INTEGRATION OF DATA QUALITY, KINETICS AND MECHANISTIC MODELLING INTO TOXICOLOGICAL ASSESSMENT OF COSMETIC INGREDIENTS

Fabian P. Steinmetz, M.Sc.

December 2015

A thesis submitted in partial fulfilment of the requirement of
Liverpool John Moores University for the degree of
Doctor of Philosophy

Acknowledgments

There are so many people I have to thank, so I would like to apologise to those I might have forgotten. First I would like to thank my director of studies, Prof. Mark T. D. Cronin, who is the best supervisor I could imagine. His support over the last three years was crucial for this project and is much appreciated. Furthermore I would like to thank my other supervisors, Dr. Judith C. Madden and Dr. Steven J. Enoch, for giving me insight into their fields and so helped me to steer my project into useful and applied science.

I must also thank my colleagues, in particular Dr. Richard Marchese Robinson, Dr. Mark D. Nelms, Dr. Claire L. Mellor, David J. Ebbrell, Iva Lukač, Dr. Katarzyna Przybylak and Antonio Cassano, who supported me during all challenges of this project. Additional thanks go to Prof. Vera Rogiers and Gamze Ates from the Vrije Universiteit Brussel. Furthermore I would like to thank the EU COSMOS project, in particular the European Community's 7th Framework Program and Cosmetics Europe for the financial support, and my colleagues from all over Europe.

Last but not least I would like to thank my family, Ulrich, Marlene and Claudia Steinmetz and my better half, Britta Steinmetz… and of course thanks to Jockel and Carina for keeping us safe and fit.

Abbreviations

3Rs:	Replacement, reduction and refinement of animal testing
AC_{50}:	Concentration causing 50% of activity
ADI:	Acceptable daily intake
ADME:	Absorption, distribution, metabolism and excretion (pharmacokinetics)
AOP:	Adverse outcome pathway
BfR:	Bundesinstitut für Risikobewertung (German Federal Institute for Risk Assessment)
BSEP:	Bile salt export pump
CDK:	Chemistry Development Kit (software library)
COACH:	COordination of projects on new Approaches to replace current repeated dose systemic toxicity testing of cosmetics and CHemicals
COSMOS:	Integrated *in silico* models for the prediction of human repeated dose toxicity of COSMetics to Optimise Safety
CS:	Confidence score
DBD:	DNA-binding domain
DETECTIVE:	Detection of endpoints and biomarkers of repeated dose toxicity using *in vitro* systems
EC:	European Commission
EC_{50}:	Concentration causing 50% of the stated effect
EChA:	European Chemicals Agency
ECI:	Eccentric connectivity index
EEA:	European Environment Agency

Abbreviations

EMA:	European Medicines Agency
EPA:	US Environmental Protection Agency
EU:	European Union
f:	Predicted value
F:	F-value (linear regression)
FDA:	US Food and Drug Administration
FP7:	Seventh Framework Programme
FXR:	Farnesoid X receptor
GLP:	Good Laboratory Practice
HeMiBio:	Hepatic Microfluidic Bioreactor
hERG:	Human ether-à-go-go-related gene
HSP90:	Heat shock protein 90
K_d:	Dissociation constant
K_i:	Binding constant
K_{OW}:	Octanol-water partition coefficient
k_p:	Skin permeability coefficient
KNIME:	Konstanz Information Miner (software)
LBD:	Ligand binding domain
LD_{50}:	Lethal dose for 50% of the population
LLNA:	Local lymph node assay
LOEL:	Lowest-observed-effect level
log P:	Logarithm of octanol-water partition coefficient (also known as log K_{OW})

LXR:	Liver X receptor
IC$_{50}$:	Concentration (in mmol/L) causing 50% of inhibition
InChIKey:	International chemical identifier key
MIE:	Molecular initiating event
MoA:	Mode of action
MoS:	Margin of safety
MW:	Molecular weight (in Da)
n:	Number of data points / test values
NASH:	Non-alcoholic steatohepatitis
NOAEL:	No-observed-adverse-effect level
NOEL:	No-observed-effect level
NOTOX:	Predicting long-term toxic effects using computer models based on systems characterisation of organotypic cultures
NR:	Nuclear receptor
OECD:	Organisation for Economic Co-operation and Development
PBPK:	Physiologically based pharmacokinetic (type of model)
PETA:	People for the Ethical Treatment of Animals
PPAR:	Peroxisome proliferator-activated receptor
pT:	Negative decadic logarithm of EC$_{50}$ for toxicity
qHTS:	Quantitative high-throughput screening
QSAR:	Quantitative structure-activity relationship
QSPR:	Quantitative structure-permeability relationship

R^2_{adj}:	Coefficient of determination adjusted for degrees of freedom
RAR:	Retinoic acid receptor
RB:	Number of rotational bonds
REACH:	Registration, Evaluation, Authorisation and Restriction of CHemicals
rLLNA:	Reduced local lymph node assay
RMSE:	Root mean square error
RMSE_{CS}:	CS-adjusted RMSE
RXR:	Retinoid X receptor
RSD:	Relative standard deviation (also known as coefficient of variation)
S:	Standard error (linear regression)
SCCS:	Scientific Committee on Consumer Safety
SCR&Tox:	Stem Cells for Relevant efficient extended and normalised Toxicology
SD:	Standard deviation
SED:	Systemic exposure dosage
SEURAT-1:	Safety evaluation ultimately replacing animal testing
SME:	Small- and medium-size enterprise
SMILES:	Simplified Molecular-Input Line-Entry System
SMARTS:	SMiles ARbitrary Target Specification
t:	t-value (linear regression)
ToxBank:	Supporting integrated data analysis and servicing of alternative testing methods in toxicology
TPSA:	Topological polar surface area

Abbreviations

TTC:	Toxicological threshold of concern
VAIM:	Vertex adjacency information magnitude
$\bar{x}$:	Arithmetic mean
XlogP:	CDK's octanol-water partition coefficient (K_{OW})
y:	Experimental value

Abstract

In our modern society we are exposed to many natural and synthetic chemicals. The assessment of chemicals with regard to human safety is difficult but nevertheless of high importance. Beside clinical studies, which are restricted to potential pharmaceuticals only, most toxicity data relevant for regulatory decision-making are based on *in vivo* data. Due to the ban on animal testing of cosmetic ingredients in the European Union, alternative approaches, such as *in vitro* and *in silico* tests, have become more prevalent.

In this thesis existing non-testing approaches (*i.e.* studies without additional experiments) have been extended, *e.g.* QSAR models, and new non-testing approaches, *e.g. in vitro* data supported structural alert systems, have been created. The main aspect of the thesis depends on the determination of data quality, improving modelling performance and supporting Adverse Outcome Pathways (AOPs) with definitions of structural alerts and physico-chemical properties. Furthermore, there was a clear focus on the transparency of models, *i.e.* approaches using algorithmic feature selection, machine learning *etc.* have been avoided. Furthermore structural alert systems have been written in an understandable and transparent manner. Beside the methodological aspects of this work, cosmetically relevant examples of models have been chosen, *e.g.* skin penetration and hepatic steatosis.

Interpretations of models, as well as the possibility of adjustments and extensions, have been discussed thoroughly. As models usually do not depict reality flawlessly, consensus approaches of various non-testing approaches and *in vitro* tests should be used to support decision-making in the regulatory context. For example within read-across, it is feasible to use supporting information from QSAR models, docking, *in vitro* tests *etc*. By applying a variety of models, results should lead to conclusions being more usable/acceptable within toxicology.

Abstract

Within this thesis (and associated publications) novel methodologies on how to assess and employ statistical data quality and how to screen for potential liver toxicants have been described. Furthermore computational tools, such as models for skin permeability and dermal absorption, have been created.

Table of Contents

1. Introduction

1.1. Safety assessment of chemicals

As modern society demands a safer environment, the assessment of the harmful effects of chemicals has become a crucial element of today's toxicology. The desire for the use of safe chemicals arises, in part, from a consumer perspective, which includes the requirement for safety of pesticides, foods, pharmaceuticals, cosmetics, industrial chemicals *etc.* It also arises within occupational health, *i.e.* for the assessment of workers who are exposed to chemicals in an industrial setting, or it may come from an environmental perspective, for example the effects of chemicals on *flora*, *fauna*, and ultimately humans, via the food chain. Advances in technology in the last century have led to an intensive production of drugs, cosmetics, food products, pesticides, munitions, synthetic fibres and industrial chemicals (Gallo, 2001; Cronin, 2013). There are many other, often less recognised, sources of chemicals to which we are exposed; substances may originate, for example, from fungi, botanicals or from the petro-chemical industry (including their combustion products). Chemicals exposure may be localised or might be long-range, for instance transportation via air pollution (Alam *et al.*, 2013) or via the food chain (Cheng *et al.*, 2015a).

The number of potential organic chemical compounds is almost unimaginably large, and it is impossible to limit exposure of chemicals *per se* as we live in a world made of chemicals. Despite this, the toxicological/pharmacological effects are known for only a very small proportion of chemicals, even for commonly used compounds (Cronin, 2013). As a result it is difficult to ensure the safety of chemicals, or the absence of associated adverse effects. Exposure to many chemicals, pharmaceuticals and pesticides being the most active/extreme examples, is controlled through national regulations. It is the purpose of regulatory authorities, such as the US Food and Drug Administration (FDA), to assess chemicals for their safety by gathering knowledge from experimentation (or non-test methods), feedback from human exposure and utilising human expertise to evaluate the data enabling risk assessment. Thus regulators use toxicological information on specific substances to define thresholds, which are thought to be safe to man and the environment for the intended use of the chemical (Merrill, 2001). It is important to distinguish between contaminants and compounds used deliberately in food or cosmetic products, *e.g.* preservatives. For the latter, there is usually more information available, for example how to detect them analytically or the principal effects they have on biological systems. Contaminants or impurities, on the other hand, vary widely in terms of chemistry and are often dependent from educts used and/or the manufacturing process (Gallo, 2001; Merrill, 2001; Feigenbaum *et al.*, 2015), they may also be as a result of compound degradation (biotic or abiotic). Despite the many substances created unknowingly within the manufacturing process, mostly in low concentrations or even traces, a great number of chemicals (and mixtures) are produced wittingly. These substances can be ingredients for consumer products or intermediates made for further chemical engineering (Faustman and Omenn, 2001).

In the context of better risk assessment for the higher production volume chemicals, the Registration, Evaluation, Authorisation and restriction of CHemicals (REACH) regulation, which is enforced by the European Union and regulated by the European Chemicals Agency (EChA), has been an immense undertaking for over a decade. REACH addresses production quantity and use of chemicals and their potential impacts on human and environmental health. Here the focus lies on evaluating chemicals, and particularly those produced in large quantities, *e.g.* greater than ten tonnes per year (EC, 2006). For the appropriate evaluation of chemicals, not only within the context of REACH, a sound understanding of toxicology and exposure is needed.

1.2. Toxicology and its application

Toxicology is the science of poisons. It is a centuries old scientific discipline, with the thoughts of the Swiss physician/alchemist Paracelsus from the 16th century still relevant today. Paracelsus stated that any substance can be poisonous, only the dose differentiates the non-poisonous from the poisonous. This statement gives rise to our understanding that the intrinsic risk of a chemical is a function of its implicit hazard and the exposure scenario, as such toxicology reveals itself as a non-trivial task (Gallo, 2001). Historically toxicology is an experimental and observational science with the use of animals to identify hazard at the heart of most studies. Some of the more common, and relevant to this thesis, toxicological procedures that have been used for hazard identification are introduced below.

1.2.1. Acute toxicity

A standard way to measure acute toxicity (normally associated with lethality) is the LD_{50} (the single dose of a substance that causes death in 50% of an animal population). These values, usually extracted from a dose-response-curve, allow for the differentiation of those chemicals with high acute toxicity from chemicals with low acute toxicity. However, acute toxicity may be influenced by many (non-chemically related) factors such as dosing regime and route, species, age, weight and sex. However lethality, as in LD_{50}, is not the only endpoint of interest when performing acute toxicity animal testing. Doses and exposure patterns, where for example blood chemistry or kidney and liver histology is pathologically changed, can reveal potential risks of chemicals and form the basis of long-term studies (Barile, 2004).

1.2.2. Chronic toxicity

In addition to testing for acute toxicity (*i.e.* short-term, high dose exposure, with the aim of identifying a lethal dose), there is also a great interest in non-lethal effects and toxicity caused by repeated dose exposure. Chronic or repeated dose toxicity is caused by long-term, continuous or fluctuating sublethal exposure of a toxicant. Typical repeated dose exposure scenarios are subacute, subchronic or chronic, and definitions and exposure guidelines for these may vary. This may be particularly important for substances which have a long half-life and a tendency to be accumulated and reveal adverse effects after long-term exposure (Barile, 2004). In addition, identifying the relevant mechanistic pathways and kinetics associated with chronic toxicity is a complex, but increasingly important, task. Further, there are compounds which interfere at low doses with the human hormonal system, such as endocrine disruptors, which are likely to cause pathological changes when exposed chronically (Fuhrman *et al.*, 2015).

1.2.3. Other effects

Further genotoxicity (*e.g.* due to mutagenicity) and immunotoxicological responses are of great interest for applied toxicology and risk assessment as these can be triggered by very low doses (Faustman and Omenn, 2001; Barile, 2004). These may be identified by specific, often mechanistically derived, tests such the Ames test for genotoxic mutagenicity. These effects are outside the scope of this thesis and are not considered further.

1.2.4. Experimental considerations when testing

When assessing the experimental determination of toxicity, the route of administration, *e.g.* topical, oral, inhalation or subcutaneous, intravenous and intramuscular injection, and the formulation of the drug/toxicant (*i.e.* particle size, excipients *etc*.) are important parameters. While an intravenous dose of a drug is usually systemically available, an oral dose is most often absorbed more slowly through the gastrointestinal tract and eventually metabolised, for example by the first-pass effect (and further metabolism by repeated passing through the liver). Further novel formulations can deliver drugs to specific target tissues, *e.g.* drug-loaded nanoparticle systems. Hence, both route of administration and formulation play important roles for the absorption, distribution and finally the biological effects of a drug/toxicant. These lessons, which were mainly learnt in the field of pharmacology, are of great importance for understanding and applying modern toxicology (Rang *et al.*, 2007a; Barile 2004). The correct dosing route is also essential to understand the effects of a particular exposure scenario, *i.e.* a pharmaceutical applied as an oral dosage form should be tested orally in an attempt to mimic (within reason) the kinetics of uptake, distribution and metabolism.

1.2.5. Data availability

It has historically been a great problem to obtain appropriate, high quality and relevant toxicity data for a variety of chemicals. It is assumed that many of the available toxicity data are in private hands, *e.g.* within pharmaceutical companies (Cases *et al.*, 2013; Tralau *et al.*, 2015). While there are often many data for acute toxicity, *e.g.* LD_{50} values for rodents, and local (adverse) reaction, such as skin and eye irritation, there is a lack of publically available data and hence a shortcoming in understanding of biological responses associated with chronic toxicity. The ever increasing number of chemicals produced and the introduction of expectations, such as reducing or replacing animal testing, is the setting for 21st Century Toxicology (Cronin, 2013; Groh *et al.*, 2015; Vinardell, 2015). Within this thesis (*e.g.* Chapters 2 and 3) efforts to supplement the availability of chronic toxicity data, as well as understand the quality of data, are presented in response to this need.

1.2.6. Application of toxicological information

There are many applications of toxicological information, but for the purposes of this thesis risk assessment, relating to regulatory acceptance of a chemical, is considered in more focus, as introduced in Section 1.3. A further application of toxicological information is the requirement to find ethical, cost-effective and scientifically valid alternatives to animal testing that has been a driving force behind the research reported in this thesis; more details on this are given in Section 1.4.

1.3. Regulatory toxicology

Regulatory toxicology is a discipline which is intended to ensure that the world, in chemical terms, becomes a safer place. It combines toxicological expertise, *i.e.* knowledge of exposure, kinetics and mechanisms, with risk assessment approaches to create regulations. Regulatory toxicological activities take place partially in industry, *e.g.* companies launching new consumer products, and partially in governmental institutions. Taking the US as an example, the FDA is the responsible governmental institution for licensing food and medical products, such as drugs. Cosmetics do not need a license *per se*, but still they need to be regarded as safe before launching a product, *i.e.* potentially hazardous chemicals have to be excluded by regulatory toxicologists. With regard to human and environmental health, the US Environmental Protection Agency (EPA) is responsible for the assessment of the impact of pesticides or industrial chemicals and their exposure to man and environmental species (Merrill, 2001). In Europe there are institutions, such as the European Medicines Agency (EMA) and the European Environment Agency (EEA), with similar responsibilities as FDA and EPA respectively. National regulations of EU member states still may vary, for example due to recommendations from domestic institutes, *e.g.* the German Federal Institute for Risk Assessment (Bundesinstitut für Risikobewertung; BfR).

1.3.1. Safety and legislation

When assessing safety for different types of products, it is important to consider individual risk-benefit ratios. Drug safety for example, often dictated by therapeutic necessities, has a quite unique and complex way of considering the risk-benefit ratio for potential patients. This consideration is inevitably associated with the process of regulatory approval of pharmaceutical drugs. Generally, pharmaceutical drugs have to pass many stages before being launched for therapeutic purposes. After many years of drug development and preclinical testing (based on animal trails), only a few drug candidates will enter the clinical phases. All three clinical phases need to be passed consecutively, whereupon the last phase (clinical phase III) would be a multi-centred trial with 1,000 to 3,000 patients. Only after that, can a drug be submitted to regulatory authorities for licensing. Following the launch of a new drug post-market surveillance (often referred to as pharmacovigilance or phase IV) is the responsibility of the pharmaceutical company; this means monitoring the drug for adverse drug reactions and side effects, and withdrawing a drug from the market, should the need arise (Merrill, *et al.* 2001; Barile, 2004; Rang *et al.*, 2007b).

Legal requirements for non-pharmaceutical products, such as biocidal and plant protection products (*e.g.* pesticides), food products and diverse consumer products (*e.g.* toys, textile products) can be quite different. In food safety, for example, food additives, flavouring substances, novel ingredients, genetically modified organism-based products and contaminants need to be assessed as being safe before bringing them onto the market. The legal responsibility for consumer goods usually lies with the producing company. With regard to the safety assessment of biocidal and plant protection products, the legal focus lies on metabolites in food, feed and groundwater, and the assessment of cumulative effects in organism and soils (Tralau *et al.*, 2015).

1.3.2. The cosmetics legislation

According to the European Cosmetics Regulation, a cosmetic product made available on the market has to be safe for humans when used normally and reasonably. Hence, the Cosmetic Directive places the responsibility for product safety clearly on the company. Nevertheless, the Scientific Committee on Consumer Safety (SCCS) provides the European Commission (EC) with scientific advice on the safety of cosmetic products (EC 2009; Vinardell, 2015).

Generally cosmetic ingredients should be inert, *i.e.* they should not have any significant pharmacological or toxicological properties. Of course, there are exceptions such as, for example zinc pyrithione, a fungi-/bacteriostatic substance used in antidandruff shampoo (Marks *et al.*, 1985), or hair dyes, such as aromatic azo dyes, which can act as mitochondrial toxins (Nelms *et al.*, 2015). Compounds of concern are often found in the functional classes associated with colourants, preservatives and UV filters. However, exposure, *i.e.* dose and type of application, plays an important role for the risk assessment of a cosmetic ingredient. In other words, considering the quantity and type of usage is also the responsibility of industry and regulatory authorities (*e.g.* through the SCCS in the EU) (Vinardell, 2015).

Since March 2013, European legislation has banned animal testing for any cosmetic ingredient marketed within the EU (EC, 2009). Proposed alternative testing methods include *in vitro* tests (*e.g.* mechanistically based mutagenicity assays) and *in silico* approaches (such as computational methods, often based on historical *in vivo* toxicity data). Current challenges revolve around the need for alternatives, especially for chronic and reproductive toxicity (Adler *et al.*, 2011). Examples where successfully validated alternatives are present are the local lymph node assay (LLNA) used for skin sensitisation and diverse alternatives for the Draize rabbit eye test to evaluate of eye irritation.

However, current alternative methods still have potential for optimisation, particularly as some of the methods are *ex vivo* assays (*i.e.* using tissues of animals) and hence animals are still used for testing (AltTox, 2015; Roberts, 2015; Vinardell, 2015).

Whilst modern animal welfare is often considered to be driven by companies, such as Lush (Lush, 2015) and organisations, such as the People for the Ethical Treatment of Animals (PETA, 2015), there have been attempts to reduce the number of animals used and to generally increase animal welfare standards for more than fifty years. For example Russell and Burch's "3Rs", which refers to the replacement, reduction and refinement of animal tests, is a paradigm, which has existed since the late 1950s (Russell and Burch, 1959). Considerable success regarding the 3Rs has already been seen within the approaches adopted by institutions such as the Organisation for Economic Co-operation and Development (OECD). The OECD aims to stimulate economic progress and world trade within a democratic and capitalistic framework, coined by its European and American member states. They play an important role by providing international guidelines, not only in the field of toxicology. For instance, included within the OECD test guideline for acute eye irritation/corrosion (OECD, 2012), the usage of topical anaesthetics and systemic analgesics is described with the aim to decrease animal suffering. Furthermore, tests such as reduced LLNA (rLLNA), which uses fewer experimental animals as compared to the conventional LLNA (itself a less invasive test using fewer animals than the guinea pig maximisation tests), are promoted (Roberts, 2015). However, despite the importance of assessment of skin sensitisation for dermally applied products, it must be pointed out that the rLLNA is an *in vivo* assay, *i.e.* novel cosmetic ingredients tested with the rLLNA would not be allowed for sale in the EU according to the EC's Cosmetics Directive (EC, 2009).

1.3.3. *The European Commission's role*

The EC plays an important role in the advance of alternative testing methods, most notably with the SEURAT-1 "Safety Evaluation Ultimately Replacing Animal Testing" cluster. SEURAT-1 is a 50 million euro project funded by EC's Seventh Framework Programme (FP7) and Cosmetics Europe, the European trade association for cosmetic, toiletry and perfumery industry. Within SEURAT-1, for five years (2011-2015), research facilities from industry and over 70 European universities and SMEs have been developing non-animal test methods for systemic toxicity following repeated exposure *etc*. The SEURAT-1 cluster is divided into seven distinct projects (Gocht and Schwarz, 2014; SEURAT-1, 2015):

- SCR&Tox, "Stem Cells for Relevant efficient extended and normalised Toxicology"
- HeMiBio, "Hepatic Microfluidic Bioreactor"
- DETECTIVE, "Detection of endpoints and biomarkers of repeated dose toxicity using *in vitro* systems"
- COSMOS, "Integrated *in silico* models for the prediction of human repeated dose toxicity of COSMetics to Optimise Safety"
- NOTOX, "Predicting long-term toxic effects using computer models based on systems characterisation of organotypic cultures"
- ToxBank, "Supporting integrated data analysis and servicing of alternative testing methods in toxicology"
- COACH, "COordination of projects on new Approaches to replace current repeated dose systemic toxicity testing of cosmetics and CHemicals"

One of the research projects within the SEURAT-1 cluster is the COSMOS Project. As the full title and semi-acronym, “Integrated *in silico* models for the prediction of human repeated dose toxicity of COSMetics to Optimise Safety” indicates, the project aims to develop computational methods. Computational methods demand sufficient data, the requirement for which is often neglected, but is actually a crucial part for any data-driven, scientific approach. Therefore, building a database relevant for cosmetics was one of the major objectives of COSMOS. The so-called “COSMOS DB” was released online as a freely available resource in December 2013. This database can be regarded as the backbone for modelling and read-across approaches used to assess cosmetic ingredients within COSMOS (Richarz *et al.*, 2014).

In addition, the refinement of the Toxicological Threshold of Concern (TTC) approach and its extension to cosmetics ingredients is a major aim of the COSMOS project. As humans are likely to be exposed to thousands of chemicals in their life-time, and it is impossible to test every compound against every possible endpoint, feasible and pragmatic approaches for risk assessment are necessary. Originally deriving from the food industry, the TTC approach applies margins of safety based on no-observed-effect levels (NOELs), *i.e.* the highest concentration of a substance not causing any toxic effects *in vivo*. The so calculated acceptable daily intake (ADI) should ensure consumer safety (Munro *et al.*, 1996; Richarz *et al.*, 2014; Feigenbaum *et al.*, 2015).

Many challenges of the COSMOS project lie within the field of kinetics, *i.e.* predicting dermal absorption and modelling the distribution of chemicals (refer to physiologically-based pharmacokinetic models). The distribution of a chemical within tissues is of particular interest regarding potential target organ toxicity. Besides kinetic aspects, specific mechanisms of toxicity were also investigated within COSMOS (Richarz *et al.*, 2014). For example, many mechanisms of toxicity have been identified lately within the

Adverse Outcome Pathway (AOP – described in Section 1.4.2) framework (Vinken *et al.*, 2013). AOP-related QSAR models, for example for compounds causing fatty liver (hepatosteatosis) via agonism of the liver X receptor (LXR; a nuclear receptor responsible for lipid regulation amongst others), are under development (Fioravanzo *et al.*, 2013; Richarz *et al.*, 2014). Further screening tools for hepatotoxicity based on structural alerts and/or physico-chemical properties have been developed within the COSMOS project (*e.g.* Nelms *et al.*, 2015, Steinmetz *et al.*, 2015a).

1.4. 21st Century Toxicology

There are many international endeavours, including projects within the Horizon2020 funding programme in Europe and Tox21 in the US, to elucidate toxicity pathways at a molecular, cellular and histological level. By employing systems biology, *i.e.* genomics, proteomics and metabolomics, and robot-supported quantitative high-throughput screening (qHTS), a large amount of data are, and will be, generated (Attene-Ramos *et al.*, 2013; Gaspar *et al.*, 2012). Tox21, for example, screens chemicals using over 75 biochemical and cell-based assays resulting in information for different perturbations of signalling pathways, inflammatory response induction, DNA damage, general cytotoxicity *etc*. If these, and the associated existing *in vivo* / clinical, data are interpreted well, a significant wealth of knowledge could be created and exploited for pharmacological research and toxicological risk assessment. One such example is the work of Attene-Ramos and colleagues regarding mitochondrial toxicity; they defined chemical (sub)structures responsible for decreasing mitochondrial membrane potential (Attene-Ramos *et al.*, 2015).

Overall the pharmacological and toxicological knowledge obtained from research projects, such as Tox21, will lead to new biomarkers, safer drugs and, in general, a deeper understanding of biochemical interactions *in vivo*, which would benefit many life-science disciplines and regulatory bodies (Gaspar *et al.*, 2012; Attene-Ramos *et al.*, 2013).

1.4.1. Alternatives to animal testing

Animal testing is usually a means to obtain information regarding the safety (or specific effects) of chemicals relevant to humans. Human trials, which *per se* would provide more relevant data, are generally regarded as unethical and limited to clinical trials, patch tests *etc.*, where mostly non-toxic doses are administered. Within 21st Century Toxicology animal testing is becoming regarded as unethical, and even lacking scientific credibility, leading to alternative methods being investigated (Russell and Burch, 1959). SEURAT-1 is a good example how toxicological research can be conducted without animal testing, *i.e.* batteries of different *in vitro* tests on the one hand and computational modelling on the other.

Similar to species differences in susceptibility towards different chemicals, *in vitro* to *in vivo* extrapolation is a difficult challenge. Approaches for alternative test methods principally come from *in vitro* studies, *e.g.* the hepatic microfluidic bioreactor – a simulation of the human liver, and *in silico* studies, mainly in the form of physiologically-based pharmacokinetic (PBPK) models predicting target organ concentrations of chemicals (Gocht and Schwarz, 2014; SEURAT-1, 2015). Furthermore there are non-testing approaches aiming to make predictions of toxicity directly from chemical structure and property, mainly based on QSAR, read-across and expert opinions, which are sometimes summarised under the banner of predictive toxicology.

1.4.2. Predictive toxicology

Predicting the toxicity of an untested chemical is of great interest for many different reasons, such as animal welfare, or simply to save the costs of testing and resources involved. Whatever the motivation is, similar methods are applied. Better known methods include Quantitative Structure-Activity Relationship (QSAR) models and the read-across (also known as the category formation) approach. While QSAR models are usually based on one or more mathematical equation(s) exploiting physico-chemical and other descriptors to predict toxic effects, read-across, as the name suggests, is a direct extrapolation of toxicological effects from structurally similar compounds, usually performed by experts (Cronin, 2004; Cronin, 2013a; Schultz *et al.*, 2015). Further the concept of the Adverse Outcome Pathway (AOP), *i.e.* describing a sequence of causally linked events at different biological levels, is increasingly used to predict toxicity (Vinken *et al.*, 2013; Vinken, 2015). An AOP is shown schematically in Figure 1.1; the first key event of an AOP, the molecular initiating event (MIE) is followed by cellular and tissue responses, which may ultimately result in an adverse effect to an organ, organism or population (Ankley *et al.*, 2010). The MIE represents the initial interaction between molecule and the target and hence represents a significant source of information to develop structure-activity relationships (SARs) as part of mechanistically based computational profilers for toxicity. Examples of MIEs include covalent binding to DNA and receptor binding (Gutsell and Russell, 2013; Allen *et al.*, 2014). The AOP framework is, for example, used in Chapter 5 and 6 to predict potentially toxic compounds.

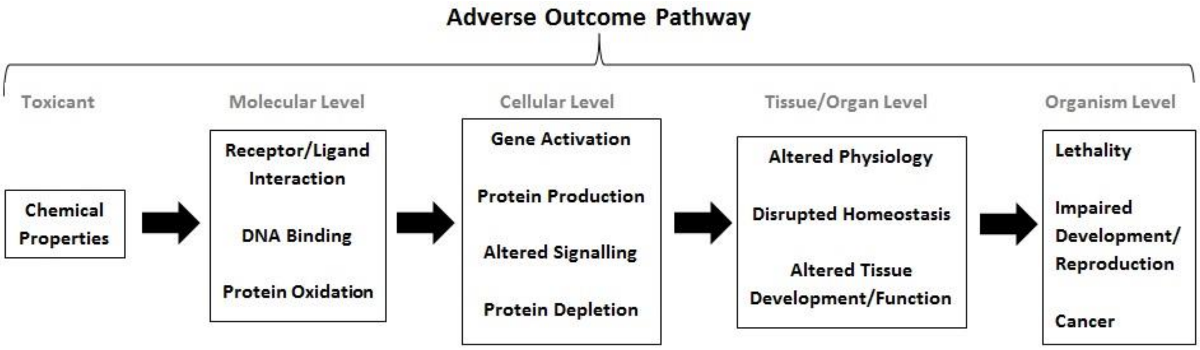


Figure 1.1: Schematic view of the Adverse Outcome Pathway framework (adapted from Ankley *et al.*, 2010)

Computational approaches, sometimes referred to as *in silico* testing or virtual screening, are often based on QSAR models, which incorporate physico-chemical and structural features in a mathematical context towards an endpoint. Endpoints which have been successfully modelled by predictive QSAR models include acute aquatic toxicity (Könemann, 1981; Verhaar *et al.* 1996), hERG (human Ether-à-go-go-Related Gene)-related toxicity (Gavaghan *et al.*, 2007), mutagenicity (Benigni and Giuliani, 1994), skin permeability (Potts and Guy, 1992) and skin sensitisation (Roberts and Williams, 1982). Naturally QSAR models have been applied to numerous other endpoints and in many other disciplines too, *e.g.* receptor binding within drug development. However what makes the models developed for aquatic toxicology, hERG-related binding, mutagenicity, skin sensitisation and permeability so significant, is that the models are robust and applicable to a large variety of chemical compounds. QSAR models within a specific class of compounds, *i.e.* having a narrow applicability domain, are often referred to as local QSARs.

The logarithm of the octanol-water partition coefficient (log P) plays an important role in many QSAR models, such as in aquatic toxicology and skin permeability, (refer to Potts and Guy, 1992; Verhaar *et al.*, 1996). Log P, also known as log K_{OW} (particularly in environmental sciences), is a measure of lipophilicity and hence is assumed to be an excellent surrogate for the partitioning and uptake of a compound through a biological membrane. As well as experimentally measuring log P, it is well predicted from topological descriptors or structural fragments (Ognichenko *et al.*, 2012).

As recognised in the late 19th Century by Charles Richet who stated "plus ils sont solubles, moins ils sont toxiques" (the more soluble they are, the less toxic they are), the relationship of biological effects and water-insolubility of chemical compounds is very well established (Richet, 1893). With regard to pharmacology and what characterises an orally bioavailable drug, log P is an often mentioned parameter. The rationale is mostly based on simple kinetics, such as passive diffusion, *e.g.* log P is used to describe the ability of a compound to pass through a biological membrane (Lipinski *et al.*, 2001), but there are mechanistic rationales too, such as the binding affinity towards receptors and transporters (refer to hydrophobic binding pockets) (Caron and Ermondi, 2008). QSAR approaches, particularly involving log P, are used, for example, in Chapters 2 and 3.

Predictive toxicology does not have limitations regarding the techniques applied in order to obtain the predictions of toxicity. That is why a wide range of methods, from machine learning to local QSARs and expert systems, is applied and investigated (*nota bene:* many approaches are combinations of different methods).

1.5. Aims of this work

Two worlds often collide in modern society; the commercial and the consumer world. Here the commercial world, predominantly backed by large global industries, is supplying the consumer's demand for innovative, safe and affordable consumer products. This includes products such as pharmaceuticals, cosmetics, clothing, toys and food. Within a capitalistic competitive environment it is important to deliver a product with an acceptable safety profile without compromising costs and/or quality. From a chemical perspective, new compounds, ideally cheap to produce and functional, need to be assessed regarding their safety. As explained above, this is no easy task. To add a further challenge to this task; animal tests are banned for new cosmetic ingredients marketed within the European Union. In a nutshell, there is a high demand for *in vitro* and *in silico* methods applicable for safety and risk assessment. Regarding *in silico* methods, the work presented in this thesis is a contribution to the current state-of-the-art. The following topics are addressed within this thesis.

- **Data quality:** Data are often erroneous for many reasons. However, having the appropriate quality of data is crucial for modelling and read-across. In particular, biological data (in comparison to physical or chemical data) are often associated with considerable error due to the complexity of assays and the difficulty of assigning endpoints. For example, a pharmacological dose-response relationship involves the formulation, dosing and administration of a substance to a group of animals, measuring a biological endpoint and applying statistical analysis to obtain an ED_{50} value. To overcome these potential pitfalls, large datasets have been investigated by statistical means to build tools for an unbiased way to assess data quality (refer to Chapter 2 and 3).
- **Kinetics:** Pharmaco- and toxicokinetics, which by definition encompass absorption, distribution, metabolism and elimination of a xenobiotic substance, are of great importance for the assessment of safety of chemical compounds. For cosmetics, in particular, skin permeability and dermal absorption are of great interest as many products are applied dermally, *e.g.* shampoo, skin cream, make-up. Approaches, such as those proposed by Lipinski *et al.* (2001) for oral drug absorption and Potts and Guy (1992) for skin permeability, have been refined and

adjusted towards 21st Century Regulatory Toxicology challenges, to support regulatory decision-making (refer to Chapters 3 and 4).

- **Mechanistically based modelling:** There is a myriad of different modes of action in the area of toxicology. For example genotoxicity, including mechanisms such as DNA adduct formation, and different enzyme- and receptor-mediated toxicities, very often extensions of pharmacological research, have been investigated in the last century. Highly relevant to cosmetics is hepatotoxicity caused by chronic exposure. Different mechanisms of toxicity have been studied in this thesis, especially nuclear receptor interaction associated with hepatosteatosis. Ligands for these receptors can lead to adverse effects even if absorbed in small quantities – particularly if absorbed over an extended time period. Therefore large *in vitro* datasets have been investigated additionally to *in vivo* and clinical data to develop screening tools for potential hepatotoxicants (refer to Chapters 5, 6 and 7).

Beyond the models and tools built, an overall aim is to propose ideas how to build and interpret new models, and of course how to use them in combination to support safety assessment in the consumer care industry. This research has been undertaken within the COSMOS project and hence, it is funded by the European Commission and Cosmetics Europe.

2. Methods for assigning confidence to toxicity data with multiple values – identifying experimental outliers*

2.1. Introduction

High quality data are preferred for model development in predictive toxicology. They are also required as a benchmark in the assessment of alternative assays and to enable analysis of toxicological pathways. Recently, further toxicity data have become available through sources such as the OECD QSAR Toolbox, release of information from dossiers submitted to the ECHA, the OECD eChemPortal and many other sources (Cronin and Schultz, 2003; Fourches *et al.*, 2010; Przybylak *et al.*, 2012; Péry *et al.*, 2013).

When using these expanding resources of toxicity data for risk assessment purposes and modelling, the quality and reliability of the data must be assessed. For instance, a given dataset could be too "poor" in terms of quality for QSAR modelling but still satisfactory for the prioritisation of chemicals for testing or regulatory classification and labelling. Whilst QSAR modelling is dependent on a sensitive statistical analysis, *e.g.* multivariate regression, to define reasonable descriptors, regulatory use of toxicity data may only need a rough estimation of hazard as a worst-case assumption, with extrapolation factors being applied (Nendza *et al.*, 2010).

Reliability is the measure of the extent of repeatability and reproducibility of a toxicity test for a particular chemical (OECD, 2003). As repeatability and reproducibility are not known for most data, a variety of approaches to assign reliability and confidence are used. Assessing data quality in predictive and computational toxicology is, however, a difficult task (Klimisch *et al.*, 1996; Przybylak *et al.*, 2012; Yang *et al.*, 2013). There are a number

* This chapter is based on a published article (Steinmetz *et al.,* 2014)

of established criteria to ascertain the reliability of toxicity data. The most commonly applied are those proposed by Klimisch *et al.* (1996). These authors discussed data attributes such as reliability, relevance and adequacy and provided a scoring system to categorise data into reliability classes:

1. reliable without restriction
2. reliable with restrictions
3. not reliable
4. not assignable

Przybylak *et al.* (2012) applied the Klimisch scheme and an updated scoring approach, based on ECHA guidance on information requirements and chemical safety assessment, to "real life" problems of toxicity data harvesting. In this work, the focus was on availability of information, consistency of study design, adherence to Good Laboratory Practice (GLP), test chemical identity and toxicological data.

Whilst reliability (the backbone of an experiment and the resulting toxicity data), and relevance (the usefulness of the resulting data for the desired purpose such as risk assessment) in principle require interpretation by experts, the determination of the reliability of data can be as well supported by methods of "weighting" the data (Klimisch *et al.*, 1996; Przybylak *et al.*, 2012; Yang *et al.*, 2013). When dealing with large sets of toxicity data, from multiple sources, there is often more than a single data entry for each compound. In this investigation these data entries are referred to as "conflicting data". Even for a well-defined assay such as the acute fish toxicity test, considerable variability in potency is seen within the results for the same compound (Hrovat *et al.*, 2009). If toxicity data are to be extracted for modelling from the increasing number of databases then criteria to identify reliable values are required. In particular, it would be helpful to be able to score data for reliability. In this way, it may be possible to rationally combine what may be considered to be low quality data to obtain a more reliable score.

Another interesting aspect of the quality control and assurance of toxicity data was investigated by Ruusmann and Maran (2013) who undertook an extended data harvest for the *Tetrahymena pyriformis* inhibition of growth assay (the Tetratox assay). They analysed the "timelines" associated with the reporting of chemical structures and experimental data and so examined when, and how, certain data were reported in the scientific literature over time. These authors came to the conclusion that mathematical manipulation (rounding, building averages *etc.*) and, of course, human error has led to differences in the data reported. For some compounds, there are many toxicity data from the same test; there is, however, no unified strategy to select which of the data to use, or how to use them. Often these toxicity data for the same compound have a normal distribution that makes it relatively easy to define a representative value via the median or arithmetic mean. Data which fall outside the normal distribution may be termed "data outliers", *i.e.* they may be subject to considerable experimental error. Figure 2.1 illustrates the issues of the presence of a data outlier in reducing certainty in the calculation of the mean or median.

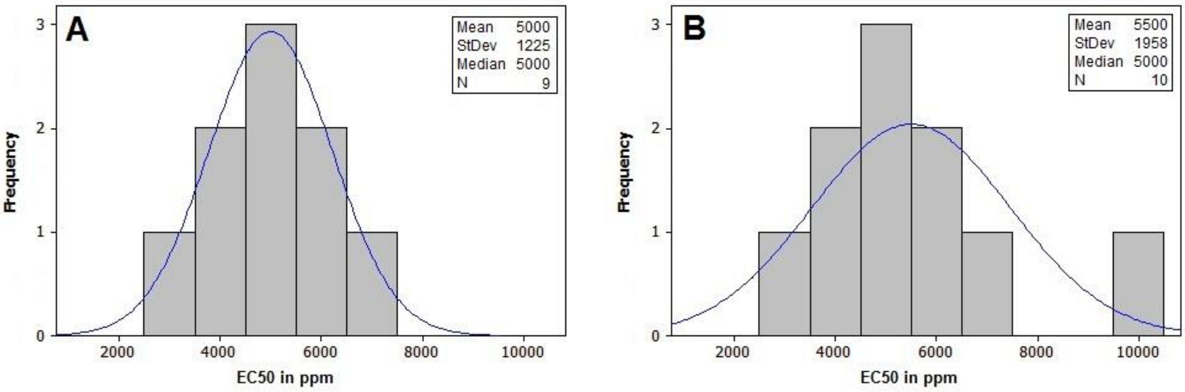


Figure 2.1: Normal bell-shaped distribution bell for a sample dataset (representative EC_{50} values for different sources for one compound) with an "optimal" normal distribution (A) and with a dataset containing an outlier in the upper range (B) demonstrating the skew it may bring to the distribution in addition to the elevated Standard Deviation (StDev)

In principle, the arithmetic mean is a good way to consolidate associated data points to a single value. Here, every data point is taken into consideration, in equal parts, to build a new value – the arithmetic mean. In contrast the median is the middle value of a distribution. When dealing with high individual spreads, the median is the more stable approach (Rowe, 2007).

Confidence scoring is based on the number and variability of conflicting data. In this context, the relative standard deviation (RSD; sometimes referred as coefficient of variation), as a quotient of standard deviation and arithmetic mean, expresses the variability of a dataset of toxicity values for one compound (Rowe, 2007). Thus a high number of entries per compound and a low RSD lead to high confidence and *vice versa*.

In order to investigate the role of variability in toxicity databases and explore the possibility of applying statistical approaches to identify reliable toxicity data, historical toxicity data, measured in the Microtox assay (and its precursors), were considered. Such data have been published since the early 1980s (*e.g.* Dutka and Kwan, 1981; Chang *et al.*, 1981; Bulich *et al.*, 1981; King and Painter, 1981; Curtis *et al.*, 1982; Yates and Porter 1982; DeZwart and Slooff, 1983; Ribo and Kaiser, 1984) and by the company Beckman Instruments, Inc. (now Beckman Coulter, Inc.). The *Aliivibrio fischeri* toxicity assay (Microtox) is a standardised aquatic toxicity test based on the marine bacterium *A. fischeri* (also known as *Photobacterium fischeri* and *Vibrio fischeri*). The photo-luminescent bacteria are exposed to a chemical at different concentrations with the reduction of light emitted being regarded as the effect. The results from the Microtox assay include the concentration of a compound where light intensity is reduced by 50% (EC_{50}). The pT value is the negative logarithm of the EC_{50}, for the purposes of this chapter the units are in mmol L^{-1}, and the measurement has historically been taken at different exposure times (5, 15 and 30 minutes) (Kaiser and Palabrica, 1991). As the *A.*

fischeri toxicity assay is a well-standardised study, little experimental variability is assumed. However, there are some data, which can be regarded as low quality, which may be attributed to inter-laboratory variation and experimental error. Cronin and Schultz (1997) furthermore suggested that there is no significant influence of exposure times (5, 15 and 30 minutes) on the toxicity of compounds, which act by non-polar narcosis. In this study non-polar narcosis is taken to be a non-specific mechanism of acute toxicity brought about by membrane perturbation (van Wezel and Opperhuizen, 1995; Ellison *et al.*, 2008). As such, in aquatic toxicology, it is well established that the logarithm of the octanol-water partition coefficient (log P) is strongly related to the toxic potency of such compounds (Verhaar *et al.*, 1992; Cronin *et al.*, 1998; Zhao *et al.*, 1998).

The aim of this investigation was to develop methods and criteria to quantify the reliability of toxicity data when multiple values from different experimental determinations are available for the same chemical. To achieve this, historical literature data, measured in the *A. fischeri* assay were used. The effect of data quality was assessed by analysing log P-based QSARs for non-polar narcosis. Specifically, this involved: updating the Microtox data compilation of Kaiser and Palabrica (1991); identifying the non-polar narcotics within those data; developing statistical criteria for determining data reliability and; the development of log P-based QSARs for non-polar narcosis with different levels of quality / confidence score.

2.2. Methods

2.2.1. Toxicity data

A literature search was conducted to obtain a large *A. fischeri* toxicity dataset (refer to Appendix A.1). The toxicity data of the compilation published by Kaiser and Palabrica (1991), consisting of 1350 *A. fischeri* EC_{50} entries, were supplemented with more recently published *A. fischeri* assay data. Subsequently salts, mixtures, polymers, organometals and duplicates (data from the same study) were removed from the dataset. In the literature where different values for different exposure times were given, the result for the longest exposure time within the interval of 30 min ≥ t ≥ 5 min) was taken. Toxicity data corresponding to shorter or longer exposure times (t < 5 min; t > 30 min) were not considered. All EC_{50} values were converted to the negative logarithm of the EC_{50} in mmol L^{-1}. In common with the original terminology of Kaiser and Palabrica (1991), this was termed "pT".

2.2.2. Chemical structures

Simplified Molecular Input Line Entry Specification (SMILES) strings for all compounds from the *A. fischeri* toxicity dataset were retrieved from the ChemSpider and ChemIDplus online chemical databases. Furthermore InChIKeys were created with the OpenBabel software (OpenBabel, 2013) using SMILES strings as input to identify identical compounds. Subsequently the dataset was cleaned, *i.e.* salts, polymers, inorganics and redundant data (data already identified in other literature) were omitted from the dataset, so that only organic compounds with at least one unique toxicity value per compound remained.

2.2.3. Assignment to mechanism of action for acute aquatic toxicity

To obtain information about the mechanism of action, the SMILES strings of the toxicity data were entered into the Toxtree software v2.5.0 (IDEAconsult, 2013). Toxtree holds a variety of toxicologically relevant decision trees used either for classification of chemicals or the elucidation of potential MoAs. The main purpose of the software is the classification of chemicals in the area of human and aquatic toxicology (IDEAconsult, 2014a). In this context the modified Verhaar algorithm (Verhaar *et al.*, 1992; Verhaar *et* *al.*, 2000; Enoch *et al.*, 2008) was applied to identify compounds acting as non-polar narcotics (Class 1). These compounds were subsequently extracted from the *A. fischeri* dataset.

2.2.4. Exposure times

The influence of exposure times (5, 15 and 30 minutes) on the toxicity value obtained was compared for the compounds identified as acting by non-polar narcosis. According to Cronin and Schultz (1997), there was little or no influence of exposure time on the toxicity value, such that it is justified to merge *A. fischeri* toxicity data from 5, 15 and 30 minute time points. To verify this, triplets of data, *i.e. A. fischeri* toxicity values for each compound at 5, 15 and 30 minutes were identified and compared via linear regression (refer to Appendix A.1).

2.2.5. Calculation of physico-chemical properties

Calculated log P values were obtained from KOWWIN v.1.68 from the freely available EPI Suite 4.11 (EPA, 2013) software. Molecular weights were calculated from the MOE 2011.10 software (MOE, 2013).

2.2.6. QSAR analysis

The relationships between pT and log P for non-polar narcotics were examined using linear regression analysis in Minitab (Minitab, 2013). The data were plotted and a linear equation including n (number of data points), S (standard error), R^2_{adj} (coefficient of determination, adjusted for the number of degrees of freedom) and F and t statistics were generated (Livingstone, 2004).

2.2.7. *Statistical analysis of toxicity data*

2.2.7.1.Omission of toxicity data outliers

If the total number of different EC_{50} values for one compound was greater than five ($n > 5$), and single entries were outside of the range of ±50% of the median, then these so-called “data outliers” were omitted from the dataset. A truncated mean was calculated from the remaining EC_{50} values and the hence pT value was re-calculated (as shown in Figure 2.3). If $n \leq 5$ for one compound, the arithmetic mean of all EC_{50} values was used to calculate the final pT value for a compound. As a result, there is a single pT value per compound which was used for QSAR modelling.

2.2.7.2.Confidence scoring

A confidence score (CS) was assigned to the pT value for each compound. A compound with a single entry ($n = 1$) was assigned a confidence score of one ($CS = 1$). For $n > 1$ the confidence scores were calculated from the number of entries per compound (n) divided by the relative standard deviation (RSD), where RSD is the ratio of the standard deviation (SD) and arithmetic mean ($\bar{x}$):

$$RSD = \frac{SD}{\bar{x}} \qquad \text{(Eq. 2.1)}$$

$$CS = \frac{n}{RSD} \qquad \text{(Eq. 2.2)}$$

For the toxicity data, two confidence score (CS) thresholds were investigated here, *i.e.* where $CS(1) > 5$ and $CS(2) > 15$ respectively.

2.3. Results

The outcome of the retrieval and the cleaning of the *A. fischeri* toxicity data and subsequent identification of non-polar narcotics are described below. Furthermore, analysis of toxicity values with respect to exposure time, to investigate the significance of the duration of the assay, as well as the analysis with respect to data quality from statistical assessment, is reported.

2.3.1. Retrieval and cleaning of A. fischeri toxicity data

A literature review revealed many sources of toxicity data from the *A. fischeri* assay. These supplemented the compilation of Kaiser and Palabrica (1991), which provided references to 28 papers and approximately 1350 data. A further 600 data were obtained giving a total of 1944 *A. fischeri* toxicity entries for over 1300 compounds. After cleaning the data to remove toxicity values for ambiguous structures, salts, polymers etc. a total of 1813 entries for 1227 compounds were obtained. This complete dataset is available as an Excel table in Appendix A.1.

2.3.2. Identification of non-polar narcotics

Of the cleaned 1813 *A. fischeri* toxicity data for the 1227 different chemical compounds, 203 of these compounds were identified as having a very strong probability of acting by the non-polar narcosis mechanisms of action. The initial assignment was performed using the modified Verhaar rules, which provide a robust starting point (Verhaar *et al.*, 1992; IDEAConsult, 2013). It is appreciated that this may be a conservative approach and that more compounds in the dataset may fall within the non-polar narcosis domain, however, the Verhaar rules were utilised to provide a defensible and repeatable method for the selection of non-polar narcotics.

2.3.3. Analysis of A. fischeri toxicity data with respect to exposure time

For the purposes of considering data quality it would be highly beneficial to be able to combine data from the different time points. In order to assess the feasibility of this, 99 of the non-polar narcotic triplets, *i.e.* compounds identified as being non-polar narcotic, with data for all three exposure times, were considered. Figure 2.2 and Equations 2.3 to 2.5 show the relationships between toxicity data from all three time points. There are no significant differences between the toxicity data as illustrated by an intercept approaching zero and a slope of unity for the regression equations between the exposure times. As a result of this analysis, confirming insignificant differences between the toxicity data for different exposure times, the data for different time points were combined, with the longest exposure being used by preference:

$$\mathrm{pT(30min)} = -\,0.04 + 0.97\ \mathrm{pT(5min)} \qquad \text{(Eq. 2.3)}$$

$$n = 99,\ R^2_{adj} = 0.99,\ S = 0.14,\ t = 90.0,\ F = 8020$$

$$\mathrm{pT(15min)} = -\,0.02 + 0.99\ \mathrm{pT(5min)} \qquad \text{(Eq. 2.4)}$$

$$n = 99,\ R^2_{adj} = 1.00,\ S = 0.09,\ t = 133,\ F = 17700$$

$$\mathrm{pT(30min)} = -\,0.02 + 0.98\ \mathrm{pT(15min)} \qquad \text{(Eq. 2.5)}$$

$$n = 99,\ R^2_{adj} = 1.00,\ S = 0.08,\ t = 154,\ F = 23600$$

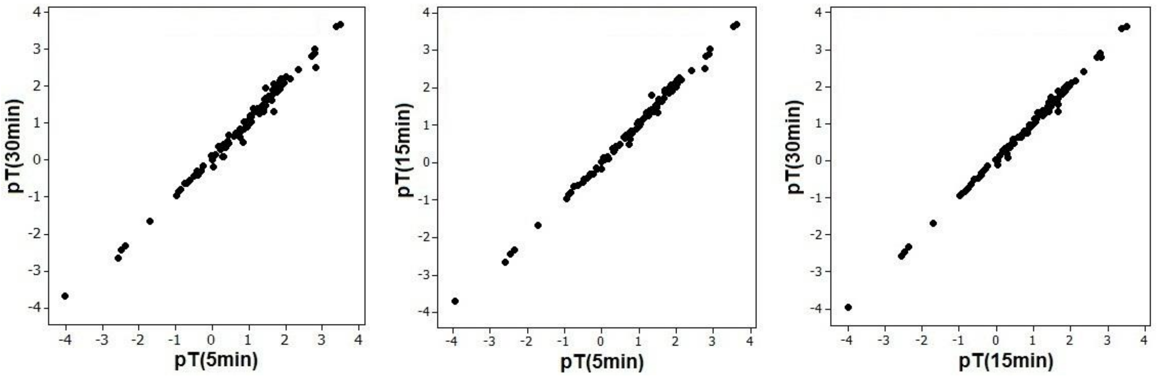
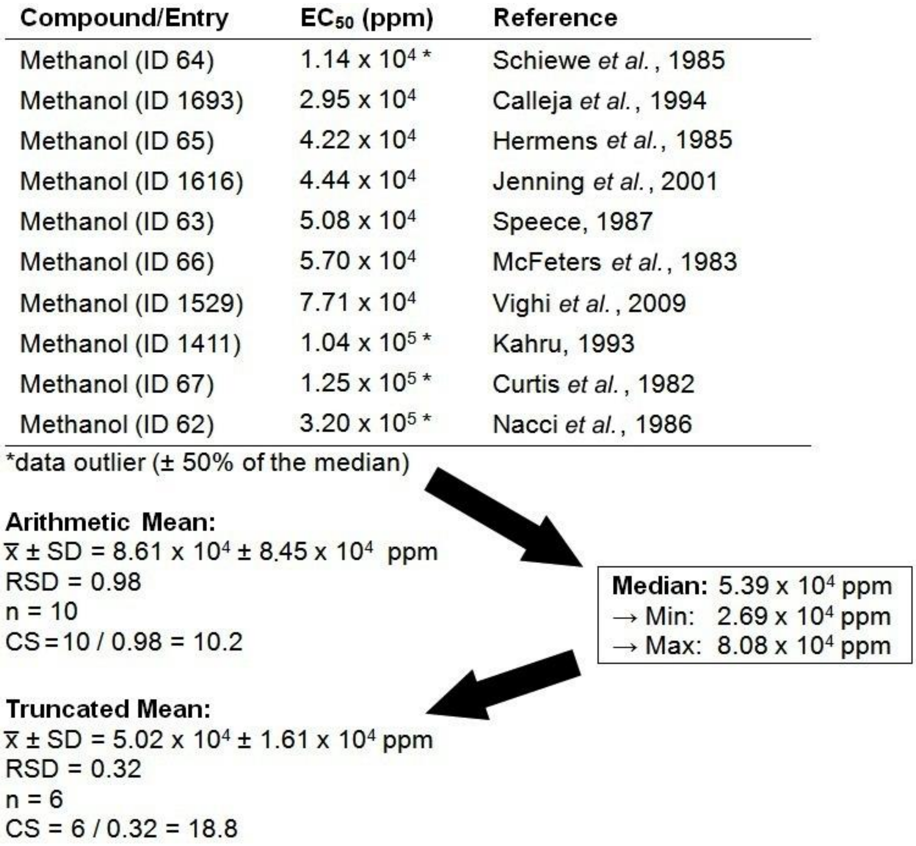


Figure 2.2: Comparison of the effect of exposure times (5, 15 and 30 min) of the pT values for 99 non-polar narcotics (refer to Appendix A.1)

2.3.4. *Analysis of A. fischeri toxicity data with respect to data quality*

Many compounds had more than a single toxicity value. In order to investigate the concept of data reliability and quality, the (truncated) mean was calculated for these compounds. Taking methanol as an example, toxicity test data are available from ten separate publications. The full set of test results and data are shown in Figure 2.3. Consideration of all ten data points gives a mean of 8.61 x 10^4 ppm and SD of 8.45 x 10^4 ppm. When data outliers were omitted, with a tolerance of ± 50% of the median (Min, Max), this resulted in only six mid-range entries with a mean of 5.02 x 10^4 ppm and SD of 1.61 x 10^4 ppm remaining for methanol (refer to Figure 3). As a consequence of the removal of the "data outliers" there is a predictable reduction in (relative) standard deviation and increase in the confidence score (CS). The ID numbers in Figure 3 are identifiers used during the data collection (refer to Appendix A.1).



Compound/Entry	EC_{50} (ppm)	Reference
Methanol (ID 64)	1.14 x 10^4 *	Schiewe *et al.*, 1985
Methanol (ID 1693)	2.95 x 10^4	Calleja *et al.*, 1994
Methanol (ID 65)	4.22 x 10^4	Hermens *et al.*, 1985
Methanol (ID 1616)	4.44 x 10^4	Jenning *et al.*, 2001
Methanol (ID 63)	5.08 x 10^4	Speece, 1987
Methanol (ID 66)	5.70 x 10^4	McFeters *et al.*, 1983
Methanol (ID 1529)	7.71 x 10^4	Vighi *et al.*, 2009
Methanol (ID 1411)	1.04 x 10^5 *	Kahru, 1993
Methanol (ID 67)	1.25 x 10^5 *	Curtis *et al.*, 1982
Methanol (ID 62)	3.20 x 10^5 *	Nacci *et al.*, 1986

*data outlier (± 50% of the median)

Arithmetic Mean:
$\overline{x} \pm SD = 8.61 \times 10^4 \pm 8.45 \times 10^4$ ppm
RSD = 0.98
n = 10
CS = 10 / 0.98 = 10.2

Median: 5.39 x 10^4 ppm
→ Min: 2.69 x 10^4 ppm
→ Max: 8.08 x 10^4 ppm

Truncated Mean:
$\overline{x} \pm SD = 5.02 \times 10^4 \pm 1.61 \times 10^4$ ppm
RSD = 0.32
n = 6
CS = 6 / 0.32 = 18.8

Figure 2.3: Microtox toxicity data for methanol and analysis to identify "data outliers" and calculate the confidence score

2.3.5. *Investigation of the effect of data quality on QSARs*

The same principle of data outlier omission and confidence scoring, as undertaken for methanol, was applied to the whole *A. fischeri* dataset. Therefore, for compounds with more than five data points ($n > 5$), truncated means with a reduced influence of outliers and higher confidence scores (CS) were created where such outliers were identified.

Figure 2.4 shows the relationship of toxicity and log P for non-polar narcotics, where the effect of data outlier omission and confidence scoring, have been applied: In the left column (Figure 2.4A, 2.4C and 2.4E) no data outlier omission was applied and in the right column (Figure 2.4B, 2.4D and 2.4F) data outlier omission was applied. In first row (Figure 2.4A and 2.4B) no confidence score threshold was applied, in the second row (Figure 2.4C and 2.4D) a confidence score threshold of CS(1) > 5 was applied and in the third row (Figure 2.4E and 2.4F) a confidence score threshold of CS(2) > 15 was applied.

The linear correlations between pT and log P in the first row of Figure 2.4A and 2.4B are weak (R^2_{adj} of 0.50 and 0.51 respectively). The confidence score thresholds for the second and the third row (C, D and E, F) allowed only “high quality” data to be plotted. Figures 2.4C and 2.4D (CS(1) > 5) and Figures 4E and 4F (CS(2) > 15) show stronger linear correlations ($R^2_{adj} > 0.79$) than Figures 2.4A and 2.4B. As the confidence score threshold of CS(2) is stricter than CS(1), the second row (Figure 2.4C and 2.4D) contains more compounds ($n = 40$ and $n = 43$ respectively) than the third row (Figure 2.4E and 2.4F) with $n = 12$ and $n = 17$ respectively. Overall the following six QSAR equations (2.6 to 2.11), referring to Figure 2.4A to 2.4F, were developed:

A: $$pT = 0.68 \log P - 1.14 \qquad \text{(Eq. 2.6)}$$
$$n = 203,\ R^2_{adj} = 0.50,\ S = 0.95,\ t = 14.3,\ F = 204.0$$

B: $$pT = 0.68 \log P - 1.11 \qquad \text{(Eq. 2.7)}$$
$$n = 203,\ R^2_{adj} = 0.51,\ S = 0.93,\ t = 14.5,\ F = 211.0$$

C: $$pT = 1.08 \log P - 2.21 \qquad \text{(Eq. 2.8)}$$
$$n = 40,\ R^2_{adj} = 0.81,\ S = 0.65,\ t = 13.0,\ F = 168.0$$

D: $$pT = 1.08 \log P - 2.20 \quad \text{(Eq. 2.9)}$$

$n = 43$, $R^2_{adj} = 0.81$, $S = 0.63$, $t = 13.6$, $F = 185.0$

E: $$pT = 1.12 \log P - 1.92 \quad \text{(Eq. 2.10)}$$

$n = 12$, $R^2_{adj} = 0.79$, $S = 0.80$, $t = 6.5$, $F = 42.8$

F: $$pT = 1.23 \log P - 2.31 \quad \text{(Eq. 2.11)}$$

$n = 17$, $R^2_{adj} = 0.83$, $S = 0.75$, $t = 9.0$, $F = 81.3$

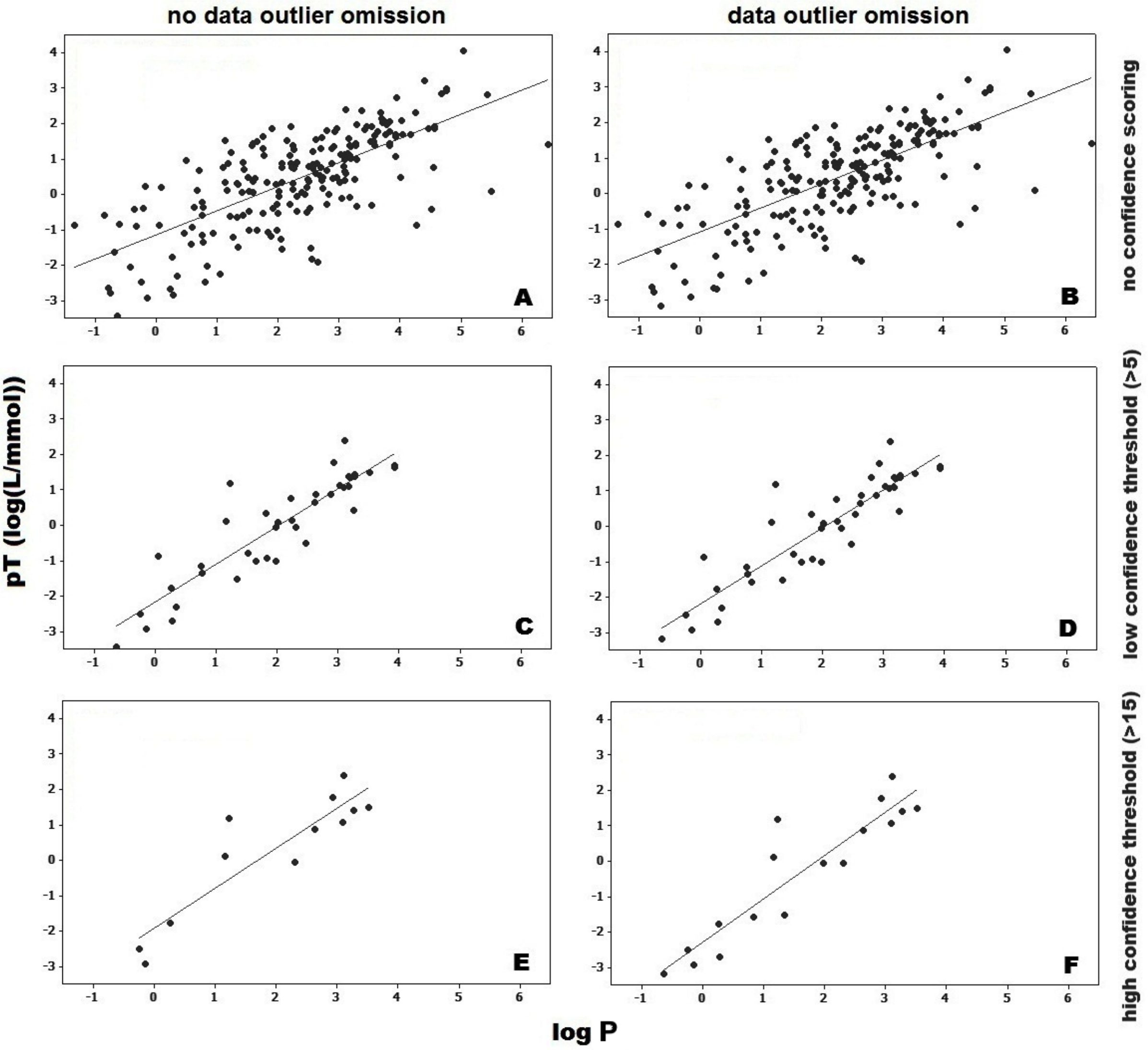


Figure 2.4: Relationship between *A. fischeri* toxicity (pT) and log P: Figure 2.4A containing pT arithmetic means (n = 203); Figure 2.4B containing pT arithmetic means after an median-based data outlier omission (n = 203); Figure 2.4C containing pT arithmetic means with a confidence filter (CS(1) > 5; n = 40), Figure 2.4D containing pT arithmetic means with a confidence filter (CS(1) > 5; n = 43) after an median-based data outlier omission; Figure 2.4E containing pT arithmetic means with a confidence filter (CS(2) > 15; n = 12); Figure 2.4F containing pT arithmetic means with a confidence filter (CS(2) > 15; n = 17) after median-based data outlier omission

2.4. Discussion

There is an increasing availability of toxicity data from various relatively standardised assays, which have been brought together in REACH submissions, the OECD QSAR Toolbox, eChemPortal and a variety of other freely available and accessible resources *etc*. An often asked question is which value is representative when multiple data points are available for the same compound from the same test. This analysis has provided a means to evaluate multiple data entries and thus, in part at least, begins to answer that question as well as supporting regulatory decisions and the creation of robust datasets for model building and read-across. The *A. fischeri* assay is a well-standardised technique; it is essentially a simplistic cytotoxicity assay, meaning variability of measurements within and between laboratories should be low. Within the compiled dataset, the non-polar narcosis-associated toxicity data from the *A. fischeri* assay were chosen due to the number of available data.

For compounds with multiple toxicity data points, statistical analysis was undertaken (refer to Figure 2.3). The purpose of this was to identify, and hence remove, data outliers. These outliers were selected on an empirical and statistical basis. It was not possible to determine if there were experimental anomalies as the original "study reports" were not available, as is common for online databases of toxicity values. Figure 2.3 illustrates this concept; the approach of removing statistical outliers is transparent and clear. It provides a useful analysis of the data, especially when combined with the confidence score (CS) discussed below. In this analysis an arbitrary cut-off of 50% of SD was applied. This was identified following a process of trial and error (results not shown in this analysis) but could be adapted *e.g.* if more or less variability was considered acceptable for a test.

The relationship between toxicity (pT) and lipophilicity (log P) for non-polar narcotics is shown in Figure 2.4. The resulting QSAR equations are not significantly different from those published previously (Cronin and Schultz, 1997; Vighi *et al.*, 2009). When no data restrictions are selected (Figures 2.4A and 2.4B) the resultant QSARs are less precise in terms of statistical fit and robustness as compared to those developed using certain data selections (from Figure 2.4C, 2.4D, 2.4E and 2.4F). Therefore an assessment and quantification of data quality will assist in the development of more robust QSARs and computational models. To select the data in an objective manner, arbitrary confidence scores (CS(1) > 5 and CS(2) > 15) were taken as thresholds. Principally, the higher the CS the more evidence, in terms of similar results for one compound is available, *i.e.* the result is regarded as more trustworthy than a compound with a low CS. Even with low numbers of data (n) and high relative standard deviation (RSD), it is impossible for CS to fall below one (CS < 1). Usually CS = 1 means that there is only one data point per compound. This can be regarded as the lowest (statistical) confidence that can be attached to a datum point. At this point it must be stressed that confidence does not necessarily relate to reliability. For many compounds the results of a single toxicity test may be highly reliable, it is simply that there is lower confidence as the value has not been replicated. As such, the plots in Figure 4C to 4F show a smaller number of data points (n = 40, 43, 12 and 17 respectively) due to a filtration process based on confidence scoring. In Figures 2.4C and 2.4E only data with high confidence scores (CS > 5 and > 15 respectively) were considered. Both show the strong linear relationship that is fundamental to non-polar narcosis (Ellison *et al.*, 2008). The inclusion of data with the lower confidence threshold (refer to Figure 2.4D) allows more data points (compounds) to be considered in the QSAR.

The data outlier omission on its own leads to more centralised (closer to the median) values, it also reduces the variability/spread of an associated dataset and so increases the corresponding CS value. This leads to more confidence being associated with the data points in Figures 2.4D and 2.4F compared to Figures 2.4C and 2.4E respectively. It is open for discussion which of Figures 2.4D (equation 2.9) or 2.4F (equation 2.11) is the better model for non-polar narcosis: on the one hand Figure 4F has a better statistical fit ($R^2_{adj} = 0.83$), in contrast Figure 2.4D incorporates more contributing data points.

The effect of the data outlier omission between Figures 2.4A and 2.4B is only marginal (R^2_{adj} 0.50 and 0.51 respectively). Both Figures contain 203 data points, but in Figure 2.4B fewer data points are orientated towards the line of best fit than in Figure 2.4A, based on the stabilising effects of the data outlier omission and truncated mean respectively. The effect is negligible as most compounds have only one data point, *i.e.* one single EC_{50} entry. The strength of the QSARs reported in Figures 2.4D and 2.4F is that data outlier omission incorporates more data points (than Figures 2.4C and 2.4E). The greater the number of data points contributing to a correlation or QSAR, the greater the weight of evidence for a correlation and QSAR respectively. This confirms that data outlier omission is a useful tool, particularly in combination with CS thresholds.

The confidence scoring, particularly when combined with the median-based data outlier omission, is a mathematical method to assess reliability of toxicity data where there are multiple entries for a single value. The metric confidence scores provided by this method can be used as thresholds or as pre-factors for weighting as described by Przybylak *et al.* (2012) or Yang *et al.* (2013). The statistical tools used, *i.e.* data outlier omission and confidence scoring, show an improvement of the model by defining “reproduced” (multiple and similar) data to be more reliable than single or non-reproducible data.

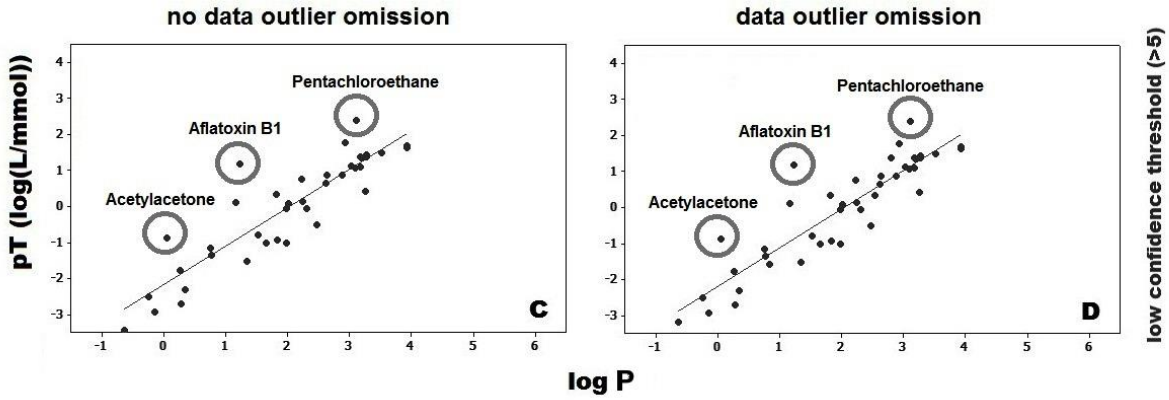


Figure 2.5: Detailed examination of Figures 2.4C and 2.4D showing compounds with excess toxicity

The investigation of Figures 2.4C and 2.4D (refer to Figure 2.5) revealed three data points with a high pT to log P ratio, indicating excess toxicity above non-polar narcosis.

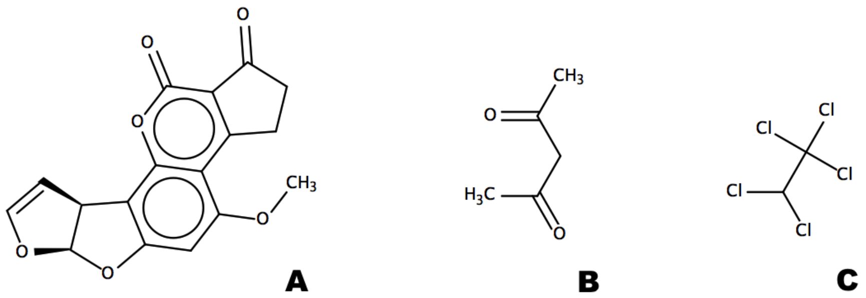


Figure 2.6: Chemical structures of aflatoxin B1 (A), acetylacetone (B) and pentachloroethane (C)

These three compounds are aflatoxin B_1, acetylacetone and pentachloroethane which are known to have toxicity-related modes of action other than non-polar narcosis (refer to Fig. 2.6). Aflatoxins are well known for their mutagenic, teratogenic, carcinogenic and in higher doses hepatotoxic effects. Typically aflatoxin is activated by enzymes of the cytochrome P450 family to an epoxide, which reacts with macromolecules including DNA, RNA and proteins. This mechanism is relevant not only for cancer and liver disease but also for direct cytotoxic effects (Wehner *et al.*, 1978; McLean an Dutton, 1995; Frisvad *et al.*, 2006). In 2004, acetylacetone (pentane-2,4-dione) was identified as genotoxic and subsequently banned by the EC as a food additive (EC, 2005). According to the mechanistic studies of Enoch *et al.* (2011) it is likely to act as a Schiff base and protein binder. Pentachloroethane on the other hand showed no protein binding potential or any other related toxicity according to Enoch *et al.* (2011) but nevertheless most safety sheets label this substance as toxic. Pentachloroethane and hexachloroethane administered orally showed adverse effects, such as renal inflammation, hepatocellular carcinoma and increased lethality, in rats (Mennear *et al.*, 1982). Classifying these compounds as non-polar narcotics (Class 1) might not be an adequate decision by the Toxtree software and this provides some indication as to how to improve/modify the Verhaar rules in the Toxtree software. The excess toxicity, compared to the non-polar narcosis-associated toxicity, could be explained by these mechanisms (Lipnick *et al.*, 1987).

2.5. Conclusions

A transparent method to identify reliable toxicity data and values for modelling, as well as providing confidence for the use of multiple entries has been developed. This will assist in the harvesting of reliable toxicity values from what may be considered as variable quality data. The ability to assess conflicting toxicity data is important not only for developing models in computational toxicology, but also for the use of the increasing number of toxicity databases available. It should be remembered that even toxicity data with a low confidence score may be highly reliable (as a single, measured data point can be accurate *per se*), however the approaches proposed in this study will be beneficial to analysing some of the larger datasets that are increasingly becoming available. The analysis confirms that data with higher confidence, as defined in this study, produce more robust QSARs.

The results from Chapter 2 show that a novel method to assess data quality from a statistical perspective has been developed (and published as Steinmetz *et al.* 2014). This provides a means to evaluate the information from the increasing number of databases with multiple data for the same compound. This study has been extended in Chapter 3 by the use of CS-weighted regression to build QSAR models. Additionally a second dataset with high relevance to the assessment of cosmetic ingredients, *i.e.* a compilation of skin permeability coefficient data, has been investigated.

3. Using statistical confidence scoring to improve QSAR/QSPR modelling*

3.1. Introduction

As already explained in detail in Chapter 2, the assessment of biological/toxicological, data quality is crucial for many disciplines, *e.g.* QSARs, grouping and read-across (Nendza *et al.*, 2010; Przybylak *et al.*, 2012; Péry *et al.*, 2013). There are two general approaches to assess the quality of biological/toxicological data; based on the assessment of the reported testing information (GLP *etc.*) and based on statistical data quality (CS *etc.*) if multiple and comparable data are available (Steinmetz *et al.*, 2014).

As the assignment of CS values to toxicological data is not a common method to date, some theoretical examples are given to facilitate interpretation. Examples of calculations of CS values are provided in Table 3.1, illustratively representing scenarios of increasing CS values. Compound A is the default scenario (the most common occurrence whereby a compound has only a single experimental value), the CS is 1. Compound B has two relatively divergent data values, differing by an order of magnitude. Clearly there will be greater confidence for the toxicity value than for compound A, but the significant difference in the values introduces some uncertainty, raising CS marginally to 1.73 – in this way there is slightly greater confidence associated with two relatively different values than a single value. More data points are considered for compounds C and D, with increasing precision of the data values. Whilst compound C (n = 4) has more data than compound D (n = 3), the values are more divergent for C (represented by a higher RSD), thus the highest CS is calculated for compound D for which there are three data points, all relatively consistent in the light of the experimental error that might be associated with an experimental test. As such, compound D has the highest CS value.

*This chapter is based on a published article (Steinmetz *et al.,* 2015b)

Table 3.1: Four examples of compounds with multiple data in the same toxicity test (EC_{50}), along with statistical criteria and CS (refer to Appendix A.2)

Compound	EC_{50} (mol/L)	$\bar{x} \pm SD^a$	RSD[b]	n[c]	CS[d]
A	10	10 ± n/a	n/a	1	1[e]
B	1 10	5.50 ± 6.36	1.16	2	1.73
C	1 80 50 100	57.75 ± 43.05	0.75	4	5.37
D	1 2 1.4	1.47 ± 0.50	0.34	3	8.74

[a]mean and standard deviation

[b]relative standard deviation

[c]number of data

[d]confidence score

[e]CS of a compound with n = 1 is defined as 1 is the minimum value

As there is growing interest in techniques such as read-across to fill data gaps for regulatory purposes, and there is increasing accessibility to toxicity data through resources such as the OECD QSAR Toolbox to perform read-across, there are more possibilities to apply approaches such as the confidence scoring to improve the robustness of modelling. In this study the relevance of the CS approach has been assessed with regard to established QSARs for two endpoints, namely skin permeability coefficients and cytotoxicity for which large compilations of historical data are available.

3.1.1. Skin permeability

There have been many efforts to develop Quantitative Structure-Permeability Relationship (QSPR) models to predict various measures of dermal absorption (Scheuplein and Blank, 1971; Potts and Guy; 1992; Abraham *et al.*, 1997; Magnusson *et al.*, 2004; Dancik *et al.*, 2013; Khajeha and Modarress, 2014). The most recognised and applied QSPR to predict the skin permeability coefficient (k_p) is that developed by Potts and Guy in 1992 (Eq. 3.1). They used the molecular weight (MW), to account for the size of a permeant and log P, as a descriptor for lipophilicity, as parameters to model k_p following an analysis based on the Flynn data compilation (Flynn, 1990). The mechanistic explanation is that small, lipophilic compounds pass through the *stratum corneum*, the outermost layer of the skin, more easily than larger, more hydrophilic compounds. As shown in Figure 3.1, the transport termed diffusion occurs within the lipophilic phase between keratinocytes (Potts and Guy, 1992; Mitragotri *et al.*, 2011).

$$\log k_p \text{ (cm/h)} = -2.7 + 0.71 \log P - 0.0061 \text{ MW} \qquad \text{(Eq. 3.1)}$$

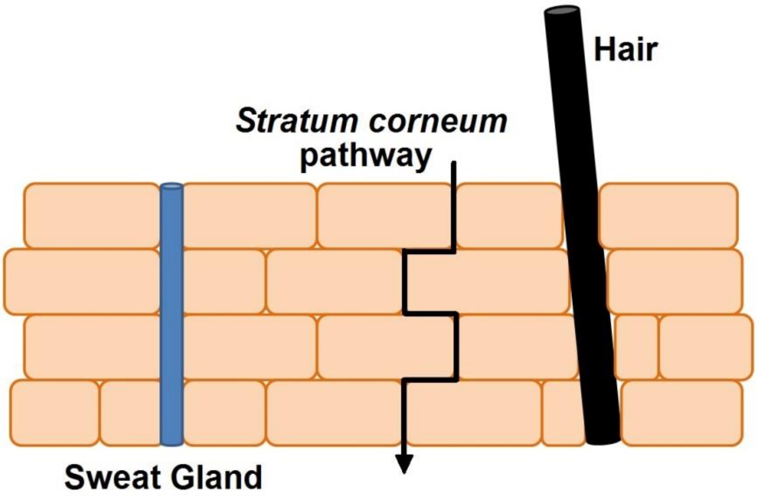


Figure 3.1: Diagrammatic representation of the skin showing the inter-cellular transport of xenobiotics through the *stratum corneum* (black arrow)

Despite the significance of Eq. 3.1, the quality of data compiled from the literature by Flynn, and hence the robustness of the Potts and Guy QSPR, has been the subject of considerable debate (Moss and Cronin, 2002; Johnson *et al.*, 1995). More human *in vitro* k_p data have inevitably become available in the two and half decades since Flynn's seminal publication (Moss and Cronin, 2002; Chauhan and Shakya, 2010; Chen *et al.*, 2013; ten Berge, 2014), thus the QSPR can be reassessed and rebuilt with a greater consideration and understanding of data quality.

3.1.2. Aquatic toxicology

As described more in detail in Chapter 2, there is a myriad of publically available eco-toxicological data, accessible for example via EPA's ECOTOX database (EPA, 2015b). The multitude of published *A. fischeri* data (as compiled in Steinmetz *et al.* (2014) and Chapter 2 respectively) was used within this study.

These two examples are illustrative of the possibilities of applying confidence scoring metrics to historical compilations of toxicity information. There are many open-access resources such as ChEMBL (2015), PDSP (2015), ACToR (EPA, 2015a), eChemPortal (OECD, 2015), TOXNET (NIH, 2015), so the life sciences, and in particular toxicology, has to deal increasingly with large and complex datasets (Zhu *et al.*, 2014). However, the task of assessing the toxicity data for quality, particularly when contradicting data are present, has not yet been accomplished. Any indication of the quality of data would be very helpful for purposes such as risk assessment, but more crucially for modelling including QSARs and read-across prediction (Przybylak *et al.*, 2012; Steinmetz *et al.*, 2014).

Therefore, the aim of this study was to investigate how using approaches for statistical data quality, *i.e.* CS, improves the development of QSAR/QSPR models. Specifically, the effect of directly incorporating the CS into the training and testing of the models was considered. To achieve this, the two endpoints described above were chosen for analysis, namely human *in vitro* skin permeability coefficients and the acute toxicity of compounds acting by a non-polar narcotic mechanism of action to *A. fischeri*. The reasons for choosing these endpoints included the fact that there were many historical data of variable and unknown quality, many compounds had been tested multiple times (a pre-requisite of applying the CS) and that there were simple, robust and mechanistically interpretable QSAR models for them. Thus, for both datasets, QSARs were constructed with and without reference to the CS.

3.2. Methods

3.2.1. Data harvest

In vitro skin permeability coefficients (k_p) were collected from the literature by compiling and subsequently merging four of the most comprehensive datasets of human skin k_p values (Moss and Cronin, 2002; Chauhan and Shakya, 2010; Chen *et al.*, 2013; ten Berge, 2014). All k_p values were converted to a standard unit (cm/h). Duplicate log k_p values (and those within ± 0.01 cm/h) were removed as they are most likely to be derived from the same source. SMILES and InChIKey strings were obtained for each compound from the ChemSpider website (RSC, 2014). The Flynn (1990) dataset contained k_p values for 94 compounds, however, 11 compounds (all substituted steroids) could not be identified by ChemSpider (RSC, 2014) or ChemIDplus (US NIH, 2014) and hence no SMILES were available to calculate descriptors. Since the structure of these compounds could not be completely verified they were excluded from subsequent analysis.

The *A. fischeri* data compilation from Chapter 2 (Steinmetz *et al.*, 2014) was used as the resource for the aquatic toxicology dataset. The chemical structures (as SMILES strings) of the comprised 1227 compounds were run through IDEAconsult’s Toxtree v2.6.6

(modified Verhaar) and non-polar narcotics were identified as being Class 1 according to the Verhaar scheme (Verhaar *et al.*, 1992; IDEAconsult, 2014a).

3.2.2. Descriptor generation

Log P and molecular weight (MW) were calculated for compounds in both datasets. The SMILES strings were used as the input format for all calculations. Log P was calculated with KOWWIN v1.68 within EPI Suite 4.11 (estimated values exclusively) (EPA, 2014). MW was calculated with the CDK node “molecular properties” within KNIME 2.9 (KNIME, 2014).

3.2.3. Calculation of confidence score (CS)

Confidence scores were calculated for the compounds in both datasets with regard to their k_p and EC_{50} values respectively. For compounds with more than a single experimental value, the arithmetic mean ($\overline{x}$), number of data points (n), SD and RSD were calculated with reference to data in the units stated in Section 3.2.1 and before logarithmic transformation. A CS was assigned to the arithmetic mean of the experimental values for each compound. Compounds with a single entry ($n = 1$) were assigned a confidence score of one ($CS = 1$). For compounds with $n > 1$, the CS was calculated as in Eq. 2.2.

3.2.4. Development of QSARs/QSPRs

Uni- and multivariate linear regression was performed on the datasets using R Studio 0.98.501.19 (R, 2014). Linear equations were generated and the following statistical, and other, criteria recorded: n (number of data points), S (standard error), R^2_{adj} (coefficient of determination, adjusted for the number of degrees of freedom), t statistics for the descriptors and F statistics for the equation. Regression analysis was performed to develop the QSARs for both datasets with and without weighting. Non-weighted regression analysis and weighted regression analysis was performed by applying CS values as weights in R using the default package lm {stats}. Weighting in linear regression means that each datum point is associated with a weight. A high weight strengthens, and a low value weakens, the impact of the data point towards the linear regression. In this manner, data for compounds associated with a high confidence score would be more heavily weighted in the regression analysis than compounds with a lower confidence score. Comparison of the statistics of the weighted and unweighted regression analysis provides an indication of whether CS is able to improve the robustness of models.

3.2.5. Evaluation of the predictivity of the QSARs/QSPRs

Statistical evaluation of the predictive capability of the CS-weighted QSAR and the CS-weighted QSPR was performed using 10-fold cross-validation, *i.e.* the compounds were ordered by k_p and pT respectively and every 10th compound was removed in turn leading to 10 training and validation sets. After applying the CS-weighted linear regression, the 10 datasets were investigated by the root mean square error (RMSE); predicted (f_i) versus experimental (y_i) values. Additionally the root mean square error adjusted for CS ($RMSE_{CS}$) was calculated (Eq. 3.2). It is expected that during the validation process, the $RMSE_{CS}$, which incorporates CS-weighting, will be lower than the standard RMSE. As the residuals (f_i - y_i) of the compounds with low CS values are weakened and the residuals of high CS compounds are strengthened, the sum of (squared) errors of the $RMSE_{CS}$ should be reduced in comparison to the conventional RMSE. The R script for $RMSE_{CS}$ cross-validation and the equations are available in Appendix C.2.

$$\mathrm{RMSE_{CS}} = \sqrt{\frac{\sum_i \mathrm{CS_i}(\mathrm{f_i} - \mathrm{y_i})^2}{\sum_i \mathrm{CS_i}}} \qquad \text{(Eq. 3.2)}$$

3.3. Results

Names of compounds, their InChIKeys and SMILES strings along with all k_p and pT values including references are available for the two datasets in Appendix A.2. Furthermore a glossary of relevant statistical equations is attached. In addition the R script for $RMSE_{CS}$ cross-validation is available in Appendix C.2.

3.3.1. Data harvest

The compilation of human *in vitro* k_p data resulted in 342 values for 226 different compounds. 55 of these compounds have more than a single k_p value. The log k_p values covered a broad range from -6.10 to 0.16. The structures included in the dataset were diverse in terms of physico-chemical properties and structure, *e.g.* solvents, alkaloids, steroids, sugars, nonsteroidal anti-inflammatory drugs *etc*. The solvents, sugars and steroids in particular, had many multiple data points. Water, with 13 different data points, had the most k_p values. The range of CS values is from 1 (for single entries) to 76.8 for chlorphenamine (based on two data points). Illustrating the capability of the CS approach, two compounds have moderately high CS values: the synthetic opioid sufentanyl with a CS value of 9.97 (based on two data points) and the cytostatic drug 5-fluorouracil with a CS value of 5.00 (based on four data points).

From the complete dataset of acute toxicity values to *A. fischeri*, comprising 1227 compounds, 203 were identified as potentially acting as non-polar narcotics according to the Verhaar scheme as implemented in Toxtree v2.6.6 (IDEAconsult, 2014a). A total of 418 different pT values were available for these compounds, with 71 of the 203 compounds having more than a single experimental value. pT values covered a broad range from -4.00 to 4.12. The structures included in the dataset were conservative in their structural diversity as they had been selected to represent the non-polar narcosis domain, including mainly solvents and medium- and long-chained alkanes, partly branched and halogenated, with only a few functional groups, such as hydroxyl- and amino-groups. The compounds investigated have a moderate spread of MW and log P and can generally be regarded as lipophilic (refer to Table 3.2). The CS spread shows the diversity between high confidence compounds, such as methyl isobutyl ketone (CS of 205 with 3 data points) and acetone (CS of 43.7 with 14 entries) and the single entry low confidence compounds (defined as CS = 1).

Table 3.2: Ranges of properties and CS for the two datasets considered in the analysis

	Human *in vitro* skin permeability coefficients	pT of non-polar narcotics to *A. fischeri*
MW (Da)	18.01 to 764.44	32.04 to 342.43
Log P	-6.76 to 8.39	-1.34 to 6.43
CS	1 to 76.8	1 to 205

3.3.2. Development of QSARs/QSPRs

QSAR/QSPR models were developed using linear regression with the experimental log k_p and pT as the dependent variables and log P and MW (for k_p only) as descriptors. Linear regression analysis was performed on both datasets, the resultant QSPRs for skin permeability coefficients based on the Potts and Guy approach (Eq. 3.3 (unweighted), Eq. 3.4 (weighted), Fig. 3.2) and the log P-based QSARs for the acute toxicity of non-polar narcotics to *A. fischeri* (Eq. 3.5 (unweighted), Eq. 3.6 (weighted), Fig. 3.3) are reported below.

3.3.2.1.QSPR: Modelling of the skin permeability coefficient

The unweighted QSPR for the dataset of skin permeability coefficients, using the Potts and Guy approach, was:

$$\log k_p = -2.45 + 0.40 \log P - 0.0045\, MW \qquad \text{(Eq. 3.3)}$$

$$n = 226,\ R^2_{adj} = 0.48,\ S = 0.82,\ t_{logP} = 13.3,\ t_{MW} = -8.97,\ F = 105$$

The reanalysis using CS-weighted k_p provided the following, similar, equation with improved statistical fit:

$$\log k_p = -2.51 + 0.50 \log P - 0.0051 \text{ MW} \qquad \text{(Eq. 3.4)}$$

$$n = 226,\ R^2_{adj} = 0.61,\ S = 1.39,\ t_{logP} = 18.7,\ t_{MW} = -9.25,\ F = 177$$

Experimental k_p values are plotted against predicted values from Eq. 3.4 in Figure 3.2, demonstrating good overall predictivity. In particular, there is a good fit about the line of unity, with a significant trend for compounds with the highest CS (represented by larger circles) to be well predicted, and the significant outliers tending to be compounds with low CS, *i.e.* single data points.

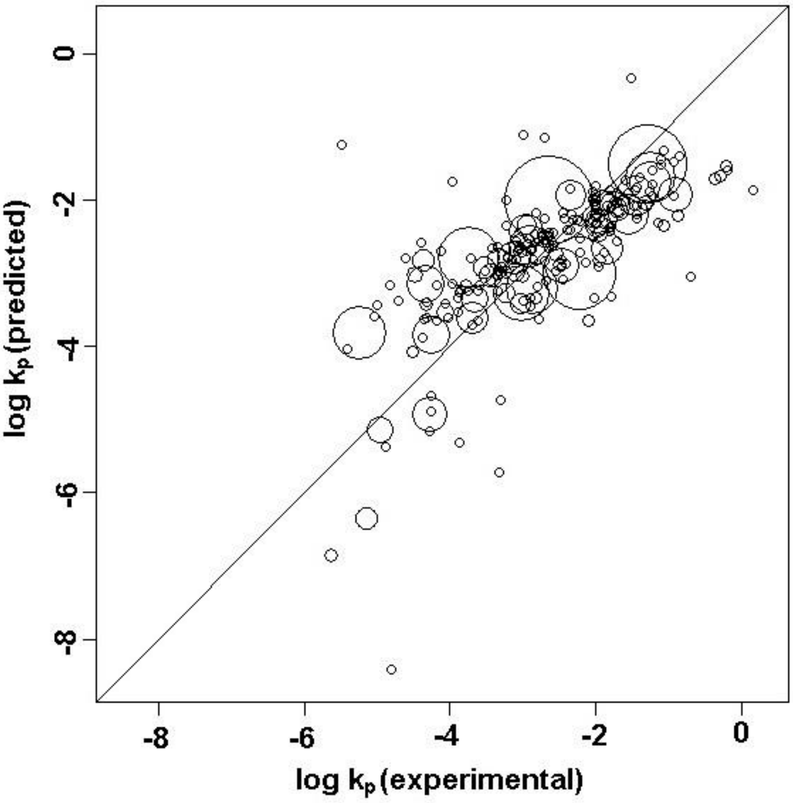


Figure 3.2: Experimental log k_p versus predicted log k_p from Eq. 3.4; the area of circles corresponding to the CS value; the larger the CS, the greater the area of the circle; the solid line indicating a slope of unity and an intercept of zero

The QSPR model represented by Eq. 3.4 was tested using 10-fold cross-validation. The statistical summary is presented in Table 3.3. Notably the $\mathrm{RMSE_{CS}}$ is lower than the RMSE.

Table 3.3: Statistical summary of 10-fold cross-validation based on Eq. 3.4 (skin permeability)

Training				Test	
Intercept	Log P	MW	R^2_{adj}	RMSE	$\mathrm{RMSE_{CS}}$
-2.51 ± 0.09	0.497 ± 0.026	-0.0051 ± 0.0004	0.61 ± 0.02	0.83 ± 0.21	0.79 ± 0.21

3.3.2.2.QSAR: Modelling of A. fischeri non-polar narcosis

The unweighted QSAR for the non-polar narcotics in the Microtox dataset, using a log P-based linear regression was:

$$pT = -1.14 + 0.68 \log P \qquad \text{(Eq. 3.5)}$$

$$n = 203,\ R^2_{adj} = 0.50,\ S = 0.95,\ t_{logP} = 14.3,\ F = 204$$

The reanalysis using CS-weighted pT provided the following equation with improved statistical fit:

$$pT = -1.67 + 0.92 \log P \qquad \text{(Eq. 3.6)}$$

$$n = 203,\ R^2_{adj} = 0.68,\ S = 1.77,\ t_{logP} = 20.9,\ F = 478$$

Figure 3.3 demonstrates the relative predictivity of Equation 3.6. There is a good fit about the line of unity, with a significant trend for compounds with the highest CS (represented by larger circles) to be well predicted, and the significant outliers tending to be compounds with low CS, *i.e.* single values (similar to k_p modelling).

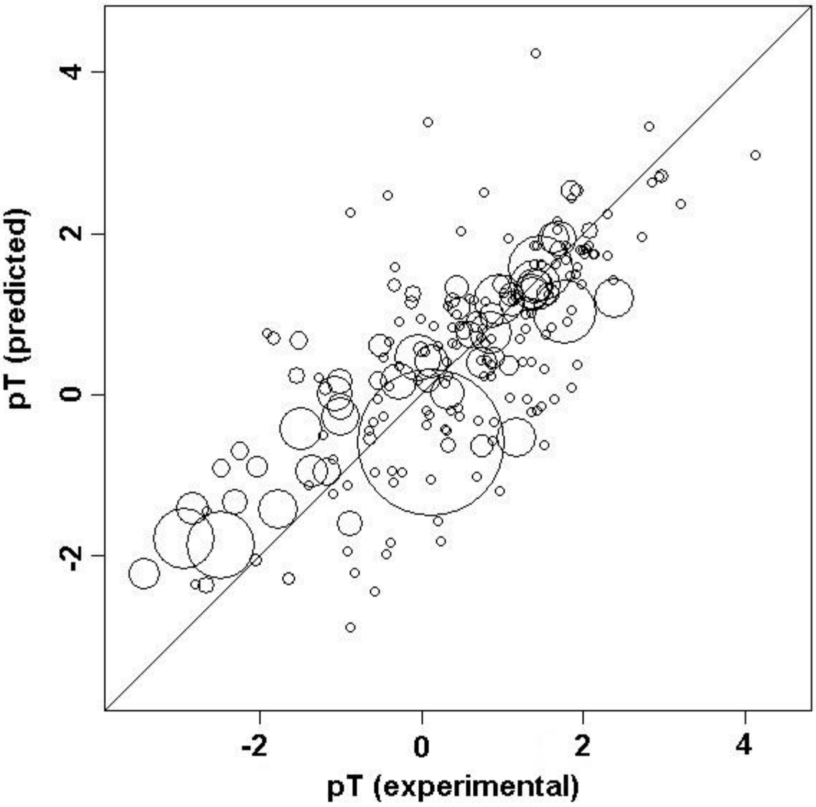


Figure 3.3: Measured pT versus predicted pT from Eq. 3.6; the area of circles corresponding to CS value; the larger the CS, the greater the area of the circle; the solid line indicating a slope of unity and an intercept of zero

The QSAR model Eq. 3.6 was assessed with 10-fold cross-validation. The summary of the statistics for Eq. 3.6 is presented in Table 3.4. The $RMSE_{CS}$ is lower than the RMSE.

Table 3.4: Statistical summary of 10-fold cross-validation based on Eq. 3.6 (aquatic toxicity)

Training			Test	
Intercept	Log P	R_{adj}^2	RMSE	$RMSE_{CS}$
-1.67 ± 0.14	0.92 ± 0.04	0.68 ± 0.03	0.99 ± 0.12	0.87 ± 0.13

3.4. Discussion

There are many future challenges in human and environmental health sciences which require the use of adequate and reliable data, these include toxicological risk assessment for occupational health and consumer goods. As the quality of toxicological data is variable and often not stated, practical and feasible methods to overcome this issue are crucial to many scientific and regulatory fields. Beside approaches such as Klimisch scoring (Klimisch *et al.*, 1997), a purely statistics-based method to support modelling approaches was proposed in Chapter 2 and expanded within this Chapter. It is difficult to determine the extent to which such a statistically-driven approach could be used for regulatory purposes, but neglecting the information multiple data hold for the same substance is not recommended if such data are available.

The aim of this work was not to build new QSAR/QSPR models, but to make two existing models more robust using independent, heterogeneous datasets. The two QSARs and associated datasets chosen are well established. In this study the datasets have been extended by further data harvesting and collection. As part of the data collection activity, multiple data were compiled for the same chemical, thus allowing for the application of the CS approach to determine the reliability of the data. This approach has not been applied formally in the development of QSARs and there are no clear guidelines on how to develop QSARs when multiple data are available for the same chemicals (*i.e.* use of the mean, most conservative value *etc*.). In addition, there appear to be few, if any, attempts to include information such as data quality as a metric or criterion for QSAR development, this being despite it being logical and acknowledged that data quality will affect the robustness of a QSAR (Wenlock and Carlsson, 2015). It should also be noted that current means of documenting QSARs provide little opportunity for assessing the quality of data. Therefore approaches that allow us to identify data quality quantitatively and without subjective bias are of value in the development of *in silico* models.

Skin permeability is often assessed by *in vitro* experimentation, but also some *in vivo* work is undertaken. *In silico* models are increasingly desirable in areas such as risk assessment where there is dermal exposure (*e.g.* for cosmetics) and for assessing adverse effects to the skin, *e.g.* skin sensitisation. Since the publication of the Flynn dataset (1990), there have been a number of QSPR analyses of skin permeability coefficients including refinements and extensions to the database (Mitragotri *et al.*, 2011). The Potts and Guy (1992) approach, based on fundamental and mechanistically comprehensible descriptors, is one of the more commonly utilised QSPR modelling methodologies. This study has derived a Potts and Guy equation for a larger dataset not only increasing the coverage of the model (*i.e.* greater chemical space) but also incorporating multiple data points for the same chemical and allowing for an assessment of quality through CS. It is noted that published skin permeability coefficients are highly variable, due in no small part to high experimental error arising from the variable nature of the (human) skin utilised and test protocols, *e.g.* use of solvents, enhancers, finite doses, vehicles, solvents *etc.* (Moss and Cronin, 2002; Johnson *et al.*, 1995). As such, it is to be expected that models will not have a very significant statistical fit (*i.e.* a high R^2) and this is borne out by many of the published models (Potts and Guy, 1992; Moss and Cronin, 2002), indeed models with significant fit should be treated with some caution as they may be overfitted.

Whilst high statistical fit was not achieved for the skin permeability QSPRs, the results show a significant relationship between log k_p and log P and MW with both variables having high t-values. The new QSPR has moderately improved statistical fit as compared to that of Potts and Guy (1992). It should be noted that some values within the Flynn dataset were subsequently shown to be incorrect and would have increased the error in the Potts and Guy QSPR (Johnson *et al.*, 1995). The novel QSPR model (refer to Eq. 3.4 and Fig. 3.2) derived from the skin permeability data has some advantages over the original Potts and Guy model. First of all increased robustness, due to model development incorporating statistical data quality (refer to Table 3.3); secondly a greater applicability domain due to implementing a dataset with greater chemical diversity (in terms of properties and structure) than Flynn (1990); and thirdly due to the usage of calculated log P (whereas the original model used measured values which are more difficult to obtain consistently). Nevertheless the differences between Potts and Guy's Eq. 3.1 and Eq. 3.4 are only marginal. It is recognised that there are many limitations to this use of this model. For example it does not predict the effects of mixtures and formulations on the penetration of single compounds, which could be of great importance for risk assessment of products and dermal drug delivery (Samaras *et al.*, 2012). However, the QSPR approach allows for a "relative" estimation of skin permeability which may be useful to rank compounds, or identify compounds with a high probability of dermal absorption and hence prioritise such compounds in the risk assessment process (*e.g.* for skin sensitisation).

Non-polar narcosis in the context of the *A. fischeri* assay was discussed in Chapter 2 (Cronin *et al.*, 1991; Steinmetz *et al.*, 2014). Even if the QSAR models of Chapter 2 and 3 are slightly different, they show the same strong relationship between hydrophobicity (log P) and toxicity as described for many species (Könemann, 1981; Verhaar *et al.*, 1991). In both cases CS, used as a threshold (Chapter 2) and as used here (weighted regression), improved the aquatic toxicology QSAR.

Consideration of the QSAR/QSPR models developed in this study shows an improvement in the models when CS-weighted regression was utilised. The improvement is in both the statistical fit as well as the slope for log P which approaches one when employing CS-weighting, *i.e.* from 0.68 to 0.90 (refer to Eq. 3.5 to 3.6). A slope of one is the theoretical optimum, which is commonly associated with models for simple unicellular organisms, *i.e.* the absorption of the compound alone directly into the cellular membrane is responsible for narcosis, whereas in higher organisms other factors such as distribution and clearance become important. The improvements following the application of CS are consistent with the notion that some historical data are of poor quality (Cronin and Schultz, 1996) and demonstrate the utility of an approach such as this, where generalistic QSARs are being developed for datasets from various sources and of unknown quality. The importance of the compounds with high CS values can be seen in Fig. 3.3, when considering that all large CS-circles (representing compounds with higher CS) are close to the line of best prediction. The quantity of data and the incorporation of statistical data quality make a robust equation with an extensive applicability domain – for non-polar narcotics. Clearly this approach could be extended to other data compilations for aquatic acute toxicity (Martin *et al.*, 2015).

The identification of compounds acting by the non-polar narcotic mechanism of action is essential to the development of models. Various approaches have been applied to identify mechanisms of action including analysis of molecular descriptor space (Schultz *et al.*, 1997), multivariate analysis of mode and mechanism of action space (Aptula *et al.*, 2002), definition of molecular fragments (Ellison *et al.*, 2008) as well as the Verhaar classification scheme that was applied in this study due to its ease of use following coding in the Toxtree software. However, there appear to be a number of anomalies in the definition of the non-polar narcosis domain in the Toxtree software. For example, aflatoxins (cf. Chapter 2; Fig. 2.5 and 2.6) are identified by the Toxtree software as being Verhaar Class 1 compounds (non-polar narcotic) but, in reality, they are potent, specifically acting, toxins and therefore do not act as non-polar narcotics, *e.g.* aflatoxin B2 has $pT_{experimental} = 1.17$ (CS = 15.4) whereas Equation 3.7 calculates $pT_{predicted} = 0.54$ (Steinmetz *et al.*, 2014). This emphasises that continual development is required of decision criteria presented in approaches such as the Verhaar scheme as new knowledge and understanding becomes apparent.

Overall for both datasets, applying CS as a weighting tool improves the training and validation of the QSAR/QSPR models. The improvements are demonstrated as increases in R^2 (Eq. 3.3 to 3.4 and Eq. 3.5 to 3.6) as a direct result of CS-weighting. Whereas increasing t and F values show improvements in the models as a result of weighting by CS, the S value does not incorporate weights and so only indicates absolute, unweighted error thus it actually increases when the non-weighted regression is compared to the weighted regression. Generally the higher the CS for the data associated with a compound, the greater the evidence is, in terms of similar results for that compound (refer to Fig. 3.2 and 3.3). In the validation process, the $RMSE_{CS}$, which incorporates CS-weighting, is lower than the standard RMSE. As residuals (f_i - y*i*) of low CS compounds are weakened and residuals of high CS compounds are strengthened, the sum of (squared) errors of the $RMSE_{CS}$ becomes lower than in the conventional RMSE. Therefore this approach could be used even for the validation of models where any metric could be applied to imply confidence, *i.e.* without calculating CS. For example a reversed Klimisch score (4 as the most reliable; 1 the least) could be used as a weight similar to the fuzzy logic approach of Yang *et al.* (2013). In the context of validation these weights then determine to what extent residuals should have impact on the RMSE.

The CS-weighting approach, whether in model development or validation, is limited by the presence of multiple entries for one compound. Thus, if multiple values are available for the dataset, more robust models may potentially be built (Steinmetz *et al.*, 2014). This robustness and the associated confidence are helpful in reducing uncertainty and hence increasing acceptance for regulatory decisions. For example in the context of REACH, there is a demand for robust QSAR models to support the toxicological assessment of chemicals. The approach described herein could thus be used to support read-across- and QSAR-based predictions (Cronin, 2013; Patlewicz *et al.*, 2014).

3.5. Conclusions

The assessment of data quality is not trivial. This study has shown that CS provides a means of assessing confidence in data when there are more than a single datum point. The CS scores can be applied to develop QSAR models through the use of weighted regression, as demonstrated in this study for historical data compilations with known variability in the quality of the data. Additionally cross-validation with RMSE_{CS} provides a measure of the robustness of an equation utilising metrics (here CS) for weighting.

The results from Chapter 3 show that a novel method, which is applying statistical data quality within modelling, leads to robust QSAR/QSPR models (as published in Steinmetz *et al.* 2015b). Beside the methodological value, particularly the QSPR model is very useful in the context of risk assessment of cosmetic ingredient – hence the relevance for the COSMOS project. Chapter 4, which deals with dermal absorption, applies similar principles of physico-chemical properties. The two main differences are Chapter 4 deals with a set of rules (similar to an expert systems) and only specific applicability domain of substances (*i.e.* hair dyes and associated compounds). Even if the data and the applied methods are different, both chapters share the same biochemical principles of skin permeation.

4. Classifying dermal absorption of cosmetic ingredients based on physico-chemical properties to facilitate safety assessment*

4.1. Introduction

As described in Chapter 1, the European Cosmetic Regulation (EC 1223/2009) requires that the ingredients in cosmetic products, as well as the formulation itself, need to be safe for human usage and it is the responsibility of the manufacturer to ensure this. The safety assessment of a product is generally based on individual safety assessments of the product's ingredients. This requires knowledge of individual ingredients (particularly those in significant concentration) in a product as well as knowledge about toxicological profiling. Furthermore use scenarios and hence exposure patterns of the product are required to allow safety evaluation / risk assessment. This information can subsequently be used for the calculation of the margin of safety (MoS). The MoS is the ratio of the no-observed-adverse-effect level (NOAEL) and the systemic exposure dosage (SED), which can, for example, be dermal absorption per skin surface and time according to use patterns (refer to Equation 4.1). Whereas the NOAEL is typically obtained from repeated dose / reproductive toxicity animal trials, the SED can be obtained from *in vivo* or *in vitro* tests (SCCS/1501/12).

$$MoS = \frac{NOAEL}{SED} \qquad \text{(Eq. 4.1)}$$

Due to the ban on animal testing in the European cosmetic legislation (refer to Chapter 1) and the absence of validated *in vitro* alternatives, it is no longer possible to obtain NOAEL values from *in vivo* experimentation to calculate the MoS for newly developed cosmetic ingredients. However, considerations of exposure may be relevant for instance

*This chapter is based on my contribution to Ates *et al.* (2015)

cosmetic ingredients with negligible dermal absorption may not require systemic toxicological assessment. Hence there is great interest in identifying compounds with low dermal absorption. If this is the case, then systemic toxicological assessment can effectively be waived and safety assessment may be based on local toxicity, *e.g.* skin irritation, corrosion and sensitisation as well as mutagenicity/genotoxicity.

Dermal absorption means the uptake of chemical substances via the skin, sometimes also referred to as percutaneous absorption. This includes skin permeation as described in Chapter 3. However, the main difference in the context of this thesis is that dermal absorption data encompasses data on the absorbed quantities of dermally applied substances, whereas skin permeability (refer to Chapter 3) exclusively describes the permeation through the *stratum corneum* (Rang *et al.*, 2007a).

The dermal absorption dataset investigated in this study is based on information harvested from the expert opinions of the European Commission's Scientific Committee on Consumer Safety (SCCS). It has a clear focus on hair dyes and associated compounds due to their potential toxicity, *e.g.* adverse effects on mitochondria (Nelms *et al.*, 2015). SCCS opinions are publically available and contain summaries of studies on different toxicological endpoints, as well as information on dermal absorption and physico-chemical properties. The information is intended to support product development, including internal safety assessment, and regulatory decision-making in the field of personal care products.

The aim of this study was to classify the dermal absorption of cosmetic ingredients. Hence rule sets are proposed, which have the potential to support regulatory safety/risk assessment. Therefore physico-chemical properties which may affect dermal absorption, such as the logarithm of the octanol-water partition coefficient (log P), molecular weight

(MW), topological polar surface area (TPSA) and the melting point (MP) were investigated. Through many studies in the field of skin permeability, *e.g.* Potts and Guy (1992), Magnusson *et al.* (2004), and oral absorption/bioavailability, *e.g.* Lipinski *et al.* (2001), Newby *et al.* (2015), the relationship between physico-chemical properties and permeation through relevant biological barriers has been investigated and discussed thoroughly. There is a strong consensus that large, hydrophilic and ionic molecules permeate membranes to a lesser extent than small, (moderately) lipophilic and uncharged molecules – similar to the skin permeability QSPR in Chapter 2. Regarding the additional descriptors; TPSA expresses the polar surface of a molecule, *i.e.* it correlates with hydrogen bonding ability and water solubility, and furthermore MP holds additional information on thermodynamical stability (solid-liquid phase change) of a substance. This information, which is easily measured (MP) and calculated (TPSA) respectively, might support the prediction of *in vivo* permeation and absorption (Pugh *et al.*, 2000; Magnusson *et al.*, 2004).

4.2. Methods

The dermal absorption data from the SCCS opinions, based on reports from 2000 to 2014, were provided by the *In Vitro* Toxicology and Dermato-Cosmetology research group of the Vrije Universiteit Brussel (Ates *et al.*, 2015) as part of a co-operative study in the SEURAT-1 Cluster. Two datasets were constructed as follows: dataset A summarises all the data without any information on MP and dataset B summarises the data, which include measured MP values within the reports. Regarding the classification of dermal absorption, *i.e.* defining the absorption threshold for potential adverse effect, the empirically derived values of 1.3% and 2% respectively were chosen (private communication with Prof. Rogiers from the Vrije Universiteit Brussel). The dataset is attached in Appendix A.3.

4.2.1. Data treatment and descriptor calculation/retrieval

The SCCS dermal absorption data are derived from various methods, *e.g.* different species, varying exposure scenarios *etc*. Measurements using rat skin were discarded from the dataset because of the relatively high uptake when compared to human or porcine skin (Ravenzwaay and Leibold, 2004). For compounds with more than one measurement per compound arithmetic means were calculated. Descriptors were calculated for the parent form of the compound, therefore SMILES strings were first neutralised and desalted within MOE (MOE, 2013). Subsequently TPSA and MW were calculated using CDK's molecular properties node within KNIME (KNIME, 2015) and log P was calculated using KOWWIN v1.68 within EPI Suite (EPA, 2013). MP was extracted from SCCS opinions if available.

4.2.2. Decision trees, clustering and modifying rules

For each dataset a set of rules, similar to Lipinski's rule of five, has been created in order to classify compounds as being associated with a toxicologically significant level dermal absorption, *i.e.* above or lower than the thresholds of 1.3% and 2%. Beside empirically derived approximations based on the literature (Potts and Guy, 1992; Magnusson *et al.*, 2004; Lipinski *et al.*, 2001; Newby *et al.*, 2015; Pugh *et al.*, 2000), the KNIME's decision tree learner (KNIME, 2015) employing log P, MW, TPSA and MP (only in dataset B) was used to determine relevant combinations of descriptor cut-offs. The decision tree learner splits classes in a binary manner by minimising differences towards split points. However, to avoid overfitting, final rule sets were defined manually by adjusting rules iteratively. Granularity and statistical performance were analysed according to Cooper *et al.* (1979), *i.e.* comparing sensitivity and specificity.

4.3. Results and discussion

The dermal absorption dataset has been enriched by physico-chemical descriptors. Furthermore the data have been split in dataset A (encompassing 116 compounds without MP) and dataset B (encompassing 70 compounds including MP).

4.3.1. Results dataset A

The following physico-chemical cut-offs (based on the KNIME decision tree learner and empirical refinement) were defined and applied. These cut-offs represent thresholds for increased permeability, hence they are referred to as 'alerts' within the context of the following rule-based models.

- MW < 180 Da
- log P ≥ 0.3

This implies that compounds with MW < 180 Da and/or log P ≥ 0.3 are more likely to be dermally absorbed in a greater magnitude. The results are illustrated in Figure 4.1, which shows that as the number of alerts increases the dermal absorption increases.

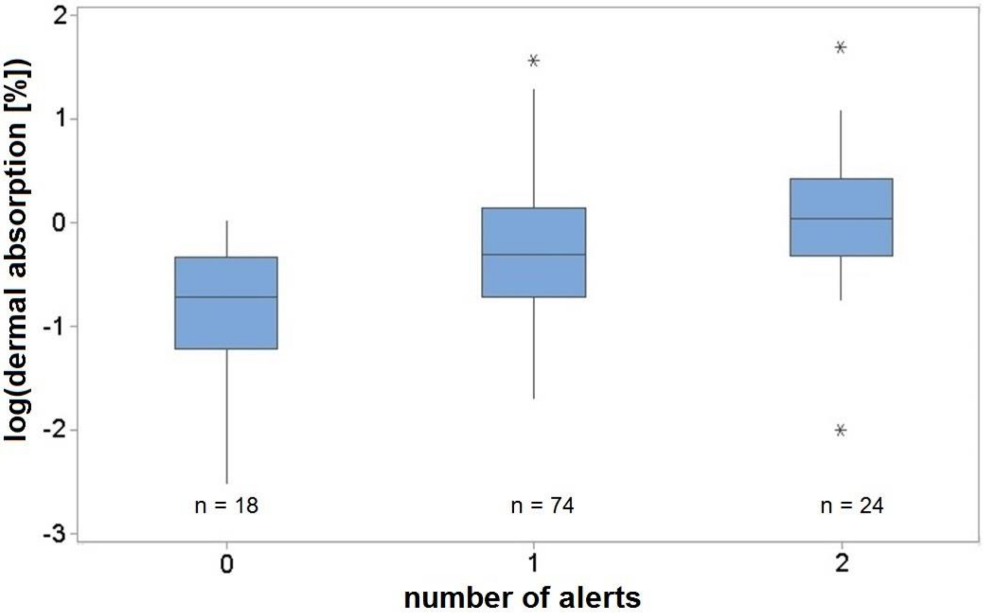


Figure 4.1: Boxplot of log (%) dermal absorption versus number of physico-chemical alerts for dataset A (n = 116); the *-symbol describing outliers

If any violation of the rules, *i.e.* MW < 180 Da or log P is ≥ 0.3, is found, the compound will be predicted as potentially highly absorbed (≥1.3%). Table 4.1 shows the results of applying the rules to dataset A. The calculation of Spearman's rank correlation led to ρ = 0.38 (S = 162320, p < 0.001), which indicates a weak (but statistically significant) correlation between number of alerts and logarithm of dermal absorption.

Table 4.1: Performance of the rules set for dataset A

Dataset A	Predicted highly absorption	Predicted low absorption	Total
High absorption (≥1.3%)	30 (25.9%)	0 (0%)	30 (25.9%)
Low absorption (<1.3%)	68 (58.6%)	18 (15.5%)	86 (74.1%)
Total	98 (84.5%)	18 (15.5%)	116 (100%)

Sensitivity = 100% **Specificity = 20.9%**

On the one hand, the rule set shows high sensitivity, *i.e.* all 30 compounds with true high absorption have been identified, on the other hand specificity is poor; 68 compounds with low absorption are predicted incorrectly to have high absorption. However, the design of the rule set is beneficial for regulatory purposes due to its cautious/restrictive design.

4.3.2. Results dataset B

The following physicochemical alerts were defined and applied:

- MW < 180 Da
- log P ≥ 0.3
- MP < 100°C
- TPSA < 40 Å^2

This implies that compounds with MW < 180 Da and/or log P ≥ 0.3 and/or MP < 100°C and/or TPSA < 40 Å^2 are more likely to be dermally absorbed. The results are illustrated in Figure 4.2, which shows that as the number of alerts increases the dermal absorption increases.

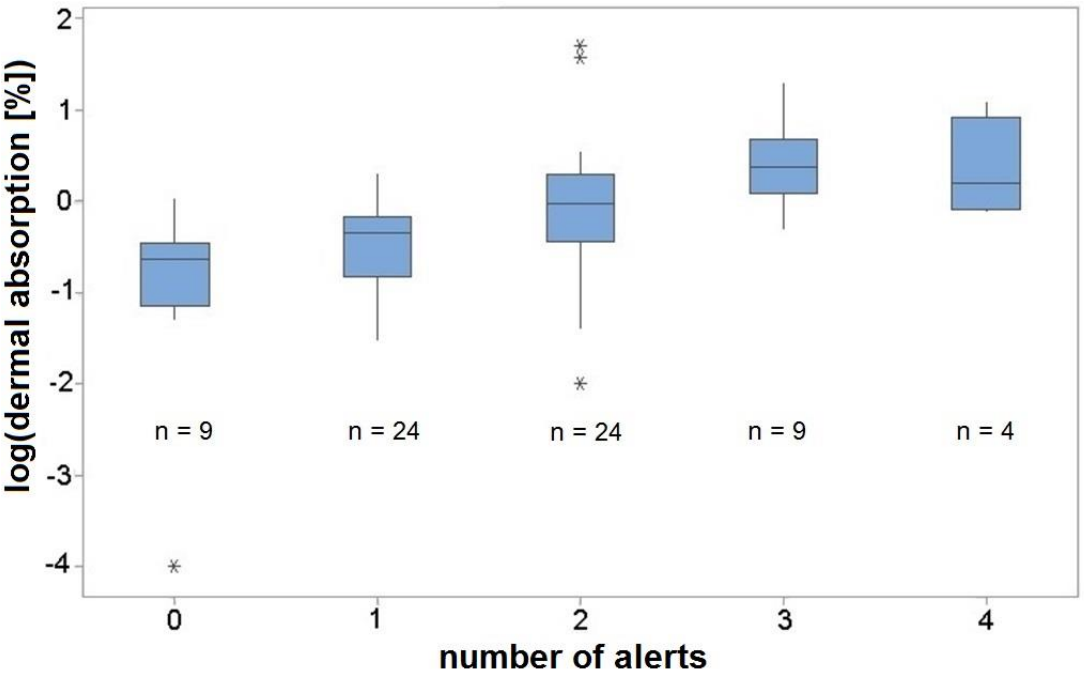


Figure 4.2: Boxplot of log (%) dermal absorption versus number of physico-chemical alerts for dataset B (n = 70); the *-symbol describing outliers

In contrast to the rule set of dataset A, there are two different ways to use the rule set for dataset B. In the first scenario, the conservative approach, if any violation of the rules, *i.e.* MW < 180 Da, log P ≥ 0.3, MP < 100°C or TPSA < 40 Å^2, is identified, the compound will be predicted as having potentially high absorption (≥1.3%). Table 4.2 shows the results of applying the rules to dataset B. The calculation of Spearman's rank correlation led to ρ = 0.60 (S = 23111, p < 0.001), which indicates a moderate, statistically significant correlation between number of alerts and logarithm of dermal absorption.

Table 4.2: Performance of the rules set for dataset B (scenario 1; "conservative")

Dataset B	Predicted high absorption	Predicted low absorption	Total
High absorption (≥1.3%)	23 (32.9%)	0 (0%)	23 (32.9%)
Low absorption (<1.3%)	38 (54.3%)	9 (12.9%)	47 (67.1%)
Total	61 (87.1%)	9 (12.9%)	70 (100%)

Sensitivity = 100% **Specificity = 19.1%**

In the second scenario, the realistic approach, violation of none or only one rule is allowed, meaning that more than one violation of the rule leads to the prediction of high absorption for a compound. The statistical performance of scenario two is expressed in Table 4.3.

Table 4.3: Performance of the rules set for dataset B (scenario 2; "realistic")

Dataset B	Predicted high absorption	Predicted low absorption	Total
High absorption (≥1.3%)	19 (27.1%)	4 (5.7%)	23 (32.9%)
Low absorption (<1.3%)	18 (25.7%)	29 (41.4%)	47 (67.1%)
Total	37 (52.9%)	33 (47.1%)	70 (100%)

Sensitivity = 82.6% **Specificity = 61.7%**

Scenario 1 and 2 can be directly compared as they use the same dermal absorption cut-off. On the one hand the model based on scenario 2 is a better overall prediction with moderate sensitivity (82.6%) and specificity (61.7%) respectively (refer to Table 4.3), on the other hand the model based on scenario 1 is very conservative with a maximum sensitivity (100%) but poor specificity (19.1%) (refer to Table 4.2). Therefore the "conservative" model might be more favourable for regulatory decision-making, due to high certainty of practically no dermal absorption when "low absorption" is predicted.

However, in the third scenario, the flexible approach, a different threshold for dermal absorption was taken (2%). The 2% threshold is empirically more favourable than 1.3% threshold to classify the dataset B. The violation of one or more rules leads to the prediction of a compound having high absorption. All other compounds are classified as having low absorption. The performance of the rule set in this scenario is shown in Table 4.4.

Table 4.4: Performance of the rules set for dataset B (scenario 3; "flexible")

Dataset B	Predicted high absorption	Predicted low absorption	Total
High absorption (≥2%)	13 (18.6%)	0 (0%)	13 (18.6%)
Low absorption (<2%)	24 (34.3%)	33 (47.1%)	57 (81.4%)
Total	37 (52.9%)	33 (47.1%)	70 (100%)

Sensitivity = 100% **Specificity = 57.9%**

When shifting the dermal absorption cut-off to 2%, it is possible to achieve maximum sensitivity (100%) while still having moderate specificity (57.9%) (refer to Table 4.4). While the assignment of the new dermal absorption cut-off seems to be favourable for the model, it is questionable if 2% *in vivo* dermal absorption is within an acceptable margin regarding regulatory assessment (*nota bene:* the initially suggested cut-off, before modelling, was 1.3% based on private communication with Prof. Rogiers).

Effects of physico-chemical properties on dermal absorption have been confirmed with similar concepts as the literature proposes, *i.e.* that small, uncharged and (moderately) lipophilic compounds pass easier through the skin (Potts and Guy, 1992; Magnusson *et al.*, 2004; Lipinski *et al.*, 2001; Newby *et al.*, 2015; Pugh *et al.*, 2000). Furthermore similar physico-chemical relationships on passing through the *stratum corneum* were confirmed by the QSPR models presented in Chapter 3. The official opinions of the SCCS do offer only limited descriptions of the testing protocols, *i.e.* it would be nearly impossible to differentiate high and low data quality based on testing protocols. However, differences in the dermal absorption testing methodology are likely to have an impact on the potentially poor data quality of some data. Therefore the focus of this study lies, as often in applied sciences, on the overall picture by accepting the potential low quality associated with the data (refer to Chapter 2 and 3).

Most cosmetic products are applied topically, which makes dermal absorption the main route of exposure. Of course, dermal absorption is only one factor within safety and risk assessment, however it is relevant for the calculation of the MoS (as described in Eq. 4.1). For compounds with marginally low dermal absorption values (*e.g.* <0.01%), SED values are very low, what may increase the MoS quite dramatically. When additionally considering the usage of the uncertainty factor, for example for the animal to human extrapolation, an experimental NOAEL value might not contribute as much as expected to the final safety evaluation. Furthermore it must be noted that rodent data, from which NOAEL values are based on, are not always of that precise. When Gottmann and colleagues investigated two datasets with experimental rodent carcinogenicity data, they only found a concordance of 57% between duplicates from different resources (Gottmann *et al.*, 2001). Keeping this mind, read-across and local QSARs (*e.g.* within one functional/chemical class of cosmetic ingredients) may be excellent tools to allow for the assessment of NOAELs – particularly as experimental testing to establish a NOAEL is no longer feasible for cosmetic ingredients.

4.4. Conclusions and perspectives

In this study an *in silico* approach to predict (or to better classify) dermal absorption of chemicals was developed. Several models were developed with differing sensitivity and specificity depending on the dermal absorption thresholds defined for classification and the availability of melting point data. It must be pointed out that, as shown by the performance of the models (Tables 4.2 and 4.3), high sensitivity usually compromises specificity and *vice versa*. It is common practice in risk assessment and regulatory affairs to consider, or plan for, worst-case scenarios by assuming “conservative” numbers in cases of doubt, for example when few or no adequate data are available. However, a conservative approach at multiple levels can cumulatively add up to an overly cautious number, *e.g.* a very low MoS value. Generally, from a scientific point of view, the most realistic equation, model or even “educated guess” should be used at every step (exposure, absorption, MoA *etc.*) within any risk assessment approach. Rounding up/off to a conservative, regulatory acceptable value should be done exclusively at the end of the approach. It is more transparent to increase the uncertainty factor at the end of the mathematical part of the risk assessment than using skewed equations and models, for example with 100% sensitivity and poor specificity (refer to Tables 4.1 and 4.2).

Nevertheless, all tools presented in this study have the potential to support risk assessment, at least on the SED side of the equation. According to the challenge, *i.e.* dealing with hair dyes *etc.* and splitting the data at 1.3% dermal absorption, models are presented with attributes such as “conservative”, “realistic” and “flexible” (threshold at 2%).

Beyond the concrete dermal absorption classification models of hair dyes *etc.*, this study serves as well as a demonstration of how to create simple classification models for dermal absorption to support non-testing approaches in the consumer and personal care industry. Both of these interpretations of this study, as a model and as a blueprint for other classification models, are relevant for the assessment of cosmetic ingredients, and hence relevant for the COSMOS project. However as dermal absorption is only one pillar of the assessment of cosmetic ingredients, toxicity-driving mechanisms need to be investigated as well. Therefore Chapter 5, 6 and 7 deal with mechanistically-based modelling with a specific focus on liver toxicity (as a relevant example for cosmetic ingredients).

5. Screening chemicals for receptor-mediated toxicological and pharmacological endpoints: Using public data to build screening tools within a KNIME workflow*

5.1. Introduction

The assessment of potential toxicants is a multidisciplinary task. Whilst the previous chapters dealt with issues, such as data quality (Chapters 2 and 3) and kinetics (Chapters 3 and 4), the following Chapters (5 to 7) deal with mechanistically based modelling and the identification of the molecular initiating events of untested compounds. Hence the following chapters provide innovative tools and methodologies for hazard identification.

Generally speaking, predicting and understanding the properties of new chemical entities is not trivial, whether in the development of novel pharmaceuticals or in assessing potential toxicity. However, *in silico*, QSAR and read-across approaches provide a means of rapidly obtaining information (Blackburn and Stuard, 2014; Cronin *et al.*, 2013; Patlewicz *et al.*, 2013). Such models can be supported by, or developed from, mechanistic understanding (Zhu *et al.*, 2014). Additionally the concept of the AOP, *i.e.* describing a sequence of causally linked events at different biological levels, is increasingly being applied to investigate adverse effects (Vinken *et al.*, 2013). As described more in detail in Chapter 1 (Section 1.4.2), models may be developed from knowledge of the first key event of an AOP, the molecular initiating event (MIE). In AOP terminology the MIE is followed by cellular and organ responses, which may ultimately result in an adverse effect to an organ, organism or population (Ankley *et al.*, 2010). The MIE represents the initial interaction between a molecule and a target. Examples of MIEs include covalent binding to DNA and, of relevance for this study, receptor binding

* This chapter is based on a published article (Steinmetz *et al.*, 2015a)

(Gutsell and Russell, 2013; Allen *et al.*, 2014). In pharmacology the mode of action, similar to an AOP, incorporates a MIE which describes how a compound interacts with specific proteins, *e.g.* receptors, carriers and enzymes. However, rather than providing the framework for describing the processes behind an adverse effect, the aim in pharmacology is to achieve a beneficial effect, such as the prevention or treatment of a disease (OECD, 2012; FDA, 2013).

Analogous to pharmacology, toxicity may also be brought about by interactions with specific proteins, such as receptors. Endocrine disruptors, for example, are a class of toxicants known to cause their effects by receptor-mediated mechanisms. As such, models for endocrine disruption are usually built around knowledge of receptor interactions, *e.g.* binding to the oestrogen receptor. For instance, one approach to modelling these effects has been proposed recently by Kolšek and colleagues (2014) who developed a tool to identify nuclear receptor ligands based on AutoDock Vina; a freeware to investigate ligand-protein-interactions (Molecular Graphics Laboratory, 2014). Limitations of this type of approach are associated with several of the typical issues of docking. First, nuclear receptors, particularly the non-steroid receptors, are considered to be flexible (Nettles *et al.*, 2007). An inflexible docking model, such as AutoDock Vina, is unlikely to cope with the diversity of ligands including, for instance, full and partial binding modes as well as inverse agonists and antagonists. The second limitation, when docking is applied on its own, is that kinetics (on a cellular level) are systemically ignored, which might be vital for *in vivo* biological activity. The physico-chemical properties of the ligand play an important role, particularly for absorption and distribution at a histological and cellular level, which may all eventually contribute to, or define, target-organ-toxicity (Campbell, 1983; Davis and Riley, 2004).

The current study focuses on the retinoic acid receptor (RAR), a target relevant for pharmacology and toxicology in equal measure. The RAR is a nuclear receptor which can be divided into three subtypes, RAR-α, RAR-β and RAR-γ. Bound together with the retinoid X receptor (RXR) as a heterodimer, RAR regulates genetic expression. All three subtypes of the RAR are activated by *all-trans* retinoic acid and 9-*cis* retinoic acid, which are derivatives of vitamin A (Liu et al., 2014). Ligands are used in the treatment of dermal diseases, *e.g. Acne vulgaris*, *Psoriasis vulgaris*, *Keratosis pilaris* and specific types of cancer, such as acute promyelocytic leukaemia (Alizadeh *et al.*, 2014; Allen and Bloxham, 1989; Dicken, 1984; Leyden *et al.*, 2005).

The toxicological effects of RAR agonists include changes in lipid metabolism, which may cause hepatosteatosis leading to liver inflammation, fibrosis and eventually liver failure. Teratogenic effects and neural disorders, such as nausea and headache, have been also reported from retinoids (Adams, 1993; Biesalski, 1989; Moya *et al.*, 2010; Shalita, 1988). There is, therefore, a great need to develop tools to identify compounds which show these effects.

There are many open source software applications and open access databases supporting modern life sciences and informatics. A number of these open access/source technologies can be utilised to develop tools and approaches for predictive and/or computational toxicology. Some technologies relevant to this study are described below.

The KoNstanz Information MinEr (KNIME) technology is a freely available software to analyse and mine data, as well as to build and evaluate predictive models. The software is based on a graphical user interface utilising so called “nodes” as key units to alter and process data in a “workflow”. The basic KNIME workflow technology, as well as many nodes and add-ons for chemo-informatics, is available from www.knime.org. Many types of data can be handled, including chemical formats, such as the Simplified Molecular Input Line Entry System (SMILES) and SMiles ARbitrary Target Specification (SMARTS) (Daylight, 2014). KNIME has a strong community of developers building additional nodes for chemo-informatics applications (amongst others), to edit data, calculate physico-chemical properties, analyse structural features *etc*. It has been shown to be useful in developing workflows for screening tools in the context of predictive toxicology (Saubern *et al.*, 2011; KNIME, 2013). Furthermore, many other programming languages, such as R, Python or Perl, can be used within a KNIME workflow (Berthold *et al.*, 2007; KNIME, 2014; Richarz *et al.*, 2013).

With regard to biological activity, there are an increasing number of resources available to retrieve information. For instance, ChEMBL is a database of bioactive molecules comprising over 1.5 million compounds and over 9,000 biological targets. Activity values are reported for a variety of endpoints including K_i, K_d, AC_{50}, IC_{50}, and EC_{50}. The database is curated manually and maintained by the European Molecular Biology Laboratory (ChEMBL, 2014). A good example of the application of ChEMBL and the utilisation of its resources was published by Czodrowski (2013). In that study, a detailed analysis of ChEMBL hERG assay data was used to build classification models relevant for drug development and demonstrated the applicability of these data for modelling and the value that may result from data mining.

Another valuable resource is the Protein Data Bank (PDB, 2014) which contains over 100,000 crystallographic structures of proteins such as receptors, transporters and enzymes. A quarter of these protein structures are of human origin, the other structures are from other mammals (mainly rodents) and bacteria. For some proteins, such as the RAR, there are data for several subtypes, species and ligands (Berman *et al.*, 1999). Besides the linked publications for every entry, ligand-protein-interactions can be investigated with specific software, for example PyMOL (2014). Visualisation of protein structures of targets, such as receptors, transporters and enzymes, and their corresponding ligands, helps to understand ligand-protein-interactions, *e.g.* hydrogen bonds between the ligand and the ligand-binding-domain of the protein.

Whilst there is a growing number of computational resources, some of which have been developed for computational toxicology, up until now there has been little effort, and few publications, demonstrating the utility of combining these disparate sources of information. The aim of this investigation, therefore, was to present a hands-on approach to develop screening tools applicable for many pharmacological and toxicological challenges. The methods applied are based firstly on gathering publically available data on RAR ligands (from ChEMBL and PDB) and secondly extracting information on physico-chemical space and structural features that are relevant to activity. Thirdly, this information was used to build a rule-based screening tool within KNIME. The purpose of the screening tool in this study was to identify potential RAR ligands. RAR is only one example target, *i.e.* this approach was designed to provide a framework that can, in principle, be used to create screening tools for other receptors should sufficient data be available.

5.2. Methods

The RAR and its ligands were investigated solely using freeware (*nota bene:* PyMOL is free for academic users only) and open access databases.

5.2.1. Analysis of RAR ligands using the PDB

The PDB 3.3 was searched for human RAR structures, *i.e.* RAR-α, RAR-β and RAR-γ (PDB, 2014). The structures obtained were investigated visually with regard to their ligand-protein-interaction within PyMOL 1.3 (PyMOL, 2014). Common structural features of the ligands, particularly when apparently responsible for similar ligand-protein-interactions, were extracted manually. The extracted structural features combined information about molecular distances and molecular electronic forces, which may be responsible for hydrogen bonding or the occupation of lipophilic pockets. Subsequently the structural features were coded manually into SMARTS strings. These SMARTS strings were later used in the rule-based workflow to predict potential RAR ligands.

5.2.2. Extracting data from ChEMBL

The ChEMBL_19 database was searched for the target "RAR" (ChEMBL, 2014). Human data from compounds with K_i (binding affinity), K_d (dissociation constant), AC_{50} (50% activity in molar units) and EC_{50} (50% effect concentration in molar units) values towards RAR-α, RAR-β and RAR-γ were downloaded, combined and sorted by the pChEMBL value. The pChEMBL value is an approach to standardise different types of activity values (Bento *et al.*, 2013). Every compound with a value of five or greater was regarded as being active due to binding towards RAR. This is consistent with the activity interpretations of the ChEMBL database.

5.2.3. Physico-chemical property calculation

The physico-chemical properties of RAR ligands were calculated using the CDK node for molecular properties within KNIME 2.9.4 (including community contributions) (KNIME, 2014). Ranges (*i.e.* minimum and maximum values) for different types of calculated descriptors for the active ligands were studied including: vertex adjacency information magnitude (VAIM) for structural complexity, number of rotational bonds (RB) for flexibility, molecular weight (MW) for molecular size and the logarithm of the octanol-water partition coefficient (XLogP; CDK's version of the log P) for lipophilicity. These four descriptors and their calculated property ranges were utilised to give an insight into the physico-chemical applicability domain (or chemical property space) of active RAR ligands.

5.2.4. Building rules for the screening workflow

The analysis of the PDB has provided structural features coded as SMARTS strings; whilst the analysis of the ChEMBL dataset provided physico-chemical property ranges. Both describe the necessary features for compounds to be active RAR ligands. These features can be interpreted as rules, where compliance and violation will distinguish between RAR ligands and non-ligands respectively. These rules, characterising the physico-chemical space (CDK node for molecular properties) and structural features (Indigo substructure matcher), were written into a KNIME workflow. When executed, this KNIME "screening workflow" can identify potential RAR ligands.

5.2.5. Testing the screening workflow

The RAR ligands, identified from the ChEMBL dataset, were used to test if all active compounds were identified by the "screening workflow". Since no external validation dataset was available, the dataset of hepatotoxicants provided by Fourches and colleagues was screened. The Fourches dataset is a large, chemically diverse dataset (951 compounds), which contains hepatotoxic and non-hepatotoxic drug molecules, including several RAR ligands (Fourches *et al.*, 2010). As the number of RAR ligands is unknown, the performance statistics (sensitivity, specificity *etc.*) of the screening workflow cannot be calculated; thus the predictions for the Fourches dataset are for illustration only. This approach cannot be considered a full validation as the Fourches data could include liver damage by a number of mechanisms not restricted to RAR binding.

5.3. Results

This study utilised a number of data sources, such as the PDB for ligand-protein-interactions and the ChEMBL database for chemical structures of active compounds against RAR.

5.3.1. Ligand-protein-interaction in RAR

20 human RAR protein structures bound to different ligands were retrieved from the PDB. These were 4JYG, 4JYH, 4JYI, 4DQM, 4DM6, 4DM8, 3KMR, 3KMZ, 1XAP, 1FD0, 1FCX, 1FCY, 1FCZ, 1DSZ, 1EXA, 1EXX, 3LBD, 4LBD, 2LBD and 1HRA (PDB, 2014). Independent of receptor subtype and ligand, as proposed by Klaholz and colleagues (2000), the hydrogen bond between an oxygen (most often from a carboxylic group) and the arginine (here: R278) was found to be of great importance for the ligand-protein-interaction. Figure 5.1, for example, indicates the carboxylic acid of retinoic acid binding to amino acid R278.

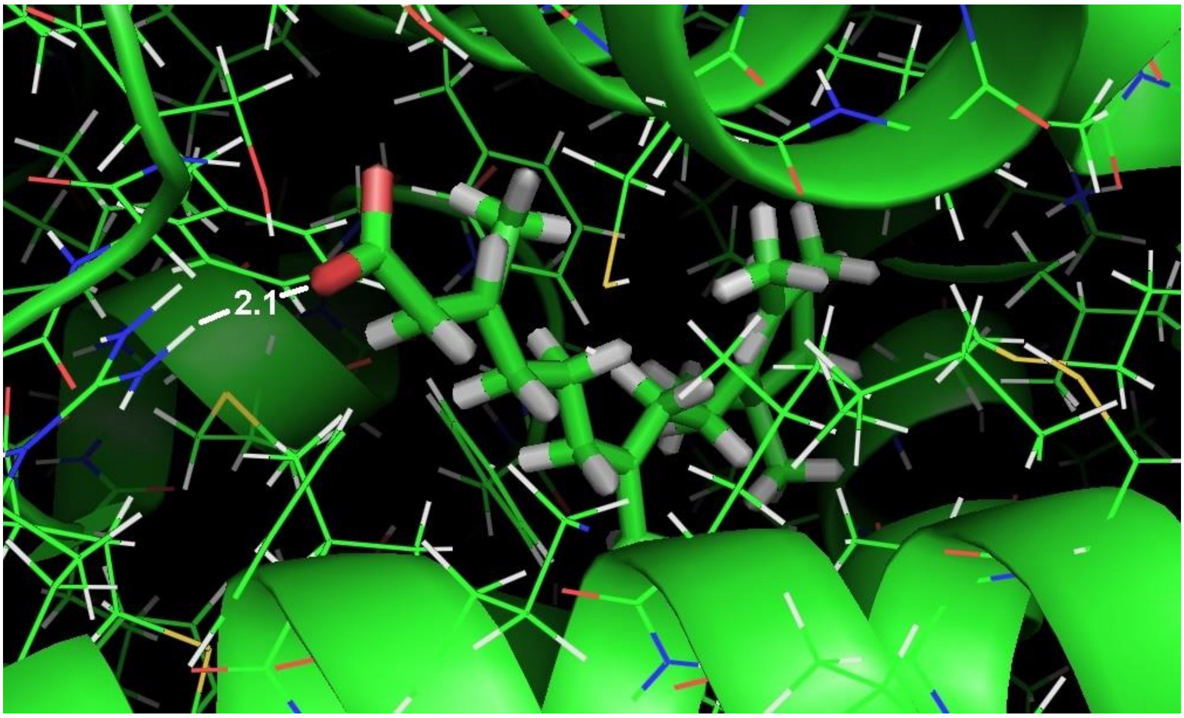
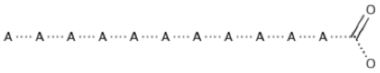
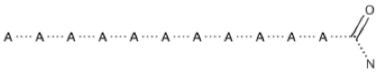
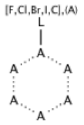


Figure 5.1: Retinoic acid binding to human RAR gamma (3LBD), highlighting the distance of 2.1 Å between R278 and an oxygen of the carboxylic group of retinoic acid (investigated with PyMOL 1.3)

5.3.2. Substructures extracted from the ChEMBL database

251 active RAR ligands (pChEMBL ≥ 5) were identified from the ChEMBL database and these are recorded in Appendix A.4. Common structural features to the ligands, as identified from analysis of the chemical properties and visual appearance, were flexibility, a lipophilic scaffold and a terminal hydrogen acceptor (*e.g.* the carbonyl of a carboxylic group). This information about essential molecular substructures and properties was coded as SMARTS strings, as shown in Table 5.1. The first rule is for a carboxylic group, an amide or a ring structure derived from these structures, *e.g.* 1,2,4-oxadiazol-5-one, which is required to be at the end of a predominantely aliphatic chain. Specific aromatic-containing scaffolds are possible too (refer to Fig. 5.2), and these are also recognised by the substructures identified in Table 5.1. Regarding the second rule, the ring structure, *e.g.* cyclohexene in retinoic acid, can be methylated or halogenated, as the ChEMBL dataset of active RAR ligands revealed.

Table 5.1: Structural features of ligands converted to rules for the KNIME workflow

Rule	SMARTS string	Structural feature
Arginine (R278) Binder	*~*~*~*~*~*~*~*~*~*~*~[#6](=O)~[#8]	
	or	
	~~*~*~*~*~*~*~*~*~*~[#6](=O)~[#7]	
and		
Methylated or halogenated ring-system	*1~*([F,Cl,Br,I,C])~*~*~*~*~1	

“A” or “*” is a wild card, *i.e.* it could represent any heavy atom

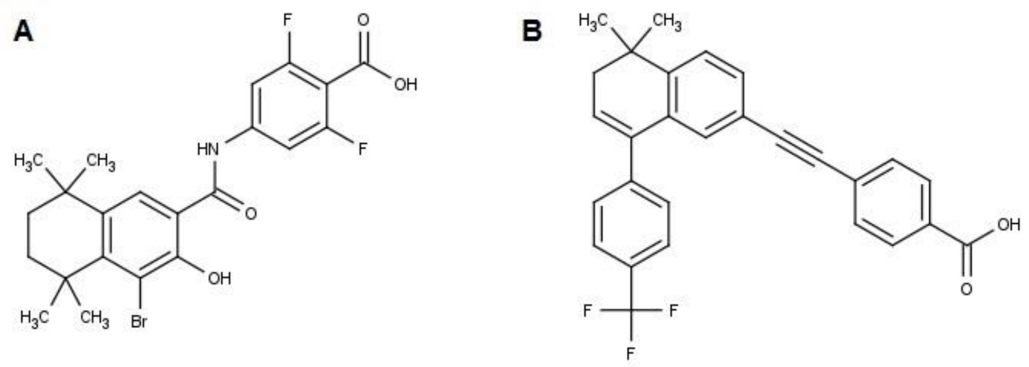


Figure 5.2: Structures of 4-{[(4-bromo-3-hydroxy-5,5,8,8-tetramethyl-5,6,7,8-tetrahydro-2-naphthalenyl)carbonyl]amino}-2,6-difluorobenzoic acid (A) and 4-({5,5-dimethyl-8-[4-(tri-fluoromethyl)phenyl]-5,6-dihydro-2-naphthalenyl}ethynyl)benzoic acid (B) illustrating the flexible nature, lipophilic character and terminal hydrogen bonding group of two chemically diverse potent RAR ligands

5.3.3. Physico-chemical properties

The ranges of the physico-chemical properties calculated for the 251 ChEMBL-derived RAR ligands are shown in Table 5.2. The ranges were converted into rules which can be used as exclusion critera, *i.e.* if a compound has a MW of greater than or equal to 500 Da, then it is, according to the retrieved data, unlikely to be a RAR ligand. The rules have some structural basis, *i.e.* VAIM and MW express the complexity and the size of the molecule respectively, and XLogP describes overall molecular lipophilicity. Besides this basic information, RB indicates the required flexibility of the (lipophilic) chain. Generally speaking, the chemical space covers small, lipophilic molecules with certain degrees of flexibility within the lipophilic scaffold. This is consistent with our understanding of the properties of the ligands and their impact on receptor binding. When dealing with continuous data, margins of error have been applied manually to the rules, *e.g.* a lower limit for XLogP being 2.00 instead of 2.03 (refer to Table 5.2) was chosen. Whilst these are arbitrary, they provide a usable buffer.

Table 5.2: Physico-chemical property ranges of the RAR ligands and derived rules

Descriptor	Min	Max		Rule
RB:	4	23	→	≥ 4
VAIM:	5.46	6.40	→	5 to 6.5
MW:	278.13	488.25	→	< 500
XLogP:	2.03	10.18	→	≥ 2.00

5.3.4. Building the KNIME workflow

A KNIME workflow, which can be downloaded from Appendix C.1, was created combining structural features based on the information from PDB and physico-chemical rules based on the ChEMBL dataset. The workflow is shown diagrammatically in Figure 5.3. The workflow takes the compound of interest through molecular input, implementation of physico-chemical and structural rules in turn, resulting in an output of whether the compound is in or out of "binding space". In more detail, the chemical structure of interest is imported as a SMILES string. Subsequently, physico-chemical properties are calculated and the exclusion criteria (refer to Table 5.2) are applied. Following this, the structural rules from Table 5.1 are applied. In this part of the workflow, the input SMILES strings, which have already passed the physico-chemical rules, are run against a set of SMARTS strings, looking for matches regarding rule 1, the arginine binder, and rule 2, the methylated/halogenated ring-system (refer to Table 5.1). If a compound's calculated physico-chemical properties are within the defined ranges (refer to Table 5.2), *i.e.* it lies within the applicability domain, and contains the relevant structural features (refer to Table 5.1), then the compound is classed as having the possibility of being an active RAR ligand. If a compound is outside the calculated physico-chemical ranges of Table 5.2 or does not contain the structural features (refer to Table 5.1), it is classified as being inactive towards RAR. Finally the workflow, as it is built in Figure 5.3, exports a csv-file gathering the potential RAR ligands.

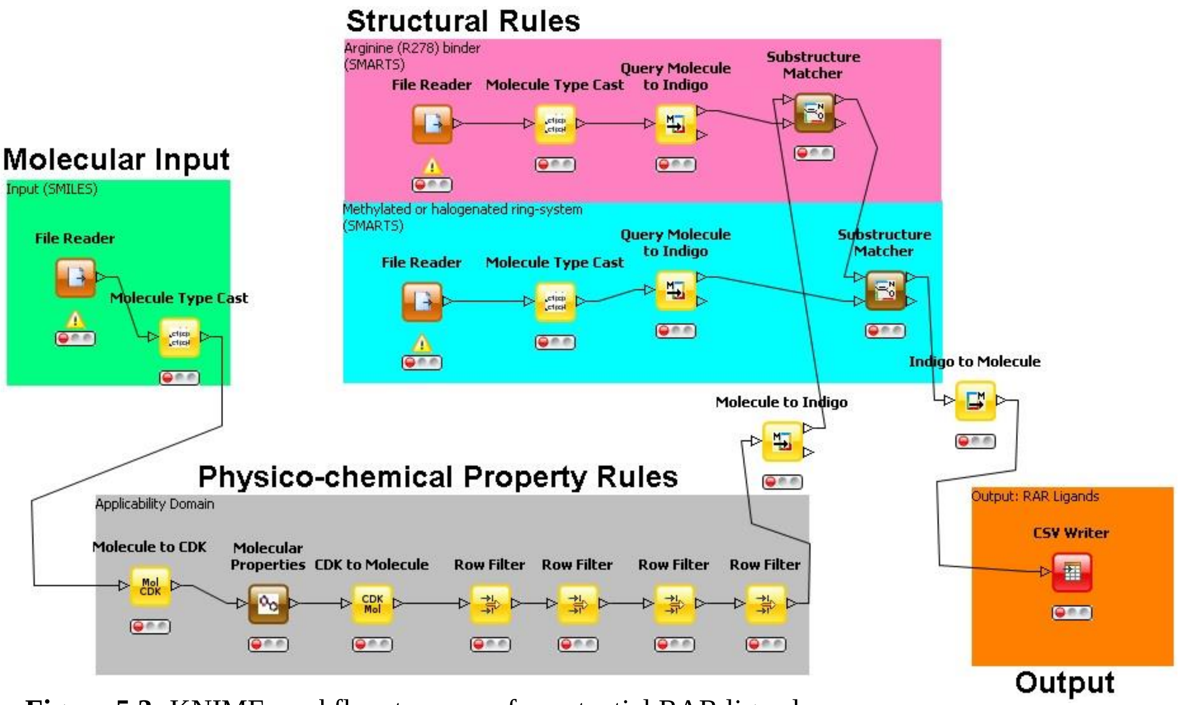


Figure 5.3: KNIME workflow to screen for potential RAR ligands

5.3.5. *Evaluating the workflow: Screening two datasets*

The workflow was used to screen the 251 active compounds identified from the ChEMBL dataset with all compounds being identified as RAR ligands. 109 of 951 compounds in the Fourches dataset (Fourches *et al.*, 2010) were identified as being RAR ligands. Beside retinoids and retinoid-similar structures, some steroids and structurally diverse drugs, such as amineptine (tricyclic antidepressant) and cocaine (tropane alkaloid) were identified as potential RAR binders. The Fourches dataset does not contain information on RAR activity, so performance statistics, such as Cooper statistics (Cooper *et al.*, 1979), *i.e.* false positive ratio, sensitivity *etc.*, are not meaningful in this context.

5.4. Discussion

Extrapolating from chemistry to pharmacology/toxicology is a non-trivial, often even impossible, task. However, it is recognised that assessing chemicals for their pharmacological and toxicological properties is of great importance for industry and regulatory agencies. The AOP framework is increasingly seen as providing usable information for modelling as it describes the linkage between the (bio)chemistry of the MIE and the potential adverse effect on individuals and populations (Gutsell and Russell, 2013). A key challenge remains in the prediction of chronic toxicity, particularly modes of action relating to organ level toxicity. New technologies have the potential to exploit the wealth of data that will be delivered from modern database approaches such as ChEMBL and increasing reporting of information from molecular biology. To exploit these data, tools and strategies, such as data mining, knowledge extraction techniques and (chemo-)informatics tools, are required. Particularly in risk assessment, the identification, characterisation and application of chemistry from the MIE of an AOP is an increasingly commonly used method to "group" or form categories of similar compounds (Vinken *et al.*, 2013; Ankley *et al.*, 2010).

Grouping is a crucial element in the further use of predictive toxicology approaches, such as read-across or QSAR and is best undertaken from a mechanistic standpoint (Blackburn and Stuard, 2014; Patlewicz *et al.*, 2013; Cronin *et al.*, 2013; OECD, 2012). One of the key challenges for grouping compounds is the definition of similarity. The mechanistic framework provided by the AOP paradigm gives a rational basis to developing chemistry based alerts (from the MIE) for grouping and ultimately confirming group membership using data from assays representing key event(s).

This study has applied innovative methods to obtain structural information relating to an important MIE. This has been achieved by investigating protein-ligand binding data and physico-chemical properties. Thus, screening a toxicity dataset with the RAR ligand workflow may help to identify compounds acting by the same mechanism and therefore belonging in the same group. For such a group of compounds there is a greater likelihood of developping mechanistically valid, robust QSARs (OECD, 2014; OECD, 2012; Enoch *et al.*, 2011; Patlewicz *et al.*, 2013). In drug design, there is an interest in identifying potent RAR agonists to address several types of cancer and skin diseases (Alizadeh *et al.*, 2014; Leyden *et al.*, 2005; Allen and Bloxham, 1989; Dicken, 1984). The interest may lie in advances towards the receptor-specificity (Vaz and de Lera, 2012; Schinke *et al.*, 2010), *i.e.* significant activity for certain receptor subtypes, or pharmacokinetics (el Mansouri *et al.*, 1995), *e.g.* targeted drug localisation. Both strategies may lead to RAR agonists with fewer side effects or better risk-benefit ratios.

In this study information from a set of 251 active RAR ligands from ChEMBL and 20 crystal structures of ligand-protein-interactions from the PDB was extracted and investigated to build a screening workflow predicting potential RAR ligands. The set of active RAR ligands is based on K_i, K_d, AC_{50} and EC_{50} values, which means beside agonists, the dataset is also likely to contain antagonists. However, structural and physico-chemical information on antagonists is regarded as beneficial to predict agonists, as both share many chemical features. The disadvantage of this procedure is a higher likelihood of predicting false positives, *i.e.* predicting antagonists as being active. However as a result of the precautionary nature of this approach, potential drug candidates in drug discovery and potential toxicants should be identified by the screening workflow.

As proposed by Klaholz and colleagues (2000) and confirmed by this study, all ligands are small, flexible compounds with lipophilic (mostly aliphatic) scaffolds and a (more or less) terminal polar functional group, for example, an amide or a carboxylic acid, which creates a hydrogen bond with arginine, *e.g.* R278 (PDB, 2014; Klaholz *et al.*, 2000). Potent ligands contain at least one ring structure in the aliphatic scaffold. Furthermore, ring structures may be halogenated, as this does not decrease lipophilicity, such as the compounds illustrated in Figure 5.2, which are highly potent RAR-α binders (Beard *et al.*, 2002; Johnson *et al.*, 1999).

Figure 5.2 also illustrates the lipophilic (mostly aliphatic) scaffold. As long as flexibility and lipophilicity are not greatly impaired, compounds with aromatic rings and amides within their scaffold are potential ligands. This explains the large number of wild cards within the SMARTS strings (refer to Table 5.1). These wild cards, which are expressed with a "*", represent any heavy atom and the wild card bond expressed with a "~" represents any type of bond. On its own the SMARTS strings developed seem not to be very specific, however due to the rule-based combination of SMARTS strings and the applicability domains defined by physico-chemical attributes, the RAR ligands can be identified with a certain degree of specificity. The exact degree of specificity cannot be calculated, but when observing the predictions for the Fourches dataset (Fourches *et al.*, 2010), where 109 potential RAR ligands out of 951 drug-like compounds were predicted, the outcome implies a certain degree of specificity – or better, selectivity. According to the analysis of the Fourches dataset, 85 compounds of the 109 predicted RAR ligands are hepatotoxic. The RAR actives from the ChEMBL dataset were all correctly predicted, which again confirms the indication of high sensitivity.

A screening workflow, as designed as in this study, is assumed to be more sensitive than specific, according to the terminology of Cooper and co-workers (1979). However, as "conservativeness" is relative, it should be pointed out that KNIME allows for the easy adjustment of workflows – without mastering computer language; parameters, thresholds and alerts can be changed intuitively. Furthermore it should be observed that the purpose of these kind of screening tools is not to replace *in vitro* assays or any other *in silico* investigation. The main application lies in tasks, such as prioritisation, or as a valuable part of an elaborated consensus model (refer to an integrated testing strategy) and it can also assist in the rational grouping of compounds assisting in read-across to predict activity and fill data gaps. It is noted that placing this knowledge in the context of the AOP framework allows for grouping and read-across to be supported with evidence from assay for other key events (Tollefsen *et al.*, 2014).

5.5. Conclusions

A novel approach to build screening tools solely with freeware (at least for academia) and open access databases has been described. The flexible design within KNIME allows for adjustment and combination of workflows individually regarding their purpose and their specific endpoints. Furthermore a prediction tool for RAR ligands, as an example for toxicology and pharmacology in equal measure, is presented, which may help to identify potential new drugs and toxicants. This study has also provided new, transparent, knowledge regarding the binding of ligands to RAR which may be useful in a number of contexts.

This study of a novel methodology to identify ligands has been published (Steinmetz *et al.*, 2015a). The Appendix contains the dataset (A.4), the rule set (B.2) and the KNIME workflow (C.1). Of course, in terms of risk assessment and the identification of potential toxicants RAR ligands are only one of a myriad of chemical groups. To include some further relevant groups, an extension of this approach towards further NRs is presented in Chapter 6. While the methods of Chapter 5 and 6 are very similar, a different method (using a different type of data) has been used in Chapter 7 – there, structural alerts for liver toxicity have been created.

6. Identification of nuclear receptor ligands associated with hepatotoxicity*

6.1. Introduction

Chapter 5 provided an illustration of how innovative methods could be applied to develop workflows to predict potential RAR binders. The work in Chapter 6 extends this approach to further nuclear receptors (NRs). NRs belong to a large group of ligand-inducible transcription factors highly relevant to toxicology. The expression of target genes is mediated and inhibited by the presence or absence of endogenous ligands respectively. The regulated genes are associated with a variety of physiological processes, such as metabolism, development and reproduction (Wang and LeCluyse, 2003). NRs are well characterised with regard to their protein structure; at the N-terminal there is a DNA-binding domain (DBD), and at the C-terminal there is a ligand binding domain (LBD). NRs exist as monomers and (hetero- and homo-)dimers, for example the liver X receptor (LXR) is bound to the retinoid X receptor (RXR) (Maglich *et al.*, 2001; Wang and LeCluyse, 2003).

There are two main types of NRs: Type I receptors are usually found in the cytoplasm and are linked to the heat shock protein 90 (HSP90) which assists in the folding of the NR and protects it from heat stress. Type I ligands travel to their specific NR target through the bloodstream bound to a steroid binding globulin. Ligand binding occurs within the cytoplasm before entering the nucleus. After dissociation of HSP90, homodimerisation and binding to the hormone response element is initiated, which finally promotes the target gene expression via TATA (5'-TATAAA-3') binding protein, transcription factor II B, RNA polymerase II *etc*. (Kersten *et al.*, 2000; Maglich *et al.*,

* This chapter is based on my contribution to Mellor *et al.* (2015)

2001). Type II ligands bind to the NR complex within the nucleus. Ligand binding causes dissociation of co-repressors and association of co-activators, for example to inhibit histone deacetylases whose enzymatic activity is responsible for coiling DNA – which is unfavourable for protein expression (Sonoda *et al.*, 2008). On a molecular level, the DBD consists of two zinc fingers that recognise specific DNA sequences and the LBD contains the ligand-dependent activation function - responsible for ligand-protein-interactions similar to other pharmacologically relevant receptors. NR ligands are usually small, moderately lipophilic compounds, *e.g.* steroids, retinoids, fatty acids (Maglich *et al.*, 2001; Moya *et al.*, 2010). Many NR ligands are associated with liver toxicity, in particular fatty liver (hepatosteatosis) and the pathways responsible have been demonstrated for many NRs (refer to Table 6.1) (Moya *et al.*, 2010; Vinken, 2013; Sahini and Borlak, 2014). In this study there is a focus on ligands of the RAR (refer to Chapter 5), the peroxisome proliferator-activated receptor (PPAR), the liver X receptor (LXR) and the retinoid X receptor (RXR) which are associated with genetic expression towards hepatic lipid accumulation and *de novo* fatty acid synthesis (Moya *et al.*, 2010; Vinken, 2013; Sahini and Borlak, 2014). Beside the significant hepatosteatotic effects of many NRs, the farnesoid X receptor (FXR) is also investigated in this study. The exact role of FXR within lipid homeostasis is not clear yet, but due to its cholestasis promoting effects, it is relevant for hepatotoxicity as well (Vinken, 2015). Similar to the approach described for RAR ligands in Chapter 5 (Steinmetz *et al.*, 2015a), it is possible to build KNIME workflows for other NR targets relevant for hepatosteatosis (Moya *et al.*, 2010; Sahini and Borlak, 2014).

Table 6.1: Summary of studied NRs and relevance to hepatosteatosis and hepatotoxicity

NR name	NR subtypes	Relevance for hepatosteatosis	References

Retinoic acid receptor (RAR)	NR1B1 (α-RAR)
	NR1B2 (β-RAR)
	NR1B3 (γ-RAR)
		All three subtypes are present in the liver. There are multiple theories for hepatic lipid accumulation.	BioGPS, 2015;
			Moya *et al.*, 2010;
			Hewitt *et al.*, 2013

Peroxisome proliferator-activated receptor (PPAR)	NR1C1 (α-PPAR) NR1C2 (β/δ-PPAR) NR1C3 (γ-PPAR)	α-PPAR is most present in the liver. Increased hepatic peroxisome expression is associated with lipid accumulation.	BioGPS, 2015; Wang *et al.*, 2002
Liver X receptor (LXR)	NR1H3 (α-LXR) NR1H2 (β-LXR)	Both subtypes present in the liver; increasing cholesterol/ lipid synthesis.	BioGPS, 2015; Schultz *et al.*, 2000
Retinoid X receptor (RXR)	NR2B1 (α-RXR) NR2B2 (β-RXR) NR2B3 (γ-RXR)	Heterodimerisation with RAR, PPAR and LXR, *i.e.* ligands support above described NR mechanisms.	Pérez *et al.*, 2012; Sahini and Borlak, 2014
Farnesoid X receptor (FXR)	NR1H4 (α-FXR) NR1H5 (β -FXR)*	Mostly present in liver and adrenal cortex, FXR activation is not associated with hepatosteatosis, but associated with jaundice and cholestasis.	BioGPS, 2015; Vinken, 2015

*No data on β-FXR found; α-FXR and FXR often used as synonyms

It must be pointed out that the different NR subtypes (refer to Table 6.1) are structurally very close to each other and therefore often bind to the same ligands with similar affinity (Steinmetz *et al.*, 2015a). RAR and RXR share the same ligands, whereas RXR appears to be more ligand-specific (Minucci *et al.*, 1997). Furthermore, it must emphasised that the role of RAR/RXR agonists regarding their hepatosteatotic mechanisms is not fully understood and has promoted a much, sometimes controversial, discussion in the scientific literature (Moya *et al.*, 2010; Sahini and Borlak, 2014).

Hepatosteatosis is a multifactorial condition, which is usually triggered by a combination of drugs and (fatty) diet. Depending on the degree of lipid accumulation, the condition may lead to a non-alcoholic steatohepatitis (NASH), a chronic liver inflammation which may lead to fibrosis and eventually liver failure. Prevalence in the European population is about 30% for hepatosteatosis with 10% to 20% developing NASH, *i.e.* there is public interest in identifying substances contributing to hepatosteatosis (Schattenberg and Schuppan, 2011; Sahini and Borlak, 2014). Hence the aim of this chapter was to build a tool which enables the identification of potential ligands for NR-associated hepatotoxicity ligands based on the ligands' inherent chemical information.

6.2. Methods

Due to the functional similarity of NR-subtypes and the opportunity to compile large datasets from ChEMBL and PDB, four NR-workflows have been developed:

- RAR/RXR
- PPAR
- LXR
- FXR (only α-FXR)

6.2.1. Analysis of ligand-protein interaction

The PDB 3.3 was searched for human NR structures for all relevant subtypes. The structures obtained were investigated visually with regard to their ligand-protein-interaction within PyMOL 1.3 (PyMOL, 2014). Structural features of the ligands, particularly when apparently responsible for similar ligand-protein-interactions, were extracted. These structural features were coded manually into SMARTS strings. Subsequently the SMARTS strings were used within a rule-based workflow to test the predictive potential toward NR ligands.

6.2.2. Analysis of physico-chemical properties

Similar to Chapter 5, ligands for all subtypes of LXR, PPAR, RAR/RXR and FXR were compiled from the ChEMBL database (ChEMBL, 2014). Their pChEMBL values were used as a means to order activity. Only chemicals with a pChEMBL values > 5 were considered as being active. CDK's molecular properties node was used to calculate all relevant descriptors within KNIME (KNIME, 2014): MW for molecular size, VAIM and eccentric connectivity index (ECI) for structural complexity, RB for molecular flexibility and topological polar surface area (TPSA) and XLogP for hydro- and lipophilicity respectively.

6.3. Results

The investigation of physico-chemical properties of the ChEMBL ligands and the analysis of the protein-ligand interaction of the PDB data resulted in the information presented below. Further details, such as activity data, SMARTS strings and the rules for the incorporation of SMARTS strings and property ranges within a screening tool (relevant for writing codes / rebuilding the workflow) are presented in appendix A.5 and B.3.

6.3.1. RAR/RXR

After observing the RAR and RXR receptors separately it was noted that their actives had very similar binding patterns and it was decided to combine them into one workflow. A total of 958 RAR actives and 1188 RXR actives were extracted from the ChEMBL database. 20 human RAR structures bound to different ligands were retrieved from the PDB. The RXR had a further 64 different human protein structures within the PDB.

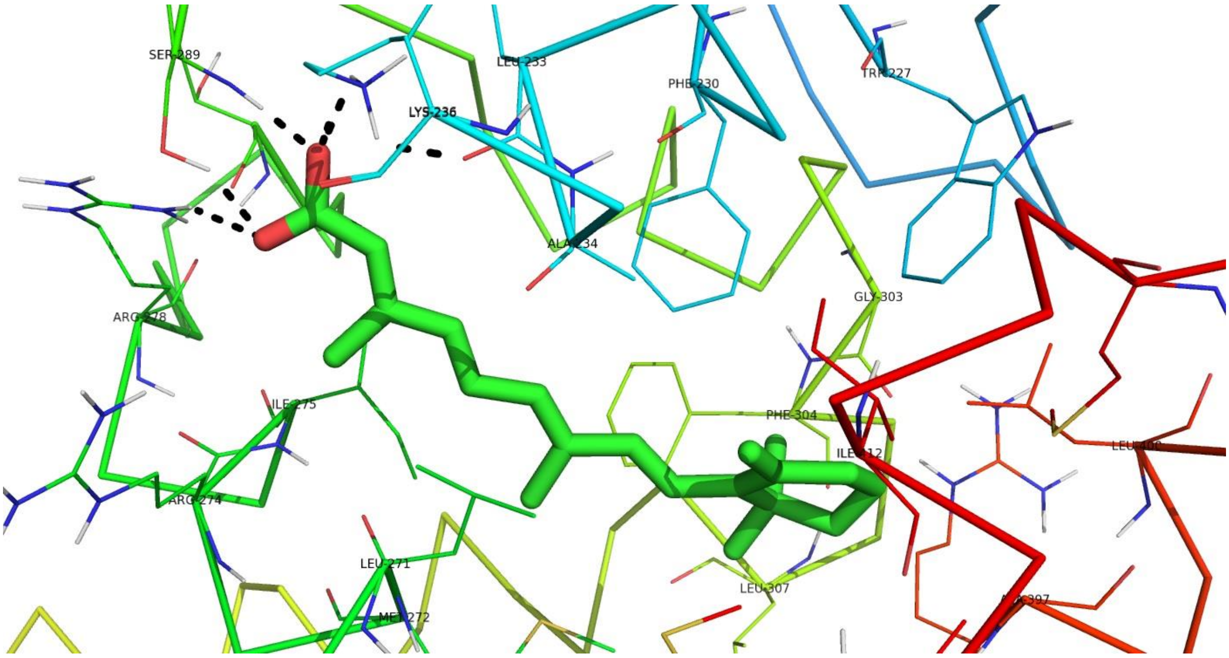


Figure 6.1: Ligand-protein-interaction of the γ-RAR (2LBD) indicating hydrogen bonds of retinoic acid within the receptor pocket (PDB, 2015)

The information obtained from observing RAR/RXR ligand-protein-interactions, as well as common substructures and physico-chemical properties were combined to form a rule-based screening workflow. The physico-chemical properties of the RAR/RXR actives were observed and the ranges that chemicals must fall within to be active were defined. The physico-chemical properties selected were MW, VAIM, RB and XLogP. It can be concluded that RAR/RXR ligands have a generally flexible, lipophilic and mostly aliphatic scaffold in common, as described in Table 6.2.

Table 6.2: Physico-chemical property ranges for RAR/RXR actives

Physico-chemical property	Value
MW	≤ 550
VAIM	5 - 7
RB	3 - 30
XLogP	2.2* - 20

* with 3 outliers < 2.2 (refer to RAR/RXR rules in Appendix B.3)

The rules in Table 6.2 can be similarly interpreted to Chapter 5’s Table 5.2. In other words, the values and ranges of Table 6.2 can be seen as single and double cut-offs for each property. Generally RAR/RXR ligands are lipophilic, however there are a small number of compounds which are active without being lipophilic (XLogP < 2.2), *e.g.* n-phosphono-L-phenylalanyl-L alanylglycinamide with an XLogP of -2.4. As these compounds have peptide-like bonds, XLogP exception rules were created. To narrow down the compounds passing through this physico-chemical filter, such as inactive amino sugars, a further filter was used. As shown in Figure 6.1, there are certain groups in the ligand (in particular double-bond oxygens), binding to one or two arginine residuals of the receptor, *e.g.* the hydrogen bond between arginine R278 and an oxygen of a ligand's carboxylic group within the RAR domain. In addition, serine S289 seems to support this functional group with another hydrogen bond (refer to Fig. 6.1). The responsible structural features are described in a structural alert system. Furthermore, RAR/RXR ligands contain at least one ring structure, which could be aromatic or aliphatic, *e.g.* cyclohexene of retinoic acid, as expressed in the structural alert system (refer to Appendix B.3) (PDB, 2015; Klaholz *et al.*, 2000; Steinmetz *et al.*, 2015a).

6.3.2. PPAR

A total of 8548 PPAR actives were extracted from the ChEMBL database. The following summarises the ligand-binding interactions of the PPAR actives observed within the PDB. In total 175 human PPAR structures were found.

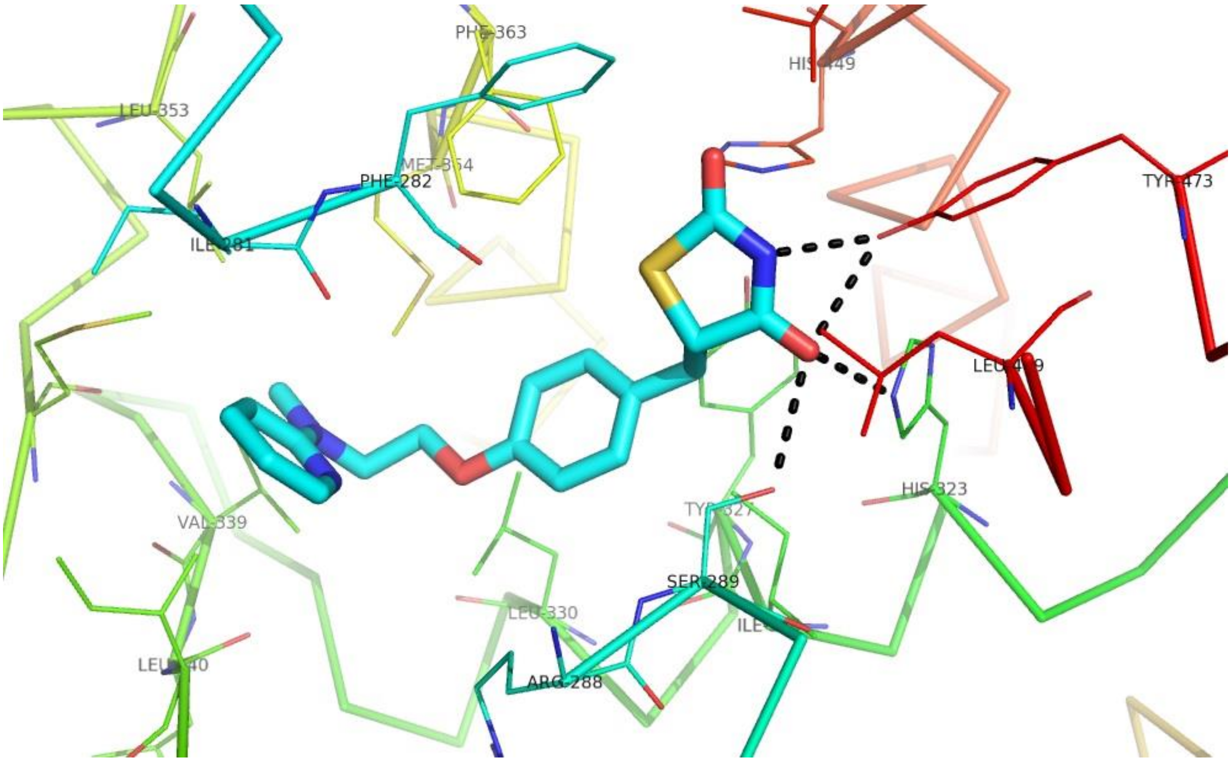


Figure 6.2: Ligand-protein-interaction of the γ-PPAR (4O8F) indicating hydrogen bonds with the polar ring system of a ligand and amino acid residues (PDB, 2015)

The information obtained from observing PPAR ligand-protein-interactions, common substructures and physico-chemical properties were combined to form a rule-based screening workflow. The physico-chemical properties of the PPAR actives were observed and ranges that chemicals must fall within to be active were defined (Table 6.3). The physico-chemical properties chosen as filters were MW, VAIM, TPSA and XLogP.

Table 6.3: Physico-chemical property ranges for PPAR actives

Physico-chemical property	Value
MW	≤ 800
VAIM	5 - 7
TPSA	20 - 300
XLogP	1.2 - 20

PPAR actives were studied and the substructural features relevant to activity were coded into SMARTS (as described in Appendix B.3). As none of the 8548 actives has a steroid-typical tricyclic backbone, which is in itself rather unusual for most NR ligands, an exclusion filter was created, *i.e.* the chemical of interest must not contain a steroid backbone to be classified as active. Further, the compound must contain one of the specific “diaromatic” scaffolds and one of the specific functional groups, which arise from hydrogen bonds to relevant amino acid residues (mainly from tyrosine, serine or histidine), in order to be classified as an active (refer to Fig. 6.2). An additional alert system describing fatty acid- and retinoid-like compounds (similar to RAR/RXR), indicates moderate PPAR affinity so triggers an alert (refer to Appendix B.3). This extension of the PPAR workflow is similar to the RAR/RXR workflow, *i.e.* it is searching for an identifying for mostly aliphatic, flexible chains with terminal polar groups (preferably carboxylic groups). Only one PPAR ligand from ChEMBL with a low activity (pChEMBL = 6.09) was not identified.

6.3.3. LXR

A total of 1721 LXR actives were extracted from the ChEMBL database. The following summarises the ligand-binding interactions of the LXR actives observed within the 16 human structures found within the PDB (PDB, 2015).

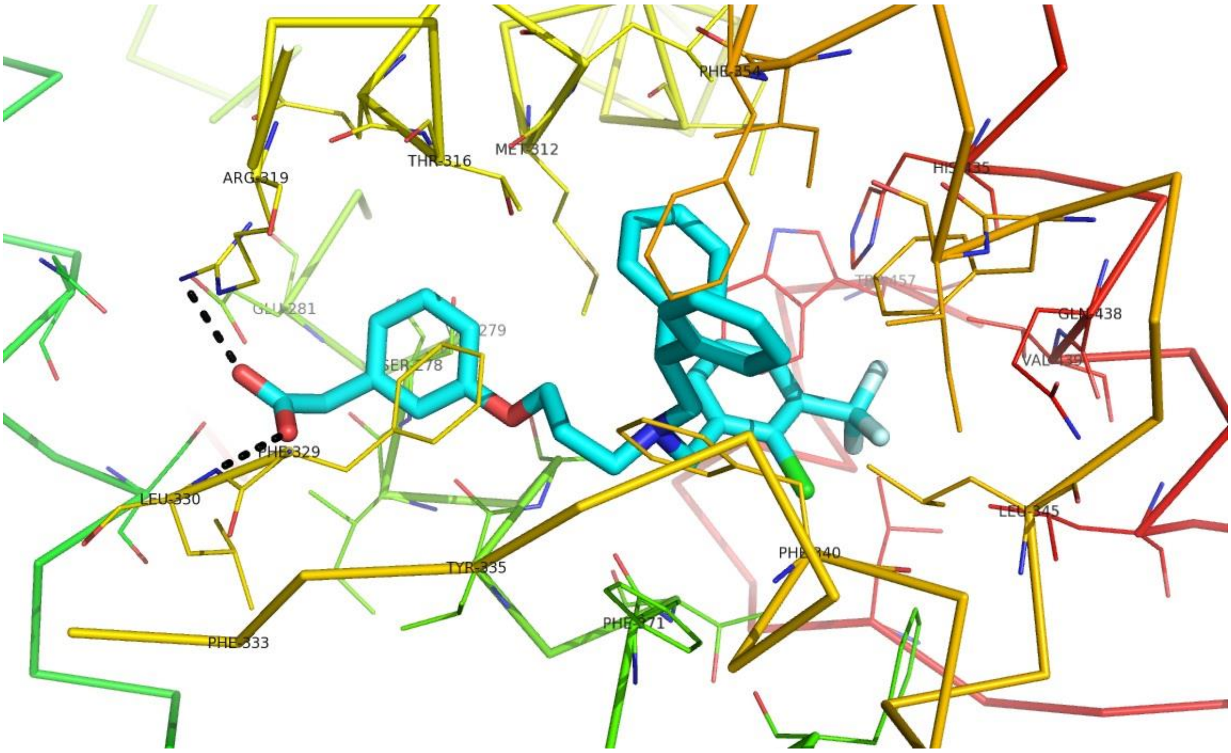


Figure 6.3: Ligand-protein-interaction of the β-LXR (4NQA) indicating hydrogen bonds and π-system-interactions as relevant for receptor binding (PDB, 2015)

The information obtained from observing LXR ligand-protein-interactions, common substructures and physico-chemical properties were combined to form a rule-based screening workflow. The physico-chemical properties of the LXR actives were observed and ranges that chemicals must fall within to be active were defined (Table 6.4). The physico-chemical properties chosen as filters were MW, VAIM and XLogP and TSPA.

Table 6.4: Physico-chemical property ranges for LXR actives

Physico-chemical property	**Range**
MW	≤ 750
VAIM	4.7 - 7
TPSA	5 - 150
XLogP	≥ 2

LXR actives were studied and the substructural features were coded into SMARTS (refer to Appendix B.3). A potential ligand contains ring backbones, which have certain interactions with arginine residues or secondary amine of leucine, particularly with a carboxylic group, aromatic methoxy groups and other “terminal” oxygens (refer to PDB, 2015: 3LOE, 4NQA, 4DK7). On the halogenated, particularly fluorinated, side of the ligand there might be interactions with histidine as well as shown in protein-ligand-structure 3FAL (H419). Additionally, π-stacking and similar interactions of the ligand’s aromatic rings to phenylalanine and tryptophan residues may be relevant for the ligand-receptor-binding. An example ligand, mostly known as GW-3965, binding to β-LXR is presented in Figure 6.3.

6.3.4. *FXR*

Although FXR agonists are not associated with NASH directly, their contribution towards cholestasis *etc.* makes them worth identifying. Hence, a total of 26 human FXR structures were found in the PDB and further 715 active ligands where retrieved from ChEMBL.

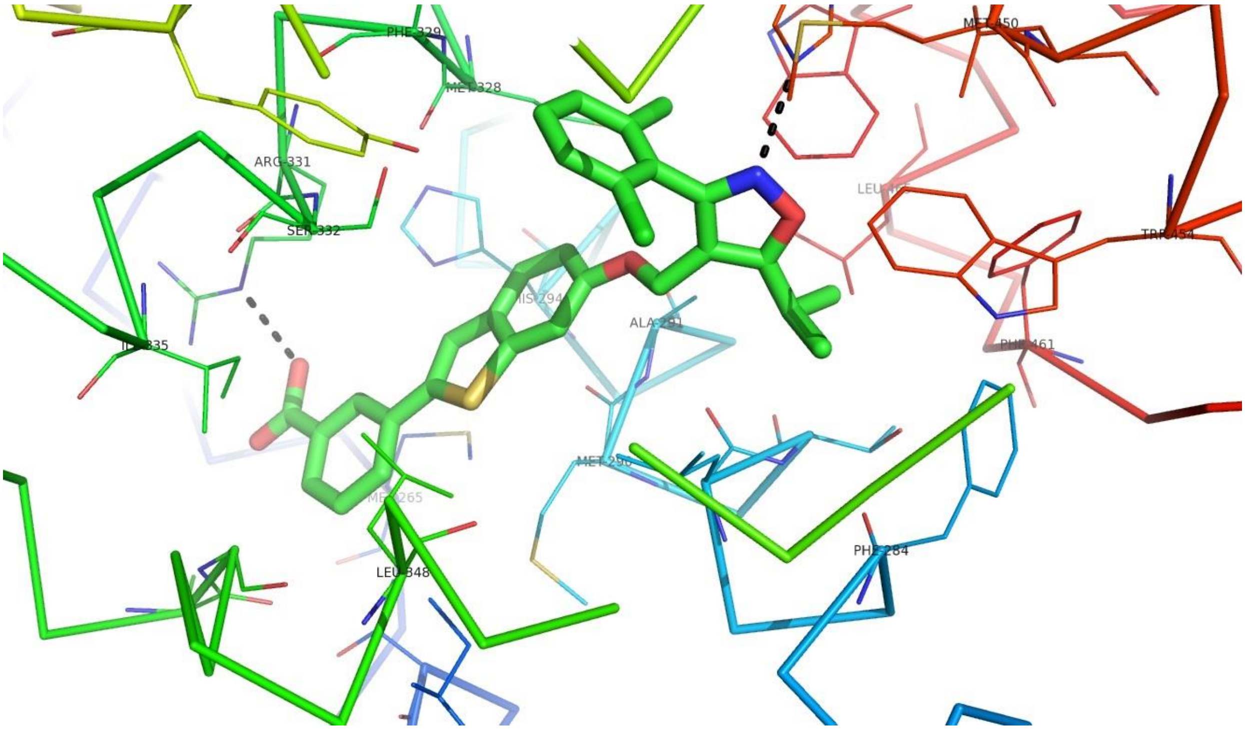


Figure 6.4: Ligand-protein-interaction of the FXR (3HC5) indicating hydrogen bonds on two ends of a lipophilic **“**tunnel” (PDB, 2015)

The information obtained from observing FXR ligand-protein-interactions, common substructures and physico-chemical properties were combined to form a rule-based screening workflow. The physico-chemical properties of the FXR actives were observed and ranges that chemicals must fall within to be active were defined (Table 6.5). The physico-chemical properties chosen as filters were MW, ECI, RB and TPSA.

Table 6.5: Physico-chemical property ranges for FXR actives

Physico-chemical property	Range
MW	≤ 900
ECI	150 - 2400
RB	≥ 2
TPSA	15 - 200

Different active ligands of the FXR were investigated (refer to PDB, 2015: 4WVD (former 4II6), 3HC5, 4QE6) and structural features relevant regarding shape and functionality, in particular formation of hydrogen bonds, were coded in SMARTS (refer to Appendix B.3). Beside the fit of the molecule within the lipophilic “tunnel”, which is determined by residues of leucine, isoleucine *etc.*, hydrogen bonds to polar residues of amino acids, such as serine, tyrosine, histidine and arginine, seem to be important for binding. Both interactions are depicted in Figure 6.5, where a synthetic ligand binds within the receptor pocket. Natural ligands are often steroid compounds, *e.g.* chenodeoxycholic acid (refer to bile salts). 11 of the 715 compounds from ChEMBL did not pass the identification criteria of the FXR-workflow. However, these outliers had only low activity (pChEMBL ≤ 5.56), such as clotrimazole, an antifungal pharmaceutical, and JHW-015, a synthetic cannabinoid – both with a pChEMBL value of 5.49.

6.4. Discussion

The results from within Chapter 6, *i.e.* the classification rules for each NR, can be used to build a workflow to identify potential receptor binders (as achieved in Chapter 5). The information/codes necessary for this type of workflow are gathered in Appendix B.3. The workflow based on these rules is then able to identify compounds potentially acting as NR ligands, *i.e.* RAR/RXR, PPAR, LXR and FXR ligands respectively. As potential NR ligands are likely to trigger a MIE associated with toxic effects, such as hepatosteatosis or cholesatasis, this information is highly relevant for risk assessment within an AOP or read-across concept (Vinken, 2013; Schultz *et al.*, 2015; Patlewicz *et al.*, 2014).

The predictive power of the workflow is difficult to measure as there are no test/validation data available. As the SMARTS strings and rules are created manually all data need to be considered as "training data". Without approaches such as cross-validation or bootstrapping, it is impossible to express statistical performance. However, what can be observed is that the training data, the ChEMBL datasets, were all identified by the rules, except one PPAR and eleven FXR ligands with low activity (pChEMBL ≤ 6.09). In other words, only twelve out of 13130 actives were not assigned correctly. This indicates a certain degree of sensitivity, but it is likely that specificity is only moderate to low, *i.e.* false positives are likely. To confirm suggested activity, it is recommended to support the NR-workflow results with additional *in vitro* assays and/or further *in silico* tests, *e.g.* docking, QSAR models.

The concept described here (and in Chapter 5) can be extended to many other receptor-mediated toxicity endpoints and so support regulatory toxicology and, thus, tasks such as risk and safety assessment. Of course, specificity of such a model is always dependent on the actual content of databases, *e.g.* ChEMBL and PDB. Further it must be noted that other hepatosteatosis and cholestasis pathways exist, for example via mitochondrial toxicity (Patel and Sanyal, 2013), *i.e.* this concept cannot to be extended to cover all relevant MIEs for hepatosteatosis and cholestasis. On the other hand “ligands for NRs” is an important class of toxicants.

Regarding the FXR, a recent illustration of an AOP for the bile salt export pump (BSEP) has been described which states how FXR agonists may significantly induce cholestasis and jaundice (Vinken, 2015). However, with regard to FXR’s role in lipid homeostasis, it is still unclear whether FXR ligands promote hepatosteatosis (refer to Moya *et al.*, 2010). Liu *et al.* (2014b) have actually postulated that the agonist may counteract hepatosteatotic effects. This leads to the question of whether FXR antagonists may trigger hepatosteatosis. In particular, as the screening workflows developed within this thesis are not likely to differentiate between competitive agonists and antagonists (refer to Chapter 5). Nevertheless it could be worth screening for FXR ligands with regard to hepatosteatosis. It must also be pointed out that adverse effects which expand onto a histopathological level, *e.g.* fibrosis, steatosis, cholestasis, may contribute to similar clinical symptoms and manifestations.

Even if this thesis is more focussed on deriving knowledge from hepatotoxic AOPs, it should be emphasised that adverse effects of NR ligands are not only restricted to hepatotoxicity. Beside clinical pharmacology and toxicology, the identification of NR ligands and prediction of their potential targets is of great interest for ecotoxicology and environmental health as well, *e.g.* the identification of endocrine disruptors in wastewater. In modern risk assessment, particularly when avoiding animal testing, it is important to use synergies from these different disciplines.

6.5. Conclusions

Identification of AOPs and MIEs is a growing topic in many sub-disciplines of toxicology, *e.g.* regulatory toxicology and risk assessment. NR ligands trigger toxicological effects such as hepatosteatosis and NASH respectively. A workflow predicting NR-associated hepatotoxicity (predominately hepatosteatosis) has been developed. Combined with further NR-workflows (*e.g.* glucocorticoid, oestrogen and vitamin D receptor) the workflow is available in COSMOS Space, a user interface hosted by the COSMOS project (Mellor *et al.*, 2015; COSMOS, 2015). A screenshot of the principle design of a screening KNIME workflow is shown in Chapter 5 (Fig 5.3) and a summary of rules (including SMARTS strings) is presented in Appendix B.3.

This chapter extends the results of Chapter 5, but it is also an excerpt of Mellor *et al.* (2015), which deals with additional NRs. While many computational toxicology tools focus on reactivity (*e.g.* Enoch *et al.*, 2011), receptor-mediated toxicity is an essential addition in the greater picture of toxicology and risk assessment. This shows the relevance of this work for the COSMOS project and for non-testing approaches (*e.g.* read-across) in the regulatory context.

7. *In silico* prediction of liver toxicity: The development of novel structural alerts*

7.1. Introduction

Liver toxicity is an important endpoint for human toxicology in general. The liver plays an important role in many metabolic pathways, for example, in the digestion of food or the clearing and transforming of xenobiotics. The organ is well perfused, with arterial and (portal) venous blood. Damage of the liver can result to liver failure and even cause death (Chen *et al.*, 2011). There are different types of pathologies of the liver that can be promoted by chemical insult, such as inflammation, fibrosis, steatosis, cholestasis and cancer (Fourches *et al.*, 2010; Hewitt *et al.*, 2013; Przybylak and Cronin, 2012). MIEs relevant for hepatosteatosis and cholestasis are discussed in Chapter 6. Particularly relevant for hepatotoxicity are compounds taken orally (*e.g.* via food, drugs *etc.*) that may enter the liver after passing through the gastro-intestinal tract. If compounds, so absorbed, are potentially toxic, *e.g.* inducing mitochondrial toxicity, the liver is often the first organ to suffer from their potential adversity (Mennecozzi *et al.*, 2012). Furthermore metabolic activation of compounds to more toxic forms might take place in the liver that leading to subsequent toxic effects, particularly occurring in adjacent organelles, cells and tissues (Barile, 2004; Rang *et al.*, 2007a). Systemically available compounds which have not been altered metabolically (for example by the first-pass effect) can also damage the liver. They will pass through the liver at high volume and may continuously harm the liver via the systemic circulation. Chronic exposure of toxicants may also manifest in a substance-induced liver injury. As cosmetic ingredients are often intended to be applied regularly, safety assessment must encompass hepatotoxicity as well. This is the rationale behind SEURAT-1's focus on finding alternative testing methods to identify hepatotoxicity, that

* This chapter is an extension of the work of Hewitt *et al.* (2013)

could otherwise only be identified in animal intensive (sub)chronic assays (Yamada *et al.*, 2013; Tralau *et al.*, 2015).

Besides *in vitro* and *in vivo* data dealing with mechanisms and pathologies of liver toxicants, human liver toxicity data are rare and of poor quality respectively. In particular chronic hepatotoxicity studies are mostly based on pharmacovigilance and other epidemiological studies, for example the harvesting of clinical literature data by Fourches *et al.* (2010) and Chen *et al.* (2011). Regarding pharmacovigilance data, it must be stressed, that adverse effects to the liver are a leading cause of pharmaceuticals failing during clinical trials and being withdrawn from the market. These adverse effects may vary in severity and type of liver injury, *e.g.* steatosis, cholestasis or acute liver failure (Chen *et al.*, 2011). As liver injury is not exclusively relevant to pharmaceuticals, methods to identify liver toxicants (*e.g.* structural alerts) may be used for the risk assessment of other classes of compounds, for example, cosmetics, food additives.

Prediction of liver toxicity may be interpreted differently according to the respective field. For example in the clinical environment biomarkers, such as elevated aminotransferases or bilirubin (refer to jaundice) or the decrease of specific proteins (*e.g.* kallistatin), are used as indicators for liver toxicity (Cheng *et al.*, 2015b; Przybylak and Cronin, 2012). However, these biomarkers are associated with existing liver injury; hence they play an important role in clinical liver function tests. The question rises to what extent these biomarkers can be used in *in vitro* assays and if they could be applied to the entirety of hepatotoxic mechanisms. In practice, there is no single *in vitro* hepatotoxicity assay, but there are assays for different modes of action / mechanisms, *e.g.* mitochondrial dysfunction, protein adduct formation (Mennecozzi *et al.*, 2012). This scenario is similar for *in silico* approaches for hepatotoxicity in the context of predictive toxicology (refer to Chapter 1). Generalised computational hepatotoxicity models do not work sufficiently and further they do not differentiate mechanisms in a transparent way. Nevertheless, there are promising models within mechanistic categories (*i.e.* groups of similarly acting compounds). This is a strong argument for the development of tools to identify mechanisms and MIEs respectively (as undertaken in Chapter 5 and 6). Additionally it should be pointed out that there are further factors, such as polymorphism and environmental factors, which impede the creation of computational models (Przybylak and Cronin, 2012).

Hewitt and colleagues (2013) defined structural alerts for liver toxicity based on structural similarity of pharmaceutical liver toxicants of the Fourches *et al.* (2010) database. They defined 16 structural alerts (glucocorticoid steroids, nitrogen mustards, catechols *etc.*) and investigated the mechanisms through an intensive literature research. In this study additional structural alerts were created from the investigation of the Chen database (Chen *et al.*, 2011). A comparison of the Fourches and Chen databases is shown in Table 7.1. Here it must be noted that a binary system (hepatotoxic and non-hepatotoxic) was used by Fourches, however a ternary system of classification was used by Chen. Hence, this does not allow for an exact comparison of the statistics. However, compounds of both datasets can be compared and investigated on a structural level. Structural alerts, for example, are a way to capture substructures (and related structural information) and further to screen compounds of interest for these substructural features. The structural alerts were coded in SMARTS strings (Daylight, 2014; Sushko *et al.*, 2012). The aim of this chapter was to create new structural alerts for liver toxicity (complementing the already existing Hewitt alerts) and so build a “finer mesh” to screen compounds for the potential substance-induced liver injury. Furthermore this work suggests a way to progress with structural alerts in general, *i.e.* first using already identified structural alerts on a new database to check validity and second, identifying new structures for potential adverse effects.

Table 7.1: Comparison of the liver toxicity databases of Fourches and Chen

	Fourches *et al.* (2010)		Chen *et al.* (2011)		
Origin of data	Literature research		Literature research		
Applicability	Human, pharmaceutical		Human, pharmaceutical		
Number of compounds	951		287		
Hepatotoxic categories	2 (0,1) 0 being no hepatotoxic concern 1 being high hepatotoxic concern		3 (0,1,2) 0 being no hepatotoxic concern 1 being low hepatotoxic concern 2 being high hepatotoxic concern		
	1 (n = 650)	0 (n = 301)	2 (n = 137)	1 (n = 85)	0 (n = 65)

7.2. Methods

7.2.1. Data collection and curation

The Chen *et al.* (2011) dataset contains information on 278 pharmaceuticals including descriptions of potential hepatotoxicity (refer to Table 7.1). This information describes on the severity based on the actual adverse effects, and in addition to the concern based on pharmacovigilance data, *i.e.* incorporating statistical evidence and severity – similar to the work of Fourches *et al.* (2010), it is derived from the scientific literature. SMILES strings for each compound were taken from Chemspider (RSC, 2014). After screening all 278 structures with the 16 structural alerts of Hewitt *et al.* (2013) with KNIME (refer to section 7.2.4), the hepatotoxic compounds, which have not been classified as such (*i.e.* false negatives), were investigated for molecular similarity.

7.2.2. Molecular similarity

To identify new categories of structurally similar compounds of the false negative hepatotoxicant from section 7.2.1, the freeware Toxmatch v1.07 (IDEAconsult, 2014b) was used. While the compounds have been tested against each other for atom environment similarity, the similarity matches were defined as being greater than or equal 60%. Compounds with multiple matches have been extracted from the dataset to investigate common substructures and scaffolds.

7.2.3. Creating structural alerts for liver toxicity

After the structures were investigated regarding their similarity, they were visually/manually grouped based on common structural features. After the grouping process, novel structural alerts were created by embedding relevant structural information into code. The novel structural alerts were coded in SMARTS strings, so they can be used to identify unknown compounds of the same category. The purpose of the new structural alerts is to complement the Hewitt *et al.* (2013) alerts, as the correctly identified hepatotoxic compounds (true positives) were not considered. In other words, this study focusses on the hepatotoxicants, which cannot be identified by Hewitt *et al.* (2013) alerts alone.

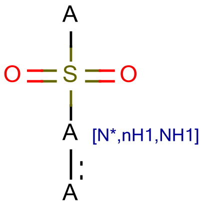
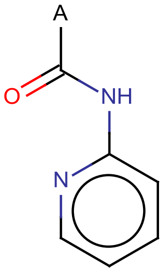
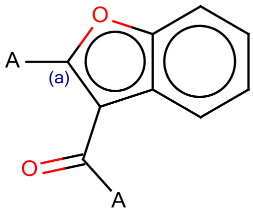
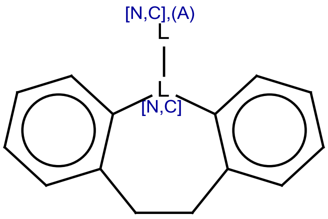
7.2.4. Screening and validation of structural alerts

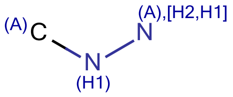
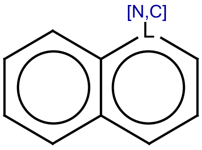
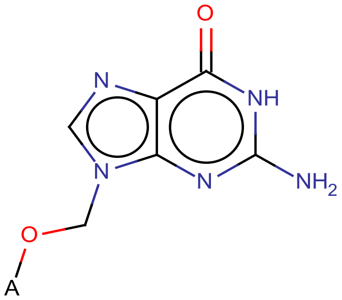
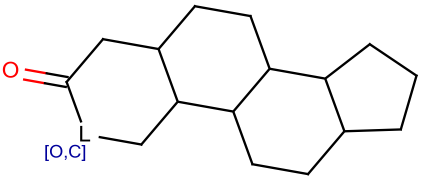
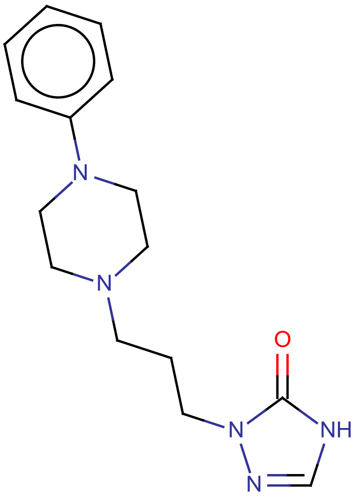
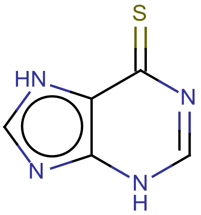
The screening itself (employing the respective structural alerts) was done within KNIME 2.7.2 (KNIME, 2014) using the Indigo substructure matcher. The screening performance of the Hewitt *et al.* (2013) alerts and of a combination of the Hewitt *et al.* (2013) and the novel alerts was investigated using the Chen *et al.* (2011) and the Fourches *et al.* (2010) datasets. For the novel structural alerts, the Chen *et al.* (2011) dataset can regarded as training dataset and the Fourches *et al.* (2010) dataset can be regarded as validation dataset. However it must be stressed that there is a large overlap of compounds in both datasets and even some with conflicting information on hepatotoxicity.

7.3. Results and discussion

The Chen *et al.* (2011) dataset holds 196 compounds of moderate to high hepatotoxicological concern, which were not identified by the Hewitt *et al.* (2013) alerts. Based on the outcome of the structural similarity analysis with Toxmatch, ten novel structural alerts (presented as SMARTS strings in Table 7.2). Each of the substructures describes a group of compounds with a potential for hepatotoxicicity.

Table 7.2: Ten novel structural alerts for liver toxicity[x]

1	*[N*,nH,NH]S(*)(=O)=O	
2	[*]C(=O)Nc1ccccn1	
3	[*]c1oc2ccccc2c1C([*])=O	
4	[N,C][N,C]1c2ccccc2CCc2ccccc12	

5	C[NH1][NH2,NH1]	
6	c1ccc2[n,c]cccc2c1	
7	Nc1nc2n(CO*)cnc2c(=O)[nH]1	
8	O=C1CC2CCC3C4CCCC4CCC3C2C[O,C]1	
9	O=C1NC=NN1CCCN1CCN(CC1)c1ccccc1	
10	S=C1N=CNc2[n]cnc12	

ˣa complete list of the combined 26 hepatotoxicity alerts (Hewitt *et al.* and Steinmetz) is attached to Appendix B.1

The principle idea of the creation of the novel structural alerts from this analysis (refer to Table 7.2) was to combine them with the 16 structural alerts of Hewitt *et al.* (2013). In the Chen *et al.* (2011) database the Hewitt *et al.* (2013) structural alerts identified 14 of 137 compounds with high hepatotoxic concern, 12 of 85 compounds with low hepatotoxic concern and 4 of 65 compounds with no hepatotoxic potential. Combining the Hewitt *et al.* (2013) alerts with the novel alerts of Table 7.2 led to the identification of 43 of 137 compounds with high hepatotoxic potential, 27 of 58 low hepatotoxic concern and 5 of 60 no hepatotoxic concern. So while increasing the identification of compounds with high hepatotoxic potential by 21.2%, the incidence of false positives (no hepatotoxic concern) has only been increased by 1.5 % (refer to Table 7.3).

When applying the same comparison to the Fourches database as a test set, similar differences can be seen. The combined 26 alerts identified 169 of 650 compounds with high hepatotoxic potential, whereas the Hewitt *et al.* (2010) alerts on their own only identified 108 of 650 compounds with high hepatotoxic potential. This means nearly 9.4% more potentially hepatotoxic compounds have been identified with the combined 26 alerts. On the other hand, the number of false positives increased from 41 to 56 from a total of 301 non-hepatotoxic compounds. This is an increase of false positives of 5.0% (refer to Table 7.4).

Table 7.3: 16 and 26 structural alerts tested on Chen *et al.* (2011) database

Using 16 Alerts (Hewitt *et al.*, 2013)				**Using 26 Alerts (combined)**			
Hepatotoxic category	Match	No match	Total	Hepatotoxic category	Match	No match	Total
2	14	123	137	2	43	94	137
1	12	73	85	1	27	58	85
0	4	61	65	0	5	60	65
Total	30	257	287	Total	75	212	287

Table 7.4: 16 and 26 structural alerts tested on Fourches *et al.* (2010) database

Using 16 Alerts (Hewitt *et al.*, 2013)				**Using 26 Alerts (combined)**			
Hepatotoxic category	Match	No match	Total	Hepatotoxic category	Match	No match	Total
1	108	542	650	1	169	481	650
0	41	260	301	0	56	245	301
Total	149	802	951	Total	225	726	951

The hepatotoxic categories from Table 7.1 express hepatotoxic concern, *i.e.* there are the binary Fourches (0,1) and the ternary Chen (0,1,2) dataset. The Fourches and the Chen datasets have an overlap of 107 compounds. The majority of the compounds defined as hepatotoxic by Chen (1,2) are hepatotoxic by Fourches as well (1) and *vice versa*. One exception would be the vasodilator benziodarone (Chen: 2), which was classified by Fourches as non-hepatotoxic (0). Conversely there are exceptions, such as the anticonvulsant pyrimidinedione primidone (Chen: 0) or the sugar alcohol *D*-(-)-mannitol (Chen: 0), which are both classified as hepatotoxic by Fourches (1).

As hepatotoxicity is such a complex, multifactorial phenomenon, it is difficult to improve performance of screening models based on structural alerts. Different pathways relevant for metabolic activation or deactivation, different genetic heritages leading to enzymatic polymorphism and different environmental factors, such as beverages, foods, drugs *etc.*, play an important role in the manifestation of liver toxicity. Of course a screening tool raising alerts might be useful for drug development, but gaining knowledge of the exact mechanisms and kinetics relevant for histopathologies associated with hepatotoxicity is of greater importance (Przybylak and Cronin, 2012; Chen *et al.*, 2011).

7.4. Conclusions and perspectives

Identifying compounds which are likely to be liver toxicants is crucial in drug design but also convenient in other areas such as safety and risk assessment. Defining more structural alerts for liver toxicity could be beneficial for many areas; from occupational health to assessing consumer products, *e.g.* cosmetics. The development of the novel structural alerts within this work led to the identification of 61 more potential liver toxicants from the databases, with only a very small increase in false positives (15), within the Fourches dataset (refer to Table 7.4).

Principally the approach of this work is that it can be repeated with every new pharmacovigilance liver toxicity dataset, and every time new structural alerts would be created. Over time this could lead to a “finer mesh” of alerts (or a “tighter sieve”) enabling the screening of a wide range of chemical compounds for substance-induced liver injury. Future work may also include the refinement of alerts, for example the steroid alerts of Hewitt *et al.* (2013), which do not identify all hepatotoxic steroids of Chen *et al.* (2011). As this work proposes another steroid alert (refer to SMARTS string no. 8 in Table 7.2), it is an interesting question if a simpler, broader steroid alert might be of more value, particular for the alerts’ predictive power.

As liver toxicity is a complex phenomenon, which can be triggered by many different mechanisms (Mennecozzi *et al.*, 2012), it is unlikely that liver toxicity could be determined by structural alerts alone – at least not as an accurate predictive tool. Particularly as the absence of an alert does not make a compound non-toxic (Przybylak and Cronin, 2012), there would be a danger of combining structural alerts to increase sensitivity until specificity is lost (and false positives become rather rule than exception).

Beside the refined methodology, the current structural alerts combined with the 16 liver toxicity alerts of Hewitt *et al.* (2013), as summarised in Appendix B.1, are a tool able to support current non-testing approaches and *in silico* risk assessment. However, it must be pointed out that they represent only a small excerpt of potential hepatotoxicants. The approach, even while using highly relevant data, is less mechanistically driven than the approach in Chapter 5 and 6. However, in the context of the AOP framework and the usage of the methods suggested in Chapter 5 and 6, results could become more refined – particularly when considering the relevance of clinical/pharmacovigilance data to today's risk assessment approaches.

8. Discussion

8.1. Summary of work

The assessment of the risk following exposure to any chemical, and hence its safety, is a non-trivial task and dependent on various types of information such as exposure, toxicokinetics and mode of action (MoA). When dealing with untested compounds, predictions based on data for tested compounds have to be made. This is usually performed with predictive toxicology tools, *e.g.* QSAR models or read-across. One problem with these approaches is that they are highly dependent on the data quality of already tested compounds. In Chapters 2 and 3 it has been demonstrated that existing, historical biological data are often of poor or unknown quality. Confidence in the use of such data can only be achieved by repetition of tests by independent researchers, with examples provided from the areas of skin penetration and aquatic toxicology where many historical data are available (Steinmetz *et al.*, 2014; Steinmetz *et al.*, 2015b). Similar problems regarding irreproducibility were investigated by Gottmann and colleagues (2001) who, following an investigation of carcinogenicity data, concluded that two large rodent datasets were only 57% concordant. The differences in bioassay results were not explainable by sex, species, strain or target organ. The problems with data quality raises the question of how many predictive toxicology models and expert systems are based on reliable data and assumptions of data, and how they might change when only high quality data are used. One reason for this lack of knowledge is that it is difficult to determine the impact on existing models since there are few duplicate experimental data; this is because it is usually considered as a waste of money and resources, in addition to the issues relating to ethics, to re-test substances. It should also be recognised that further technical replicates are not likely to be the key to solve this problem, as most sources of error cannot be excluded by immediately/simultaneously repeating an assay with the same sample, reagent, apparatus, staff *etc*. (Madden *et al.*, 2012). Generally accuracy and precision of experimental values are important factors associated with repeatability and experimental error. On the one hand, consistency in the execution of experiments, for example by the usage of standard operating procedures (SOPs), can lower experimental errors. On the other hand, biological data are dependent of too many variables to address issues with accuracy and precision properly. These circumstances lead to a general uncertainty of biological data. One solution is to apply the confidence score (CS) within QSAR modelling, development of structural alerts, read-across or validation approaches to improve predictive toxicology. Even if the rebuilding and renewing of well used models incorporating more reliable data and/or confidence (as presented in this thesis) improves models only slightly, ultimately they provide a more realistic (weight of) evidence-based approach, as demanded by the users of predictive toxicology tools. In the long run, being able to assess the uncertainty of data could enable a measure of confidence to be assigned to predictions from QSAR models *etc.* For instance, greater confidence could be achieved by focusing model development and validation on the most reliable data points (for example applying the methods described in Chapters 2 and 3).

The assessment of data quality is obviously vital for their use in risk assessment and modelling. However, data quality is not the only criterion required to perform risk assessment; it is essential to obtain information on exposure, kinetics and, if available, MoA as well. With regard to cosmetic ingredients in particular, knowledge of skin permeability and dermal absorption is crucial. Whilst skin permeability only describes the passage of a molecule through the uppermost skin layers, such as the *stratum corneum*, dermal absorption data describe the complete process from the dermal administration of a compound to its detection in the blood. Whereas data for skin permeability have the advantage that they can be generated using an *in vitro* assay (using, for instance, human skin), the investigation of dermal absorption is typically dependent on *in vivo* tests. It must be remembered that dermal metabolic activity is neglected in skin permeability tests as opposed to dermal absorption tests, as in rare cases results and conclusions may differ. In addition the dermal absorption of chemicals by rodents is likely to be higher than for humans due to their higher pore density (simply due to rodents having fur) compared to human skin. Even if animal tests on cosmetic ingredients are not allowed anymore, this must be factored into *in silico* predictions employing *in vivo* data (Mitragotri *et al.*, 2011; Ravenzwaay and Leibold, 2004; Hughes and Edwards, 2010).

Regardless of the data type indicating dermal absorption, without systemic availability it is unlikely for a substance to reveal systemic toxic effects. The skin permeability of a compound is determined by many factors such as its concentration, volatility, effect of the formulation or delivery vehicle and the substance-specific k_p value. This k_p value is one of the key (physico-)chemistry-dependent properties reliant on molecular size and lipophilicity (Potts and Guy, 1992; Mitragotri *et al.*, 2011). A validated, robust QSPR model to predict the k_p value of an untested compound is reported in Chapter 3 (Eq. 3.5). Dermal absorption, similar to oral bioavailability, is dependent on a compound's physico-chemical properties. The principle concept is that small, moderately lipophilic (*i.e.* uncharged) compounds pass through relevant membranes more easily than ionic or large compounds (Lipinski *et al.*, 2001; Mitragotri *et al.*, 2011). Based on the same scientific background, a rule-based prediction system has been built to classify dermal absorption of hair dyes (refer to Chapter 4). The rules are based on the same idea of small, moderately lipophilic compounds passing through relevant membranes. Beside MW and log P, information on MP and TPSA improved the performance of the rule-based prediction models. Many hair dyes are known for being toxic, for example due to protein and DNA binding. This indicates a principle hazard for mutagenicity and eventually carcinogenicity (Vinardell *et al*, 2015; Nelms *et al.*, 2015). It is important for consumers not to absorb these compounds in quantities where they might cause systemic toxicity, *e.g.* geno-, hepato- or nephrotoxicity. Regarding risk assessment, it is assumed that low dermal availability is a key factor for safety. Both methods, the QSPR model (refer to Chapter 3) and the rule-based prediction system (refer to Chapter 4), can support risk assessment of cosmetic ingredient regarding their dermal and systemic availability. Whilst the QSPR is designed for a broad applicability domain of organic chemicals, the rule-based prediction system is designed for hair-dyes and associated substances. Another big difference is that the QSPR model calculates continuous data points, *i.e.* absolute k_p values, the rule-based prediction system orders compounds into two classes according to an internally decided safety threshold for hair dyes and associated substances.

As well as kinetics, MoA and, in particular, MIEs are of great interest for predictive toxicologists. As MIEs represent the initial chemical interaction of a chemical compound with a biological target (*e.g.* protein), every toxicological effect is based on at least one MIE. Predictive toxicology tries to associate chemical information with adverse effects, hence the prediction of initial chemical interactions is probably the most obvious approach to this field. This thesis (and also the COSMOS Project and indeed the SEURAT-1 Cluster) has had a clear focus on hepatotoxicity. As a typical pathology caused by chronic, systemic toxicity, adverse effects to the liver are relevant for consumer goods including cosmetic products (Tralau *et al.*, 2015; Vinardell, 2015). Many different mechanisms are known to cause hepatotoxicity however, with regard to cosmetic ingredients, there is great interest, and indeed need, to identify compounds with the ability to cause adverse effects at low doses and following repeated exposure. Amongst other mechanisms, NR ligands play an important role in liver toxicity. Since many NR ligands have the potential to trigger cholestasis and hepatosteatosis (Vinken, 2015; Mellor *et al.*, 2015). An *in silico* method to screen for potential agonists has been developed and described in Chapter 5 (Steinmetz *et al.*, 2015a). This methodology, which employs calculated physico-chemical properties and combinations of structural features, was extended in Chapter 6 to be applicable to more NRs (Mellor *et al.*, 2015). The *in silico* screening tools described in Chapters 5 and 6 are developed by using *in vitro* datasets, *i.e.* clinical relevance for every prediction cannot be assured to the same extent as for models exclusively based on *in vivo* data. It must be noted that *in vitro* and *in vivo* data always embody limitations towards the prediction of clinical outcomes, and that even clinical data are not always suitable (refer to polymorphism). Nevertheless, the *in silico* screening tools built within this thesis can be used for the generation of new leads in drug design and as a tool in risk assessment (*e.g.* for category formation) due to their conservative and generalistic nature of the model. Additionally prioritisation, for example in ecotoxicology, to identify hazardous agents would be another application for the screening tool.

The results described in Chapter 7 are more clinically relevant as they address clinical (pharmacovigilance) data from Fourches *et al.* (2010) and Chen *et al.* (2011). It must be pointed out that cosmetic ingredients should generally (with a few exceptions) not have any clinical relevance, *i.e.* pharmacological/toxicological effects. The liver toxicity data have been used to create structural alerts to identify potential hepatotoxicants. The work described in Chapter 7 is principally an extension of the approach taken by Hewitt *et al.* (2013), who presented a comprehensive review on mechanisms of liver toxicity and assigned a small number of structural alerts to these mechanisms. The focus of this study was not to extend the review but develop further structural alerts for hepatotoxicity. The structural alerts defined in Chapter 7 represent substructures associated with hepatotoxicity under chronic administration; ten new alerts have been added to the existing sixteen of Hewitt *et al.* (2013). Since the *in silico* screening tools (refer to Chapters 5 and 6) predict only the potential to elicit a MIE, they do not predict toxicological effect directly. Whereas the structural alerts of Chapter 7 predict clinical toxicity directly. Even if the answers of the structural alerts model seem to be more relevant, it must be noted that the data were very limited regarding structural diversity and data quality. Practically this means, that the combined structural alerts of Steinmetz and Hewitt together do not capture the entirety of hepatotoxicants. However, as shown in Chapter 7, the extension enables to identify many more hepatotoxicants without compromising integrity. All workflows and codes needed to build the described screening tools are attached in Appendices.

This thesis provides a number of different approaches relevant to risk assessment, *e.g.* models for hazard identification and exposure, so consideration is required as to how this body of work contributes to the toxicological assessment of novel cosmetic ingredients. The integration of data quality, kinetics and mechanistic modelling is a challenge, of course, and, furthermore, the precise manner in which this is carried out will depend strongly on individual scenarios. In general, the aim of predictive toxicology strategies is to combine all available knowledge to obtain the most plausible prediction for the relevant compound, so issues such as toxicity, safety, risk and exposure can be addressed properly, *e.g.* bans/limits of chemicals for certain applications. Risk assessment uses a variety of data from hazard and exposure, as well as use case, to make a decision. Traditionally these data include information from animal tests (*e.g.* NOAEL). As this type of testing is not acceptable anymore, at least not for novel cosmetic ingredients, new strategies will need to be applied. The development of a model predicting NOAEL values for all relevant biological endpoints and all types of organic chemicals in a near future is rather unlikely. Nevertheless, when going more into detail, plenty of useful tools and strategies have been proposed – not only within this thesis. Since statistical and individual assessment of data quality contributes to higher certainty, more robust models and read-across approaches can be built and employed. This benefits the confidence in the obtained prediction. Furthermore, predictive models within a specific applicability domain of interest, *e.g.* a chemical family and their relation towards biological effects, can be expressed in a mechanistically more transparent fashion. Added to this, expert knowledge and computational tools dealing with ADME properties will predict uptake (*e.g.* dermal absorption), distribution, potential metabolites and elimination rates. Last, but not least, category formation, supported by screening tools, docking *etc.*, may help to build local QSARs or binary data systems (*e.g.* toxic, non-toxic). Whilst consensus approaches employing different prediction tools (for example as demonstrated in Norlen *et al.* (2014)) can be used to predict the same endpoint, integration in the context of this discussion rather means the combination of different endpoints or effects, such as dermal absorption, mitochondrial toxicity, NR affinity *etc.*, to support or reject a complete toxicological hypothesis, *e.g.* low doses of the compound administered dermally may trigger lipid accumulation in hepatocytes. The integration of multiple datasets and methods and the interpretation of the predictions are complex endeavours, which still demand toxicological expertise. Therefore the integration, as it is discussed in this thesis, is far from an automated predictive toxicology tool, which just needs to be fed with data and provides results ready for the use of regulatory authorities.

8.2. Future of risk assessment

Due to the complex chemical exposure scenarios of modern humans, risk and safety assessment is – and most likely will stay – a challenging task. Consumer goods of all kinds are designed and marketed worldwide. Beside a general lack of (high quality) toxicological and pharmacological data, it is difficult to consider individual contributing factors, such as potential adverse effects due to polymorphisms or due to synergy of different ingredients. With the new legislation, the lack of traditional toxicology data is becoming greater, which creates a big challenge for regulatory toxicologists. What initially appears to be a setback for regulators, the ban of animal tests for cosmetics and cosmetics ingredients in the European market, appears to be an advantage on further consideration. Whilst NOAEL and LOEL data are used to determine toxicity (with huge variability according to Gottmann *et al.*, 2001), mechanistic knowledge is often not incorporated. Furthermore Lewis *et al.* (2002) described many potential flaws in the way toxicity data, such as NOAEL values, are generated. These flaws are, for example, a lack of definitions of toxicology terms and inconsistent interpretations of such. Even histopathological data, if retrieved at all, are often not enough to identify or comprehend toxicity-driving MoAs. Since the information historically obtained by *in vivo* testing must be replaced somehow, it is necessary to focus on the advances of *in vitro* and *in silico* technologies.

Of course in this thesis the focus lies on computational, non-testing strategies. Modern *in silico* tools often directly, or indirectly, employ, or can be used to gain, insights into mechanistic knowledge, *i.e.* information regarding enzymatic inhibition, receptor binding, electrophilic attack of DNA or proteins *etc*. Whereas local QSARs, docking, structural alerts and other screening tools often have direct associations with a single mechanism, machine learning approaches may be used to gain insights into mechanisms of action if their predictions can be interpreted in terms of the contributions of individual molecular features, as described by Palczewska *et al.* (2013). Furthermore, ADME- and PBPK-modelling provides answers regarding the localisation and enrichment of compounds (refer to target organ toxicity) and the nature of the chemical species reaching the site of biological action (which may be affected by metabolism of the original chemical). Hence, integration of kinetic and mechanistic modelling may benefit regulatory toxicology, which has been up until now dominated by *in vivo* testing. In non-testing strategies, regulatory toxicology is dominated by read-across approaches, at least for systemic toxicity. In addition to the purely computational tools described, *in vitro* assays may benefit predictions and hence toxicological assessments. By integration of multiple *in vitro* and *in silico* tests (refer to Integrated Testing Strategies) the joined information may support kinetic or mechanistic hypotheses. On the one hand it is a philosophical question to what extent tests need to be conducted to obtain satisfactory information, on the other hand it is a quite practical question as well. However, without neglecting the limitations of *in vitro* assays as compared to *in vivo* tests, it should be pointed out that the similar considerations are applicable to *in vivo* testing, for example when deciding on species, sex, dose pattern *etc.*, and, as just as importantly, what analysis and histopathological investigations to perform (Hartung *et al.*, 2013). Driven by the costs of testing and the plethora of chemicals and mixtures, the philosophical question has to turn into a pragmatic question to obtain a pragmatic answer regarding potential risks.

Beyond setting the scene of the current, the alleged turning point of risk and safety assessment influenced by 21st Century Toxicology and the increase of ethical concerns, it is not clear to what extent risk and safety assessment will change and if the new paradigm will be sustainable. However, given the ever increasing number of new chemicals submitted to regulators for market approval, even if risk assessment processes have a very low error rate, it seems inevitable that some high risk chemicals may still make it onto the market. Particularly under capitalism, which urges the launch of novel products, it may be only a question of time when the next “chemical catastrophe” happens. Insufficiently assessed drugs, such as thalidomide (Dally, 1998) or fenfluramine/phentermine (Surapaneni *et al.*, 2011), or food adulterants, such as melamine in milk (Wu and Zhang, 2013) or tricresyl phosphate in ginger extract (Morgan and Penovich, 1978), may change the attitude of the public (and hence politicians), which may lead to more toxicological vigilance regarding risk and fewer ethical concerns regarding animal trials. Whatever will happen in 21st century regarding regulatory toxicology is difficult to predict. Of course politics play an important role regarding toxicological decision-making. However, it us up to scientists, particulary young scientists from the "digital native" generation, to face current challenges and influence policy makers and public opinion towards careful, but also feasible and sensible, assessment of chemicals, such as cosmetic ingredients. Technological development is moving on a fast pace – many tools we use nowadays naturally have been advanced technologies, with limited access, a few decades ago. The trend of freeware, open-access databases and transparent programming is contributing *in silico* toxicology massively. 21st Century Toxicology has already brought many new insights, not only to me, but to the whole scientific community, and there are still some years to come.

8.3. Conclusions

The work presented in this thesis on statistical data quality, kinetics, *i.e.* skin permeability and dermal absorption, and mechanistically based modelling (hepatotoxicity alerts and prediction of NR-associated hepatosteatosis) are pieces designed to be integrated into modern toxicological risk and safety assessment. In particular, the work in this thesis should support risk assessment in the field of cosmetic ingredients. As opposed to many current achievements in predictive toxicology, transparency and flexibility of models and tools are strengths of this thesis as compared to statistical performance. Tools, methods and strategies are published and proposed respectively within this thesis and the associated scientific articles (refer to Appendix D). Beyond cosmetics, beneficial usage within other disciplines, for example pharmacology, is discussed in many chapters. The overall aim was to propose ideas regarding how to build and interpret new models, and of course how to use them in combination (refer to consensus approach) with a focus on safety assessment in the consumer care industry. More specifically, a method for calculating statistical data quality (to improve QSAR modelling and other computational approaches), new models for skin permeability / dermal absorption, screening tools for NR-binding and hepatotoxicity structural alerts have been created. All these models and tools have the potential to support toxicological evaluations and regulatory decision-making directly. However, the methodologies used in development of the tools and models might support regulatory toxicology even more indirectly. The transparent design of workflows (mostly within KNIME) and statistics, the usage of freeware (*e.g.* R, Toxmatch) and the usage of open-access data (*e.g.* ChEMBL, PDB) will give toxicologists and regulators the opportunity to adjust tools and models towards their relevant problems. In a nutshell, the aim of novel but transparent *in silico* tools and models to support the assessment of cosmetics ingredients can be regarded as accomplished on multiple levels.

8.4. Future work

As described earlier (refer to Chapter 1), toxicology is a great field with far too many chemicals, endpoints *etc.* to ever fully elucidate this discipline. Even restricting ourselves to particular aspects of this field, such as the topics investigated within this thesis, it seems to be an almost endless story. From protein interactions to pathological pathways, a lot of things are not currently as clear and understandable as they should be. Biological data are often not of the necessary data quality (refer to Chapters 2 and 3), probably due to the complexity of biomolecular assays, which makes toxicological interpretation rather difficult. Tendencies as described in skin permeability / dermal absorption (Chapters 3 and 4) or in NR-related mechanistic modelling (Chapters 5, 6 and 7) are used for predictions. When considering all those possible influences, such as metabolism or active transport, predictions led purely by toxicologically relevant properties cannot always be correct. So, there is still plenty room to improve, but this is difficult due to a lack of data and hence a lack of understanding. Nevertheless, predictive toxicology is of great importance for many different areas, although there are limiting factors, such as the many unknown mechanisms and, of course, the limited availability of data. The most limiting factor probably is, and always will be, the complexity of the human body. However, even if not perfectly, it is likely that QSARs and AOPs will join to quantitative AOPs ready for endpoint- and MoA-specific predictions in the near future. They will have the great advantage of transparency compared to global, endpoint-specific QSAR models, which do not account for mechanistic information.

As mentioned earlier, it is unlikely that automated model development will deal with toxicological issues in the near future. The development of a predictive toxicology tool, which just needs to be fed with data and spits out results ready for use, is not foreseeable. But, that is not all bad news. The good news is that many computational toxicologists can follow up in this interesting field, probably for many decades, or even centuries, to come. Particularly as data, descriptors and knowledge constantly increase (and even sometimes change), there could be many full-time jobs created for computational toxicologists, just for updating and refining already existing models. Despite this endeavour, there is still the potential for real novelties which might be based on new pharmacological/toxicological insights (*e.g.* new AOPs) or the combination of different technologies (*e.g.* combining QSPRs with PBPK-models). However, all these approaches have the potential to create game-changing models and affect the way regulatory toxicologists approach the challenges of tomorrow.

9. References

Abraham M.H. *et al.* (1997) Algorithms for skin permeability using hydrogen bond descriptors: the problem of steroids. *J. Pharm. Pharmacol.* 49: 858-865

Adams J. (1993) Structure-activity and dose-response relationships in the neural and behavioral teratogenesis of retinoids. *Neurotoxicol. Teratol.* 15: 193-202

Adler S. *et al.* (2011) Alternative (non-animal) methods for cosmetics testing: current status and future prospects-2010. *Arch. Toxicol.* 85: 367-485

Alam M.S. *et al.* (2013) Application of 2D-GCMS reveals many industrial chemicals in airborne particulate matter. *Atmos. Environ.* 65: 101-111

Alizadeh F. *et al.* (2014) Retinoids and their biological effects against cancer. *Int. Immunopharmacol.* 18: 43-49

Allen J.G. and Bloxham D.P. (1989) The pharmacology and pharmacokinetics of the retinoids. *Pharmac. Ther.* 40: 1-27

Allen T.E.H. *et al.* (2014) Defining molecular initiating events in the adverse outcome pathway framework for risk assessment. *Chem. Res. Toxicol.* 27: 2100-2112

AltTox (2015) Table of validated and accepted alternative methods. http://alttox.org/mapp/table-of-validated-and-accepted-alternative-methods/ (accessed 24.08.2015)

Ankley G.T. *et al.* (2010) Adverse outcome pathways: a conceptual framework to support ecotoxicology research and risk assessment. *Environ. Toxicol. Chem.* 29: 30-41

Aptula A.O. *et al.* (2002) Multivariate discrimination between modes of toxic action of phenols. *Quant. Struct. Act. Rel.* 21: 12-22

Aruoja V. *et al.* (2011) Toxicity of 58 substituted anilines and phenols to algae *Pseudokirchneriella subcapitata* and bacteria *Vibrio fischeri*: comparison with published data and QSARs. *Chemosphere* 84: 1310-1320

Ates G. *et al.* (2015) Linking existing *in vitro* dermal absorption data to physicochemical properties: contribution to the design of a weight-of-evidence approach for the safety evaluation of cosmetic ingredients with low dermal bioavailability. Submitted to *Regul. Toxicol. Pharm.*

Attene-Ramos M.S. *et al.* (2013) The Tox21 robotic platform for the assessment of environmental chemicals - from vision to reality. *Drug Discov. Today* 18: 15-16

Attene-Ramos M.S. *et al.* (2015) Profiling of the Tox21 chemical collection for mitochondrial function to identify compounds that acutely decrease mitochondrial membrane potential. *Environ. Health Perspect.* 123: 49-56

Backhaus T. *et al.* (1997) Toxicity testing with *Vibrio fischeri*: a comparison between long term (24 h) and the short term (30 min) bioassay. *Chemosphere* 35: 2925-2938

Barile F.A. (2004) Clinical toxicology - principles and mechanisms. CRC Press, Boca Raton, Florida, USA, pp. 35-75

Beard R.L. *et al.* (2002) Synthesis and biological activity of retinoic acid receptor-specific amides. *Bioorg. Med. Chem. Lett.* 12: 3145-3148

Benigni R. and Giuliani A. (1994) Quantiative structure-activity relationship (QSAR) studies in genetic toxicology: mathematical models and the “biological activity” term of the relationship. *Mutat. Res.* 306: 181-186

Bento A.P. *et al.* (2013) The ChEMBL bioactivity database: an update. *Nucleic Acids Res.* 42: 1-8

Berman H.M. *et al.* (1999) The Protein Data Bank. *Nucleic Acids Res.* 28: 235-242

Berthold M. *et al.* (2007) Knime: the Konstanz Information Miner, studies in classification, data analysis, and knowledge organization (GfKL), Springer-Verlag, Freiburg, Germany, pp. 1-8

Biesalski H.K. (1989) Comparative assessment of the toxicology of vitamin A and retinoids in man. *Toxicology* 57: 117-161

BioGPS (2015) http://www.biogps.org (accessed 17.07.2015)

Blackburn K. and Stuard S.B. (2014) A framework to facilitate consistent characterization of read across uncertainty. *Regul. Toxicol. Pharmacol.* 68: 353-362

Bláha L. *et al.* (1998) QSAR for acute toxicity of saturated and unsaturated halogenated aliphatic compounds. *Chemosphere* 36: 1345-1365

Bulich A.A. *et al.* (1981) Reliability of the bacterial luminescence assay for determination of the toxicity of pure compounds and complex effluents. *Aquatic Toxicology and Hazard Assessment, ASTM STP* 737: 338-47

Burden N. *et al.* (2015) Aligning the 3Rs with new paradigms in the safety assessment of chemicals. *Toxicology* 330: 62-66

Calleja M.C. *et al.* (1994) Human acute toxicity prediction of the 50 MEIC chemicals by a battery of ecotoxicological tests and physicochemical properties. *Food Chem. Toxicol.* 32: 173-187

Campbell T.J. (1983) Importance of physico-chemical properties in determining the kinetics of the effects of Class I antiarrhythmic drugs on maximum rate of depolarization in guinea-pig ventricle. *Br. J. Pharmac.* 80: 33-40

Caron G. and Ermondi G. (2008) Lipophilicity: chemical nature and biological relevance. In: Mannhold R. (editor). Molecular drug properties - measurement and prediction. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, pp. 326-327

Cases M. *et al.* (2013) The eTOX library of public resources for *in silico* toxicity prediction. *Mol. Inform.* 32: 24-35

Chang J.C. *et al.* (1981) Use of Microtox assay system for environmental samples. *Bull. Environ. Contam. Toxicol.* 26: 150-156

Chauhan P. and Shakya M. (2010) Role of physicochemical properties in the estimation of skin permeability: *in vitro* data assessment by Partial Least-Squares regression. *SAR QSAR Environ. Res.* 21: 481-494

ChEMBL (2014) Database version 19. http://www.ebi.ac.uk/chembl (accessed 07.07.2014)

ChEMBL (2015) Database version 19, http://www.ebi.ac.uk/chembl/ (accessed 11.02.2015)

Chen M. *et al.*, (2011) FDA-approved drug labelling for the study of drug-induced liver injury. *Drug Discov. Today* 16: 697-703

Chen L. *et al.* (2013) Recent advances in predicting skin permeability of hydrophilic solutes. *Adv. Drug Deliv. Rev.* 65: 295-305

Cheng Z. *et al.* (2015a) Replacing fish meal by food waste to produce lower trophic level fish containing acceptable levels of polycyclic aromatic hydrocarbons: health risk assessments. *Sci. Total Environ.* 523: 253-261

Cheng Z. *et al.* (2015b) Kallistatin, a new and reliable biomarker for the diagnosis of liver cirrhosis. *Acta Pharm. Sin.* B 5: 194-200

Cooper J.A. *et al.* (1979) Describing the validity of carcinogen screening tests. *Br. J. Cancer* 39: 87-89

COSMOS (2015) COSMOS Space, cosmosspace.cosmostox.eu/app/home (accessed 22.07.2015)

Couling D.J. *et al.* (2006) Assessing the factors responsible for ionic liquid toxicity to aquatic organisms via quantitative structureÐproperty relationship modelling. *Green Chem.* 8: 82-90

Cronin M.T.D. and Schultz T.W. (1996) Structure-toxicity relationships for phenols to *Tetrahymena pyriformis*. *Chemosphere* 32: 1453-1468

Cronin M.T.D. and Schultz T.W. (1997) Validation of *Vibrio fischeri* acute toxicity data: mechanism of action-based QSARs for non-polar narcotics and polar narcotic phenols. *Sci. Total Environ.* 204: 75-88

Cronin M.T.D. and Schultz T.W. (1998) Structure-toxicity relationships for three mechanisms of action of toxicity to *Vibrio fischeri*. *Ecotox. Environ. Saf.* 39: 65-69

Cronin M.T.D. and Schultz T.W. (2003) Pitfalls in QSAR. *J. Mol. Struct. (Theochem)* 622: 39-51

Cronin M.T.D. *et al.* (1991) QSAR studies of comparative toxicity in aquatic organisms. *Sci. Total Environ.* 109/110: 431-439

Cronin M.T.D. *et al.* (2013) Applying read-across for quantitative chronic toxicity prediction. Abstract of 49th congress of the European Societies of Toxicology (EUROTOX), *Toxicol. Lett.* 221, pp. 52

Cronin M.T.D. (2004) Predicting chemical toxicity and fate - an introduction. In: Cronin M.T.D. and Livingstone D.J. (editors). Predicting chemical toxicity and fate (1st edition). CRC Press, Boca Raton, Florida, USA, pp. 3-12

Cronin M.T.D. (2013a) An introduction to chemical grouping, categories and read-across to predict toxicity. In: Cronin M.T.D. *et al.* (editors). Chemical toxicity prediction - category formation and read-across (1st edition). RSC Publishing, Cambridge, UK, pp. 1-2

Cronin M.T.D. (2013b) Evaluation of categories and read-across for toxicity prediction allowing for regulatory acceptance. In: Cronin M.T.D. *et al.* (editors). Chemical toxicity prediction - category formation and read-across (1st edition). RSC Publishing, Cambridge, UK, pp. 155-167

Curtis C. *et al.* (1982) Evaluation of a bacterial bioassay as a method for predicting acute toxicity of organic chemicals to fish. *Aquatic Toxicology and Hazard Assessment: ASTM STP* 766: 170-178

Czodrowski P. (2013) hERG Me Out. *J. Chem. Inf. Model.* 53: 2240-2251

Dally A. (1998) Thalidomide: was the tragedy preventable? *Lancet* 351: 1197-1199

Dancik Y. *et al.* (2013) Design and performance of a spreadsheet-based model for estimating bioavailability of chemicals from dermal exposure. *Adv. Drug Deliv. Rev.* 65: 221-236

Davis A.M. and Riley R.J. (2004) Predictive ADMET studies, the challenges and the opportunities. *Curr. Opin. in Chem. Biol.* 8: 378-386

Dawson D.A. *et al.* (2006) Chemical mixture toxicity testing with *Vibrio fischeri*: combined effects of binary mixtures for ten soft electrophiles. *Ecotox. Environ. Saf.* 65: 171-180

Daylight (2014) http://www.daylight.com (accessed 07.07.2014)

DeZwart D. and Slooff W. (1983) The Microtox as an alternative assay in the acute toxicity assessment of water pollutants. *Aquat. Toxicol.* 3: 129-138

Dicken C.H. (1984) Retinoids: a review. *J. Am. Acad. Dermatol.* 11: 541-552

Docherty K.M. and Kulpa C.F. (2005) Toxicity and antimicrobial activity of imidazolium and pyridinium ionic liquids. *Green Chem.* 7: 185-189

Dutka B.J. and Kwan K.K. (1981) Comparison of three microbial toxicity screening tests with the Microtox test. *Bull. Environ. Contam. Toxicol.* 27 (1981), 753-757

EC (2005) European Commission, regulation no. 389/2005 of the European Parliament and the council of 18 May 2005 (accessed 23.04.2014)

EC (2006) European Commission, regulation no. 1907/2006 of the European Parliament and the council of 18 December 2006 (accessed 23.04.2014)

EC (2009) European Commission, regulation no. 1223/2009 of the European Parliament and the council of 30 November 2009 (accessed 23.04.2014)

el Mansouri S. *et al.* (1995) Time- and dose-dependent kinetics of all-trans-retinoic acid in rats after oral or intravenous administration(s). *Drug Metab. Dispos.* 23: 227-231

Ellison C.M. *et al.* (2008) Definition of the structural domain of the baseline non-polar narcosis model for *Tetrahymena pyriformis*. *SAR QSAR Environ. Res.* 19: 751-783

Enoch S.J. *et al.* (2011) A review of the electrophilic reaction chemistry involved in covalent protein binding relevant to toxicity. *Crit. Rev. Toxicol.* 41: 783-802

Enoch S.J. *et al.* (2008) Classification of chemicals according to mechanism of aquatic toxicity: an evaluation of the implementation of the Verhaar scheme in Toxtree. *Chemosphere* 73: 243-248

EPA (2013) US Environmental Protection Agency. EPI Suite 4.1 software. http://www.epa.gov/oppt/exposure/pubs/episuite.htm (accessed 17.01.2013)

EPA (2014) EPI Suite 4.11, US Environmental Protection Agency. http://www.epa.gov/opptintr/exposure/pubs/episuite.htm (accessed 20.08.2014)

EPA (2015a) ACToR, US Environmental Protection Agency. http://www.epa.gov/actor/ (accessed 20.02.2015)

EPA (2015b) ECOTOX Database, US Environmental Protection Agency. http://cfpub.epa.gov/ecotox/ (accessed 20.02.2015)

Faustman G.M. and Omenn G.S. (2001) Risk assessment. In: Klaassen C.D. (editor). Casarett and Doull’s toxicology - the basic science of poisons (6th edition). McGraw-Hill, New York, USA, pp. 69-95

FDA (2013) US Food and Drug Administration, code of federal regulations title 21, April 2013. http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?fr=3.2 (accessed 03.08.2013)

Feigenbaum A. *et al.* (2015) Reliability of the TTC approach: learning from inclusion of pesticide active substances in the supporting database. *Fd. Chem. Toxicol.* 75: 24-38

Fioravanzo E. *et al.* (2013) Molecular modelling of LXR binding to evaluate the potential for liver steatosis, Abstract of 49th congress of the European Societies of Toxicology (EUROTOX), *Toxicol. Lett.* 221, pp. 83

Flynn G.L. (1990) Physicochemical determinants of skin absorption. In: Gerrity T.R. and Henry C.J. (editors). Principles of route-to-route extrapolation for risk assessment (1st edition). Elsevier New York, USA (1990), pp. 93-127

Fourches D. *et al.* (2010) Cheminformatics analysis of assertions mined from literature that describe drug-induced liver injury in different species. *Chem. Res. Toxicol.* 23: 171-183

Frank R.A. *et al.* (2010) Use of a (quantitative) structure-activity relationship [(Q)SAR] model to predict the toxicity of naphthenic acids. *J. Tox. Env. Health* A 73: 319-329

Freitag D. *et al.* (1994) Structural configuration and toxicity of chlorinated alkanes. *Chemosphere* 28: 253-259

Frisvad J.C. *et al.* (2006) Important mycotoxins and the fungi which produce them. *Adv. Exp. Med. Biol.* 571: 3-31

Froehner K. *et al.* (2000) Bioassays with *Vibrio fischeri* for the assessment of delayed toxicity. *Chemosphere* 40: 821-828

Futran Fuhrman V. *et al.* (2015) Why endocrine disrupting chemicals (EDCs) challenge traditional risk assessment and how to respond. *J. Hazard Mater.* 286: 589-611

Gälli R. *et al.* (1994) Toxicity of organophosphate insecticides and their metabolites to the water flea *Daphnia magna*, the Microtox test and an acetylcholinesterase inhibition test. *Aquat. Toxicol.* 30: 259-269

Gallo M.A. (2001) History and scope of toxicology. In: Klaassen C.D. (editor). Casarett and Doull’s toxicology - the basic science of poisons (6th edition). McGraw-Hill, New York, USA, pp. 15-20

Garcia M.T. *et al.* (2001) Fate and effect of monoalkyl quaternary ammonium surfactants in the aquatic environment. *Environ. Pollut.* 111: 169-175

Garcia M.T. *et al.* (2005) Biodegradable ionic liquids (Part II). Effect of the anion and toxicology. *Green Chem.* 7: 9-14

Garcia M.T. *et al.* (2008) Fate and effects of amphoteric surfactants in the aquatic environment. *Environ. Int.* 34: 1001-1005

Gaspar R. *et al.* (2012) Towards a European strategy for medicines research (2014-2020): the EUFEPS position paper on Horizon 2020. *Eur. J. Pharm. Sci.* 47: 979-987

Gavaghan C.L. *et al.* (2007) Development, interpretation and temporal evaluation of a global QSAR of hERG electrophysiology screening data. *J. Comput.-Aided Mol. Des.* 21: 189-206

Gocht T. and Schwarz M. (2014) Introduction. In: Gocht T. and Schwarz M. (editors). SEURAT-1: towards the replacement of *in vivo* repeated dose systemic toxicity testing, Vol. 4. European Commission for Research & Innovation, Imprimerie Mouzet, France, Paris, pp. 23-30

Gottmann E. *et al.* (2001) Data quality in predictive toxicology: reproducibility of rodent carcinogenicity experiments. *Environ. Health Perspect.* 109: 509-514

Groh K.J. *et al.* (2015) Development and application of the adverse outcome pathway framework for understanding and predicting chronic toxicity: I. Challenges and research needs in ecotoxicology. *Chemosphere* 120: 764-777

Gutsell S. and Russell P. (2013) The role of chemistry in developing understanding of adverse outcome pathways and their application in risk assessment. *Toxicol. Res.* 2: 299-307

Hartung T. *et al.* (2013) Food for thought … integrated testing strategies for safety assessments. *Altex* 30: 3-18

Hermens J. *et al.* (2005) Quantitative structure-activity relationships and mixture toxicity or organic chemicals in *Photobacterium phosphoreum*: the Microtox test. *Ecotoxicol. Environ. Saf.* 9: 17-25

Hewitt M. *et al.* (2013) *In silico* prediction of liver toxicity: chemical category formation, structural alert development and mechanism of action elucidation. *Crit. Rev. Toxicol.* 43: 537-58

Holdway D.A. *et al.* (1991) The acute toxicity of pulse-dosed, para-substituted phenols to larval American flagfish (*Jordanella floridae*): a comparison with toxicity to photoluminescent bacteria and predicted toxicity using log K_{OW}. *Sci. Total Environ.* 104: 229-237

Hrovat M. *et al.* (2009) Variability of *in vivo* fish acute toxicity data. *Regul. Toxicol. Pharm.* 54: 294-300

Hughes M.F. and Edwards B.C. (2010) *In vitro* dermal absorption of pyrethroid pesticides in human and rat skin. *Toxicol. Appl. Pharm.* 246: 29-37

IDEAconsult (2013) Toxtree v.2.5.0, http://www.ideaconsult.net/web/ngn/blogs/-/blogs/ new-toxtree-2-5-0-release (accessed 13.06.2013)

IDEAconsult (2014a) Toxtree v2.6.6. http://toxtree.sourceforge.net/ (accessed 03.12.2014)

IDEAconsult (2014b) Toxmatch v1.07 https://eurl-ecvam.jrc.ec.europa.eu/laboratories-research/predictive_toxicology/qsar_tools/toxmatch / (accessed 11.07.2014)

Jennings V.L.K. *et al.* (2001) Assessing chemical toxicity with the bioluminescent photobacterium (*Vibrio fischeri*): A comparison of three commercial systems. *Water Res.* 35: 3448-3456

Johnson M.E. *et al.* (1995) Permeation of steroids through human skin. *J. Pharm. Sci.* 84: 1144-1146

Johnson A.T. *et al.* (1999) Chandraratna. High affinity retinoic acid receptor antagonists: analogs of AGN 193109. *Bioorg. Med. Chem. Lett.* 9: 573-576

Kahru A. (1993) *In vitro* toxicity testing using marine luminescent bacteria *Photobacterium phosphoreum*: the Biotox^{TM} test. *ATLA-Altern. Lab. Anim.* 21: 210-215

Kaiser K.L.E. and Palabrica V.S. (1991) *Photobacterium phosphoreum* toxicity data index. *Water Poll. Res. J. Can.* 21: 361-431

Kersten S. *et al.* (2000) Roles of PPARs in health and disease. *Nature* 405 (2000), 421-424

Khajeha A. and Modarress H. (2014) Linear and nonlinear quantitative structure-property relationship modelling of skin permeability. *SAR QSAR Environ. Res.* 25: 35-50

King E.F. and Painter A.H. (1981) Assessment of toxicity of chemicals to activated sludge microorganisms. Acute Aquatic Ecotoxicological Tests, *INSERM* 10, 143-153

Klaholz B.P. *et al.* (2000) Structural basis for isotype selecitivity of the human retinoic acid nuclear receptor. *J. Mol. Biol.* 302: 155-170

Klimisch H.-J. *et al.* (1997) A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data. *Regul. Toxicol. Pharmcol.* 25: 1-5

KNIME (2013) Newsletter, September 2013. http://www.knime.com/files/knime newsletter_vol3no3_2013.pdf (accessed 08.08.2014)

KNIME (2014) Version 2.7. http://www.knime.org/ (accessed 13.01.2014)

KNIME (2015) Version 2.10. http://www.knime.org/ (accessed 27.01.2015)

Kolšek K. *et al.* (2014) Endocrine disruptome - an open source prediction tool for assessing endocrine disruption potential through nuclear receptor binding. *J. Chem. Inf. Model.* 54: 1254-1267

Könemann H. (1981) Quantitative structure-activity relationships in fish toxicity studies. Part 1: relationship for 50 industrial pollutants. *Toxicology* 19: 209-221

Lewis R.W. *et al.* (2002) Recognition of adverse and nonadverse effects in toxicity studies. *Toxicol. Pathol.* 30: 66–74

Leyden J.J. *et al.* (2005) Topical retinoids in inflammatory acne: a retrospective, investigator-blinded, vehicle-controlled, photographic assessment. *Clin. Ther.* 27: 216-224

Lin Z. *et al.* (2005) A simple hydrophobicity-based approach to predict the toxicity of unknown organic micropollutant mixtures in marine water. *Mar. Pollut. Bull.* 50 (2005), 617-623

Lipinski C.A. *et al.* (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* 46: 3-26

Lipnick R.L. *et al.* (1987) A QSAR study of the acute toxicity of some industrial organic chemicals to goldfish. Narcosis, electrophile and proelectrophile mechanisms. *Xenobiotica* 17: 1011-1025

Liu Z. *et al.* (2014a) Complexity of the RAR-mediated transcriptional regulatory programs. *Subcell. Biochem.* 70: 203-225

Liu X. *et al.* (2014b) Activation of farnesoid X receptor (FXR) protects against fructose-induced liver steatosis via inflammatory inhibition and ADRP reduction. *Biochem. Biophys. Res. Commun.* 450: 117-123

Livingstone D.J. (2004) Building QSAR models: a practical guide. In: Cronin M.T.D. and Livingstone D.J. (editors). Predicting chemical toxicity and environmental fate. CRC Press, Boca Raton, Florida, USA, pp. 161-164

Lush (2015) http://www.lushprize.org/ (accessed 19.03.2015)

Madden J.C. *et al.* (2012) Strategies for the optimisation of in vivo experiments in accordance with the 3Rs philosophy. *Regul. Toxicol. Pharmacol.* 63: 140-154

Maglich J.M. *et al.* (2001) Comparison of complete nuclear receptor sets from the human, *Caenorhabditis elegans* and *Drosophila* genomes. *Genome Biol.* 2: research0029.1-0029.7

Magnusson B.M *et al.* (2004) Molecular size as the main determinant of solute maximum flux across the skin. *J. Invest. Dermatol.* 122: 993-999

Marks R. *et al.* (1985) The effects of a shampoo containing zinc pyrithione on the control of dandruff. *Br. J. Dermatol.* 112: 415-22

Martin T.M. *et al.* (2015) Comparison of global and mode of action-based models for aquatic toxicity. *SAR QSAR Environ. Res.* 26: 245-262

McFeters G.A. *et al.* (1983) A comparison of microbial bioassays for the detection of aquatic toxicants. *Water Res.* 17: 1757-1762

McLean M. and Dutton M.F. (1995) Cellular interactions and metabolism of aflatoxin: an update. *Pharmac. Ther.* 65: 163-192

Mellor C.L. *et al.* (2015) The identification of nuclear receptors associated with hepatic steatosis to develop and extend Adverse Outcome Pathways. *Crit. Rev. Toxicol*. (early online: 1-15)

Mennear J.H. *et al.* (1982) Studies on the carcinogenicity of pentachloroethane in rats and mice. *Fund. Appl. Toxicol.* 2: 82-87

Mennecozzi M. *et al.* (2012) Hepatotoxicity screening taking a mode-of-action approach using HepaRG cells and HCA. *Altex Proc.* 1: 193-204

Merrill R.A. (2001) Regulatory toxicology. In: Klaassen C.D. (editor). Casarett and Doull's toxicology - the basic science of poisons (6th edition). McGraw-Hill, New York, USA, pp. 759-767

Minitab (2013) Version 16.2.2.0. http://www.minitab.com

Minucci S. *et al.* (1997) Retinoid X receptor (RXR) within the RXR-retinoic acid receptor heterodimer binds its ligand and enhances retinoid-dependent gene expression. *Mol. Cell. Biol.* 17: 644-655

Mitragotri S. *et al.* (2011) Mathematical models of skin permeability: an overview. *Int. J. Pharm.* 418: 115-129

MOE (2013) Version 2011.10, http://www.chemcomp.com/MOE-Molecular_Modeling_and_Simulations.htm (accessed 27.09.2013)

Molecular Graphics Laboratory (2014) http://www.autodock.scripps.edu (accessed 07.07.2014)

Morgan J.P. and Penovich P. (1978) Jamaica ginger paralysis. Forty-seven-year follow-up. *Arch. Neurol.* 35: 530-532

Mortimer M. *et al.* (2008) High throughput kinetic *Vibrio fischeri* bioluminescence inhibition assay for study of toxic effects of nanoparticles. *Toxicol. In Vitro* 22: 1412-1417

Moss G.P. and Cronin M.T.D. (2002) Quantitative structure-permeability relationships for percutaneous absorption: re-analysis of steroid data. *Int. J. Pharm.* 238: 105-109

Moya M. *et al.* (2010) Enhanced steatosis by nuclear receptor ligands: a study in cultured human hepatocytes and hepatoma cells with a characterized nuclear receptor expression profile. *Chem. Biol. Interact.* 184: 376-387

Munro I.C. *et al.* (1996) Correlation of structural class with No-Observed-Effect Levels: a proposal for establishing a threshold of concern. *Fd. Chem. Toxicol.* 34: 829-867

Nacci D. *et al.* (1986) Comparative evaluation of three rapid marine toxicity tests: sea urchin early embryo growth test, sea urchin sperm cell toxicity test and Microtox. *Environ. Toxicol. Chem.* 5: 521-525

Nelms M.D. *et al.* (2015) Proposal of an *in silico* profiler for categorisation of repeat dose toxicity of hair dyes. *Arch. Toxicol.* 89: 733-741

Nendza M. *et al.* (2010) Data quality assessment for *in silico* methods: A survey of approaches and needs. In: Cronin M.T.D. and Madden J.C. (editors). *In silico* toxicology: Principles and applications (1st edition). RSC Publishing, Cambridge, UK, pp. 59-69

Nettles K.W. *et al.* (2007) Structure plasticity in the oestrogen receptor ligand-binding domain. *EMBO reports* 8: 563-568

Newby D. *et al.* (2015) Decision trees to characterise the roles of permeability and solubility on the prediction of oral absorption. *Eur. J. Med. Chem.* 90: 751-765

NIH (2014) ChemIDplus, US National Institutes of Health. http://chem.sis.nlm.nih.gov/chemidplus (accessed 01.09.2014)

NIH (2015) TOXNET, US National Institutes of Health. http://toxnet.nlm.nih.gov/ (accessed 20.02.2015)

Norlen H. *et al.* (2014) A tutorial for analysing the cost-effectiveness of alternative methods for assessing chemical toxicity: the case of acute oral toxicity prediction. *Altern. Lab. Anim.* 42: 115-127

OECD (2003) Organisation for Economic Co-operation and Development, description of selected key generic terms used in chemical hazard/risk assessment. http://www.oecd-ilibrary.org/environment/descriptions-of-selected-key-generic-terms-used-in-chemical-hazard-risk-assessment_9789264079120-en (accessed 17.05.2013)

OECD (2012a) Organisation for Economic Co-operation and Development, guideline for the testing of chemicals - acute eye irritation/corrosion 405. http://iccvam.niehs.nih.gov/SuppDocs/FedDocs/OECD/OECD-TG405-2012-508.pdf (accessed 06.06.2015)

OECD (2012b) Organisation for Economic Co-operation and Development, collection of working definitions, appendix 1. http://www.oecd.org/chemicalsafety/testing/49963576.pdf (accessed 03.08.2014)

OECD (2014) Organisation for Economic Co-operation and Development, guidance on grouping of chemicals, 2nd edition. http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono%282014%294&doclanguage=en (accessed 03.08.2014)

OECD (2015) Organisation for Economic Co-operation and Development, eChemPortal. http://www.echemportal.org/ (accessed 20.02.2015)

Ognichenko L.N. *et al.* (2012) QSPR prediction of lipophilicity for organic compounds using random forest technique on the basis of simplex representation of molecular structure. *Mol. Inform.* 31: 273-280

OpenBabel (2013) Version 2.3.2. http://openbabel.org/wiki/Main_Page (accessed 16.04.2013)

Palczewska A. *et al.* (2013) Interpreting random forest models using a feature contribution method. Information Reuse and Integration (IRI), IEEE 14th International Conference: 112-119

Parvez S. *et al.* (2008) Toxicity assessment of organic pollutants: reliability of bioluminescence inhibition assay and univariate QSAR models using freshly prepared *Vibrio fischeri*. *Toxicol. Vitro* 22: 1806-1813

Patel V. and Sanyal A.J. (2013) Drug-induced steatohepatitis. *Clin. Liver Dis.* 17: 533-546

Patlewicz G. *et al.* (2013) Use of category approaches, read-across and (Q)SAR: General considerations. *Regul. Toxicol. Pharmacol.* 67: 1-12

Patlewicz G. *et al.* (2014) Read-across approaches - misconceptions, promises and challenges ahead. *Altex* 31: 387-396

PDB (2014) Protein Data Bank. http://www.rcsb.org/pdb/home/home.do (accessed 07.07.2014)

PDB (2015) Protein Data Bank. http://www.rcsb.org/pdb/home/home.do (accessed 13.8.2015)

PDSP (2015) Psychoactive Drug Screening Program. http://pdsp.med.unc.edu/pdsp.php (accessed 11.02.2015)

Pérez E. *et al.* (2012) Modulation of RXR function through ligand design. *Biochim. Biophys. Acta* 1821: 57-69

Péry A.R.R. *et al.* (2013) Perspectives for integrating human and environmental risk assessment and synergies with socio-economic analysis. *Sci. Total Environ.* 456-457: 307-316

PETA (2015) People for the Ethical Treatment of Animals. http://www.peta.org.uk/ features/new-cosmetics-law/ (accessed 19.03.2015)

Potts R.O. and Guy R.H. (1992) Predicting skin permeability. *Pharm. Res.* 9 (1992), 663-669

Przybylak K.R. and Cronin M.T.D. (2012) *In silico* models for drug-induced liver injury – current status. *Expert Opin. Drug Metab. Toxicol.* 8: 201–217

Przybylak K.R. *et al.* (2012) Assessing toxicological data quality: Basic principles, existing schemes and current limitations. *SAR QSAR Environ. Res.* 23: 435-459

Pugh W.J. *et al.* (2000) Epidermal permeability-penetrant structure relationships: 4, QSAR of permeant diffusion across human *stratum corneum* in terms of molecular weight, H-bonding and electronic charge. *Int. J. Pharm.* 197: 203-211

PyMOL (2014) Version 1.3. http://www.pymol.org (accessed 07.07.2014)

R (2014) The R Project, R Studio 0.98.501.19. http://www.r-project.org/ (accessed 20.08.2014)

Rang H.P. *et al.* (2007a) Rang and Dales's pharmacology (6th edition). Churchill Livingstone, Edinburgh, UK, pp. 98-112

Rang H.P. *et al.* (2007b) Rang and Dales's pharmacology (6th edition). Churchill Livingstone, Edinburgh, UK, pp. 781-786

Ravenzwaay B. van and Leibold E. (2004) A comparison between *in vitro* rat and human and *in vivo* rat skin absorption studies. *Hum. Exp. Toxicol.* 23: 421-430

Ribo J.M. and Kaiser K.L.E. (1984) Toxicities of aniline derivatives to *Photobacterium phosphoreum* and their correlations with effects to other organisms and structural parameters. In: Kaiser K.L.E. (editor). QSAR in environmental toxicology (1st edition). D. Reidel Publishing Company, Dordrecht, Netherlands, pp. 319-336

Richarz A.-N. *et al.* (2013) *In silico* workflows for toxicity prediction implemented into KNIME. Abstract of the 49th congress of the European Societies of Toxicology (EUROTOX), *Toxicol. Lett.* 221, pp. 81

Richarz A.-N. *et al.* (2014) COSMOS: integrated *in silico* models for the prediction of human repeated dose toxicity of COSMetics to Optimise Safety. In: Gocht T. and Schwarz M. (editors). SEURAT-1: towards the replacement of *in vivo* repeated dose systemic toxicity testing, Vol. 4. European Commission for Research & Innovation, Imprimerie Mouzet, Paris, France, pp. 186-209

Richet C.M. (1893) Note sur le rapport entre la toxicité et les propriétés physiques des corps. *Compt. Rend. Soc. Biol.* 45: 775-776

Roberts D.W. and Williams D.L. (1982) The derivation of quantitative correlations between skin sensitisation and physico-chemical parameters for alkylating agents and their application to experimental data for sultones. *J. Theor. Biol.* 99: 807-825

Roberts D.W. (2015) Estimating skin sensitization potency from a single dose LLNA. *Regul. Toxicol. Pharm.* 71: 437-443

Rowe P.H. (2007) Essential statistics for the pharmaceutical sciences. John Wiley & Sons, Chichester, UK, pp. 13-20

RSC (2014) Royal Society of Chemistry, ChemSpider. http//:www.chemspider.com/ (accessed 08.08.2014)

Russell W.M.S. and Burch R.L. (1959) The Principles of humane experimental technique. Methuen, London, UK, pp. 1-238

Ruusmann V. and Maran U. (2013) From data point timelines to a well curated data set, data mining of experimental data and chemical structure data from scientific articles, problems and possible solutions. *J. Comput. Aided Mol. Des.* 27: 583-603

Sahini N. and Borlak J. (2014) Recent insights into the molecular pathophysiology of lipid droplet formation in hepatocytes. *Prog. Lipid Res.* 54: 86-112

Samaras E.G. *et al.* (2012) The effect of formulations and experimental conditions on *in vitro* human skin permeation - data from updated EDETOX database. *Int. J. Pharm.* 434: 280-291

Saubern S. *et al.* (2011) KNIME Workflow to assess PAINS filters in SMARTS format. Comparison of RDKit and Indigo cheminformatics libraries. *Mol. Inform.* 30: 847-850

SCCS, 2012. The SCCS's notes of guidance for the testing of cosmetic substances and their safety evaluation - 8th revision (SCCS/1501/12), http//:ec.europa.eu/health/ scientific_committees/consumer_safety/docs/sccs_s_006.pdf (accessed 27.07.2015)

Schattenberg J.M. and Schuppan D. (2011) Nonalcoholic steatohepatitis: the therapeutic challenge of a global epidemic. *Curr. Opin. Lipidol.* 22: 479-488

Scheuplein R.J. and Blank I.H. (1971) Permeability of the skin. *Physiol. Rev.* 51: 702-747

Schiewe M.H. *et al.* (1985) Use of a bacterial bioluminescence assay to assess toxicity of contaminated marine sediments. *Can. J. Fish Aquat. Sci.* 42: 1244-1248

Schinke C. *et al.* (2010) Design and synthesis of novel derivatives of *all-trans* retinoic acid demonstrate the combined importance of acid moiety and conjugated double bonds in its binding to PML-RAR-α oncogene in acute promyelocytic leukemia. *Leuk. Lymphoma* 51: 1108-1114

Schultz T.W. *et al.* (1997) Identification of mechanisms of toxic action of phenols to *Tetrahymena pyriformis* from molecular descriptors. In: Chen F. and Schüürmann G. (editors). Quantitative structure-activity relationships in environmental sciences - VII (1st edition). SETAC Press, Pensacola, USA, pp. 329.342

Schultz T.W. *et al.* (2000) Role of LXRs in control of lipogenesis. *Genes Dev.* 14: 2831-2838

Schultz T.W. *et al.* (2015) A strategy for structuring and reporting read-across prediction of toxicity. *Regul. Toxicol. Pharmacol.* 72: 586-601

SEURAT-1 (2015) Safety Evaluation Ultimately Replacing Animal Testing. http//:www. seurat-1.eu/ (accessed 20.03.2015)

Shalita A.R. (1988) Lipid and teratogenic effects of retinoids. *J. Am. Acad. Dermatol.* 19: 197-198

Sonoda J. *et al.* (2008) Nuclear receptors: decoding metabolic disease. *FEBS Lett.* 582: 2-9

Speece R. (1987) Drexel University, Pittsburgh, PA, private communication (refer to Kaiser and Palabrica, 1991)

Steinmetz F.P. *et al.* (2014) Methods for assigning confidence to toxicity data with multiple values - identifying experimental outliers. *Sci. Total Environ.* 482-483: 358-365

Steinmetz F.P. *et al.* (2015a) Screening chemicals for receptor-mediated toxicological and pharmacological endpoints: using public data to build screening tools within a KNIME workflow. *Mol. Inf.* 34: 171-178

Steinmetz F.P. *et al.* (2015b) Data quality in the human and environmental health sciences: using statistical confidence scoring to improve QSAR/QSPR modeling. *J. Chem. Inf. Model.* 55: 1739-1746

Surapaneni P. *et al.* (2011) Valvular heart disease with the use of fenfluramine-phentermine. *Tex. Heart. Inst. J.* 38: 581-583

Sushko I. *et al.* (2012) ToxAlerts: a web server of structural alerts for toxic chemicals and compounds with potential adverse reactions. *J. Chem. Inf. Model.* 52: 2310-2316

Sütterlin H. *et al.* (2008) The toxicity of the quaternary ammonium compound benzalkonium chloride alone and in mixtures with other anionic compounds to bacteria in test systems with *Vibrio fischeri* and *Pseudomonas putida*. *Ecotox. Environ. Saf.* 71: 498-505

ten Berge W. (2014) Homepage of Wil ten Berge. http://home.wxs.nl/~wtberge/skin perm2013a.zip (accessed 01.03.2014)

Thumm W. *et al.* (1992) Toxicity tests with luminescent photobacterium and quantitative structure activity relationships for nitroparaffins. *Chemosphere* 24: 1835-1843

Tollefsen K.E. *et al.* (2014) Applying Adverse Outcome Pathways (AOPs) to support Integrated Approaches to Testing and Assessment (IATA). *Regul. Toxicol. Pharmacol.* 70: 629-640

Tralau T. *et al.* (2015) Regulatory toxicology in the twenty-first century: challenges, perspectives and possible solutions. *Arch. Toxicol.* 89: 823-850

van Wezel A.P. and Opperhuizen H. (1995) Narcosis due to environmental pollutants in aquatic organisms: residue-based toxicity, mechanisms, and membrane burdens. *Crit. Rev. Toxicol.* 25: 255-279

Vaz B. and de Lera Á.R. (2012) Advances in drug design with RXR modulators. *Expert Opin. Drug Discov.* 7: 1003-1016

Verhaar H.J.M. *et al.* (1992) Classifying environmental pollutants. 1: structure-activity relationships for prediction of aquatic toxicity. *Chemosphere* 25: 471-491

Verhaar H.J.M. *et al.* (1996) Classifying environmental pollutants 2: Separation of class 1 (baseline toxicity) and class 2 (polar narcosis) based on chemical descriptors. *J. Chemometr.* 10: 149-162

Verhaar H.J.M. *et al.* (2000) Classifying environmental pollutants: Part 3. External validation of the classification system. *Chemosphere* 40: 875-883

Vighi M. *et al.* (2009) Toxicity on the luminescent bacterium *Vibrio fischeri* (Beijerinck). I: QSAR equation for narcotics and polar narcotics. *Ecotox. Environ. Saf.* 72: 154-161

Vinardell M.P. (2015) The use of non-animal alternatives in the safety evaluations of cosmetics ingredients by the Scientific Committee on Consumer Safety (SCCS). *Regul. Toxicol. Pharmacol.* 71: 198-204

Vinken M. *et al.* (2013) Development of an adverse outcome pathway from drug-mediated bile salt export pump inhibition to cholestatic liver injury. *Toxicol. Sci.* 136: 97-106

Vinken M. (2013) The adverse outcome pathway concept: a pragmatic tool in toxicology. *Toxicology* 312: 158-165

Vinken M. (2015) Adverse Outcome Pathways and Drug-Induced Liver Injury testing. *Chem. Res. Toxicol.* 28: 1391-1397

Wang H. and LeCluyse E.L. (2003) Role of Orphan nuclear receptors in the regulation of drug metabolising enzymes. *Clin. Pharmacokinet.* 42: 1331-1357

Wang N. *et al.* (2002) Constitutive activation of peroxisome proliferator-activated receptor-gamma suppresses pro-inflammatory adhesion molecules in human vascular endothelial cells. *J. Biol. Chem.* 277: 34176-34181

Wehner F.C. *et al.* (1978) Mutagenicity to *Salmonella typhimurium* of some *Aspergillus* and *Penicillium* mycotoxins. *Mutat. Res.* 58 (1978), 193-203

Wenlock M.C. and Carlsson L.A. (2015) How experimental errors influence drug metabolism and pharmacokinetic QSAR/QSPR models. *J. Chem. Inf. Model.* 55: 125-134

Wu Y. and Zhang Y. (2013) Analytical chemistry, toxicology, epidemiology and health impact assessment of melamine in infant formula: recent progress and developments. *Fd. Chem. Toxicol.* 56: 325-335

Yamada T. *et al.* (2013) A category approach to predicting the repeated-dose hepatotoxicity of allyl esters. *Regul. Toxicol. Pharmacol.* 65: 189-195

Yang L. *et al.* (2013) Towards a fuzzy expert system on toxicological data quality assessment. *Mol. Inform.* 32: 65-78

Yates I.E. and Porter J.K. (1982) Bacterial bioluminescence as a bioassy for mycotoxins. *Appl. Environ. Microbiol.* 44: 1072-1075

Zhao Y.H. *et al.* (1998a) QSAR study of the toxicity of benzoic acids to *Vibrio fischeri*, *Daphnia magna* and carp. *Sci. Total Environ.* 216: 205-215

Zhao Y.H. *et al.* (1998b) Quantitative structure-activity relationships of chemicals acting by non-polar narcosis - theoretical considerations. *Quant. Struct.-Act. Relat.* 17: 131-138

Zhu H. *et al.* (2014) Big data in chemical toxicity research: the use of High-Throughput Screening assays to identify potential toxicants. *Chem. Res. Toxicol.* 27: 1643-1651

Ziehl T.A. and A. Schmitt (2000) Sediment quality assessment of flowing waters in South-West Germany using acute and chronic bioassays. *Aquat. Ecosys. Health Manage.* 3: 347-357

10.Appendices

The appendices contain data tables (A), rules and SMARTS patterns (B), codes and workflows (C) and published works (D). Except the published works, *i.e.* relevant authored and co-authored articles and abstracts in appendix D, the appendix is not paper-based. However, all appendices can be found in the attached USB card.

A Data tables

All datasets used in these studies are available on the USB card attached to this thesis.

A.1 Microtox dataset (including exposure comparison)

(refer to attached USB card)

A.2 Microtox and skin permeability dataset (including statistical equations)

(refer to attached USB card)

A.3 Dermal absorption dataset

(refer to attached USB card)

A.4 RAR ligand dataset

(refer to attached USB card)

A.5 NR ligand dataset (including RAR/RXR, PPAR, LXR and FXR)

(refer to attached USB card)

B Rules and SMARTS patterns

All rules and SMARTS patterns used in these studies are available electronicallly.

B.1 Hepatotoxicity structural alerts “Hewitt & Steinmetz”

(refer to attached USB card)

B.2 Rules for RAR ligand screening

(refer to attached USB card)

B.3 Rules for NR ligand screening (RAR/RXR, PPAR, LXR and FXR)

(refer to attached USB card)

C Codes and workflows

All codes and workflows used in these studies are available electronically.

C.1 KNIME workflow for RAR ligand screening

(refer to attached USB card)

C.2 R-code for RMSE$_{CS}$ (10-fold crossvalidation for CS-weighted multivariate linear regression)

(refer to attached USB card)

D Published Works

All relevant articles and conference abstracts published within this PhD as main author (or co-author with significant contribution) are presented on the following pages.

D.1 Steinmetz *et al.* Sci. Total Environ. 482 (2014), 358-365

D.2 Steinmetz *et al.* Mol. Inform. 34 (2015), 171-178

D.3 Steinmetz *et al.* J. Chem. Inf. Model. 55 (2015), 1739-1746

D.4 Mellor *et al.* Toxicology Letters 229 (2014), 162

D.5 Cronin *et al.* Altex Proceedings 1/14 (2014), 69

D.6 Cronin *et al.* Toxicology Letters S238 (2015), 166

D.7 Steinmetz *et al.* Toxicology Letters S238 (2015), 166

D.8 Richarz *et al.* Toxicology Letters S238 (2015), 166

D.9 Richarz *et al.* Toxicology Letters S238 (2015), 170

D.10 Tsakovska *et al.* Toxicology Letters S238 (2015), 173

D.11 Mellor *et al.* (2015) Crit. Rev. Toxicol. (early online: 1-15)