (Balls et al. 2006). The ITS that were recently developed within the OSIRIS project (Vermeire et al. 2013) are an example of tools equipped with a decision theory framework that includes alternative methods. These ITS provide an intelligent, substance-tailored approach for assessing a chemical hazard in terms of skin sensitisation, repeated dose toxicity, mutagenicity & carcinogenicity, bioconcentration factor, and aquatic toxicity (Vermeire et al. 2013).

Non-testing structure-based approaches are a common component in all ITS (Worth 2010). Information gaps are addressed by predicting a property of interest from known information about similar substances. The data can be derived from one substance (read-across), a group of substances that constitute a category (trend analysis, also termed as internal QSAR), or from QSARs. QSARs use statistical methods to identify correlations between molecular features and the response for a particular endpoint (Kruhlak et al.

2012). Computational tools provide rapid results and can be applied to a large number of chemicals; thus, QSARs usually support a priority testing procedure. Furthermore, QSARs can supplement experimental data in WoE approaches (Hartung et al. 2010) or be used as a replacement. However, the use of QSAR as a standalone replacement is seldom seen, especially for human health endpoints, as seen from the industrial chemicals registered within REACH up until 2013 (ECHA 2011b; ECHA 2014b). In contrast, the standalone use of QSARs is more prevalent for environmental endpoints, as the use in the high-tonnage category of chemicals reached 39% for studies of bioaccumulation in fish and approximately 4% for short- and long-term studies of toxicity to fish (ECHA 2014b).

To increase the regulatory acceptance of QSAR methods, the Organisation for Economic Co-operation and Development (OECD) has defined validation principles (OECD 2007a) and developed a QSAR Toolbox (OECD 2015i) intended for use by governments, the chemical industry, and other stakeholders to fill gaps in toxicological and ecotoxicological data needed in hazard assessment. Additional freely available or commercial tools also exist (ANTARES 2011). These QSAR models cover a vast number of ecotoxicological (aquatic toxicity and effects on terrestrial organisms) and toxicological (skin irritation, eye irritation, skin sensitisation, mutagenicity, acute toxicity, repeated dose toxicity, reproductive toxicity, toxicokinetics and carcinogenicity) endpoints.

3.1.1.2. ADME and computational tools

Although toxicokinetics, which describe the absorption, distribution, metabolism, and excretion (ADME) of a substance after entering the body, are not required within REACH, their assessment is encouraged as a means to interpret data or assist testing strategy and study design (ECHA 2014a). Certain information regarding toxicokinetic behaviour might also deem the further testing of properties unnecessary, for example, if a query compound violates more than one of Lipinski's rules (Lipinski 2004), such as that $\log P$ is greater than 5 and that the compound has more than ten hydrogen bond acceptors, then it will most likely not be orally active in humans. During the derivation of dose-response curves different assessment factors need to be involved in the extrapolation of experimental data to the human situation, such as aspects related to interspecies variability and different experimental exposure duration (ECHA 2012a). These assessment factors can be determined through physiologically based pharmacokinetic (PBPK) modelling.

QSAR models play an important role in predicting the properties and potencies that can serve as inputs in ADME modelling. QSAR models cover mainly factors related to absorption (such as intestinal permeability), distribution (blood-brain barrier permeability, plasma protein binding, steady-state volume of distribution), and metabolism (cytochrome P450 catalysed biotransformations and UDP-glucuronosyltransferases metabolism) (Sridhar et al. 2012; Chohan et al. 2008; M. Honorio et al. 2013). QSAR models can also predict physicochemical properties such as octanol-water partition coefficient (K_{ow}), solubility, and logarithmic acid dissociation constant (pKa) to provide information about membrane permeability.

3.1.2. Exposure assessment

During the exposure assessment step (Figure 2) risk assessors estimate both the occupational exposure and exposure related to consumers and the environment to determine either the predicted environmental concentration (PEC) or human intake (Leeuwen 2007; ECHA 2012b). The exposure is calculated based on the emissions and pathways of the substance at any stage in its life cycle (Figure 1). Once a substance is released into the environment it enters a certain compartment and then diffuses between compartments (air, water, soil, sediment) to attain equilibrium. Organic acids and bases can be present in ionic form depending on a compartment's pH, which greatly affects both their fate and toxicity, and challenges scientists to develop tools that can estimate the behaviour of such chemicals (Franco et al. 2010). The substances can be consumed by organisms, which leads to accumulation, and they can be transformed into other substances (metabolites). During biotic degradation (biodegradation) various biological organisms, such as bacteria and fungi, dissolve and decompose the chemical materials present in water, soil and sediment. On the other hand, abiotic degradation processes use physical and/or chemical mechanisms to transform compounds into simpler products, with the most important processes being hydrolysis, oxidation, reduction and photochemical degradation. A CSA, however, only takes into account hydrolysis in surface water, as well as photolysis in surface water and the atmosphere, when assessing abiotic degradation (ECHA 2012a). Biodegradation in surface water, soil, and sediment, which is already required at lower tonnages, should also be included.

The fate and distribution of the released substance in the different environmental compartments is estimated based on the mapping of uses and relevant substance properties (ECHA 2012a). Both measurements and modelling can be used to assess exposure concentration. The latter is more

common since it reflects the “average” concentration, whereas the location and time of measurements can introduce bias (Meent & Bruijn 2007). In every compartment chemical concentration is governed by the principle of mass conservation, i.e. only mass flows of a substance into and out of a compartment can cause the appearance or disappearance of mass in the compartment. Entire environmental media can be represented through multimedia fugacity models, in which the lifetime of chemicals partitioned among air, soil, sediment, and water is predicted based on mass balance equations for each phase (Wania & Mackay 1999).

SimpleBox, which was developed by the European Chemical Agency for the calculation of PEC and human exposure, is an example of an environmental fate model included in the European Union System for the Evaluation of Substances (EUSES) (ECHA 2012a). The fate of a substance in a WWTP provides essential information about the concentrations of chemicals entering the aquatic system, which is particularly important for pharmaceuticals that are constantly present at low concentrations in the environment. SimpleTreat (ECHA 2012a) was developed for this purpose, and has been recommended for estimating the fate of substances in a WWTP.

QSPRs are utilized in the exposure assessment. QSPR predictions of various physicochemical properties, such as K_{ow} , Henry’s law constant, and soil-water/sediment-water partition coefficients, can serve as inputs for multimedia models. Additionally, QSARs can be used to estimate the bioconcentration factor (BCF) in fish when there is no data available, for example, if the chemical has low solubility. EPI Suite is an example of software that allows the user to estimate physicochemical and environmental fate properties (EPA 2016).

3.1.3. Risk characterisation

During the final stage of CSA, the hazard and exposure assessments are integrated and the risk is characterized (Figure 2) by estimating the incidence and severity of the adverse effects and the likeliness that they will occur in the human population or an environmental compartment.

In the risk assessment for humans, the long-term and acute effects are determined for each exposure pattern, that is, inhalation, dermal and oral exposure routes for the exposed population (workers, general population, consumers and humans exposed via the environment) (ECHA 2012c). And in particular for pesticides and food additives, an Acceptable or Tolerable Daily Intake (ADI and TDI) is derived (OECD 2014a; OECD 2004a).

In the risk assessment for the environment, the PEC and PNEC estimates are compared for each of the inland environmental protection targets (aquatic and terrestrial ecosystem, atmosphere, predators, and micro-organisms in sewage treatment plants) and each of the marine environmental protection targets (aquatic ecosystem as well as predators and top predators)(ECHA 2012c).

Once the risk is characterized, appropriate measures can be taken for how to address the risk. If an unacceptable risk is identified, then risk management measures will aim at reducing the highest exposures. These measures may comprise introducing appropriate labelling and instructions, redesigning the production process, limiting the marketing and use of the substance, and/or improving waste water treatment options (START 2008; Engelen et al. 2007).

4. Chemometrics

Chemometrics utilise informatics methods to study chemical problems (Varmuza & Filzmoser 2008). Usually, different methods are used to transform input data and extract the most significant chemical information. Depending on the aim of the analysis, that is, if the variation among samples should be related to a certain response, modelling can either be unsupervised (independent of a response), such as principal component analysis (PCA), or supervised (dependent on a response), such as partial least squares (PLS) regression. QSAR and QSPR modelling are forms of supervised models.

4.1. PCA

PCA can be used to study the similarities, or dissimilarities, among n objects, based on a matrix \mathbf{X} that contains m variables (Figure 3). The objects can be, for example, chemicals or samples, whereas the variables can comprise spectral data, concentration levels, or various physicochemical properties. The multivariate environment is represented in a new coordinate system formed by k orthogonal latent variables (components) that describe the maximum possible variance and separate significant information from noise (matrix \mathbf{R}). Data values are projected onto the latent variables by means of scores (matrix \mathbf{T}) that reflect the distances between objects, where a larger distance means less similarity. The first latent variable (called the first principal component in PCA) describes the maximum variance of the scores and each consecutive component is orthogonal to the preceding components (the maximum number of k components equals the number of original m variables). Usually, the first two or three principal components are used for scatter plots as they contain most of the variance (Varmuza & Filzmoser 2008). The corresponding loading vectors (matrix \mathbf{P}) describe the importance of every original variable in a component, and hence, characterise the objects. For example, water samples could be grouped according to similar concentrations of particular pollutants. Scores mapped in the space of k components would show which of the n samples share similar chemical content, whereas the loading vectors would describe which of the m pollutants are the most important for every component, i.e. which pollutants are present at the highest concentrations for every group of samples.

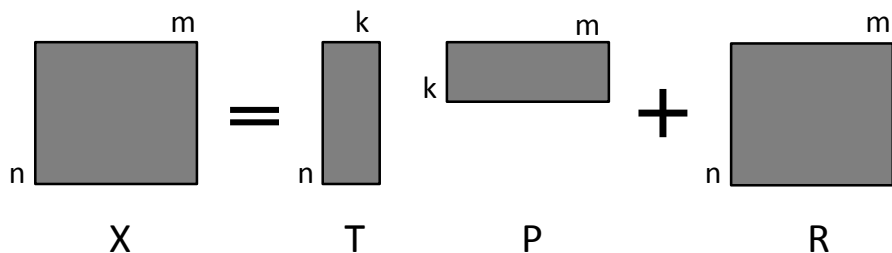


Figure 3: Reconstruction of the X -matrix from PCA scores, T , and the loading matrix P using k components; R is the residual matrix.

PCA is commonly used to visualise multivariate data (through scatter plots), or to decrease the number of variables to a set of uncorrelated variables that can be used in further analyses, for example, in an applicability domain assessment (section 4.2.3. ‘AD assessment’). When a number of chemicals are entered as n variables and their structural characteristics (descriptors), described numerically, are entered as m variables, PCA can give a good overview of the variation among chemicals. Additionally, PCA is particularly useful for identifying objects that behave differently from the rest of the set, i.e. that are potential outliers; thus, it is a good starting point for further QSAR/QSPR development.

4.2. QSAR/QSPR

QSAR/QSPR methods model the mathematical relationship $y=f(X)$, where X data describe so-called molecular descriptors and y describes the quantitative or qualitative response, for example, LD_{50} or binary information about an adverse effect (Figure 4). The molecular descriptor is “the final result of a logical and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment” (Todeschini & Consonni 2000). Based on the collected data (training set), a relationship between X and y is modelled and internally validated through methods such as leave-more-out cross-validation (Efron & Gong 1983) or y permutation test (Wold et al. 1995). Additional data (validation set) that are not used in developing the $y=f(X)$ relationship are usually used for the external validation of a model. The model can be applied to new compounds (Figure 4) once its predictive accuracy has been shown to be at an acceptable level through statistical methods. However, the limits of the model need to be determined (applicability domain estimation) before its application. A

prediction for a query compound is reliable only if the chemical falls within the applicability domain of the model.

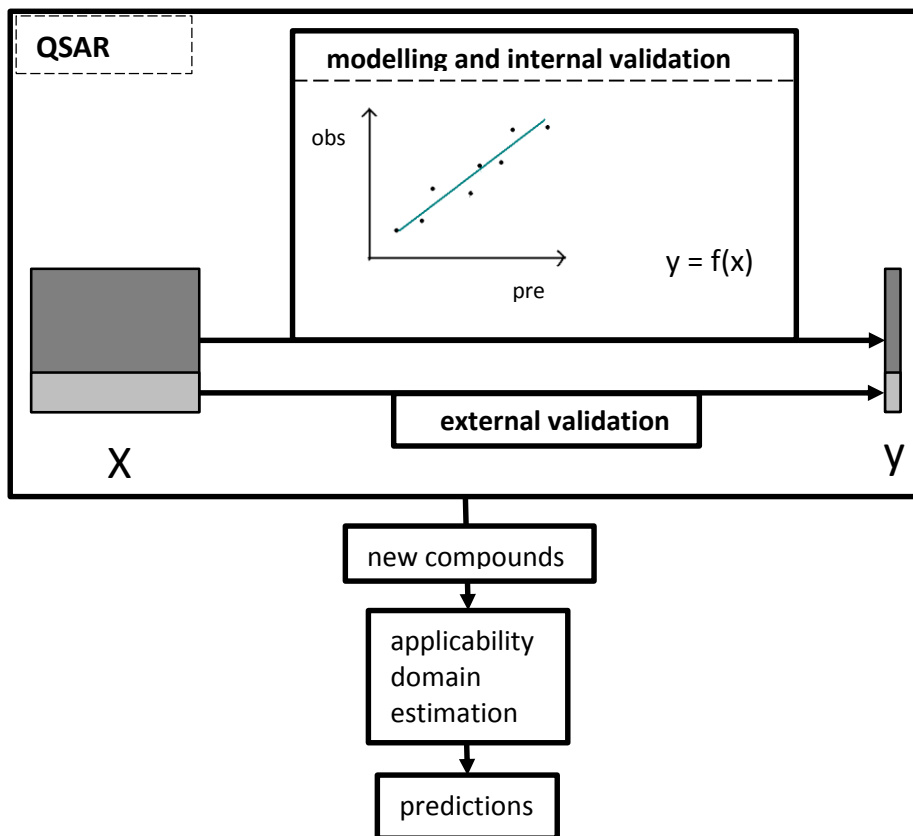


Figure 4: An overview of QSAR modelling. The dark grey data was used for modelling (training set) and light grey data was used for validation (validation set).

The OECD countries established a set of five principles that characterise a reliable model to facilitate the use of QSARs for regulatory purposes (OECD 2007a). These principles state that a QSAR model should be associated with the following information:

- 1) A defined endpoint
- 2) An unambiguous algorithm
- 3) A defined domain of applicability
- 4) Appropriate measures of goodness-of-fit, robustness and predictivity
- 5) A mechanistic interpretation, if possible

The five steps of QSAR modelling, with respect to the OECD principles, are discussed below.

4.2.1. A defined endpoint

Ideally, for a QSAR model to be adequate it should not only be scientifically valid and applicable to the query chemical, but also relevant for regulatory purposes (Figure 5.). For this reason, although models that predict undefined endpoints, such as hepatotoxicity (Valerio 2009), identify potential toxicity, they should be considered vague, and hence, not acceptable from a regulatory point of view. From scientific point of view, however, the model may aid the understanding of studied phenomena or can be used for advancing QSAR methodology.

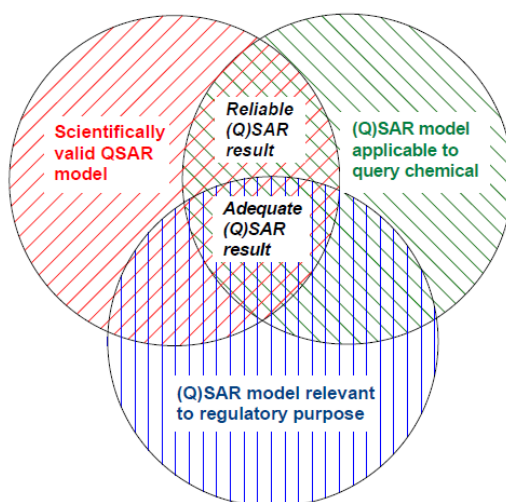


Figure 5: The overlapping considerations of validity, applicability, and relevance, all of which are necessary to demonstrate the adequacy of a QSAR model (Worth 2010).

To facilitate the use of QSARs for regulatory decisions, it would be ideal to build models that predict endpoints defined by harmonised test protocols (Zvinavashe et al. 2008). Such endpoints are associated with OECD test guidelines and can comprise:

- physicochemical properties (for example, partition coefficient (n-octanol/water) (OECD 2006) or boiling point (OECD 1995b))
- effects on biotic systems (e.g. acute toxicity for daphnids (OECD 2004b) or fish (OECD 1992b) by means of effect (EC₅₀) or lethal concentration (LC₅₀) for 50% of the species)

- degradation and accumulation (e.g. bioaccumulation in fish (OECD 2012a) by means of bioconcentration factor or readily biodegradability (OECD 1992c))
- health effects (e.g. genotoxicity in bacterial cells (OECD 1997a), lethality in rodents (OECD 2015e)).

4.2.2. An unambiguous algorithm

To ensure that the model is transparent, the algorithm, descriptors, and modelled chemicals that were used should be well-defined so that the user understands how the value was estimated and can reproduce the calculations, if needed.

Two classes of molecular descriptors exist: 1) experimental measurements, for example, $\log K_{ow}$, dipole moment, polarizability, and other physicochemical properties, and 2) theoretical molecular descriptors derived from a symbolic representation of the molecule (Consonni & Todeschini 2009). The structural information descriptors can be divided into groups based on their dimensions, as follows:

- 0D descriptors – based only on the chemical formula, for example, a number of atoms.
- 1D descriptors (substructure list representation) – based on the structural formula, for example, a list of structural fragments/functional groups.
- 2D descriptors (topological representation) – based on the two-dimensional representation of a molecule, for example, molecular graph or Simplified Molecular Input Line Entry Specification (SMILES). These descriptors define the connectivity of atoms in a molecule.
- 3D descriptors (geometrical descriptors) – based on the three-dimensional representation of a molecule. These descriptors need to be calculated by means of quantum mechanics. They describe the bulk and steric properties, surface area, volume, and energy of the molecule.
- 4D descriptors – characterised by a scalar field associated with the 3D molecular geometry. These descriptors are rather seldom used in typical QSAR/QSPR modelling.

Machine learning methods construct algorithms that can learn from, and make predictions based on, data, as well as calculate the $y=f(X)$ relationship. The first step is usually the construction of a linear regression, for example, a

multi-linear regression (MLR) (eq. 1) as this model is relatively easy to interpret.

$$y = b_0 + b_1x_1 + b_2x_2 + \dots + b_kx_k + e \quad (1)$$

MLR is simple to use and usually explains the studied phenomena through only a few descriptors. However, when too many descriptors are used there is a risk of both over-fit and a higher degree of error. Therefore, the modeller should ensure that the descriptors are not highly correlated, which can be evaluated by means of a F-test, for example. MLR is commonly used when the mechanism underlying a particular phenomenon is known prior to the modelling so that the focus is on choosing the appropriate descriptors. In other cases, partial least squares (PLS) regression may be a more useful alternative. While MLR describes the maximum correlation by relating the response (y) with a carefully chosen set of descriptors (\mathbf{X}), PLS correlates the response with latent variables (LV) created from the \mathbf{X} data. The framework underlying PLS is similar to that of PCA: the \mathbf{X} data are represented as a linear combination of the original variables (see section 4.1 'PCA'). The difference is that instead of principal component scores (that are derived solely from \mathbf{X}), the components in PLS are related to y . The modelled property (y) is a linear combination of these new components, which are constructed by means of MLR. In this way, PLS is insensitive to collinear variables and it captures the high variance of \mathbf{X} and the correlation between \mathbf{X} and y , resulting in high covariance between \mathbf{X} and y (Varmuza & Filzmoser 2008). The classification of compounds according to this method is commonly referred to as PLS discriminant analysis (PLS-DA).

QSAR modelling can also use non-linear functions, such as artificial neural networks (ANNs), k -nearest neighbour (k -NN), support vector machines (SVMs) and random forest (RF). ANNs employ the idea of interconnected "neurons", which are mathematical functions conceived as a model of biological neurons. These neurons consist of different linear combinations of x -variables that are then applied to a nonlinear function to create z -variables. These z -variables can be then used in different ways to produce the final y : a) as inputs of a neuron with output y , b) in a linear regression model, and c) in a nonlinear regression model (Varmuza & Filzmoser 2008). In k -NN, the closest k objects in chemical space (in PCA, for example) are used to determine the group membership of a new object, which is expected to behave similarly to neighbouring objects. The sample is then classified using similarity measures, such as the Tanimoto index (Willett 1987), which is used for binary variables. SVM separates objects by using linear boundaries that are produced in a transformed space of the x -variables which, contradictory to PCA, is at a higher dimension than the original descriptor

set (Varmuza & Filzmoser 2008). RF is an ensemble learning method in which random decision trees are created to classify instances with similar values in subsets of the data set.

4.2.3. Applicability domain

The application of QSARs to new chemicals is limited “in terms of the types of chemical structures, physicochemical properties and mechanisms of action for which the models can generate reliable predictions” (OECD 2007a). However, no specifications have been given for how the AD should be assessed. AD assessment methods focus on ways to describe the similarity of a query compound to the set of chemicals that a QSAR model is based upon. In this way, a high degree of similarity to the chemicals in the training set should imply that the prediction for a new chemical also has a high degree of accuracy. Some of approaches for defining the AD are described below.

The simplest approach defines AD with boundaries of the descriptors that were used in the training set. This approach, termed Boundary Box, can be limited directly by values of the used descriptors or by other means, such as LVs from PCA (Sahigara et al. 2012). When a query compound is present inside of this m -dimensional hyper-rectangle, then it is considered to be inside AD. This approach cannot, however, identify potential internal empty regions that exist within the interpolation space (Netzeva et al. 2005).

Distance-based methods calculate the distance between a query compound and a defined point within the descriptor space. The distance can be calculated by different measures, such as Mahalanobis, Euclidean and City Block distances (Jaworska et al. 2005), after which it is compared to a pre-defined threshold. The ‘leverage’ measure based on this principle is commonly used by many QSAR modellers (Gramatica 2007). Chemical similarity can also be assessed by means of k-NN approach (see previous section for more details). The main pitfall of distance-based approaches is that the “distance” is open to interpretation by the QSAR modeller. Various numbers of correctly predicted chemicals can be obtained depending on the value chosen for threshold and the distance measure used (Sahigara et al. 2012).

More advanced methods include using the standard deviations of predicted values on the basis of ensemble learning (Kaneko & Funatsu 2014), probability density distribution-based methods (Netzeva et al. 2005; Sahigara et al. 2012), stepwise approaches using multiple distance-based

methods in parallel (Dimitrov et al. 2005) or combining similarity, ensemble learning by means of RF and the predicted value itself (Sheridan 2012).

Knowledge of the mechanism of action can also aid in defining the AD of the model (Dimitrov et al. 2005). For example, predicting the aquatic toxicity of compounds within the AD of a QSAR model based on substituted mononitrobenzenes can lead to an underestimation of certain $\log K_{ow}$ ranges; in such cases, the toxicity is better explained by a QSAR model based on nonpolar narcotics (Zvinavashe et al. 2008).

The recently developed concept of conformal predictions (CP) (Norinder et al. 2014) incorporates confidence metrics into predictions, which makes it easier to choose an appropriate similarity threshold. In the CP method, the prediction regions are determined by a prediction algorithm. This process uses a conformity measure that measures how similar a query compound is to the training set, which can be defined as the probability that a prediction is equal to the percentage of correct predictions given by the machine learning method. Another aspect of CP is validity, which is similar, but not identical, to accuracy and is related to the chosen confidence level. Based on the confidence level used, chemicals can be classified as ‘empty’ (out of domain) or ‘both’ (the model is unable to assign a class) when the validity is calculated. The research underlying this thesis studied this approach, and its application is discussed further in the section 5.5 ‘AD assessment’.

4.2.4. Validation

Every QSAR model should provide parameters that describe performance, which are usually based on the difference between observed and predicted y . The goodness-of-fit-measure describes how well a model explains variation in the response and is usually characterised by the coefficient of determination (R^2), where n is number of chemicals in the training set (eq. 2). Internal validation, on the other hand, describes the robustness of a model by measuring stability. It can be defined as Q_{cv}^2 (eq. 3), which is derived from a cross-validation procedure in which the model is recalculated with either one or a few compounds omitted in every step. Internal validation can also utilise the bagging (bootstrap aggregating) procedure, in which an ensemble of models based on random training sets is created from the original training set by sampling with replacement (Breiman 1996). Predictivity is usually measured by using an additional external set to calculate the externally validated coefficient of determination (R_{pred}^2) (eq. 4), where n is number of chemicals in the test set. For qualitative responses, the statistics for goodness-of-fit, robustness, and predictivity are based on the number of true positives (TPs), true negatives (TNs), false positives (FPs)

and false negatives (FNs) predicted by a model that may include precision (eq. 5), specificity (eq. 6), sensitivity (eq. 7), balanced accuracy (eq. 8), and F-measure (F)(eq. 9).

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i^{obs} - y_i^{pred})^2}{\sum_{i=1}^n (y_i^{obs} - \bar{y}_i^{obs})^2} \quad (2)$$

$$Q_{cv}^2 = 1 - \frac{\sum_{i=1}^n (y_i^{obs} - y_{i-i}^{pred})^2}{\sum_{i=1}^n (y_i^{obs} - \bar{y}_i^{obs})^2} \quad (3)$$

$$R_{pred}^2 = 1 - \frac{\sum_{i=1}^n (y_i^{obs} - y_i^{pred})^2}{\sum_{i=1}^n (y_i^{obs} - \bar{y}_i^{obs})^2} \quad (4)$$

$$precision = \frac{TPs}{TPs+FPs} \quad (5)$$

$$specificity = \frac{TNs}{TNs+FPs} \quad (6)$$

$$sensitivity = \frac{TPs}{TPs+FNs} \quad (7)$$

$$balanced\ accuracy = \frac{sensitivity+specificity}{2} \quad (8)$$

$$F = \frac{2 \times precision \times sensitivity}{precision + sensitivity} \quad (9)$$

4.2.5. Interpretation

Linear methods make it easier to understand models because they allow the descriptors of highest importance to be investigated. For this reason, understandable descriptors aid model interpretation and contribute to the understanding of mechanisms that govern a particular phenomenon. However, a response is often the result of a series of complex biological or physicochemical mechanisms, which can make ascribing mechanistic meaning to the molecular descriptors very difficult, as well as reductionist (Gramatica 2010). Some scientists (Livingstone 2000; Zefirov & Palyulin 2001) even differentiate between predictive QSARs, which are designed to maximise predictivity, and descriptive QSARs, which aim for descriptor interpretability, as often models that use non-linear machine learning

methods and/or contain difficult-to-interpret descriptors attain the highest predictivity. Therefore, fulfilling this principle depends on a model's aim.

5. Hazard assessment of industrial chemicals – identification of SVHCs

This section presents the results from papers I-IV. The research underlying these papers focused on using QSARs to identify CMR chemicals and potential endocrine disruptors. The results were split into five subsections, which focused on:

1. An evaluation of individual expert systems as a way to identify CMR chemicals (paper I)
2. Development of QSAR models to identify potential endocrine disrupting chemicals (EDCs) among high production volume chemicals (HPVCs) and low production volume chemicals (LPVCs) (paper II)
3. Studying how the use of consensus and WoE approaches in QSAR modelling can aid the identification of SVHCs (papers I and III)
4. Studying how the combination of metabolite simulators and QSAR modelling can facilitate the identification of SVHCs (papers I and II)
5. Studying the AD issues faced in papers I-III, and how the CP method can address these issues (paper IV)

5.1. CMR properties

Carcinogens are agents that are directly involved in causing cancer and can be classified as either genotoxic or non-genotoxic depending on their specific pathogenic mechanism. Genotoxic carcinogens can cause irreversible damage or mutations to genetic material by binding to DNA, whereas non-genotoxic carcinogens can stimulate cell growth in other ways, for example, by disrupting cellular structures, changing the rate of cell proliferation, or other processes that can lead to genetic error (Lee et al. 2013). The OECD has validated guidelines for carcinogenicity studies that are based either on rodents (OECD 2009b)(OECD 2009c) or non-rodents (OECD 1998), for example, dogs or swine. The complexity of the endpoint makes it difficult to develop adequate *in vitro* alternatives (Adler et al. 2011), and this has prompted the development of integrated approaches that involve multiple *in vitro* models, or methods that can model an *in vivo* response. According to CLP regulation, carcinogens can be classified as category 1A (known carcinogen) or 1B (presumed carcinogen) based on scientific evidence (CLP 2008). In cases of equivocal data or results from only *in vitro* data, chemicals can be classified as suspected human carcinogens (category 2). QSAR models for carcinogenicity are mostly based on *in vivo* rodent data and they include both statistics-based and rule-based models that comprise a

set of structural alerts such as the Benigni/Bossa rules, which were based on *Salmonella thyphimurium* (Benigni & Bossa 2008b).

Mutagens can cause both direct and indirect damage to DNA, resulting in mutations (genetic alterations). A number of OECD guidelines exist for *in vitro* bacterial (OECD 1997a), *in vitro* mammalian (OECD 2015g; OECD 2015d; OECD 2014b; OECD 2014e), and *in vivo* mammalian assays (OECD 1986; OECD 1997b; OECD 2015e; OECD 2015f; OECD 2014f; OECD 2014c; OECD 2014d; OECD 2013) that can be used for testing potential mutagens. Several QSARs that can be used for prioritising chemicals for further testing have been developed based on these assays. As for carcinogenicity, chemicals can be classified as mutagens of category 1A or 1B if they are known, or presumed, to induce hereditary mutations in human germ cells, or as mutagens of category 2 if there is a concern that they could induce these mutations.

Another classification is toxicity to reproduction, which includes compounds that impair the reproductive system and/or cause adverse effects in a developing embryo or foetus. The OECD guidelines for tests of toxic effects to reproductive systems and fetuses basically only cover *in vivo* tests (OECD 2015a; OECD 2015b; OECD 2007b; OECD 2012b; OECD 2001a; OECD 1983; OECD 2001b), however, *in vitro* tests related to endocrine disruption, such as estrogenic and androgenic activity (OECD 2015h; OECD 2015c; OECD 2012c) or steroidogenesis (OECD 2011), may also identify chemicals that are potential reproductive toxicants. Furthermore, short-term *in vivo* tests, such as the Hershberger (OECD 2009a) and uterotrophic (OECD 2007c) assays, as well as *in vivo* reproduction/developmental toxicity screening tests updated with endocrine disruptor endpoints (OECD 2015a; OECD 2015b), could be used for this purpose. Since reproductive toxicity is a complex endpoint, QSAR models are usually developed for particular *in vitro* responses (Matthews et al. 2007) and are often related to endocrine disruption (Cronin & Worth 2008). QSAR models for developmental toxicity have been developed, which is as well a complex endpoint (Cassano et al. 2010).

QSAR models can be implemented in expert systems that emulate the decision-making ability of a human expert; apart from statistics-based models, these systems can also include 'if-then-else' rules based on structural alerts defined by human experts, which has been done in models for carcinogenicity and mutagenicity (Benigni & Bossa 2008b). A list of available models that may be applicable for REACH was compiled within the ANTARES project (ANTARES 2011). Out of these, we studied 47 QSAR

models that had been implemented in expert systems to predict CMR effects (paper I). These expert systems were:

- statistics-based
 - QSAR Toolbox (20 models)(<http://www.qsartoolbox.org/>),
 - LAZAR (8 models)(<http://lazar.in-silico.de/>),
 - TEST (2 models)
(<https://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test>),
- rule-based
 - Toxtree (5 models)(<http://toxtree.sourceforge.net/>)
 - Derek Nexus (8 models)(<https://www.lhasalimited.org/>)
- hybrid
 - VEGA (4 models)(<http://www.vega-qsar.eu/>)

This array of QSAR models was evaluated against a set of organic compounds with well-described risk assessment reports. Before the introduction of REACH, the national authority for chemical safety of each European Union member state held the burden of risk identification and management. During this time the EU assessed 141 chemicals out of the chemicals produced or imported in the EU in high volumes (quantities of 1000 tonnes or more per year). These chemicals were carefully evaluated over a range of endpoints, including CMR effects, and thus, the results provide a good dataset for evaluating QSAR models. Salts and inorganics were omitted, which resulted in a total of 91 chemicals (the RAR set). In addition, information about the CMR effects of 152 compounds was extracted from Annex VI in CLP (CLP 2008) and these chemicals were then used as the validation set (the CLP set).

The workflow of the study was as follows (shown in Figure 6):

- the chemical variation for the RAR and CLP sets was examined (step 1),
- QSAR models with the RAR set (step 2) and the CLP set (step 3) as inputs were evaluated by comparing the predictions with the CLP classification based on CMR properties,
- the results for the RAR and CLP sets were compared (step 4),
- chemicals commonly predicted as false negatives (FNs) were identified (step 5),
- the predictions from all models using the RAR set were combined in a WoE approach (step 6),
- an ITS tool and metabolism simulator were applied to the common FNs (step 7).

Steps 6 and 7 were investigated in the sections 5.3. ‘WoE approach and consensus modelling’ and 5.4. ‘Metabolism simulator’, respectively.

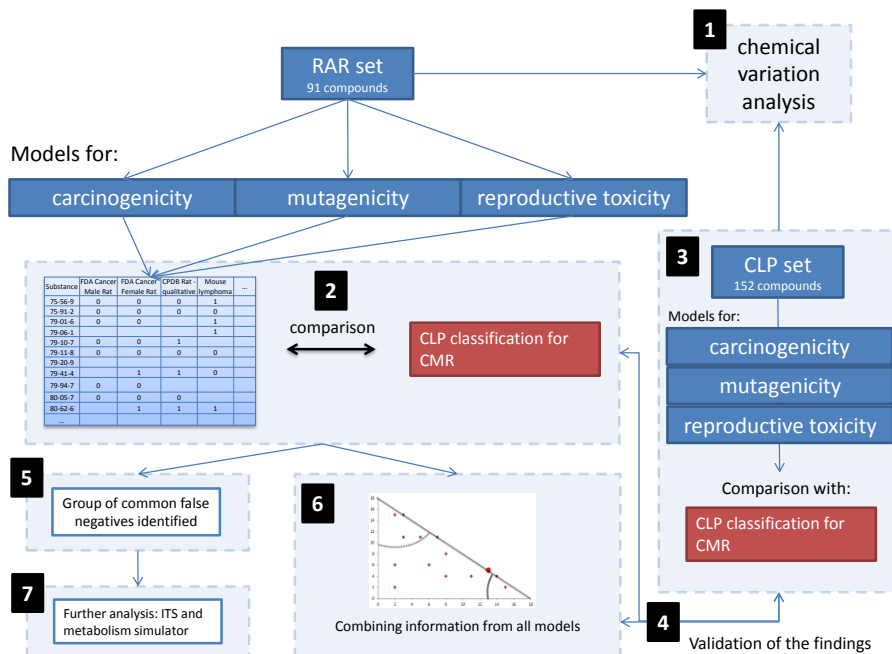


Figure 6: Workflow of the study presented in paper I. The numbers in black boxes represent steps in the study, which are described in different sections of this thesis (see the main text for more details).

Step 1

To investigate how well the studied chemicals (RAR and CLP sets) represent commonly used chemicals, they were mapped using PCA based on calculated chemical descriptors, with HPVCs and LPVCs as representatives of commonly used industrial chemicals including 6655 chemicals (Figure 7). Both sets showed variation when compared to the HPVCs and LPVCs on the basis of their chemical properties. However, the CLP set showed a better spread than the RAR set and included more hydrophilic compounds (upper half of Figure 7B), and the latter difference was shown to be statistically significant when the results of models for mutagenicity were evaluated (see below).

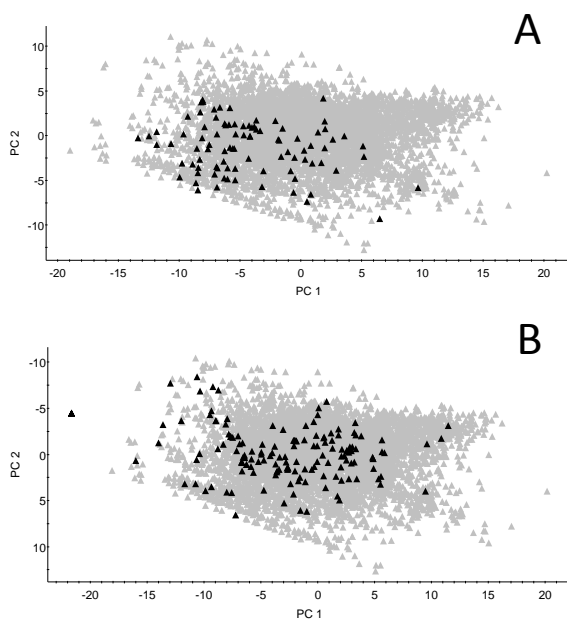


Figure 7: Score plots based on the RAR set (black triangles in plot A) and the CLP set (black triangles in plot B), mapped through PCA with HPVCs & LPVCs as background (grey triangles). Abbreviations are explained in the text.

Steps 2-4

The QSAR model that assessed carcinogenicity yielded the most accurate results, demonstrated by an F-measure of 0.86 for female rat carcinogenicity (from the QSAR Toolbox). However, the QSAR models that assessed mutagenicity performed much worse, as the most accurate model, which was based on the Ames assay (from the QSAR Toolbox), only reached an F-measure of 0.58. This poor result may have stemmed from a low number of evaluated mutagenic chemicals ($n=15$), since the F-measure of a model based on Ames assay validating the CLP set was 0.75, and this measure was even better (0.83 F-measure) for the DNA reactivity model based on Ashby fragments (Ashby & Tennant 1991) from the QSAR Toolbox. The models that assessed reproductive toxicity performed the worst, which was expected, since, as mentioned before, reproductive toxicity is a complex endpoint and the models were focused on only one aspect, such as estrogen binding or developmental toxicity. According to a recent report from ECHA (ECHA 2016b) the use of QSARs as stand-alone information for such complex endpoints is unacceptable. The reproductive toxicity model that performed best in terms of F-measure was the estrogen receptor binding (ER) model

(0.71), but this model provided predictions for only eight compounds from the 91 RAR set chemicals.

In general, the statistics-based models performed the best. However, some of them (ER model and Toxtree QSAR model for carcinogenicity) had very narrow applicability domains and were only applicable to a limited set of compounds. Rule-based models performed noticeably worst.

Step 5

We were able to identify compounds that are commonly misclassified as toxic through all of the models (Figure 6). These were mostly small structures with low complexity, which contain a single benzene ring with a hydroxyl, amino, or nitro motif. Earlier research had already proposed that these compounds are problematic for QSAR models (Hillebrecht et al. 2011), which can be partially explained by the lack of structural alerts for non-genotoxic carcinogens in rule-based models (Benigni & Bossa 2008a).

5.2. Potential EDCs

The endocrine system is a collection of glands that regulates the secretion of almost all the hormones that influence the vital functions of an organism, such as growth and the development of the brain, nervous system, and reproductive system. The main endocrine glands consist of the hypothalamus, pineal gland, pituitary gland, thyroid gland, adrenal gland, pancreas, ovaries, and testicles. Any molecular initiating event can lead to an adverse outcome. EDCs can alter the functioning of the endocrine system, which may lead to an adverse effect (WHO/IPCS 2002). Substances can alter endocrine activity by binding to a hormone receptor or transport proteins, interfering with hormone synthesis, metabolism, and/or clearance, or by affecting the neuro-endocrine signalling involved in the regulation of endocrine function (EFSA 2013). Alterations of the endocrine system can cause an array of diseases, including obesity, hyper- or hypothyroidism, diabetes, and sex hormone disorders that reduce fertility.

The OECD proposed a suite of standardised *in vitro* test guidelines for the identification of EDCs (OECD, 2012) that include the estrogen-androgen-thyroid (EAT) and steroidogenesis pathways. Positive results from *in vivo* reproductive toxicity tests can also indicate endocrine disruption. More information about these tests can be found in the previous section (5.1. 'CMR properties').

The focus of paper II was to develop QSAR models that can identify potential EDCs that demonstrate an E, A or T mechanism (steps 1-3 in Figure 8). Within this paper we also investigated the use of a metabolite simulator (steps 4 and 6) and assessed the AD of the models (step 5). In this section only steps 1-3 are discussed. Steps 4 and 6 are discussed in section 5.4. 'Metabolism simulator', whereas step 5 is discussed in section 5.5. 'AD assessment'.

To aid the identification of potential EDCs, we collected *in vitro* data for EAT pathways (step 1 in Figure 8) and developed a set of QSAR models (step 3 in Figure 8) that could be used as first tier prioritisation tools. The models were then applied to HPVCs & LPVCs (paper II), and the E-pathway model was applied to over 32,000 chemicals that humans may be exposed to (paper III). Additionally, the EA models were remodelled using random forest combined with the CP method for the assessment of AD (paper IV) (more details in section 5.5. 'AD assessment').

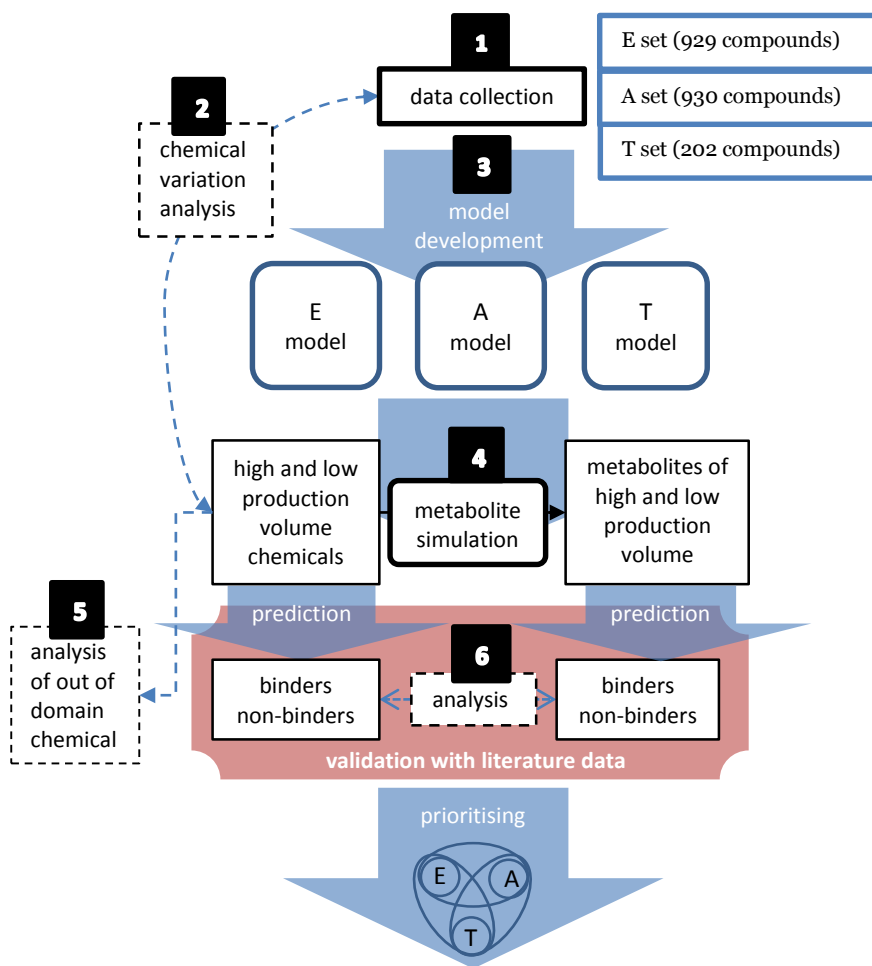


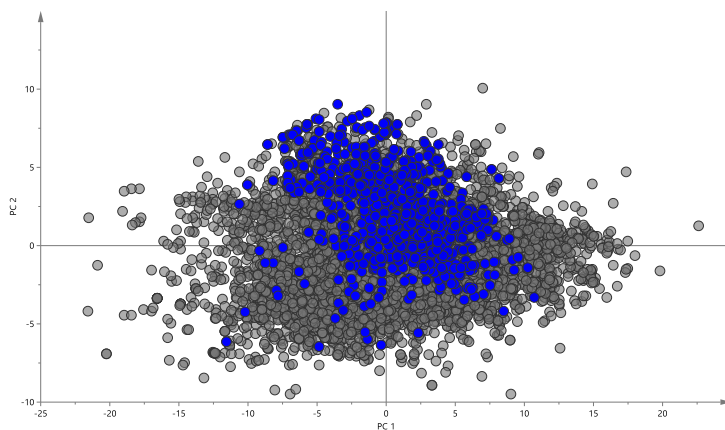
Figure 8: Workflow of the study presented in paper II. The numbers in black boxes represent steps in the study, which are described in detail in different sections of this thesis (see the main text for more details).

Steps 1-2

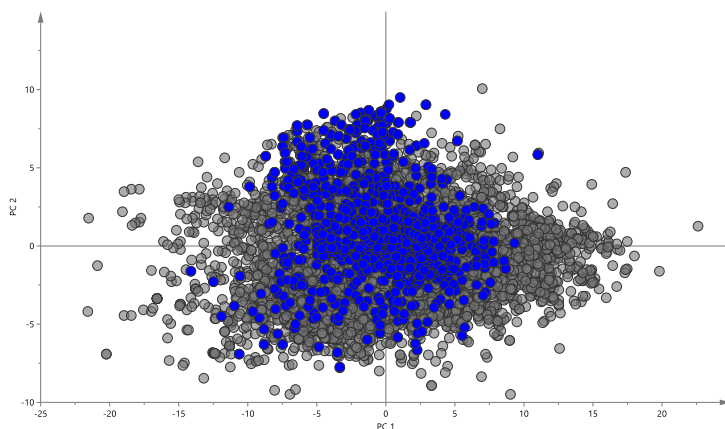
Data, which were collected from *in vitro* assays that measured binding to the estrogen receptor and transthyretin (carrier of the thyroid hormone – thyroxine) in human cells and to the androgen receptor in hamster cell (step 1), were plotted against HPVCs & LPVCs (step 2). This was similar to the process described in paper I (see previous section for more details). PCA clearly showed that compounds used to develop the E and A models (the E and A sets, respectively) represented a large share of the chemical variation

(Figure 9A-B), and that there were fewer examples of small polar compounds in the E set (left lower corner of Figure 9A). The T set, however, showed noticeably less variation and could be separated into two clusters: halogenated aromatic compounds (upper group in Figure 9C) and per- and poly-fluorinated compounds (lower group in Figure 9C).

A



B



C

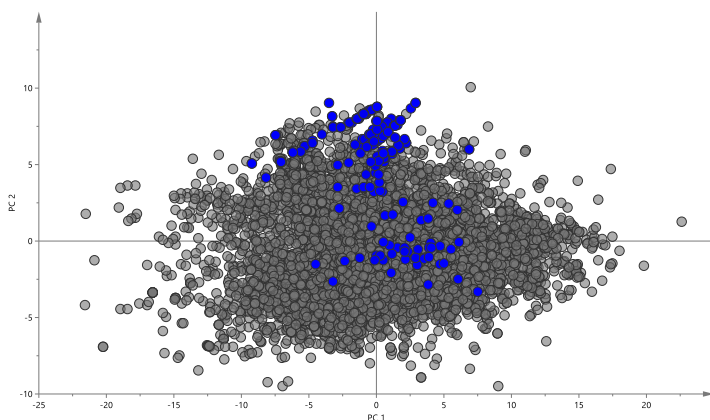


Figure 9: PCA plots of training and test sets (blue circles) for estrogen (A), androgen (B), and transthyretin (C) models, mapped with HPVCs & LPVCs (grey circles) as background. Principal components PC1 and PC2 explain 54% of the variance, and the model has a total of 9 PCs that explain 84% of the variance (Paper II).

Step 3

We developed a number of models that covered different descriptor sets and machine learning methods. These models were based on five descriptor sets, seven types of machine learning methods (both linear and non-linear), and two types of internal validation methods (cross-validation or bagging). Prior to modelling, the datasets were split into training and test sets in a ratio of 4:1.

The differences in balanced accuracy between the developed models spanned from 61 to 89%, but, apart from some of the linear models, that is, MLR and PLS, no large differences between the chosen descriptors and validation methods (cross-validation or bagging) were noticed. Similar observations were noted in both papers IV, in which the EA models included RF combined with CP method for AD assessment, and III, in which the E models were developed by different scientific groups and evaluated with experimental data for around 1500 chemicals. All of the E models from study III had comparable performances independent of the methods they used. These results indicate that model performance depends more on the quality of the collected data than the chosen machine learning method.

We chose the most accurate (highest balanced accuracy for the training set) and most cautious (lowest number of false negatives) models and applied these to the HPVCs and LPVCs (step 3 in Figure 8). These models were based on Dragon descriptors and were built with Associative Neural Networks (ASNNs) that combine an ensemble of feedforward neural networks and the k-nearest neighbour technique. As the T set comprised almost five times less compounds than the E and the A sets (Figure 9) it was not surprising that the predicted response of the T model included only 38% of the HPVCs and LPVCs (Figure 10). 9% of the HPVCs & LPVCs were predicted to be E and/or A binders (Figure 10A-B), whereas only 1% were potential T binders. This predicted proportion of E binders was similar to the result (8.2%) obtained from applying a consensus E model to a set of 32k chemicals (paper III).

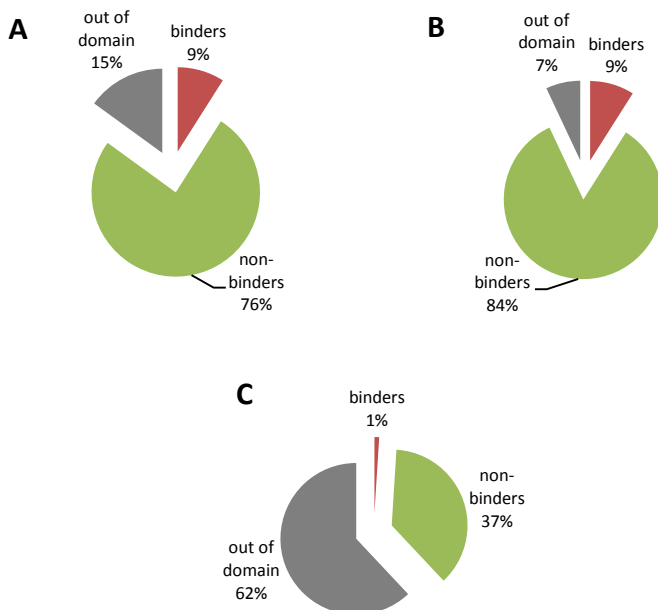


Figure 10: The percentages of high and low production volume chemicals that are predicted binders, non-binders and out of the model's applicability domain in the models for estrogen (A) and androgen (B) receptors, and transthyretin (C) binding.

According to the OECD principles on validation of QSAR models (OECD 2007a), the mechanistic interpretation of a model should be provided, if

possible. However, this requirement can be difficult to fulfil in the case of non-linear models, which are often considered black boxes. Furthermore, certain descriptors may be complicated and, hence, challenging to explain. To ease interpretation, we developed PLS models based on simple constitutional descriptors, related to functional group counts, for example, similar to Vorberg & Tetko (Vorberg & Tetko 2014). Although these models demonstrated a 4-9% lower balanced accuracy, they were valuable in characterising typical E, A, and T binders. Briefly, E and T binders mostly consist of lipophilic compounds with hydroxyl groups attached to an aromatic ring. Usually, they have numerous N and O atoms, which can serve as hydrogen bond donors (as found also by Papa et al. (Papa et al. 2013)). This finding indicates a possibility of polarisability, which was proven to have a significant impact on estrogenic activity (Liu et al. 2008). Similar to E and T binders, A binders are also lipophilic and contain aromatic rings; however, they do not contain hydroxyl groups but have characteristics rather atypical for E and T binders, such as nitro-groups, as well as aliphatic secondary and tertiary amines, as also observed by Jensen et al. (Jensen et al. 2011).

5.3. WoE approach and consensus modelling

The WoE approach can be described as the organisation of existing information for further use in additional tests and studies within the decision-making process (Rovida et al. 2015). This decision-making process commonly involves ITS. ITS combine available data in a WoE approach, and when there is a lack of experimental data for a query compound *in silico* tools may have a decisive impact on whether the compound should be prioritised for further testing. However, the application of individual models for prioritisation may not provide optimal results if each model has low accuracy or is too specific. This is because the various models may each describe slightly different aspects of hazard, for example, models based on *in vitro* data may measure different cell alterations and impairments. Furthermore, models based on *in vivo* data may not be able to identify a hazard due to species differences. This causes certain challenges; first, how can models that describe the same mechanism reach a consensus prediction (consensus modelling), and second, how can the predictions of models that measure the same response through different assay be best managed. The latter case poses a large problem for decision-making, which we tried to address in paper I. Consensus modelling was the subject of investigation in paper III and is also discussed in this section.

In paper I we wanted to identify simple ways for combining the outcomes of multiple models that predict a particular response to avoid potential false negatives and hence, prioritise certain chemicals for further testing. We combined the results from the C, M, and R models, and since the C models produced the most satisfactory results, only the results obtained from the C models are exemplified in this thesis (the combined results for M and R can be found in paper I). We plotted the number of predictions implying that a query compound is carcinogenic ('yes' answers) against the number of predictions implying that a chemical is non-carcinogenic ('no' answers) (Figure 11). The compounds that were experimentally measured to be non-carcinogenic (62 compounds) were plotted in Figure 11A, whereas the compounds that were experimentally measured to be carcinogenic (29 compounds) were plotted in Figure 11B. Every diamond represents one compound (or multiple compounds, depending on the size or shape of the diamond, see figure description). Non-carcinogenic chemicals received a maximum of 11 'yes' answers and 16 'no' answers, whereas carcinogenic received a maximum of 16 'yes' answers and 10 'no' answers. Clearly, none of the carcinogenic compounds were predicted to be positive by all models (Figure 11B). However, some of the non-carcinogenic compounds were predicted to be negative by all models (if they were not outside of the applicability domain) (Figure 11A). In both plots, we set a threshold for the

number of 'yes' and 'no' answers that would separate the regions where predictions were more unanimous and as a result, we were able to distinguish areas that could be defined as true negative (TN), true positive (TP), false positive (FP), and false negative (FN). After defining these areas, we noticed that many carcinogenic compounds were in the TP area and none were in the FN area (Figure 11B), which demonstrated that in such way the models had followed the precautionary principle. In the case of non-carcinogenic compounds, a majority were located in the TN area, and only seven were in the FP area (Figure 11A). However, a large share of chemicals landed in the 'grey zone' present between thresholds; these chemicals received an approximately equal number of 'yes' and 'no' answers, and hence, could be regarded as equivocal. Clearly, the hazard decision-making process for these chemicals requires more advanced ITS methods that consider more data.

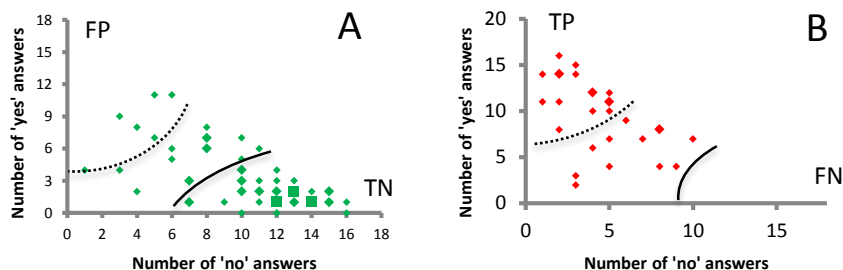


Figure 11: The number of 'yes' and 'no' answers regarding carcinogenicity provided by the C models, assessing the RAR set, for compounds that have been experimentally determined to be non-carcinogenic (A) or carcinogenic (B). A small diamond corresponds to a one chemical, a bigger diamond to 2-3 chemicals, and a square to 4-5 chemicals. The dotted line represents a threshold above which chemicals received at least 65% of 'yes' answers from all the models, whereas the normal line is a threshold under which chemicals received more 75% of 'no' answers.

A similar integration of predictions, which included an array of QSAR models that predict estrogen receptor binding, was covered in paper III. Multiple filtering steps were applied to improve the accuracy of individual models to reach consensus. The filters included, similar to paper I, setting a threshold for the number of 'yes' answers that confirm a chemical as a binder, as well as strict rules for defining a binder, which were based on a threshold for potency level, as it was found that binders that form a weak association with the estrogen receptor are the main source of discordance among different QSAR models. The addition of these filtering steps

increased the models' balanced accuracy (reaching about 90%) and the percentage of predicted binders (from a prediction set comprising 32,464 chemicals) from 8.2% to 12.3%. The filtering steps were included to decrease the number of potential false negatives, hence, they fulfilled the precautionary principle. On the other hand, the strict definition of a binder may have caused the models to miss certain weak binders. This may be unacceptable if the aim is to find any potential estrogen binders, irrespective of their potency. Allowing more disagreement between models increases the amount of chemicals that can be prioritised. Although this may reduce the number of potential FNs, it may also increase the number of FPs.

5.4. Metabolism simulator

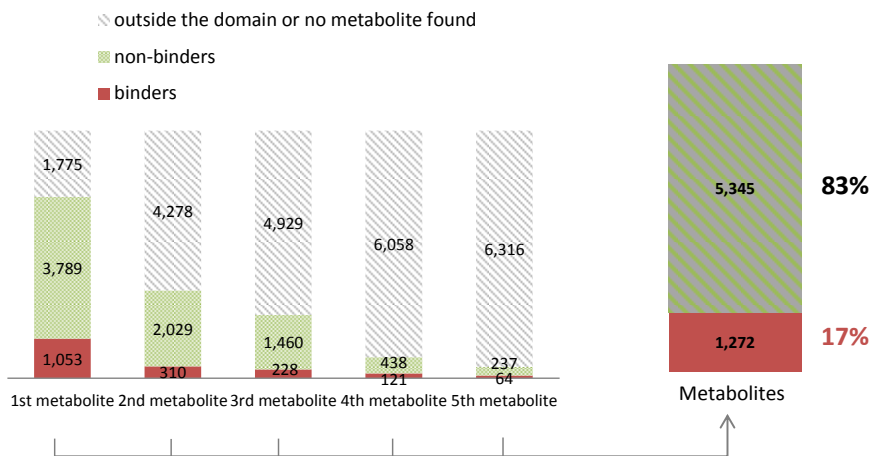
Various metabolite simulators, which cover a vast number of isoforms of the CYP P450 system and differ in accuracy (Zhou et al. 2006; Trunzer et al. 2009; G. Shin et al. 2011; T'jollyn et al. 2011; Liu et al. 2012; Sridhar et al. 2012), are available for use mainly in drug discovery and development. Moreover, the QSAR Toolbox (OECD 2015i), which is tailored to help companies and authorities assess the (eco)toxicity hazards of chemicals to be registered under REACH, includes the rat liver S9 metabolism simulator. Once potential metabolites have been predicted by these metabolism simulators, they can be further assessed for possible adverse effects, for example, through QSAR modelling.

We used the QSAR Toolbox metabolism simulator to generate metabolites of commonly identified FNs (paper I) (section 5.1. 'CMR properties'). CMR models were then applied to the metabolites. Out of 13 FNs, only one chemical (1,3-butadiene) produced a metabolite that was predicted to be mutagenic by several models. The rest of the metabolites were either non-mutagenic, like their parent compounds, or received equivocal results. We applied a recently developed ITS tool for mutagenicity (Vermeire et al. 2013) to these FNs. This tool is in agreement with REACH in terms of the amount of data that is required for chemicals of every tonnage that are subject to regulation. However, after this ITS tool had been applied the incorrect assessment persisted, which implies that only the bacterial mutagenic test (required for low tonnage (1-10 ton) chemicals, as stated in REACH) may be insufficient for prioritisation purposes. There were chemicals (phenol, benzene, aniline) among the frequent false negatives that, according to European Union Risk Assessment Reports (EC 2006; EC 2008a; EC 2004), were measured as negatives in bacterial tests but as positives (mutagenic) in mammalian tests. For example, 2-nitrotoluene is non-mutagenic *in vitro*, but mutagenic *in vivo*, which shows the importance of including information about potential toxic metabolites in a hazard assessment.

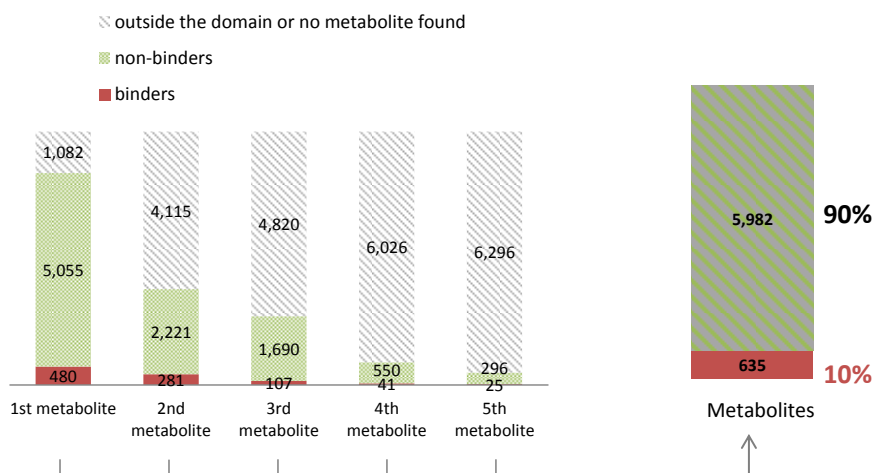
We studied the use of a metabolite simulator in more detail by investigating the significance of metabolism for endocrine disruption (paper II) (see section 5.2. 'Potential EDCs' for more details). A maximum of 5 major metabolites per compound were generated for the HPVC and LPVC set (Figure 12) by the liver consensus model in MetaSite (Discovery, 2014), which covers pathways mediated by the CYP2C9, CYP2D6, and CYP3A4 isoforms of the CYP P450 system. An evaluation of experimentally determined metabolites showed that MetaSite correctly simulated metabolites for 16 out of 32 chemicals (50% accuracy). The E, A, and T models were applied to every metabolite, as seen in Figures 12A, 12B, and

12C, respectively. The predictions were combined on the basis of a precautionary principle, which stated that if one metabolite of a parent compound was predicted to be active, then the compound was considered a bioactivated E, A or T binder. With this strategy, 17% of the metabolites within the HPVC & LPVC set were predicted to be E binders (Figure 12A), 10% to be A binders (Figure 12B), and 2% to be T binders (Figure 12C). When these results were compared to the results for the parent compounds (see section 5.2. 'Potential EDCs'), we found that the number of E binders had doubled, the number of A binders had slightly increased, and the number of T binders had even tripled (from 50 (1% of HPVC & LPVC set) to 165 (2%) chemicals (Figures 10 and 12)). We believe that the deamination of secondary amines was the reason for the large increase in the number of E binders. This reaction is typically catalysed by cytochrome P450 in phase I metabolism, and amines are atypical for E binders (see section 5.2 'Potential EDCs'). On the other hand, amines are a typical structural feature of A binders, and hence, the deamination leads to a deactivation of compounds, which may be the reason for the relatively small increase in the number of bioactivated A binders. As mentioned in the section 5.2 'Potential EDCs', the T set included a large proportion of polyhalogenated aromatics, and the hydroxylated aromatic ring is a typical structural feature of T binders. Therefore, hydroxylation, which is a common biotransformation pathway for aromatic compounds, was the major metabolic pathway behind the significant increase in the number of T binders. This transformation mechanism has already been proven to generate potential endocrine-disrupting metabolites from brominated aromatic flame retardants (Hamers et al. 2008).

A



B



C

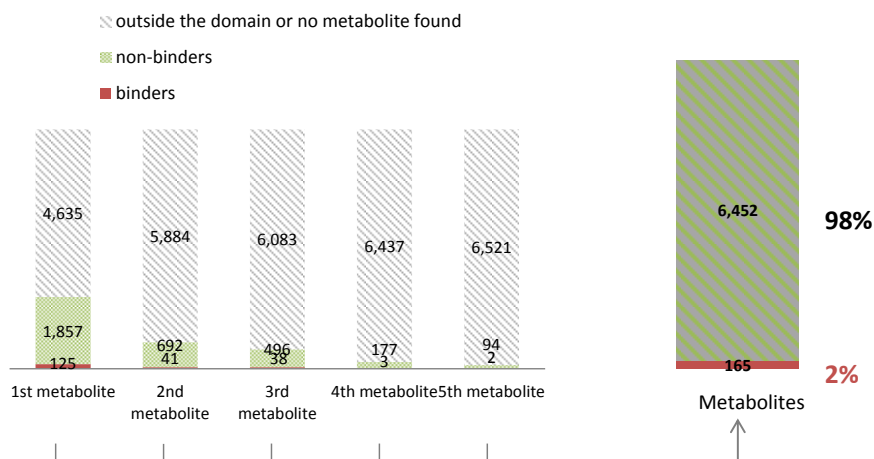


Figure 12: The ratios of predicted binders, non-binders, and compounds outside of the applicability domains for the estrogen (A), androgen (B), and transthyretin (C) models when tested against the 5 most commonly (according to MetaSite) formed metabolites of the HPVCs & LPVCs. The final numbers of compounds that were binders (red) and non-binders or outside of the applicability domain (green dots/grey stripes) are shown to the right of each graph (Paper II).

Although the accuracy of the metabolite simulator predictions was only 50%, we were able to identify several potential endocrine disruptors, for example hexachlorobenzene, 2,2',4,4',5-pentabromodiphenyl ether (BDE-99), 1,2-dichlorobenzene, and 4-(2,4-dichlorophenoxy)butyric acid. Initially, these

compounds were predicted to be non-binders by the T-model, but their correctly simulated metabolites (verified against experimental data) were identified as potential T binders. Earlier experiments have shown that these chemicals alter thyroxine levels *in vivo* (Berg et al. 1991; den Besten et al. 1991; Branchi et al. 2005).

5.5. AD assessment

The assessment of a QSAR model's application limits is one of the most important parts of guaranteeing the accuracy of provided predictions. A vast number of methods exist for this purpose, making it difficult to choose the most suitable approach, particularly when none of the methods seem ideal due to their unique advantages and limitations. In recent guidance (ECHA 2016b) for how to use and report QSARs, ECHA suggests checking a compound's similarity to the training set during AD assessment, that is, does the substance fall within the descriptor ranges of the model and does it have structural fragments or analogues that are represented in the training set. However, no details are given regarding how a compound's similarity should be assessed. Additionally, it has been advised that the mechanistic and metabolic domains of the compound should be investigated, i.e. does the target chemical show relevant mechanisms of action that are not covered by the model or should some related metabolites be considered. The latter aspect highlights the importance of considering a substance's ADME when using QSAR models, which was also discussed in the section 5.4. 'Metabolism simulator'. We faced AD issues in all of the studies included in this thesis and the problems we faced are discussed below.

The application limits of a QSAR model are inevitably related to the size of the training set; if the training set includes a large number of compounds with a high degree of chemical variation, then the model will also be applicable to a large number of compounds. This, however, does not imply that a model based on a low number of chemicals will be unsatisfactory. It can have high local performance, as was found in study I (section 5.1. 'CMR properties') for: 1) the estrogen receptor binding model from the QSAR toolbox used for identifying potential reproductive toxicants and 2) the aliphatic amine Toxtree QSAR model used for identifying potential mutagenic chemicals. Although the models were applicable to a low number of chemicals, they still provided highly accurate predictions, as is commonly noticed when the predictions from local and global QSAR models are compared (Yuan et al. 2007; Puzyn et al. 2011) and which is particularly important when using QSAR for a quantitative response.

The results of study I also emphasised the importance of defining the AD for rule-based models. These models identify structural alerts that are responsible for a particular effect, but, if there is an absence of an alert, a chemical is considered unknown rather than 'safe'. For this reason, rule-based models are not a reliable way to confirm a chemical's safety. On the other hand, in terms of prioritisation purposes these models may be of

higher value. However, in our study the rule-based models were found to have lower accuracy than the statistics-based models.

Papers II and IV, which studied different AD approaches, investigated compounds outside of the defined AD. In paper II, the AD of the EA models was assessed in PCA space using a distance measure (DModX), whereas in paper IV the AD was defined by the CP method. In the first method, chemicals with a DModX larger than the chosen threshold were considered different from the training set. The chemical characteristics of HPVCs & LPVCs defined as outside of the AD with this method can be split into four groups: large and hydrophobic structures with many single bonds; chemicals with many hydrogen bond acceptors, double bonds and rings, and often with sulphuric fragments; small and volatile structures such as acetylene; and phosphate-substituted amino acids. The CP method for the RF model identified similar structural characteristics as outside of the AD, which implies shared features of AD concepts. We believe that large, long-chained structures are a typical example of compounds outside of the descriptor ranges, whereas other chemicals can be members of empty regions within the interpolation space of the training set. Hence, the latter chemicals may represent certain missing chemical structures that the models (ASNN and RF in papers II and IV, respectively) could not find the neighbours, i.e. analogues, of.

Among the compounds that were often defined as outside of the model's AD in paper IV were aromatic structures that include iodine. Tiraticol (Figure 13) was within the AD of all remodelled E models (listed in Table 1) and was predicted correctly to be an E-binder. On the other hand, thyropropic acid (measured E-binder), which differs from tiraticol by one carbon atom (Figure 13), was defined as outside of the AD. These examples show that a small change in the chemical structure can lead to a prediction outside of the AD which is considered of lower reliability.

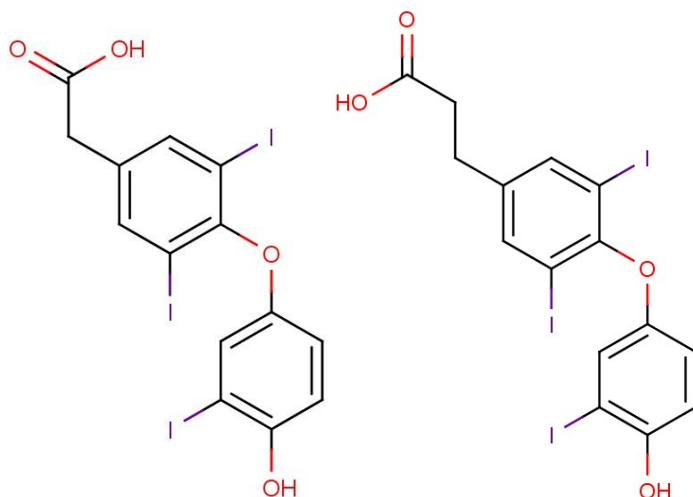


Figure 13: The structures of tiratricol (3,3',5-triiodothyroacetic acid) (left) and thyropropic acid (3,3',5-triiodothyropropionic acid) (right).

In paper III, we investigated various E models developed by several research groups. These models had their AD defined by different distance-based approaches when applied to a set of more than 32k chemicals. Regardless of the chosen AD method the balanced accuracy did not increase after removing predictions outside of the AD, which is in agreement with observations of Sahigara et al. (Sahigara et al. 2012) that predictions outside of the AD are not necessarily less accurate than those within the limits.

The CP method was applied to the EA models (built in paper II, see section 5.2. 'Potential EDCs') and added confidence metrics to provide better insights about the reliability of predictions and quality of the models (paper IV). In CP, the user determines a significance level for the validity of a prediction. CP categorises compounds as 'empty' (error) or 'both' (equivocal), depending on the chosen significance level. Hence, these chemicals migrate between these two classes and often share similar structural characteristics. The models were found to have varying accuracies depending on the chosen significance level, which was expected (Tables 1 and 2). A higher significance level implies that the rules for defining the similarity of a query compound to one in the training set are stricter. Therefore, setting the significance level to 0.25 led to more compounds being defined as 'empty' in both the E and A models (Tables 1 and 2) than when the significance level was set to 0.20. Furthermore, narrowing the predictive boundaries (higher significance level) of the E models increased accuracy (sensitivity and specificity) (Table 1). For the A models, however, a higher

significance level did not always lead to higher accuracy, which also depended on the descriptor set used (Table 2). The results from the A models raised questions about setting an appropriate significance level when the AD is defined. This decision also depends on the purpose of the model. If the purpose is risk assessment, then the model should be more cautious (sensitive); hence, a higher significance level during AD determination, which should improve accuracy, seems reasonable. On the other hand, if the QSAR is a prioritisation tool, then the model should be applicable to a larger number of chemicals; hence, the significance level can be decreased to a point where the accuracy of the model is still satisfactory. This reasoning fits well with the results from the E models (Table 1). However, based on the results from the A models, choosing the optimal significance level can be quite complicated.

Table 1: Statistics for the model based on the estrogen dataset

set	sign-level	validity	%cmpds in class 'empty'	%cmpds in class 'both'	sensitivity	specificity	precision
Training ¹	0.20	0.81	5.57	0.14	0.85	0.86	0.86
Training ¹	0.25	0.77	14.01	0	0.87	0.92	0.91
Training ²	0.20	0.80	8.55	0	0.86	0.89	0.88
Training ²	0.25	0.76	15.72	0	0.89	0.91	0.90
Training ³	0.20	0.81	6.35	0	0.88	0.86	0.85
Training ³	0.25	0.75	15.31	0	0.89	0.88	0.87
Test ¹	0.20	0.84	5.41	0	0.92	0.87	0.87
Test ¹	0.25	0.80	12.97	0	0.95	0.89	0.89
Test ²	0.20	0.84	5.95	0	0.94	0.86	0.86
Test ²	0.25	0.81	11.35	0	0.96	0.87	0.87
Test ³	0.20	0.87	4.30	0	0.93	0.88	0.88
Test ³	0.25	0.82	9.68	0	0.94	0.89	0.88

¹dragon descriptors were used in the modelling

²signatures descriptors were used in the modelling

³rdkit descriptors were used in the modelling

Table 2: Statistics for the model based on the androgen dataset

set	sign-level	validity	%cmpds in class 'empty'	%cmpds in class 'both'	sensitivity	specificity	precision
Training ¹	0.20	0.81	0	9.81	0.79	0.79	0.65
Training ¹	0.25	0.77	2.02	0.67	0.78	0.79	0.63
Training ²	0.20	0.81	0.14	3.72	0.82	0.79	0.65
Training ²	0.25	0.76	6.06	0	0.84	0.79	0.65
Training ³	0.20	0.80	0	4.55	0.82	0.78	0.63
Training ³	0.25	0.76	5.10	0.14	0.83	0.78	0.63
Test ¹	0.20	0.82	0	8.11	0.79	0.81	0.64
Test ¹	0.25	0.76	3.24	0	0.80	0.78	0.61
Test ²	0.20	0.81	0	1.62	0.81	0.81	0.66
Test ²	0.25	0.76	5.41	0	0.80	0.81	0.66
Test ³	0.20	0.82	0	3.23	0.84	0.81	0.66
Test ³	0.25	0.76	7.53	0	0.85	0.81	0.68

¹dragon descriptors were used in the modelling

²signatures descriptors were used in the modelling

³rdkit descriptors were used in the modelling

The E and A models classified chemicals into 'empty' and 'both' groups differently. The E models classified more compounds as 'empty' when the significance level grew and did not classify almost any compounds as 'both', irrespective of significance level (Table 1). On the other hand, the A models classified chemicals as 'empty' when the significance level was high (0.25) and as 'both' when the significance level was low (0.2) (Table 2). We believe that this difference is related to the confidence of the models to define binders and also indicates that the A set includes a more diverse group of chemicals, as seen in Figure 9 in section 5.2. 'Potential EDCs'. The A models are less certain in defining A binders, shown by lower precision and a higher number of 'both' compounds when a lower significance level was chosen (Table 2). In contrast, the E models seem to distinguish well between binders and non-binders, and the significance level has a noticeable effect on the size of the AD space.

6. Exposure assessment of pharmaceuticals – environmental fate at a WWTP

Chemicals emitted from products and materials used by humans, e.g. in private households, public buildings, and industry, may reach WWTPs through direct disposal or via dust. Information about the fate of chemicals at a WWTP provides essential information about the concentrations of chemicals that enter the aquatic system. A good understanding of the wastewater treatment process is an important element of assessing the exposure to a particular chemical in the environment. The removal of a chemical from wastewater occurs mainly through three processes: biodegradation, adsorption to sludge, and volatilisation (Meent & Bruijn 2007). However, some WWTPs use additional treatment methods, such as ozonation (Ternes et al. 2004).

Sorption is characterised by the distribution coefficient (K_D), which describes the partitioning of a particular chemical between water and solids. The sorption of organic substances to solids, such as soil, is mainly driven by hydrophobic interactions; hence, a substance with a higher $\log K_{ow}$ will be more attracted to solids. Other molecular interactions may also occur between a substance and a solid, for example, hydrogen bonding, van der Waals forces, and/or ion-dipole interactions (Yaron & Saltzman 1978; Bailey et al. 1968; Tülp et al. 2009). The latter interactions are particularly important for pharmaceuticals, which are often acids, bases, or ampholytes, and thus, may be present in ionised form depending on the pH of the environment, a factor that can significantly affect their fate and toxicity.

In paper V, we studied how a pharmaceutical compound's ionisation state affects sludge-water partitioning. The study focused on: 1) the importance of hydrophobic bonding as the main mechanism driving sorption to sludge; 2) developing separate QSARs for chemicals that are neutral, positively-, and negatively-charged at an experimental pH; 3) assessing if descriptors calculated for ionised structures improve model performance compared to the typical modelling situation, in which the uncharged structures are used as inputs, and 4) using QSARs to characterise the types of interactions that occur between neutral, positively-, and negatively-charged chemicals and sludge.

We collected the distribution coefficients (K_D) for secondary treatment sludge for 110 pharmaceutical compounds at a pH ranging from 6.7 to 7.6.

After plotting chemicals according to their $\log K_D$ and $\log K_{OW}$ values (Figure 14), we found that a compound's degree of hydrophobicity drove partitioning to sludge, and that many of the hydrophobic compounds had a low number of hydrogen bond acceptors per atom. However, compounds with higher number of acceptors, for example, tetracycline and pipemidic acid, had higher $\log K_D$ values than were expected based on their $\log K_{OW}$, which indicates that they may interact with sludge via hydrogen bonding. A group of chemicals with $\log K_{OW}$ values ranging from 3 to 6 seemed to deviate from the main trend (Figure 14). These chemicals were mostly derivatives of acetic and propionic acid and were predominantly negatively charged at pH 7. Since sludge is generally considered to have an overall negative charge (Hyland et al. 2012), ionic repulsion could explain why the sorption of these compounds was lower than expected from their $\log K_{OW}$ value.

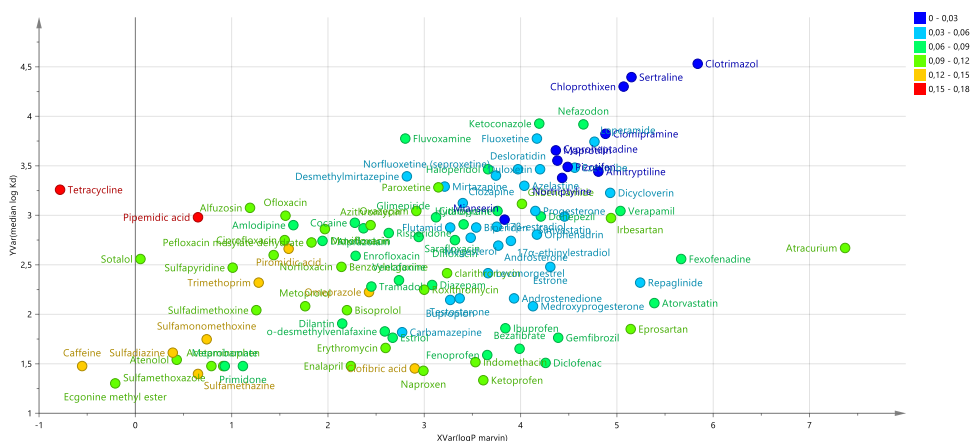


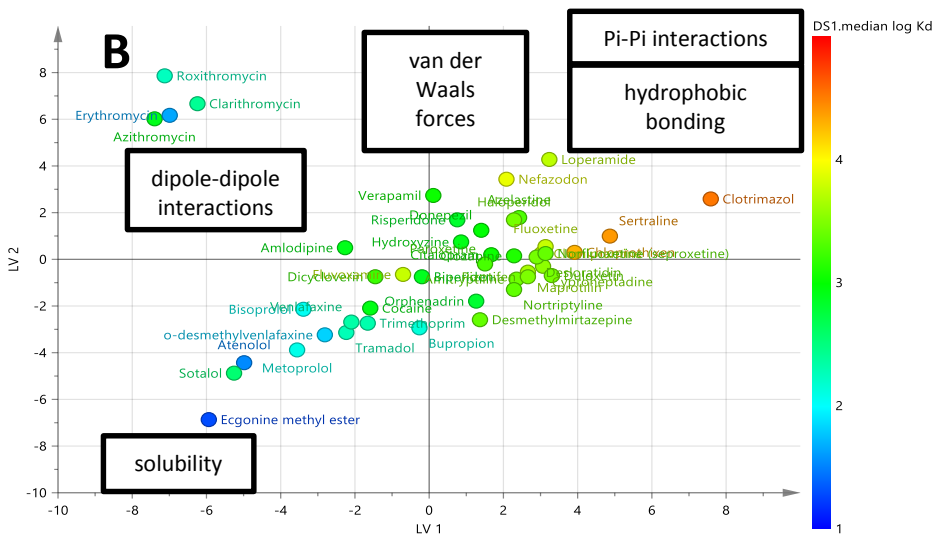
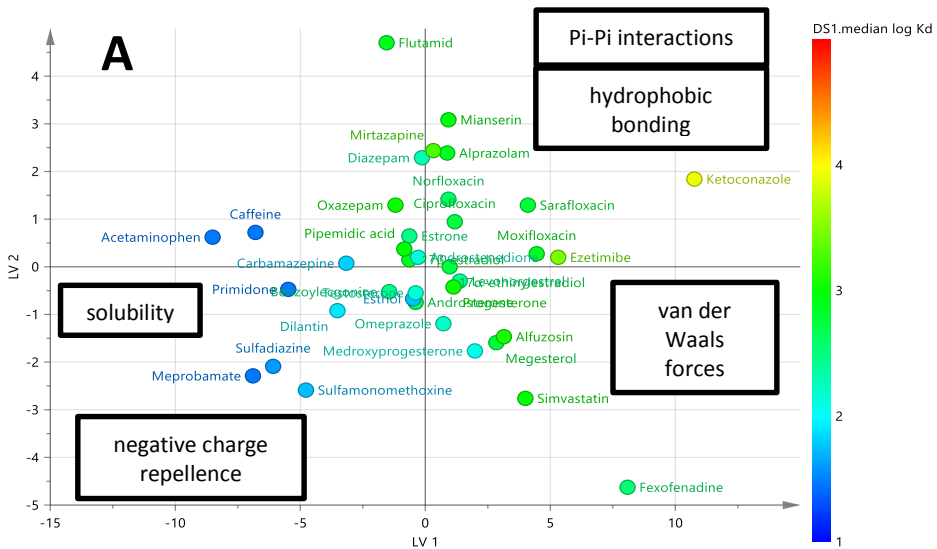
Figure 14: Visualisation of the $\log K_D$ to $\log K_{OW}$ correlation (calculated by Marvin software (Chemaxon, Budapest)) of the studied chemicals. The compounds are coloured according to their number of hydrogen bond acceptors per atom (paper V).

The approach presented above included only hydrophobicity and the number of atoms that could participate in hydrogen bonding; hence, it was not sufficient for characterising all of the chemical variation of the studied compounds. For this reason, a multivariate approach was used to investigate other interactions. Chemicals were split according to their dominant charge at the experimental pH (pH 7 was used as the threshold), which resulted in 36 neutral, 45 positively-, and 28 negatively-charged compounds. Four models were created for every group. These models differed in terms of descriptor types and the number of chemicals included (certain chemicals were moved to other groups because their pKa value was close to the

experimental pH, and hence, they were present in two or more forms). The best performing models had R^2Y values between 0.75-0.77 and Q^2 values between 0.65-0.73 for all three groups. These models outperformed models for positively- and negatively-charged chemicals built by Sathyamoorthy & Ramsburg (Sathyamoorthy & Ramsburg 2013) and also included more compounds.

We found that using descriptors calculated for ionised structures (instead of uncharged structures) did not improve the performance of models for positively- and negatively-charged chemicals. However, neutral models showed a 5% improvement in performance when descriptors representing ionised structures were included, which indicates the importance of describing ampholytes accurately.

We highlighted the primary interactions occurring between each of the three groups and sludge based on the PLS loading plots used in the modelling (Figure 15). All of the chemicals could interact with sludge via van der Waals forces and Pi-Pi interactions. Apart from these interactions, hydrophobic bonding was the primary mechanism driving the sorption of neutral and positively-charged chemicals (Figure 15A-B). Negatively-charged chemicals, on the other hand, interacting with sludge in a different way. These chemicals mainly interacted with sludge through covalent bonding, as well as ion-ion and dipole-dipole interactions, while hydrophobic bonding was of little importance (Figure 15C). We expected to identify an ionic attraction between the positively-charged chemicals and the negatively-charged sludge, but, surprisingly, no descriptors related to ion-ion interactions were present in the final model. On the contrary, neutral compounds showed a degree of negative charge repulsion (Figure 15A), which can be justified by the presence of zwitterions.



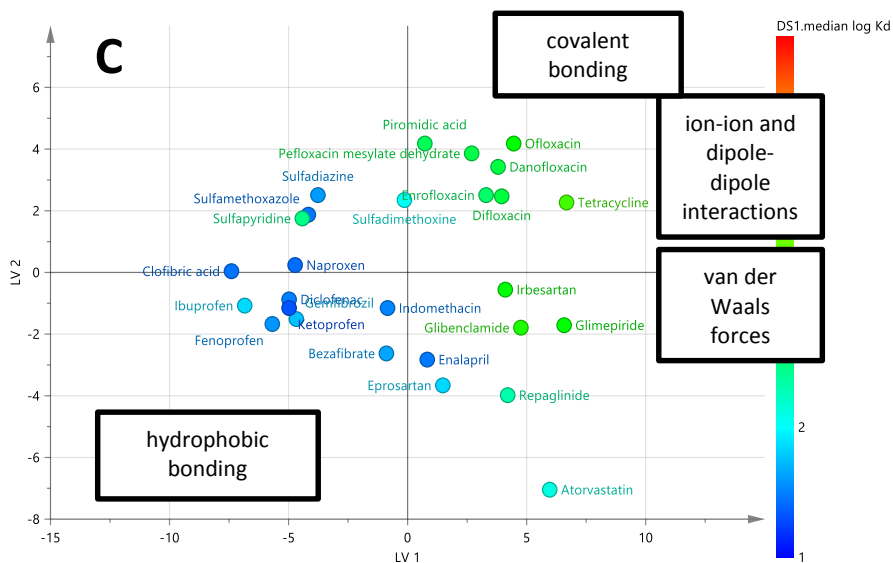


Figure 15: Score plots for the final PLS models for neutral (A), positively- (B), and negatively- (C) charged chemicals, developed in paper V. The most significant interactions between each group of chemicals and sludge, derived from the most significant descriptors in the loading plots, are shown in the boxes. Compounds are coloured according to the value of the response (median value of log_D).

7. Conclusions and future prospects

The research presented in **paper I** focused on evaluating the use of QSARs as a 1st tier tool for the prioritisation of potential carcinogens, mutagens and reproductive toxicants for further testing. It was concluded that the available models cannot correctly predict reproductive toxicity because the underlying biological and chemical mechanisms are too complex. On the other hand, when QSARs were applied to carcinogenicity, even a single model could reach a high hit rate in identifying carcinogens. If QSARs are applied to mutagenicity, we recommend combining the results from a few models, as the majority of the models are based on *in vitro* data that can stem from different species and/or describe slightly different molecular events. The Ames test, which is based on bacterial cells and demanded by REACH for all chemicals produced in quantities of more than 1 ton, performed well. However, we found that the assessment of some chemicals through the Ames test alone may be insufficient, as certain chemicals were negative in *in vitro* bacterial tests but positive in *in vitro* mammalian tests or *in vivo* tests. Also, special attention must be given to small, simple organic chemicals, for example, those containing a benzene ring with a hydroxyl motif, as they could not be identified as toxic by any of QSAR models, or the ITS tool. Overall, the array of QSAR models was useful in identifying potential carcinogens and mutagens; however, the evaluation of common false negatives revealed the limitations of the available tools and the need to include information about a chemical's metabolism in modern ITS tools.

In **paper II**, we developed various models that could be used for identifying potential EDCs during 1st tier screening based on alterations relating to binding to transthyretin and the estrogen and androgen receptors. The performance of these models was verified through statistical tests and was concluded high. The models are also easily accessible for use by authorities, producers, importers, the scientific community, and other interested parties. The EA models, in particular, can be applied to a wide range of HPVCs & LPVCs. **Paper II** also presented results that highlighted an important role for metabolism in the prioritisation process, particularly for chemicals demonstrating E- and T- driven alterations in the endocrine system. At least one metabolite was predicted correctly in 50% of the cases when a metabolism simulator was used for predicting phase I metabolites of HPVCs & LPVCs. Even though the accuracy of the simulations was not high, when the predictions were combined with predictions from EAT models the metabolism simulator was found to be useful in identifying chemicals that demonstrate *in vivo* activity related to E- and T- binding. Overall, these findings reveal the limitations of currently available metabolite simulators,

especially for industrial chemicals, which were predicted noticeably worse than pharmaceuticals.

Paper III presented the results of a combination of predictions from a range of estrogen receptor binding QSAR models that had been developed with various datasets and machine learning methods. When the results of this paper were compared to those from **paper I**, it was shown that consensus modelling provides a better overview of a chemical's potential toxicity. When the outcomes from all available models are integrated in an intelligent manner, for example, by setting an appropriate threshold for the number of toxic/non-toxic responses, the accuracy of predictions can increase. The focus can be either a decrease in the number of false negatives (if the precautionary principle needs to be considered) or a decrease in the number of false positives, depending on the application of such a consensus model.

Papers II and IV investigated AD approaches for models that include a set of chemical compounds with a high degree of chemical variation, for example, EA models for HPVCs & LPVCs. Furthermore, these papers included characterisations of chemicals that were defined as outside of the AD. Both methods for assessing the AD (distance-based and CP) identified similar chemical properties. The CP method, however, allowed for a better understanding of the models in terms of data quality and the chemical variation within the training set. Because CP combines the distance of a chemical to the training set and confidence metrics, this method, when compared to other AD methods, makes it easier to choose an appropriate threshold for reliable outcomes.

Paper V focused on using QSAR tools to analyse the sorption of various compounds to sludge in a WWTP. The results demonstrated that various compounds interact through different mechanisms with sludge. Multivariate models built for neutral, positively- and negatively charged compounds in pH 7 reached high statistical significance and were able to characterize the predominant chemical-sludge interactions for each group. Neutral and positively-charged compounds shared a typical hydrophobicity-driven mechanism, but also demonstrated Pi-Pi and dipole-dipole interactions with the sludge surface. In contrast, the negatively-charged compounds predominantly interacted with the sludge via covalent bonding and ion-ion, ion-dipole, and dipole-dipole forces. The study demonstrated that molecules interact differently with sludge depending on their charge, and hence, chemicals should be modelled separately, especially when they exist in a negatively-charged state. On the other hand, the use of descriptors calculated for ionized structures did not improve the performance of a model when the

overall charge of a molecule was negative or positive. Such descriptors were, however, important in describing neutral ampholytes.

The research underlying this thesis has suggested that it is impossible to predict complex endpoints, such as reproductive toxicity, directly from the chemical structure. However, computational tools can be helpful in elucidating the steps that lead to a specific adverse outcome, for example, by identifying a molecular initiating event (such as estrogen binding) and/or toxic metabolites. Therefore, it is my belief that future work should focus on combining predictions from computational tools that model metabolism, molecular initiating events, and other key events at the subcellular, cellular, tissue, or even organ level. These events can then be used to characterise the adverse outcome pathway (AOP), as proposed by Patlewicz & Fitzpatrick (Patlewicz & Fitzpatrick 2016), and as part of ITS. For this reason, specific assays need to be developed, or mapped, if they already exist, for revealing the key events within the AOP. These assays can be then used to develop models, as was recently done by Cox et al. (Cox et al. 2014), who derived endocrine prediction models from a set of high-throughput screening assays. There are plans to phase out *in vivo* testing, so future work will inevitably focus on predicting *in vivo* effects through *in vitro* tools. Patlewicz et al. have already discussed how an AOP could involve QSARs to predict E binding (estrogen receptor binding assay) and E-induced *in vitro* cellular response (estrogen receptor transactivation assay), which could replace *in vivo* test measuring estrogenic activity (uterotrophic assay) and eventually *in vivo* test regarding adverse effects (mammalian two generation assay) (Patlewicz et al. 2015).

There is a large potential for using computational tools to identify potentially toxic chemicals, but it is mainly limited by the availability of high-quality data. From a methodological point of view, I believe that QSARs provide sufficiently good results to replace assays that are used for model building. It is, however, important that the models are accurate and have well-defined applicability domains, as currently, this seems to be the main flaw of QSAR modelling. If confidence metrics were provided for every step in a AOP that could be used in ITS, then the uncertainty present in the final judgment of a compound's toxicity could be better defined. A recently developed ITS tool for skin sensitisation (Jaworska et al. 2015), which involves a combination of *in vitro* assays and *in silico* tools to predict bioavailability-related properties and metabolism and also provides the probability for every step through Bayesian networks, has a bright future and could help introduce similar approaches for other responses.

Recent efforts show a new trend in the application of QSARs to hazard and exposure assessments that includes the mechanistic explanation of the modelling property or activity, as well as the introduction of metrics to determine the uncertainty involved in applying the model to new compounds. I think the understanding of QSARs methodology can guarantee their future utility as commonly used tools for predicting the fate, toxicity, and ecotoxicity of emerging chemicals. The improving performance and understanding of QSAR models can increase the awareness and credibility of non-testing tools, so that one day the scientific community, regulatory bodies and industry can reach an agreement for how to utilise QSARs instead of animal testing for any type of toxic response and live happily ever after.

Acknowledgements

Welcome to the most readable part of the thesis! On this occasion I would like to thank you for taking the effort to read the thesis from the beginning all through to the end. If you haven't done it yet, I greatly encourage you to do so! Reaching this part after going through all the pages is much more satisfying.

The first thanks I would like to give to Patrik Andersson. Thank you Patrik for all those evenings you spent on reading my poor writing which I did the night before. I hope my more than four-year experience as a PhD student made your reading of my thesis easier. Thank you as well for showing the way in doing research and all the other support that you gave me during this time. While we are in this paragraph I'll use the opportunity to thank also for all the funding that you got, so that this position could happen. Thanks Swedish Research Council! I am also grateful for the travelling support from the Kempe foundations and the Wallenberg foundations.

I would like to thank Christina Rudén for support especially in the first part of my PhD studies and Anna Linusson for help on the thesis. I would also like to thank Ulf Norinder for all the input, it was great to do research with you! Thanks also go to the other side of the ocean, Kamel Mansouri and Richard Judson, it was a big pleasure to be a part of your ambitious project. I would also like to thank all the people in the Environmental Chemistry group for the inspiring environment and help, as well as the administration and other workers at KBC in Umeå who made my work easier.

I would like to thank Igor Tetko for the project in Munich and help and also all the people from Helmholtz Centre in Munich for their scientific input as well as for making my project stay memorable. Special thanks go to Elena Salmina, Sasha Safanyaev, Milan Voršilák, Roman Schneidermann, Eva Schlosser and Stefan Brandmaier.

I would like to thank the physics group for their inspiring lunch discussions and both entertaining and useful scientific input. Big thanks to Johan and Daniel Zakrisson, Narges Mortezaei, Pontus Svenmarker, Thomas Hausmaninger, Daniel Vågberg and Isak Silander.

I would also like to thank the following people: Malin Larsson and Eddie Wadbro for scientific input, support and just entertaining moments; Niklas Blomqvist for pretty much EVERYTHING including language corrections for

both languages, tips and mental support; Bobby Norgren for taking care of my physical health.

I would like to thank all the former and current members of the Polish “mafia” in Umeå for all the help in keeping my mental balance. Dzięki kochani za Wasze wsparcie i za te niezapomniane imprezy, które pozwoliły mi utrzymać równowagę psychiczną ;) Dziękuję Wam wszystkim razem i każdemu z osobna: Marta Andruszkiewicz, Basia Brzozowska, Magda Chelińska, Paulina Czichos, Piotr Jabłoński, Michał Janasik, Michał Kluczek, Michał Mazurkiewicz, Agata Pawlak, Artur Skwarek, Anna Strzelczyk i Michał Zglejc.

Honourable mentions. Piotr Rybacki oraz Alicja już teraz Rybacka ;) Wam dziękuję za to, że byliście zawsze, kiedy wracałam do domu. Anna Newlaczyl. Aniuś, mało tu byłaś, ale też Ci dziękuję za to, że jednak byłaś. Oraz Piotr Ciepłiński. Dziękuję Tobie Dzióbek za to, że byłeś przez cały ten czas.

Last and most importantly, I would like to thank my beloved parents and this part I'll translate so that not only my parents will know how great they are.

Beloved parents, thank you for all the support during these years and faith in me defending. There are no words that are able to express my gratitude, but I'll do my best. Mom, thank you for constantly taking care of me even though I'm an adult now and for motivating me in going up the ladder of scientific career. Dad, thank you for reminding me how important it is to follow your dreams and that life sometimes is a joke that you just have to laugh at and go forward. Thank you both for all you did for me, for your time and effort, for teaching me rules and values and shown love, which made me the person I am today. I hope you're proud. I love you. ♪ ♪♪...*some emotional melody in the background...* ♪ ♪♪

Kochani rodzice, dziękuję za Wasze wsparcie przez te lata i wiarę w to, że mi się uda obronić. Nie istnieją słowa, które są w stanie wyrazić moją wdzięczność, ale się postaram. Mamo, dziękuję za Twoje nieprzerwane sprawowanie pieczy nade mną pomimo tego, że jestem już dorosłą osobą oraz za motywację we wspinianiu się po drabinie kariery naukowej. Tato, dziękuję za przypomnianie mi jak ważne jest podążanie za marzeniami oraz, że życie to czasem żart, z którego trzeba się po prostu śmiać i iść dalej. Dziękuję za wszystko, co dla mnie zrobiliście, za Wasz czas i trud, za wpojone zasady i wartości oraz okazywaną miłość, co sprawiło, że dzisiaj jestem taką osobą, a nie inną. Mam nadzieję, że jesteście dumni. Kocham Was. ♪ ♪♪... *jakaś wzruszająca melodia w tle...* ♪ ♪♪

References

- Adler, S. et al., 2011. Alternative (non-animal) methods for cosmetics testing: current status and future prospects---2010. *Archives of Toxicology*, 85(5), pp.367–485.
- ANTARES, 2011. Alternative Non-Testing methods Assessed for REACH Substances.
- Ashby, J. & Tennant, R.W., 1991. Definitive relationships among chemical structure, carcinogenicity and mutagenicity for 301 chemicals tested by the U.S. NTP. *Mutation Research/Reviews in Genetic Toxicology*, 257(3), pp.229–306.
- Bailey, G.W., White, J.L. & Rothberg, T., 1968. Adsorption of Organic Herbicides by Montmorillonite: Role of pH and Chemical Character of Adsorbate. *Soil Science Society of America Journal*, 32, pp.222–234.
- Balls, M. et al., 2006. The Principles of Weight of Evidence Validation of Test Methods and Testing Strategies: The Report and Recommendations of ECVAM Workshop 58. *Alternatives to laboratory animals : ATLA*, 34(6), pp.603–620.
- Benigni, R. & Bossa, C., 2008a. Predictivity and Reliability of QSAR Models: The Case of Mutagens and Carcinogens. *Toxicology Mechanisms and Methods*, 18(2-3), pp.137–147.
- Benigni, R. & Bossa, C., 2008b. Structure alerts for carcinogenicity, and the Salmonella assay system: a novel insight through the chemical relational databases technology. *Mutation research*, 659(3), pp.248–61.
- Berg, K.J. et al., 1991. Interactions of halogenated industrial chemicals with transthyretin and effects on thyroid hormone levels in vivo. *Archives of Toxicology*, 65(1), pp.15–19.
- den Besten, C. et al., 1991. The liver, kidney, and thyroid toxicity of chlorinated benzenes. *Toxicology and Applied Pharmacology*, 111(1), pp.69–81.
- Boxall, A.B.A. et al., 2012. Pharmaceuticals and Personal Care Products in the Environment: What Are the Big Questions? *Environmental Health Perspectives*, 120(9), pp.1221–1229.
- Boxall, A.B.A., 2004. The environmental side effects of medication. *EMBO Reports*, 5(12), pp.1110–1116.
- Branchi, I. et al., 2005. Early developmental exposure to BDE 99 or Aroclor 1254 affects neurobehavioural profile: interference from the administration route. *Neurotoxicology*, 26(2), pp.183–92.
- Breiman, L., 1996. Bagging predictors. *Machine Learning*, 24(2), pp.123–

140.

- Cassano, A. et al., 2010. CAESAR models for developmental toxicity. *Chemistry Central Journal*, 4(1), pp.1–11.
- Chemsec, 2015. SIN List. Available at: <http://chemsec.org/what-we-do/sin-list> [Accessed February 2, 2016].
- CHMP, 2006. Guideline on the environmental risk assessment of medicinal products for human use.
- Chohan, K.K., Paine, S.W. & Waters, N.J., 2008. Advancements in Predictive In Silico Models for ADME. *Current Chemical Biology*, 2(3), pp.215–228.
- Christensen, F.M. et al., 2011. European Experience in Chemicals Management: Integrating Science into Policy. *Environmental Science & Technology*, 45(1), pp.80–89.
- CLP, 2008. Regulation (EC) No 1272/2008.
- Consonni, V. & Todeschini, R., 2009. Molecular descriptors. In *Recent Advances in QSAR Studies: Methods and Applications*. Springer Verlag, pp. 29–102.
- Cox, L.A. et al., 2014. Applying a scientific confidence framework to a HTS-derived prediction model for endocrine endpoints: lessons learned from a case study. *Reg. Toxicol. Pharmacol*, 69, pp.443–450.
- Cronin, M.T.D. & Worth, A.P., 2008. (Q)SARs for Predicting Effects Relating to Reproductive Toxicity. *QSAR & Combinatorial Science*, 27(1), pp.91–100.
- CVMP, 2004. Guideline on environmental impact assessment for veterinary medicinal products.
- Dimitrov, S. et al., 2005. A Stepwise Approach for Defining the Applicability Domain of SAR and QSAR Models. *Journal of Chemical Information and Modeling*, 45(4), pp.839–849.
- EC, 2004. *European Union Risk Assessment Report: aniline*. EUR 21092 EN.
- EC, 2006. *European Union Risk Assessment Report: phenol*. EUR 22522 EN/2.
- EC, 2008a. *European Union Risk Assessment Report: benzene*. EUR 19011 EN.
- EC, 2008b. Regulation (EC) No 1333/2008.
- EC, 2009a. Regulation (EC) No 1107/2009.
- EC, 2009b. Regulation (EC) No 1223/2009.

- EC, 2012. Regulation (EC) No 528/2012.
- EC, 2016a. Candidate list of suspected endocrine disruptors. Available at: http://ec.europa.eu/environment/chemicals/endocrine/strategy/substances_en.htm [Accessed February 3, 2016].
- EC, 2016b. Endocrine Disruptors. Available at: http://ec.europa.eu/environment/chemicals/endocrine/index_en.htm [Accessed February 3, 2016].
- EC, 2016c. Eudralex, The Rules Governing Medicinal Products in the European Union.
- ECHA, 2011a. *Guidance on information requirements and chemical safety assessment Part B: Hazard assessment.*
- ECHA, 2011b. *The Use of Alternatives to Testing on Animals for the REACH Regulation,*
- ECHA, 2012a. *Guidance on information requirements and chemical safety assessment Chapter R. 16: Environmental Exposure Estimation.*
- ECHA, 2012b. *Guidance on information requirements and chemical safety assessment Part D: Exposure Scenario Building.*
- ECHA, 2012c. *Guidance on information requirements and chemical safety assessment part E: Risk Characterisation.*
- ECHA, 2014a. Guidance on information requirements and chemical safety assessment Chapter R. 7C: Endpoint specific guidance.
- ECHA, 2014b. *The Use of Alternatives to Testing on Animals for the REACH Regulation,*
- ECHA, 2016a. Candidate List of substances of very high concern for Authorisation. Available at: <http://echa.europa.eu/en/candidate-list-table> [Accessed April 11, 2016].
- ECHA, 2016b. *Practical Guide 5 - How to use and report (Q)SARs.*
- Efron, B. & Gong, G., 1983. A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation. *The American Statistician*, 37(1), pp.36–48.
- EFSA, 2013. Scientific Opinion on the hazard assessment of endocrine disruptors: Scientific criteria for identification of endocrine disruptors and appropriateness of existing test methods for assessing effects mediated by these substances on human health and the env. *EFSA journal*, 11(3), p.83 pp.
- Engelen, J.G.M. Van et al., 2007. Risk Assessment of Chemicals: An Introduction. In C. J. van Leeuwen & T. G. Vermeire, eds. Dordrecht: Springer Netherlands, pp. 195–226.
- EPA, 2016. Estimation Programs Interface Suite™ for Microsoft® Windows.

- Franco, A. et al., 2010. An unexpected challenge: ionizable compounds in the REACH chemical space. *The International Journal of Life Cycle Assessment*, 15(4), pp.321–325.
- G. Shin, Y. et al., Comparison of Metabolic Soft Spot Predictions of CYP3A4, CYP2C9 and CYP2D6 Substrates Using MetaSite and StarDrop. *Combinatorial Chemistry & High Throughput Screening*, 14(9), pp.811–823.
- Gramatica, P., 2007. Principles of QSAR models validation: Internal and external. *QSAR and Combinatorial Science*.
- Gramatica, P., 2010. Recent Advances in QSAR Studies: Methods and Applications. In T. Puzyn, J. Leszczynski, & T. M. Cronin, eds. Dordrecht: Springer Netherlands, pp. 327–366.
- Hamers, T. et al., 2008. Biotransformation of brominated flame retardants into potentially endocrine-disrupting metabolites, with special attention to 2,2',4,4'-tetrabromodiphenyl ether (BDE-47). *Molecular Nutrition & Food Research*, 52(2), pp.284–298.
- Hansch, C. et al., 1962. Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients. *Nature*, 194(4824), pp.178–180.
- Hartung, T. et al., 2010. First alternative method validated by a retrospective weight-of-evidence approach to replace the Draize eye test for the identification of non-irritant substances for a defined applicability domain. *ALTEX*, 27(1), pp.43–51.
- Hillebrecht, A. et al., 2011. Comparative Evaluation of in Silico Systems for Ames Test Mutagenicity Prediction: Scope and Limitations. *Chemical Research in Toxicology*, 24(6), pp.843–854.
- Hyland, K.C. et al., 2012. Sorption of ionized and neutral emerging trace organic compounds onto activated sludge from different wastewater treatment configurations. *Water research*, 46(6), pp.1958–68.
- Jaworska, J., Aldenberg, T. & Nikolova, N., 2005. Review of methods for QSAR applicability domain estimation by the training set. , (January 2005), pp.445–459.
- Jaworska, J.S. et al., 2015. Bayesian integrated testing strategy (ITS) for skin sensitization potency assessment: a decision support system for quantitative weight of evidence and adaptive testing strategy. *Archives of Toxicology*, 89(12), pp.2355–2383.
- Jensen, G.E. et al., 2011. QSAR models for anti-androgenic effect – a preliminary study. *SAR and QSAR in Environmental Research*, 22(1-2), pp.35–49.
- Jones, O.A., Voulvoulis, N. & Lester, J.N., 2003. Potential impact of

- pharmaceuticals on environmental health. *Bulletin of the World Health Organization*, 81(10), pp.768–769.
- Kaneko, H. & Funatsu, K., 2014. Applicability Domain Based on Ensemble Learning in Classification and Regression Analyses. *Journal of Chemical Information and Modeling*, 54(9), pp.2469–2482.
- Krewski, D. et al., 2010. Toxicity Testing in the 21st Century: A Vision and a Strategy. *Journal of Toxicology and Environmental Health, Part B*, 13(2-4), pp.51–138.
- Kruhlik, N.L. et al., 2012. (Q)SAR Modeling and Safety Assessment in Regulatory Review. *Clinical Pharmacology & Therapeutics*, 91(3), pp.529–534.
- Lee, S.J. et al., 2013. Distinguishing between genotoxic and non-genotoxic hepatocarcinogens by gene expression profiling and bioinformatic pathway analysis. *Scientific Reports*, 3, p.2783.
- Leeuwen, C.J. Van, 2007. Risk Assessment of Chemicals: An Introduction. In C. J. van Leeuwen & T. G. Vermeire, eds. Dordrecht: Springer Netherlands, pp. 1–36.
- Lipinski, C.A., 2004. Lead- and drug-like compounds: the rule-of-five revolution. *Drug discovery today. Technologies*, 1(4), pp.337–41.
- Liu, H., Papa, E. & Gramatica, P., 2008. Evaluation and QSAR modeling on multiple endpoints of estrogen activity based on different bioassays. *Chemosphere*.
- Liu, R. et al., 2012. 2D SMARTCyp Reactivity-Based Site of Metabolism Prediction for Major Drug-Metabolizing Cytochrome P450 Enzymes. *Journal of Chemical Information and Modeling*, 52(6), pp.1698–1712.
- Livingstone, D.J., 2000. The Characterization of Chemical Structures Using Molecular Properties. A Survey. *Journal of Chemical Information and Computer Sciences*, 40(2), pp.195–209.
- M. Honorio, K., L. Moda, T. & D. Andricopulo, A., 2013. Pharmacokinetic Properties and In Silico ADME Modeling in Drug Discovery. *Medicinal Chemistry*, 9(2), pp.163–176.
- Matthews, E.J. et al., 2007. A comprehensive model for reproductive and developmental toxicity hazard identification: II. Construction of QSAR models to predict activities of untested chemicals. *Regulatory toxicology and pharmacology : RTP*, 47(2), pp.136–55.
- McClellan, R.O., 1999. Keynote address: human health risk assessment: A Historical Overview and Alternative Paths Forward. *Inhalation Toxicology*, 11(6-7), pp.477–518.
- Meent, D.D. Van & Bruijn, J.H.M. De, 2007. Risk Assessment of Chemicals: An Introduction. In C. J. van Leeuwen & T. G. Vermeire, eds.

- Dordrecht: Springer Netherlands, pp. 159–193.
- Netzeva, T.I. et al., 2005. Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. The report and recommendations of ECVAM Workshop 52. *Alternatives to laboratory animals : ATLA*, 33(2), pp.155–73.
- Norinder, U. et al., 2014. Introducing Conformal Prediction in Predictive Modeling. A Transparent and Flexible Alternative to Applicability Domain Determination. *Journal of Chemical Information and Modeling*, 54(6), pp.1596–1603.
- NRC, 2007. *Toxicity Testing in the 21st Century: A Vision and a Strategy*, Washington, DC: National Research Council, The National Academies Press.
- OECD, 1992a. OECD Environment Monographs No. 58. Report of the OECD workshop on quantitative structure activity relationships (QSARs) in aquatic effects assessment.
- OECD, 1983. *Test No. 415: One-Generation Reproduction Toxicity Study*, OECD Publishing.
- OECD, 1986. *Test No. 485: Genetic toxicology, Mouse Heritable Translocation Assay*, OECD Publishing.
- OECD, 1992b. Test No. 203: Fish, Acute Toxicity Test.
- OECD, 1992c. Test No. 301: Ready Biodegradability.
- OECD, 1993a. OECD Environment monograph No. 67. Application of structure-activity relationships to the estimation of properties important in exposure assessment.
- OECD, 1993b. OECD Environment Monograph No. 68. Structure-activity relationships for biodegradation.
- OECD, 1995a. OECD Environment Monographs No. 92. Guidance document for aquatic effects assessment.
- OECD, 1995b. Test No. 103: Boiling Point.
- OECD, 1997a. *Test No. 471: Bacterial Reverse Mutation Test*, OECD Publishing.
- OECD, 1997b. *Test No. 486: Unscheduled DNA Synthesis (UDS) Test with Mammalian Liver Cells in vivo*, OECD Publishing.
- OECD, 1998. *Test No. 409: Repeated Dose 90-Day Oral Toxicity Study in Non-Rodents*, OECD Publishing.
- OECD, 2001a. *Test No. 414: Prenatal Development Toxicity Study*, OECD Publishing.
- OECD, 2001b. *Test No. 416: Two-Generation Reproduction Toxicity*, OECD

- Publishing.
- OECD, 2004a. Descriptions of Selected Key Generic Terms used in Chemical Hazard/Risk Assessment.
- OECD, 2004b. Test No. 202: *Daphnia* sp. Acute Immobilisation Test.
- OECD, 2006. Test No. 123: Partition Coefficient (1-Octanol/Water): Slow-Stirring Method.
- OECD, 2007a. Guidance document on the validation of (quantitative) structure-activity relationships [(Q) SAR] models. *OECD Series on Testing and Assessment No. 69. ENV/JM/MONO (2007) 2*, 154.
- OECD, 2007b. *Test No. 426: Developmental Neurotoxicity Study*, OECD Publishing.
- OECD, 2007c. *Test No. 440: Uterotrophic Bioassay in Rodents*, OECD Publishing.
- OECD, 2009a. *Test No. 441: Hershberger Bioassay in Rats*, OECD Publishing.
- OECD, 2009b. *Test No. 451: Carcinogenicity Studies*, OECD Publishing.
- OECD, 2009c. *Test No. 453: Combined Chronic Toxicity/Carcinogenicity Studies*, OECD Publishing.
- OECD, 2011. *Test No. 456: H295R Steroidogenesis Assay*, OECD Publishing.
- OECD, 2012a. Test No. 305: Bioaccumulation in Fish: Aqueous and Dietary Exposure.
- OECD, 2012b. *Test No. 443: Extended One-Generation Reproductive Toxicity Study*, OECD Publishing.
- OECD, 2012c. *Test No. 457: BG1Luc Estrogen Receptor Transactivation Test Method for Identifying Estrogen Receptor Agonists and Antagonists*, OECD Publishing.
- OECD, 2013. *Test No. 488: Transgenic Rodent Somatic and Germ Cell Gene Mutation Assays*, OECD Publishing.
- OECD, 2014a. Guidance Document 116 on the Conduct and Design of Chronic Toxicity and Carcinogenicity Studies, Supporting Test Guidelines 451, 452 and 453.
- OECD, 2014b. *Test No. 473: In Vitro Mammalian Chromosomal Aberration Test*, OECD Publishing.
- OECD, 2014c. *Test No. 474: Mammalian Erythrocyte Micronucleus Test*, OECD Publishing.
- OECD, 2014d. *Test No. 475: Mammalian Bone Marrow Chromosomal Aberration Test*, OECD Publishing.

- OECD, 2014e. *Test No. 487: In Vitro Mammalian Cell Micronucleus Test*, OECD Publishing.
- OECD, 2014f. *Test No. 489: In Vivo Mammalian Alkaline Comet Assay*, OECD Publishing.
- OECD, 2015a. *Test No. 421: Reproduction/Developmental Toxicity Screening Test*, OECD Publishing.
- OECD, 2015b. *Test No. 422: Combined Repeated Dose Toxicity Study with the Reproduction/Developmental Toxicity Screening Test*, OECD Publishing.
- OECD, 2015c. *Test No. 455: Performance-Based Test Guideline for Stably Transfected Transactivation In Vitro Assays to Detect Estrogen Receptor Agonists and Antagonists*, OECD Publishing.
- OECD, 2015d. *Test No. 476: In Vitro Mammalian Cell Gene Mutation Tests using the Hprt and xpvt genes*, OECD Publishing.
- OECD, 2015e. *Test No. 478: Rodent Dominant Lethal Test*, OECD Publishing.
- OECD, 2015f. *Test No. 483: Mammalian Spermatogonial Chromosomal Aberration Test*, OECD Publishing.
- OECD, 2015g. *Test No. 490: In Vitro Mammalian Cell Gene Mutation Tests Using the Thymidine Kinase Gene*, OECD Publishing.
- OECD, 2015h. *Test No. 493: Performance-Based Test Guideline for Human Recombinant Estrogen Receptor (hrER) In Vitro Assays to Detect Chemicals with ER Binding Affinity*, OECD Publishing.
- OECD, 2015i. The OECD QSAR Toolbox.
- Papa, E., Kovarich, S. & Gramatica, P., 2013. QSAR prediction of the competitive interaction of emerging halogenated pollutants with human transthyretin. *SAR and QSAR in Environmental Research*, 24(4), pp.333–349.
- Patlewicz, G. et al., 2015. Proposing a scientific confidence framework to help support the application of adverse outcome pathways for regulatory purposes. *Regulatory toxicology and pharmacology : RTP*, 71(3), pp.463–77.
- Patlewicz, G. & Fitzpatrick, J.M., 2016. Current and Future Perspectives on the Development, Evaluation, and Application of in Silico Approaches for Predicting Toxicity. *Chemical Research in Toxicology*, p.acs.chemrestox.5b00388.
- Puzyn, T. et al., 2011. Global versus local QSPR models for persistent organic pollutants: balancing between predictivity and economy. *Structural Chemistry*, 22(4), pp.873–884.

- REACH, 2006. Regulation (EC) No 1907/2006.
- Rovida, C. et al., 2015. Integrated Testing Strategies (ITS) for safety assessment. *ALTEX-Alternatives to Animal Experimentations*, 32(1), pp.25–40.
- Sahigara, F. et al., 2012. Comparison of different approaches to define the applicability domain of QSAR models. *Molecules (Basel, Switzerland)*, 17(5), pp.4791–810.
- Sathyamoorthy, S. & Ramsburg, C.A., 2013. Assessment of quantitative structural property relationships for prediction of pharmaceutical sorption during biological wastewater treatment. *Chemosphere*, 92(6), pp.639–46.
- Sheridan, R.P., 2012. Three Useful Dimensions for Domain Applicability in QSAR Models Using Random Forest. *Journal of Chemical Information and Modeling*, 52(3), pp.814–823.
- Sridhar, J. et al., 2012. Insights on Cytochrome P450 Enzymes and Inhibitors Obtained Through QSAR Studies. *Molecules*, 17(12), pp.9283–9305.
- START, 2008. Pharmaceuticals for Human Use: Options of Action for Reducing the Contamination of Water Bodies. ISOE GmbH: Frankfurt, Germany, 52.
- T'jollyn, H. et al., 2011. Evaluation of Three State-of-the-Art Metabolite Prediction Software Packages (Meteor, MetaSite, and StarDrop) through Independent and Synergistic Use. *Drug Metabolism and Disposition*, 39 (11), pp.2066–2075.
- Ternes, T.A., Joss, A. & Siegrist, H., 2004. Peer Reviewed: Scrutinizing Pharmaceuticals and Personal Care Products in Wastewater Treatment. *Environmental Science & Technology*, 38(20), p.392A–399A.
- Todeschini, R. & Consonni, V., 2000. *Handbook of molecular descriptors*, Wiley-VCH.
- Trunzer, M., Faller, B. & Zimmerlin, A., 2009. Metabolic Soft Spot Identification and Compound Optimization in Early Discovery Phases Using MetaSite and LC-MS/MS Validation. *Journal of Medicinal Chemistry*, 52(2), pp.329–335.
- Tülp, H.C. et al., 2009. pH-Dependent Sorption of Acidic Organic Chemicals to Soil Organic Matter. *Environmental Science & Technology*, 43(24), pp.9189–9195.
- Waldron, H.A., 1983. Did the Mad Hatter have mercury poisoning? *British Medical Journal (Clinical research ed.)*, 287(6409), p.1961.
- Valerio, L.G., 2009. In silico toxicology for the pharmaceutical sciences. *Toxicology and applied pharmacology*, 241(3), pp.356–70.

- Wania, F. & Mackay, D., 1999. The evolution of mass balance models of persistent organic pollutant fate in the environment. *Environmental Pollution*, 100(1-3), pp.223–240.
- Varmuza, K. & Filzmoser, P., 2008. Introduction to multivariate statistical analysis in chemometrics.
- Vermeire, T. et al., 2013. OSIRIS, a quest for proof of principle for integrated testing strategies of chemicals for four human health endpoints. *Regulatory toxicology and pharmacology : RTP*, 67(2), pp.136–45.
- WHO/IPCS, 2002. *Global Assessment of the State-of-the-science of Endocrine Disruptors*.
- Willett, J. 1953-, 1987. *Similarity and Clustering in Chemical Information Systems*, New York, NY, USA: John Wiley & Sons, Inc.
- Wold, S., Eriksson, L. & Clementi, S., 1995. Statistical validation of QSAR results. *Chemometric methods in molecular design*, pp.309–338.
- Vorberg, S. & Tetko, I. V., 2014. Modeling the Biodegradability of Chemical Compounds Using the Online CHEmical Modeling Environment (OCHEM). *Molecular Informatics*, 33(1), pp.73–85.
- Worth, A.P., 2010. Recent Advances in QSAR Studies: Methods and Applications. In T. Puzyn, J. Leszczynski, & T. M. Cronin, eds. Dordrecht: Springer Netherlands, pp. 367–382.
- Yaron, B. & Saltzman, S., 1978. Soil-parathion surface interactions. In F. Gunther & J. Gunther, eds. *Residue Reviews SE - 1*. Residues of Pesticides and Other Contaminants in the Total Environment. Springer New York, pp. 1–34.
- Yuan, H., Wang, Y. & Cheng, Y., 2007. Local and Global Quantitative Structure–Activity Relationship Modeling and Prediction for the Baseline Toxicity. *Journal of Chemical Information and Modeling*, 47(1), pp.159–169.
- Zefirov, N.S. & Palyulin, V.A., 2001. QSAR for Boiling Points of “Small” Sulfides. Are the “High-Quality Structure-Property-Activity Regressions” the Real High Quality QSAR Models? *Journal of Chemical Information and Computer Sciences*, 41(4), pp.1022–1027.
- Zhou, D. et al., 2006. Comparison of methods for the prediction of the metabolic sites for CYP3A4-mediated metabolic reactions. *Drug Metabolism and Disposition*, 34 (6), pp.976–983.
- Zvinavashe, E., Murk, A.J. & Rietjens, I.M.C.M., 2008. Promises and Pitfalls of Quantitative Structure–Activity Relationship Approaches for Predicting Metabolism and Toxicity. *Chemical Research in Toxicology*, 21(12), pp.2229–2236.