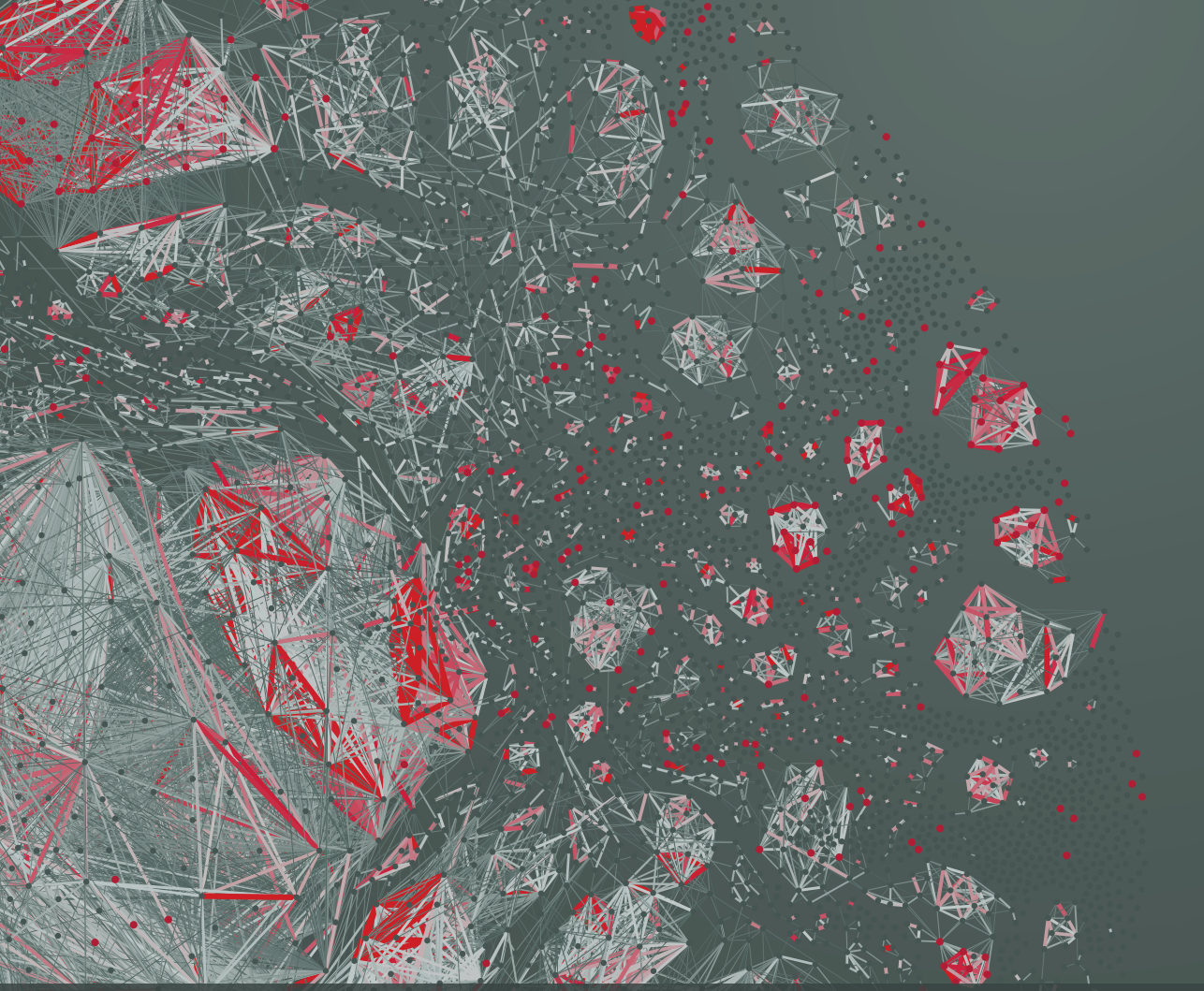
Chemical similarity: structuring risk and hazard assessment

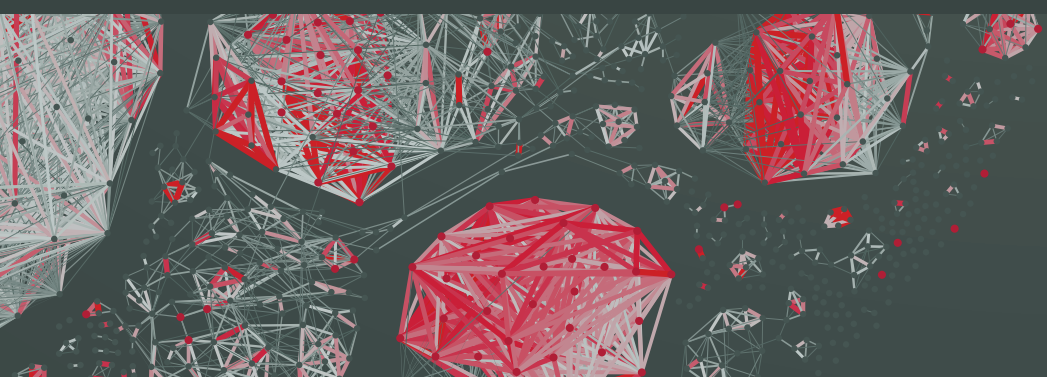
Wassenaar, P.N.H.



CHEMICAL SIMILARITY

STRUCTURING RISK AND HAZARD ASSESSMENT

Pim N.H. Wassenaar



Chemical Similarity

Structuring Risk and Hazard Assessment

Pim N.H. Wassenaar

Chemical Similarity: Structuring Risk and Hazard Assessment

PhD thesis, Leiden University, The Netherlands

The research as described in this thesis was conducted at the Institute of Environmental Sciences (CML), Leiden University, The Netherlands.

ISBN: 978-90-5191-700-0
Cover: Pim N.H. Wassenaar
Design and layout: Pim N.H. Wassenaar
Printed by: ProefschriftMaken

Chemical Similarity

Structuring Risk and Hazard Assessment

Proefschrift

ter verkrijging van

de graad van doctor aan de Universiteit Leiden,

op gezag van rector magnificus prof.dr.ir. H. Bijl,

volgens besluit van het college voor promoties

te verdedigen op dinsdag 19 april 2022

klokke 13.45 uur

door

Pim Nicolaas Hubertus Wassenaar

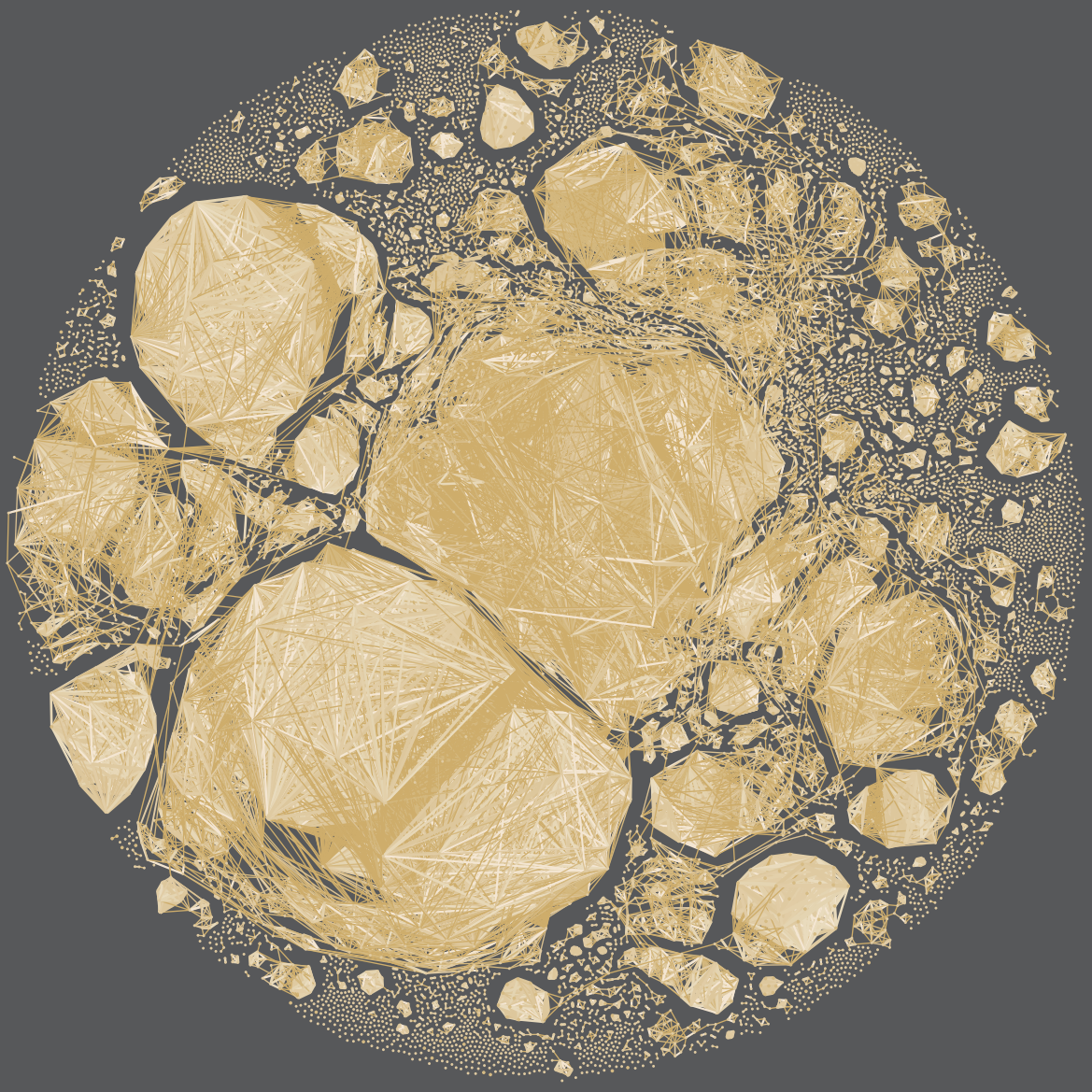
geboren te Beverwijk

in 1993

Promotores:
Prof.dr.ir. W.J.G.M. Peijnenburg
Prof.dr.ing. M.G. Vijver

Promotiecommissie:
Prof.dr. A. Tukker
Prof.dr.ir. P.M. van Bodegom
Prof.dr. S. Suh
Dr. J.N. van Rijn
Prof.dr. E. Benfenati (Mario Negri Institute for Pharmacological Research)
Dr. E. Papa (University of Insubria)

Table of contents



1

General Introduction

1.1 Chemical universe

We are surrounded [...]. They are in front of us, behind us, and we are flanked on both sides […].

Two quotes by Lewis Burwell Puller, who was a US marine from 1918-1955, inspired me [1]. Although not thinking of military action, but talking about chemicals. Chemicals are everywhere. We see, feel, inhale, drink and eat them, and there is no way to avoid them all.

Chemicals are involved in most of our activities during daily life and fulfil a fundamental role in our society. Since the 19th century, the chemical industry evolved rapidly, including the development of synthetic fertilizers, plastics, dyes, surfactants and pharmaceuticals [2]. These advancements significantly influenced and formed modern society as we know it.

The indispensability of chemicals is particularly evident from the large number of over 350,000 chemicals and chemical mixtures that are registered worldwide for production and use [3]. These chemicals, that represent the so-called chemical universe, can be divided in several categories based on various aspects including the type of substance, chemical structure, and environmental source or type of application (see Textbox A).

Textbox A: Characterizing the chemical universe

The chemical universe, as illustrated in Figure A.1, can be categorized based on various aspects including the type of substance, chemical structure and environmental source or type of application.

Type of substance
In general, three types of substances can be identified [8]. The first two types are the mono- and multi-constituent substances. Like the names suggest, both types contain one or several main constituents (i.e. structures/components) and potentially some impurities that make up the composition. These two categories are also described as well-defined substances, as the composition is (or can be) well characterized. The third type concerns UVCB substances, which stands for substance of Unknown or Variable composition, Complex reaction products or Biological materials. These substances contain many different constituents of which some can be (partially) unknown and/or the exact composition can be variable or difficult to predict. Besides these substance types, other categories can be defined, like polymers and fibers (including plastics and nanomaterials).

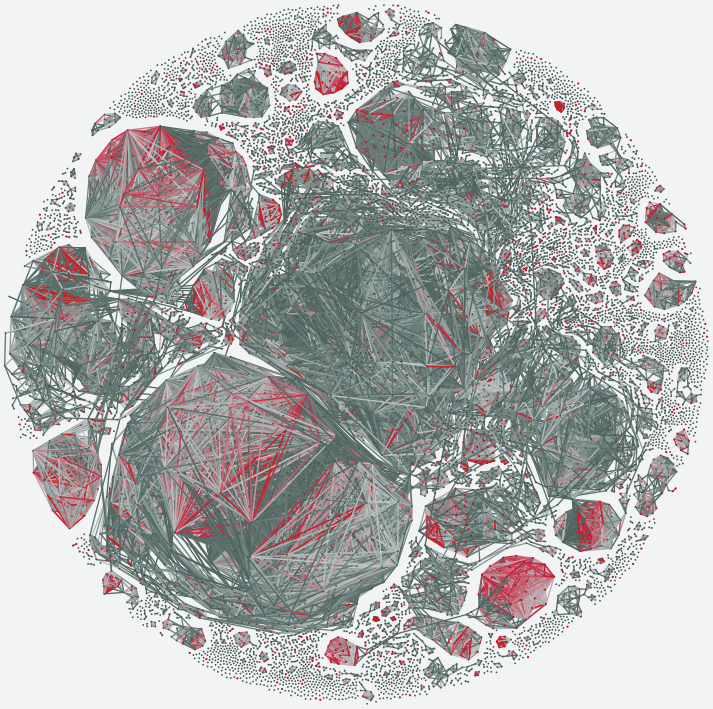


Figure A.1. Illustration of (a part of) the chemical universe in a chemical similarity network (generated with Gephi-v0.9.2, ForceAtlas-2 [4,5]). Within this figure, dots/nodes represent chemical structures and lines/edges represent chemical similarity between two connected nodes. Edges are thicker and colored in red between nodes with a higher structural similarity. Chemical similarity is calculated using the CDK Extended fingerprint from PaDEL-Descriptor [6] and the Jaccard-Tanimoto coefficient [7], with a similarity cut-off value of 0.7. Red colored nodes represent substances of very high concern (see section 1.4 for more details) and gray colored nodes represent REACH registered substances (see Textbox B for more details).

Chemical structure

A wide variety in chemical structures exist, that originate from the available chemical elements and fragments. These elements and fragments can be considered as the building blocks for bigger, more complex fragments/chemicals that are (or theoretically can be) present in the chemical universe. The structures within the chemical universe can be grouped based on these building blocks at several levels. Generally, chemicals can be divided in organic and inorganic compounds in a first step (i.e. chemicals that contain at least one or no carbon atom, respectively). In subsequent steps, chemicals can be divided based on other aspects, including general atom-types (e.g. hydrocarbons, organometallic compounds and organohalogen compounds), more specific structural features (e.g. saturated, unsaturated and aromatic hydrocarbons) and similar core structures (i.e. scaffolds/parent structures) (e.g. alkanes or phenanthrenes and derivatives) [9].

Environmental source and type of application

The chemical universe can also be divided based on the source of the chemicals. Chemicals can be divided based on general sources, like chemicals that are derived from a biological source (e.g. derived from plant or animal species) versus a non-biological source (e.g. chemical or mineral sources), or can be based on the production process (e.g. synthesized versus refined chemicals) [ECHA, 2017a]. In addition, chemicals can be categorized based on more specified sources or types of application. Some broad categories that can be defined include agricultural chemicals (e.g. biocides and pesticides), chemicals in consumer products (e.g. chemicals in toys, electronics, household supplies, cosmetics and food contact materials), industrial chemicals (e.g. lubricants and chemicals used in building and construction), pharmaceuticals and chemicals used in health care, and intermediates (e.g. human metabolites or industrial intermediates) [10]. As several chemicals can have multiple use patterns, they can belong to multiple categories.

Substance identifiers

Several substance or chemical identifiers are in use to distinguish the substances in the chemical universe. Substance identifiers can be categorized in systematic chemical names (e.g. International Union of Pure and Applied Chemistry (IUPAC)-name), registration numbers (e.g. Chemical Abstracts Service (CAS)-number or European Community (EC)-number) and structural identifiers (e.g. Simplified Molecular Input Line Entry System (SMILES) or IUPAC International Chemical Identifier (InChi)) [3,11]. Some examples of these substance identifiers are provided in Table A.1. Each (type of) identifier has its own 'pros and cons' (with its own specific ambiguities). For instance, systematic names provide information on the structure and/or source of the chemical, but are relatively more susceptible to typographical errors or variation. Registration numbers are easily verifiable, but do not have a direct physical meaning. Structural identifiers on the other hand, provide a molecular representation but cannot (easily) be applied to substances that contain multiple constituents (e.g. UVCBs).

Table A.1. Examples of several substance identifiers for two substances, a mono-constituent and a UVCB substance.

Substance identifier	**Substance A (mono-constituent substance)**	**Substance B (UVCB)**
IUPAC name	1,4-diisopropylbenzene	Naphtha (petroleum), catalytic reformed
CAS-number	100-18-5	68955-35-1
EC-number	202-826-9	273-271-8
SMILES	CC(C)C1=CC=C(C=C1)C(C)C	No single structure can be assigned
InChi	InChI=1S/C12H18/c1-9(2)11-5-7-12(8-6-11)10(3)4/h5-10H,1-4H3	No single structure can be assigned

1.2 Safe and sustainable chemicals

Although the use of chemicals provides numerous benefits and our daily life is drowned with it, their application may harm human health and the environment. In the first place, chemicals may cause adverse effects on humans or the environment upon exposure or emission. The extent of adverse effects is related to the inherent chemical properties and the type and degree of exposure (this will be discussed in more detail in section 1.3). To manage the safety of chemicals, particular chemical legislations are in place in numerous countries and regions around the world. Within Europe the main chemical legislation is REACH (Regulation on Registration, Evaluation, Authorisation and Restriction of Chemicals; EC/1907/2006) [12], which is applicable to most chemicals that are produced and/or imported into Europe (see Textbox B for more details on the REACH legislation). In addition, for some specific applications, separate regulations are in place, including cosmetics, pesticides, biocides and food contact materials.

Secondly, the use and production of chemicals may also harm human health and the environment by the impact of their environmental footprint. The chemical industry is one of the most polluting, energy and resource-intensive sectors [13] and, from a lifecycle perspective, effects on climate change, resource use, ecosystems and biodiversity can particularly be expected [13].

In order to (further) minimize the adverse effects of chemicals and to protect human health and the environment in a holistic manner, the European Commission (EC) recently launched a chemical strategy for sustainability [13]. The general aim of this strategy is to stimulate a transition to more safe and sustainable chemicals and is part of the European Green Deal, which sets an overarching objective for the EU to become a sustainable climate neutral and circular economy by 2050. Within their strategy, the EC strives for a 'toxic-free environment' (sometimes also defined as a 'non-toxic environment'). Pragmatically, this means a substitution or minimization of the use of substances of concern as far as possible. For instance, by phasing-out the most harmful substances for non-essential societal uses, and (where possible) by replacing substances of concern with chemicals that are (inherently) safe and sustainably by design. However, at the moment, the exact interpretation and implementation of this concept needs to be further defined [14]. In addition, from a sustainability perspective, the EC strives for cleaner production and recycling processes in which the whole life cycle of substances, materials and products should be considered. This is an important precondition for reaching the EU ambition of a clean (toxic-free) circular economy.

These recent developments emphasize the importance and relevance of the transition to safe and sustainable chemicals. Within this thesis, my main focus is on the safety aspects of chemicals (i.e. toxicity) and is not focusing on sustainability aspects in particular.

Textbox B: REACH legislation

In order to facilitate safe use of chemicals, the REACH regulation came into force in the EU in 2007 (EC/1907/2006) [12], which also established an European Chemicals Agency (ECHA) which is responsible for managing this regulation. In principal, REACH applies to all chemical substances, but excludes radioactive substances, substances subject to customs supervision, non-isolated intermediates, the transport of dangerous substances and wastes (REACH article 2). Furthermore, some substances are exempted from key provisions of REACH, including polymers and substances that are already sufficiently regulated under other regulations (e.g. medicinal products, food products and pesticides; REACH article 2). Like the name suggests, the main provisions of REACH include registration, evaluation, authorization and restriction of substances on the European market, each of which are briefly discussed within this section.

Registration

Under the REACH regulation, manufacturers and importers of substances are responsible for the registration of substances that are produced or imported in the EU above one ton per year via a registration dossier. The registration is based on a 'one substance, one registration' principle, meaning that manufacturers and importers of the same substance should submit a joint registration. If a substance is not registered, it is not allowed on the European market. The registration dossiers include a technical dossier (for substances ≥ 1 ton/y) and a chemical safety report which contains a hazard assessment (for substances ≥ 10 ton/y). The hazard assessment should include the minimum standard information requirements to meet the registration obligations of REACH. These requirements are dependent on the quantity of the substance that is manufactured or imported and follows the 'no data, no market' principle (REACH Annex VII-X). When registration dossiers are submitted by manufacturers or importers, they are checked on technical completeness by ECHA whereupon the substance is allowed on the European market.

Evaluation

Although it is the responsibility of the registrant to ensure safe use of their chemicals, the registration dossiers can be evaluated by authorities in order to analyze whether the registration dossiers are in line with the REACH requirements and whether safe use is plausible. A selection of registration dossiers are evaluated for compliance with the legal requirements by ECHA with support from EU Member State authorities. Following such

a dossier evaluation, additional data could be requested from a registrant. In addition, Member States may perform so-called substance evaluations with the aim to request additional information from the registrant(s) of a substance to address a specific concern. Substance evaluations typically request information which is not available as part of the standard REACH information requirements but is needed to clarify a concern.

Authorization and restriction

When there are serious concerns with the use of a substance for human health or the environment, the substance can be regulated within REACH via two procedures: authorization and/or restriction.

The authorization process consists of two phases, a hazard identification phase targeted towards inclusion of a substance on the so-called candidate list for authorization and a prioritization towards inclusion on the authorization list. A substance is placed on the candidate list for authorization when it is identified as a substance of very high concern (SVHC) following a regulatory decision process. This process includes chemicals that are of very high concern due to their hazardous properties (such as carcinogenicity) as defined in REACH article 57 (see section 1.4 for more details). Periodically, substances on the candidate list are prioritized for inclusion on the authorization list (REACH Annex XIV). When a substance is included on the authorization list, its use is no longer allowed in the EU unless manufacturers and/or importers request an exemption (authorization) for a specific application. Authorizations are only accepted when the manufacturers and/or importers can prove that the substance can be used under adequate control of risks (in case of a threshold of toxicological concern) or when there are no alternatives and the socio-economic benefits of the use outweighs the risks. The authorization title stimulates the substitution to less hazardous alternative substances or processes.

In addition, if a substance poses an EU-wide unacceptable risk to human health or the environment, its manufacture, placing on the market and use can be restricted (REACH Annex XVII). Member States and the European Commission have the right to propose restrictions targeted at SVHCs or any other substances. Restrictions can be broad covering all uses of a substance or can target only some or a specific use. When a substance is restricted, industry has no possibility to ask for an exemption, in contrast to the process of authorization.

Other chemical regulations and directives

The REACH legislation is closely connected with the European regulation on the Classification, Labelling and Packaging of substances and mixtures (CLP; EC/1272/2008) [15]. This regulation requires manufacturers, importers or downstream users of substances

or mixtures to classify, label and package their hazardous chemicals appropriately before placing them on the market, and is probably best known for the hazard pictograms that can be found on the packaging of substances and mixtures (e.g. consumer formulations). In addition, there are many other legislations in Europe that regulate chemicals, like the Industrial Emissions Directive (which specifically focusses on industrial emissions instead of substances that are placed on the market), worker protection legislations and cosmetic products regulation.

1.3 Risk and hazard assessment

Generally, a risk assessment is conducted to evaluate whether a substance can be used in a safe way. This assessment can be divided in environmental and human risk assessments. Although there are many similarities between environmental and human risk assessment, there are also some key differences. Most importantly, environmental risk assessment deals with millions of species and aims to protect populations and ecosystems, whereas human risk assessment aims to protect all individuals of a certain species (i.e. humans). Despite these differences, all risk assessments principally consist of two main aspects, an exposure assessment and a hazard (or effect) assessment (Figure 1.1). Ultimately, the results of the exposure and hazard assessment are combined to characterize the risk [16].

Within the exposure assessment, the concentration of a chemical in an environmental compartment or the extent of exposure of a chemical to an organism of interest is determined. For environmental risk assessment this can be a measured or predicted concentration in water, sediment, soil or air, and is generally defined as the predicted environmental concentration (PEC) (e.g. in mg/L). For human risk assessment, various exposure routes are generally combined in order to determine a total daily intake (e.g. in mg/kg body weight/d). This can include direct environmental exposure via inhalation, soil ingestion and dermal contact, but can also include indirect environmental exposure via food products and drinking water [17]. In order to determine a total dose for humans, information on the chemical concentration as well as on the daily intake of these components is required. In addition, the exposure assessment might be refined by considering additional factors, like bioavailability (which depends on the route of uptake) or by also considering exposure at work or from consumer products [17].

The aim of a hazard assessment is to determine a reference level, which represents a concentration or dose below which no or very limited adverse effects are expected. There are many different types of reference levels and they are generally based on toxicity measures that are derived from concentration/dose-response curves of experimental studies [17]. Within environmental risk assessment procedures, this typically concerns No-Observed

Effect Concentrations (NOEC) or effect concentrations affecting 10% of the population (EC10). For human risk assessment, this may concern a No or Lowest-Observed Adverse Effect Level (NOAEL/LOAEL) or benchmark dose (BMD). Depending on the protection goal and regulatory framework, the most critical toxicity measures are converted to safe reference levels by applying assessment factors. These assessment factors account for the uncertainty when extrapolating experimental data to humans or ecosystems (e.g. to account for interspecies differences, intraspecies variation, and to extrapolate short-term to long-term exposures) [16]. Within environmental risk assessments, a compartment specific Predicted No Effect Concentration (PNEC) is used as reference level, which is often based on single species test data from varying trophic levels (e.g. algae, water fleas and/or fish for the aquatic compartment). Within human risk assessment, a Derived No Effect Level (DNEL) is frequently determined as reference level for threshold substances.

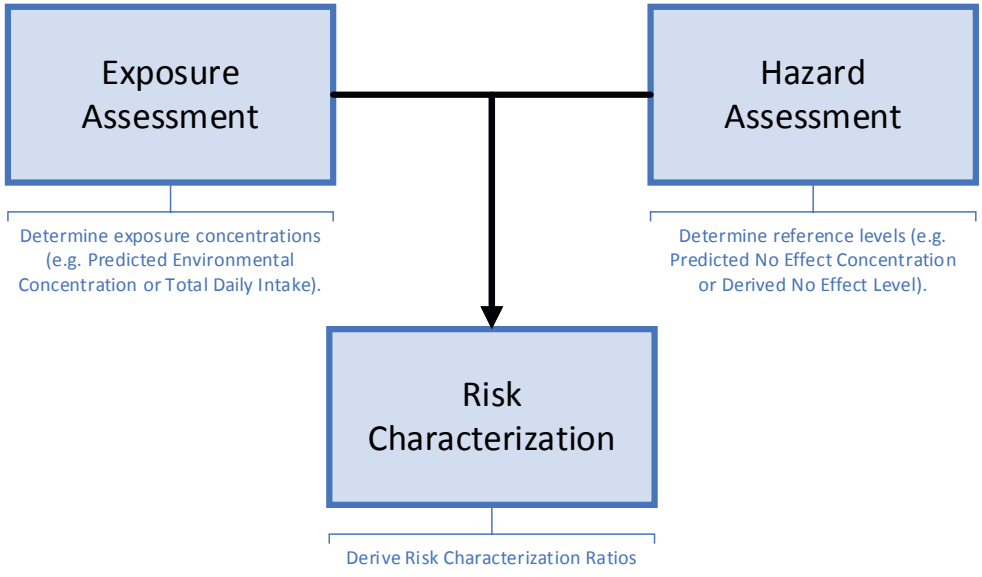


Figure 1.1. Overview of the main steps within risk assessment.

Subsequently, the results of the exposure assessment are compared with the results of the hazard assessment in order to derive a Risk Characterization Ratio (RCR) (Figure 1.1). The RCR provides an estimate of the likelihood that adverse effects occur due to actual or predicted exposure to a chemical [16]. For environmental risk assessment the PEC/PNEC ratio is generally applied and for human risk assessment the total daily intake/DNEL ratio is frequently used (for threshold substances). An RCR below 1 indicates safe use, whereas an RCR above 1 indicates a concern. When the assessment indicates a concern, this may require a refinement of the assessment (when many conservative assumptions have been made) or

may require the implementation of risk management measures to reduce the risks. There are several categories of risk management measures, including restrictions on the use/marketing of the substance, the implementation of technical measures to minimize emissions of the substance, or by implementing personal protection measures [16]. Within the risk assessment, RCRs have to be evaluated for all endpoints and for each protection goal (i.e. compartment/population) to guarantee safe use of a substance.

1.4 Substances of very high concern

Generally, regulatory measures are only taken when the risk assessment indicates a concern (RCR ≥ 1; see section 1.3). However, for some substances only the hazards are already sufficient to trigger further regulatory action. This particularly applies to substances with carcinogenic (C), mutagenic (M) or reprotoxic (R) properties, substances with persistent, bioaccumulative and toxic (PBT) or very persistent and very bioaccumulative (vPvB) properties, or substances with an equivalent level of concern, like endocrine disrupting (ED) substances. Substances with such hazard properties are considered of very high concern, as even the lowest amount of exposure may cause serious and often irreversible adverse effects.

Within REACH, substances that meet specific hazard-based criteria for these abovementioned endpoints can be identified as a Substance of Very High Concern (SVHC) following a regulatory decision process. The SVHC criteria are discussed in more detail in Textbox C. The ultimate aim of SVHC identification is to substitute the use of SVHCs by safer ('non-regrettable') alternatives (see also Textbox B).

Besides REACH, the SVHC criteria are also applied in other (national) chemical regulations, like the Dutch national policy on 'Zeer Zorgwekkende Stoffen' (ZZS; which is literally translated as substances of very high concern) [18]. This policy aims to prevent or minimize industrial emissions of ZZS. ZZS cover a broader range of chemicals than the SVHCs under REACH, but are identified on the same hazard criteria as SVHCs.

Textbox C: SVHC criteria

Substances with the following hazard properties can be identified as Substances of Very High Concern (SVHCs) according to the REACH legislation article 57:

- Substances with carcinogenic, mutagenic or reprotoxic (CMR) properties.
- Substances with persistent, bioaccumulative and toxic (PBT) or very persistent and very bioaccumulative (vPvB) properties.
- Substances with properties that cause an equivalent level of concern as CMR or PBT/vPvB substances.

CMR

CMR substances are of particular concern to human health. Carcinogenic chemicals have the potential to induce or increase the incidence of cancers, whereas mutagenic chemicals have the potential to cause heritable gene mutations (including heritable structural and numerical chromosome aberrations). Reprotoxic chemicals, on the other hand, can cause adverse effects on sexual function and fertility, and/or cause developmental toxicity in the offspring [15]. A substance can be identified as SVHC based on CMR properties, when it meets the criteria for classification as carcinogenic, mutagenic or reprotoxic category 1A or 1B in accordance with the CLP regulation. A category 1A classification can be provided when there are known effects on humans based on scientific evidence, and a category 1B classification can be provided when a substance has presumed effects on humans. In addition, a category 2 classification can be provided when a substance is suspected to cause such effects in humans, but for which the evidence is not sufficiently convincing. Category 2 classifications are insufficient to identify a substance as SVHC based on CMR properties.

PBT/vPvB

PBT/vPvB substances are generally of particular concern for the environment, but may also affect human health. Once emitted, PBT/vPvB substances cannot easily be removed from the environment by biotic and abiotic degradation processes (persistent), and are likely to reach high and potential toxic concentrations in organisms or humans upon continued emission (bioaccumulative). In principle, the effects of such accumulation are unpredictable in the long-term, because long-term exposures are expected that affect multiple life-cycles of species. In addition, exposure concentrations are unpredictable, but could be very high. Consequently, there is also a concern for vPvB substances, even when no toxicity is demonstrated in laboratory tests [19]. Substances can be identified as SVHC based on PBT or vPvB properties when they meet the REACH Annex XIII criteria. The PBT/vPvB criteria are shown in Table C.1.

Table C.1. PBT/vPvB criteria according to the REACH regulation (Annex XIII).

Criteria	PBT	vPvB
Persistence	In any of the following situations: - Degradation half-life in fresh or estuarine surface water is higher than 40 days. - Degradation half-life in marine surface water is higher than 60 days. - Degradation half-life in soil, freshwater sediment or estuarine water sediment is higher than 120 days. - Degradation half-life in marine sediment is higher than 180 days.	In any of the following situations: - Degradation half-life in marine, fresh or estuarine surface water is higher than 60 days. - Degradation half-life in soil, marine water sediment, freshwater sediment or estuarine water sediment is higher than 180 days.
Bioaccumulation	In the following situation: - BCF in aquatic species is higher than 2000.	In the following situation: - BCF in aquatic species is higher than 5000.
Toxicity	In any of the following situations: - Chronic NOEC or EC10 for marine or freshwater organisms is less than 0.01 mg/L. - The substance meets the criteria for classification as carcinogenic (cat. 1A or 1B), mutagenic (cat. 1A or 1B), or reprotoxic (cat. 1A, 1B or 2). - Other evidence of chronic toxicity (e.g. STOT RE cat. 1 or 2).	

Abbreviations: BCF – Bioconcentration factor; NOEC - No observed effect concentration; EC10 - Effect concentration affecting 10% of the population; STOT RE - Specific target organ toxicity after repeated exposure.

Equivalent level of concern

In addition, substances with properties that cause an equivalent level of concern as CMR or PBT/vPvB substances can be identified as SVHC on a case-by-case basis. There are no specific criteria for this category, because it is primarily set-up as a safety net for very hazardous substances that do not meet the standard SVHC criteria. Examples of hazard categories that might cause an equivalent level of concern could include endocrine disrupting (ED) substances, substances with specific target organ toxicity after repeated exposure (STOT-RE), substances that cause serious (respiratory) sensitization or substances with persistent, mobile and toxic (PMT) properties.

1.5 Challenges in current risk and hazard assessment

The European chemical regulations have been refined and improved significantly in the recent decades, particularly with the implementation of the REACH and CLP regulations (see also Textbox B). However, despite the progress that has been made, several challenges remain ahead of us in terms of regulating and ensuring safe use of chemicals [13]. Within this thesis I will specifically focus on two main challenges, including i) a lack of (reliable) data, and ii) a relative slow or inefficient evaluation/regulation process:

i) Despite the implementation of new regulations, there is a lack of (reliable) data for risk and hazard assessments for many individual substances. Due to the principles of REACH (Textbox B), no full toxicity profile is available for substances that are produced or imported at lower volume levels. Also, no toxicity data are available for many emission products, because they do not have a REACH registration obligation. In addition, not all data that are provided are of sufficient quality. For instance, two third of the REACH registration dossier appeared to be incompliant with the REACH data requirements [13]. As a consequence, it can be difficult to reliably determine the actual risks and hazards of substances based on the available data. Nevertheless, adequate risk and hazard assessments are essential prior to the use and release of chemicals in order to be able to ensure safe use. This is particularly evident for some (legacy) chemicals like polychlorinated biphenyls (PCBs), dichlorodiphenyltrichloroethane (DDT) and perfluorooctanesulfonic acid (PFOS) [20]. Exposure to these chemicals appeared to be of greater concern than previously anticipated. At the time that safety concerns are raised, widespread exposure has often already occurred, and typically the set of available toxicity data is inadequate to introduce risk management measures immediately. Consequently, chemicals of potential concern continue to be emitted, with a risk of significant effects on human health and the environment in the long-term. Accordingly, it is important to be able to signal potential concerns and improve the early identification and regulation of hazardous chemicals before widespread exposure occurs.

ii) The identification of substances of concern and the implementation of risk management measures is a time-consuming process and it can take several years before risk management measures are implemented for a single substance. This is particularly caused by relative slow procedures and a limited available evaluation capacity. As a consequence, only a limited number of substances can be evaluated at a time. Eventually, once a substance is regulated it is often observed that it is substituted by substances with comparable or unknown properties. Accordingly, this substance-by-substance evaluation approach can be considered as a not fully effective

and efficient process, and might be unsustainable to ensure safe use of chemicals in the long-term. Especially given the large number of (new) substances that are produced and imported each year. Therefore, it is important that new approaches are developed and implemented to improve the effectiveness and speed of chemical risk and hazard assessments. This is necessary to ensure that once chemical concerns are raised and confirmed, chemicals can also be quickly regulated accordingly.

1.6 Optimizing risk and hazard assessment

Recent developments in the field of (computational) toxicology provide opportunities to address the two main challenges as highlighted in section 1.5. Particularly, progress has been made in combining available toxicity data in (large) toxicological datasets [21]. This does not only include *in vivo* toxicity data, but also (high throughput) *in vitro* data and various omics-data (of which some can be referred to as 'big data'). At the moment, over 900 toxicity databases exist [22], including several databases that merge multiple data sources, like eChemPortal and PubChem. At the same time, fast progress has been made in computational toxicology, specifically including the developments in chemoinformatic toolkits (e.g. RDkit, CDK, ChemmineR, OpenBabel), but also for instance the availability of machine learning methods [23–25].

These advancements particularly provide opportunities to develop new *in silico* models to improve early signaling of chemicals of potential concern. The specific benefits of *in silico* models is that they provide fast (and inexpensive) predictions of chemical toxicity, and only require limited input information that is generally widely available (e.g. chemical structure or CAS-number). Therefore, such models could particularly be used to screen and prioritize substances for further evaluation, to ensure that available evaluation capacity is invested in substances that matter most (like suspected SVHC substances). By using additional early-warning triggers, there is an increased chance to identify and regulate substances of concern before wide-spread exposure has occurred.

In addition, to further increase the efficiency of risk and hazard assessments, there is a need to shift from substance-by-substance assessment to group assessment approaches. Within group assessment approaches similar substances (called chemical analogues) are evaluated together, with the underlying assumption that similar substances have similar properties and effects, and therefore also similar concerns. In order to define a category of substances for a group assessment, several aspects need to be fulfilled. The most fundamental aspect (and starting point) of any grouping approach is structural similarity among the group members. In addition, substances within a group must have comparable biological activity (e.g. physicochemical properties, absorption, distribution, metabolism, and excretion

(ADME) properties, and mechanism of action), in which the properties are similar across all group members or follow a regular predictable pattern in relation to the differences in chemical structures [26,27]. This information must scientifically justify the evaluation of the substances as a group. By applying group assessment approaches, available data can be used more efficiently, as data from one or multiple substances within the group could potentially be used to fill in data gaps of other substances within the group (this approach is also known as read-across) [26]. Consequently, group assessment approaches can reduce the need of (animal) tests for individual substances. In addition, group evaluations have the potential to speed up risk and hazard assessment as the evaluations are more resource efficient, and importantly may also prevent regrettable substitution to comparable substances. Within regulatory frameworks, first-steps have already been made in the transition from substance-by-substance assessment to group assessment approaches, indicating the potential efficiency of regulating groups of chemicals [28]. Nevertheless, this transition is still in its early stages and needs to be further optimized at all levels (screening, data generation and assessment) to ensure that substances of concern are progressed to regulatory risk management steps as efficient as possible [29].

1.7 Chemical similarity

Chemical similarity is considered a valuable target to further improve and optimize risk and hazard assessment approaches on the abovementioned aspects (see section 1.5 and 1.6). Throughout this thesis, chemical similarity is generally defined as the total overlap between molecular structures of two chemicals considering all structural features, and thus is not necessarily restricted to a common core structure (i.e. scaffold) or a common functional group. As (structurally) similar chemicals are likely to have similar properties (i.e. the similar property principle) [30], chemical similarity can be used for screening activities. Particularly, chemicals of potential concern could be identified based on chemical similarity to known substances of concern. In addition to screening purposes, chemical similarity can also be used for defining and evaluating groups of comparable substances, as chemical similarity is one of the prerequisites for grouping chemical structures. Accordingly, more extensive and targeted use of chemical similarity within risk and hazard assessment has the potential to improve the early signaling of concerns and stimulate the transition from substance-by-substance assessment to group assessment approaches, and therefore is a key topic within this thesis.

The assessment of chemical similarity between two substances is rather subjective, as "similarity like beauty is more or less in the eye of the beholder" [31]. It can be challenging to state with certainty that two substances are structurally similar to each other and what their degree of similarity might be [31]. Therefore, computational methods have been developed to support similarity assessments in a consistent manner. These methods are generally based

on a similarity measure, which consists of a chemical descriptor (also known as a fingerprint) and a similarity coefficient [32]. More details on chemical similarity calculations are provided in Textbox D.

Textbox D: Illustration of chemical similarity calculations

A similarity measure is used to calculate the structural similarity between two chemicals (Figure D.1). The two main elements of a similarity measure are a descriptor (or representation) of the chemical structures and a similarity coefficient [32].

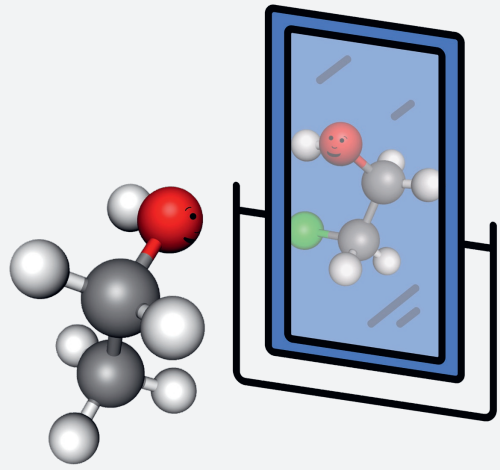


Figure D.1. Illustration of chemical similarity. Chemicals are drawn with MolView.

Descriptor

First, descriptors are used to characterize the molecules by assigning numerical values to the structures. Several types of descriptors are available to represent chemical structures, of which chemical fingerprints are most commonly applied for similarity searches [32]. A chemical fingerprint focusses on the presence or absence of substructures and can be calculated from the 2D representation of molecules. Generally, fingerprints are expressed in fixed-length binary bit-strings, in which each bit represents the presence (1-score) or absence (0-score) of a structural feature. The type and number of fragments/features that are analyzed depends on the fingerprint [32,33]. Several types of fingerprints exist, including dictionary-based fingerprints, path-based fingerprints and circular-based fingerprints (see Figure D.2) [34]. Within dictionary-based fingerprints, each bit relates to the presence or absence of a predefined fragment. Within path- and circular-based fingerprints, all fragments present within a structure are characterized (according to specified rules that are dependent on the fingerprint) and are assigned to a bit in the fingerprint using a hashing algorithm.

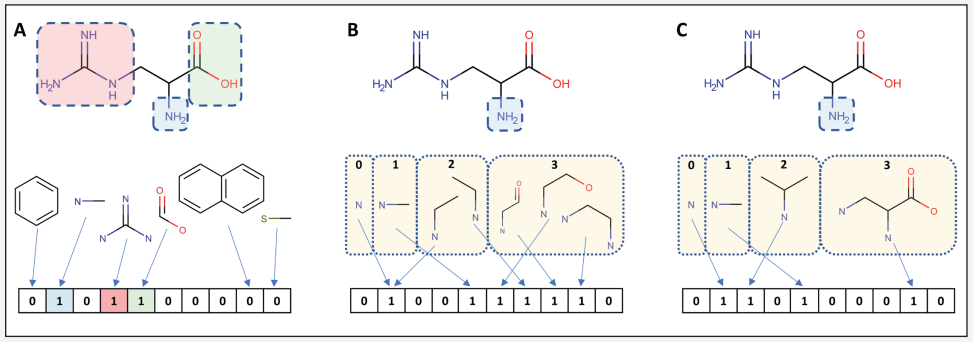


Figure D.2. Examples of chemical fingerprints. A) Dictionary based fingerprint: only the presence/absence of specific predefined fragments is analyzed. B-C) Path-based fingerprint and circular-based fingerprint, respectively: Specific fragments are extracted from the structure and are assigned to a bit in the fingerprint. Extracted fragments start at a specific atom (as highlighted in the structure) and follow a path/radius through the structure with a specific path-length/radius. Generally, fragments up to a path-length of seven or radius of three are explored and assigned. For illustration purposes only fragments from one starting atom (with limited path-length) are shown. Note that bit-collision may occur when two fragments are assigned to the same bit.

Similarity coefficient

Secondly, similarity coefficients are used to quantitatively express the similarity between two chemical descriptors [7,32]. When comparing two binary fingerprints, four different bit-combinations can be identified, denoted as *a*, *b*, *c* and *d*-bits (see Figure D.3):

- *a*-bits: represent the counts that a feature is present in the first and absent in the second structure ("x=1 and y=0").
- *b*-bits: represent the counts that a feature is absent in the first and present in the second structure ("x=0 and y=1").
- *c*-bits: represent the counts that a feature is present in both structures ("x=1 and y=1").
- *d*-bits: represent the counts that a feature is absent in both structures ("x=0 and y=0").

These four bit-counts are combined in similarity coefficients to quantify chemical similarity, in which either *c*- and/or *d*-bits represent the structural overlap between two chemicals (i.e. the common presence or absence of features). Over 40 different similarity coefficients are available to calculate similarity values between binary fingerprints [7], of which the Jaccard-Tanimoto coefficient is most familiar (Equation D.1).

$$s = \frac{c}{a+b+c} \qquad (D.1)$$

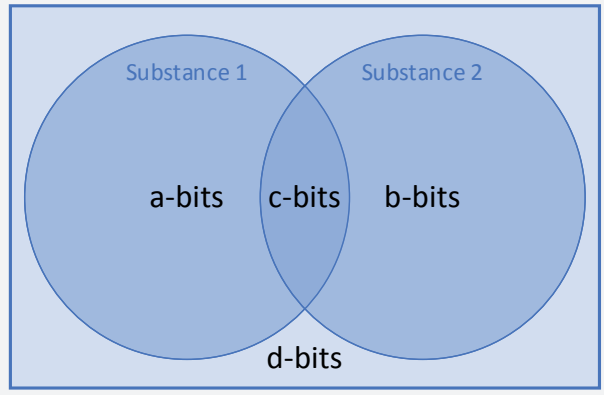


Figure D.3. Representation of the four different types of bits that can be identified when analyzing the overlap in the chemical fingerprints of two chemicals.

The similarity scores (also known as similarity values) that are calculated by the similarity coefficients generally range between 0 and 1. A similarity value of 0 means that two structures are totally different, and a similarity score of 1 means that two structures are (nearly) identical.

Similarity evaluation

Similarity assessments generally use a similarity threshold to classify chemicals as structurally similar or dissimilar. There is however no uniform similarity threshold, as the similarity values depend on the fingerprint (which all consider a specific set of fragments) and the similarity coefficient (which all have varying functional shapes/distributions). There are many types of descriptors and similarity coefficients available and there is no similarity measure that consistently is most effective. The most optimal similarity measure for instance depends on the type of data and on the goal of the analysis [32].

1.8 Aims and outline of this thesis

This thesis primarily focusses on (aspects related to) chemical similarity and its (potential) use within risk and hazard assessment approaches. The separate sections specifically target chemical similarity in relation to SVHC-properties, and address screening, data generation and evaluation of substances. The overall aim of this thesis is to develop similarity-based models that enhance the identification of chemicals of potential concern, and stimulate the transition from substance-by-substance assessment towards group assessment approaches. In addition, this thesis aims to investigate how (biological) similarity and variability could influence and could be applied in risk and hazard assessment.

This research is done within the scope of a science-based assessment of the safe and sustainable use of chemicals within the chemical universe of modern society. The scientific results of this thesis will contribute to future (regulatory) frameworks, including the EU-wide ambition for a toxic-free environment. It provides opportunities and directions to tackle today's challenges associated with the lack of (reliable) data and the relatively inefficient evaluation and regulation process. To enable the direct communication with risk assessors, this research is performed at Leiden University and the National Institute for Public Health and the Environment (RIVM) at the Centre for Safety of Substances and Products (RIVM-VSP). The scientific knowledge is gained at both research institutes, whereas the added value of the RIVM allows to act directly at the science-policy interface.

In the first part of this thesis, **chapters 2-4**, I focus on the development and improvement of chemical similarity models. The aim of the similarity models is to identify potential SVHC substances for screening and prioritization processes. The availability and use of more specific early-warning triggers may increase the chance to identify and regulate substances of concern before widespread exposure has occurred. Chapter 2 specifically focusses on the development of similarity models, whereas chapter 3 focusses on the application of these models on the broader universe of chemicals. Within chapter 4, the developed similarity models are further optimized based on the results and conclusions from chapters 2 and 3. The scientific knowledge gained in these chapters is immediately incorporated into an online instrument that is made freely available and can be used for the regulation of chemicals in modern society as it helps to identify chemicals of potential concern (https://rvszoeksysteem.rivm.nl/ZzsSimilarityTool).

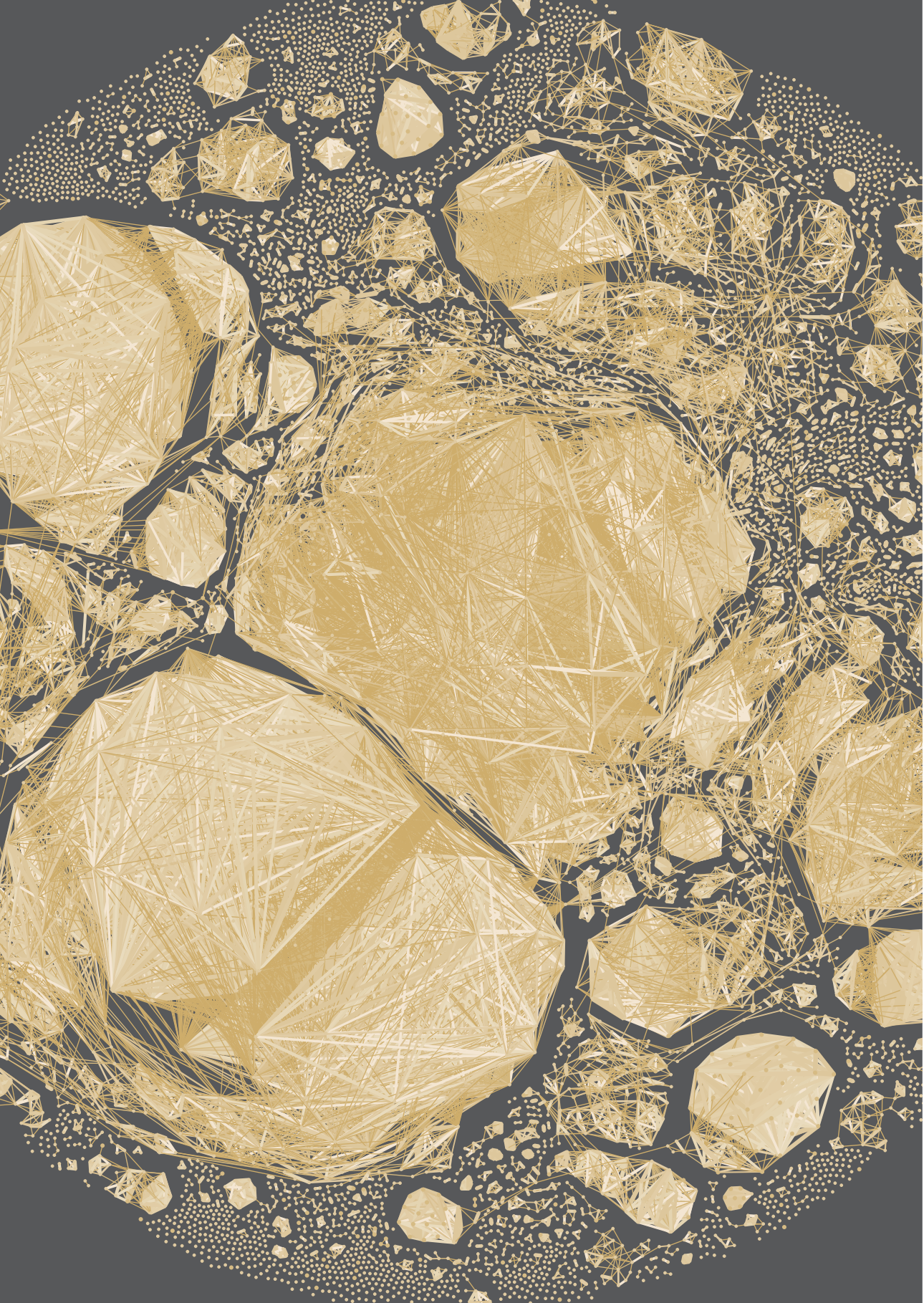
Chapter 5 covers another aspect of similarity and focusses on (biological) variability and uncertainty. Variability is an important aspect in relation to similarity. Generally, two substances can be considered as similar when they are structurally and biologically similar (i.e. they have the same or a predictable trend in biological activity). However, in order to conclude that substances are biologically similar/dissimilar, we need to know the individual

variation in biological activity. To gain additional insight in the potential impact of variability on similarity assessments, we analyzed and evaluated the variation in fish bioconcentration factors (BCF) for single substances. Within this chapter, BCFs were selected as an example, as BCFs are important for the identification of PBT/vPvB SVHCs (see Textbox C).

Within **chapter 6**, the use of chemical similarity for evaluation purposes is investigated. This chapter specifically focusses on the PBT-assessment of petroleum UVCBs, and aims to derive a conclusion of the PBT/vPvB properties of a group of 884 chemicals (constituents) within one assessment. Alkylated three-ring polycyclic aromatic hydrocarbons (PAHs) constitute the chemicals of interest.

Finally, **chapter 7** synthesizes the overall findings within the previous chapters and puts them in perspective with the rapidly changing societal challenges. In addition, chapter 7 provides recommendations and future perspectives on how to improve risk and hazard assessment.

1



2

Chemical Similarity to Identify Potential Substances of Very High Concern – an Effective Screening Method

Pim N.H. Wassenaar, Emiel Rorije, Nicole M.H. Janssen, Willie J.G.M. Peijnenburg and Martina G. Vijver

Published in Computational Toxicology 12 (2019), 100110.

Abstract

There is a strong demand for early stage identification of potential substances of very high concern (SVHC). SVHCs are substances that are classified as carcinogenic, mutagenic or reprotoxic (CMR); persistent, bioaccumulative and toxic (PBT) or very persistent and very bioaccumulative (vPvB); or as substances with an equivalent level of concern, like endocrine disruption (ED). The endeavor to improve the identification of potential SVHCs is also acknowledged by the European Commission, in their long-term vision towards a non-toxic environment. However, it has been shown difficult to identify substances as potentially harmful.

With this goal in mind, we have developed a methodology that predicts whether a substance is a potential SVHC based on chemical similarity to chemicals already identified as SVHC. The approach is based on the structural property principle, which states that structurally similar chemicals are likely to have similar properties.

We systematically analyzed the predictive performance of 112 similarity measures (i.e. all different combinations of 16 binary fingerprints and 7 similarity coefficients) classifying the substances in the dataset as (potential) SVHC or non-SVHC. The outcomes were analyzed for 546 substances that we collected within the Dutch SVHC database – with identified CMR, PBT/vPvB and/or ED properties – and 411 substances that lack these hazardous properties. The best similarity measures showed a high predictive performance with a balanced accuracy of 85% correct identifications for the whole dataset of SVHC substances, and 80% for CMR, 95% for PBT/vPvB and 99% for ED subgroups.

This effective screening methodology showed great potential for early stage identification of potential SVHCs. This model can be applied within regulatory frameworks and safe-by-design trajectories, and hence can contribute to the EU goal of achieving a non-toxic environment.

2.1 Introduction

In recent decades, exposure to specific chemicals appeared of greater concern than previously anticipated, including concerns for polychlorinated biphenyls (PCBs), dichlorodiphenyl-trichloroethane (DDT) and perfluorooctanesulfonic acid (PFOS) [20]. In many cases, when safety concerns are raised, widespread exposure has often already occurred, and typically the set of available toxicity data is inadequate to introduce risk management measures immediately. Consequently, chemicals of potential concern continue to be emitted, with the risk of significant effects on human and environmental health in the long-term. Therefore, it is important to signal emerging concerns and improve the early stage identification of hazardous chemicals before widespread exposure occurs. This endeavor is also acknowledged by the European Commission in their long-term vision towards a non-toxic environment [35,36]. In particular, high priority is given to so-called substances of very high concern (SVHC), which include substances with carcinogenic, mutagenic or reprotoxic (CMR) properties, substances with persistent, bioaccumulative and toxic (PBT) or very persistent and very bioaccumulative (vPvB) properties, or substances with endocrine disrupting (ED) properties [12]. Substances can be identified as SVHC following a regulatory decision process in which all available data is evaluated.

To improve the identification of potential SVHCs, it is essential to make efficient use of the limited amount of available (fate and toxicity) data. Several models have been described in the literature that predict hazard properties of chemicals from simple properties, like aquatic toxicity based on the octanol/water partition coefficient (K_{ow}) and/or structural alerts [37–39], or based on more complex algorithms [40–45]. Many of these models are (at least partially) based on the structural property principle, which assumes that (structurally) similar chemicals are likely to have similar properties [30]. Although these models are very useful to predict the effect of a chemical on a specific endpoint, their applicability to identify potential SVHC substances is limited. This is a consequence of the fact that the group of SVHC substances covers a broad range of different toxicological endpoints and mode of actions - and are only identified following a regulatory decision process. Within current models it is difficult to simulate such a regulatory weight-of-evidence approach. Potentially, total chemical similarity to known SVHC substances can be a useful way to estimate (potential) SVHC status, as such a method might be able to cover more information on SVHC identification properties.

To our knowledge, only two models, both with the aim of prioritization, attempt to identify potential SVHCs directly based on structural similarity to substances already identified as being SVHCs, including the SINimilarity tool developed by ChemSec [46], and screening scenarios as applied by the European Chemical Agency (ECHA) within the SVHC Roadmap program [47]. However, these methods do not provide optimized and cross-validated

methodologies, resulting in an unknown predictive performance. If a high predictive accuracy could be achieved using only chemical similarity information, the lack of toxicity information can be bypassed, and those substances of potential SVHC concern, that are currently deemed "safe" in the absence of toxicity information, can be prioritized for further follow-up action. In addition, the chemical similarity information also provides a clear follow-up direction, as the potential concern is directly related to the concern of the most similar SVHC substance.

The aim of the present study was to evaluate the efficiency of a broad set of similarity measures for the identification of potential SVHCs, with a specific focus on separately identifying CMR, PBT/vPvB and ED concerns. We built upon the knowledge gained (see e.g. [32]) for calculating chemical similarity, that generally consists of two main elements: a descriptor (or representation) of the chemical structure and a similarity coefficient. First, descriptors are used to characterize the molecules that are compared by assigning numerical values to structures [32,33,48]. These values are in most methods related to the absence or presence of specific chemical substructures and are often encoded in fixed-length bit-strings (consisting of zeros and ones) [49]. These bit-strings are also known as fingerprints. Secondly, similarity coefficients are used to quantitatively express the similarity between two chemical descriptors [7,32,48]. For our purpose, the similarity between two fingerprints can be used to quantify the structural overlap between a chemical with unknown hazardous properties and known SVHCs. Many types of descriptors and similarity coefficients are available and there is no similarity measure that consistently is most effective (i.e. there is no single best "fingerprint - coefficient" combination for all applications) [32,49,50]. Our study outcome provides the most optimal set of similarity measures as a first screening model to identify substances of potential SVHC concern.

2.2 Methods

The study approach consists of four general steps (Figure 2.1). First, a dataset of substances with and without CMR, PBT/vPvB and/or ED properties was constructed (paragraph 2.2.1). Secondly, binary fingerprints were generated for all substances in the datasets (paragraph 2.2.2). Thirdly, similarity values (i.e. quantitative values of chemical similarity) were calculated between substances by comparing the fingerprints with similarity coefficients (paragraph 2.2.3). Only the extent of similarity to substances with identified CMR, PBT/vPvB and/or ED properties leading to the SVHC status was investigated. Finally, we determined an optimal similarity threshold and the predictive performance of each "fingerprint-coefficient" combination (paragraph 2.2.4). Steps two to four were reiterated for multiple "fingerprint-coefficient" combinations, as well as for different SVHC subgroups (i.e. for CMR, PBT/vPvB and ED separately and together), in order to identify the optimal model(s) based on balanced accuracy. A more elaborate description of these steps is provided in the following paragraphs.

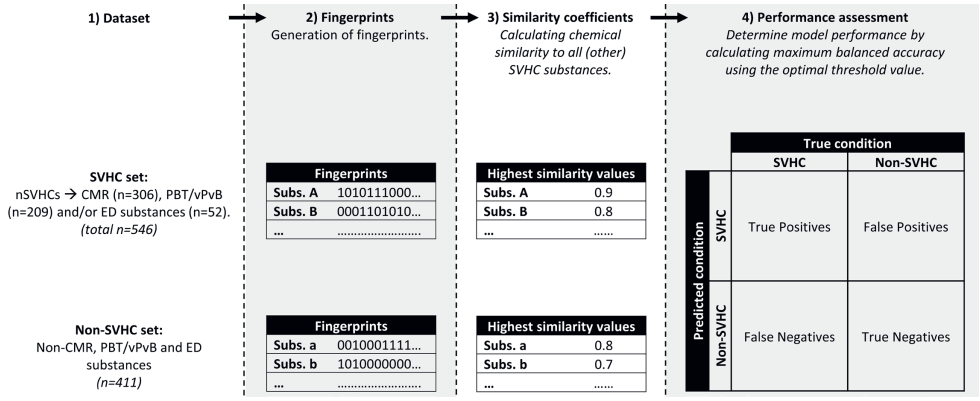


Figure 2.1. Overview of the methodology divided into four steps. Steps two to four were reiterated for multiple fingerprint-coefficient combinations.

2.2.1 Dataset

In order to identify chemicals of (potential) concern based on structural similarity to known toxicants, a set of known CMR, PBT/vPvB and ED substances is required. For this purpose, a Dutch list of substances of very high concern was selected, as all substance on this list have CMR, PBT/vPvB and/or ED properties (see [51]; extracted on 01-03-2018). This list covers a broader range of chemicals than the EU-SVHC list under REACH, but are identified based on the same hazard criteria as the EU-SVHC substances (i.e. REACH article 57 [12]). The generation and composition of this list of substances is more elaborately described in Supplemental Material S.1.

In addition, for modelling purposes we also compiled a list of substances that are known not to have CMR, PBT/vPvB and/or ED properties. All substances on the REACH Annex IV – which lists chemicals that are considered to be inherently safe – were selected for this purpose, as well as all approved biocides and pesticides (see [52,53]; extracted on 23-05-2018). The list of biocides and pesticides is suited for our purpose as all substances approved for introduction on the European market have been tested experimentally and are negative for CMR, PBT/vPvB and ED endpoints, according to the SVHC criteria.

Several adjustments were made to the compiled substance lists, as chemical similarity searches require a specific and unambiguous chemical structure as input information. In cases that a group of substances was included in one of the above-mentioned lists (e.g. polychlorinated naphthalenes), representative chemical structures were generated and selected for inclusion in order to ensure that the structures represent the varying types of branching and/or substituents (e.g. tri- up till octachloro naphthalene, with two isomers per chlorine-atom

count). When a substance is a mixture or a UVCB (Substances of Unknown or Variable composition, Complex reaction products or Biological materials), only the (representative) chemical structures of those components causing the concern were included (e.g. benzene in some of the UVCBs). When a substance is considered a non-SVHC substance, the main constituent(s) were included. Each unique chemical structure was included once in the final list. In addition, specific metal-complexes (i.e. based on arsenic, beryllium, cadmium, chromium, lead, mercury, nickel and cobalt) and fibers were excluded. For these metal-based complexes, it is generally the metal atom causing the concern, irrespective of the organic counterparts. In case of fibers, the toxicity is (also) determined by physical aspects other than their chemical structure (e.g. diameter, length and shape). In addition, all inorganic substances were removed from the list of non-SVHC substances.

In total, a dataset of 546 SVHC and 411 non-SVHC single chemical structures was compiled (see Supplemental Material Excel). Of the 546 SVHC substances, 306 are known to have CMR properties, 209 to have PBT/vPvB properties, and 52 are known to have ED properties. All chemical structures were represented by a (single) SMILES code [54] and all charged structures were converted to their neutral counterparts, where possible (Supplemental Material S.2). These SMILES codes were used for the analyses.

2.2.2 Fingerprints

We restricted this study to binary fingerprints based on 2D-fragments, as they tend to be more selective than whole molecule descriptors. Moreover, 2D-fragments descriptors are (computationally) easier to handle than 3D-fragment descriptors [32]. The fingerprints were selected in such a way to ensure maximum diversity and include dictionary-based, path-based, circular-based and pharmacophore-based fingerprints (Table 2.1) [34]. The fingerprints were generated using freely available resources, including the software packages RDkit and PaDEL-Descriptor (based on the Chemistry Development Kit (CDK) libraries) [6,55]. For all non-dictionary based fingerprints, a string length of 1024 bits was used. More details on the generation of the fingerprints are given in Supplemental Material S.3.

2.2.3 Similarity coefficients

The similarity between two 2D-binary fingerprints of known SVHCs and non-SVHC substances can be computed by using various formulas, the so-called similarity coefficients. When comparing two binary fingerprints, four different bit-combinations could be identified - denoted as *a*, *b*, *c* and *d*. *A*, *b*, *c* and *d* represent the counts that a feature is present in one structure and absent in the other ("x=1 and y=0"), absent in the first and present in the second structure ("x=0 and y=1"), present in both ("x=1 and y=1") and absent in both

("x=0 and y=0"), respectively. These four numbers are combined in similarity coefficients to quantify chemical similarity. In total, 44 different similarity coefficients are available to calculate similarity values between binary fingerprints [7]. We selected seven coefficients for our analysis based on diversity and based on their performance as observed by Todeschini et al. (2012) and Floris et al. (2014) [7,56] (see Table 2.2). Similarity coefficients "SS1", "Ja" and "Gle" all showed a high performance within Todeschini et al. 2012, but have an exactly similar performance as the JT-coefficient. Therefore, it has been decided to only include the JT-coefficient within this study. All included similarity coefficients were rescaled to provide similarity values between 0 and 1 using Equation 2.1, similar to Todeschini et al. (2012) [7].

$$s' = \frac{s+\alpha}{\beta} \qquad (2.1)$$

Where s is the original similarity value (Table 2.2), s' is the rescaled function in the range [0, 1], and α and β are numerical parameters whose values are reported in Table 2.2. When α = 0 and β = 1, this means that no transformation has been applied [7].

Table 2.1. Binary fingerprints included in this study.

Name	Number of bits	Type of fingerprint	Source
Substructure Fingerprints	307	Dictionary based fingerprints	PaDEL-Descriptor [6]
MACCS Fingerprints	166		
E-State Fingerprints	79		
PubChem Fingerprints	881		
Klekota-Roth Fingerprints	4860		
CDK Extended Fingerprints	1024	Topological or Path-based fingerprints	RDkit [55]
Atom Pairs Fingerprints	1024		
Topological Torsion Fingerprints	1024		
Extended Connectivity Fingerprints (diameter = 0) (ECFP0)	1024	Circular fingerprints *	
Extended Connectivity Fingerprints (diameter = 2) (ECFP2)	1024		
Extended Connectivity Fingerprints (diameter = 4) (ECFP4)	1024		
Extended Connectivity Fingerprints (diameter = 6) (ECFP6)	1024		
Functional-Class Fingerprints (diameter = 0) (FCFP0)	1024	Circular/pharmacophore fingerprints *	
Functional-Class Fingerprints (diameter = 2) (FCFP2)	1024		
Functional-Class Fingerprints (diameter = 4) (FCFP4)	1024		
Functional-Class Fingerprints (diameter = 6) (FCFP6)	1024		

**Morgan fingerprints were calculated using RDkit with radius of 0, 1, 2 and 3; which is roughly equivalent to ECFP and FCFP0, 2, 4, and 6.*

Table 2.2. Similarity coefficients included in this study (obtained from [7]).

Name	Formula	α	β	Class	Conditions
Jaccard-Tanimoto (JT)	$s = \frac{c}{c+a+b}$	0	1	A	c=0 → s=0
Harris-Lahey (HL)	$s = \frac{c(2d+a+b)}{2(c+a+b)} + \frac{d(2c+a+b)}{2(a+b+d)}$	0	p	S	c=p or d=p → s=1; den=0 → s=0
Consonni-Todeschini 4 (CT4)	$s = \frac{\ln(1+c)}{\ln(1+c+a+b)}$	0	1	A	None
Sokal-Sneath 3 (SS3)	$s = \frac{1}{4}\left[\frac{c}{c+a} + \frac{c}{c+b} + \frac{d}{a+d} + \frac{d}{b+d}\right]$	0	1	S	c=p or d=p → s=1; c=0 and d=0 → s=0
Cohen (Coh)	$s = \frac{2(cd-ab)}{(c+a)(a+d)+(c+b)(b+d)}$	+1	2	Q	c=p or d=p → s=1; den=0 → s=0
Simple Matching (SM)	$s = \frac{c+d}{c+a+b+d}$	0	1	S	None
Yule 2 (Yu2)	$s = \frac{\sqrt{cd} - \sqrt{ab}}{\sqrt{cd} + \sqrt{ab}}$	+1	2	Q	c=p, d=p or ab=0 → s=1

Names of the coefficients are provided as in accordance to Todeschini et al. 2012 [7], though the definition of a and c are switched in Todeschini et al. 2012 [7]. The column "Class" represents the type of coefficient: S = symmetric coefficient (counts a and d are considered equally); A = asymmetric coefficient (only count a is considered); Q = correlation based coefficients that are transformed to obtain a value between zero and one. The column "conditions" represents conditions that were assumed in order to avoid singularities. Den = denominator; p = a + b + c + d.

2.2.4 Performance assessment

Performance statistics

In total, 112 different similarity measures were selected (i.e. all different combinations of 16 fingerprints and 7 similarity coefficients) and we analyzed their predictive performance on classifying the substances in the dataset as (potential) SVHC or non-SVHC. For non-SVHC substances, similarities were calculated to all substances in the SVHC set based on the fingerprint-coefficient combination. Similarities for SVHC substances were calculated to all other substances on the SVHC set. Iteratively, one SVHC molecule at a time was left out of the dataset and compared to the other SVHC substances. For each substance, only the highest similarity value was retained.

For each fingerprint-coefficient combination, we determined the maximum balanced accuracy (Equation 2.2), by selecting the optimal threshold (i.e. a value between 0 and 1) to predict (potential) SVHC status versus non-SVHC status. Substances with a similarity value equal to or above this threshold are predicted to be structurally similar to a substance with CMR, PBT/vPvB or ED properties to such an extent that they are potential CMR, PBT/vPvB or ED themselves (and vice versa). When using a threshold value, the number of 'True Positives

(TP)', 'False Positives (FP)', 'False Negatives (FN)' and 'True Negatives (TN)' predictions can be determined for a fingerprint-coefficient combination, as well as the balanced accuracy (Equation 2.2). By iteratively assessing the fingerprint-coefficient performance for all distinguishing threshold values (ranging from 0-1), the optimal threshold, with maximum balanced accuracy could be determined. The optimal threshold was selected for each specific fingerprint-coefficient combination to ensure equal model comparisons.

$$Balanced\ Accuracy = \frac{Sensitivity+Specificity}{2} = \frac{\frac{TP}{TP+FN}+\frac{TN}{TN+FP}}{2} \qquad (2.2)$$

Best model selection

In addition to the overall performance (with all CMR, PBT/vPvB and ED substances together in the reference set), also the predictive performance of all fingerprint-coefficient combinations for specific subgroups were analyzed (i.e. for the subgroups of CMR, PBT/vPvB and ED substances separately). The whole set of non-SVHC substances was used as truly negative data in each case. The best performing model was selected based on the balanced accuracy.

Best model evaluation

Within the best performing models, we analyzed whether potential bias was introduced by the optimal similarity coefficient. Specifically, symmetric similarity coefficients may tend to predict small substances - with many '0-bits' - as similar to small SVHC substances, because of common absence of many features (i.e. *d*-fragments). Although such a model could be considered most optimal based on statistical performance of the dataset, the occurrence of this type of similarities is undesirable, as upon application many small substances will incorrectly be classified as (potential) SVHC. Therefore, when potential symmetric coefficient bias was identified in a best performing model, we decided to use an asymmetric similarity coefficient for substances with a low number of '1-bits' (i.e. JT or CT4, which only considers *c*-fragments as similar). The most optimal fragment count cut-off was analyzed based on balanced accuracy.

Furthermore, we analyzed the robustness of the best performing models by assessing the performance after two different robustness checks. Within the first robustness check, we extended the non-SVHC dataset by adding the substances of the "non-relevant" SVHC subgroup to the non-SVHC dataset. To illustrate, for the CMR-model, all PBT/vPvB and ED SVHC substances that do not have CMR properties were considered as not-CMR, and thus added to the non-SVHC set for this robustness check. This robustness check could not have been conducted on the overall model, as in this case all SVHC subgroups are relevant. Within a second robustness check, we reduced the number of representative structures of

group entries that were included within the SVHC as well as within the non-SVHC set to generally two structures (see Supplemental Material Excel). In addition, some structurally similar substances are represented various times in the SVHC or non-SVHC datasets, including a large number of individual PCB isomers, chlorinated dibenzofurans, chlorinated dibenzodioxins and polybrominated diphenyl ethers on the PBT/vPvB dataset. To determine the robustness of the best performing models, such groups have also been reduced to a representation of generally two representative structures (see Supplemental Material Excel). The performance of the adjusted datasets within the different robustness checks was assessed similarly as described above, using the optimal threshold of the best-performing model.

In addition, hierarchical cluster diagrams were generated for the different SVHC subgroups in order to analyze the diversity within the subgroups. Hierarchical clusters were based on the similarity matrix of the subgroup, using single-linkage method.

The performance of the best predictive models was also compared to existing methodologies – using the SVHC dataset – including Toxtree (i.e. Benigni/Bossa rulebase for mutagenicity and carcinogenicity), DART and the PB-score tool [38,39,57]. For this analysis, the presence of a structural alert from Toxtree and/or DART was interpreted as a prediction of SVHC status based on CMR properties.

Besides performance evaluation, also applicability domain was analyzed by determining the 95th percentile of molecular weight, log K_{ow} [37], number of atoms, number of rings and number of aromatic rings within the applied datasets.

All data was analyzed in R (version 3.5.1) [58], using *caret*, *ChemmineR*, *caTools*, *ROCR* and *rcdk* [59–63].

2.3 Results

2.3.1 Best model selection

Overall model performance

Table 2.3 shows the ten best performing models when all CMR, PBT/vPvB and ED substances are taken together in a single SVHC dataset. A wide variety of fingerprints was identified in the top ten models, including dictionary-based, path-based, circular-based and pharmacophore-based fingerprints. In contrast, one similarity coefficient, the Simple Matching (SM), is dominating the top ten models. Furthermore, it can be observed that relatively high optimal similarity thresholds are determined. The height of the threshold is highly related to the used similarity coefficient, and is specifically high for the SM coefficient (Figure S.1). This is a

consequence of the fact that c and d variables are treated as similar in this coefficient (Table 2.2).

The overall best performing model, PubChem-SM combination, has an overall balanced accuracy of 0.846. However, this specific combination is not the most optimal for the specific subgroups, having different (toxicological) concerns. Therefore, we also analyzed model performances for the CMR, PBT/vPvB and ED groups separately.

Subgroup model performance

The best performing similarity models optimized for the separate CMR, PBT/vPvB and ED subgroups are shown in Table 2.4 (in row one till three, respectively). For the ED subgroup, 30 out of the 112 tested different similarity measures showed similar predictive performance, but the rank of the fingerprints and coefficients separately shows a highest rank for the FCFP4 fingerprint and the SS3 similarity coefficient. The best performing combination of fingerprint and similarity coefficient is different for the different subgroups, and a (slightly) higher balanced accuracy is obtained when compared to the best performing overall model (Table 2.3).

2.3.2 Best model evaluation

Symmetric coefficient bias

By applying the "Extended fingerprint – SM coefficient" combination for the CMR dataset, with a 0.944 similarity threshold, all substances with less than 63 fingerprint bits were considered to be similar to CMR-SVHCs (Figure 2.2A). This coefficient bias is also observed upon visual inspection of the FP-substances, perceiving a better similarity assessment with increased number of fingerprint bits (e.g. 'Methyl octanoate' and '3-propanolide'; or 'Captan' and 'Captafol'; Figure 2.2B).

Table 2.3. Ten best performing fingerprint-coefficient combinations for the dataset with all CMR, PBT/vPvB and ED substances included. Also specific subgroup performances – in balanced accuracy – are provided based on the optimal overall threshold values. The numbers represent the number of SVHC substances, 411 non-SVHC substances were included. Highest balanced accuracies are given in italic bold. AUC is the area under the curve of ROC-plot.

Model		Threshold	Overall model performance (n=546 SVHC)					Balanced accuracy of subgroups using overall threshold value		
Fingerprint	Coefficient		Sensitivity	Specificity	Precision	AUC (ROC)	Balanced accuracy	CMR (n=306 SVHC)	PBT/vPvB (n=209 SVHC)	ED (n=52 SVHC)
Pubchem	SM	0.985	0.810	0.883	0.902	0.904	***0.846***	0.801	0.929	0.988
Extended	SM	0.957	0.806	0.878	0.898	0.897	0.842	***0.811***	0.889	0.981
MACCS	SM	0.970	0.734	0.946	0.948	0.897	0.840	0.760	***0.951***	0.960
FCFP4	SM	0.991	0.835	0.842	0.875	0.893	0.839	0.802	0.911	***0.990***
KlekotaRoth	SM	0.998	0.773	0.898	0.909	0.889	0.835	0.777	0.921	0.942
ECFP2	SM	0.992	0.852	0.813	0.858	0.900	0.832	0.798	0.925	0.987
ECFP4	SM	0.984	0.832	0.832	0.868	0.882	0.832	0.791	0.900	***0.990***
Extended	SS3	0.895	0.714	0.942	0.942	0.888	0.828	0.775	0.902	0.971
Extended	Coh	0.884	0.711	0.934	0.935	0.887	0.822	0.769	0.899	0.981
MACCS	SS3	0.923	0.716	0.922	0.924	0.875	0.819	0.739	0.924	0.969

Table 2.4. Best performing fingerprint-coefficient combination for the CMR, PBT/vPvB and ED subgroups, including balanced accuracies after robustness checks (see section 2.3.2). The CMR model was improved by combining a symmetric and asymmetric coefficient in order to prevent symmetric coefficient bias (see section 2.3.2). In robustness check 1, the SVHC substances that did not belong to the subgroup of concern were added to the dataset as non-SVHCs. In robustness check 2, the number of representative structures for group entries and structurally similar substances were reduced to generally two structures in the SVHC and non-SVHC set. The numbers represent the number of SVHC substances. The number of non-SVHC substances varies between the full model assessment (n=411) and the robustness checks (see section 2.3.2). '-' means that it is not possible to calculate a single AUC for a combination of two models. AUC is the area under the curve of ROC-plot.

Subset	Model		Threshold	Sensitivity	Specificity	Precision	AUC (ROC)	Balanced accuracy	Robustness check	
	Fingerprint	Coefficient							1	2
CMR (n=306)	Extended	SM	0.944	0.784	0.854	0.800	0.859	0.819	0.735	0.799
PBT/vPvB (n=209)	MACCS	SM	0.970	0.919	0.983	0.965	0.971	0.951	0.942	0.911
ED (n=52)	FCFP4	SS3	0.866	0.981	1.000	1.000	0.984	0.990	0.969	0.917
CMR improved (n=306)	Extended	CT4 (<85) SM (≥85)	0.851 0.944	0.650	0.949	0.905	-	0.800	0.742	0.769

Based on our assessment, finding an optimal cut-off within the range of 63 to 100 fingerprint bits, the combination of the CT4 coefficient for substances with less than 85 fingerprint bits and the SM coefficient for substances with 85 or more fingerprint bits is most optimal, with a balanced accuracy of 0.800 and threshold values of 0.851 and 0.944, respectively (Table 2.4, row 4). The statistical performance of the CT4-SM combination is lower than the SM coefficient only (when looking at the balanced accuracy), due to an increase in FN-classified substances. On the contrary, also more substances are correctly classified as negative, including structures with a relative low number of fingerprint bits, like methyl octanoate and the terpenoid blend QRD-460 (Figure 2.2B; Figure S.2). This results in a much better specificity and precision (Table 2.4; Table S.1). The PBT/vPvB and ED models do not require a combination of asymmetric and symmetric coefficients as no symmetric coefficient bias was observed (Supplemental Material S.4; Figure S.2).

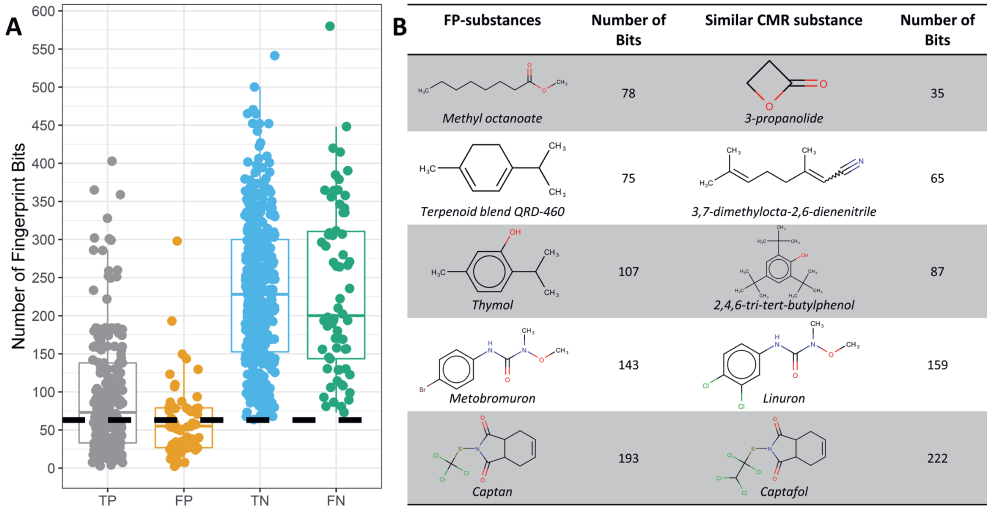


Figure 2.2. Classification of the CMR-SVHC and non-SVHC substances using the "Extended Fingerprint – SM coefficient" combination. A) Fingerprint bit count distributions across the different classifications: True Positive, False Positives, True Negatives and False Negatives. All substances with less than 63 fingerprint bits are classified as positive (dashed-line). B) Illustration of some False Positive classified substances and the most similar CMR substance. With an increase in the number of fingerprint bits, less ambiguous similarities are established.

Robustness checks

The robustness of the best-performing subgroup models was investigated via two robustness checks (Table 2.4). Within the first robustness check, the SVHC substances that did not belong to the subgroup of concern were added to the dataset as non-SVHCs (i.e. 'robustness check 1'). For the best performing CMR model, 651 non-SVHC substances were included, for the best PBT/vPvB model 748 non-SVHC substances and for the best ED model 905 non-SVHC substances. Within the second robustness check, we reduced the number of representative

structures for group entries and structurally similar substances of the SVHC and non-SVHC set to generally two structures (i.e. 'robustness check 2'). In total, 30 substances were excluded from the non-SVHC set, 35 from the CMR subset, 96 from the PBT/vPvB subset, and 34 from the ED subset.

Adding the non-target SVHC-substances to the non-SVHC set lowered the balanced accuracy and hence the predictive performance, specifically for the CMR similarity model. Conversely, removal of close structural analogues resulted in a larger decrease in predictive performance for the PBT/vPvB and ED specific models.

Single-point-of-knowledge

The CMR and PBT/vPvB subgroup have a quite broad basis with 306 and 209 substances, respectively, whereas the ED subgroup only consists of 52 substances. Within the PBT/vPvB and ED subgroups, some groups of very similar structures can be identified, and only a few single-point-of-knowledge structures (SPOKs) are included (Figure 2.3). SPOKs are substances that are not comparable to any other substance in the subgroup and thus are single-point-of-knowledges within the dataset (i.e. the FN). Within the ED substances, four groups and one distinct substance are present; in the PBT/vPvB subgroup, 15 groups and 17 distinct substances were identified (giving 1 and 17 false negatives, respectively). On the contrary, the CMR-SVHC dataset is much more diverse in chemical structures and contains much more SPOKs, reflected in the high number of FN-classified substances (n=107). For the CMR subgroup, no unambiguous hierarchical clustering can be generated as the CT4-SM coefficient combination does not fulfill the mathematical conditions for all substances (i.e. similarity between substance x and y is not necessarily similar to the similarity between y and x). Nevertheless, some groups can be identified, including polycyclic aromatic hydrocarbons, haloalkanes, cyclic and acyclic ethers, alkyl phenols, phthalates, aromatic amines, nitroaromatics and chloroaromatics. As a consequence of the high structural diversity, the calculated balanced accuracy is also lower for the CMR subgroup compared to the PBT/vPvB and ED groups. It should be noted that the SPOK false negatives will be included in the full dataset of SVHC substances when applying the model to a new substance.

Performance of existing models

The performance of a CMR model (i.e. the sum outcome from Toxtree and DART [39,57]) on the used SVHC-set was analyzed. Substances were considered as CMR by the model when a Toxtree or DART alert was identified. A balanced accuracy of 0.62 was determined, with a sensitivity of 0.78 and a specificity of 0.47. Furthermore, the performance of a PBT model was evaluated (i.e. PB-score tool [38]). For four substances no PB-score could be calculated as no log K_{aw} could be estimated. For the used dataset, a balanced accuracy of 0.73 was determined, with a sensitivity of 0.53 and a specificity of 0.93. No ED model was analyzed because of the

limitations identified in the ED-similarity model (see discussion).

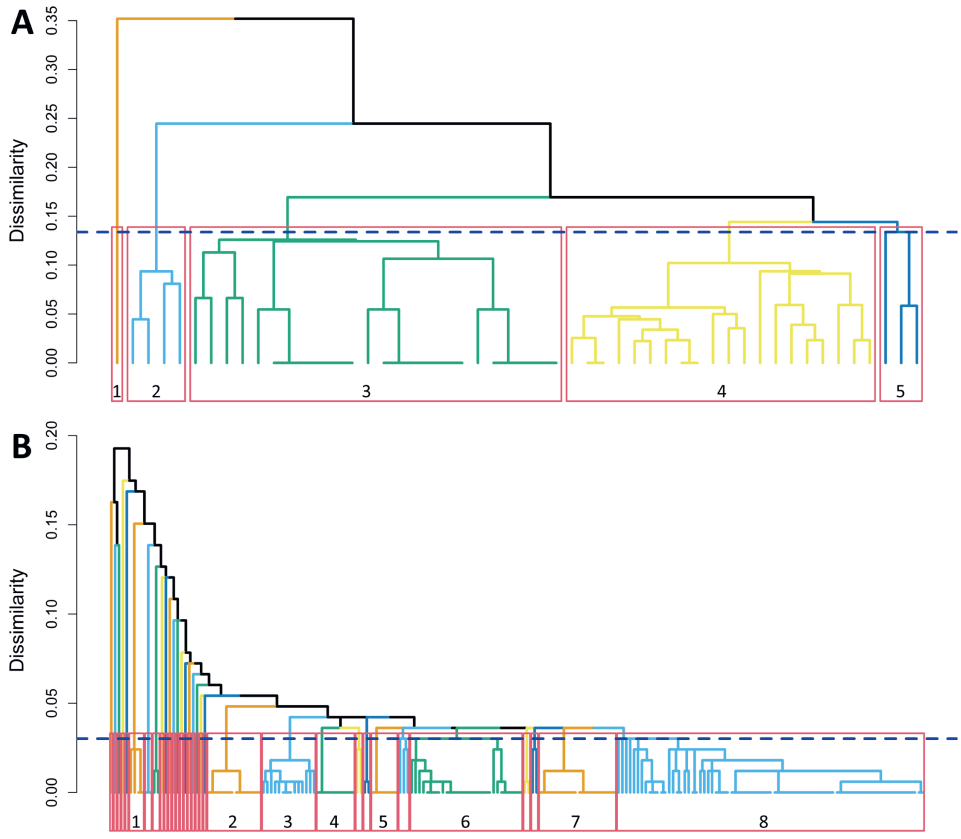


Figure 2.3. Hierarchical clustering for the ED and PBT/vPvB subgroups based on single linkage method. For ED, the FCFP4 fingerprint and SS3 coefficient are plotted, and for PBT/vPvB the MACCS fingerprint and SM coefficient. The y-axis describes the dissimilarity between the SVHC structures and is equal to 1 minus the similarity. The blue dotted line represents the used threshold (i.e. 1 minus threshold values). The red-colored boxes represent clusters of similar substances. A) ED clusters. Five different clusters can be identified: 1 = Diosgenin, 2 = Phthalates, 3 = Ethoxylated phenols, 4 = Nonyl and heptyl phenols, 5 = Octyl, pentyl and bi-phenols (Bisphenol A). B) PBT/vPvB clusters. Thirty-two different clusters can be identified, including some large clusters: 1 = Phenolic benzotriazoles, 2 = Halogenated Dioxins, 3 = Chlorinated paraffins, 4 = Brominated diphenyl ethers, 5 = Perfluorinated carboxylic acids, 6 = Polycyclic aromatic hydrocarbons, 7 = Halogenated dibenzofurans, 8 = Halogenated aromatics and cycloalkanes.

2.4 Discussion

As ever-increasing amounts of substances are produced, applied and emitted, it is important to focus attention on assessing the risks of those substances that are most likely to actually cause problems. Therefore, there is a need for efficient screening and prioritization methods

to identify chemicals with a high potential of being hazardous. Within this study we evaluated the efficiency of a set of similarity measures for the identification of (potential) SVHCs. Based on our approach, we identified the three best performing models for CMR, PBT/vPvB and ED subgroups, that all show a promising balanced accuracy (≥0.8) based on the used dataset.

2.4.1 Model performance

The three subgroup-specific models showed a better performance than one single overall model. This is likely related to a difference in mode(s) of action between CMR, PBT/vPvB and ED substances, and is also reflected in the most optimal fingerprints. In addition, predictive performance appeared reasonably robust with less than 10% reduction of balanced accuracy following the two robustness checks for all best performing models.

For the PBT/vPvB substances, the MACCS fingerprint performed best. The MACCS fingerprint contains only 166 predefined bits and was particularly developed to categorize substances in functional groups [64]. The PBT/vPvB dataset has a low structural diversity, with many substances sharing common structural features (Figure 2.3), including aromatic-rings and high levels of halogenation. In addition, small substances are often not considered PBT/vPvB, as in general a lower octanol-water-partitioning is observed for smaller substances, and this in turn is related to the bioaccumulation potential [19]. Apparently, the MACCS fingerprint is very effective in making a distinction between PBT/vPvB and non-PBT/vPvB substances based on these common features. Consequently, a high predictive performance is observed for this dataset (0.951).

The CMR substances are structurally much more diverse, with 107 SPOKs in the SVHC dataset. This diversity is also reflected in the most optimal fingerprint, the Extended Fingerprint. This path-based fingerprint, which is based on the well-known Daylight fingerprint [65], recognizes all paths within a structure consisting of 1-9 atoms (i.e. search depth of 8 bonds) and also includes some additional bits that describe ring features [6]. Compared to dictionary-based fingerprints, it is assumed that this method is more suitable to capture the broad diversity in CMR substances, as it characterizes all possible fragments within a structure.

As the balanced accuracy for the CMR subgroup was relatively low (compared to the PBT/vPvB and ED groups), we added an extra fingerprint that encodes for the presence of CMR-specific fragments identified in expert-models like Toxtree and DART [39,57]. Nonetheless, the inclusion of the mechanistically based substructures in the fingerprint did not lead to any improvement in the predictive performance (Supplemental Material S.5). Apparently, the size of the dataset and the fragments present in the optimal fingerprint already cover the specific structural features that have been linked to our collective knowledge of mechanisms of action

leading to CMR effects. The additional fingerprint is therefore excluded again.

For ED substances, the FCFP-4 is identified as best performing fingerprint. FCFP-4 identifies fragments based on functional group patterns. It recognizes atoms as hydrogen donors, hydrogen acceptors, aromatics, halogens, basic-atoms and acidic-atoms, and it identifies fragments based on patterns between these atoms (e.g. hydrogen donor – hydrogen acceptor – hydrogen donor) [55]. Endocrine disruptors generally interact with specific hormone receptors or interact with proteins in the hormone pathway [66], and such (receptor) binding properties are potentially identified best by the features covered in the FCFP-fingerprint. Furthermore, the diameter of 4 (FCFP-4) scored slightly better for the similarity search than a diameter of 2 or 6, which is in line with earlier findings [67]. Rogers and Hahn (2010) [67] concluded that a diameter of four is typically sufficient for similarity searches whereas a diameter of six or eight is best for activity learning methods.

Despite the very high performance for the ED subgroup (0.990), prediction results from this model should be interpreted with caution. The currently used ED-SVHC dataset is limited as it only consists of a few number of substances that have a large structural overlap (Figure 2.3) and consequently results in higher uncertainty around the optimal threshold value compared to the other models (Figure S.3). In addition, there is only one substance on the ED-list with a hormone backbone (i.e. Diosgenin). The reason for the low number of identified ED-SVHC substances is partially related to the fact that only those substances are identified as ED for which SVHC-identification is of added regulatory value. In addition, only recently guidance and criteria are developed for the identification of ED substances [68]. It is recommended to further develop the ED model when more substances are classified as ED-SVHC, or by including known endocrine disrupting substances such as the natural substrates (and synthetic variants derived thereof) interacting with estrogen/androgen/thyroid and steroidogenic pathways. With a broader dataset, a more sophisticated screening model will be possible. Based on the current dataset the ED-SVHC similarity model is expected to miss many (potential) ED substances.

A higher performance is observed for the best-scoring CMR and PBT/vPvB similarity models compared to existing models [38,39,57], when using the SVHC dataset. This indicates the value and relevance of the structural property principle for identifying potential SVHC substances. For the ED model, no comparison was made with existing models because of the limitations as mentioned above.

2.4.2 Focus and restriction of the modelling

We limited our assessment to the performance of 2D-binary fingerprints, and the presence

or absence of 2D-fragments. More sophisticated fingerprints are also available, including count-based fingerprints, taking into account how many times a fragment is present, or 3D-fingerprints that consider chemical conformation. Particularly, 3D-fingerprints could be relevant to identify potential ED substances, as receptor-binding properties are highly important for this group. In general, however, 2D-binary fingerprints are most popular as they are an acceptable trade-off between the wealth of (possible) information and simplicity, enabling an easy and quick comparison [32,56]. Especially for the proposed screening activities, the currently evaluated methodology is considered adequate.

In principle, all non-SVHC substances that have been used for modelling purposes within this study are tested on CMR, PBT/vPvB and ED properties. Nevertheless, it is possible that some substances are currently not identified as such, but will become a SVHC substance in future, when new information becomes available or when new evaluations are conducted. For instance, glyphosate is included in the non-SVHC list used in this study, although its carcinogenicity is currently extensively discussed [69,70]. Furthermore, as shown in Figure 2.2, Captafol is considered as CMR substance whereas its close structural analogue Captan is not (see Supplemental Material S.1). Captafol is classified as a carcinogen category 1B (leading to SVHC status), and Captan as a carcinogen category 2 [71]. Although the model identifies Captan as a false positive, the results could be very useful and may provide further arguments for (de)-classification of these substances. For instance, within European regulatory frameworks, a category 2 classification (for carcinogenicity but also for mutagenicity and reproductive toxicity) is often the highest classification that can be agreed upon when there are insufficient (experimental) data to support a category 1B classification [72].

Despite the conductance of a performance analysis, including robustness checks, we were not able to conduct a proper external validation in order to analyze the performance on an external dataset. As SVHCs are identified after a regulatory decision process in which all available data is evaluated, we are not in the position to mark substances as SVHC for external validation purposes. Similarly, non-SVHC substances are challenging to assign, as many substances are not extensively evaluated on all SVHC endpoints (i.e. CMR, PBT/vPvB and ED). A proper external validation set can therefore only be developed in future, when new SVHC and non-SVHC substances are identified. Future work will focus on the application of the developed methodology to large sets of substances to obtain a better idea of the application performance.

2.4.3 Use and applicability domain of the model

The assumption, that structurally similar substances are likely to have similar properties, seems valid based on our analysis and model performances. The proposed similarity models focus on multiple endpoints (i.e. CMR, PBT/vPvB and ED) and could be applied as a first screening

model, enabling to prioritize further follow-up analyses. The model directly highlights the most similar SVHC substance(s), which could provide additional information on the specific concerns. The absolute results should not be interpreted as a conclusive outcome. The methodology is framed to give systematic and transparent ways to identify relations that would not manually be identified. Based on the follow-up, it could be concluded that 1) the substance is likely to have similar effects, 2) that further data is required to substantiate the outcome, or 3) that the substance is not expected to have CMR, PBT/vPvB or ED properties.

Furthermore, it should also be highlighted that the developed model considers a screening model to identify whether new chemicals are structurally similar to known SVHC substances. It should be kept in mind that SVHCs are identified based on a regulatory decision process in which available data is evaluated. Consequently, a negative model results (i.e. not structurally similar to a SVHC substance) does not necessarily means that the substance for instance has no carcinogenic, or persistent properties. What it does mean is that the chemical is not structurally similar to a SVHC and that related regulatory consequence may - at the moment - not be applicable for the new chemical.

A short guide on the application of the methodology is provided in Supplemental Material S.3. With respect to the applicability domain, an increase in reliability is observed with an increase in structure complexity for all three models, especially for the CMR model (i.e. number of atoms and different atom types). The structure similarity models are not applicable to arsenic, beryllium, cadmium, chromium, lead, mercury, nickel and cobalt-metal derivatives. For these chemicals, the metal atoms (or ions) are thought to be the cause of concern, irrespective of the (organic) groups present in the inorganic molecule. These metal-based complexes are by definition predicted to be SVHC substances. However, the models can be used to generate a first prediction for non-dissociating metals (e.g. organotin substances). In principle, the chemical similarity itself is an applicability domain descriptor. If the new substance is sufficiently similar to an existing SVHC, the substance is clearly within the applicability domain of the model. Furthermore, physicochemical boundaries (i.e. 95th percentiles) have been calculated for the different models based on molecular weight, log K_{ow}, number of atoms, number of rings and the number of aromatic rings (Table S.2). The similarity methodology does not discriminate between pristine substances or environmental and/or metabolic breakdown products; this model is applicable to both. Risk assessors, we therefore advise not only to apply the predictive model to the parent substance, but also to the breakdown products as well as possible tautomers, as these may give different similarity outcomes.

This effective screening method can particularly be applied during product development and chemical synthesis. By enhancing attention on chemicals of potential SVHC concern as early as possible within regulatory frameworks and safe-by-design trajectories, this methodology

contributes to the transition towards a non-toxic environment.

2.5 Conclusions

Within this study, a systematic and transparent methodology was established that could identify potential SVHCs based on structural similarity to a known set of SVHCs. We have analyzed the influence of selected similarity characterizations (fingerprints and coefficients) on the identification of chemicals of potential SVHC concern. A good statistical performance was obtained for CMR, PBT/vPvB and ED substances, but nevertheless further work is considered necessary to improve the ED part due to the small reference dataset for this SVHC concern.

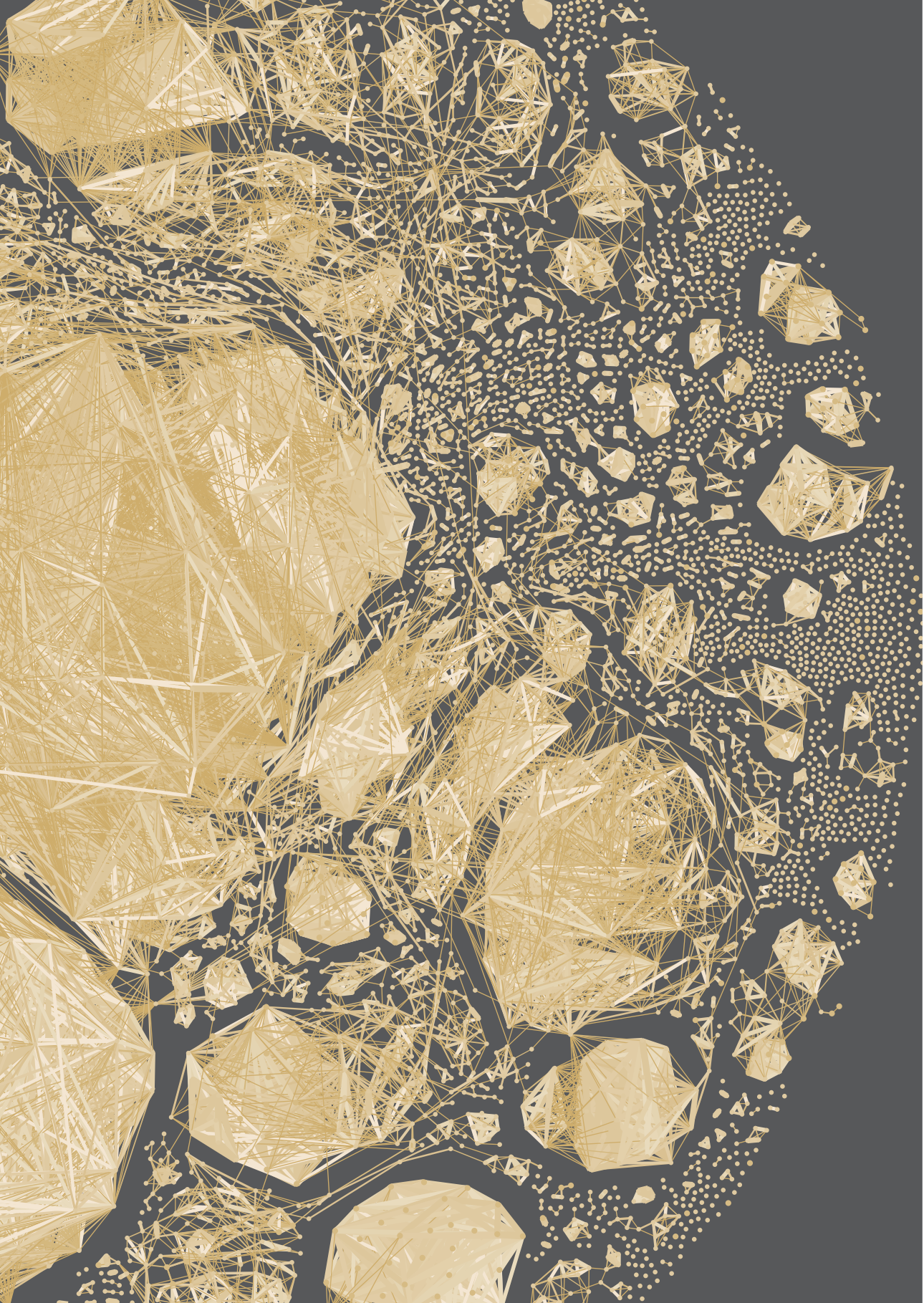
Application of the developed methodology is considered useful to identify chemicals of potential concern as early as possible, and as such may ensure that up-front more adequate risk management measures can be applied to contribute towards a non-toxic environment. It is foreseen that this scientifically-based model is beneficial to (environmental) risk assessors, industrial partners and academia.

Acknowledgements

This work was partially funded by the Dutch Ministry of Infrastructure and Water Management.

Supplemental material

Supplementary data to this chapter can be found online at https://doi.org/10.1016/j.comtox.2019.100110.



3

Evaluating Chemical Similarity as a Measure to Identify Potential Substances of Very High Concern

Pim N.H. Wassenaar, Emiel Rorije, Martina G. Vijver and Willie J.G.M. Peijnenburg

Published in Regulatory Toxicology and Pharmacology 119 (2021), 104834.

Abstract

Due to the large amount of chemical substances on the market, fast and reproducible screening is essential to prioritize chemicals for further evaluation according to highest concern. We here evaluate the performance of structural similarity models that are developed to identify potential substances of very high concern (SVHC) based on structural similarity to known SVHCs. These models were developed following a systematic analysis of the performance of 112 different similarity measures for varying SVHC-subgroups. The final models consist of the best combinations of fingerprint, similarity coefficient and similarity threshold, and suggested a high predictive performance (≥80%) on an internal dataset consisting of SVHC and non-SVHC substances. However, the application performance on an external dataset was not evaluated.

Here, we evaluated the application performance of the developed similarity models with a 'pseudo-external assessment' on a set of substances (n=60-100 for the varying SVHC-subgroups) that were putatively assessed as SVHC or non-SVHC based upon consensus scoring using expert elicitations (n=30 experts). Expert scores were direct evaluations based on structural similarity to the most similar SVHCs according to the similarity models, and did not consider an extensive evaluation of available data. The use of expert opinions is particularly suitable as this is exactly the intended purpose of the chemical similarity models: a quick, reproducible and automated screening tool that mimics the expert judgement that is frequently applied in various screening applications. In addition, model predictions were analyzed via qualitative approaches and discussed via specific examples, to identify the model's strengths and limitations.

The results indicate a good statistical performance for carcinogenic, mutagenic or reprotoxic (CMR) and endocrine disrupting (ED) substances, whereas a moderate performance was observed for (very) persistent, (very) bioaccumulative and toxic (PBT/vPvB) substances when compared to expert opinions. For the PBT/vPvB model, particularly false positive substances were identified, indicating the necessity of outcome interpretation. The developed similarity models are made available as a freely-accessible online tool.

In general, the structural similarity models showed great potential for screening and prioritization purposes. The models proved to be effective in identifying groups of substances of potential concern, and could be used to identify follow-up directions for substances of potential concern.

3.1 Introduction

Worldwide, more than 350,000 chemicals and chemical mixtures are registered for production and use [3]. Due to this large amount of substances, screening and prioritization are essential in order to focus chemical evaluation on those chemicals of highest concern. Chemical regulations particularly aim to minimize exposures and emissions of chemicals with serious and irreversible effects on human health or the environment as much as possible. In Europe, this specifically includes substances that are carcinogenic, mutagenic or reprotoxic (CMR); persistent, bioaccumulative and toxic (PBT) or very persistent and very bioaccumulative (vPvB); or substances with an equivalent level of concern, like endocrine disruption (ED). Substances that meet specific criteria for these endpoints of concern are identified via a hazard-based approach as Substances of Very High Concern (SVHC) within the REACH regulation (Registration, Evaluation, Authorization and Restriction of Chemicals; EC/1907/2006). The ultimate aim is to substitute these substances by safer (non-regrettable) alternatives.

To facilitate the identification of potential (new-)SVHC substances, we recently developed a chemical similarity methodology that assesses whether a new chemical is structurally similar to a known SVHC [73]. A high resemblance in chemical structure might be an indication of comparable effects ('similar property principle' [30]), and therefore could be a trigger for further evaluation. The developed methodology is based on a Dutch list of SVHCs [51], which covers a broader range of chemicals than the EU-SVHC list under REACH, but are identified based on the same hazard criteria (see Supplemental Material S.1 for more details). The final models suggested a good performance on a dataset with non-SVHC and SVHC substances with CMR, PBT/vPvB and ED properties during 'internal' validation (balanced accuracies ≥ 80 %) and outperformed several well-known predictive models when applied to this dataset [73]. Accordingly, these results are promising for further application within screening and prioritization activities on potential SVHCs.

It should be noted that the developed similarity models were not evaluated on their application performance with an external dataset. Ideally, an external validation using a new set of SVHCs is conducted to further assess and evaluate the model's predictive performance. However, we are currently lacking datasets of new SVHCs and non-SVHCs. It is not possible to pre-classify substances as SVHC for external validation purposes, as SVHCs are identified after a regulatory decision process in which all available data are evaluated. Similarly, non-SVHC substances are challenging to assign, as many substances are not extensively evaluated on all SVHC endpoints (i.e. CMR, PBT/vPvB and ED). The limitation is that a proper external validation set can therefore only be developed in future, when new SVHC and non-SVHC substances are identified. To overcome this current limitation, we aimed to evaluate and assess the application performance of the developed similarity methodology on a large set

of substances via quantitative and qualitative analyses, using expert elicitation and group evaluations.

Within this study we applied the newly developed chemical similarity models to the list of all registered substances under REACH, for which we do not (yet) have specific knowledge on potential concerns. We compared the chemical similarity as computed by the models with expert judgement classifications in order to assess the developed similarity models. The use of expert opinions is not uncommon in the field of predictive toxicology. For instance, expert classifications have been used within the development of the widely applied biodegradation models Biowin3 and Biowin4 – where the entire training dataset is formed by expert elicitation [37] – and expert classifications are also applied within specific machine learning algorithms (i.e. active learning approaches) [74,75]. Furthermore, within this study, illustrations are given aimed to show the model's potential for screening purposes (including single-substances and groups of substances). Specific examples that are discussed include phenolic benzotriazoles and bisphenol analogues.

3.2 Methods

3.2.1 REACH dataset

To investigate and assess model applicability, a dataset consisting of all REACH registered substances was used. For these substances we did not evaluate specific knowledge on potential concerns, which might be available for a subset of these substances in their REACH registration dossiers. The REACH registered substances were extracted from the webpage of the European Chemicals Agency (ECHA) on registered substances ([76]; extracted on 17-05-2019). In total, this list consisted of 24,694 entries representing 22,180 unique substances with chemical names and CAS-numbers. Based on this information, SMILES were generated using the KNIME (v3.7) workflow as developed by Gadaleta et al. [77], which connects SMILES from different data sources to CAS-numbers and/or chemical names. For cases where multiple CAS-numbers were available per substance, only the first CAS-number was used for the KNIME input (n=439).

Following the first part of the KNIME workflow, all entries are divided in three groups: Maintained, Rejected and Manual check. A substance is maintained when the retrieved SMILES from different sources are consistent. Substances are rejected when the retrieved SMILES are highly discordant or information is totally missing, whereas a further assessment is necessary when some identical and different/missing SMILES are retrieved (i.e. a manual check). Upon manual check some concordant SMILES had to be retrieved from other datasets. The following data sources were considered consecutively: ECHA dissemination

site (https://echa.europa.eu/nl/search-for-chemicals; primary source of the substances), ChemicalBook (https://www.chemicalbook.com/; suggested by [77]) and Molbase (http://www.molbase.com/; suggested by [77]). When no SMILES was retrieved via the above-mentioned sources, google searches were performed. In addition, substances that could not be represented by a single SMILES were removed during the manual check. This included substances with chemical names that describe mixtures, chemical substances of unknown or variable composition, complex reaction products and biological materials (UVCBs; including petroleum, extracts, fatty acids, glycerides, hydrocarbons, oil, residues, resins and rosins), reaction masses, reaction products (including products) or polymers [77]. Furthermore, ionic substances that have large (organic) counter ions were excluded as they cannot be represented by a single structure.

The information of the manual check is used in the second part of the KNIME workflow. This results in a list of maintained substances with corresponding CAS-numbers and SMILES and a list of rejected substances for which no reliable SMILES could be (automatically) retrieved. Subsequently, all substances that are on a Dutch list of Substances of Very High Concern ([51]; extracted on 01-03-2018) were excluded from the maintained substances, as those substances are in the training dataset of the structural similarity models (see section 3.2.2).

3.2.2 Structural similarity screening

Subsequently, the dataset was screened with the structural similarity models as described by Wassenaar et al. (2019). Within these models, the structure of a chemical is compared to known CMR, PBT/vPvB and ED substances included on a Dutch list of SVHCs ([51]; extracted on 01-03-2018; Table S.1). This list covers a broader range of chemicals than the EU-SVHC list under REACH, but are identified based on the same hazard criteria as the EU-SVHC substances (i.e. REACH article 57 [12]). The generation and composition of this list of substances is more elaborately described by Wassenaar et al. [73] and in the Supplemental Material S.1. Throughout the text, these substances are referred to as SVHCs.

Within the structural similarity models, first a fingerprint is generated for a substance based on its chemical structure. Secondly, the similarity of the fingerprint to the fingerprints of all SVHCs is expressed using a similarity coefficient. This results in similarity values to all SVHC substances, ranging from 0 (i.e. structures are considered as totally different) to 1 (i.e. structures are considered as identical). Thirdly, the similarity values are compared to a similarity threshold (i.e. a specific value between 0 and 1). Above the threshold, the substance is considered to be sufficiently structurally similar to assume comparable toxicological effects/concerns. The type of fingerprint, the similarity coefficient and the threshold applied in the structural similarity models, were determined in an optimization process (Wassenaar et al.

2019), and differ for the various SVHC-subgroups (Table 3.1).

The results of the structural similarity models were visualized within chemical similarity networks using Gephi (v0.9.2) [4] for the different subsets (CMR, PBT/vPvB and ED). Within the similarity networks, only chemical similarities above the model threshold values were included (see Table 3.1). In addition, Gephi was used to visually cluster the substances according to 'Modularity Class' following the algorithm of Blondel et al. [78].

Table 3.1. Overview of the characteristics of the structural similarity models [73].

Subset	Model		Threshold	Number of substances	Balanced accuracy	Sensitivity	Specificity	Precision
	Fingerprint	Coefficient						
CMR	CDK Extended	CT4 (<85*) SM (≥85*)	0.851 0.944	306	0.80	0.65	0.95	0.90
PBT/vPvB	MACCS	SM	0.970	209	0.95	0.92	0.98	0.96
ED	FCFP4	SS3	0.866	52	0.99	0.98	1.00	1.00

** A different similarity coefficient is used in the CMR similarity model for substances that have less than 85 fragments identified in the fingerprint (<85) and substances with 85 or more fragments identified in the fingerprint (≥85). CMR = carcinogenic, mutagenic or reprotoxic substances; PBT/vPvB = persistent, bioaccumulative and toxic / very persistent and very bioaccumulative substances; ED = endocrine disrupting substances. The MACCS and CDK Extended fingerprint are generated using PaDEL-Descriptor [79] and the FCFP4 fingerprint (i.e. Functional-Class Fingerprints with a diameter of 4) with RDkit using Morgan fingerprints [55]. Names of the coefficients are provided as in accordance to [7]: CT4 = Consonni-Todeschini 4; SM = Simple Matching; SS3 = Sokal-Sneath 3.*

3.2.3 Expert elicitation

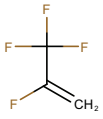
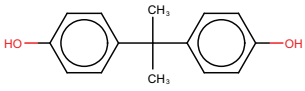
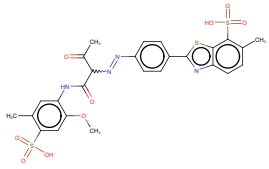
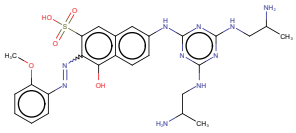
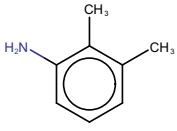
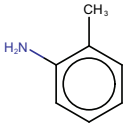
The results as computed by the chemical similarity models were compared with scorings performed by a group of chemists/toxicologists. First, a pilot phase with four experts was conducted to optimize the exercise of scoring chemical pairs. In addition, the results of the pilot phase were used to perform a power analysis, in order to provide an indication of the minimum number of experts necessary for the expert judgement survey in the assessment phase (see Supplemental Material S.2 for more details on the pilot phase and power analysis). Subsequently, in the assessment phase, a survey – consisting of non-SVHC/SVHC-pairs – was distributed among a group of participants working in the field of toxicology. A list of substance-pairs was provided to each expert, consisting of a chemical with unknown SVHC properties (taken from the REACH dataset) and the most similar SVHC (according to the chemical similarity model; either with a similarity above or below the threshold). Two questions related to toxicological and chemical similarity were asked for each substance-pair (see Table 3.2 for an example):

1) Toxicological similarity: *Do you expect similar toxicological effects/concerns for the unknown chemical based on chemical similarity, when compared to the chemical of known toxicological concern?* The scoring was a binary answer 'Yes' or 'No' executed by 30 participants.

2) Chemical similarity: *To which extent do you consider the two substances as structurally similar?* The scoring was based on a 5-point Likert scale [80] executed by 10 participants (a guide for scaling has been provided to the experts; see Supplemental Material S.2).

The results of the assessment phase were used to provide a statistical assessment of the ability of the structure similarity based computational models to reproduce the consensus expert elicitations regarding toxicological effects/concerns.

Table 3.2. An example of the expert judgement exercise. Substance-pairs consisting of a substance with unknown properties and the most similar SVHC were provided to the experts, who had to answer a question on toxicological concern ('yes/no'-score) and a question on chemical similarity (5-point Likert scale [80]: 1 = strongly disagree, 2 = disagree, 3 = neither agree/disagree, 4 = agree, 5 = strongly agree). During the assessment phase, both questions were asked separately.

Substance with unknown properties	**SVHC**	**1) Similar concern expected?**	**2) Structurally similar?**
		Yes/No	**5-point Likert scale**
		No	1 *(strongly disagree)*
		No	3 *(neither agree/ disagree)*
		Yes	5 *(strongly agree)*

Assessment phase

The assessment phase consisted of 256 substance-pairs, consisting of a substance with unknown properties from the REACH dataset and its most similar SVHC according to the similarity models. In total, 96 substance-pairs represented the CMR model(s), 100 the PBT/vPvB model, and 60 the ED model (see Supplemental Material S.3 for more details on the selection). The inclusion of substances-pairs was according to stratified random sampling based on computer-generated similarity values. In other words, the REACH dataset was divided in bins of similarity scores from which one or multiple substance-pairs were randomly selected. Substance-pairs were selected in such a way to ensure balanced groups of similar and non-similar substance-pairs (i.e. above and below the similarity model thresholds, Table 3.1). The procedure used for selection of substance-pairs for the CMR-CT4 model differed slightly from the described procedure (see Supplemental Material S.3). The substance-pairs were provided in random order to the experts, without showing the computer calculated similarity values in order to avoid any influence on the expert opinions. In addition, one random substance-pair of the CMR, PBT/vPvB and ED subset was included three times in order to investigate the consistency in scoring.

Two groups of participants were requested to fill in the survey. One group consisted of toxicologists in training and were only requested to answer the first question (i.e. toxicological concern) for the 256 substance-pairs (n=20 experts). All participants in this group have a background in chemistry and/or toxicology and are working in a related field as risk assessor, researcher or PhD-candidate (either in academics, government or industry). The other group consisted of direct colleagues, including experts in CMR, PBT and ED-assessments (n=10 experts). This group was requested to answer both questions (i.e. toxicological concern and chemical similarity) for all substance-pairs separately (i.e. 512 questions in total). Besides the scoring of the substance-pairs, participants were also requested to score their own expertise with respect to toxicity assessment and/or knowledge of molecular structures (range of 1-10). For both groups, three versions of the survey were generated with a random order of substance-pairs and different order of the models (e.g. first non-SVHC/CMR pairs followed by non-SVHC/ED pairs, etc.). In addition, four experts, which were also involved in the pilot phase, filled in the survey for a second time, two to three months after their first scored submissions, in order to investigate the scoring consistency over time.

Data analysis

The performance of the models is analyzed by comparing the predictions by the models to the predictions by a group of experts. In other words, we analyze whether the similarity models do highlight those chemicals that would also be selected as substances of potential concern by a group of experts. Within this study, the use of expert elicitation thus considers a direct expert response on chemical similarity and related concerns, and cannot be considered

as an extensive (expert) evaluation of all available data on a specific chemical. The use of expert opinions is particularly suitable as this is exactly the intended purpose of the chemical similarity models: a quick, reproducible and automated screening tool that mimics the expert judgement that is frequently applied in various screening applications.

The results of the binary scoring on toxicological similarity were analyzed by using a confusion matrix (i.e. expert judgement vs similarity model prediction), in which the expert judgement scores were considered as the 'true'-effects. Following expert judgement scoring, a substance with unknown SVHC-effects was considered as potential SVHC based on majority voting (i.e. >50% 'yes'-score). Majority voting on chemical similarity is regularly applied in several setting, e.g. at the European Medicines Agency [81]. Within this analysis, the predictions of assessors from both groups were combined. The results of the scoring of chemical similarity using the 5-point Likert scale, were used to analyze the relation between toxicological similarity (i.e. the results from the first question) and chemical similarity. In addition, the chemical similarity scores of the participants were compared to the computer-generated similarity values. When an expert did not answer the toxicological or chemical similarity question for a specific substance-pair, the observation was excluded from analysis (n=14 and n=3, respectively). All data was analyzed in R (v3.6) [58].

3.2.4 Illustrative cases

The results of the model application to the REACH dataset and the results of the expert scoring exercise were also analyzed and interpreted qualitatively. Specific groups of substances from the dataset are highlighted as illustrative cases, in order to indicate the potential of the models for screening purposes. In addition, specific limitations of the screening models based on chemical similarity are identified and discussed.

3.3 Results

3.3.1 REACH dataset

A reliable unique SMILES could be assigned to 9593 chemicals out of the 22,180 REACH registered substances. In Table 3.3 an overview of the physicochemical and structure properties of those substances is provided, as well as for the SVHC subsets as used in the similarity models.

Table 3.3. Physicochemical and structural properties of the substances in the REACH dataset and the expert elicitation dataset. In addition, the properties of the SVHCs as included within the structural similarity models are provided for the different models (i.e. CMR, PBT/vPvB and ED). The ranges represent the 2.5th and 97.5th percentiles of the properties. Log K_{ow} was predicted according to EpiSuite [37].

Properties	REACH dataset	Expert elicitation dataset	Structural similarity models		
			CMR	PBT/vPvB	ED
Number of substances	9593	256	306	209	52
Molecular weight	86 – 740	90 – 834	50 – 686	156 – 734	192 – 549
Log K_{ow}	-2.58 – 10.77	-2.1 – 12.45	-1.59 – 10.41	3.26 – 11.18	3.53 – 6.24
Number of atoms (incl. H)	10 – 98	10 – 102	6 – 84	10 – 81	33 – 92
Number of rings	0 – 6	0 – 6	0 – 6	0 – 6	1 – 6
Number of aromatic rings	0 – 5	0 – 6	0 – 6	0 – 6	1 – 2

3.3.2 Structural similarity screening

The chemical structures of the REACH dataset were compared to the SVHC substances by using the similarity models. Of the 9593 REACH substances, 1485 (15.5% of total) were considered to be sufficiently structurally similar to at least one CMR-SVHC and therefore predicted to be potential CMR substances. Of those 1485, 883 had less than 85 fragments identified in the fingerprint and were compared with the CT4 similarity coefficient. This is 29.5% of all substances in the REACH dataset with less than 85 fragments identified in the fingerprint. The other 602 substances were predicted as potential CMR according to the SM similarity coefficient, which is 9.1% of all the REACH dataset substances with 85 or more fragments identified in the fingerprint. The PBT/vPvB similarity model considered 533 substances as sufficient structurally similar to classify as a potential PBT/vPvB-SVHC (5.6%). For two substances of the REACH dataset, the MACCS fingerprint could not be generated, and thus no comparison could be made to PBT/vPvB-SVHCs. According to the ED model, 113 substances (1.2%) were considered to be sufficiently structurally similar to an ED-SVHC.

3.3.3 Expert elicitation

Assessment phase

The results for the binary question (toxicological concern), indicated that 102 of the 256 substances are considered potential SVHC following majority voting, with an average 'yes-voting' of 74% (±14% standard deviation) (Table S.2). In total, 154 substances were not considered potential SVHC based on expert elicitations with an average 'yes-voting' of 24% (±15% standard deviation) (Table S.2). Results for the second question (chemical similarity), indicate that the moderate values (i.e. 2-4) are selected more frequently – in 76% of the cases – compared to the extremes (i.e. 1 and 5). In addition, the average spread (i.e. standard error) around the mean of the chemical similarity score is lower for the extremes (Figure S.1),

indicating slightly less variation among the experts. The individual scores as provided by the experts are shown in Figure S.2 and S.3 and summarized in Table S.2. On average, the experts scored their own expertise to toxicity assessment and/or knowledge of molecular structures at 7.1 (n=10 extended survey) and 5.5 (n= 20 short survey) out of a max of 10, respectively. No relation was observed between the expertise of the participants and their provided scores, and also not between the participants of the different groups (see Figure S.2 and S.3).

The relation between toxicological similarity (n=30 experts) and chemical similarity (n=10 experts) as assessed by the experts is shown in Figure 3.1. For all subsets – CMR, PBT/vPvB and ED – there is a clear relationship observed (R^2 ranging from 0.84-0.89), and indicates the importance of chemical similarity for toxicological concern. Based on this relationship, 50% of the experts expect a comparable toxicological concern for a substance-pair with an average chemical similarity of around 3 in the Likert scale used for chemical similarity.

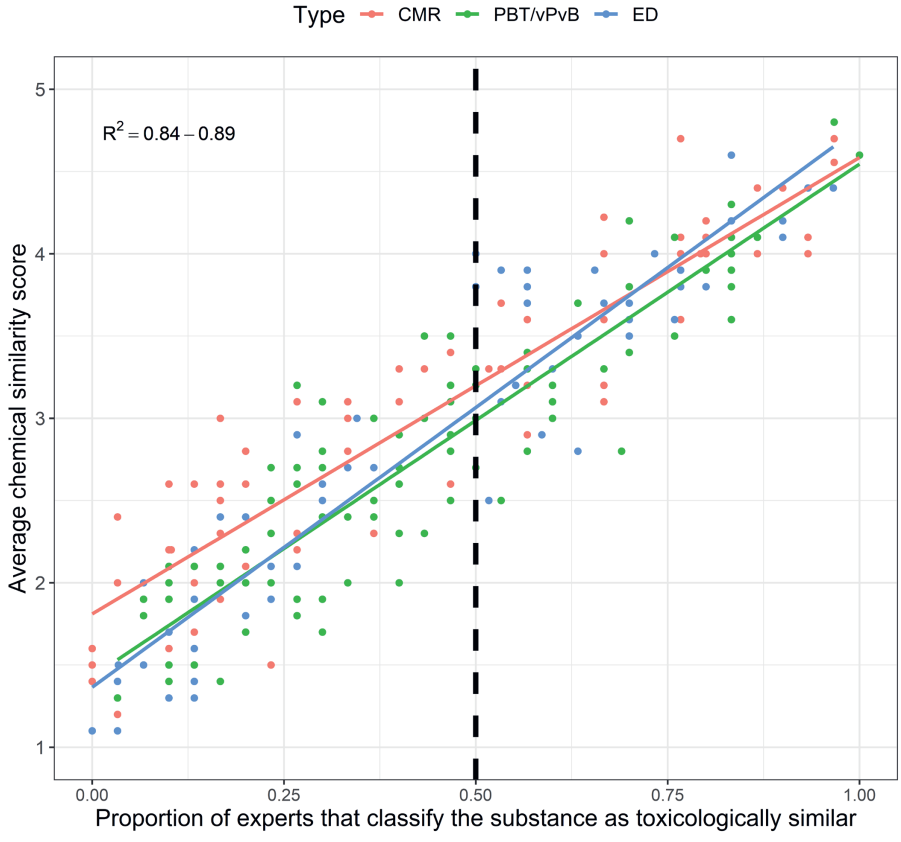


Figure 3.1. Relation between toxicological similarity (n=30 experts) and chemical similarity (n=10 experts) as assessed by the experts for the 256 substance-pairs. R^2 quantifies the goodness of fit.

Besides the average expert scores and the variation between experts, we also investigated the variation for a single expert by including a substance-pair three times within the CMR, PBT/vPvB and ED subset. With respect to the toxicological similarity (i.e. 'Yes' or 'No'), 73-83% of the assessors provided three times the exact same answer for the substance-pair, for the different subsets. In addition, in 37% of the cases the assessors provided three-times the same chemical similarity score for the substance-pairs. When they provided a different score, they varied on average with a score of 1.21. In addition, four experts repeated the full exercise two to three months after their first submission, in order to investigate the consistency over time. On average, the experts scored 83% of the substance-pairs similar as to their first application with respect to toxicological similarity. With respect to chemical similarity, the experts provided the same answer in 59% of the cases. In cases they provided a different score, they varied on average with a score of 1.16. Furthermore, most variation was observed around substance-pairs with higher uncertainty (i.e. substance-pairs with around 50% 'yes'-scores for toxicological similarity, or an average chemical similarity around 3).

Computer model performance compared to expert elicitation

Table 3.4 shows the expert scores for the binary question (toxicological concern) in comparison to the computer predictions (also visualized in Figure S.4). The scores from the experts are taken as the 'true'-values based on majority voting (i.e. > 50% 'yes'-scores). It can be observed that the similarity models follow the expert opinions with a balanced accuracy between 0.69 – 0.87. The performance of the CMR model – when compared to the expert judgement exercise – is comparable to the performance obtained from an internal ('training') dataset (see Table 3.4) [73]. The predictive performance of the ED model is slightly below the performance observed during internal validation, whereas the model application performance for the PBT/vPvB subset is much lower when predicting expert opinions. Specific examples of substances that are differently classified by the similarity models, when compared to the expert judgement scores, are discussed in section 3.3.4.

Table 3.4. Cooper statistics for the computer similarity model classifications of potentially SVHC or non-SVHC when compared to the majority opinion of a group of human experts. The reference balanced accuracy represents the model performance following internal validation as analyzed by Wassenaar et al. [73]. TP = true positives, FP = false positives, TN = true negatives, FN = false negatives.

Subset	Number of substance-pairs	TP	FP	TN	FN	Sensitivity	Specificity	Balanced accuracy	Reference balanced accuracy
Overall	256	85	43	111	17	0.83	0.72	0.78	-
CMR	96	37	11	41	7	0.84	0.79	0.81	0.80
PBT/vPvB	100	23	27	43	7	0.77	0.61	0.69	0.95
ED	60	25	5	27	3	0.89	0.84	0.87	0.99

3.3.4 Illustrative cases

The results of the structural similarity models, after application to the REACH dataset, were visualized by using chemical similarity networks (see Figure 3.2 and Figure S.5-S.6). Based on these networks, groups of substances can be identified that are all predicted to be sufficiently structurally similar to one or multiple existing SVHCs, and therefore can be considered as potential SVHC. Within this section, several specific groups are highlighted to indicate varying applications of the models. In addition, general and substance-specific model limitations – as apparent from the REACH dataset screening and expert elicitations – are discussed.

Group screening

To illustrate the use of the models for group screening, we here provide an example of an identified group consisting of phenolic benzotriazoles present in the REACH dataset (see Figure S.5). Four phenolic benzotriazoles are currently identified as PBT/vPvB and five REACH registered substances are considered to be structurally similar to those substances (see Table 3.5). In addition, we identified nine additional chemicals with the phenolic benzotriazole structure in the REACH dataset, which are not considered to be sufficiently structurally similar to the four SVHC phenolic benzotriazoles according to the similarity model.

All phenolic benzotriazoles that are considered structurally similar to a known SVHC, meet the P-screening criteria, and are close to or above the B-screening criteria [19,37] (Table 3.6). Follow-up PBT/vPvB analyses are already being conducted within the REACH framework for three of these five substances (Table 6). Considering the chemical – and potentially the biological – similarity between these substances, this could be a trigger for further evaluation or assessment of this specific group of substances.

Furthermore, besides PBT/vPvB concern, the ED structural similarity model identifies substance nr.7 (in Table 3.5-3.6) as structurally similar to an ED substance (i.e. 4-(1,1,3,3-tetramethylbutyl)phenol). This type of additional triggers, obtained via the similarity models, could lead to new hypotheses on chemical effects/concerns that could be further investigated.

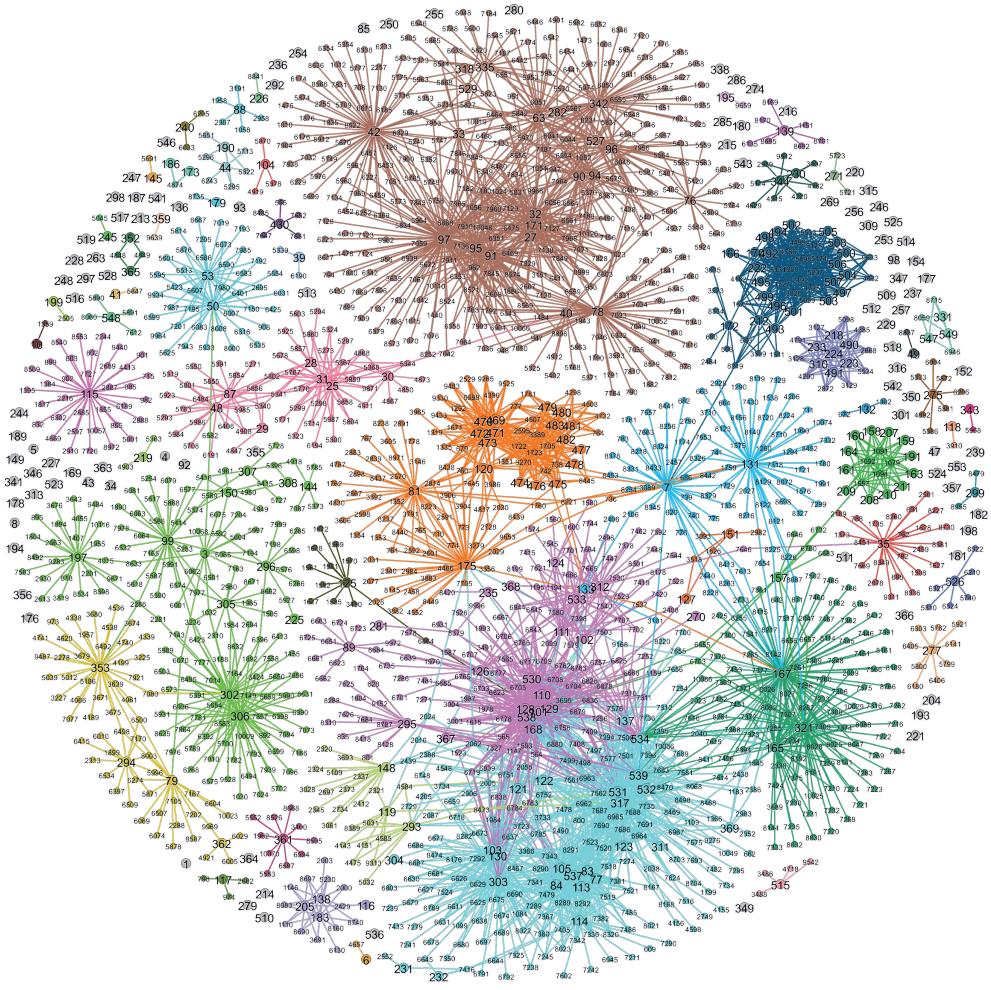


Figure 3.2. Chemical similarity network of the REACH dataset substances and CMR-SVHC substances. The chemical similarity network was generated using Gephi by using Fruchtmann-Reingold layout and the similarity thresholds of the structural similarity model (see Table 3.1). Each node represents a chemical with corresponding ID-number. The numbers with a large font size represent the ID-numbers of CMR-SVHC substances (Table S.1) and the numbers with smaller font size represent the ID-numbers of REACH dataset substances. The lines represent a chemical similarity (i.e. similarity value above the model threshold) between the REACH dataset substance and a CMR-SVHC substance as predicted by the structural similarity model. The length of the lines (i.e. the distance between two connected nodes) does not represent the extent of similarity (i.e. the height of the similarity values as predicted by the similarity model). The colors represent clusters of substances that are considered to be structurally similar to the same SVHC substance(s). The clusters are predicted by Gephi using 'Modularity Class'. Some examples of classes: Brown – small organic oxygen compounds (e.g. nr. 32); Orange – phenols (e.g. nr.175), Blue – aromatic amines and nitro-aromatic compounds (e.g. nr. 303 and 317); Green – polycyclic aromatic hydrocarbons (e.g. nr. 158); Dark blue and purple – phthalates (e.g. nr. 224 and 501); Light pink – small chlorinated and brominated organic compounds (e.g. nr. 25); Light green – diphenyl methane-backbones (e.g. nr. 293).

Table 3.5. Example of a group of phenolic benzotriazoles present in the REACH dataset (see Figure S.5). Four phenolic benzotriazoles are classified as PBT and/or vPvB. Five phenolic benzotriazoles in the REACH dataset are considered structurally similar to these SVHC by the PBT/vPvB structural similarity model, whereas nine phenolic benzotriazoles are not considered to be sufficiently structurally similar by the model.

SVHC	REACH substance *Structurally similar*	REACH substance *Not structurally similar*
1)	5)	10)
2)	6)	11)
3)	7)	12)
4)	8)	13)
	9)	14)
		15)
		16)
		17)
		18)

Table 3.6. Additional substance information for the group of phenolic benzotriazoles as depicted in Table 3.5 (linked based on ID number). REACH-ID represents the number of the substances as depicted in Figure S.5. Substance properties were predicted with EpiSuite [37].

ID	REACH-ID	CAS	PBT/vPvB SVHC status	Degradability (Biowin2 v4.10)	Degradability (Biowin3 v4.10)	Log K_{ow} (KOWWIN v1.68)
1	51	25973-55-1	PBT/vPvB	**0.011**	**2.05**	**7.25**
2	52	36437-37-3	vPvB	**0.139**	**2.25**	**6.31**
3	86	3846-71-7	PBT/vPvB	**0.016**	**2.12**	**6.27**
4	82	3864-99-1	vPvB	**0.001**	**1.83**	**6.91**
5	4989	70321-86-7	- [1,2]	**0.092**	**1.89**	**7.67**
6	4822	73936-91-1	- [1]	**0.003**	**1.67**	**8.82**
7	3970	3147-75-9	- [1]	**0.016**	**2.12**	**6.21**
8	3071	3147-76-0	-	**0.168**	**2.45**	4.36
9	3321	3896-11-5	- [3]	**0.024**	**2.06**	**5.55**
10	2049	2440-22-4	- [4,5]	0.785	**2.68**	3
11	3072	92484-48-5	-	**0.187**	**2.56**	1.24
12	3502	96478-09-0	-	0.982	**2.61**	3.93
13	3757	84268-36-0	-	**0.197**	**2.58**	3.3
14	3952	84268-33-7	-	0.862	**2.33**	**4.94**
15	3971	3147-77-1	-	0.960	2.75	**5.97**
16	4588	107479-06-1	-	**0.057**	**1.85**	**6.13**
17	5475	103597-45-1	-	**0.000**	**0.93**	12.46
18	9684	84268-08-6	-	0.936	**2.47**	**7.39**

Conducted or ongoing activities/evaluations within REACH: 1) Undergoing PBT assessment. 2) Regulatory management option analysis (RMOA) on persistence and ED. 3) RMOA on CMR. 4) RMOA on persistence, human health and reprotoxicity. 5) Substance evaluation on sensitization (concluded: no follow-up).

Dissimilarity screening

The similarity models can also be used to evaluate dissimilarity. More specifically, it could be tested when a chemical is not considered to be sufficiently structurally similar to a known SVHC. Such evaluations are of particular interest to prioritize potential alternatives to known SVHCs (e.g. within safe-by-design development processes). Although the similarity models classify substances as potentially-SVHC versus non-SVHC, the difference between the similarity value and the established model thresholds (Table 3.1) could be an indication of the certainty for the classifications.

To illustrate the use for non-similarity screening, we here provide an example of bisphenol A (BPA) analogues. BPA is acknowledged as being a reprotoxic chemical with ED properties (nr.1 Figure 3.3) [82]. In Figure 3.3, a sequence of BPA analogues is shown with slight changes in the BPA structure from chemical nr.2 up to larger changes in structure nr.8. For all those structures the chemical similarity to BPA is analyzed using both the CMR and ED structural similarity models. According to these models, substance nr.7 (tetramethyl bisphenol F; TM-BPF) and substance nr.8 (tetra(tertbutyl) bisphenol F; TTB-BPF) are considered to be

structurally too dissimilar to BPA to be classified as potential SVHC, and could potentially be given higher priority within a safe-by-design development process. TM-BPF seems to show lower estrogenic activities compared to BPA [83], though further data analysis on TM-BPF is still ongoing [84], in which also other modes of action need to be considered. For instance, BPA also shows anti-androgenic effects [85], which are also predicted for TM-BPF [86]. Furthermore, the ED properties of TTB-BPF are currently under investigation via a substance evaluation within REACH [87].

The absence of chemical similarity to BPA does not by definition mean no concerns. For instance, bisphenol S (BPS) is not considered to be structurally similar to BPA by the models (i.e. substance nr.9 in Figure 3.3). However, biological analysis indicates that BPS – although via different pathways – has the potential to interfere with the endocrine system [88]. Therefore, prioritization of alternatives ideally consists of a combination of chemical similarity with biological similarity, as for instance conducted by the NTP for several bisphenols [88]. Within such evaluations, also chemical similarities to other SVHCs should be considered in order to prevent regrettable substitution. For instance, when BPS (substance nr.9 in Figure 3.3) will be identified as ED-SVHC in future, substance nr.10 will be considered as a potential SVHC by the model (when applying current threshold values).

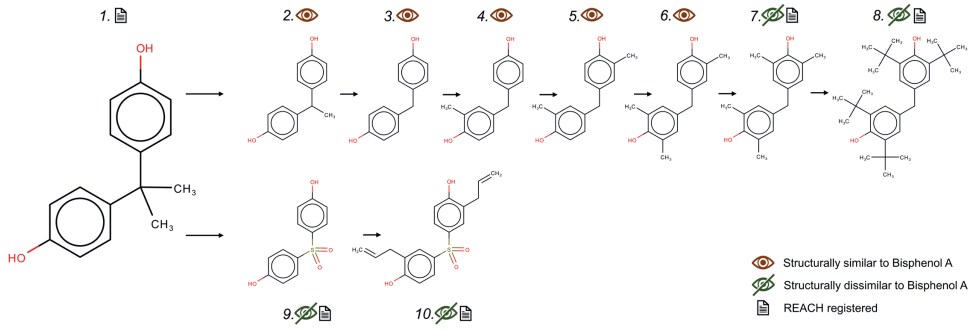


Figure 3.3. Structural similarity of several bisphenol A (BPA) analogues to BPA according to the CMR and/or ED models. Model thresholds are 0.944 and 0.866 for the CMR and ED model, respectively (Table 3.1). 1 = BPA (CAS: 80-05-7); 2 = bisphenol E (CMR model 0.99, ED model 0.89); 3 = bisphenol F (BPF; CMR model 0.98, ED model 0.87); 4 = methyl-BPF (CMR model 0.95, ED model 0.82); 5 = dimethyl-BPF (CMR model 0.95, ED model 0.80); 6 = trimethyl-BPF (CMR model 0.95, ED model 0.79); 7 = tetramethyl-BPF (CAS: 5384-21-4; CMR model 0.94, ED model 0.75); 8 = tetra(tertbutyl)-BPF (CAS: 118-82-1; CMR model 0.94, ED model 0.76) ; 9 = bisphenol S (CAS: 80-09-1; CMR model 0.89, ED model 0.86); 10 = 2,2'-diallyl-4,4'-sulfonyldiphenol (CAS: 41481-66-7; CMR model 0.84, ED model 0.76). Note that the shown examples are a subset of registered BPA analogues.

Interpretation of model results

As the results of the models are solely based on overlap in chemical structure, they should be interpreted and weighed accordingly for follow-up assessment, as model predictions might be false positives or false negatives. Although a low amount of false classified substances was identified for the CMR and ED models, especially the number of false positives for the PBT/vPvB subset – when compared to expert solicitation – indicates the necessity of interpretation.

For instance, chemical similarity does not mean that there is always a hazard concern. It is possible that a chemical has a high similarity with a SVHC, but that the specific functional group causing the concern is missing (see Table 3.7, Chemical nr.1 – the aromatic amine is the reason for the carcinogenic effects). On the other hand, absence of similarity does also not guarantee absence of toxicological concern. A chemical could exert specific effects via different mechanisms than the currently known SVHCs, or the structural overlap might just be too low according to the model (see Table 3.7, Chemical nr.2).

Nevertheless, our earlier work showed that the structural similarity model for identifying carcinogenic/mutagenic SVHCs performs better than a well-known structural-alert screening model applied to the same dataset [73]. This indicates the relevance of full chemical overlap for toxicity prediction, and might be explained by a closer relationship with partitioning properties of substances (as also illustrated for the phenolic benzotriazole backbone in the previous section). In addition, other fragments present in the substance may function as a (steric) shield, adjusting the stability or reactivity of specific fragments/substances [89].

Specific model limitations were identified upon application to the REACH dataset, and upon comparison to the expert judgement scores. Some substances that were classified differently by the structural similarity models and the expert pool are shown in Table 3.7. In cases where the computer model predicts non-SVHC, whereas the experts see a toxicological concern, the substances generally have several functional fragments in common with the SVHC substance, that – according to the expert – are potentially responsible for the effects. However, the models do not regard them as similar, as their total structural overlap is considered insufficient. This is for instance due to differences in the linkage of atoms or the presence of other functional groups (e.g. Chemical nr.3-5, Table 3.7).

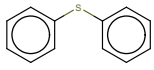
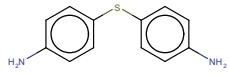
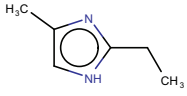
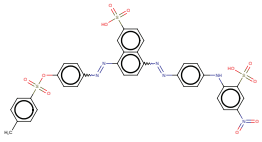
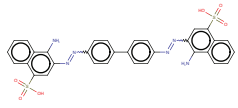
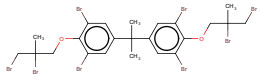
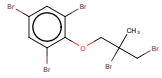
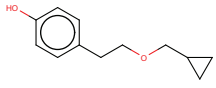
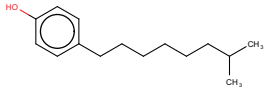
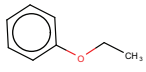
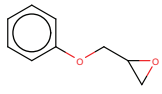
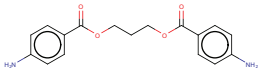
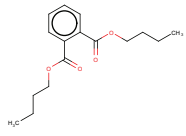
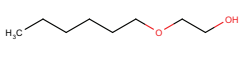
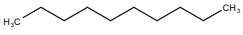
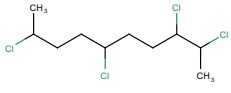
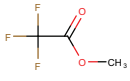
The cases for which the model predicts SVHC and the experts see no concern, differ per model. For the CMR-CT4 and ED model, these substances generally miss a specific fragment that is considered important for the concern by the experts (e.g. Chemical nr.6-7, Table 3.7). For the CMR-SM model, and partially also for the CMR-CT4 model, also other differently classified substances are identified, that are related to the presence and absence of ring-structures. Due to the use of the CDK Extended fingerprint in the CMR model – which is a path-based

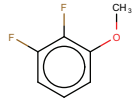
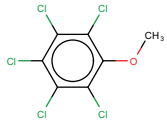
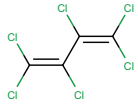
fingerprint – not many additional fragments are identified for substances with a straight-chain of (carbon) atoms or when these atoms are structured in a ring. Consequently, substances with this kind of variation (i.e. linear versus ring) could be considered as similar by the model, whereas this is not perceived as similar by the experts (e.g. Chemical nr.8, Table 3.7). For the PBT/vPvB model on the other hand, additional differences in classifications are related to the disregard of counts of specific fragments (e.g. counts of halogen substituents – multiple halogens make a substance more PBT-like; counts of aromatic structures – polyaromatic hydrocarbons are considered PBT/vPvB, where monoaromatic hydrocarbons are normally not PBT/vPvB, etc.; e.g. Chemical nr.9-10, Table 3.7). Furthermore, the type of halogen (i.e. F, Br, Cl, I) is not considered within the fragments as defined by the MACCS fingerprint. The type of halogen is regularly considered as important for the PBT/vPvB properties by the experts (e.g. Chemical nr. 11, Table 3.7), but is not always decisive (e.g. Chemical nr.12, Table 3.7). As a consequence of the underlying methodology (i.e. the applied fingerprint), a lower balanced accuracy is observed for the PBT/vPvB model when predicting the expert elicitation results. The abovementioned classification errors were also observed in the different clusters of REACH dataset substances visualized in Figure 3.2 and Figure S.5-S.6.

3.4 Discussion

The goal of this study was to investigate the application performance of the newly developed structural similarity models [73] on the broader universe of chemicals. As currently no external validation set could be developed based upon toxicological studies and regulatory decisions, we used expert judgement scores regarding the toxicological similarity between known SVHCs and chemicals with unknown SVHC properties, to derive a pseudo-external validation set. The use of expert opinion was particularly suitable as the ultimate goal of the computer similarity models is to provide an automated, fast and reproducible alternative to expert opinion, as expert consultation requires much more time, manpower and therefore money. Based on our analyses, comparable performance statistics were observed for the CMR model (balanced accuracy of 0.81 reproducing the expert elicitation), when compared to the predictive performance previously determined during internal validation (balanced accuracy of 0.80) [73]. For the ED model a relative high balanced accuracy was observed (0.87) reproducing the expert elicitation (compared to a balanced accuracy of 0.99 during internal validation), whereas a moderate balanced accuracy was observed for the PBT/vPvB model (0.69, as compared to a balanced accuracy of 0.95 during internal validation). In addition, we provided several examples for application and result interpretation of the models.

Table 3.7. Examples of substance-pairs that are differently classified by the computer similarity models when compared to the expert elicitation results. The first two substance-pairs are illustrative cases, which were not part of the expert elicitation.

ID	REACH substance	Most similar known SVHC	SVHC subset	Model prediction (sim. value)	Expert prediction (proportion [1])
1			CMR (CT4)	SVHC (0.905)	- [2]
2			CMR (SM)	Non-SVHC (0.925)	- [3]
3			CMR (SM)	Non-SVHC (0.794)	SVHC (0.933)
4			PBT/vPvB	Non-SVHC (0.964)	SVHC (0.759)
5			ED	Non-SVHC (0.847)	SVHC (0.633)
6			CMR (CT4)	SVHC (0.906)	Non-SVHC (0.433)
7			ED	SVHC (0.873)	Non-SVHC (0.2)
8			CMR (SM)	SVHC (0.964)	Non-SVHC (0.4)
9			PBT/vPvB	SVHC (0.97)	Non-SVHC (0.133)
10			PBT/vPvB	SVHC (0.976)	Non-SVHC (0.467)

11			PBT/vPvB	SVHC (0.988)	Non-SVHC (0.5)
12			PBT/vPvB	SVHC (0.988)	SVHC (0.567)

1) Proportion of experts voting that the substance is toxicologically similar to the SVHC. 2) Above similarity threshold, but no concerns for carcinogenicity as it does not contain aromatic amines. Currently there are no classifications or processes ongoing for the non-SVHC. 3) Below similarity threshold, but potentially CMR as under investigation for classification within REACH-CLP.

3.4.1 Application performance

Variation in expert elicitation

The expert scores showed a clear relationship between chemical similarity and toxicological similarity (Figure 3.1). This indicates that, in general, a high chemical similarity is expected to be related to comparable toxic concerns. Accordingly, the key assumption of the structural similarity models seems valid, and is reproducing the assumptions used by toxicological experts. In addition, the relation between the expert SVHC predictions and the data as available within REACH is illustrated in Supplemental Material S.4 for several substances in the dataset.

When looking in more detail to the expert scores, it can be observed that there is some variation in scores across experts (Figure S.2-S.3). Variation between expert similarity scores has been observed and described in earlier studies, and is suggested to be related to intuition, perception and experience of the assessor [31,81,90,91]. Intuitively and unconsciously, assessors reduce the complexity of chemical structures and score chemical similarity based on only a few structural features or patterns that are perceived as most essential [31,92]. The essentiality of the structural features or patterns may differ per individual based on their scientific experience [31,81]. In addition, it has been suggested that the alignment of chemicals as provided to the experts (e.g. in which rotation/angle the non-SVHC and SVHC structures were shown) may influence the perception of similarity across experts to a different extent [81]. Also, the similarity scale may not be interpreted in a uniform manner by the different assessors [31]. Although measures have been taken to provide a guide for scaling, slight differences in applied scales were also observed in this study (Figure S.3; i.e. some assessors only provided scores in the range of 2-4). Furthermore, it has been suggested that similarity scorings are context-dependent (i.e. dependent on the order of substances in the survey) [31]. Therefore, we provided the experts with different (random) orders of the substances and subsets. Despite the variation among experts, the results indicate that averages

are clearly related to chemical similarity. Therefore, the group averages can be considered as much more valuable than single expert scores, as in line with previous conclusions (the wisdom-of-crowds principle) [91,93], and were therefore applied in this research.

Besides variation among experts, variation within the scores of a single expert are expected. In order to also quantify the variation within expert similarity scores (e.g. signs of fatigue or training effects), we tested the internal consistency during the assignment and the consistency over time. Both analyses showed comparable results, with ~80% of the assessors providing consistent 'yes/no'-scores with respect to toxicological similarity, and ~50% with respect to chemical similarity (1-5). In cases experts provided a different chemical similarity score, they varied on average with a score of 1 on the Likert-scale. Accordingly, the experts could be considered relatively consistent. In addition, the amount of variation in the experts scores – and particularly the 'yes/no'-scores – does not merely represent a confounding factor, it also reflects the (un)certainty of the toxicological similarity for each substance-pair. The amount of (un)certainty in the 'yes/no'-scores seems to be clearly related to the computer-generated similarity values, with less uncertainty for extreme similarity values (see Figure S.4).

Performance considerations

The CMR- and the ED-similarity models were able to predict expert opinions to a relative high extent (balanced accuracy >80%). The results of these models might be considered even more robust than single expert opinions, as the computer model consistently derives structural features from substances and systemically calculates chemical similarity, without applying biased or context-dependent deviation. The PBT/vPvB model predictions, on the other hand, only resemble the expert elicitations to a moderate extent.

Differences in substance classifications – between the models and the experts – were particularly related to the absence or presence of a specific functional fragment, as several fragments are being related to a specific effect [39,94]. Although the absence or presence of a single fragment could influence the toxicological concern, total chemical similarity may not be affected significantly, and therefore could result in different classifications between the experts and the models [50]. Vice versa, the absence or presence of a functional fragment does not necessarily mean that a specific effect will occur. The advantages of using full chemical overlap over structural alerts, is the closer relation to partitioning properties of substances and therefore also potentially to toxicokinetic and toxicodynamic processes. This is also reflected in the predictive performance of the models, according to internal and external validation, and suggests that equal treatment of mechanistically relevant and irrelevant fragments by the similarity models may not be a huge problem in practice. Acknowledging the limitations, these models show great potential to be applied in screening and prioritization approaches.

3.4.2 Advances in screening and prioritization

Within the risk assessment of chemicals, there is a general transition from substance-by-substance assessments to group assessment approaches, in which assessments for some individual substances could be made based upon their similarity to other tested chemicals within the group (read-across) or based on simple trends observed within the group [27,29]. As illustrated in Figure 3.2 and section 3.3.4, the structural similarity models could be used to identify relevant groups of chemicals that are structurally similar to one or more SVHCs. Such groups could be selected for further evaluation, in which then also biological similarity needs to be considered, including bioavailability, degradation, bioaccumulation, physicochemical properties and toxicity [26]. In addition, the structural similarity models could also be used to fine tune read-across in groups that are predefined based on their biological mechanism, as proposed by Mellor er al. [50] and illustrated in Figure 3.3.

Grouping of chemicals is of particular interest in terms of effective use of available information, thereby aiming to reduce animal testing and potentially speeding up risk assessment and management, and, ultimately, increasing the level of protection for human health and the environment [27,29]. In addition, group regulations could prevent regrettable substitution to a close structural analogue with comparable technical functioning and toxic properties [95].

Currently, several group prioritizations and evaluations are already being conducted by ECHA, based on their in-house screening methodology [29]. Examples of concluded, ongoing or test-cases for group evaluations include non-branched aliphatic fatty acids [29], per- and polyfluoroalkyl substances (PFAS) [96], organotin compounds and polyol acrylates [97]. Within their recent report, ECHA highlights the importance of further optimization of the screening of groups of substances [29]. The developed similarity models could potentially contribute to such advancements. In addition, to further contribute to current ongoing activities on the identification of groups of chemicals of high concern, the screening of emission and monitoring data is highly encouraged. Specifically, because this kind of data provides different insights in substances of potential concern, as not all substances that are emitted to the environment are registered within specific legislations.

In addition, the developed structural similarity models could help define follow-up directions. As the models analyze chemical similarity to known SVHCs, the specific concern of the most similar SVHC(s) provides a relevant trigger and direction for follow-up analysis. Information on the specific concern could be combined with integrated testing and assessment strategies (ITS) – as included in several regulatory guidelines [19] – to define directions for further analysis. For example, when the most similar SVHC is considered a mutagen based on point-mutations, results of relevant Ames tests could be evaluated first. When a SVHC is

considered as PBT with specific toxicity to algae, aquatic algae tests could be given higher priority compared to invertebrate or fish toxicity data.

3.4.3 Notes on application and future recommendations

The methodology as analyzed and evaluated in this research is made available in the form of a web-based tool at https://rvszoeksysteem.rivm.nl/ZzsSimilarityTool or in the form of a R-script in the supplemental material of Wassenaar et al. (2019). The tool should be applied as a first screening model and could help in prioritization and grouping of substances. It should be noted that similarity to the broader list of Dutch SVHC is investigated (extracted on 01-03-2018), rather than the smaller list of EU-SVHC under REACH (see Wassenaar et al. [73] and Supplemental Material S.1 for more details). Furthermore, it should be highlighted that several SVHCs that are classified as such based on an 'equivalent level of concern', are not yet included in the models. This includes substances that are considered SVHC based on sensitizing properties, SVHCs with specific target organ toxicity after repeated exposure (STOT-RE), and SVHCs with persistent, mobile and toxic (PMT) properties. Additionally, the number and variation in ED substances that are currently classified as SVHC – and thus included in the model – is limited (n=52). Therefore, substances with a steroid backbone will currently not yet be identified as similar to an existing ED substance, although they can be expected to have endocrine disrupting properties. Furthermore, we advise to not only apply the predictive model to the parent substance, but also to the breakdown products, as this may give different similarity outcomes. In addition, we noted that the SMILES standardization step, as included in the workflow of the models, does not consistently apply to phthalates in the CMR model (i.e. some phthalate SMILES are adjusted, whereas others were not adjusted). This is likely related to a bug in the SMILES-standardization step of the PaDEL-Descriptor, in line with earlier reported bugs [79,98]. Exclusion of this step from the model workflow will not result in significant changes and conclusions (model conclusions of less than 1% of the substances in the REACH dataset change), but for individual (phthalate) substances it can make a difference. In addition, exclusion of the SMILES standardization step does not necessarily result in more reliable predictions for substances with an altered conclusion. For the PBT/vPvB model on the other hand, which is also based on a fingerprint derived from PaDEL-Descriptor, only marginal differences were observed (model conclusions of less than 0.1% of the substances in the REACH dataset change, n=4).

Additionally, several future adjustments could be made to further improve the performance of the models. Based on the conducted analysis, the models seem to incorrectly classify substances in the direction of false positives (i.e. higher sensitivity then specificity, Table 3.4), though most false classified substances have a similarity to a SVHC close to the model's threshold (Figure S.4). Especially for the PBT/vPvB model – where many false positives were

close to the threshold – adjustment of the threshold could be considered, depending on the application purpose of the model. For instance, adjustment of the threshold to 0.971 results in a balanced accuracy of 0.74, with markedly less false positives. Nevertheless, for screening applications false positives might be preferred in a regulatory context over false negatives. In addition, for the PBT/vPvB model specifically, future adjustments could consider an update of the underlying fingerprint. Inclusion of counts and types of halogen containing fragments will potentially improve the performance. Furthermore, the models in general could potentially be improved by expressing the results more quantitatively, instead of 'yes/no'-scores. For instance, by providing a probability score with the 'yes/no'-classification. This may improve result interpretation, as it helps to separate and identify borderline-cases from clear-cases. In addition, overall screening performance might be improved by combining the results of multiple screening models. Generally, an improved performance is observed when a consensus model is applied [99,100], as underlying methods are generally based on varying types of information (e.g. structural features and physicochemical properties).

3.5 Conclusions

Within this study, the performance of newly developed structural similarity models to identify potential SVHCs was investigated. The models were applied to a large dataset, and predictions were evaluated with a set of substances that were putatively assessed as SVHC or non-SVHC based upon consensus scoring using expert elicitations. The use of expert opinions was particularly suitable as this is exactly the intended purpose of the chemical similarity models: a quick, reproducible and automated screening tool that mimics the expert judgement that is frequently applied in various screening applications. The results indicate a good statistical performance for CMR and ED substances, whereas a moderate performance was observed for PBT/vPvB substances when compared to expert opinions. For the PBT/vPvB model, particularly false positive substances were identified, which indicates the necessity of outcome interpretation.

In general, the structural similarity models showed great potential for screening and prioritization purposes. The models provide an automated, fast and reproducible alternative to expert opinions, and the results are more consistent compared to direct expert reactions, which can be prone to biased or context-dependent deviations. The models provide clear follow-up directions for substances of potential concern, and could particularly be used to identify groups of substances of potential concern. By this, it could further contribute to the transition from substance-by-substance assessments to group assessment approaches.

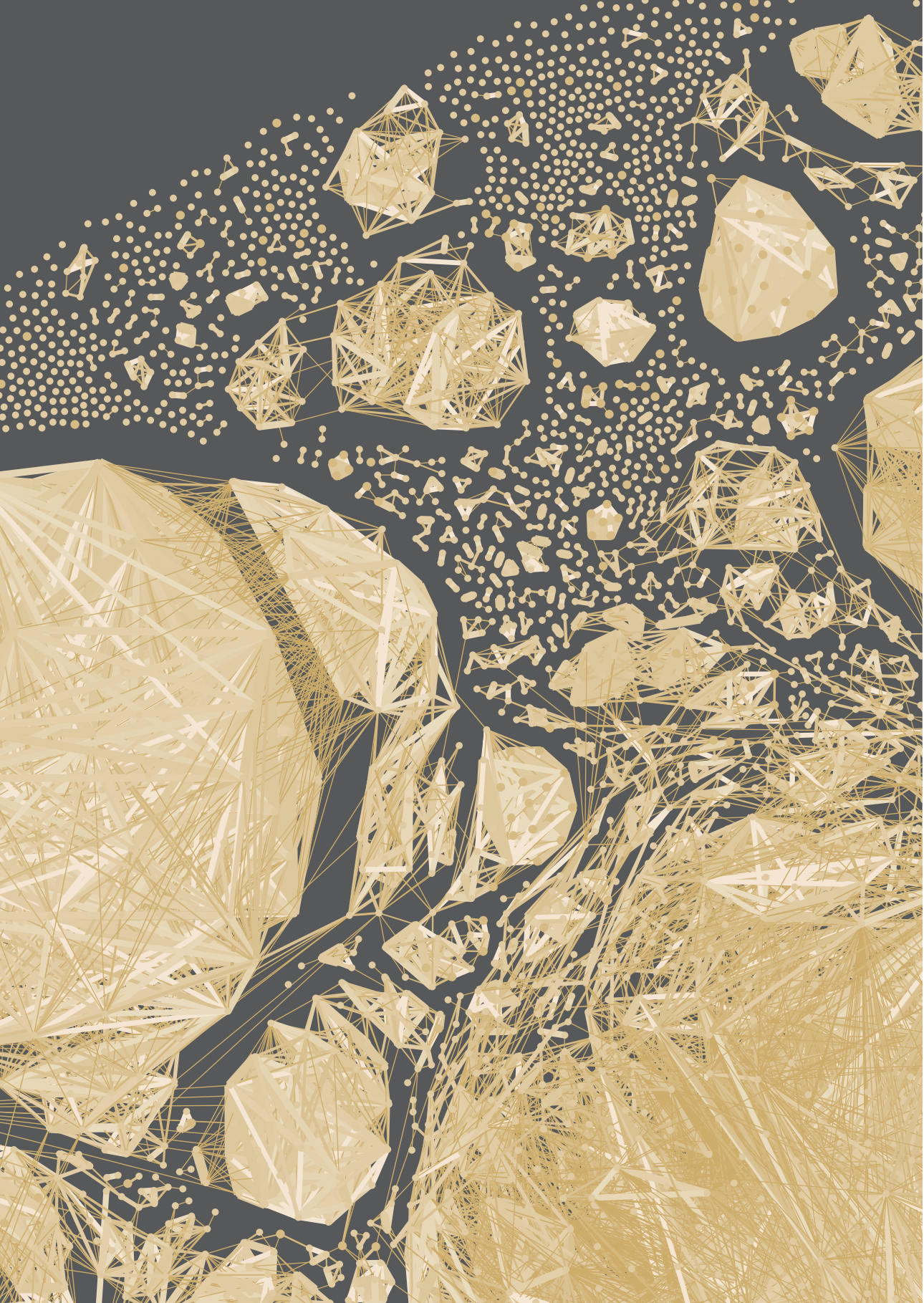
Acknowledgments

We highly appreciate and acknowledge the assistance of Dr. Domenico Gadaleta (Istituto di Ricerche Farmacologiche Mario Negri IRCCS) with the KNIME-workflow. We thank ECHA for providing a list of SMILES of registered substances, we thank Dr. Ellen Cieraad (Leiden University) for advice on the expert judgement survey, and we gratefully acknowledge all participants who filled in the expert judgement survey. In addition, we would like to thank all colleagues who internally reviewed the manuscript. This work was partially funded by the Dutch Ministry of Infrastructure and Water Management.

Supplemental material

Supplementary data to this chapter can be found online at https://doi.org/10.1016/j.yrtph.2020.104834.

3



4

ZZS Similarity Tool: the Online Tool for Similarity Screening to Identify Chemicals of Potential Concern

Pim N.H. Wassenaar, Emiel Rorije, Martina G. Vijver and Willie J.G.M. Peijnenburg

A version of this chapter has been submitted.

Abstract

Screening and prioritization of chemicals is essential to ensure that available evaluation capacity is invested in those substances that are of highest concern. We therefore recently developed structural similarity models that evaluate the structural similarity of substances with unknown properties to known Substances of Very High Concern (SVHC), which could be an indication of comparable effects. In the current study the performance of these models is improved by (1) separating known SVHCs in more specific subgroups, (2) (re-)optimizing similarity models for the various SVHC-subgroups, and (3) improving interpretability of the predicted outcomes by providing a confidence score. The improvements are directly incorporated in a freely accessible web-based tool, named the ZZS similarity tool: https://rvszoeksysteem.rivm.nl/ZzsSimilarityTool. Accordingly, this tool can be used by risk assessors, academia and industrial partners to screen and prioritize chemicals for further action and evaluation within varying frameworks, and could support the identification of tomorrow's substances of concern.

4.1 Introduction

Evaluation and regulation of chemical substances is crucial to ensure safe production and use of chemicals. For substances that are of concern regulatory measures can be implemented that assure a minimization of emissions and exposure, and/or could stimulate the substitution by safer (non-regrettable) alternatives. Such actions contribute to the European ambitions of a toxic-free environment [13]. However, as available evaluation capacity is limited, it is essential to first evaluate (and subsequently regulate) those substances that are of highest concern. To facilitate the identification of substances of potential concern, we recently developed structural similarity models that evaluate the structural similarity of substances with unknown hazard properties to known Substances of Very High Concern (SVHC) [73]. Structural similarity is considered an important descriptor in various research fields, including toxicology (e.g. for read-across [50]) and pharmacology (e.g. for virtual screening [101–103]), as a high resemblance in chemical structure could be an indication of comparable properties and effects ('similar property principle') [30]. Therefore, substances that are structurally similar to known SVHCs might be selected for further evaluation.

The SVHC similarity models are based on chemical fingerprints and similarity coefficients [7,32] and the workflow of the models is illustrated in Figure 4.1. Separate similarity models have been developed for three groups of SVHCs, including (1) SVHCs with carcinogenic (C), mutagenic (M) or reprotoxic (R) properties (i.e. CMR), (2) SVHCs with persistent, bioaccumulative and toxic (PBT) or very persistent and very bioaccumulative (vPvB) properties (i.e. PBT/vPvB), and (3) SVHCs with endocrine disrupting (ED) properties. These models showed promising performance statistics (with balanced accuracies of 0.80-0.99) [73], and showed a reasonable performance on a broader universe of chemicals as analyzed by a pseudo-external validation (with balanced accuracies of 0.69-0.87) [104]. In addition, the model predictions appear to be more robust than expert judgements [104]. To enable the use of the similarity models by risk assessors, academia and industrial partners, we have made the models publicly available via a freely accessible web-based tool, named the ZZS similarity tool: https://rvszoeksysteem.rivm.nl/ZzsSimilarityTool (ZZS = 'Zeer Zorgwekkende Stoffen' [in Dutch], which is literally translated as substances of very high concern). Accordingly, this tool can be used to screen and prioritize chemicals for further action and evaluation within varying frameworks, and is already applied in various screening activities [105,106].

Upon obtaining more experience with the application of the similarity models, we identified several methodological aspects that could be further optimized to improve the performance of the models [73,104]. Particularly, the PBT/vPvB model misclassified various substances due to amongst others insufficient consideration of the type and number of halogenated fragments and aromatic structures. Moreover, the SVHC-categorization insufficiently reflected the

current SVHC status, and the binary nature of the predictions limited the interpretation of the results. Therefore, the current study aims to improve the performance of the models by (1) separating the known SVHCs in more specific subgroups, (2) (re-)optimizing similarity models for the various SVHC-subgroups, and (3) improving interpretability of the predicted outcomes by providing a confidence score. In addition, the underlying reference dataset of SVHC substances was updated. The improvements as described in this study are directly incorporated in the ZZS similarity tool, and enhance the applicability of the models.

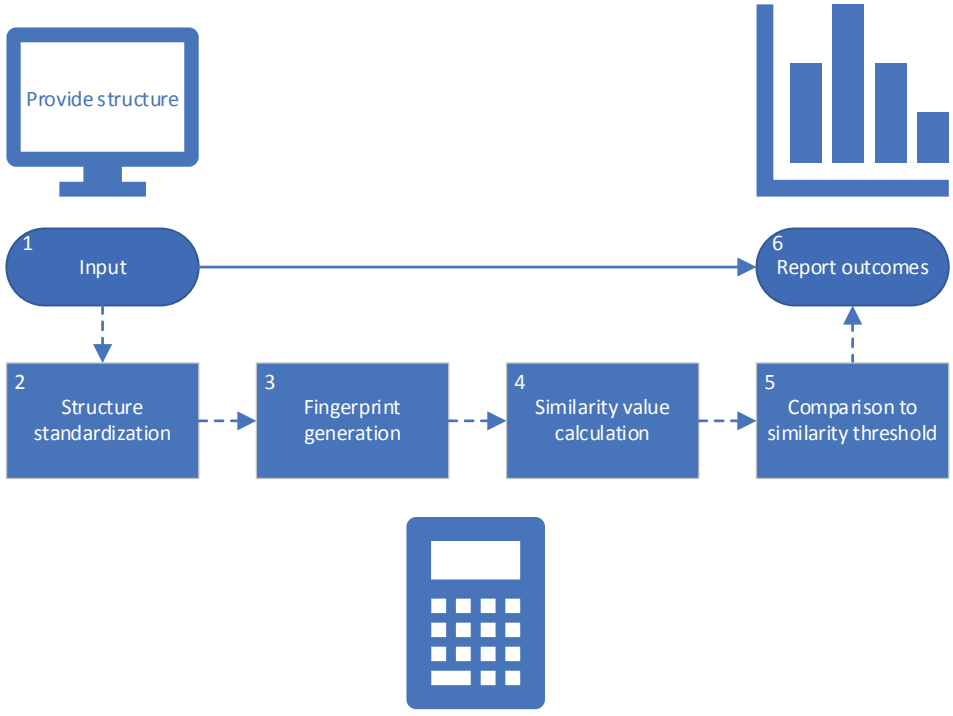


Figure 4.1. Illustration of the general workflow of the similarity models that are incorporated in the ZZS similarity tool. Step 1 and 6 consider the input and output as shown by the ZZS similarity tool, and step 2-5 are used to calculate and predict the structural similarity. An input structure can be provided as SMILES or CAS-number (step 1), which is converted to a standardized SMILES to ensure equal comparison to SVHC structures (step 2). The standardized SMILES is used to generate chemical fingerprints using PaDEL-Descriptor [6] (step 3). The fingerprint of the input structure is compared to the fingerprints of all SVHCs to calculated similarity values by using a similarity coefficient (step 4). The calculated similarity values are compared to a similarity threshold to predict whether the input structure is considered sufficiently structurally similar to an SVCH (step 5), and the results are reported (step 6). For each SVHC-category a specific model was developed and optimized, that consists of a unique fingerprint, coefficient and threshold combination.

4.2 Methods

The methodological aspects of the similarity models that are adjusted in this study are shown in Table 4.1, and include an update of the underlying SVHC dataset, a re-categorization of the SVHCs into subgroups, a (re-)optimization of the similarity models, and the addition of a quantitative outcome score.

Table 4.1. Aspects of the structural similarity models that are adjusted within the current study.

Adjusted aspects	Description and motivation
Dataset	Update of the underlying SVHC dataset.
Model-separation	Separation of CM and R concerns, as these effects are often exerted via different mode of actions. Improved distinction between European SVHCs (including CLP classifications and POP identifications) and Dutch SVHCs.
Model (re-)optimization	Optimization of the sub-models. Specifically necessary for the PBT/vPvB category, for which a moderate performance on the broader universe of chemicals was observed.
Outcome interpretation	Addition of a quantitative confidence score, besides the qualitative conclusion (sufficiently similar: yes/no), to support better outcome interpretation.

Abbreviations: SVHC - Substances of Very High Concern. CMR - Carcinogenic (C), Mutagenic (M) or Reprotoxic (R) properties. PBT/vPvB - very (v) Persistent (P), Bioaccumulative (B) and Toxic (T) properties. CLP: Classification, Labelling and Packaging of substances and mixtures. POP: Persistent Organic Pollutants.

4.2.1 Dataset

The dataset of SVHCs was updated between 2018 to 2021 based on the substances that were included on a Dutch list of SVHCs (25-01-2021; [107]) following the same refinement procedure as previously described [73]. This list includes substances that are identified based on the same hazard criteria as the European SVHCs, but are derived from various additional sources and therefore cover a slightly broader range of chemicals than the SVHCs under REACH (see Supplemental Material S.1 for a detailed description of the Dutch list of SVHCs). In addition, the substances included on the final SVHC list were categorized based on their hazard class (in which a chemical can belong to multiple hazard classes). Distinctions were made between the 'classical' SVHC hazard categories, including C, M, R, PBT and vPvB. Substances were added to these categories when they were considered to have such specific effects according to their inclusion on the European SVHC list, the European CLP list Annex VI, or on the list of Persistent Organic Pollutants (POPs). All POPs were considered as PBT and/or vPvB within the dataset. In addition, as within our previous work, a specific ED

category was used. Substances were added to this category when they were included on the European SVHC list based on ED effects. Substances on the Dutch list of SVHCs that do not belong to any of the abovementioned categories were included in the 'Other'-category, like substances on the European SVHC list with persistent, mobile and toxic (PMT) properties, specific target organ toxicity after repeated exposure (STOT-RE) or sensitizing properties. In addition, substances were also included in the 'Other'-category when they were only included on the Dutch list of SVHCs based on other sources, like substances on the OSPAR list for priority action [108] or priority hazardous substances according to the Water Framework Directive (see Supplemental Material S.1 for more details about the Dutch list of SVHCs).

For modelling purposes, also a list of non-SVHCs was required. We used the same list as used by Wassenaar et al. (2019), but excluded the substances that were now included on the 'new' list of SVHCs (resulting in a total of 406 substances).

As chemical similarity evaluations require unambiguous chemical structures as input information, we normalized and standardized all SMILES to QSAR ready structures with a Kekulé representation [109]. This was done by extracting the QSAR ready structures from a CAS-SMILES list from the US-EPA [110] or by generating QSAR ready structures with a KNIME workflow [111]. In exceptional cases, where no QSAR ready SMILES could be generated, the most uniform representation was manually selected (e.g. from PubChem or ECHA dissemination site).

4.2.2 Models

The SVHC dataset was separated into five different hazard classes, including CM, R, PBT/vPvB, ED and Other, and all used the same set of non-SVHCs for model optimization.

The selection of the best performing similarity models was specifically restricted to fingerprints that could be generated with PaDEL-Descriptor (as those are incorporated in the online ZZS similarity tool) [6]. This includes the Substructure, MACCS, E-State, PubChem, Klekota-Roth and CDK Extended fingerprint. Fingerprints were generated for all QSAR ready SMILES of the SVHC and non-SVHC substances with PaDEL-Descriptor, enabling PaDEL to remove salts, detect aromaticity, and standardize tautomers and nitro groups [6]. These fingerprints were all tested in combination with the JT, HL, CT4, SS3, Coh, SM and Yu2 similarity coefficients [7,73]. More details on the fingerprints and similarity coefficients are provided in Supplemental Material S.2.

We analyzed the predictive performance of the varying fingerprint-coefficient combinations for classifying the substances in the dataset as (potential) SVHC or non-SVHC per SVHC

category (i.e. CM, R, PBT/vPvB, ED and Other). For each fingerprint-coefficient combination similarity values were calculated. Non-SVHC substances were compared to all SVHCs, whereas SVHCs were compared to all other SVHCs (excluding itself), and per substance only the highest similarity value was retained. Next, the maximum balanced accuracy was determined (Equation 4.1), by selecting the optimal threshold (i.e. a value between 0 and 1) to predict (potential) SVHC status versus non-SVHC status. Details are according to Wassenaar et al. (2019).

Selection of (or adjustments to) the best performing models focus on quantitative performance statistics (i.e. balanced accuracy), but also included qualitative selection criteria (which could vary between hazard classes), where necessary. For instance, in the case of a symmetric similarity coefficient (i.e. coefficients in which absence and presence of features that are in common between two structures contribute equally to the determined similarity), specific care was given to symmetric coefficient bias (i.e. the phenomenon where chemicals with less than a specific number of fragment features are always predicted to be structurally similar to an SVHC due to high overlap in absent features) (see Wassenaar et al. (2019) for a more detailed description). Furthermore, for the PBT/vPvB-model (as well as the 'Other'-model) specific attention was given to the performance on the broader universe of chemicals, as these models had a relative low external performance in a previous evaluation [104].

$$balanced\ Accuracy\ (bAcc) = \frac{Sensitivity + Specificity}{2} = \frac{\frac{TP}{TP+FN} + \frac{TN}{TN+FP}}{2} \qquad (4.1)$$

4.2.3 Outcomes

A quantitative confidence score was added to the binary model predictions (i.e. the yes or no prediction on sufficient structural similarity). The confidence scores represent the confidence in the structural similarity between a chemical and an SVHC, and are derived from the similarity values. The following stepwise procedure was followed for each similarity model. First, we iteratively assessed the performance for all distinguishing similarity values (i.e. threshold values) based on the subgroup specific SVHC and non-SVHC datasets, and derived balanced positive predictive values (bPPV) for each similarity threshold value (see Equation 4.2). Second, the bPPVs were min-max normalized to confidence values ranging from 0 to 1 (i.e. 0-100%), in which the model's optimal threshold value was set to a confidence value of 0.5 (i.e. 50%). Third, we fitted two functions through these normalized bPPVs. One function is fitted to the similarity values ranging from 0 to the model's optimal threshold (with confidence scores ranging from 0-50%), and the other function is fitted to the similarity values from the model's optimal threshold till 1 (with confidence scores ranging from 50-100%). Depending on the distribution of the bPPV for all similarity thresholds values, a corresponding function

was selected (e.g. exponential or sigmoidal function). In cases where no clear distribution pattern was observed, a linear trend was used. The fitted functions at least had to cover the confidence ranges from 0.5-49.5% and 50.5-99.5%, and must sufficiently represent the derived bPPV points, where possible. When necessary, the fit was manually optimized to meet these conditions, by for instance constraining the bottom or top of the curves at specific similarity values, or by providing additional weight to specific datapoints. A visual example of the fitting through a distribution of bPPV values as a function of similarity threshold values is given in the results section (Figure 4.2).

$$balanced\ Positive\ Predictive\ Value\ (bPPV) = \frac{Sensitivity}{Sensitivity + (1 - Specificity)} = \frac{\frac{TP}{TP+FN}}{\frac{TP}{TP+FN} + (1 - \frac{TN}{TN+FP})} \quad (4.2)$$

All analyses within this study were performed in R (unless otherwise specified) [112] using *caret*, *ChemmineR*, *caTools*, and *ROCR* [60,62,113,114].

4.3 Results & Discussion

4.3.1 Dataset

The new dataset consists of 621 substances, of which 80 structures were not yet included in the previous dataset. In addition, eight structures were removed (e.g. as they do not meet the SVHC criteria anymore), or were represented by newly included (n=3) or already existing structures (see Supplemental Material Excel for more details). Furthermore, we re-categorized the substances across the hazard classes to better reflect the current SVHC status and thereby improve the interpretability (e.g. distinction between EU-based SVHCs versus SVHCs that are only identified as a Dutch SVHC; and distinction between CM- and R-concerns). The distribution of substances within this updated dataset across the different hazard categories is shown in Table 4.2, and the individual substances are included in Supplemental Material Excel.

Table 4.2. Overview of the new dataset and the distribution over hazard categories, in comparison to the previous dataset as included in Wassenaar et al. (2019).

Hazard class	Previous dataset	New dataset
Total	546	621
CM	150 [1]	153
R	166 [1]	178
PBT/vPvB	209	137
ED	52	51
Other	- [2]	131 [3]

1) In the previous work, CM and R were combined as one class (n=306). 2) In the previous work, no 'Other'-category was included. 3) The 'Other'-category consists of 10 substances that are identified as EU-SVHC based on PMT (n=3) or respiratory sensitizing properties (n=7). All others are not identified as EU-SVHC, EU-CLP or POP, but are included on the Dutch list of SVHCs based on specific concerns related to similar endpoints (C: n=3, M: n=1, R: n=14, PBT: n=64, PBT/vPvB: n=29, ED: n=6, PMT: n=2, and others: n=2) from other sources (e.g. OSPAR [108]; in which PBT/vPvB concerns are dominating).

4.3.2 Models

Specifically the PBT/vPvB model required improvement according to the performance on the broader universe of chemicals [104]. Despite the excellent performance on classifying substances in the original SVHC dataset, the PBT/vPvB model misclassified many substances due to amongst others insufficient consideration of the type and number of halogenated fragments and aromatic structures. In addition, the performance of the other similarity models (i.e. CMR and ED models) were reanalyzed as several adjustments have been made, including an update of the SVHC dataset and a new categorization of substances (i.e. CM, R, PBT/vPvB, ED and 'Other'-models).

Optimization of the CM- and R-models based on the new datasets indicated that the CDK Extended fingerprint with SM-coefficient was the best or second best performing fingerprint-coefficient combination for the CM- and R-dataset, respectively. For the R-dataset, the Extended-Coh combination scores best followed closely by the Extended-SM fingerprint-coefficient combination (with balanced accuracies of 0.814 and 0.808, respectively). These results are comparable to the results from our previous study, in which the Extended-SM fingerprint-coefficient combination outperformed all other combinations for the CMR-dataset with comparable optimal similarity thresholds (i.e. a threshold of 0.946 for the CM-dataset, 0.944 for the R-dataset and 0.944 for the combined dataset in the previous study) [73]. We decided to additionally use an asymmetric similarity coefficient (i.e. JT or CT4 coefficient) for substances with a low number of fingerprint bits, as symmetric coefficient bias was observed. Statistical derivation of an optimal cut-off value (i.e. below which number of fingerprint bits the JT or CT4 coefficient should ideally be used) resulted in broad uncertainty ranges due to a limited number of substances in the subsets. As comparable best-performing

models were derived for the new CM- and R-dataset as previously determined, we decided to retain the CMR-model. The established optimal threshold and cut-off specifications are given in Table 4.3, and showed to be robust to minor changes in the dataset and do not specifically require an adjustment of the optimized parameters. Moreover, this decision was justified by the fact that besides an update of the dataset there was no specific incentive to improve the performance of the CMR-model based on the previous evaluations.

Revision of the PBT/vPvB model using the MACCS-SM fingerprint-coefficient combination was required considering its performance on the broader universe of chemicals [104]. In addition, as many not (yet) EU-recognized PBT/vPvB chemicals were reallocated to the 'Other'-category, also specific attention was given to the optimization of the similarity models for this group. The Klekota-Roth, PubChem and CDK Extended fingerprint were identified as best performing fingerprints based on performance statistics for the PBT/vPvB-SVHCs and non-SVHCs. However, upon a more in-depth analysis of the predicted similarities (including false positives and false negatives) and its applicability on the broader universe of chemicals, it could be concluded that the Klekota-Roth fingerprint is not suitable to predict structural similarity amongst PBT/vPvB chemicals. The Klekota-Roth fingerprint provides a lot of emphasis to (small) linear chains of varying sizes and to relatively large fragments, but insufficiently weighs typical PBT-related fragments like aromatic-ring structures. In addition, for relatively many chemicals only a limited number of fragments are identified, and accordingly such chemicals are more easily (but often incorrectly) predicted as structurally similar to a PBT/vPvB-SVHC. The PubChem and Extended fingerprints have their own strengths and limitations. The PubChem fingerprint specifically weighs aromatic structures and halogens, but does not systematically cover the whole chemical structure. The Extended fingerprint specifically considers all fragments present within a chemical, but focusses specifically on path-based fragments which may insufficiently describe ring-structures. As both fingerprints have their own unique flaws, they were combined to form the final PBT/vPvB model. The corresponding best performing coefficients included the JT, CT4 and SS3 coefficients. As the observed differences between these coefficients are in the details (see Supplemental Material S.2) and are partially related to the determined optimal threshold values, we selected the fingerprint-coefficient combinations with best performance on the expert judgement dataset (given preference to high bPPVs and few false positives, to ensure confidence in model predictions). The final PBT/vPvB model uses both the PubChem-JT and Extended-CT4 fingerprint-coefficient combinations, and only predicts that a chemical is structurally similar to a PBT/vPvB-SVHC when both models support this conclusion (see Table 4.3).

In the ED-dataset, four new ED-SVHCs were added and five SVHCs were allocated to the 'Other'-category. The results of the model optimization indicate that multiple fingerprint-

coefficient combinations can be considered as the best-performing model, all with a balanced accuracy of 0.99 (including models based on the CDK Extended fingerprints and Klekota-Roth fingerprints). Although previously the RDkit based FCFP4 fingerprint with the SS3 coefficient was considered most optimal (with an equal performance statistics), we now pragmatically selected a PaDEL-based fingerprint as those were incorporated in the online ZZS similarity tool. We selected the CDK Extended fingerprint with JT-coefficient from the best performing fingerprint coefficient combinations. The Extended fingerprint was chosen above the Klekota-Roth fingerprint, as the Klekota-Roth fingerprint only considers a specific number of pre-specified fragments and therefore might be less specific when applied to a broader universe of chemicals. In addition, the JT-coefficient was selected as this coefficient uses an asymmetric function for which there is no risk of symmetric coefficient bias, and was preferred above the CT4 coefficient (see Supplemental Material S.2).

The 'Other'-category consists of SVHC substances whose properties are different from the abovementioned categories, or whose properties are not universally recognized as such. These substances were separated from the CM, R, PBT/vPvB and ED-dataset to better reflect the current SVHC status and thereby improve the interpretability. For the 'Other'-dataset very comparable observations and conclusions were made as compared to the PBT/vPvB-dataset, and this might not be a surprise considering the broad representation of PBT/vPvB related chemical concerns in the 'Other' category (>70%, see Table 4.2). The only differences between both models are the most optimal threshold values that are derived from the dataset (see Table 4.3).

Although we specifically assessed the performance of PaDEL-based fingerprints within this study, comparable or lower performances were observed for the RDkit related fingerprints that were previously tested as well [73,115].

Table 4.3. Overview of the final models – including performance statistics – to predict structural similarity to SVHCs.

Subset	Fingerprint	Coefficient	Threshold	#SVHCs	#non-SVHCs	TP	FP	TN	FN	Sens	Spec	bAcc	bPPV
CM <85	CDK Extended	CT4	0.851	89	63	55	6	57	34	0.618	0.905	0.761	0.866
CM ≥85	CDK Extended	SM	0.944	64	343	36	5	338	28	0.563	0.985	0.774	0.975
CM-combined [1]	-	-	-	153	406	91	11	395	62	0.595	0.973	0.784	0.956
R <85	CDK Extended	CT4	0.851	57	63	36	7	56	21	0.632	0.889	0.760	0.850
R ≥85	CDK Extended	SM	0.944	121	343	77	7	336	44	0.636	0.980	0.808	0.969
R-combined [1]	-	-	-	178	406	113	14	392	65	0.635	0.966	0.800	0.948
PBT-1	PubChem	JT	0.774	137	406	130	2	404	7	0.949	0.995	0.972	0.995
PBT-2	CDK Extended	CT4	0.887	137	406	124	1	405	13	0.905	0.998	0.951	0.997
PBT-combined [2]	-	-	-	137	406	123	1	405	14	0.898	0.998	0.948	0.997
ED	CDK Extended	JT	0.693	51	406	50	0	406	1	0.980	1.000	0.999	1.000
Other-1	PubChem	JT	0.818	131	406	87	15	391	44	0.664	0.963	0.814	0.947
Other-2	CDK Extended	CT4	0.901	131	406	76	17	389	55	0.580	0.958	0.769	0.933
Other-combined [2]	-	-	-	131	406	73	10	396	58	0.557	0.975	0.766	0.958

1) Substance is either assessed on structural similarity according to model 1 or model 2, depending on its number of fragment features. 2) Substance is assessed based on model 1 and model 2, and is only considered as structurally similar to an SVHC when it meets the criteria of both models.

4.3.3 Outcomes

Within the previous models, only a dichotomous, qualitative, prediction of the concern was made for the structural similarity of a chemical to an SVHC. Based on the similarity score and model specific threshold, the models predicted whether or not a chemical is sufficiently structurally similar to an SVHC (and thus predicted to be a potential SVHC). To support a better interpretation of the outcomes for prospective model users, a quantitative confidence score is added to this binary prediction. The developed quantitative scores describe the confidence in structural similarity between a chemical and an SVHC, with a higher confidence for higher structural similarity. This supports the intuitive interpretation that a substance that is more similar to an existing SVHC is also predicted with more certainty to have SVHC properties. The confidence score functions were derived separately for each model and are based on the normalized bPPV for substances in the SVHC and non-SVHC dataset. A similarity value equal to a model's optimized threshold was given 50% confidence, with a maximum confidence of 100% (in case of a similarity score of 1) and a minimum confidence of 0% (in case of a similarity score of 0). An example of such a function is shown in Figure 4.2, and a detailed overview of all derived confidence functions (including figures showing the bPPV as a function of the similarity value) is provided in Supplemental Material S.3. The functions do not aim to provide an exact confidence trigger, but are meant to provide additional (data-driven) information that could guide interpretation and follow-up evaluation. We specifically did not include a predictive score for non-similarity to an SVHC (for instance based on negative predictive values), as the models only make statements about the similarity and not the absence of similarity to SVHCs. This is related to the fact that it cannot be concluded that a substance is not a potential SVHC based on a lack of structural similarity, as a substance might exert effects through different (yet unknown) modes of action.

4.3.4 Application to a broader universe of chemicals

To illustrate the effects of the model adjustments, we applied the newly optimized similarity models to a dataset of REACH registered substances that was used by Wassenaar et al. (2021). The REACH dataset was slightly adjusted, by converting the SMILES to QSAR ready SMILES, similarly as performed for the SVHC dataset (see section 4.2.1). This resulted in a dataset of in total 9456 REACH registered substances. The results of the screening are shown in Table 4.4, in which also the results of the previous similarity models are included (using the newly updated SVHC-dataset).

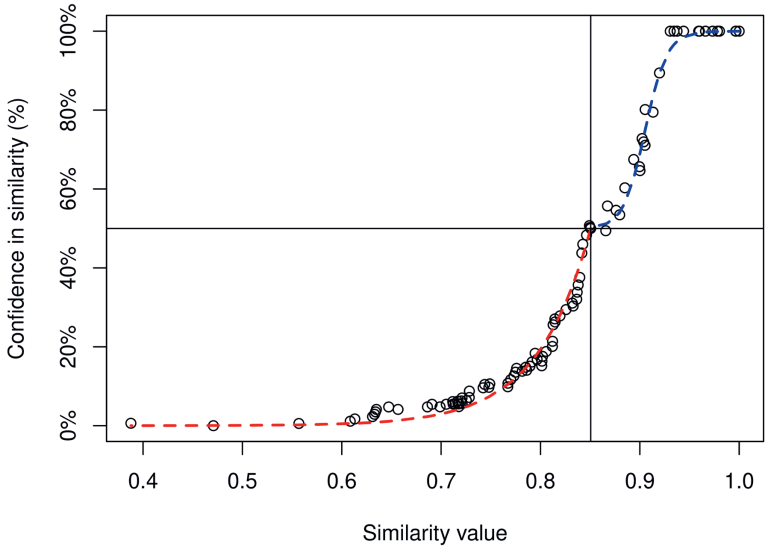


Figure 4.2. Relation between the structural similarity value and the confidence in the predicted structural similarity between a chemical and a Reprotoxic (R)-SVHC based on the CDK Extended-CT4 fingerprint-coefficient combination. The fitted curves describe the normalized bPPV as a function of the similarity value used as a threshold value, and are derived from the R-SVHC and non-SVHC datasets (for substances with less than 85 fragment features, i.e. bits in the CDK Extended fingerprint). The vertical line represents the model's optimized threshold value (0.851) giving the best balanced accuracy, and the horizontal line represents the 50% confidence score. More details are presented in Supplemental Material S.3.

Based on the results as shown in Table 4.4, it can be observed that far more substances are identified as potential CM or R, compared to the other three categories. This difference can likely be explained by a larger diversity in SVHC structures within the CM- and R-categories. As previously shown, these categories have much more 'single-point-of-knowledge' structures, compared to PBT/vPvB and ED-SVHCs which can be divided into relatively few groups of chemicals [73]. Therefore, it cannot simply be concluded that the PBT/vPvB-, ED- and 'Other'-models are more strict compared to the CM- and R-models. The addition of confidence scores, however, allows for a better interpretation of the predicted results (with a higher structural similarity resulting in a higher confidence in predicted results).

Furthermore, Table 4.4 and Supplemental Material S.4 indicate very comparable distributions in confidence scores across the varying categories. The results for the 'Other'-SVHC category are the only exception, for which a relative high amount of structures are identified that are structurally very similar to an 'Other'-SVHC. This is, at least partially, related to the steep increase in confidence scores in relation to the similarity values, which follows from the model's derived bPPV (see Supplemental Material S.3). In addition, this might also be related to a coincidental high representation of structurally very similar substances in the screened dataset.

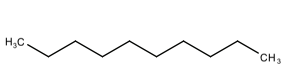
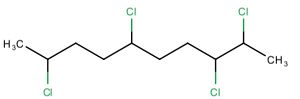
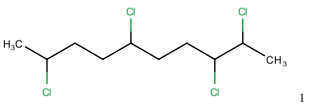
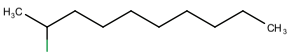
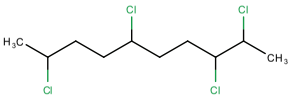
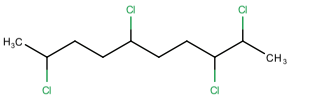
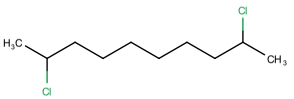
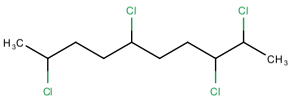
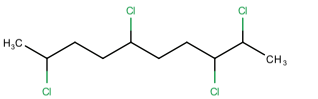
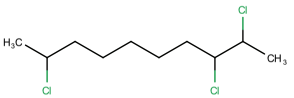
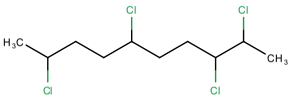
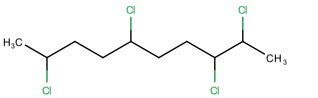
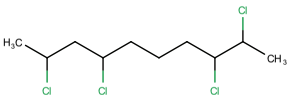
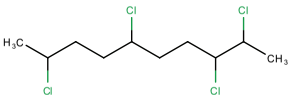
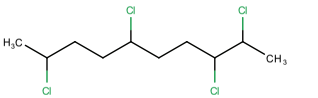
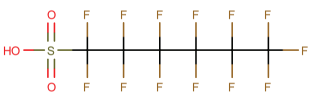
Table 4.4. Application of the newly optimized similarity models to a dataset of 9456 REACH registered substances. The confidence-bins represents the number of substances that are predicted to be structurally similar to an SVHC with a specific confidence in the structural similarity. The previously used similarity models are described by Wassenaar et al. (2019).

Model	Similar substances	Similar substances by previous models	50-75% Confidence	75-90% Confidence	≥90% Confidence
CM-combined [1]	1060	- [3]	701	149	210
CM<85	688	- [3]	466	76	118
CM≥85	372	- [3]	235	73	92
R-combined [1]	936	- [3]	729	98	109
R<85	522	- [3]	376	60	68
R≥85	414	- [3]	353	38	41
PBT/vPvB	53 [2]	360 [4]	38	12	2
ED	109	139 [5]	86	13	10
Other	129 [2]	554 [4,6]	32	46	51

1) Combination of two sub-models. 2) For two chemicals the PubChem fingerprint could not be generated (total = 9454). 3) The CM- and R-models were not adjusted. 4) For one chemical the MACCS fingerprint could not be generated (total = 9455). 5) For 82 chemicals the RDkit equivalent FCFP4-fingerprint could not be generated (total = 9374). 6) The previously derived PBT/vPvB model was applied to the 'Other'-dataset (as the 'Other'-SVHCs mainly consists of SVHCs previously included in the PBT/vPvB-SVHC dataset).

Predictions for the PBT/vPvB- (and 'Other'-) model have much improved (with a much lower number of substances that are predicted as potential SVHCs; Table 4.4) and can be better interpreted using the additional confidence scores (see Table 4.5) as compared to predictions reported in Wassenaar et al. (2021). The added value of the confidence scores is particularly evident from examples 1-5 in Table 4.5, where an increased confidence in structural similarity is observed with an increase in halogenated fragments. In addition, these confidence scores are very useful when interpreting the similarity amongst groups of structurally similar substances, as illustrated in Supplemental Material S.5 where we present the confidence scores for previously discussed case-studies [104]. Despite the many improvements, also a deficiency of the models can be observed which particularly relates to the use of the CDK Extended fingerprint (see examples 14-15, Table 4.5). As this fingerprint considers a path-based fingerprint, not many additional fragments are identified for substances with a straight-chain of (carbon) atoms or when these atoms are structured in a ring.

Table 4.5. Specific examples of predictions by the PBT/vPvB- (IDs 1-8, 10, 11, 13, 14) and 'Other'-model (IDs 9, 12, 15), including confidence scores in structural similarity. The examples illustrate the model's improved consideration of the number of halogenated fragments (examples 1-7), type of halogenated fragments (examples 8-9), and backbone/aromatic structures (examples 10-13). A few deficiencies of the new models are illustrated with examples 14 and 15.

ID	Substance with 'unknown' properties	Previous most similar known SVHC	Previous model prediction	New most similar known SVHC	New model prediction	New model confidence in structural similarity
1			**SVHC**		Non-SVHC	43%
2			**SVHC**		**SVHC**	75%
3			**SVHC**		**SVHC**	83%
4			**SVHC**		**SVHC**	93%
5			**SVHC**		**SVHC**	96%
6			**SVHC**		Non-SVHC	23%

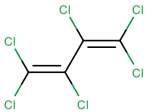
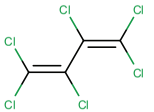
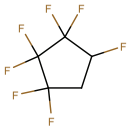
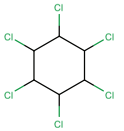
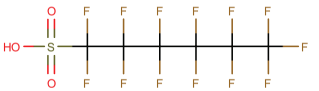
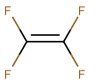
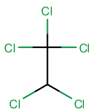
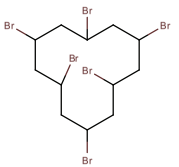
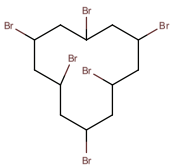
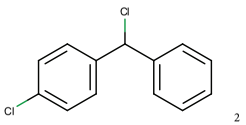
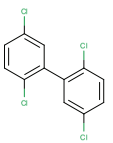
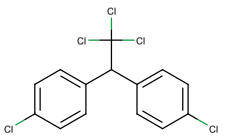
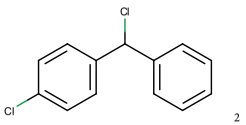
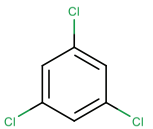
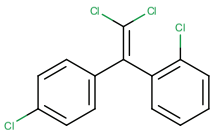
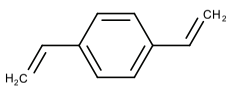
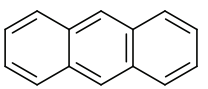
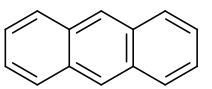
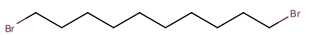
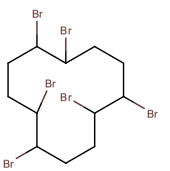
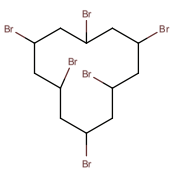
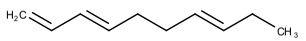
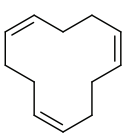
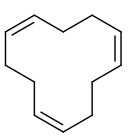
7			**SVHC**		Non-SVHC	7%
8			**SVHC**		Non-SVHC	1%
9			**SVHC**		Non-SVHC	6%
10			**SVHC**		Non-SVHC	0%
11			**SVHC**		**SVHC**	75%
12			**SVHC**		Non-SVHC	49%

Table continued on next page.

Table 4.5 (continued).

ID	Substance with 'unknown' properties	Previous most similar known SVHC	Previous model prediction	New most similar known SVHC	New model prediction	New model Confidence in structural similarity
13			**SVHC**		Non-SVHC	2%
14			Non-SVHC		**SVHC**	81%
15			Non-SVHC		**SVHC**	80%

1) This SVHC considers the third most similar SVHC, with a comparable similarity value for the two other most similar SVHCs (i.e. all with 43% confidence in structural similarity). This structure is included in this table as illustrative example in relation to examples 2-5. 2) This structure is provided as input to the PBT/vPvB-model (ID 11) and the 'Other'-model (ID 12).

4.3.5 ZZS similarity tool

The updated dataset and re-optimized similarity models were incorporated in the online ZZS similarity tool (https://rvszoeksysteem.rivm.nl/ZzsSimilarityTool). Also the user-interface was improved by adding the possibility to use a CAS-input as well as a batch-job possibility, besides the already existing SMILES-input option (see Figure 4.3). The CAS-search feature was included by adding a list of >700.000 CAS-SMILES combinations, originating from the US-EPA [110]. We refined this list by removing entries without a CAS-number or a QSAR ready SMILES, and removed any chirality description within the SMILES (as chirality has not been used in the similarity model optimization). Furthermore, we ensured that SMILES from substances in the final updated SVHC dataset (for which a CAS-number is available; see section 4.2.1) were consistent or included. Some more details on the implementation of the similarity models and use of the ZZS similarity tool are provided in Supplemental Material S.6.

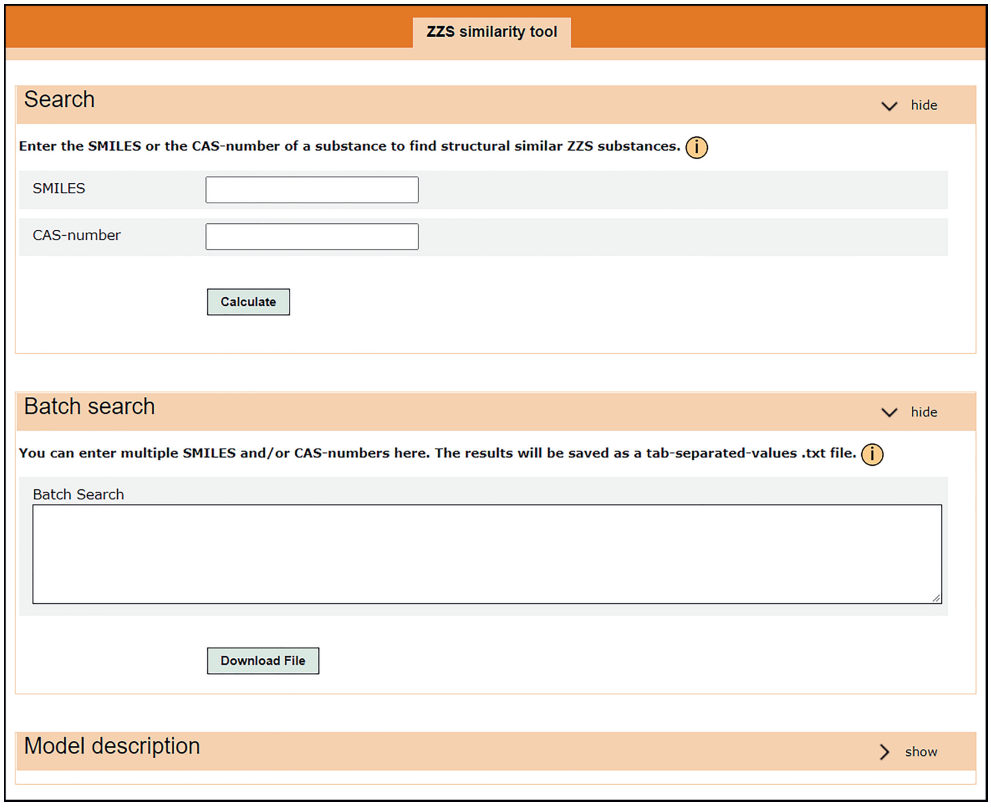


Figure 4.3. The ZZS similarity tool main web-page with the input modes: single search and batch search (using SMILES and/or CAS-numbers).

4.4 Conclusions

Within this study similarity models were extended and optimized to improve the identification of substances with potential SVHC properties. We specifically (1) accounted for differences in mode of action, (2) upgraded the PBT/vPvB sub-models, and (3) added quantitative confidence scores. In addition, the models were extended by using more data. The revised similarity models have been incorporated in the online freely available ZZS similarity tool (https://rvszoeksysteem.rivm.nl/ZzsSimilarityTool), with an user-friendly interface both enhancing interpretability and input options. Application of these models by risk assessors, academia and industrial partners will result in faster, easier and more reliable identification of substances that are potentially of very high concern, and as such can contribute to the transition to a toxic-free environment.

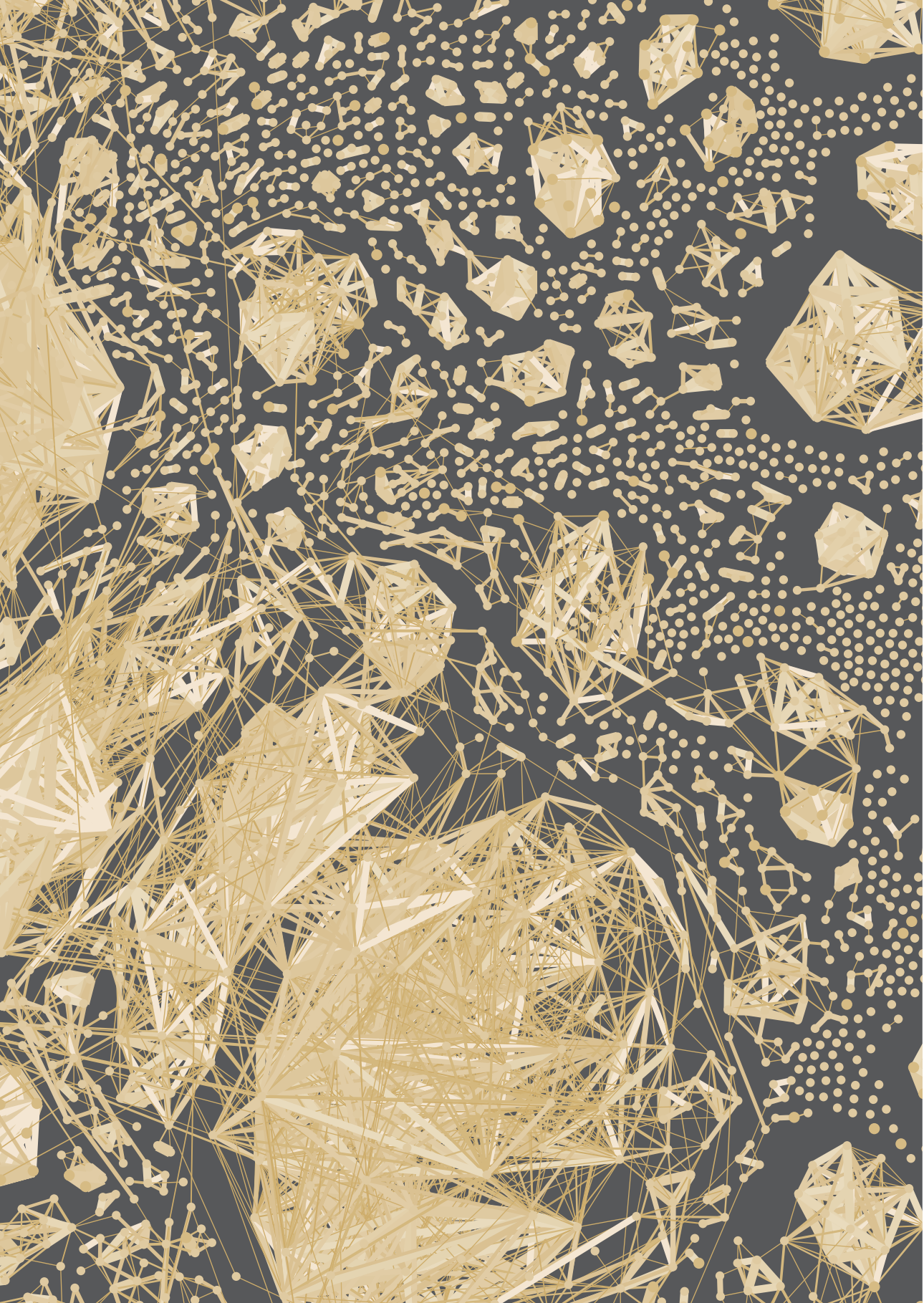
Acknowledgements

The authors would like to thank and acknowledge Roel Schreurs, Rudy Otzen, Birgit ter Horst and Martin van den Berg (RIVM) for incorporating the similarity models into the ZZS similarity tool web application. This work was partially funded by the Dutch Ministry of Infrastructure and Water Management.

Supplemental material

Supplementary data to this chapter can be obtained by the first author via e-mail (pim.wassenaar@rivm.nl).

4



5

Variability in Fish Bioconcentration Factors: Influences of Study Design and Consequences for Regulation

Pim N.H. Wassenaar, Eric M.J. Verbruggen, Ellen Cieraad, Willie J.G.M. Peijnenburg and Martina G. Vijver

Published in Chemosphere 239 (2020), 124731.

Abstract

The fish bioconcentration factor (BCF) is an important aspect within bioaccumulation assessments. Several factors have been suggested to influence BCF values – including species, developmental stage, mixture exposure, and calculation method. However, their exact contribution to variance in BCF values is unknown. Within this study we assessed the relative impact of these test characteristics on BCF values and analyzed the reproducibility of aquatic exposure bioconcentration tests.

Linear mixed effects analyses were performed on a newly develop database to investigate the relationship between the response variable (i.e. lipid normalized log BCF values) and several test characteristics as fixed effects.

Lower BCF values were observed for substances that were simultaneously applied with high molecular weight polycyclic aromatic hydrocarbons compared to single substance exposure (with an average difference of -0.81 log BCF). Also, lower BCFs upon kinetic determination were observed compared to steady-state BCFs (log BCF -0.27), and lower BCFs for species from the Ostariophysi subcohort level (log BCF -0.17 to -0.15). In addition, data analysis showed high variation within BCF values for single substances (average SD = log BCF 0.21), which questions the robustness of the current bioaccumulation assessments. For example, the 95% confidence range of a BCF value of 2500 ranges from 953 ('not-bioaccumulative') to 6561 ('very bioaccumulative').

Our results show that the use of one single BCF leads to a high uncertainty in bioaccumulation assessments. We strongly recommend that within future bioconcentration studies, the used experimental design and test conditions are described in detail and justified to support solid interpretation.

5.1 Introduction

The bioaccumulation potential of chemicals is an important factor within risk assessment. Accumulation may result in high internal concentrations leading to toxicity, even when external concentrations are low [116]. Therefore, substances with a high bioaccumulation potential are of concern, with even higher concerns for substances that – besides being (very) bioaccumulative –are also (very) persistent in the environment and/or toxic to humans or biota (i.e. PBT/vPvB-assessment). From a regulatory point of view, emissions of such substances should be minimized as much as possible, as their effects are unpredictable in the long-term, and as it is very difficult to remove the substances from the environment [19].

International regulatory criteria on bioaccumulation assessment (B-assessment) are mainly based on bioconcentration factors (BCF) in aquatic species [117]. BCFs represent the accumulation of a substance via aquatic exposure, and can be determined under laboratory-controlled conditions via OECD Test Guideline 305, ASTM E1022-94 or OPPTS Test Guideline 850.1730. Within this test, the BCF is determined at steady-state conditions (i.e. the ratio of the substance concentration in fish, C*f*, to the water concentration, C*w*, at steady state) or via kinetic determination (i.e. the ratio of the uptake rate constant, k_1, to the depuration rate constant, k_2). For very hydrophobic substances the BCF could alternatively be determined via dietary exposure [118]. In principle, a substance is considered to be bioaccumulative when the BCF value exceeds a specific threshold. Depending on the regulatory framework, the bioaccumulation cut-off value ranges from 500-5000 (Table 5.1). In addition, within some legislations, bioaccumulation factors (BAFs) or octanol-water partitioning coefficients (log K_{ow}) can also be considered within the B-assessment [117]. The consequences of B-classification varies from product labeling, restrictions in use, to minimization of emissions, with the ultimate aim of chemical substitution (e.g. for PBT/vPvB substances).

Within current regulatory frameworks, one BCF value is generally sufficient to conclude on the bioaccumulative properties. Hence, the variation (i.e. reproducibility) of this value is usually not considered. Several biotic and abiotic factors have been suggested to influence BCF values – including species, developmental stage, exposure method, calculation method, and various others [119]. And although known, the accepted experimental designs often do not specify or take into account such factors, as their exact contributions are unknown. Only some guidance and advice is provided within test guidelines with respect to preferences and/or reporting of these factors [19,118]. However, because of the importance of the B-assessment within chemical safety assessment – as indicated by the relative high number of test requests in Europe [120] – it is considered relevant to analyze the contribution of the factors that are suggested to affect the bioaccumulation potential of chemicals.

Table 5.1: BCF threshold values as applied in several international regulations [117].

Regulation	Assessment type	Bioaccumulative	Very bioaccumulative
POPs UNEP Stockholm Convention	POPs identification	5000	-
OSPAR Convention	PBT substances identification	500	-
Canadian Environmental Protection Act (CEPA)	PBT substances identification	5000	-
US Toxic Substances Control Act (TSCA)	PBT/vPvB substances identification	1000	5000
Australian National Industrial Chemicals Notification and Assessment Scheme (NICNAS)	PBT/vPvB substances identification	2000	5000
EU REACH Regulation (1907/2006)	PBT/vPvB substances identification	2000	5000
EU Plant Protection Product Regulation (1107/2009)	PBT/vPvB substances identification	2000	5000
EU Biocidal Products Regulation (528/2012)	PBT/vPvB substances identification	2000	5000
UN Globally Harmonised System (GHS)	Hazard classification and labelling	500	-
EU CLP Regulation (1272/2008)	Hazard classification and labelling	500	-

On top of that, in recent years, there has been an increasing interest in the development of alternative bioaccumulation tests, as the fish bioaccumulation studies are time consuming, expensive and animal demanding. Several new models include *in silico*, *in vitro* and invertebrate or early-life-stage *in vivo* test systems [121–124]. In order to evaluate their performance, performance information on the reference benchmark, i.e. the aqueous OECD 305 test, is necessary.

In this study we analyze and evaluate the reproducibility and influential factors for the bioconcentration test via aquatic exposure. Using a newly developed database of bioconcentration values, we assessed the impact of different test characteristics (e.g. combination exposure, calculation methodology, species and life stage of the fish) on BCF values and their variation. These test characteristics were selected specifically, because of their potential influence on BCF values and the availability of relevant information in reported studies.

5.2 Methodology

5.2.1 Data selection

Experimental BCF values were selected from the databases as developed by Arnot and Gobas [125] and the Japan METI-NITE database [126] (data extracted on 19-03-2018). Data were restricted to aquatic exposure experiments with fish, only considering direct exposure (i.e. excluding studies investigating bioconcentration in the second generation) and limited to laboratory-derived data. For each experiment only one overall BCF value was included, thus excluding all intermediate measurements. In case of steady-state BCF values, the included value involves the reported BCF value or the average of all BCF values at steady state. In addition, reported BCF values below or above a certain value (i.e. '<' or '>') were excluded as no absolute value was derived. Identified data were scored on reliability based on criteria related to substance concentration, reported BCFs, and general test conditions. The following substance based criteria were used: 1) the water exposure concentration should be measured and not nominal; and 2) water exposure concentrations should be below water solubility limits (as estimated by WSKOW v1.42 from EPISuite [37]). With respect to the reported BCF values, the following criteria were applied: 3) reported BCF values should be substance specific (e.g. not based on total radiolabeled content); 4) when BCF steady-state values are reported, exposure duration needs to be sufficient to reach steady-state conditions (this aspect was analyzed similar as assessed by Arnot and Gobas [125]: when "steady state" was declared by the authors, or when time was sufficient to reach 80% of steady state according to model estimations [125]); 5) the BCF should be based on whole body content; and 6) lipid content of the fish should be reported. In addition, several experimental test conditions should be met. Total organic carbon content must be lower than 2 mg/L, pH should be between 6 and 8.5 at the start of the experiment, temperature must be close to the recommended ranges as reported in the OECD TG 305 [118] and must not be below 3°C or above 30°C, the dissolved oxygen concentration must be above 60% of saturation and no toxicity should be observed during the test. Data was included in case that ranges of organic carbon, pH or oxygen concentrations were reported that partially meet the criteria, or when these parameters were not reported. These quality screening criteria are comparable to those suggested in previous studies [125,127].

For substances with at least one reliable BCF value, we gathered additional data via the OECD QSAR Toolbox [128] and the US-EPA ECOTOX database [129] (data extracted on 02-04-2018). Retrieved BCF values were scored on reliability, similar as described above. Ultimately, only substances with three or more unique BCF values were used for further analysis, and substances with less BCF values were excluded.

5.2.2 Data extraction

For the included data, we collected parameters related to bibliographic data, chemical descriptors, test conditions and endpoint information (see Table S1). *Bibliographic data* includes the first-author, reference and year of publication. Information on the *chemical descriptors* consists of CAS number, substance name, SMILES, functional group (based on ECOSAR classifications), water solubility estimates, and log K_{ow} estimates [37]. The *test conditions* includes mean measured water concentration, radiolabeled substance (i.e. yes or no), exposure duration, temperature, pH, dissolved oxygen concentration, organic carbon content, lipid fraction, calculation method (i.e. steady-state or kinetic approach), combination exposure (i.e. exposure to a single substance or to a mixture), species at subcohort level, full-grown organism size (i.e. below or above 10cm [130]) and life stage. In addition, *endpoint information* includes the BCF values.

5.2.3 Data analysis

We used R [58] and the *nlme* package [131] to perform a linear mixed effects analysis of the relationship between the response variable (BCF value) and several test characteristics as fixed effects, including combination exposure, calculation method, species at subcohort level, life stage and full-grown organism size. The BCF values were lipid-normalized as advised in the OECD TG 305 and were log-transformed as standard deviation (SD) was correlated with BCF values (Figure S1). We used substances within functional classes as a random intercept in order to account for substance dependent differences within a functional class. This means that (average) BCF values are expected to differ per substance and that substances from the same functional class are expected to behave more similar than substances from a dissimilar functional class.

A three-step approach was followed. First, correlations between all fixed effects were investigated using bias-corrected Cramer's V. Of the five included variables, full-grown organism size was correlated (bias-corrected Cramer's $V > 0.7$) with life stage and combination exposure (i.e. single or mixture exposure) – and was excluded from further analysis.

Secondly, a candidate model set was constructed consisting of all possible additive combinations of fixed effects. Models with homoscedastic variances and heteroscedastic variances of the different fixed effects were included using the *varIdent* function. One model (full fixed effects and heteroscedastic variances for combination exposure, organism subcohort and calculation method) could not be run due to singularities. No interaction effects were included because of rank deficiency. All models were compared using the corrected Akaike Information Criterion (AICc), ranging from the null model (without any fixed effects) to the full model (including

all fixed effects). All models in the candidate set were fitted and then compared using AICc to determine the Kullback-Leibler (KL) best model [132]. The KL best model is the most parsimonious model (best fit to the data for the least number of parameters) given the model set. Additional models were considered to receive substantial support if the difference between model i AICc value and that of the KL best model (Δ AICc i) was < 2 [132].

Thirdly, we analyzed the contributions of the fixed effects on the means and SD for the best model and calculated marginal- and conditional-R^2. The marginal-R^2 describes the proportion of variance explained by the fixed effects and the conditional-R^2 describes the proportion of variance explained by both, the fixed and random effects. Visual inspection of residual plot of the best model did not reveal any obvious deviations of model assumptions. For relevant fixed effects, differences of the means were investigated using Tukey or Dunnett test for statistical variances.

5.3 Results

5.3.1 Included data

In total, 326 BCF values of 64 substances were included (details are given in Table S1). The BCF values ranged from log BCF 0.75 to 4.49 (i.e. BCFs ranging from 5.6 to 30625; data normalized to 5% lipid content) and estimated log K_{ow} values ranged from -2.15 to 6.79. For most substances three BCF values were included, though for some substances up to 23 BCF values were available. On average, two different references reported BCF values per included substance, with a maximum of six different studies. The substances covered eleven different functional groups (Figure 5.1A).

Different test conditions were applied in the included BCF studies (Figure 5.1B). Most BCF values were derived by a steady-state approach (n=299), whereas some were based on kinetic determinations (n=27). In addition, 149 BCF values were derived upon single substance exposure. Mixture exposures could be divided into organophosphate pesticides, halogenated organics and mixtures of polycyclic aromatic hydrocarbons (PAHs). Within the group of PAHs, two studies were included in which fish were exposed in combination with a potent mixed function oxygenase (MFO) stimulator, that mimics the metabolism induction of heavy weight PAHs (e.g. β-naphthaflavone) [133]. Furthermore, 17 different fish species were included, which could be divided into three groups based on subcohort level. The groups include the Neoteleostei, Protacanthopterygii and Ostariophysi (Figure S2). Common life stages include juveniles (n=187) and adults (n=114), though some studies used egg and/or larval stages (n=25). In addition, a clear balance was observed in the number of small (<10cm) and large species (>10cm), when considering their full-grown size. Within different

experiments, different combinations of test conditions were applied.

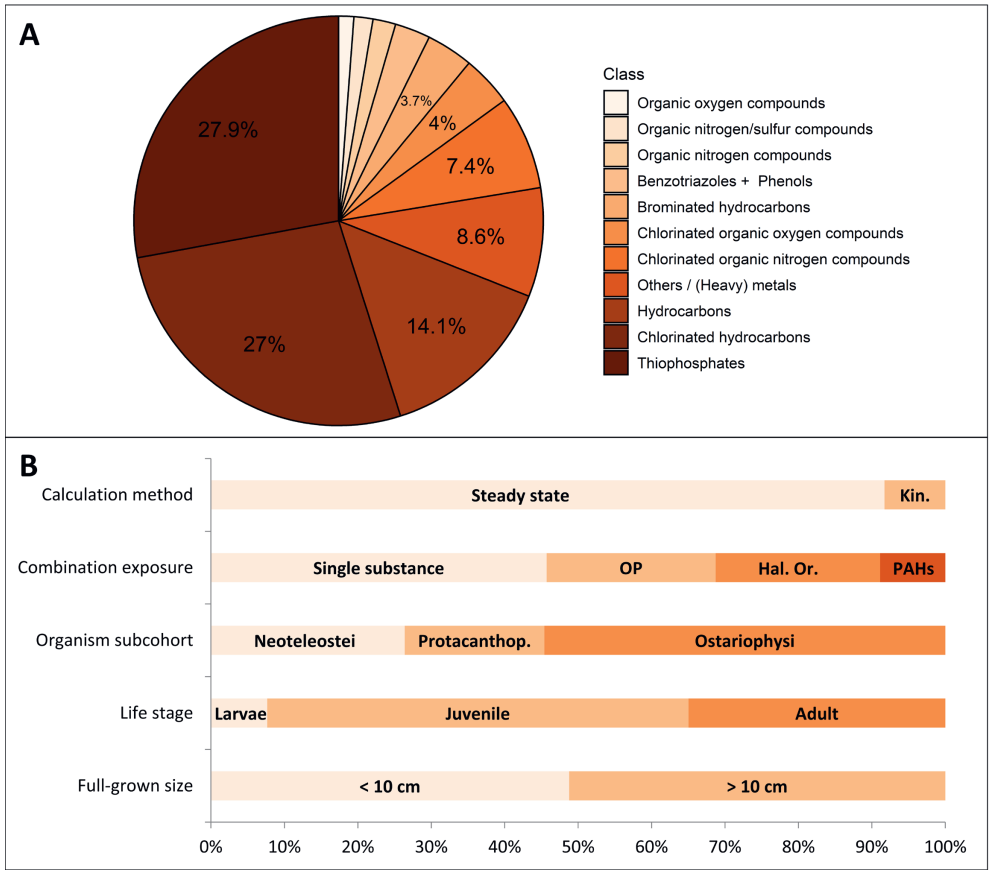


Figure 5.1: Overview of the included substances and BCF values. A) Overview of the functional classes of the different substances (n=64). B) Overview of the presence of different test conditions within the included test data (n=326). Kin. = Kinetic determination; OP = Organophosphate pesticides; Hal. Or. = Halogenated organics; PAHs = Polycyclic aromatic hydrocarbons (PAHs).

5.3.2 Explaining factors within BCF model

The results of the ten best descriptive models for bioaccumulation potential, based on AICc, are shown in Table 5.2. The top-ranked model included combination exposure, calculation method, organism subcohort, and life stage as fixed effects. Effectively, this means that those variables influence the BCF value. Furthermore, this model includes heteroscedastic variances for combination exposure, organism subcohort and life stage. Accordingly, differences in BCF variation (i.e. SD) are observed for different combinations of these variables. The top-ranked model had a marginal and conditional R^2 of 0.0974 and 0.843, respectively. Below, we discuss, for the top-ranked model, the differential effects of the included test characteristics

on obtained BCF values, and their variance.

Table 5.2: Overview of the top ten descriptive models. AICc = corrected Akaike Information Criterion. An "x" indicates the inclusion of a specific fixed effect within the model, or an allowance for heteroscedastic variances. Marginal and conditional R^2 are 0.0974 and 0.843, respectively, for the top-ranked model.

Rank	Fixed effects				Heteroscedastic variances				AICc	Δ AICc
	Combination exposure	Calculation method	Organism subcohort	Life stage	Combination exposure	Calculation method	Organism subcohort	Life stage		
1	x	x	x	x	x		x	x	282.5	0
2	x	x	x	x	x	x	x	x	285.3	2.8
3	x	x	x	x	x	x		x	290.3	7.8
4	x	x	x	x	x			x	294.6	12.1
5	x	x		x	x	x		x	294.6	12.1
6	x		x	x	x		x	x	296.4	13.9
7	x	x		x	x			x	297.9	15.3
8	x	x	x	x		x	x	x	300.5	18.0
9	x	x	x		x	x	x		303.0	20.5
10	x			x	x			x	307.9	25.4
…	…	…	…	…	…	…	…	…	…	…
80									364.9	82.4

Factors influencing bioconcentration

The set of test conditions were found to contribute differently to the BCF values. No difference in BCF value was observed when fish were exposed to a single substance or in a mixture with organophosphate pesticides or halogenated organics (Table 5.3). Substances that were tested in such mixtures were in general of the same class (i.e. organophosphate pesticides or halogenated organics, respectively). However, a significantly lower log BCF of 0.81 was observed upon exposure to a mixture of PAHs ($p < 0.0001$; Table 5.3). Further investigation revealed that PAHs were mainly tested simultaneously in combination with hydrocarbons and only ones in combination with an organic oxygen compound. For four substances, BCFs in our database had been generated upon single substance exposure as well as upon

exposure to a mixture of PAHs (Figure 5.2; including anthracene, dibenzofuran, fluorene and phenanthrene). From these substances it can be observed that the BCFs of three-ring PAHs (anthracene and phenanthrene) are much lower in case of exposure to a mixture of PAHs, whereas a small increase in BCF is observed in case of mixed exposure for the other substances.

Table 5.3: The effects on log BCF values of the test conditions that are included within the best descriptive model. Statistical analysis includes either Dunnett's test for combination exposure or Tukey's test for the other categories.

Group	**Comparison**		
	Compared to	**Effect in log BCF [SE]**	**p-value**
Halogenated organics	Single substance	0.07 [0.06]	0.2754
Organophosphate pesticides	Single substance	-0.11 [0.09]	0.2178
PAHs	Single substance	-0.81 [0.12]	*<0.0001*
Kinetic	Steady state	-0.27 [0.06]	*<0.0001*
Neoteleostei	Ostariophysi	0.15 [0.04]	*0.0003*
Ostariophysi	Protacanthopterygii	-0.17 [0.06]	*0.0022*
Protacanthopterygii	Neoteleostei	0.02 [0.05]	0.7767
Egg/larval stage	Juvenile stage	-0.08 [0.06]	0.2010
Juvenile stage	Adult stage	-0.02 [0.07]	0.7973
Adult stage	Egg/larval stage	0.10 [0.09]	0.2628

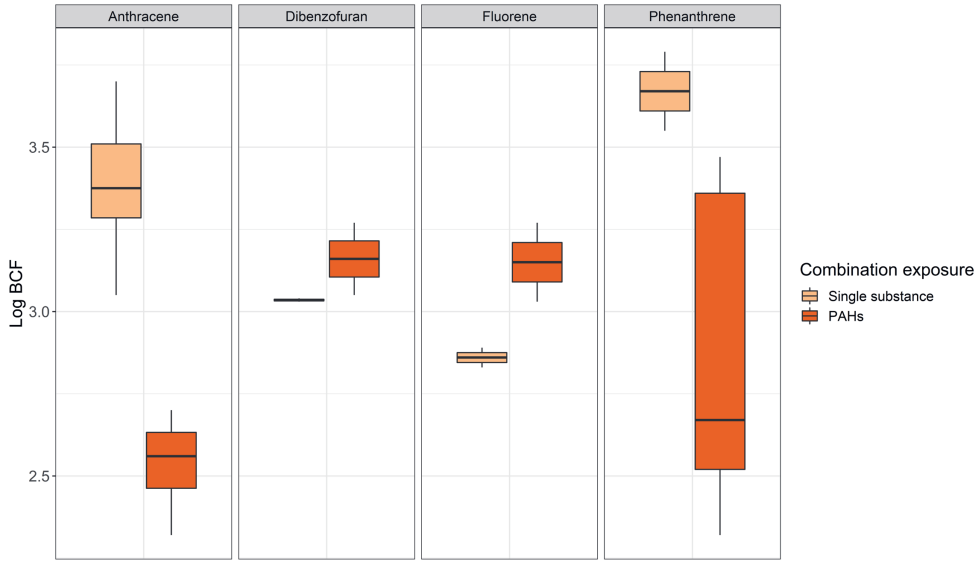


Figure 5.2: Overview of log BCF values for substances tested upon single substance exposure or upon exposure to a mixture of PAHs.

Furthermore, for 14 substances BCFs were determined via steady-state assessment as well as on the basis of kinetic approaches. The results indicate a significantly lower log BCF value of 0.27 when kinetically determined ($p < 0.0001$; Table 5.3).

The impact of subcohorts in the tests was found to result in different BCF values between organisms from the Ostariophysi as compared to the Neoteleostei and Protacanthopterygii subcohorts. For the group of Ostariophysi, which is mainly represented by the common carp (n=97; Figure S2), a lower log BCF of approximately 0.16 was observed ($p < 0.005$; Table 5.3). The Neoteleostei and Protacanthopterygii, which are mainly represented by the guppy and high-eyes medaka (n=37 and 23), and the rainbow trout (n=60), respectively, showed to have higher log BCF values.

Finally, life stage explains a certain amount of the variation in the data, as it is included as fixed effect within the top-ranked model. Lower BCF values are observed for egg and larval stages, compared to higher BCF values for adult fish (Table 5.3). However, no statistically significant differences of mean BCFs were observed between different life stages.

Variability in bioconcentration

Besides the influence of the test characteristics on the mean BCF values, also influences on SDs were estimated for different combinations of these characteristics. Within Table 5.4 all SDs are presented for groups of test characteristics with at least ten BCF values, which were corrected for dependent substance differences within functional classes. To clarify, when a substance will be tested multiple times in a BCF test using the following conditions: i) single substance exposure, ii) in an organism from the Neoteleostei subcohort, iii) at a juvenile life stage; a SD of 0.238 log BCF is expected to be observed based on available data. The observed SDs range from 0.090 to 0.343 log BCF with an average of 0.214 SD. To illustrate the average variation, 95% confidence ranges have been calculated in Table 5.5 for several BCF values, as based on 1.96 SDs of the mean.

5.4 Discussion

The aqueous exposure bioconcentration test is highly important for bioaccumulation assessments within regulatory frameworks. Nevertheless, little is known about the reproducibility and the factors within these laboratory experiments that affect the actual BCF value. Based on secondary data gathered within our database, we showed considerable impact of experimental design on the obtained BCF values and their variation. Specifically, mixture exposure, calculation method and the selected test fish species influenced the BCF values.

Table 5.4: SDs as calculated for combinations of test conditions that are considered relevant within the best descriptive model. Only groups of substances for which ten or more BCF values were available are included.

Combination exposure	Organism subcohort	Life stage	Number of BCF values	SD
Single substance	Neoteleostei	Juvenile stage	16	0.238
Single substance	Ostariophysi	Egg/Larval stage	12	0.090
Single substance	Ostariophysi	Juvenile stage	99	0.343
Halogenated organics	Neoteleostei	Adult stage	12	0.310
Halogenated organics	Protacanthopterygii	Juvenile stage	50	0.199
Organophosphate Pesticides	Neoteleostei	Adult stage	44	0.221
Organophosphate Pesticides	Ostariophysi	Adult stage	28	0.166
Organophosphate Pesticides	Ostariophysi	Adult stage	16	0.144

Table 5.5: The 95% confidence ranges of several BCF values based on the average SD of 0.214 log BCF.

BCF	Log BCF	Range ± 2xSD
100	2	38 – 262
500	2.7	191 – 1312
2000	3.3	762 – 5249
5000	3.7	1905 – 13122
10000	4	3810 – 26244

5.4.1 Influencing factors

Mixtures

A significantly lower log BCF of 0.81 was observed when the test substance was co-exposed with 4- or 5-ring PAHs. This was specifically observed for the 3-ring PAHs anthracene and phenanthrene (Figure 5.2). Earlier research indicated that single exposure to 3-ring PAHs did not stimulate the MFO system, whereas it was stimulated in combination with 4 or 5-ring PAHs [133]. Specifically, the MFO systems aryl hydrocarbon hydroxylase (AHH) and aniline hydroxylase (AH), as well as cytochrome P450 levels were induced by high molecular weight PAHs, including pyrene, chrysene and benzo(a)pyrene [133]. The MFO system is known to metabolize aromatic hydrocarbons by oxygenation and does not only act on higher weight PAHs. Consequently, lower BCF values are observed for 3-ring PAHs within a mixture of higher weight PAHs. Although no specific contributions were identified for mixtures containing organophosphate pesticides or halogenated organics, these findings suggest that results of mixture experiments should be interpreted with caution.

Calculation method

BCF values calculated based on kinetics resulted in lower log BCF values than determined by steady-state analysis. In theory, both approaches should provide similar results when uptake follows first-order kinetics and when steady-state BCFs are really based on steady-state data [19]. As it might be uncertain whether steady state is reached – especially for hydrophobic substances – kinetic BCF values are generally preferred [19]. If steady-state levels would not be achieved, one would expect to observe a lower BCF value for steady-state determinations. Nonetheless, we observed the opposite.

Potentially, the observed difference could be explained by a peak in fish concentration prior to achieving plateau levels. Such a phenomenon is regularly observed, and could be related to an interactive relationship between bioaccumulation kinetics and metabolic enzyme activities [134]. When a steady-state BCF is determined within this peak, a higher BCF value might be obtained compared to kinetic BCFs (Figure S3).

Organism subcohort

Data analysis revealed a significant difference in BCF values for species from varying subcohorts, with lower values for species from the Ostariophysi. This effect is likely related to differences in toxicokinetics.

The uptake of chemicals via the gills is generally related to the ventilation rate and the uptake efficiency [135]. The ventilation rate is described as the amount of water per time unit that is ventilated through the gills. The ventilation rate may differ across species, with higher rates for more active species [135]. The uptake efficiency, in the form of blood-water partitioning, is not assumed to vary between species for substances with a log K_{ow} above 3 [135,136].

Differences in depuration could be related to variances in metabolic activity among species, due to the presence of different biotransformation enzymes. Although many of those enzymes are very much conserved, different isoenzymes have been identified within different fish species, including different cytochrome P450 enzymes, glutathione S-transferases and ABC-transporters [137,138]. The presence and absence of many of those isoenzymes are related to the phylogeny of the species, and the activity of isoenzymes is thus likely to vary between different subcohort levels. As a consequence, varying V_{max} (i.e. the maximum reaction rate at saturating substrate concentration) and K_m levels (i.e. the substrate concentration at which the reaction rate is half of V_{max}) can be observed for different species [139,140]. For instance, differences have been observed within the metabolism of methoxychlor by the rainbow trout and the common carp, showing different metabolic profiles [141]. Only one metabolite was observed within rainbow trouts, whereas several metabolites were identified within carps. Despite information on the presence of different isoenzymes among (classical) fish species,

we lack knowledge on complex metabolic pathways of many substances and species. Better insight in these processes is considered valuable for risk management to quantify the variation across species.

Life stage

No significant effect of life stage on BCF values was seen, although a tendency of lower BCF values for the egg/larval stage, followed by juveniles and adult fishes was observed. Potentially, lower BCF values can be observed for early-life stages due to a larger growth capacity, resulting in growth dilution [19]. Furthermore, earlier research suggests that different life stages have different metabolic capacity, with varying V_{max} and K_m values [139]. However, also comparable differences in uptake rates have been observed [142], potentially resulting in comparable BCF values across life stages. Because of the comparable outcomes across life stages, the use of egg/larval stages might become of future interest to replace the standard *in vivo* bioconcentration test with non-protected *in vivo* systems [124].

5.4.2 Variability in bioconcentration

When considering the contribution of the different fixed effects, an average SD of 0.214 log BCF was determined. This variation is in line with the results of the OECD ring-test as conducted in 1985 by Kristensen and Nyholm [143]. Within this study, lindane was analyzed by 12 different laboratories testing one or two concentrations, resulting in a total of 22 BCF values. In addition, an optional chemical, 2,3,4,5-tetrachlorophenol (TeCP), was analyzed by four different laboratories, with in total seven BCF values. When normalizing the results to 5% lipid content, and only including the data that met the quality criteria of < 20% fluctuation in water concentration, a SD of 0.20 log BCF can be derived for lindane (n=19). For TeCP no reliable data could be retrieved according to the report [143]. The SD of 0.20 log BCF values, as derived under very strict conditions, is similar to our results.

While the above described test characteristics influence the BCF values, the remaining variation of 0.214 SD can be explained by other variables that were not yet considered in our analysis. Several factors have been suggested to potentially influence bioconcentration, including water-to-fish ratios [143], temperature [139,144,145], sex differences [139,144,145], feeding procedure (i.e. food item, feeding rate and feeding quantity) [146,147], and slight experimental variances in water chemistry and dissolved oxygen concentrations [148]. Most of these variables are expected to (in)directly influence the metabolic capacity of the organisms, and/or are directly related to changes in activity and oxygen consumption [149]. Indirectly, some of those factors might be partially covered by the inclusion of subcohort levels within the analysis. However, we can currently only speculate on the relative importance of all these variables, as many of them are not (consistently) reported. In addition, growth

dilution is known to significantly influence bioconcentration, especially for substances with a high bioaccumulation potential and for test organisms at early-life stages [118,150]. However, this parameter is scarcely reported and was therefore not included in the analysis. Moreover, part of the variability could potentially be related to variances in exposure concentration. As in theory the BCF is a net result of uptake and elimination rates, which are independent of exposure concentration [125], we did not consider this factor in the current analysis. However, concentration dependent BCFs could be of potential importance, specifically for polar chemicals, or for chemicals that undergo metabolic conversion when internal threshold concentration are attained [118].

In addition, it is expected that a significant amount of variation is related to intra-species differences. For instance, a two to three-fold variation is typically observed in the standard as well as maximum metabolic rate between individuals of the same fish species [151]. Individual differences are likely related to differences in genes and developmental conditions [151]. This may result in biological differences, like individual differences in isoenzyme content [139], and/or differences in behavior, like aggressiveness, boldness and (spontaneous) activity [151,152]. These factors are known to influence metabolic rates within organisms and subsequently affect ventilation rates, and thus may influence bioconcentration. A more accurate mean BCF (less influenced by the effect of individual differences) can be obtained by analyzing explicitly the biological variation within test organisms or by pooling or taking the mean of more samples [143], though sampling bias, due to behavioral differences, should be considered [153].

Besides the factors mentioned above, variation and uncertainty could also be related to laboratory practices, like fish maintenance, chemical analysis and data reporting. For instance, inadequate removal of uneaten food and/or feces may result in significant levels of organic carbon, limiting the bioavailability of the test substance [118,125,154]. Also differences in the analytical techniques (measuring chemical concentrations in water and fish), can contribute to the variation. Although it is generally assumed that the analytical methods are sufficiently optimized, variation may especially be observed for substances with a low water solubility. Moreover, we currently assumed that the selected water quality criteria (i.e. organic carbon, pH, temperature and dissolved oxygen concentration) were sufficiently strict to guarantee a limited influence on the BCF variability. Although some studies reported a range of water quality parameters that only partially met the criteria (n=5; see Table S1), exclusion of these values did not resulted in any changes on effect directions and significance levels. Nevertheless, also multiple studies did not report one or several water quality parameters and – following our approach – were included in the data analysis. This interpretation is a potential source of uncertainty, as extreme values for water quality parameters could significantly influence BCF variability [125]. We therefore encourage to report the water quality parameters in detail in

future studies.

5.4.3 Consequences for regulation and recommendations

When converting the SD to a 95% confidence range, an uncertainty of ± 0.419 log BCF is obtained (i.e. 1.96xSD; Table 5.5). This variation questions the robustness of the current B-assessment within regulatory frameworks, in which a single BCF value is generally sufficient to derive a conclusion. For example, a BCF value of 2500, which is normally interpreted as 'bioaccumulative', could also be considered as 'not bioaccumulative' and 'very bioaccumulative' based on the 95% confidence range (953-6561). The use of multiple experiments and/or species would be valuable for the B-assessment. Including more studies in order to encapture variability, has also been suggested for sediment quality assessments [155]. Potentially, new alternative bioconcentration methods based on invertebrate *in vivo* experiments could be valuable within such assessment, as they are less expensive and time consuming, and do not consider vertebrate testing [123]. The test performance of such methodologies could be compared and evaluated in the light of the performance of the current gold test standard as analyzed within this study (i.e. the aquatic exposure fish bioconcentration test). Specifically the use of alternative – non-vertebrate – bioconcentration tests should be stimulated, in order to further support the 3R principles (i.e. replacement, reduction and refinement of animal studies) [156,157]. Furthermore, we highlight that future studies should explicitly state and justify all experimental decisions and conditions, specifically also with respect to species-selection and simultaneous testing of substances. This is key, to improve the number of valid BCFs in databases.

5.5 Conclusions

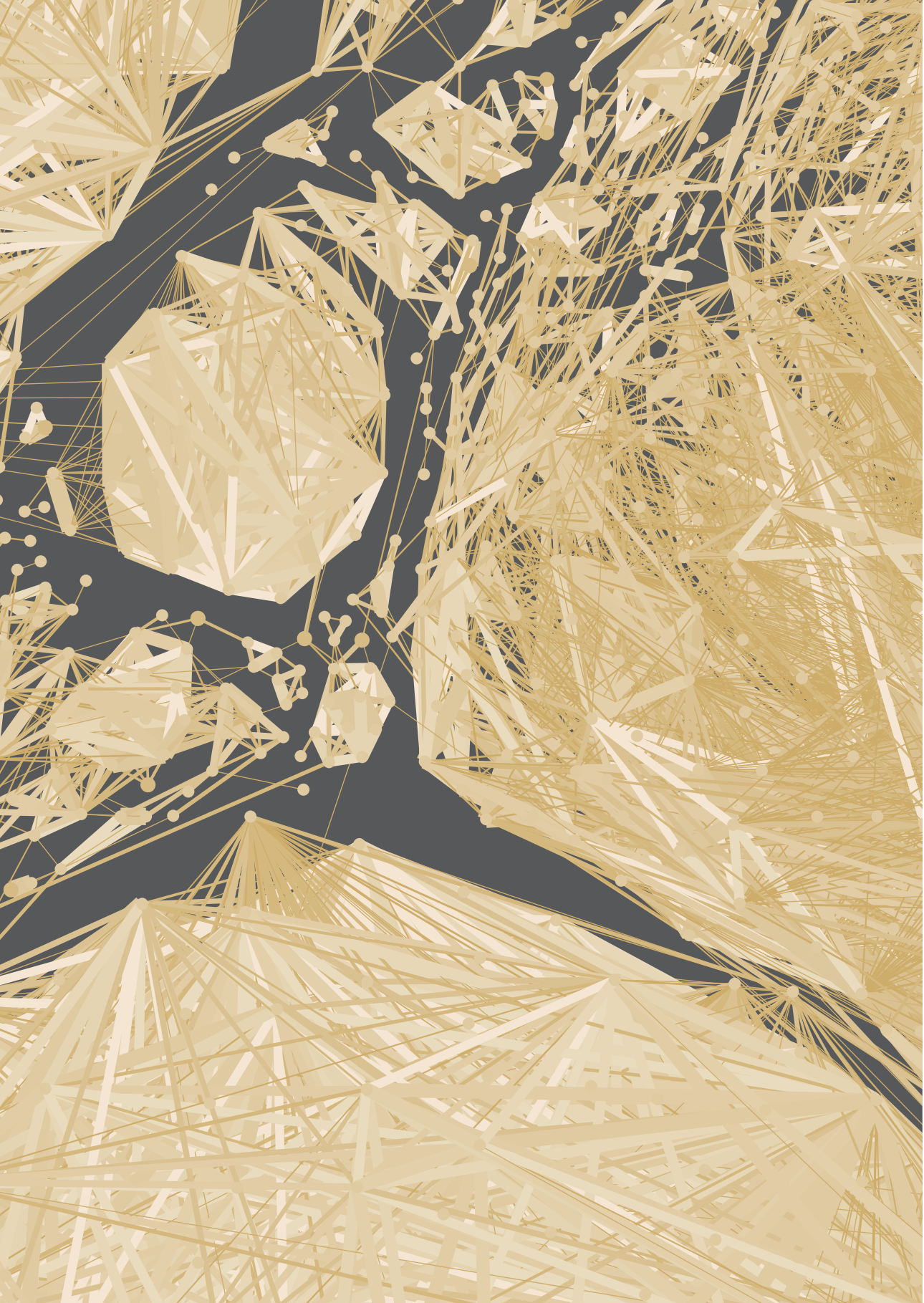
Although guidance documents on bioaccumulation studies exist for many years and many studies have been performed accordingly, a review on reproducibility was lacking. Nonetheless, there is a crucial role of bioaccumulation assessment within regulatory frameworks. Our assessment indicates that several factors are influencing the bioconcentration potential, each of which should preferably be considered when interpreting the test results. The robustness of an experimentally determined bioaccumulation potential – although following the strict guidelines – is less than expected. We revealed a high variation in BCF values, with an average SD of 0.214 log BCF, within the fish bioconcentration test. Species selection and test designs where multiple substances are tested simultaneously showed to be important aspects leading to variation. The typical variability within BCF values results in high uncertainty in the B-assessment within regulatory frameworks. We, therefore, recommend the use of test species from at least two different subcohorts, including vertebrates or invertebrates.

Acknowledgements

This work was partially funded by the Dutch Ministry of Infrastructure and Water Management.

Supplemental material

Supplementary data to this chapter can be found online at https://doi.org/10.1016/j.chemosphere.2019.124731.



6

Persistence, Bioaccumulation and Toxicity-Assessment of Petroleum UVCBs: a Case Study on Alkylated Three-Ring PAHs

Pim N.H. Wassenaar and Eric M.J. Verbruggen

Published in Chemosphere 276 (2021), 130113.

Abstract

Substances with (very) persistent, (very) bioaccumulative, and/or toxic properties (PBT/vPvB) are of environmental concern and are identified via hazard-based PBT-assessment approaches. The PBT-assessment of well-defined substances is optimized over the past decades, but is under development for substances of unknown or variable composition, complex reaction products or biological materials (UVCBs). Particularly, the large number of constituents and variable composition complicate the PBT-assessment of UVCBs. For petroleum UVCBs, the use of the hydrocarbon block method (HBM) is proposed. Within this method, groups of constituents with similar physicochemical properties and structure are treated as a single entity and are expected to have comparable environmental fate and hazard properties. So far, however, there is a lack of experience with the application of the HBM for PBT-assessment purposes.

The aim of this study is to investigate the suitability of the HBM for the PBT-assessment of petroleum UVCBs by evaluating the group of alkylated three-ring polycyclic aromatic hydrocarbons (PAHs). The presented approach is based on experimental data and model predictions and followed the guidelines of the European Chemicals Agency.

Because of a lack of relevant experimental data, relative trend analyses were applied. The results indicate that alkylated three-ring PAHs are more persistent, bioaccumulative, and toxic than the parent three-ring PAHs. As the parent three-ring PAHs are currently identified within Europe as PBT/vPvB substances, the alkylated three-ring PAHs could also be considered as PBT/vPvB. Accordingly, this case study provides the prospects for the application of the HBM for the PBT-assessment of UVCBs using trend analysis.

6.1 Introduction

Regulatory priority is given to substances that are (very) persistent in the environment, (very) bioaccumulative in organisms, and/or toxic to the environment or to humans (i.e. PBT/vPvB) [29]. Once emitted, PBT/vPvB substances cannot easily be removed from the environment, and are likely to reach high and potential toxic concentrations in organisms or humans upon continued emission [19]. Therefore, regulatory agencies try to identify PBT/vPvB substances in order to take relevant regulatory measures.

PBT/vPvB substances are identified following a hazard-based PBT-assessment [19]. Within this assessment, chemical persistence is evaluated based on environmental half-lives in different environmental compartments. Bioaccumulation is generally assessed based on bioconcentration in aquatic organisms, whereas toxicity is evaluated based on toxic effects to aquatic organisms or specific toxicity to mammalian species including humans. Within the PBT-assessment, these properties are compared to specified criteria to determine whether a substance is PBT/vPvB [19].

Within REACH, the European regulation on industrial chemicals (EC/1907/2006), the PBT-assessment principally considers a single constituent assessment. This means that a PBT-assessment needs to be conducted for all constituents within a substance that are present above a concentration of 0.1% (w/w) [19]. Although this is seemingly clear for well-defined substances, like mono-constituent and (to a lesser extent to) multi-constituent substances, it is more complex for substances of unknown or variable composition, complex reaction products or biological materials (UVCBs). UVCBs contain a large number of constituents, often ranging from hundreds to thousands in number [158], of which a significant fraction could be unknown, and/or their concentrations could be variable or unpredictable in the composition [19]. Technically it is not possible to identify, isolate and test all individual constituents [19,159], which are often individually present below 0.1% (w/w). However, as the individual constituents are generally very similar in structure to many other constituents, with sum concentrations of structurally similar constituents frequently above 0.1% (w/w), they are considered relevant for the PBT-assessment. This structural complexity complicates the PBT-assessment of UVCBs in comparison to well-defined substances.

Despite the complexity, PBT-assessment approaches for UVCBs are indispensable as approximately 40% of all REACH registered substances are considered UVCBs [160]. One specific group of UVCBs are petroleum-derived substances. Petroleum substances are considered UVCBs as their composition is highly variable (depending on the source and batch of crude oil, as well as specific production processes) and partially unknown, as it is not possible to identify each individual constituent. Petroleum UVCBs mainly contain

hydrocarbons in the form of paraffins (alkanes), naphthenes (cycloalkanes) and/or aromatics, but can also contain other hydrocarbon structures like naphthenic-aromatics (Figure 6.1A). Several of these hydrocarbon constituents are of potential PBT/vPvB concern due to their physicochemical properties [19]. In addition, these hydrocarbons are potentially emitted to a high extent as petroleum-derived substances are used in large quantities. In 2013, 971 million tons of petroleum-derived substances were manufactured or imported into the European Union. The highest fraction is applied as fuel (64%, 618 million ton) and approximately 4% (37.5 million ton) for industrial or widespread uses (i.e. professional and consumer applications). The remaining fraction is registered for intermediate uses, meaning that they are further refined on site into other product types (32%, 315 million ton) [161]. These products generally consider more refined UVCBs and are more likely to have widespread applications. Because of the widespread applications and presence of potential PBT/vPvB-constituents, PBT-assessments of petroleum UVCBs are essential.

The Technical Guidance Document of the European Chemicals Agency (ECHA) on the assessment of PBTs/vPvBs provides information on approaches to assess UVCBs [19]. This includes the 'known-constituents approach', the 'whole-substance approach' and the 'fraction profiling approach', also known as the hydrocarbon block method (HBM). The latter approach is specifically suited for petroleum substances. The HBM resolves complex petroleum substances into pseudo-constituents ('blocks') that are defined by, and assessed based on, representative hydrocarbon structures exhibiting similar physicochemical properties [19,162]. Generally, the constituents are grouped based on their chemical class and number of carbon atoms (Figure 6.1A). The underlying assumption of the HBM is that all constituents within a block have fairly similar physicochemical properties, and to a certain extent also a fairly similar biodegradability (P), bioaccumulation (B) and aquatic toxicity (T) potential. Accordingly, a block of constituents could be assessed as if it were a single constituent and the PBT properties of the block/representative constituent could be compared to the PBT criteria, similar to mono-constituent substances [19]. So far, however, there is a lack of experience with the application of the HBM for PBT-assessment purposes of petroleum UVCBs.

This study aims to investigate the suitability of the HBM for the PBT-assessment of petroleum UVCBs by evaluating the group of three-ring polycyclic aromatic hydrocarbons (PAHs). We specifically selected this group of hydrocarbons as there is a relatively large volume of data available in comparison to other hydrocarbon categories. The group of three-ring PAHs includes the parent three-ring PAHs (i.e. non-alkylated anthracene and phenanthrene) and all alkylated derivatives up to four extra carbon atoms (i.e. C14-C18/P-C4; Figure 6.1B).

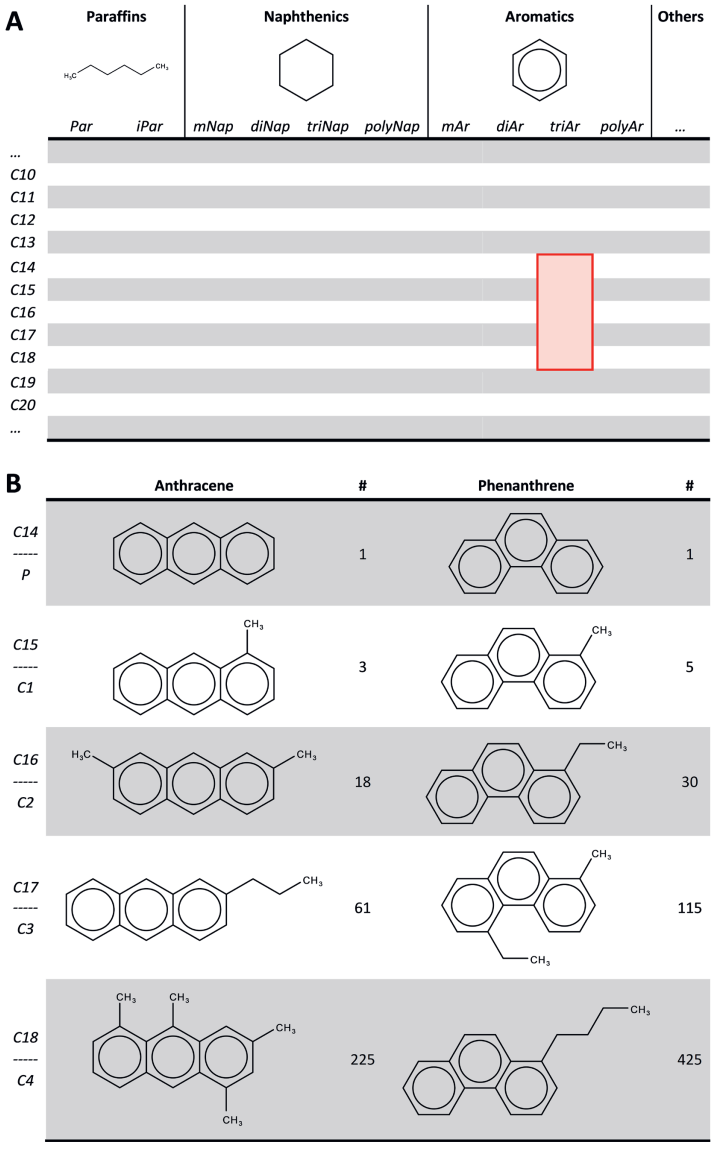


Figure 6.1. A) Hydrocarbon block method (HBM), with chemical classes in columns and number of carbon atoms in rows. The red rectangle highlights the hydrocarbon blocks that represent the three-ring polycyclic aromatic hydrocarbons (PAHs) that are investigated within this study (more details on these blocks are shown in Figure 6.1B). Par = normal alkanes or paraffins; iPar = branched alkanes or paraffins; mNap = mono-naphthenics; diNap = di-naphthenics; triNap = tri-naphthenics; polyNap = poly-naphthenics; mAr = mono-aromatics; diAr = di-aromatics; triAr = tri-aromatics; polyAr = poly-aromatics. B) Three-ring PAHs C14-C18, with representative structures of the different blocks and the total number of structures belonging to the block (#). P are the parent substances anthracene and/or phenanthrene, C1 three-ring PAHs have one extra carbon atom, C2 contain two extra carbon atoms, C3 contain three extra carbon atoms and C4 contain four extra carbon atoms. The extra carbon atoms can be present in the form of methyl, ethyl, n-propyl, isopropyl, n-butyl, sec-butyl, isobutyl and tert-butyl chains.

6.2 Methods

Within this study, a PBT-assessment on the group of three-ring PAHs was conducted following the ECHA Technical Guidance Document [19]. The PBT-properties of the parent (i.e. non-alkylated) three-ring PAHs have been investigated for decades, and anthracene and phenanthrene are currently identified within Europe as PBT and vPvB, respectively [163]. Although the PBT/vPvB-status of phenanthrene is under discussion [164,165], the current PBT/vPvB-status of anthracene and phenanthrene were considered as starting point for our assessment. The current PBT-assessment mainly focusses on the PBT-properties of the alkylated three-ring PAHs, which might be equally or more abundant in crude oils than parent three-ring PAHs [166]. We specifically focus on alkylated three-ring PAHs with up to four extra carbon atoms (in the form of methyl, ethyl, n-propyl, isopropyl, n-butyl, sec-butyl, isobutyl and tert-butyl chains), as those are most frequently encountered (i.e. C14-C18/P-C4) [167]. In total, this category includes 884 unique constituents (i.e. all hypothetical constituents with up to four extra carbon atoms with the above-mentioned alkyl substituents) which are grouped based on the number of carbon atoms (Figure 6.1B). Within this grouping (and throughout the text), P are defined as the parent substances anthracene and/or phenanthrene and contain 14 carbon atoms. The alkylated three-ring PAHs are defined as C1, C2, C3 and C4. The C1 three-ring PAHs have one extra carbon atom compared to the parent three-ring PAHs (with 15 carbon atoms in total), C2 contain two extra carbon atoms (16 carbon atoms in total), C3 contain three extra carbon atoms (17 carbon atoms in total) and C4 contain four extra carbon atoms (18 carbon atoms in total).

Within this PBT-assessment, we analyze the P-, B- and T-properties separately, using experimental data and model predictions. Model predictions were conducted to substantiate experimental data and/or to fill experimental data gaps. Experimental data were gathered via internet searches and references in relevant literature, and were evaluated for quality and relevance (see below). Searches combined substance related keywords (e.g. alkylated three-ring PAHs, alkylated anthracene, alkylated phenanthrene, substituted anthracene, substituted phenanthrene, etc.) with test related keywords (e.g. degradation, bioaccumulation, bioconcentration, toxicity, etc.), and optionally the compartment of interest (e.g. water, aquatic, fish, invertebrates, etc.).

6.2.1 Persistence

Experimental laboratory biodegradation data of alkylated three-ring PAHs in water, sediment and/or soil compartments were collected. The studies were evaluated based on their relevance for P-assessment purposes according to the ECHA Technical Guidance Document [19] and aspects from related OECD Test Guidelines (e.g. OECD TG 307, 308 and 309 [168–170]).

Environmental relevance of the reported biodegradation half-life values has been assessed based on the aspects as provided in Table 6.1. All criteria apply to aquatic biodegradation studies and where relevant also to sediment and soil biodegradation studies. When degradation data were reported in terms of depletion over time (i.e. in percentage), or in case of data fits that do not directly provide a well-defined half-life, we calculated half-lives using GraphPad Prism (v8), where possible, using first-order degradation kinetics. In addition, biodegradation half-lives were normalized to 12°C by using the Arrhenius equation [171], which generally showed to be applicable to convert the environmental half-lives of hydrocarbons [172]. Based on this analysis, studies were classified as either 'relevant to determine environmental degradation half-life values', 'not-relevant to determine environmental degradation half-life values' or 'not assignable' in case of missing data/details. Furthermore, photodegradation data on alkylated three-ring PAHs were collected and analysed. However, as the contribution of photodegradation to the overall degradation can be considered negligible in several environments [19,171], the process of photodegradation is not considered in the persistence conclusion [19].

In addition to experimental data, we predicted the degradation potential for all 884 constituents of the category by using two models, Biowin3 and BioHCwin [37]. Biowin3 aims to predict the required time for complete ultimate biodegradation in a typical aquatic environment. The results are given as a quantitative value, ranging from 1 to 5, and P-screening criteria have specifically been adopted for this model (i.e. <2.25-2.75 [19]). These semi-quantitative ratings have been transformed to half-life values using Equation 6.1, which is modified from Rorije et al. [38]. BioHCwin, on the other hand, is specifically designed for the prediction of primary biodegradation of hydrocarbons. Although this model does not provide a half-life for a specific compartment, the outcomes are generally interpreted as half-life values in freshwater compartments [173].

$$Water\ HL = 5378 \cdot e^{-1.95 \cdot Biowin3} \qquad (6.1)$$

Table 6.1. Relevance assessment of experimental biodegradation data for PBT-assessment purposes. Definitions: ↑ = The application/presence of this aspect will result in an increase in the observed degradation. ↓ = The application/presence of this aspect will result in a decrease in the observed degradation. — = No effect on the observed degradation, or no conclusion is drawn based on the application/presence of this aspect. X = This aspect is not considered relevant for determining environmental degradation half-lives (i.e. not representative of environmental conditions).

Aspect		Effect on degradation	Reasoning	Conclusion on study relevance
1	Application of high non-dissolved concentrations	↓	Non-dissolved constituents are not available for degradation.	X
2	Addition of mineral media	↑	Addition of mineral media increases observed degradation due to an increase in microbial growth. Simulation tests are performed in field sampled water, sediment or soil.	X
3	Application of crude oil or a complex mixture of hydrocarbons	↑↓	Application of crude oil influences the solubility of varying constituents, and can influence the composition of the microbial community resulting in an increase or decrease of degradation.	X
4	Addition of a dispersant	↑	Addition of a dispersant increases observed degradation due to an increase in bioavailability, and co-metabolism might occur concurrent with degradation of the dispersant.	X
5	Pre-adapted or irrelevant inoculum	↑↓	A pre-adapted inoculum increases observed degradation. An irrelevant inoculum can result in higher or lower observed degradation (e.g. artificial seawater).	X
6	Observed evaporation	↑	Inclusion of evaporation in degradation calculations will underestimate the degradation half-life.	X
7	Assumption of lag-phase	↑	Lag-phases are a sign of adaptation. If a lag-phase is assumed but does actually not occur, half-lives are underestimated. In case of no adaptation, lag-phase should be part of the half-life. If significant adaptation occurs see aspect 5.	X —
8	Normalization to an internal marker with varying physico-chemical properties	↑↓	Normalization to an internal marker may increase or decrease observed biodegradation due to differences in physicochemical properties (e.g. sorption/evaporation).	—
9	Absence of abiotic controls	↑—	Unknown whether there might be partial dissipation.	—

6.2.2 Bioaccumulation

The evaluation of experimental bioconcentration data on alkylated three-ring PAHs was limited to laboratory tests, focusing on fish or invertebrate species. Both steady state and/or kinetically determined bioconcentration factors (BCF) were extracted or, where possible, derived from raw data using GraphPad Prism. Besides aquatic exposure experiments, also dietary exposure studies were analyzed. Results of dietary exposure studies were transformed to BCFs using the OECD BCF Estimation Tool [174], as within REACH the B-criteria is based on BCF values for aqueous species. From the BCF Estimation Tool we reported the range of estimated BCF values and specifically highlight the prediction of the model by Sijm et al. [175]. The relevance of the data was scored based on criteria suggested and applied in previous studies [125,127,176] and are related to OECD TG 305 and the ECHA Technical Guidance Document [19,118]. The criteria were applied to both aqueous and dietary exposure experiments, where relevant. The following substance based criteria were used: 1) the test exposure concentration should be measured and should not be nominal; and 2) water exposure concentrations should be below three times the estimated water solubility (as predicted according to Verbruggen et al. [177]). With respect to the BCF values, the following criteria were applied: 3) reported BCF values should be substance specific (e.g. not on total radiolabeled content); 4) when BCF steady-state values are reported, exposure duration should have been sufficient to reach steady-state conditions; 5) the BCF should be based on whole body content; and 6) lipid content of the exposed species should be reported. In addition, several experimental test conditions should be met: total organic carbon content must be lower than 2 mg/L, pH should be between 6.0 and 8.5 at the start of the experiment, temperature should be within the recommended ranges (e.g. as reported in the OECD TG 305 [118]), and no toxicity should be observed during the accumulation test. Furthermore, experiments in which organisms were simultaneously exposed to chemicals that stimulate the mixed-function oxygenase (MFO) enzyme system, like four- and five-ring PAHs, were considered not relevant. Earlier research indicates that such exposures result in lower BCFs for three-ring PAHs compared to single substance exposures due to enhanced metabolic activation [133,176]. Based on this analysis, studies were classified as either 'relevant to determine the bioconcentration potential', 'not-relevant to determine the bioconcentration potential' or 'not assignable' in case of missing data/details.

In addition to the experimental data, we estimated the bioconcentration potential of this group with the EU-BCFmax and BCFBAF EpiSuite models [37,38]. The EU-BCFmax model only considers passive uptake and excretion of a chemical, and can be calculated from the octanol-water partition coefficient (K_{ow}). Using log K_{ow} predictions of EpiSuite [37], the EU BCFmax was calculated with Equation 6.2, as derived from Rorije et al. [38]. The BCFBAF EpiSuite software consists of two BCF models, the BCFBAF Arnot-Gobas model and BCFBAF

regression [37]. As opposed to the EU BCFmax model, the BCFBAF Arnot-Gobas model considers multiple factors, including uptake and elimination through gills, fecal egestion, growth and biotransformation, and other factors like bioavailability and gills absorption efficiencies. As input, the model requires log K_{ow} and first-order metabolism rate constant (kM) predictions from EpiSuite [37]. We predicted the bioconcentration for the upper trophic level, which were normalized to a 5% lipid content. The BCFBAF regression model, on the other hand, is developed based on a regression through a dataset of experimental BCF values for non-ionic substances versus log K_{ow} values from EpiSuite [37].

$$Log\ BCFmax = \frac{34.43}{2.93 \cdot (2\pi)^{0.5}} \cdot e^{-0.5 \cdot \left(\frac{Log\ Kow - 6.52}{2.93}\right)^2} \quad (6.2)$$

6.2.3 Toxicity

The toxicity assessment solely focused on aquatic toxicity and did not consider effects on human health. Aquatic toxicity data on alkylated three-ring PAHs were gathered and evaluated, focusing on acute, sub-chronic and chronic exposures to organisms from varying trophic levels. In case of sediment exposure experiments, data were converted to pore water concentrations using equilibrium partitioning in order to provide indicative water toxicity data. To apply the equilibrium partitioning method, we derived organic carbon-water partition coefficients (K_{oc}) according to Verbruggen et al. [177], and used the fraction of organic carbon as reported in the respective toxicity studies. The reliability of identified studies was scored according to the technical guidance for deriving environmental quality standards [178]. The following criteria were used: 1) exposure concentrations should be measured or stable concentrations could be guaranteed (e.g. Teflon-lined capping and renewal or flow-through exposure protocol); 2) effect concentrations should be below three times the estimated water solubility (as predicted according to Verbruggen et al. [177]); 3) effects should be exposure concentration related; and 4) effects/mortality under control conditions should not exceed recommended values from relevant guidelines. Based on this analysis, studies were classified as either 'reliable', 'not-reliable' or 'not assignable' in case of missing data/details. For studies that were considered as reliable, NOEC or EC10 values were extracted or derived with GraphPad Prism from original data, where possible. In addition, we made a distinction between studies with and without UV-exposure, as PAHs may exert a phototoxic effect [179]. UV-exposure regimes were considered relevant for exposure at the earth's surface and shallow surface water, when the applied intensities were below full-strength sunlight as described in literature [180–184].

In addition, we predicted the ecotoxicity of the group of three-ring PAHs by using the Target Lipid Model (TLM) [185] and an Internal Lipid Residues Model (ILRM) [177,186]

(see Supplemental Material S1). Both models have been developed to estimate the baseline toxicity caused by hydrocarbons/PAHs, and are based on species sensitivity distributions (SSDs) of internal membrane/target lipid concentrations against fraction of affected species. The TLM is based on acute and sub-chronic toxicity data for 54 species, mainly based on hydrocarbon toxicity (but also includes effects of several other chemicals) and results are extrapolated to chronic effects by using an acute-to-chronic (ACR) ratio of 5.22. The ILRM is based on a SSD of chronic PAH toxicity and total petroleum hydrocarbon (TPH) toxicity values for 58 different aquatic, benthic and terrestrial species. In order to estimate the fraction of species that is affected at a specific exposure concentration, only the log K_{ow} is needed. The TLM is based on log K_{ow} estimates from EpiSuite [37], and the ILRM is based on log K_{ow} estimates from ClogP Bioloom [187], which are subsequently used to predict membrane water partition coefficients (K_{mw}) [177]. For all 884 three-ring PAH constituents, we predicted the fraction of affected species at an exposure concentration of 10 µg/L (i.e. T-criteria [19]) or at maximum water solubility. For both models, water solubility was estimated based on log K_{ow} from ClogP Bioloom estimates according to Verbruggen et al. [177], which are slightly more conservative for the alkylated three-ring PAHs than the WSKOW EpiSuite estimates, which are normally applied in the TLM. For the TLM, this only results in minor differences in the predicted fraction of affected species for C4 three-ring PAHs (< 3% difference). In addition, we applied the median affected fraction per constituent, where the TLM generally uses the lower confidence limit of the affected fraction to calculate the impact [185].

6.3 Results & Discussion

6.3.1 Persistence

Experimental data

In total, 18 studies were identified that investigated degradation of alkylated three-ring PAHs, of which 16 studies analyzed biodegradation in water and one study in either sediment and soil compartments. In addition, three studies were identified that analyzed photodegradation of alkylated three-ring PAHs. The experimental data generally report individual half-life values or percentages of degradation for the parent three-ring PAHs (i.e. anthracene and/or phenanthrene) and averages for the varying alkylated three-ring PAHs (i.e. C1, C2, C3 and C4).

For all records collected from literature, the identified experimental degradation data were considered not relevant to determine environmentally realistic degradation half-lives, due to the inclusion of one or multiple aspects as described in Table 6.1 (see Table S1-3 for details). Nevertheless, the information of relative degradability of varying constituents within each study could be relevant for a qualitative assessment of the biodegradation kinetics.

The relative persistence of parent and alkylated three-ring PAHs in the aquatic compartment (for which most data are available) is shown in Figure 6.2. Only records with half-life values (Table S1) are included in this figure. In general, persistence seems to increase with the level of alkylation, with half-lives for parents < C1 < C2 < C3 < C4 three-ring PAHs. The observed trend is consistent across the studies and gives an indication of the expected relative behavior of alkylated and parent three-ring PAHs, even though the calculated degradation half-lives may not be relevant at environmentally realistic conditions. In addition, within the studies that report biodegradation in terms of depletion over time (in %) rather than half-lives, the same trend is observed (i.e. parents < C1 < C2 < C3 < C4 three-ring PAHs) (see Table S1). Slight variances between studies and overlap in persistence between adjacent blocks could potentially be explained by the presence of varying types of constituents in the different test materials. It has for instance been observed that varying methylated three-ring PAHs can have slightly different biodegradation kinetics (isomer variation) [188,189]. The exact composition of the tested material may thus slightly influence the observed degradation.

A similar trend as depicted in Figure 6.2 is observed when restricting to those studies where the influence of physicochemical properties on the observed relative biodegradation trend is minimal (Figure S1). Within this subset, studies were included that corrected or controlled for evaporation and photodegradation (i.e. closed (dark) test systems or corrected with abiotic controls). In addition, within these studies all constituents could be considered dissolved to a relative similar extent (i.e. single substance exposures or addition of a dispersant).

Furthermore, the sediment and soil biodegradation studies show a similar trend as observed in the water degradation studies (i.e. parents < C1 < C2 < C3 < C4 three-ring PAHs; Table S2-3). On the contrary, an opposite trend is observed for photodegradation, indicating faster photodegradation with an increase in alkylation for three-ring PAHs (Table S4). The trends as observed in the biodegradation studies are not considered to be influenced by photodegradation processes, as many studies corrected half-lives for abiotic degradation or were conducted in the dark (Table S1 and Figure S1).

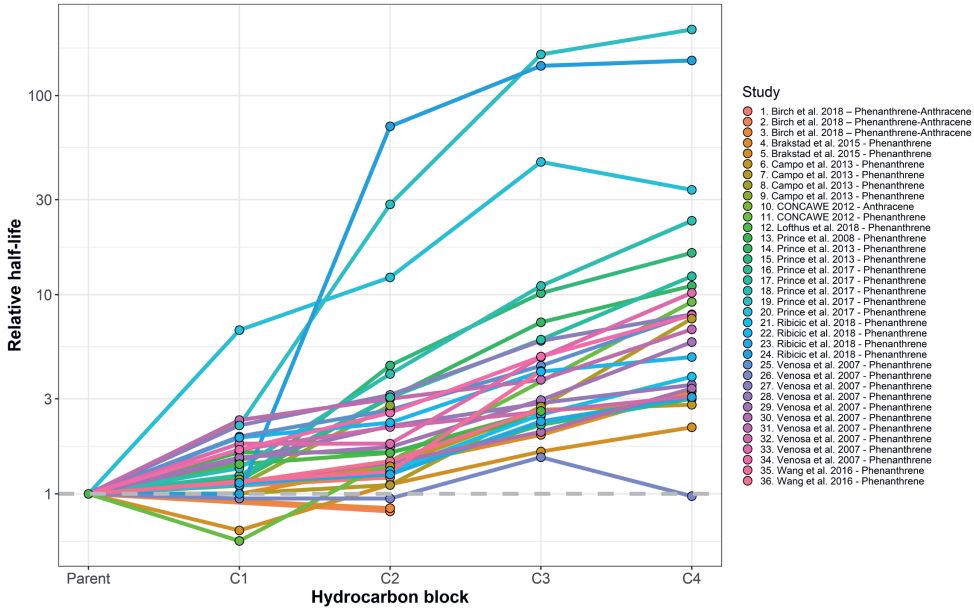


Figure 6.2. Trends in relative water biodegradation half-lives across parent and alkylated three-ring PAHs (see Table S1 and S5 for more details on test designs). Half-life values were normalized to the observed half-life for the parent substances and 'larger than values' (i.e. '>') were not included (see Table S5). Parent represents anthracene and/or phenanthrene, C1 are alkylated three-ring PAHs with one extra carbon atom, C2 contain two extra carbon atoms, C3 contain three extra carbon atoms and C4 contain four extra carbon atoms.

QSAR data

Besides experimental data, environmental biodegradation was predicted with Biowin3 and BioHCwin for all constituents belonging to the group of three-ring PAHs.

According to Biowin3 predictions, the parent three-ring PAHs are much more persistent than the alkylated three-ring PAHs (Figure 6.3A). However, when alkylated, a slight increase in persistence is observed with an increase in carbon number and/or chains. A higher half-life is particularly predicted for constituents with an increased number of alkyl-chains. Furthermore, it can be observed that there is a substantial spread in the half-life values across the total group of three-ring PAHs, of which several carbon number blocks contain non-persistent and persistent constituents.

In contrast to Biowin3, the BioHCwin model predicts different half-lives between anthracene and phenanthrene and their derivatives (Figure 6.3B). In addition, this model predicts that the alkylated three-ring PAHs are more persistent than the parent three-ring PAHs. Similar to Biowin3, the BioHCwin model also predicts a higher half-life with an increase in the number of chains for the alkylated three-ring PAHs. Overall, BioHCwin predicts a wide variation in

half-life values within carbon number groups, in which all groups contain non-persistent and very persistent constituents.

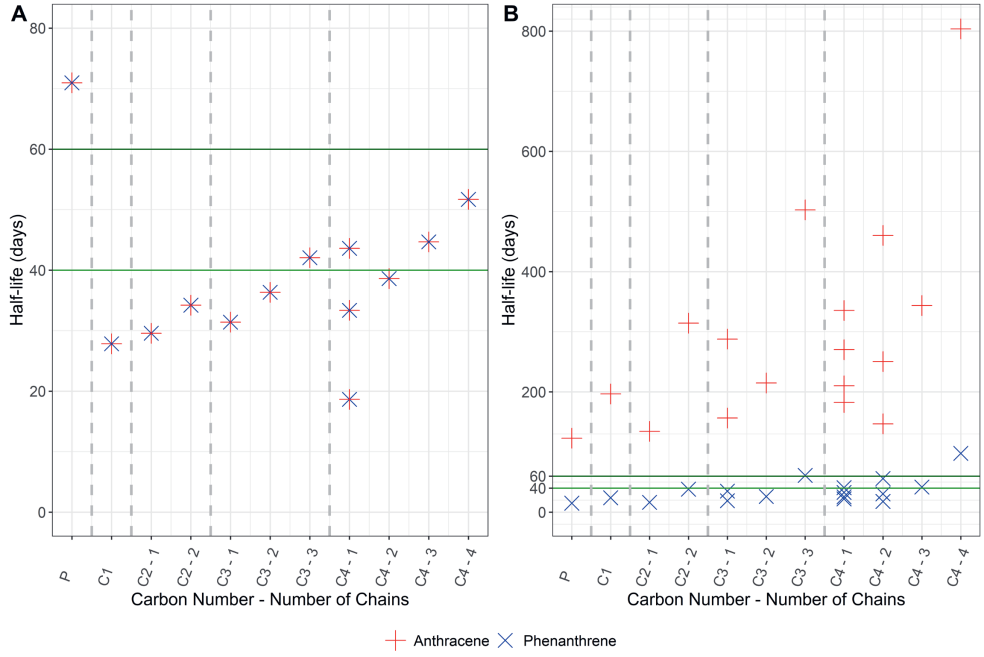


Figure 6.3. Predictions of the biodegradation half-lives for the group of three-ring PAHs. A) Predictions by Biowin3. B) Predictions by BioHCwin. P are the parent substances anthracene and/or phenanthrene, C1 are alkylated three-ring PAHs with one extra carbon atom, C2 contain two extra carbon atoms, C3 contain three extra carbon atoms and C4 contain four extra carbon atoms. In addition, the constituents within these blocks are further categorized based on the number of alkyl chains.

Persistence discussion

Based on available experimental data, a clear pattern in biodegradation was observed for parent and alkylated three-ring PAHs in water, sediment, and soil compartments, in which parent three-ring PAHs seem to be less persistent than alkylated forms (parent < C1 < C2 < C3 < C4). Less weight is given to the results of the QSAR analyses, as two fundamental contradictions between the models were observed that also deviate from the experimental data. Therefore, the QSAR predictions should be interpreted with caution (see discussion below).

The pattern as observed in the experimental data might be caused by interference of the alkyl substituents with oxidation enzymes via steric hindrance [190]. Similar conclusions are obtained in other studies. For instance, Leblond et al. [191] concluded that an increase in the number of methyl substituents and an increase in the size of a substituent, result in

a decreased primary biodegradation rate based on data on naphthalene (including varying methyl, dimethyl and ethyl derivatives). In addition, the same pattern was observed for phenanthrene in field studies by Douglas et al. [192] and Prince et al. [193]. Furthermore, the observed pattern is in accordance with the expected overall pattern of biodegradation of crude oil components as described by Prince and Walters [190].

In addition, there might be other factors that also contribute to the biodegradability of parent and alkylated three-ring PAHs that were not observed in the current evaluation. For instance, Wammer and Peters [194] concluded that naphthalene substances with a substituent in an α-position have lower biodegradation rates compared to naphthalenes without substituents or with a substituent in the β-position (see Figure 6.4A). Within the study of Wammer and Peters [194] no strong correlation was observed with the presence and length of alkyl substituents for naphthalenes.

Furthermore, the biodegradation might also be influenced by the presence/absence of specific sites or regions, like bay-regions and k-regions (see Figure 6.4B). It has been suggested that these regions can be used by (bacterial) enzymes to break down PAHs and thus potentially influence the biodegradation rate. As can be observed from Figure 6.4B, such regions can be formed by the PAH backbone (also called angular PAHs) or by alkyl substituents on specific locations [195,196]. Currently, these regions have not (extensively) been linked to biodegradation differences and are specifically related to carcinogenic effects of PAHs due to reactivity of the oxidized metabolites (i.e. epoxides) formed in a first step. Specifically, the high molecular weight PAHs with four- and five-rings can be transformed to 'bay-region' diol epoxides, which are highly reactive [197].

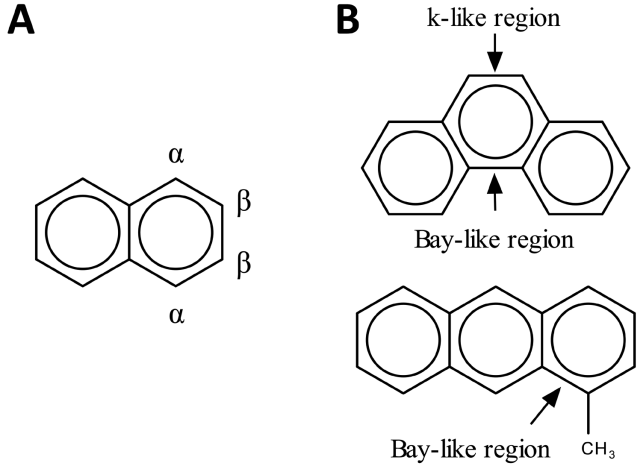


Figure 6.4. A) Characterization of substituent positions on the rings of naphthalene according to Wammer and Peters [194]. B) Bay-like regions and k-like regions in three-ring PAHs.

Within the QSAR analyses, two fundamental differences between the models were observed that also deviate from the experimental data. First, the models differently weigh the contribution of alkyl substituents to persistence. In comparison to parent three-ring PAHs, alkylation increases persistence according to BioHCwin (which is in line with experimental data) and decreases persistence according to Biowin3. These differences are not related to differences in the predicted endpoint (i.e. Biowin3 predicts ultimate biodegradation and BioHCwin predicts primary biodegradation), as Biowin4 (which predicts primary degradation) predicts a similar trend to Biowin3 (data not shown). Secondly, both models differently weigh the contribution of the parent backbone to persistence. Biowin3 makes no distinction between the parent backbones, whereas BioHCwin predicts that anthracene is very persistent and phenanthrene is not persistent. This difference is not specifically apparent from experimental data. Based on this evaluation it can be concluded that predictions of the two QSAR models (Biowin3 and BioHCwin) for the group of three-ring PAHs (and probably other alkylated PAHs) do not specifically match the experimental data, and therefore should be interpreted with caution. Apparently, the models are based on too few experimental data points to generate reliable predictions for the group of three-ring PAHs.

Besides the identified trends (as described above), there is also monitoring data available that indicates the presence of alkylated three-ring PAHs in the environment, both in populated and more remote areas. For instance, alkylated three-ring PAHs have been detected in water [198–200], sediment [201,202] and soil compartments [203,204], in which the measured concentrations of alkylated three-ring PAHs are generally comparable to the concentrations of parent three-ring PAHs. Although the sources of PAHs may vary (e.g. petrogenic and/or pyrogenic), these monitoring studies indicate that alkylated three-ring PAHs are present across the world. It should be noted, in this respect, that monitoring information on alkylated three-ring PAHs is not as extensive as monitoring data on parent three-ring PAHs, as most studies only focus on parent PAHs. To monitor the wide variety of alkylated PAHs, generally more specialized analytical techniques are required.

Persistence conclusion

In conclusion, no environmentally relevant half-lives were identified for the alkylated three-ring PAHs following the included validity criteria. However, considering the weight of evidence, independent of the test-setup and shortcomings, a clear trend is observed in experimental data across parent and alkylated three-ring PAHs, indicating an increase in persistence with the level of alkylation (i.e. parent < C1 < C2 < C3 < C4). This trend was consistent in the water, sediment and soil compartments. Therefore, it can be concluded that C1 to C4 anthracene and phenanthrene are more persistent than the parent three-ring PAHs.

6.3.2 Bioaccumulation

Experimental data

In total, seven fish and three invertebrate bioconcentration studies were identified that report BCFs for alkylated three-ring PAHs (Table S6-7). These studies mainly report BCFs for C1 and C2 alkylated three-ring PAHs, and only one study reported BCFs for C3 and C4 constituents. Furthermore, within the reported bioconcentration tests, data is generally only available for a few constituents specifically. This is in contrast to persistence tests, where often groups of constituents are analyzed and followed over time (i.e. parents, C1, C2, C3 and/or C4).

Of all available bioconcentration data on alkylated three-ring PAHs, only one relevant BCF for one specific constituent is identified (i.e. BCF of 3896 for 9-methylanthracene, C1 three-ring PAH [205]). All other data were considered not-relevant to determine the bioconcentration potential, mainly due to simultaneous exposure to four and five-ring PAHs (see section 6.2.2). We therefore restricted our analyses to bioconcentration trends across parents and alkylated three-ring PAHs that were tested within the same study.

The trends that can be observed in the individual fish bioconcentration studies, testing a parent and alkylated three-ring PAH in a mixture test design, are shown in Figure 6.5A. These data show a quite consistent BCF trend for phenanthrene, in which the BCF tends to decrease with an increase in alkylation. For anthracene on the other hand, a slight increase in BCF is observed with an increase in alkylation. However, as only one study analyzed the bioconcentration trend for anthracenes, less weight can be given to this observation, especially because for phenanthrene nearly the same trend was observed in this study.

The bioconcentration trends as observed within the three invertebrate studies are visualized in Figure 6.5B. These studies show a consistent pattern that deviates from the trend as observed in the fish bioconcentration studies with phenanthrene. Available invertebrate data indicate an increase in BCF with increases in alkylation/hydrophobicity, both for anthracene and phenanthrene.

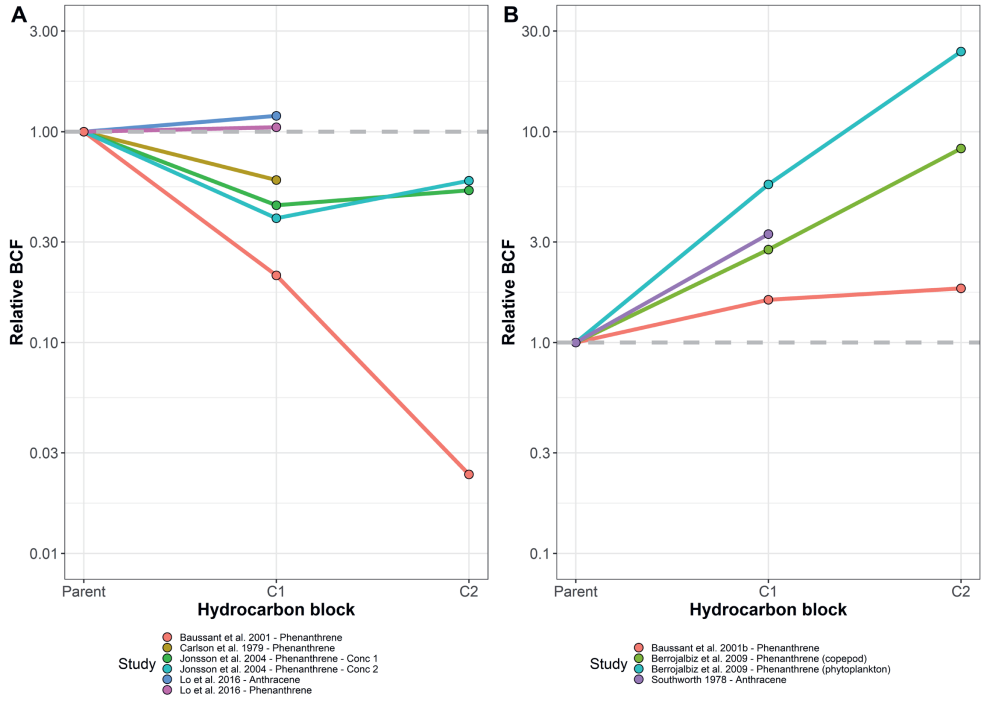


Figure 6.5. Trends in relative bioconcentration factors (BCF) across parents and alkylated three-ring PAHs. Where possible, BCFs are based on kinetic data (see Table S6-7). A) Fish data. B) Invertebrate data. Parent represents anthracene and/or phenanthrene, C1 are alkylated three-ring PAHs with one extra carbon atom and C2 contain two extra carbon atoms.

QSAR data

Besides experimental data, the bioaccumulation potential of the group of alkylated three-ring PAHs was predicted with BCF estimates of the EU-BCFmax and BCFBAF models (i.e. BCFBAF regression and BCFBAF Arnot-Gobas model).

The EU-BCFmax and the BCFBAF regression model predict a similar bioconcentration pattern across parent and alkylated three-ring PAHs (Figure 6.6A-B). Both models predict an increase in bioconcentration potential until a certain optimum (mainly related to log K_{ow}), after which a decline in BCF values is predicted. However, the absolute predicted BCF values differ between both models. For instance, the EU-BCFmax model predicts that the parent three-ring PAHs already meet the B-criteria, whereas the BCFBAF regression model suggests that the B-criteria is exceeded at a higher carbon number. In addition, the BCFBAF regression model predicts a different bioconcentration potential for anthracene and phenanthrene derivatives. The predictions by the BCFBAF Arnot-Gobas model, on the other hand, indicate a different pattern in which no optimum can be observed (Figure 6.6C). This model suggests that none of the hydrocarbon blocks meet the B-criteria.

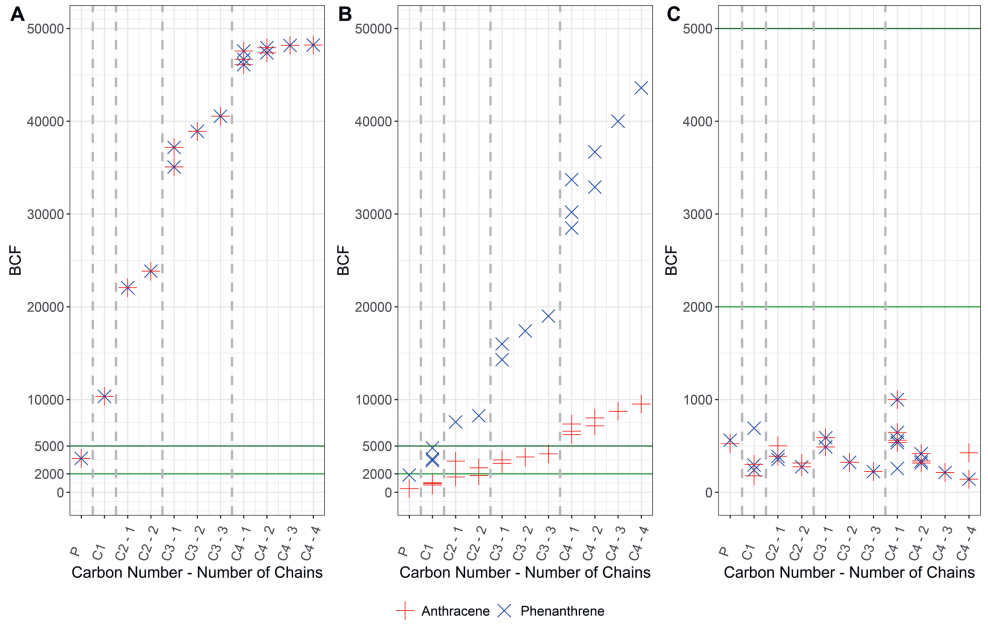


Figure 6.6. Predictions of bioconcentration factors (BCF) for the group of three-ring PAHs. A) Predictions by the EU BCFmax model. B) Predictions by the BCFBAF regression model. C) Predictions by the BCFBAF Arnot-Gobas model, upper trophic level, normalized to 5% lipid content. P are the parent substances anthracene and/or phenanthrene, C1 are alkylated three-ring PAHs with one extra carbon atom, C2 contain two extra carbon atoms, C3 contain three extra carbon atoms and C4 contain four extra carbon atoms. In addition, the constituents within these blocks are further categorized based on the number of alkyl chains.

Bioaccumulation discussion

Based on available fish bioconcentration data, a trend for phenanthrene was observed that indicates lower BCF values with increased alkylation. The steepness of the decrease, however, cannot be assessed based on the limited amount of data. According to these results, the alkylated three-ring PAHs are potentially more readily metabolized than parent three-ring PAHs. Remarkably, however, the observed bioconcentration trend for alkylated phenanthrenes in fish deviates from the observed biodegradation trends, despite commonality of reactions in microbial and fish transformations [206]. Potentially, this is related to (functional) differences between microbial and fish enzymes, but could also be related to other fate processes that take place in the degradation experiments (e.g. sorption). Moreover, a decrease in bioavailability in the fish bioconcentration experiments seems not a plausible explanation for the observed difference, as an opposite bioconcentration trend is observed for invertebrates.

For anthracene on the other hand, an opposite trend was observed in fish bioconcentration data that indicates a slight increase in BCF with an increase in alkylation. However, as only one study analyzed the trend for anthracene, less weight can be given to this observation,

especially because for phenanthrene nearly the same trend was observed in this study. Potentially, differences in fish bioconcentration potential between phenanthrene and anthracene derivatives could be related to factors as discussed in Figure 6.4, including the position of the alkyl substituent(s) and the presence of bay- and k-regions.

In general, the fish bioconcentration trends correspond to the trends as predicted by the BCFBAF Arnot-Gobas model. However, the reliability of the absolute values as estimated by this model could be questioned, as for instance the parents are not predicted to be bioaccumulative (see discussion below).

The invertebrate bioconcentration data indicate a clear trend across all available studies. These studies show increases in BCF values with increases in alkylation. The difference between fish and invertebrate bioconcentration trends could be explained by a general lower metabolic capacity in invertebrates [19]. Such differences have been observed for PAHs [207], and also for several other chemicals, like polychlorobiphenyls [208]. In addition, the trend as observed in experimental invertebrate bioconcentration data corresponds to the predictions of the EU-BCFmax model, which also does not take metabolism into account (i.e. increases in bioconcentration with increases in log K_{ow}).

The predictions by the EU-BCFmax model (and BCFBAF regression model) are very different from the predictions by the BCFBAF Arnot-Gobas model. One of the main factors contributing to the observed difference is the metabolism correction. When applying the BCFBAF metabolism correction to the EU BCFmax model, similar results are obtained as in the BCFBAF Arnot-Gobas model, and vice versa (Figure S2). Although the application of metabolism correction could be considered relevant, particularly for fish, the reliability of the first-order metabolism rate constant (kM) predictions for three-ring PAHs by the BCFBAF model could be questioned. The data that are used to develop the kM BCFBAF model were not primarily set up to determine kM values and are derived from studies with varying study set-ups and species. Metabolism rate constants are derived from these studies based on BCF and/or total elimination rate constants following several assumptions. The related three-ring PAH data that were used in the development of the kM BCFBAF model consider average kM values that are mainly derived from studies in which fish were exposed to multiple substances simultaneously (including four- and five-ring PAHs). These mixtures significantly influence the metabolism within the fish, resulting in faster transformation of three-ring PAHs compared to single substance exposures [133,176]. When only kM data were used from studies that applied single substance exposures, the derived half-lives are a factor 3 to 6 higher than current applied values (Table S8). Consequently, the kM BCFBAF corrections may overpredict the metabolism, resulting in lower predicted BCF values.

In addition to the bioconcentration trends as observed in the laboratory experiments and the predictive models, several studies have measured alkylated three-ring PAHs in organisms/food chains in the field, including in invertebrates, fish and seabirds [199,200,209,210]. These field studies, however, do not provide a clear bioaccumulation/biomagnification pattern across parent and alkylated three-ring PAHs, as no (consistent) trends can be identified. For instance, the data of Takeuchi et al. [200] indicates an increase in bioaccumulation with an increase in alkylation, whereas no clear trend can be observed from the data by Khairy et al. [199] and Nfon et al. [209] (see Table S9). In addition, the field data cannot directly be compared to the B-criteria, as the criteria is based on bioconcentration potential [19]. Nevertheless, this data indicates that alkylated three-ring PAHs are identified in organisms, and can be found at comparable concentrations as parent three-ring PAHs.

Bioaccumulation conclusion

In conclusion, alkylated three-ring PAHs can be considered more bioaccumulative than the parent three-ring PAHs in invertebrates, which seem to follow the trend of increased bioconcentration with increases in log K_{ow}. Therefore, the observed trend for parents, C1 and C2 three-ring PAHs could be extrapolated to C3 and C4 three-ring PAHs based on QSAR predictions. In contrast to invertebrates, the BCF tends to decrease in fish with an increase in alkylation for phenanthrenes. For alkylated anthracenes no conclusions can be drawn for bioconcentration in fish, mainly due to a lack of data.

6.3.3 Toxicity

Experimental data

In total, 20 toxicity studies with aquatic organisms were identified that specifically report the toxicity of alkylated three-ring PAHs (Table S10). Investigated endpoints include mortality, growth, morphology (including symptoms related to blue sac disease in fish) and hatching. Of the identified studies, ten were considered reliable. These studies are reported in Tables 6.2-6.3 for alkylated phenanthrenes and anthracenes, respectively. Within these tables, only studies on parent three-ring PAHs toxicity are included when they also analyzed toxicity of alkylated three-ring PAHs. For the parent three-ring PAHs there is much more toxicity data available, for which some critical studies are reported in Table S11.

Based on available data, it can be observed that alkylated anthracenes and alkylated phenanthrenes exert a phototoxic effect at relevant UV-exposure conditions (i.e. applied UV-intensities were below the intensities of full-strength sunlight). This effect seems to be stronger for alkylated anthracenes compared to alkylated phenanthrenes. Furthermore, it can be observed that some species (like the mysid shrimp, *Americamysis bahia*) seem to be more sensitive to (alkylated) three-ring PAHs compared to others (Tables 6.2-6.3). Differences in

sensitivity between organisms to hydrocarbon and PAH exposure have been shown in SSDs (Supplemental Material S1, [177,185,186]), and the order of species in these SSDs matches the species sensitivity as observed for three-ring PAHs in this study rather well (Table S12).

In general, only few toxicity data on alkylated three-ring PAHs are available, and only a low number of different constituents have been tested. For the lower carbon number blocks (i.e. parents, C1 and C2), one or more constituents show to be toxic to a sensitive aquatic organism below an exposure concentration of 10 µg/L (i.e. T-criteria [19]). For the higher carbon number blocks very few data are available, including only one test with a C3-constituent and four tests with the same C4-constituent (i.e. retene). For all these studies, none of the more sensitive species were used (Table S12) and it is unclear whether UV radiation has been applied within the set-up of these studies.

Table 6.2. Reliable toxicity data on alkylated phenanthrenes. Toxicity data are reported as EC10 or NOEC values for the most sensitive endpoint. When EC10 and/or NOEC values were not reported and could not be derived, reported EC50 values are presented. Bold values are toxicity values below 10 µg/L. More details on the specific tests are included in Table S10 and detailed study references are provided in the supplemental material. 1) The NOEC/EC10 values for phenanthrene and methyl-phenanthrene from Finch 2017 are around the T-threshold upon acute (48h) exposure. For phenanthrene an exposure concentration-response curve fit could be conducted to derive an EC10. 2) Derived with GraphPad Prism from original data. 3) Sediment toxicity test. 4) Only one concentration was tested, at which more than 50% effect was observed.

UV conditions	Carbon number	Substance	Species	Tox (µg/L)		Ref	Remarks
With UV	Parents	Phenanthrene	*Americamysis bahia*	12.5	EC10	Finch 2017	1, 2
		Phenanthrene	*Rhepoxynius abronius*	117.9	EC50	Boese 1998	3
	C1	3-methylphenanthrene	*Americamysis bahia*	**8.65**	NOEC	Finch 2017	1
		2-methylphenanthrene	*Rhepoxynius abronius*	36.5	EC50	Boese 1998	3
	C2	3,6-dimethylphenanthrene	*Americamysis bahia*	**3.24**	NOEC	Finch 2017	
		3,6-dimethylphenanthrene	*Rhepoxynius abronius*	>60.6	EC50	Boese 1998	3
Without UV	Parents	Phenanthrene	*Americamysis bahia*	<13.3	NOEC	Finch 2017	
		Phenanthrene	*Oryzias latipes*	>200	NOEC	Rhodes 2005	
		Phenanthrene	*Rhepoxynius abronius*	123.5	EC50	Boese 1998	3
	C1	3-methylphenanthrene	*Americamysis bahia*	21.2	EC10	Finch 2017	2
		2-methylphenanthrene	*Rhepoxynius abronius*	36.5	EC50	Boese 1998	3
	C2	3,6-dimethylphenanthrene	*Americamysis bahia*	>16.3	NOEC	Finch 2017	
		3,6-dimethylphenanthrene	*Oryzias latipes*	>200	NOEC	Rhodes 2005	
		3,6-dimethylphenanthrene	*Rhepoxynius abronius*	>60.6	EC50	Boese 1998	3
Not-specified	Parents	Phenanthrene	*Oryzias latipes*	42.8	EC10	Turcotte 2011	2
		Phenanthrene	*Oryzias melastigma*	84.5	EC10	Mu 2014	2
	C1	1-methylphenanthrene	*Oryzias latipes*	**1.8**	EC10	Fallahtafti 2012	2
		1-methylphenanthrene	*Oryzias latipes*	95.3	EC10	Turcotte 2011	2
		3-methylphenanthrene	*Raph. subcapitata*	280	EC50	Kang 2016	
	C2	1,7-Dimethylphenanthrene	*Oryzias latipes*	18.7	EC10	Turcotte 2011	2
		2,7-Dimethylphenanthrene	*Oryzias latipes*	16.1	EC10	Turcotte 2011	2
		2-Ethylphenanthrene	*Oryzias latipes*	32	EC10	Turcotte 2011	2
		3,6-dimethylphenanthrene	*Raph. subcapitata*	>37	EC50	Kang 2016	
	C3	7-Ethyl-1-methylphenanthrene	*Oryzias latipes*	>36	EC10	Turcotte 2011	
	C4	7-Isopropyl-1-methylphenanthrene	*Oryzias melastigma*	45.2	EC10	Mu 2014	2
		7-Isopropyl-1-methylphenanthrene	*Oncorhynchus mykiss*	**<9**	EC50	Brinkworth 2003	4
		7-Isopropyl-1-methylphenanthrene	*Oryzias latipes*	47.5	EC10	Turcotte 2011	2
		7-Isopropyl-1-methylphenanthrene	*Oryzias latipes*	50.7	EC50	Lin 2015	

Table 6.3. Reliable toxicity data on alkylated anthracenes. Toxicity data are reported as EC10 or NOEC values for the most sensitive endpoint. When EC10 and/or NOEC values were not reported and could not be derived, reported EC50 values are presented. Bold values are toxicity values below 10 µg/L. More details on the specific tests are included in Table S10 and detailed study references are provided in the supplemental material. 1) Derived with GraphPad Prism from original data. 2) Sediment toxicity test.

UV conditions	Carbon number	Substance	Species	Tox (µg/L)		Ref	Remarks
With UV	Parents	Anthracene	*Americamysis bahia*	**0.44**	EC10	Finch 2017	1
	C1	2-methylanthracene	*Americamysis bahia*	**0.27**	NOEC	Finch 2017	
		2-methylanthracene	*Rhepoxynius abronius*	**4.5**	EC50	Boese 1998	2
		9-methylanthracene	*Rhepoxynius abronius*	10.9	EC50	Boese 1998	2
	C2	2,3-dimethylanthracene	*Americamysis bahia*	**0.055**	NOEC	Finch 2017	
		9,10-dimethylanthracene	*Rhepoxynius abronius*	>41.9	EC50	Boese 1998	2
Without UV	Parents	Anthracene	*Americamysis bahia*	>44.2	NOEC	Finch 2017	
		Anthracene	*Oryzias latipes*	>40	EC10	Turcotte 2008	
	C1	2-methylanthracene	*Americamysis bahia*	>6.37	NOEC	Finch 2017	
		1-methylanthracene	*Oryzias latipes*	25	EC10	Turcotte 2008	1
		9-methylanthracene	*Oryzias latipes*	130	EC10	Turcotte 2008	1
		2-methylanthracene	*Rhepoxynius abronius*	>36.5	EC50	Boese 1998	2
		9-methylanthracene	*Rhepoxynius abronius*	105.5	EC50	Boese 1998	2
	C2	2,3-dimethylanthracene	*Americamysis bahia*	>0.37	NOEC	Finch 2017	
		9,10-dimethylanthracene	*Oryzias latipes*	<12	EC10	Turcotte 2008	1
		9,10-dimethylanthracene	*Rhepoxynius abronius*	>41.9	EC50	Boese 1998	2
Not specified	C1	9-methylanthracene	*Raph. subcapitata*	320	EC50	Kang 2016	
	C2	9,10-dimethylanthracene	*Raph. subcapitata*	>7.9	EC50	Kang 2016	

QSAR data

Ecotoxicity was also predicted with the TLM and ILRM for all constituents that belong to the group of alkylated three-ring PAHs. The predictions of both models are shown in Figure 6.7, and express the fraction of species that is potentially affected at an exposure concentration of 10 µg/L or at the maximum water solubility level. These predictions indicate a clear and consistent trend across both models, with higher toxicity being predicted for constituents containing more carbon atoms (i.e. parents < C1 < C2 < C3 < C4 three-ring PAHs). Only slight differences between the TLM and ILRM are observed, in which the ILRM predicts a higher fraction of affected species for the lower carbon number blocks (P, C1 and C2) and the TLM for the higher carbon number blocks (C3 and C4).

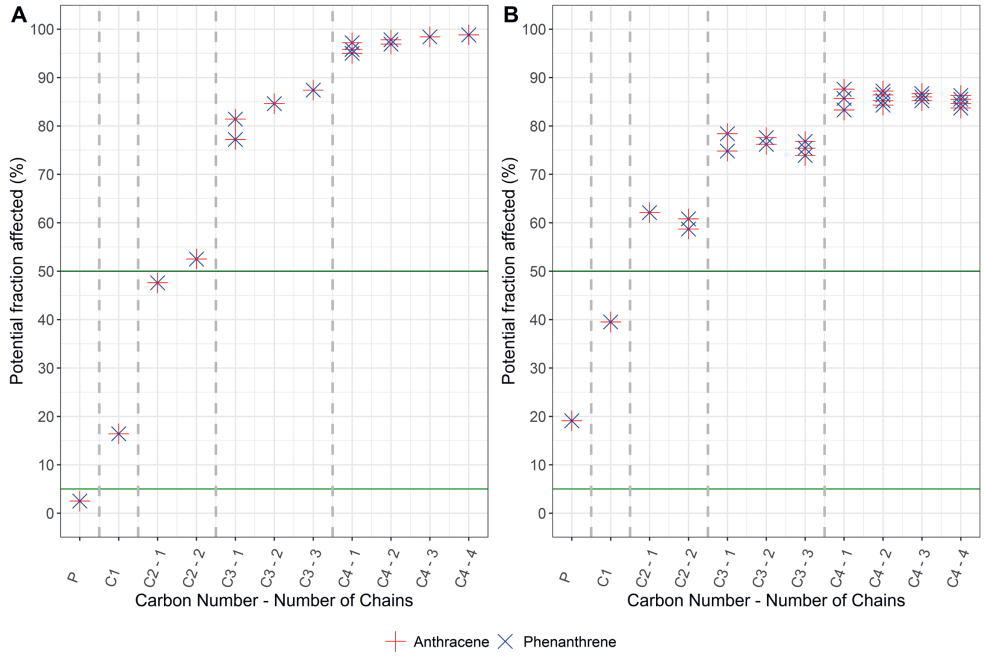


Figure 6.7. Predictions of ecotoxicity in terms of fraction of affected species at an exposure concentration of 10 µg/L or at maximum water solubility for the group of three-ring PAHs. A) Predictions by the Target Lipid Model. B) Predictions by the Internal Lipid Residues Model. P are the parent substances anthracene and/or phenanthrene, C1 are alkylated three-ring PAHs with one extra carbon atom, C2 contain two extra carbon atoms, C3 contain three extra carbon atoms and C4 contain four extra carbon atoms. In addition, the constituents within these blocks are further categorized based on the number of alkyl chains.

Toxicity discussion

Several reliable experimental toxicity studies were available, particularly for the lower hydrocarbon blocks (i.e. parents, C1 and C2). Although the identified studies mainly consider short-term or sub-chronic exposures, (photo)toxic effects are observed below the T-threshold of 10 μg/L for these lower carbon number constituents.

According to the available data, alkylated anthracenes seem slightly more toxic than alkylated phenanthrenes. In particular, alkylated anthracenes seem to have a higher phototoxic potential, which might be related to the molecule's HOMO-LUMO gap (i.e. the energy difference between the highest occupied and the lowest unoccupied molecular orbital). De Lima Ribeiro and Ferreira [211] predicted that substances with a HOMO-LUMO gap between 6.5-7.9 eV have a high phototoxic potential, whereas substances with a HOMO-LUMO gap beyond this range are predicted to have a much lower or no phototoxic potential. The HOMO-LUMO gap of anthracene is within these boundaries (7.279 eV), whereas the HOMO-LUMO gap of phenanthrene is above the upper value (8.209 eV) [211]. It has been shown that methylation generally reduces the HOMO-LUMO gap of substances. Methyl- and dimethyl-phenanthrene, for instance, have a HOMO-LUMO gap of 8.13 eV and 8.05 eV, respectively [212]. This may explain the observed phototoxic effects of alkylated phenanthrenes in comparison to the parent three-ring PAH. Furthermore, the difference in the HOMO-LUMO gap between alkylated anthracenes and phenanthrenes may explain the different phototoxic potencies as observed for these constituents.

The predictive models, which are mainly based on non-phototoxic effects, estimate a consistent increase in toxicity with an increase in alkylation. This trend is not particularly confirmed by the experimental data without UV-exposure, in which in general no toxicity below 10 μg/L is observed. However, it should be noted that available experimental data consider acute or sub-chronic exposures (which includes many larger than values: '>'), and no information on chronic effects without UV-exposure is available. Therefore, in the absence of data, it cannot be concluded that the alkylated three-ring PAHs will not exhibit any long-term effects below 10 μg/L without UV-exposure. The available experimental phototoxicity data (with mysid shrimp, *Americamysis bahia*) show a clear trend that is consistent and fits the trend as predicted by the models, indicating an increase in toxicity with hydrophobicity (see Figure S3). Accordingly, a similar trend is expected for long-term toxicity, for which the models were developed. Therefore, higher (photo)toxic effects are expected for C3 and C4 constituents, compared to the parent, C1 and C2 constituents.

Besides aquatic toxicity, alkylated three-ring PAHs may also exert toxic effect to mammalian species. However, in the current evaluation of alkylated three-ring PAHs we did not consider toxicity to mammalian species, including humans. It should be noted, in this respect, that for

the parent three-ring PAHs, anthracene and phenanthrene, there are currently no notified classifications to mammals/humans that would give rise to a T-identification within the PBT-assessment [213].

Toxicity conclusion

In conclusion, experimental data on C1 and C2 alkylated three-ring PAHs indicate aquatic toxicity below the T-threshold of 10 μg/L, and these alkylated PAHs can therefore be considered as T. For the constituents belonging to the higher hydrocarbon blocks (i.e. C3-C4), far less experimental data are available and none of the more sensitive species have been tested (Table S12). Therefore, no conclusions could be drawn based on experimental data for the C3-C4 constituents. Nevertheless, based on the predictive models, a clear trend of increasing toxicity with an increase in alkylation/hydrophobicity is predicted. Accordingly, C3-C4 constituents are expected to be more toxic than the lower carbon number constituents.

6.3.4 PBT conclusion and outlook to other UVCBs

Based on the P-, B- and T-assessment, it can be concluded that the alkylated three-ring PAHs are more, or at least equally persistent, bioaccumulative in invertebrates and toxic to aquatic organisms when compared to the parent three-ring PAHs, anthracene and phenanthrene (Table 6.4). As anthracene and phenanthrene are currently both considered as PBT and vPvB, respectively [163], the alkylated three-ring PAHs could also be considered to have PBT/vPvB properties.

It should be noted, however, that the current PBT/vPvB-status of phenanthrene is under discussion [164,165]. If these discussions would result in a removal of the PBT/vPvB-status of phenanthrene, it will be more difficult to derive a conclusion for the whole group of alkylated three-ring PAHs based on current available data. Particularly, as this would mean a diverging PBT/vPvB-starting point for the group of three-ring PAHs. It should be noted, in this respect, that a difference in PBT/vPvB-properties between the two parent three-ring PAHs would, in principle, not be in accordance with the block homogeneity assumptions of the HBM.

Application of the HBM in the PBT-assessment for other hydrocarbon blocks is likely more complicated and challenging compared to the three-ring PAHs, as for many hydrocarbon categories far less experimental data are available and/or parent structures have not been assessed (yet). For instance, difficulties arise when no (unambiguous) PBT-trends could be derived based on available experimental data. Potentially, a part of these experimental data gaps could be filled by extrapolating the results of the current evaluation of three-ring PAHs to other categories. For instance, several studies indicate comparable P-, B- and T-trends for two-ring PAHs [191,214–216] and four-ring PAHs [214,217], and may suggest a generic trend

across parent and alkylated PAHs. Nevertheless, a systematic evaluation of available data would be necessary to conclude on such trends for other categories. Furthermore, challenges with the application of the HBM arise when a relative PBT-trend could be derived, but no clear PBT/vPvB-starting point is established or the starting point is below or around the border of the PBT-criteria. In such cases, quantitative data would be necessary in order to conclude the PBT-assessment.

Several methodologies have been proposed that could be used to assess and generate data on UVCBs, including constituent-based approaches and whole substance-based approaches [19,159]. Ideally, for PBT-assessment purposes of these substances, additional data are generated via a constituent-based approach (like the HBM), as the physicochemical properties differ significantly between the varying constituents of petroleum UVCBs. Although all constituents within a petroleum UVCB could be emitted as a whole, the constituents will fractionate in the environment due to the varying physicochemical properties and persistence, resulting in different distribution, fate and exposure patterns [19,159]. Therefore, one or more (representative) constituents or groups/blocks of very similar constituents should be tested, and only in cases where all constituents are (structurally) very similar, a whole substance-based approach might be followed. Nevertheless, further scientific discussions and analyses are necessary to improve the understanding of which specific data would be sufficient to derive a PBT-conclusion on a block or UVCB when (additional) quantitative data are necessary.

Table 6.4. Conclusions of the PBT-assessment on the group of alkylated three-ring PAHs. Parent represents anthracene and/or phenanthrene, C1 are alkylated three-ring PAHs with one extra carbon atom, C2 contain two extra carbon atoms, C3 contain three extra carbon atoms and C4 contain four extra carbon atoms.

Carbon number	Persistence	Bioaccumulation	Toxicity [1]
Parent [2]	vP	B/vB	nT/T [3]
C1	More persistent than parent [4]	More bioaccumulative than parent [5]	T [6]
C2	More persistent than parent [4]	More bioaccumulative than parent [5]	T [6]
C3	More persistent than parent [4]	More bioaccumulative than parent [5]	More toxic than parent, C1 and C2 [7]
C4	More persistent than parent [4]	More bioaccumulative than parent [5]	More toxic than parent, C1 and C2 [7]

1) The toxicity assessment solely focused on aquatic toxicity and did not consider effects on human health. 2) Based on the SVHC-dossiers of anthracene, phenanthrene and coal-tar-pitch high temperature. It should be noted that the current PBT/vPvB-status of phenanthrene is under discussion see [164,165] and text. 3) Phenanthrene is currently not identified to be toxic, though data is available showing toxic effects below 10 µg/L (Table S11). 4) Based on a relative persistence trend to parent three-ring PAHs for water, sediment and soil compartments. 5) Based on a relative bioconcentration trend to parent three-ring PAHs for invertebrate data. 6) Based on experimental data indicating toxic effect below 10 µg/L. 7) Based on a relative aquatic toxicity trend to parent and alkylated three-ring PAHs according to modelled data.

6.4 Conclusion

Within this study, we applied the fraction profiling approach or so-called hydrocarbon block method (HBM) for a PBT-assessment of alkylated three-ring PAHs to investigate the suitability of the HBM for the assessment of UVCBs. Evaluation of available data revealed that the absolute degradation half-lives and BCF values from many studies are of insufficient relevance for PBT-assessment purposes. Nevertheless, by using trend analyses on a block of hydrocarbons with a known PBT/vPvB starting point, it was possible to derive a PBT-conclusion for 884 constituents in one assessment. This case study on the alkylated three-ring PAHs gives promising perspectives for other hydrocarbon blocks and possibly for other UVCBs. For these cases, further work is required to evaluate the suitability of the HBM when trend analyses are not possible and/or conclusive.

Acknowledgements

We highly appreciate and acknowledge the fruitful discussions with our colleagues Emiel Rorije, Fleur van Broekhuizen and Joop de Knecht. Furthermore, we would like to thank everyone who has provided valuable comments and we would like to thank all colleagues who have reviewed the manuscript internally. This work was funded by the Dutch Ministry of Infrastructure and Water Management.

Supplemental material

Supplementary data to this chapter can be found online at https://doi.org/10.1016/j.chemosphere.2021.130113.

7

General Discussion

We are surrounded by thousands of chemicals, which in their entirety are known as the chemical universe. On the one hand, chemicals provide many benefits to mankind, but on the other hand their application can result in unwanted emissions that harm human health and the environment. Ideally, all chemicals in the chemical universe are regulated to such an extent that safe and sustainable production and use are guaranteed. This particularly applies to chemicals that are able to exert serious and irreversible adverse effects, which are also known as substances of very high concern (SVHC). However, this safe and sustainable ambition is challenged by a general lack of (reliable) toxicity data for many chemicals which complicates the risk and hazard assessment, and is hampered by relative slow and inefficient evaluation and regulation processes. In general, this is related to time-consuming procedures, limited available evaluation capacity, and an inefficiency in the regulation of one substance at a time.

This thesis aimed to identify and provide opportunities to overcome these challenges by focusing on (aspects related to) chemical similarity. First, we developed similarity-based screening models that enhance the identification of chemicals of potential concern, and support the transition from substance-by-substance assessment towards group assessment approaches (chapters 2-4). In addition, we analyzed the potential benefits of using chemical similarity for data generation and evaluation purposes, by investigating how biological similarity and variability could influence and could be incorporated in risk and hazard assessment (chapters 5 and 6).

7.1 Early and effective signaling of concerns

It is essential to signal potential concerns as early as possible, preferably before widespread exposure occurs. Accordingly, we developed a screening methodology that predicts whether a substance is a potential SVHC based on chemical similarity to chemicals already identified as SVHC (chapters 2-4). Overall, during the development of the structural similarity models, a satisfactory predictive performance is observed (with balanced accuracies ≥ 0.75 based on the used dataset) (chapters 2 and 4). In addition, also a reasonable performance on the broader universe of chemicals is observed (chapter 3), indicating its capability to identify substances of potential concern.

7.1.1 Model performance

In an ideal world, a model provides a correct prediction for every single substance (i.e. balanced accuracy = 1). Accordingly, there seem to be opportunities to improve the similarity models. Potentially, the statistics of the developed similarity models might be improved by using other molecular descriptors or model algorithms, like i) 3D molecular descriptors and ii) machine learning methods.

i) 3D molecular descriptors define the shape of chemicals (e.g. in the form of volumes or surfaces) and could be used as an alternative to (or in combination with) the currently applied 2D descriptors [218]. Several of such shape-based 3D descriptors showed to be particularly useful when applied to computational drug discovery, as it is an important descriptor for determining binding affinity as well as several other properties, like solubility [219,220]. Potentially, the use of 3D descriptors could reduce the number of activity cliffs (i.e. two very similar chemicals which have an unexpectedly high difference in activity) [21,221].

ii) The use of (supervised) machine learning algorithms is becoming more and more common for developing predictive models (e.g. random forest and (deep) neural networks). Machine learning methods are especially useful when dealing with large and complex datasets, as these algorithms can process a wide volume and variety of data, even when the data are highly imbalanced (i.e. few active/toxic chemicals versus many inactive/non-toxic chemicals) [21,25].

Despite the bright prospects of these advanced models and descriptors, their application to the currently used dataset will unlikely result in significantly better models for the similarity tool:

- First, a comparable number of falsely classified substances are expected to be observed. This is mainly due to the fact that a comparable number of false negatives will be obtained as several SVHC substances represent single-points-of-knowledge, which are not comparable to any of the other SVHC structures (chapter 2). Although the number of false positives could potentially be reduced, the impact on the overall statistics is considered limited as the performance was mainly influenced by false negatives (chapter 4), and there is always a chance of false positives due to activity cliffs. The use of (currently available) 3D descriptors (specifically within computational toxicology) does not necessarily outperform the predictive performance of 2D descriptors [222], particularly due to challenges and uncertainties related to chemical conformations and alignments [32,223]. Moreover, the application of machine learning methods to the currently developed datasets does not necessarily provide models with a significantly improved predictive performance, as also verified within several student projects using classical machine learning models [224–226]. This is mainly related to the fact that the developed similarity models are based on a relative small dataset, as only a limited number of substances are identified as SVHC or can be considered as non-SVHCs.

- Second, and most importantly, the application of 3D descriptors and machine learning algorithms would affect the interpretability of the models. The similarity tool is developed for screening and prioritization purposes, and aims to provide an automated, fast and reproducible alternative to expert opinions. Hence, the model's predictions should not be interpreted as conclusive outcomes but should be subject to further evaluation. The strength of the similarity tool derives from its simplicity of use (i.e. online freely accessible program with user-friendly interface) and interpretation, and is considered fit-for-purpose for screening SVHCs. The models provide a systematic and transparent way to identify relations that would not manually be identified, and can visually be verified as they are based on straightforward unambiguous 2D molecular descriptors (i.e. fragments). This is in contrast to the use of 3D descriptors, which are more difficult to (visually) interpret, also considering the uncertainties in chemical conformations and alignments [32,223]. In addition, application of machine learning methods would potentially affect the interpretability of the models, as many machine learning models function as a 'black box' [227].

Accordingly, the use of 3D descriptors and machine learning algorithms within the currently developed similarity models are not likely to improve the performance significantly, as the models have been developed with a specific goal in mind. Nevertheless, considering the advancements with 3D descriptors and machine learning algorithms, it is likely that these aspects will play an increasingly important role for the signaling of emerging concerns in future [24,32,223,228], which is further discussed in section 7.3.

7.1.2 The online ZZS similarity tool

The developed similarity models are incorporated in an online publicly available instrument, named the ZZS similarity tool: https://rvszoeksysteem.rivm.nl/ZzsSimilarityTool (see Figure 7.1) (ZZS = 'Zeer Zorgwekkende Stoffen' [in Dutch], which is literally translated as substances of very high concern; see chapters 2-4 for a detailed description of the incorporated similarity models). This tool is already being applied within several frameworks and processes, and has already been viewed by more than 2500 unique visitors in the first 18 months after release. At the RIVM, the tool is particularly used to support and advise licensing authorities in the assessment of potential risks of substances [229], and has already been applied for almost hundred recorded cases. In some of these cases, the predictions by the similarity tool were decisive for the final advice (e.g. substance has an equivalent concern as potential ZZS) [230]. In addition, the tool proves to be of added value in discovering that a substance belongs to a ZZS category, and therefore must be treated as such. Furthermore, the tool has been applied to monitoring data of water bodies in the Netherlands to identify chemicals of potential concern

that may require further regulation [105], and the tool has been incorporated in several other screening models to identify emerging contaminants in water and soil [106,231]. The relevance and added value of the similarity tool is apparent from these applications, and its use has also been presented as a step towards a safer and healthier environment in a letter to the parliament by the Ministry of Infrastructure and Water Management (about actions on environmental safety and risks) [232]. Nevertheless, it is considered important to continue with the efforts of making this kind of knowledge on chemical toxicity more widely available and to raise awareness on chemicals of potential concern, especially at the places in society were direct decisions are being made (e.g. licensing authorities, research and development departments). It is considered crucial to have these actors involved when aiming for a toxic-free environment.

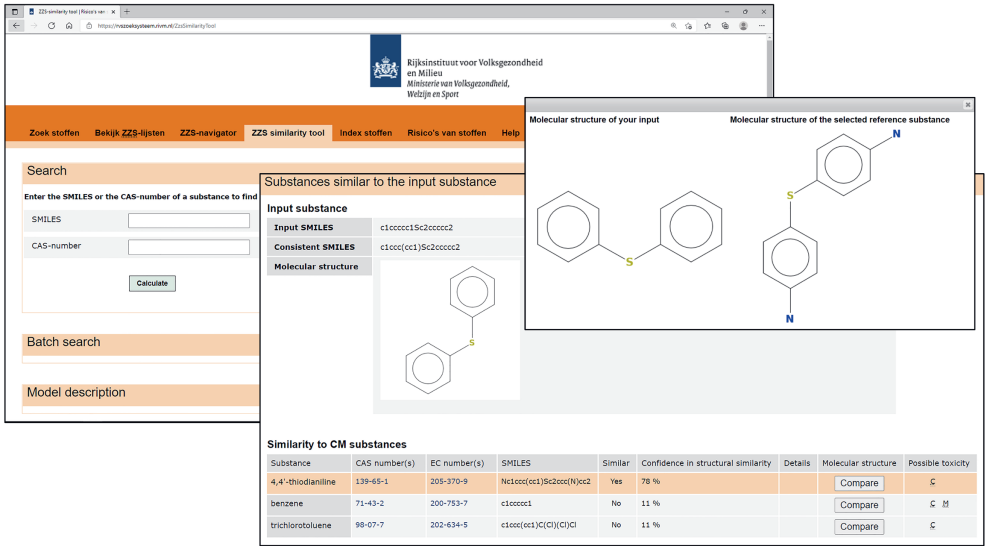


Figure 7.1. Overview of the ZZS similarity tool that is developed based on the scientific output of this thesis (chapters 2-4). In screen 1, a chemical can be provided as input (e.g. diphenyl sulphide, CAS: 139-66-2). In screen 2, the results of the ZZS similarity tool are reported. In screen 3, the structural similarity with the most similar ZZS is presented (i.e. 4,4'-thiodianiline, CAS: 139-65-1). Subsequently, the predictions should be subject to further evaluation (e.g. input structure is structurally comparable to the ZZS, but lacks aromatic amines and therefore is not likely to exert comparable ZZS-effects).

7.2 From screening to evaluation

As highlighted in chapters 2-4 and in paragraph 7.1, it is important to be able to signal concerns regarding potential SVHC properties. However, it is equally important and one of the key ambitions in the European chemical strategy [13] to also improve the early identification and regulation of hazardous chemicals before widespread exposure occurs. For instance, before

any follow-up can be given to predictions made by the similarity tool (and before definitive risk management measures can be introduced), further interpretation and evaluation of the potential concerns is required. In this case, the key question is whether chemical similarity between two chemicals does also translates into biological similarity (i.e. do they have the same or a predictable trend in biological activity)? Aspects related to this question have been touched upon in chapters 5 and 6, which specifically focused on PBT/vPvB related topics.

7.2.1 The influence of variability

Chapter 5 shows that there is a large variation in measured bioconcentration factors (BCFs) for individual substances, with an average standard deviation of 0.21 log BCF. A part of the observed variation can be explained by specific aspects in the experimental design (e.g. mixture exposure, calculation method and the selected test fish species), whereas the remaining (unexplained) variation is likely related to variation in laboratory practices and biological variation (chapter 5). Although not studied within this thesis, it is likely that comparable variations can be observed for other endpoints. For instance, Braakhuis et al. (2019) concluded that No and Low Observed Adverse Effect Levels (N(L)OAELs) for developmental toxicity in rats and rabbits may differ up to a factor of 25 for an individual substance [233]. These results indicate that significant variations might be present, despite several measures that have been taken to increase reproducibility of test results and evaluations, like the introduction of harmonized test guidelines (e.g. OECD, ASTM and OCSPP test guidelines), the availability of reliability assessment methods and technical guidance documents [19,234,235] and the use of reporting checklists [236,237]. Apparently, however, we still lack sufficient knowledge about the influence and importance of certain key factors on the test outcomes, including aspects related to test species and conditions. As also illustrated in chapter 5 - Table 5.5, this could have severe consequences (e.g. it can make the difference between SVHC identification or not) and could deteriorate the believe in – and question the robustness of – current risk and hazard assessments.

In order to account for and minimize the impact of variability, several aspects could be considered to improve the robustness of toxicity tests and evaluations. First, a number of additional studies could be considered instead of focusing on a single test outcome to encompass the diversity, as suggested in chapter 5. Accordingly, a more precise estimate of the (range of the) true effect(s) could be obtained. It should be noted however, that such an approach further expands the data needs (to which many substances currently already do not comply with), and that specific care should be taken not to hinder the transition to animal-free safety assessment, by specifically stimulating the use of alternative (non-vertebrate) test systems. For example, Mangold-Döring et al. (2021) aimed to capture the variability in the form of diversity across species by using *in silico* models [238]. They predicted a range of

BCFs for individual substances by modeling a set of ecologically plausible 'virtual species objects' that consisted of random combinations of actually occurring parameters. This type of models have the potential to provide insight in the effect size distribution, and as such could support risk and hazard assessments. Second, in order to minimize the use of (animal) tests, a more pragmatic approach could be followed by applying an uncertainty factor to the derived effect to account for the potential variability, as for instance suggested by Braakhuis et al. (2019) [233]. Third, a part of the variability could in some cases be eliminated by shifting to other (potentially less ambiguous) endpoints. For instance, Braakhuis et al. (2019) suggests to switch from N(L)OAELs to benchmark doses, which generally have a lower inherent uncertainty and this switch thus ensures a more equal evaluation [233]. And fourth, there is a trend of increasingly switching to more holistic based evaluations (and criteria) that make more efficient use of all (sources of) available information in the form of weight-of-evidence (WoE) approaches to reduce the impact of variability [239]. The reliability of conclusions derived from WoE evaluations is generally higher (i.e. uncertainties are lower) as it considers varying types of information (i.e. varying lines of evidence and metrics, like *in vivo*, *in vitro*, and *in silico* data), and weights them based on relevance, reliability and strength to create a coherent body of evidence [240]. Such a WoE approach has also been applied in chapter 6. In order to improve the transparency and consistency of WoE evaluations, aspects of systematic reviews (SR) could be applied. The strength of SR is particularly to systematically (in a planned and transparent manner) collect and evaluate all available data [240]. Although aspects of both methods (i.e. WoE and SR) are frequently combined, Suter et al. (2020) made some suggestions to integrate the best aspects of both methodologies to improve future risk and hazard assessments [240]. Nevertheless, it should be noted that for deriving reliable robust conclusions with WoE approaches a significant body of evidence must be available (see also chapter 6).

7.2.2 The use of chemical similarity

Following the results of chapter 5, it is clear that similarity cannot be evaluated independently from the potential high variability in derived effects of individual substances. Accordingly, it might be difficult to state with sufficient reliability that two or multiple structurally similar chemicals do or do not have comparable biological activity when the impact of variability might not be sufficiently quantified or reduced. Nevertheless, despite the potential presence of variability in biological activity, we were able to illustrate that chemical similarity could serve as a basis for the evaluation of PBT-properties in chapter 6. By using a WoE approach and trend analyses, we were able to derive a consistent trend and conclusion on the PBT-properties for the group of alkylated three-ring PAHs. Hence, this work confirms the underlying read-across hypothesis (for which chemical similarity forms the fundamental basis [26]), and thereby also confirms the viability and validity of chemical similarity as a screening and prioritization

feature for further chemical evaluation and regulation.

Additionally, however, this thesis specifically provides new insights into how chemical similarity can be applied for the evaluation of groups of similar substances, and as such could contribute to the transition from substance-by-substance assessments towards group assessment approaches. As described in chapter 1.6, this transition is necessary in order to make the regulatory framework more efficient, consistent and predictable [13,241]. Some explorations to stimulate this transition have already been conducted, including some initial group prioritizations and evaluations [29,96,97]. Nevertheless, further work (on case-studies) and guidance is required to identify and tackle the challenges ahead [29]. Remaining questions include: how to identify and define groups? How to streamline the process of group assessment? And how to regulate groups of substances? In chapter 3 of this thesis, we illustrate that the structural similarity models could be used to identify relevant groups of chemicals that are structurally similar to one or more SVHCs. Subsequently, such a group could be further refined and evaluated on biological similarity (i.e. bioavailability, degradation, bioaccumulation, physicochemical properties and toxicity) in order to evaluate and conclude whether the substances have the same or a predictable trend in biological activity, as exemplified in chapter 6. In chapter 6, we investigated a specific group of chemicals that resulted from the application of the hydrocarbon block method (HBM) and could all be part of a (petroleum) UVCB. Within this study, we were able to derive a conclusion for the PBT/vPvB properties of the whole group of 884 chemicals within one assessment by filling in data gaps for individual chemicals (constituents) without the need of extra tests. Accordingly, this study indicates great potential for the use of the HBM and provides valuable information on how trend analyses can be used for group evaluations.

In addition, several challenges were identified for future group evaluations, particularly when insufficient or inconsistent data are available that do not allow to derive a uniform conclusion. The type and amount of data that would be necessary in these situations will be dependent on a case-by-case basis, which will be related to the complexity of the chemical structures. Although theories have been devised on how to evaluate groups of chemicals and UVCBs [19,26], currently explicit experience and guidance is lacking for specific complex situations where insufficient data are available. Therefore, additional (follow-up) activities – in line with the study as reported in chapter 6 – are needed to investigate and realize the full potential of these theories or methodologies. For the assessment of groups of chemicals, specific attention is currently given to so-called new approach methodologies (including *in vitro* and *in silico* tools) that could potentially be used to fill in relevant data gaps necessary to justify a read-across hypothesis [242]. For the assessment of UVCBs specifically, besides insufficient data also other challenges are involved, like an unknown and variable composition [159]. Options on how to deal with these challenges are discussed at various (international) levels, including at

the Petroleum and Coal stream substances (PetCo) working group at ECHA [243]. The PetCo WG consists of member state competent authorities, the European Commission, ECHA and industry stakeholders, and works in close collaboration with other relevant groups like RIME (Risk Management and Evaluation platform) and the PBT expert group of ECHA. Their main aim is to identify and prioritize PetCo UVCBs for further work, find and apply methodologies to assess their impact, and discuss the most appropriate options to manage these substances. These and other multidisciplinary/multistakeholder initiatives (like collaborations [244] and workshops [159,245]) are necessary to understand the difficulties and offer integrated solutions for UVCB evaluations. Altogether, this will improve the risk and hazard estimation of UVCBs (by efficiently considering and evaluating the available data) and will contribute to a more targeted regulation of these substances, which is absolutely essential when realizing that approximately 40% of all REACH registered substances are considered UVCBs [160]. In order to optimize risk and hazard assessment as efficient as possible, we must ensure that the obtained knowledge on group and UVCB evaluations (as well as mixture effect evaluations) are mutually supportive and reinforcing, where possible.

7.3 The contribution of chemical similarity to future risk and hazard assessment

The results of this thesis indicate that chemical similarity can play a key role within risk and hazard assessment at the screening, data generation and evaluation phase. And as such, could contribute to the European ambitions of a toxic-free environment, as set out in the chemical strategy for sustainability [13]. Chemical similarity could be used to identify and evaluate SVHC-properties of groups of chemicals (chapters 2, 3, 4 and 6) and accordingly could identify hazardous chemicals of which the uses are preferably substituted or minimized. With this in mind, this thesis also indicates that chemical similarity is a useful metric to identify potential chemical alternatives in safe-by-design trajectories (chapter 2-4). As such, more extensive application of chemical similarity in risk and hazard assessments will contribute to a safer chemical environment as it enables more efficient use of available data and knowledge, and thereby also provides opportunities for animal-free safety testing.

Despite the potential benefits of the use of chemicals similarity and the ambitions of a toxic-free environment, we should realize that it remains unlikely that we will ever be able to assess and manage all hazards and risks before any exposure and/or effects occur. Many substances of concern of the past, present and potentially future, like polychlorinated biphenyls (PCBs), per- and polyfluoroalkyl substances (PFAS) and nanomaterials [20,246,247], are or were widely applied before adequate regulation is in place. First of all, this might be caused by the fact that these 'new-types' of substances of concern exert adverse effects in different ways that are or were not part of our universal knowledge of toxicity, and can often not be identified

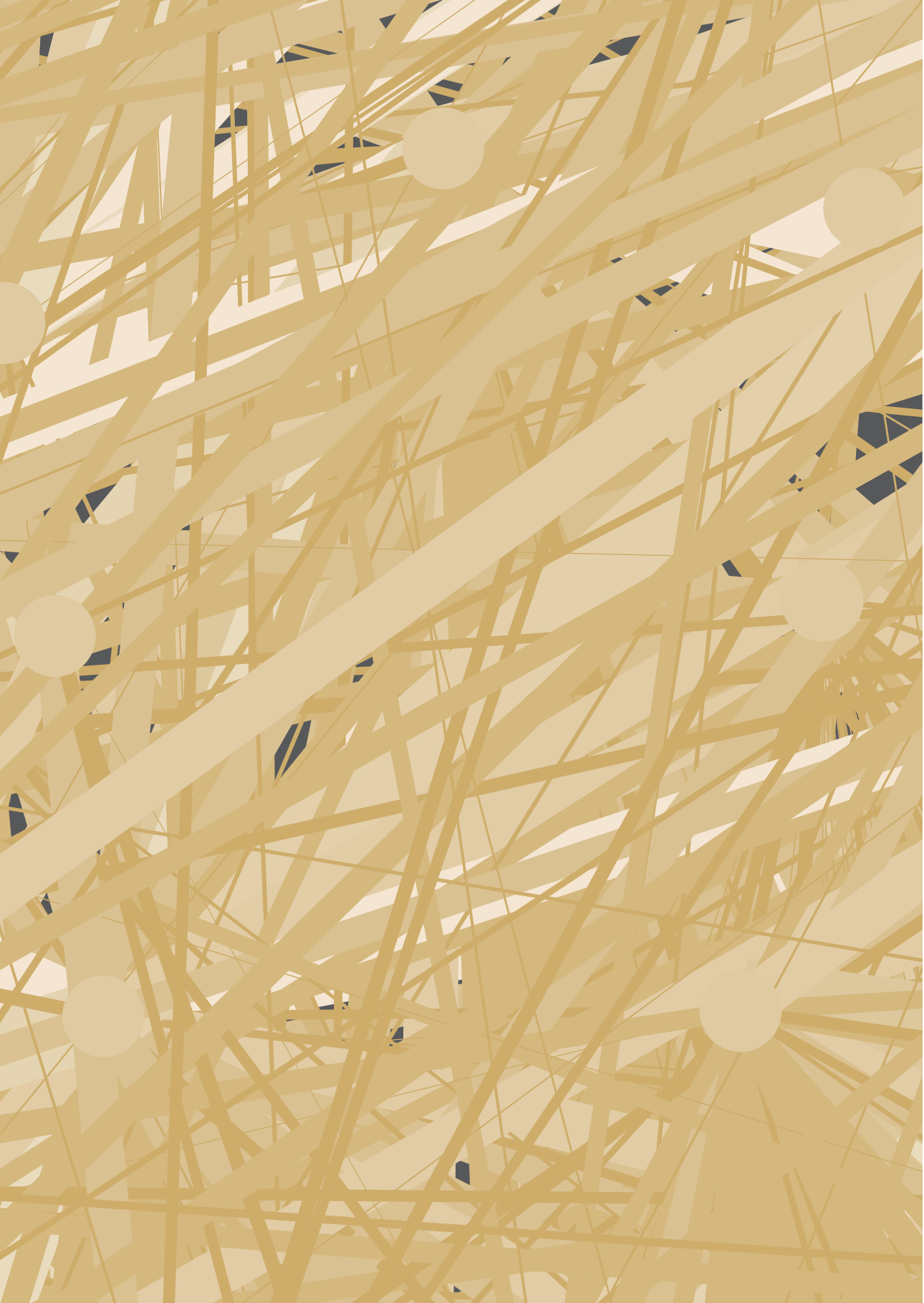
with the (at that point) standard experimental test setups. Therefore, such properties are also not (yet) included within any existing screening or prioritization model (e.g. the developed structural similarity models (chapters 2-4) will not identify any effects that are not comparable to any effects exerted by known chemicals of concern). Accordingly, such 'new' effects are generally only identified or observed after those chemicals are fully developed and applied, ultimately resulting in the implementation of appropriate regulatory measures. Second, our evaluations are challenged over the years as our scope from what we (are able to) consider to be adverse effects continuously evolves (e.g. current focus is extending to neurotoxicity, immunotoxicity, and extreme persistence and mobility). And third, some substances of very high concern, for which no suitable alternatives are available, are used for (societal) essential applications that outweigh the risks. Such cases may require further (chemical) innovations, but at least require a minimization of exposure and emissions as far as possible. In order to minimize the impact of chemicals in these inevitable situations, we need to ensure that the timelines for bringing chemicals under regulation are reduced. For instance, by streamlining the regulatory process and by reducing its workload [248], but also by emphasizing and prioritizing extensive safety testing at an early developmental stage (as for instance being aimed for with safe and sustainable-by-design [249]).

Although by definition we will never be able to know everything in advance, the use of chemical similarity proves to be a valuable aspect that should be exploited to accelerate and approach the transition to a toxic-free environment as closely as possible. It is within our capabilities to ensure that all chemicals that exert known types of concern are identified and regulated before any adverse effects from emissions or exposures occur (i.e. safe and sustainable-by-design). In order to achieve this overarching aim, it seems essential to make more efficient use of chemical similarity. Based on the results of this thesis, I want to provide several key aspects that require further emphasis to fully exploit the potential of chemical similarity to screen, prioritize and evaluate tomorrow's substances of concern. All these aspects focus on optimizing the use of available data.

First of all, I highly recommend to combine chemical similarity models with models that focus on biological activity, in order to advance to more comprehensive screening and prioritization of chemicals. This may include the use of (existing) predictive models (e.g. structural alerts and physicochemical properties) or combinations with other types of data, like high-throughput *in vitro* and omics test data. Such complex consensus models have the ability to provide more direct input for follow-up evaluations (whereas more simplistic models have their own benefits; see section 7.1). Furthermore, such models could help define more unambiguous groups of chemicals, and thereby could further stimulate the transition towards group assessment approaches, which I suggest as a second priority for future developments. Evaluations of groups of chemicals must become a standard practice within risk and hazard assessment, as

it is considered as one of the essential steps to make chemical legislation more efficient and effective. Developments in both these processes (i.e. screening and group evaluations) will be greatly stimulated by the (open access) availability of (large) toxicological datasets and advancements in computational toxicology. We already gained some experience with the use of supervised machine learning algorithms on large (imbalanced) *in vitro* datasets to predict endocrine active and disrupting chemicals, which showed promising perspectives for future applications [226]. Furthermore, more extensive use of machine learning approaches shows lots of possibilities for future risk and hazard assessments in the shorter- and long-term, both on the scientific-technical evaluation process (like gathering, extracting and organizing data as well as the detection of patterns in large datasets) and on the decision making process (like enhancement of the evaluation process and simulations of expert judgment) [24]. We should, however, ensure that results of newly developed models are transparent and interpretable, in order to guarantee community-wide (including regulatory) trust and acceptance. Moreover, in order to make use of these developments to its fullest extent and to acquire high-quality output, it is utmost essential to make use of high-quality data (as 'garbage in, is garbage out'). Data quality and reliability are still relevant and big issues nowadays [23], due to the fact that important (technical) study details are regularly not/inadequately reported or not/incorrectly applied (as also clearly observed in chapters 5 and 6). When developing models based on unreliable data, unreliable predictions are made. In addition, much can be improved with respect to data availability and accessibility [248]. Luckily clear data reporting [236,237] and responsible data management are attracting attention (including complying with the FAIR-principles: findable, accessible, interoperable and reusable) [250,251], and should become standard practice within future research to ensure that generated data can be used to its fullest extent.

In conclusion, based on the results of this thesis I promote more extensive use of chemical similarity within risk and hazard assessment as it has the ability to circumvent several issues related to a lack of data and evaluation efficiency. By further investing in the use of chemical similarity, in combination with other innovations (i.e. new approach methodologies [252]), we will be able to make the transition to a more efficient and effective chemical regulatory system. And as such, the use of chemical similarity allows us to consciously diminish unforeseen and unintended impacts of any of the substances that are part of the chemical universe, and thereby enables us to strengthen the protection of human health and the environment.



References

[1] Military quotes, USMC General Lewis B. Chesty Puller Quotes, (n.d.). https://www.military-quotes.com/chesty-puller.htm (accessed July 6, 2021).

[2] F. Aftalion, A history of the international chemical industry, Chemical Heritage Foundation, 2001.

[3] Z. Wang, G.W. Walker, D.C.G. Muir, K. Nagatani-Yoshida, Toward a Global Understanding of Chemical Pollution: A First Comprehensive Analysis of National and Regional Chemical Inventories, Environmental Science and Technology. 54 (2020) 2575–2584. https://doi.org/10.1021/acs.est.9b06379.

[4] M. Bastian, S. Heymann, M. Jacomy, Gephi: An Open Source Software for Exploring and Manipulating Networks, Third International AAAI Conference on Weblogs and Social Media. (2009).

[5] M. Jacomy, T. Venturini, S. Heymann, M. Bastian, ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software, PLoS ONE. 9 (2014). https://doi.org/10.1371/journal.pone.0098679.

[6] C.W. Yap, PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints, Journal of Computational Chemistry. (2011). https://doi.org/10.1002/jcc.21707.

[7] R. Todeschini, V. Consonni, H. Xiang, J. Holliday, M. Buscema, P. Willett, Similarity coefficients for binary chemoinformatics data: Overview and extended comparison using simulated and real data sets, Journal of Chemical Information and Modeling. 52 (2012) 2884–2901. https://doi.org/10.1021/ci300261r.

[8] ECHA - European Chemicals Agency, Guidance for identification and naming of substances under REACH and CLP, 2017. https://doi.org/10.2823/538683.

[9] Y. Djoumbou Feunang, R. Eisner, C. Knox, L. Chepelev, J. Hastings, G. Owen, E. Fahy, C. Steinbeck, S. Subramanian, E. Bolton, R. Greiner, D.S. Wishart, ClassyFire: automated chemical classification with a comprehensive, computable taxonomy, Journal of Cheminformatics. 8 (2016) 1–20. https://doi.org/10.1186/s13321-016-0174-y.

[10] J. Ravichandran, B.S. Karthikeyan, P. Singla, S.R. Aparna, A. Samal, NeurotoxKb 1.0: Compilation, curation and exploration of a knowledgebase of environmental neurotoxicants specific to mammals, Chemosphere. 278 (2021). https://doi.org/10.1016/j.chemosphere.2021.130387.

[11] F.I. Saldívar-González, C.S. Huerta-García, J.L. Medina-Franco, Chemoinformatics-based enumeration of chemical libraries: a tutorial, Journal of Cheminformatics. 12 (2020). https://doi.org/10.1186/s13321-020-00466-z.

[12] European Parliament, Council of the European Union, REACH Regulation (EC/1907/2006), 2006.

[13] European Commission, Chemicals Strategy for Sustainability - Towards a Toxic-Free Environment, 2020.

[14] J. van Dijk, A. Leopold, H. Flerlage, A. van Wezel, T.B. Seiler, M.H. Enrici, M.C. Bloor, The EU Green Deal's ambition for a toxic-free environment: Filling the gap for science-based policymaking, Integrated Environmental Assessment and Management. 17 (2021) 1105–1113. https://doi.org/10.1002/ieam.4429.

[15] European Parliament, Council of the European Union, CLP Regulation (EC/1272/2008), 2008.

[16] C.J. van Leeuwen, T.G. Vermeire, Risk Assessment of Chemicals - An Introduction, Springer Science & Business Media, 2007.

[17] C.A.M. van Gestel, F.G.A.J. van Belleghem, N.W. van den Brink, S.T.J. Droge, T. Hamers, J.L.M. Hermens, M.H.S. Kraak, A.J. Löhr, J.R. Parsons, A.M.J. Ragas, N.M. van Straalen, M.G.

Vijver, Chapter 6 'Risk Assessment & Regulation' in Environmental toxicology, an open online textbook. (online ed.), 2019. https://maken.wikiwijs.nl/120183/6__Risk_Assessment___Regulation (accessed June 9, 2021).

[18] Bureau KLB, Ameco, National Policies on Substances of Concern, 2017. www.bureauklb.nl.

[19] ECHA - European Chemicals Agency, Guidance on information requirements and chemical safety assessment (Chapter R.11: PBT/vPvB assessment), 2017. https://doi.org/10.2823/128621.

[20] S. Sauvé, M. Desrosiers, A review of what is an emerging contaminant, Chemistry Central Journal. 8 (2014) 15. https://doi.org/10.1186/1752-153X-8-15.

[21] H.L. Ciallella, H. Zhu, Advancing Computational Toxicology in the Big Data Era by Artificial Intelligence: Data-Driven and Mechanism-Driven Modeling for Chemical Toxicity, Chemical Research in Toxicology. 32 (2019) 536–547. https://doi.org/10.1021/acs.chemrestox.8b00393.

[22] G. Pawar, J.C. Madden, D. Ebbrell, J.W. Firman, M.T.D. Cronin, In silico toxicology data resources to support read-across and (Q)SAR, Frontiers in Pharmacology. 10 (2019). https://doi.org/10.3389/fphar.2019.00561.

[23] H. Yang, L. Sun, W. Li, G. Liu, Y. Tang, In Silico Prediction of Chemical Toxicity for Drug Design Using Machine Learning Methods and Structural Alerts, Frontiers in Chemistry. 6 (2018). https://doi.org/10.3389/fchem.2018.00030.

[24] C. Wittwehr, P. Blomstedt, J.P. Gosling, T. Peltola, B. Raffael, A.N. Richarz, M. Sienkiewicz, P. Whaley, A. Worth, M. Whelan, Artificial Intelligence for chemical risk assessment, Computational Toxicology. 13 (2020). https://doi.org/10.1016/j.comtox.2019.100114.

[25] Y. Wu, G. Wang, Machine learning based toxicity prediction: From chemical structural description to transcriptome analysis, International Journal of Molecular Sciences. 19 (2018). https://doi.org/10.3390/ijms19082358.

[26] ECHA - European Chemicals Agency, Read-Across Assessment Framework (RAAF), 2017. https://doi.org/10.2823/619212.

[27] OECD - Organisation for Economic Co-operation and Development, Guidance on grouping of chemicals, second edition (Series on testing and assessment No. 194), 2014.

[28] ECHA - European Chemicals Agency, Transparent progress in addressing substances of concern - Integrated Regulatory Strategy Annual Report, 2021. https://doi.org/10.2823/506792.

[29] ECHA - European Chemicals Agency, Grouping speeds up regulatory action, 2020. https://doi.org/10.2823/3244.

[30] M.A. Johnson, G.M. Maggiora, Concepts and Applications of Molecular Similarity, Wiley, New York, 1990.

[31] G. Maggiora, M. Vogt, D. Stumpfe, J. Bajorath, Molecular similarity in medicinal chemistry, Journal of Medicinal Chemistry. 57 (2014) 3186–3204. https://doi.org/10.1021/jm401411z.

[32] P. Willett, The calculation of molecular structural similarity: Principles and practice, Molecular Informatics. 33 (2014) 403–413. https://doi.org/10.1002/minf.201400024.

[33] R.D. Brown, Descriptors for diversity analysis, Perspectives in Drug Discovery and Design. 718 (1997) 31–49. https://doi.org/10.1007/BF03380180.

[34] S. Riniker, G.A. Landrum, Open-source platform to benchmark fingerprints for ligand-based virtual screening, Journal of Cheminformatics. 5 (2013) 26. https://doi.org/10.1186/1758-2946-5-26.

[35] European Parliament, Council of the European Union, Decision No 1386/2013/EU on a General Union Environment Action Programme to 2020 'Living well, within the limits of our planet,' 2013.

[36] European Commission, Fact Sheet: 7th Environmental Action Plan, (2013). http://ec.europa.eu/environment/pubs/pdf/factsheets/7eap/en.pdf (accessed February 2, 2019).

[37] US EPA - United States Environmental Protection Agency, Estimation Programs Interface

Suite for Microsoft Window, v4.1, (2012).

[38] E. Rorije, E.M.J. Verbruggen, A. Hollander, T.P. Traas, M.P.M. Janssen, Identifying potential POP and PBT substances : Development of a new Persistence/Bioaccumulation-score (RIVM Report 601356001), (2011) 1–88.

[39] R. Benigni, C. Bossa, N. Jeliazkova, T. Netzeva, A. Worth, The Benigni / Bossa rulebase for mutagenicity and carcinogenicity – a module of Toxtree, 2008.

[40] P. Banerjee, A.O. Eckert, A.K. Schrey, R. Preissner, ProTox-II: A webserver for the prediction of toxicity of chemicals, Nucleic Acids Research. (2018). https://doi.org/10.1093/nar/gky318.

[41] H. Yang, C. Lou, L. Sun, J. Li, Y. Cai, Z. Wang, W. Li, G. Liu, Y. Tang, AdmetSAR 2.0: Web-service for prediction and optimization of chemical ADMET properties, Bioinformatics. (2019). https://doi.org/10.1093/bioinformatics/bty707.

[42] BIOVIA, TOPKAT (Toxicity Prediction by Komputer Assisted Technology), (n.d.). https://www.3dsbiovia.com/products/datasheets/qsar-admet-and-predictive-toxicology-with-ds.pdf.

[43] MultiCase Inc, MultiCase, (n.d.). http://www.multicase.com/.

[44] Leadscope, Leadscope, (n.d.). http://www.leadscope.com/.

[45] DTU, Danish (Q)SAR Database, (2015). http://qsar.food.dtu.dk/.

[46] ChemSec, Methodology for grouping the SIN List and development of the SINimilarity tool, (2015). https://chemsec.org/publication/sin-list/methodology-for-grouping-the-sin-list-and-development-of-the-sinimilarity-tool/.

[47] ECHA - European Chemicals Agency, Screening Definition Document - Methodology for identifying (groups of) potential substances of concern for (further) regulatory action, 2019. https://echa.europa.eu/documents/10162/19126370/screening_definition_document_en.pdf/e588a9f8-c55e-4412-a760-49ddbf7ac687.

[48] R. Todeschini, V. Consonni, Molecular descriptors for chemoinformatics: volume I: alphabetical listing/volume II: appendices, references., John Wiley and Sons, 2009.

[49] P. Willett, J.M. Barnard, G.M. Downs, Chemical similarity searching, Journal of Chemical Information and Computer Sciences. (1998). https://doi.org/10.1021/ci9800211.

[50] C.L. Mellor, R.L. Marchese Robinson, R. Benigni, D. Ebbrell, S.J. Enoch, J.W. Firman, J.C. Madden, G. Pawar, C. Yang, M.T.D. Cronin, Molecular fingerprint-derived similarity measures for toxicological read-across: Recommendations for optimal use, Regulatory Toxicology and Pharmacology. (2019). https://doi.org/10.1016/j.yrtph.2018.11.002.

[51] RIVM - National Institute for Public Health and the Environment, List of Dutch Substances of Very High Concern [in Dutch], (2018). https://rvszoeksysteem.rivm.nl/ZZSlijst/Index (accessed March 1, 2018).

[52] ECHA - European Chemicals Agency, Biocidal Active Substances., (2018). http://echa.europa.eu/web/guest/information-on-chemicals/biocidal-active-substances (accessed May 23, 2018).

[53] European Commission, EU Pesticides Database, (2018). http://ec.europa.eu/food/plant/pesticides/eu-pesticides-database/public/?event=activesubstance.selection&language=EN (accessed May 23, 2018).

[54] Daylight, SMILES - A Simplified Chemical Language, (2008). http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html (accessed February 15, 2019).

[55] G. Landrum, RDKit: Open-source Cheminformatics and machine-learning, http://www.rdkit.org/. (2019).

[56] M. Floris, A. Manganaro, O. Nicolotti, R. Medda, G.F. Mangiatordi, E. Benfenati, A generalizable definition of chemical similarity for read-across, Journal of Cheminformatics. (2014). https://doi.org/10.1186/s13321-014-0039-1.

[57] S. Wu, J. Fisher, J. Naciff, M. Laufersweiler, C. Lester, G. Daston, K. Blackburn, Framework

for identifying chemicals with structural features associated with the potential to act as developmental or reproductive toxicants, Chemical Research in Toxicology. (2013). https://doi.org/10.1021/tx400226u.

[58] R Core Team, R: A Language and Environment for Statistical Computing, (2019).

[59] M. Kuhn, R package, caret: Classification and Regression Training. Version 6.0-81, CRAN. (2018).

[60] Y. Cao, A. Charisi, L.C. Cheng, T. Jiang, T. Girke, ChemmineR: A compound mining framework for R, Bioinformatics. (2008). https://doi.org/10.1093/bioinformatics/btn307.

[61] J. Tuszynski, caTools: Tools: moving window statistics, GIF, Base64, ROC AUC, etc. R package version 1.17.1.1, CRAN. (2018).

[62] T. Sing, O. Sander, N. Beerenwinkel, T. Lengauer, ROCR: Visualizing classifier performance in R, Bioinformatics. (2005). https://doi.org/10.1093/bioinformatics/bti623.

[63] R. Guha, Chemical Informatics Functionality in R, Journal of Statistical Software. (2007). https://doi.org/10.18637/jss.v018.i05.

[64] MACCS, Molecular ACCess System (MACCS) keys. MDL Information Systems - As interpreted by CDK, (n.d.).

[65] Daylight, Fingerprints - Screening and Similarity, (2008). http://www.daylight.com/dayhtml/doc/theory/theory.finger.html (accessed February 15, 2019).

[66] Å. Bergman, J. Heindel, S. Jobling, K. Kidd, R. Zoeller, Endocrine Disrupting Chemicals - 2012, 2013.

[67] D. Rogers, M. Hahn, Extended-connectivity fingerprints, Journal of Chemical Information and Modeling. (2010). https://doi.org/10.1021/ci100050t.

[68] ECHA - European Chemicals Agency, EFSA - European Food Safety Authority, JRC - Joint Research Centre, Guidance for the identification of endocrine disruptors in the context of Regulations (EU) No 528/2012 and (EC) No 1107/2009, EFSA Journal. (2018). https://doi.org/10.2903/j.efsa.2018.5311.

[69] International Agency for Research on Cancer, Glyphosate Monograph, 2015.

[70] J. V. Tarazona, D. Court-Marques, M. Tiramani, H. Reich, R. Pfeil, F. Istace, F. Crivellente, Glyphosate toxicity and carcinogenicity: a review of the scientific basis of the European Union assessment and its differences with IARC, Archives of Toxicology. (2017). https://doi.org/10.1007/s00204-017-1962-5.

[71] ECHA - European Chemicals Agency, Search for Chemicals, (2019). https://echa.europa.eu/en/information-on-chemicals (accessed September 20, 2019).

[72] M. Woutersen, M. Beekman, M.E.J. Pronk, A. Muller, J.A. de Knecht, B.C. Hakkert, Does REACH provide sufficient information to regulate mutagenic and carcinogenic substances?, Human and Ecological Risk Assessment. (2018). https://doi.org/10.1080/10807039.2018.1480351.

[73] P.N.H. Wassenaar, E. Rorije, N.M.H. Janssen, W.J.G.M. Peijnenburg, M.G. Vijver, Chemical similarity to identify potential Substances of Very High Concern – An effective screening method, Computational Toxicology. 12 (2019). https://doi.org/10.1016/j.comtox.2019.100110.

[74] B. Settles, Active Learning Literature Survey - Computer Sciences Technical Report 1648. University of Wisconsin–Madison, 2010.

[75] X. Yang, Y. Wang, R. Byrne, G. Schneider, S. Yang, Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery, Chemical Reviews. (2019). https://doi.org/10.1021/acs.chemrev.8b00728.

[76] ECHA - European Chemicals Agency, Registered substances, (2019). https://echa.europa.eu/information-on-chemicals/registered-substances (accessed May 17, 2019).

[77] D. Gadaleta, A. Lombardo, C. Toma, E. Benfenati, A new semi-automated workflow for chemical data retrieval and quality checking for modeling applications, Journal of

Cheminformatics. (2018). https://doi.org/10.1186/s13321-018-0315-6.

[78] V.D. Blondel, J.L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, Journal of Statistical Mechanics: Theory and Experiment. (2008). https://doi.org/10.1088/1742-5468/2008/10/P10008.

[79] C.W. Yap, PaDEL-Descriptor web page, (2014). http://www.yapcwsoft.com/dd/padeldescriptor/ (accessed May 27, 2020).

[80] R. Likert, A technique for the measurement of attitudes, Archives of Psychology. (1932).

[81] P. Franco, N. Porta, J.D. Holliday, P. Willett, The use of 2D fingerprint methods to support the assessment of structural similarity in orphan drug legislation, Journal of Cheminformatics. (2014). https://doi.org/10.1186/1758-2946-6-5.

[82] ECHA - European Chemicals Agency, Substance Infocard - 4,4'-isopropylidenediphenol, (2020). https://echa.europa.eu/substance-information/-/substanceinfo/100.001.133 (accessed May 27, 2020).

[83] A.M. Soto, C. Schaeberle, M.S. Maier, C. Sonnenschein, M. V. Maffini, Evidence of Absence: Estrogenicity Assessment of a New Food-Contact Coating and the Bisphenol Used in Its Synthesis, Environmental Science and Technology. (2017). https://doi.org/10.1021/acs.est.6b04704.

[84] ECHA - European Chemicals Agency, Substance Infocard - 4,4'-methylenedi-2,6-xylenol, (2020). https://echa.europa.eu/substance-information/-/substanceinfo/100.023.980 (accessed May 27, 2020).

[85] ECHA - European Chemicals Agency, SVHC Support Document - 4,4'-Isopropylidenediphenol (Bisphenol A), (2017). https://echa.europa.eu/documents/10162/908badc9-e65d-3bae-933a-3512a9262e59 (accessed July 17, 2020).

[86] K. Mansouri, N. Kleinstreuer, A.M. Abdelaziz, D. Alberga, V.M. Alves, P.L. Andersson, C.H. Andrade, F. Bai, I. Balabin, D. Ballabio, E. Benfenati, B. Bhhatarai, S. Boyer, J. Chen, V. Consonni, S. Farag, D. Fourches, A.T. García-Sosa, P. Gramatica, F. Grisoni, C.M. Grulke, H. Hong, D. Horvath, X. Hu, R. Huang, N. Jeliazkova, J. Li, X. Li, H. Liu, S. Manganelli, G.F. Mangiatordi, U. Maran, G. Marcou, T. Martin, E. Muratov, D.T. Nguyen, O. Nicolotti, N.G. Nikolov, U. Norinder, E. Papa, M. Petitjean, G. Piir, P. Pogodin, V. Poroikov, X. Qiao, A.M. Richard, A. Roncaglioni, P. Ruiz, C. Rupakheti, S. Sakkiah, A. Sangion, K.W. Schramm, C. Selvaraj, I. Shah, S. Sild, L. Sun, O. Taboureau, Y. Tang, I. V. Tetko, R. Todeschini, W. Tong, D. Trisciuzzi, A. Tropsha, G. Van Den Driessche, A. Varnek, Z. Wang, E.B. Wedebye, A.J. Williams, H. Xie, A. V. Zakharov, Z. Zheng, R.S. Judson, Compara: Collaborative modeling project for androgen receptor activity, Environmental Health Perspectives. (2020). https://doi.org/10.1289/EHP5580.

[87] ECHA - European Chemicals Agency, Substance Infocard - 2,2',6,6'-tetra-tert-butyl-4,4'-methylenediphenol, (2020). https://echa.europa.eu/substance-information/-/substanceinfo/100.003.891 (accessed May 27, 2020).

[88] NTP - National Toxicology Program, Biological Activity of Bisphenol A (BPA) Structural Analogues and Functional Alternatives, 2017. https://doi.org/10.22427/ntp-rr-4.

[89] R. Karaman, Prodrugs design based on inter- and intramolecular chemical processes, Chemical Biology and Drug Design. (2013). https://doi.org/10.1111/cbdd.12224.

[90] C. Lester, A. Reis, M. Laufersweiler, S. Wu, K. Blackburn, Structure activity relationship (SAR) toxicological assessments: The role of expert judgment, Regulatory Toxicology and Pharmacology. (2018). https://doi.org/10.1016/j.yrtph.2017.12.026.

[91] M.S. Lajiness, G.M. Maggiora, V. Shanmugasundaram, Assessment of the consistency of medicinal chemists in reviewing sets of compounds, Journal of Medicinal Chemistry. (2004). https://doi.org/10.1021/jm049740z.

[92] P.S. Kutchukian, N.Y. Vasilyeva, J. Xu, M.K. Lindvall, M.P. Dillon, M. Glick, J.D. Coley, N. Brooijmans, Inside the Mind of a Medicinal Chemist: The Role of Human Bias in Compound Prioritization during Drug Discovery, PLoS ONE. (2012). https://doi.org/10.1371/journal.

pone.0048476.

[93] M.D. Hack, D.N. Rassokhin, C. Buyck, M. Seierstad, A. Skalkin, P. Ten Holte, T.K. Jones, T. Mirzadegan, D.K. Agrafiotis, Library enhancement through the wisdom of crowds, Journal of Chemical Information and Modeling. (2011). https://doi.org/10.1021/ci200446y.

[94] A. Lombardo, F. Pizzo, E. Benfenati, A. Manganaro, T. Ferrari, G. Gini, A new in silico classification model for ready biodegradability, based on molecular fragments, Chemosphere. (2014). https://doi.org/10.1016/j.chemosphere.2014.02.073.

[95] KEMI - Swedish Chemicals Agency, Grouping of chemical substances in the REACH and CLP regulations - PM 2/18, 2018.

[96] ECHA - European Chemicals Agency, Five European states call for evidence on broad PFAS restriction, ECHA/NR/20/13. (2020).

[97] ECHA - European Chemicals Agency, Collaborative approach pilot projects - Final report, 2018. https://doi.org/10.2823/224234.

[98] H. Moriwaki, Y.S. Tian, N. Kawashita, T. Takagi, Mordred: A molecular descriptor calculator, Journal of Cheminformatics. (2018). https://doi.org/10.1186/s13321-018-0258-y.

[99] A. Fernández, A. Lombardo, R. Rallo, A. Roncaglioni, F. Giralt, E. Benfenati, Quantitative consensus of bioaccumulation models for integrated testing strategies, Environment International. (2012). https://doi.org/10.1016/j.envint.2012.03.004.

[100] D. Ballabio, F. Biganzoli, R. Todeschini, V. Consonni, Qualitative consensus of QSAR ready biodegradability predictions, Toxicological and Environmental Chemistry. (2017). https://doi.org/10.1080/02772248.2016.1260133.

[101] A. Cereto-Massagué, M.J. Ojeda, C. Valls, M. Mulero, S. Garcia-Vallvé, G. Pujadas, Molecular fingerprint similarity search in virtual screening, Methods. 71 (2015) 58–63. https://doi.org/10.1016/j.ymeth.2014.08.005.

[102] T.G. Kristensen, J. Nielsen, C.N.S. Pedersen, Methods for similarity-based virtual screening, Computational and Structural Biotechnology Journal. 5 (2013) e201302009. https://doi.org/10.5936/csbj.201302009.

[103] Y. Yang, J. Zhan, Y. Zhou, SPOT-Ligand: Fast and effective structure-based virtual screening by binding homology search according to ligand and receptor similarity, Journal of Computational Chemistry. 37 (2016) 1734–1739. https://doi.org/10.1002/jcc.24380.

[104] P.N.H. Wassenaar, E. Rorije, M.G. Vijver, W.J.G.M. Peijnenburg, Evaluating chemical similarity as a measure to identify potential substances of very high concern, Regulatory Toxicology and Pharmacology. (2021). https://doi.org/10.1016/j.yrtph.2020.104834.

[105] C.E. Smit, P.N.H. Wassenaar, L. de Boer, N.M.H. Janssen, Research into substances of potential concern in Dutch surface water [In Dutch], H2O. (2021). https://www.h2owaternetwerk.nl/vakartikelen/onderzoek-naar-mogelijk-zorgwekkende-stoffen-in-nederlands-oppervlaktewater (accessed December 1, 2021).

[106] J. Hartmann, E.M.J. Verbruggen, E. Rorije, M. van der Aa, P.N.H. Wassenaar, A. Bannink, Screening and prioritising PMT substances: development of a robust T-score, (2021). https://www.umweltbundesamt.de/sites/default/files/medien/362/dokumente/day_2_afternoon_02_julia_hartmann.pdf (accessed August 12, 2021).

[107] RIVM - National Institute for Public Health and the Environment, List of Dutch Substances of Very High Concern [in Dutch], (2021). https://rvszoeksysteem.rivm.nl/ZZSlijst/Index (accessed January 25, 2021).

[108] OSPAR, Chemicals for Priority Action, (2020). https://www.ospar.org/work-areas/hasec/hazardous-substances/priority-action (accessed December 1, 2021).

[109] K. Mansouri, C. Grulke, R. Judson, A. Richard, A. Williams, N. Kleinstreuer, Open-source QSAR-ready chemical structure standardization workflow, 2021. https://doi.org/10.23645/epacomptox.15070041.v1.

[110] US EPA - United States Environmental Protection Agency, Chemistry Dashboard Data:

DSSTox QSAR Ready File. , (2018). https://doi.org/10.23645/epacomptox.6253679.v1.

[111] K. Mansouri, Standardization workflow for QSAR-ready chemical structures pretreatment., (2020). https://github.com/kmansouri/QSAR-ready (accessed March 10, 2021).

[112] R Core Team, R: A Language and Environment for Statistical Computing, (2021).

[113] M. Kuhn, R package, caret: Classification and Regression Training. Version 6.0-90, CRAN. (2021). https://CRAN.R-project.org/package=caret (accessed December 6, 2021).

[114] J. Tuszynski, caTools: Tools: Moving Window Statistics, GIF, Base64, ROC AUC, etc. R package version 1.18.2, CRAN. (2021). https://CRAN.R-project.org/package=caTools (accessed December 6, 2021).

[115] G. Landrum, RDKit: Open-source Cheminformatics and machine-learning, http://www.rdkit.org/. (2021).

[116] ECHA - European Chemicals Agency, Guidance on information requirements and chemical safety assessment (Chapter R.7c: Endpoint specific guidance), 2017. https://doi.org/10.2823/43472.

[117] JRC - Joint Research Centre, Review of available criteria for non-aquatic organisms within PBT/vPvB frameworks - Part I: Bioaccumulation assessment, 2014. https://doi.org/10.2788/11347.

[118] OECD - Organisation for Economic Co-operation and Development, OECD Guidelines for the Testing of Chemicals - Bioaccumulation in Fish: Aqueous and Dietary Exposure (OECD TG 305), 2012.

[119] H.J. Geyer, G.G. Rimkus, I. Scheunert, A. Kaune, K.-W. Schramm, A. Kettrup, M. Zeeman, D.C.G. Muir, L.G. Hansen, D. Mackay, Bioaccumulation and Occurrence of Endocrine-Disrupting Chemicals (EDCs), Persistent Organic Pollutants (POPs), and Other Organic Compounds in Fish and Other Organisms Including Humans, in: Bioaccumulation – New Aspects and Developments, 1987. https://doi.org/10.1007/10503050_1.

[120] ECHA - European Chemicals Agency, Evaluation under REACH: Progress Report 2017 - 10 years of experience, 2017. https://doi.org/10.2823/76886.

[121] F. Grisoni, V. Consonni, S. Villa, M. Vighi, R. Todeschini, QSAR models for bioconcentration: Is the increase in the complexity justified by more accurate predictions?, Chemosphere. (2015). https://doi.org/10.1016/j.chemosphere.2015.01.047.

[122] J. Nichols, K. Fay, M.J. Bernhard, I. Bischof, J. Davis, M. Halder, J. Hu, K. Johanning, H. Laue, D. Nabb, C. Schlechtriem, H. Segner, J. Swintek, J. Weeks, M. Embryc, Reliability of in vitro methods used to measure intrinsic clearance of hydrophobic organic chemicals by rainbow trout: Results of an international ring trial, Toxicological Sciences. (2018). https://doi.org/10.1093/toxsci/kfy113.

[123] C. Schlechtriem, S. Kampe, H.J. Bruckert, I. Bischof, I. Ebersbach, V. Kosfeld, M. Kotthoff, C. Schäfers, J. L'Haridon, Bioconcentration studies with the freshwater amphipod Hyalella azteca: are the results predictive of bioconcentration in fish?, Environmental Science and Pollution Research. (2019). https://doi.org/10.1007/s11356-018-3677-4.

[124] J. Sanz-Landaluze, M. Pena-Abaurrea, R. Muñoz-Olivas, C. Cámara, L. Ramos, Zebrafish (danio rerio) eleutheroembryo-based procedure for assessing bioaccumulation, Environmental Science and Technology. (2015). https://doi.org/10.1021/es504685c.

[125] J.A. Arnot, F.A. Gobas, A review of bioconcentration factor (BCF) and bioaccumulation factor (BAF) assessments for organic chemicals in aquatic organisms, Environmental Reviews. (2006). https://doi.org/10.1139/a06-005.

[126] NITE - National Institute of Technology and Evaluation, The biodegradation and bioconcentration data under CSCL have been published, (2018). https://www.nite.go.jp/en/chem/kasinn/cscl_data.html (accessed April 10, 2018).

[127] T.F. Parkerton, J.A. Arnot, A. V. Weisbrod, C. Russom, R.A. Hoke, K. Woodburn, T. Traas, M. Bonnell, L.P. Burkhard, M.A. Lampi, Guidance for evaluating in vivo fish bioaccumulation data, Integrated Environmental Assessment and Management. (2008). https://doi.

org/10.1897/IEAM_2007-057.1.

[128] OECD - Organisation for Economic Co-operation and Development, QSAR Toolbox, (2019). https://qsartoolbox.org/.

[129] US EPA - United States Environmental Protection Agency, ECOTOX, (2018). https://cfpub.epa.gov/ecotox/ (accessed April 10, 2018).

[130] R. Froese, D. Pauly, FishBase, (2019). www.fishbase.org (accessed February 12, 2019).

[131] J. Pinheiro, D. Bates, S. DebRoy, D. Sarkar, the R.C. Team, The nlme Package: Linear and Nonlinear Mixed Effects Models, R-Project. (2018).

[132] K.P. Burnham, D.R. Anderson, Model selection and multimodel inference: a practical information-theoretic approach, Second Edition, 2002.

[133] R. Carlson, A. Oyler, E. Gerhart, R. Caple, K. Welch, H.B.D. and S.D. Kopperman, Implications to the aquatic environment of polynuclear aromatic hydrocarbons liberated from Northern Great Plains coal, 1979.

[134] H. Wang, Y. Li, X. Xia, X. Xiong, Relationship between metabolic enzyme activities and bioaccumulation kinetics of PAHs in zebrafish (Danio rerio), Journal of Environmental Sciences (China). (2018). https://doi.org/10.1016/j.jes.2017.03.037.

[135] W. Larisch, K.U. Goss, Uptake, distribution and elimination of chemicals in fish – Which physiological parameters are the most relevant for toxicokinetics?, Chemosphere. (2018). https://doi.org/10.1016/j.chemosphere.2018.07.112.

[136] J.W. Nichols, J.M. McKim, M.E. Andersen, M.L. Gargas, H.J. Clewell, R.J. Erickson, A physiologically based toxicokinetic model for the uptake and disposition of waterborne organic chemicals in fish, Toxicology and Applied Pharmacology. (1990). https://doi.org/10.1016/0041-008X(90)90338-U.

[137] D. Schlenk, M. Celander, E.P. Gallagher, S. George, M. James, S.W. Kullman, P. Van Den Hurk, K. Willett, Biotransformation in fishes, in: The Toxicology of Fishes, 2008. https://doi.org/10.1201/9780203647295.

[138] M. Ferreira, J. Costa, M.A. Reis-Henriques, ABC transporters in fish species: A review, Frontiers in Physiology. (2014). https://doi.org/10.3389/fphys.2014.00266.

[139] P.N. Fitzsimmons, G.J. Lien, J.W. Nichols, A compilation of in vitro rate and affinity values for xenobiotic biotransformation in fish, measured under physiological conditions, Comparative Biochemistry and Physiology - C Toxicology and Pharmacology. (2007). https://doi.org/10.1016/j.cbpc.2006.12.011.

[140] J.A. Arnot, D. Mackay, T.F. Parkerton, M. Bonnell, A database of fish biotransformation rates for organic chemicals, Environmental Toxicology and Chemistry. (2008). https://doi.org/10.1897/08-058.1.

[141] I. Bischof, J. Köster, H. Segner, C. Schlechtriem, Hepatocytes as in vitro test system to investigate metabolite patterns of pesticides in farmed rainbow trout and common carp: Comparison between in vivo and in vitro and across species, Comparative Biochemistry and Physiology Part - C: Toxicology and Pharmacology. (2016). https://doi.org/10.1016/j.cbpc.2016.05.003.

[142] G.I. Petersen, P. Kristensen, Bioaccumulation of lipophilic substances in fish early life stages, Environmental Toxicology and Chemistry. (1998). https://doi.org/10.1897/1551-5028(1998)017<1385:BOLSIF>2.3.CO;2.

[143] P. Kristensen, N. Nyholm, Bioaccumulation of Chemical Substances in Fish Flow-Through Method: Ring-Test Programme 1984 - 1985, 1987.

[144] K.M. Kleinow, M.J. Melancon, J.J. Lech, Biotransformation and induction: Implications for toxicity, bioaccumulation and monitoring of environmental xenobiotics in fish, Environmental Health Perspectives. (1987).

[145] K. Monostory, K. Jemnitz, L. Vereczkey, Xenobiotic metabolizing enzymes in fish: diversity, regulation and biomarkers for pollutant exposure, Acta Physiologica Hungarica. (1996).

[146] D. Chabot, R. Koenker, A.P. Farrell, The measurement of specific dynamic action in fishes, Journal of Fish Biology. (2016). https://doi.org/10.1111/jfb.12836.

[147] A. Tandler, F.W.H. Beamish, Apparent specific dynamic action (SDA), fish weight and level of caloric intake in largemouth bass, Micropterus salmoides Lacepede, Aquaculture. (1981). https://doi.org/10.1016/0044-8486(81)90017-X.

[148] D. Randall, Control and co-ordination of gas exchange in water breathers, in: Springer (Ed.), Vertebrate Gas Exchange, Berlin, 1990.

[149] R. Yang, C. Brauner, V. Thurston, J. Neuman, D.J. Randall, Relationship between toxicant transfer kinetic processes and fish oxygen consumption, Aquatic Toxicology. (2000). https://doi.org/10.1016/S0166-445X(99)00050-8.

[150] M. Adolfsson-Erici, G. Åkerman, M.S. Mclachlan, Measuring bioconcentration factors in fish using exposure to multiple chemicals and internal benchmarking to correct for growth dilution, Environmental Toxicology and Chemistry. (2012). https://doi.org/10.1002/etc.1897.

[151] N.B. Metcalfe, T.E. Van Leeuwen, S.S. Killen, Does individual variation in metabolic phenotype predict fish behaviour and performance?, Journal of Fish Biology. (2016). https://doi.org/10.1111/jfb.12699.

[152] C. Toms, D. Echevarria, D. Jouandot, A Methodological Review of Personality-Related Studies in Fish: Focus on the Shy-Bold Axis Behaviour, International Journal of Comparative Psychology. (2010). https://doi.org/10.1016/j.socscimed.2010.01.002.

[153] P.A. Biro, N.J. Dingemanse, Sampling bias resulting from animal personality, Trends in Ecology and Evolution. (2009). https://doi.org/10.1016/j.tree.2008.11.001.

[154] L. Böhm, R.A. Düring, H.J. Bruckert, C. Schlechtriem, Can solid-phase microextraction replace solvent extraction for water analysis in fish bioconcentration studies with highly hydrophobic organic chemicals?, Environmental Toxicology and Chemistry. (2017). https://doi.org/10.1002/etc.3854.

[155] J.L. Van Geest, D.G. Poirier, K.R. Solomon, P.K. Sibley, A comparison of the bioaccumulation potential of three freshwater organisms exposed to sediment-associated contaminants under laboratory conditions, Environmental Toxicology and Chemistry. (2011). https://doi.org/10.1002/etc.456.

[156] W. de Wolf, M. Comber, P. Douben, S. Gimeno, M. Holt, M. Léonard, A. Lillicrap, D. Sijm, R. van Egmond, A. Weisbrod, G. Whale, Animal use replacement, reduction, and refinement: development of an integrated testing strategy for bioconcentration of chemicals in fish., Integrated Environmental Assessment and Management. (2007). https://doi.org/10.1897/1551-3793(2007)3[3:AURRAR]2.0.CO;2.

[157] European Parliament, Council of the European Union, Directive 2010/63/EU on the protection of animals used for scientific purposes, Official Journal of the European Union. (2010).

[158] R.J. Quann, Modeling the chemistry of complex petroleum mixtures, in: Environmental Health Perspectives, 1998. https://doi.org/10.1289/ehp.98106s61441.

[159] D. Salvito, M. Fernandez, K. Jenner, D.Y. Lyon, J. de Knecht, P. Mayer, M. MacLeod, K. Eisenreich, P. Leonards, R. Cesnaitis, M. León-Paumen, M. Embry, S.E. Déglin, Improving the Environmental Risk Assessment of Substances of Unknown or Variable Composition, Complex Reaction Products, or Biological Materials, Environmental Toxicology and Chemistry. 39 (2020) 2097–2108. https://doi.org/10.1002/etc.4846.

[160] S.S. Kutsarova, D.G. Yordanova, Y.H. Karakolev, S. Stoeva, M. Comber, C.B. Hughes, E. Vaiopoulou, S.D. Dimitrov, O.G. Mekenyan, UVCB substances II: Development of an endpoint-nonspecific procedure for selection of computationally generated representative constituents, Environmental Toxicology and Chemistry. (2019). https://doi.org/10.1002/etc.4358.

[161] CONCAWE, REACH - Roadmap for Petroleum Substances, 2019.

[162] CONCAWE, Environmental risk assessment of petroleum substances: the hydrocarbon block

method - report no. 96/52, 1996.

[163] ECHA - European Chemicals Agency, Candidate List of substances of very high concern for Authorisation, (2020). https://echa.europa.eu/en/candidate-list-table (accessed June 11, 2020).

[164] InfoCuria, Action brought on 21 March 2019 – Exxonmobil Petroleum & Chemical v ECHA (Case T-177/19), 2019.

[165] C.B. Hughes, D.M. Brown, L. Camenzuli, A.D. Redman, J.S. Arey, D. Vione, N. Wang, E. Vaiopoulou, Can a chemical be both readily biodegradable AND very persistent (vP)? Weight-of-evidence determination demonstrates that phenanthrene is not persistent in the environment, Environmental Sciences Europe. 32 (2020). https://doi.org/10.1186/s12302-020-00427-1.

[166] D.M. Pampanin, M.O. Sydnes, Chapter 5: Polycyclic aromatic hydrocarbons a constituent of Petroleum: Presence and influence in the aquatic environment, in: Hydrocarbon, 2013.

[167] S.A. Stout, S.D. Emsbo-Mattingly, G.S. Douglas, A.D. Uhler, K.J. McCarthy, Beyond 16 Priority Pollutant PAHs: A Review of PACs used in Environmental Forensic Chemistry, Polycyclic Aromatic Compounds. (2015). https://doi.org/10.1080/10406638.2014.891144.

[168] OECD - Organisation for Economic Co-operation and Development, OECD Guidelines for the Testing of Chemicals - Aerobic and Anaerobic Transformation in Soil (OECD TG 307), 2002.

[169] OECD - Organisation for Economic Co-operation and Development, OECD Guidelines for the Testing of Chemicals - Aerobic and Anaerobic Transformation in Aquatic Sediment Systems (OECD TG 308), 2002.

[170] OECD - Organisation for Economic Co-operation and Development, OECD Guidelines for the Testing of Chemicals - Aerobic Mineralisation in Surface Water – Simulation Biodegradation Test (OECD TG 309), 2004.

[171] ECHA - European Chemicals Agency, Guidance on Information Requirements and Chemical Safety Assessment (Chapter R.7b: Endpoint specific guidance), 2017. https://doi.org/10.2823/84188.

[172] K.K. Sjøholm, H. Birch, A. Dechesne, A.P. Loibner, P. Mayer, Temperature dependency of biodegradation kinetics in environmental surface waters across Europe, in: SETAC SciCon - Poster, 2020.

[173] CONCAWE, PETRORISK version 7.0 Public: Manual, 2016.

[174] OECD - Organisation for Economic Co-operation and Development, BCF Estimation Tool v2, (2019).

[175] D.T.H.M. Sijm, M.E. Verberne, W.J. De Jonge, P. Pärt, A. Opperhuizen, Allometry in the Uptake of Hydrophobic Chemicals Determined in Vivo and in Isolated Perfused Gills, Toxicol Appl Pharmacol. 131 (1995) 130–135. https://doi.org/10.1006/taap.1995.1054.

[176] P.N.H. Wassenaar, E.M.J. Verbruggen, E. Cieraad, W.J.G.M. Peijnenburg, M.G. Vijver, Variability in fish bioconcentration factors: Influences of study design and consequences for regulation, Chemosphere. (2020). https://doi.org/10.1016/j.chemosphere.2019.124731.

[177] E.M.J. Verbruggen, M. Beek, J. Pijnenburg, T.P. Traas, Ecotoxicological environmental risk limits for total petroleum hydrocarbons on the basis of internal lipid concentrations, Environmental Toxicology and Chemistry. 27 (2008) 2436–2448. https://doi.org/10.1897/07-597.1.

[178] European Commission, Technical Guidance for Deriving Environmental Quality Standards - Guidance Document no. 27, 2018.

[179] A.P. Roberts, M.M. Alloy, J.T. Oris, Review of the photo-induced toxicity of environmental contaminants, Comparative Biochemistry and Physiology Part - C: Toxicology and Pharmacology. (2017). https://doi.org/10.1016/j.cbpc.2016.10.005.

[180] B.L. Boese, J.O. Lamberson, R.C. Swartz, R.J. Ozretich, Photoinduced toxicity of fluoranthene

to seven marine benthic crustaceans, Archives of Environmental Contamination and Toxicology. (1997). https://doi.org/10.1007/s002449900201.

[181] P.M. Allred, J.P. Giesy, Solar radiation-induced toxicity of anthracene to Daphnia pulex, Environmental Toxicology and Chemistry. (1985). https://doi.org/10.1002/etc.5620040212.

[182] J.T. Oris, J.P. Giesy, P.M. Allred, D.F. Grant, P.F. Landrum, Photoinduced toxicity of anthracene in aquatic organisms: An environmental perspective, in: Studies in Environmental Science, 1984. https://doi.org/10.1016/S0166-1116(08)72143-5.

[183] J.T. McCloskey, J.T. Oris, Effect of water temperature and dissolved oxygen concentration on the photo-induced toxicity of anthracene to juvenile bluegill sunfish (Lepomis macrochirus), Aquatic Toxicology. (1991). https://doi.org/10.1016/0166-445X(91)90069-L.

[184] R.L. Peachey, D.G. Crosby, Phototoxicity in tropical reef animals, in: Marine Environmental Research, 1996. https://doi.org/10.1016/0141-1136(95)00044-5.

[185] J.A. McGrath, C.J. Fanelli, D.M. Di Toro, T.F. Parkerton, A.D. Redman, M.L. Paumen, M. Comber, C. V. Eadsforth, K. den Haan, Re-evaluation of target lipid model–derived HC5 predictions for hydrocarbons, Environmental Toxicology and Chemistry. (2018). https://doi.org/10.1002/etc.4100.

[186] E.M.J. Verbruggen, Environmental risk limits for polycyclic aromatic hydrocarbons (PAHs) - For direct aquatic, benthic, and terrestrial toxicity (RIVM report 607711007/2012), 2012.

[187] BioByte, Bio-Loom ClogP estimates, (2006).

[188] H. Huang, B.F.J. Bowler, T.B.P. Oldenburg, S.R. Larter, The effect of biodegradation on polycyclic aromatic hydrocarbons in reservoired oils from the Liaohe basin, NE China, in: Organic Geochemistry, 2004. https://doi.org/10.1016/j.orggeochem.2004.05.009.

[189] J.M. Bayona, J. Albaigés, A.M. Solanas, R. Pares, P. Garrigues, M. Ewald, Selective Aerobic Degradation of Methyl-Substituted Polycyclic Aromatic Hydrocarbons in Petroleum by Pure Microbial Cultures†, International Journal of Environmental Analytical Chemistry. (1986). https://doi.org/10.1080/03067318608076451.

[190] R.C. Prince, C.C. Walters, Chapter 11: Biodegradation of Oil Hydrocarbons and Its Implications for Source Identification, in: Oil Spill Environmental Forensics, 2006.

[191] J.D. Leblond, T. Wayne Schultz, G.S. Sayler, Observations on the preferential biodegradation of selected components of polyaromatic hydrocarbon mixtures, Chemosphere. (2001). https://doi.org/10.1016/S0045-6535(00)00161-2.

[192] G.S. Douglas, A. Adward Bence, R.C. Prince, S.J. McMillen, E.L. Butler, Environmental stability of selected petroleum hydrocarbon source and weathering ratios, Environmental Science and Technology. (1996). https://doi.org/10.1021/es950751e.

[193] R.C. Prince, R.M. Garrett, R.E. Bare, M.J. Grossman, T. Townsend, J.M. Suflita, K. Lee, E.H. Owens, G.A. Sergy, J.F. Braddock, J.E. Lindstrom, R.R. Lessard, The roles of photooxidation and biodegradation in long-term weathering of crude and heavy fuel oils, Spill Science and Technology Bulletin. (2003). https://doi.org/10.1016/S1353-2561(03)00017-3.

[194] K.H. Wammer, C.A. Peters, Polycyclic aromatic hydrocarbon biodegradation rates: A structure-based study, Environmental Science and Technology. (2005). https://doi.org/10.1021/es048939y.

[195] P. Tuarze, Alkylated polycyclic aromatic hydrocarbons in fish and risk to human health: Implications for ecological and human health risk assessements in the Alberta oil sands, 2015.

[196] H.I. Abdel-Shafy, M.S.M. Mansour, A review on polycyclic aromatic hydrocarbons: Source, environmental impact, effect on human health and remediation, Egyptian Journal of Petroleum. (2016). https://doi.org/10.1016/j.ejpe.2015.03.011.

[197] D. Ghosal, S. Ghosh, T.K. Dutta, Y. Ahn, Current state of knowledge in microbial degradation of polycyclic aromatic hydrocarbons (PAHs): A review, Frontiers in Microbiology. (2016). https://doi.org/10.3389/fmicb.2016.01369.

[198] B. González-Gaya, A. Martínez-Varela, M. Vila-Costa, P. Casal, E. Cerro-Gálvez, N.

Berrojalbiz, D. Lundin, M. Vidal, C. Mompeán, A. Bode, B. Jiménez, J. Dachs, Biodegradation as an important sink of aromatic hydrocarbons in the oceans, Nature Geoscience. (2019). https://doi.org/10.1038/s41561-018-0285-3.

[199] M.A. Khairy, M.P. Weinstein, R. Lohmann, Trophodynamic behavior of hydrophobic organic contaminants in the aquatic food web of a tidal river, Environmental Science and Technology. 48 (2014) 12533–12542. https://doi.org/10.1021/es502886n.

[200] I. Takeuchi, N. Miyoshi, K. Mizukawa, H. Takada, T. Ikemoto, K. Omori, K. Tsuchiya, Biomagnification profiles of polycyclic aromatic hydrocarbons, alkylphenols and polychlorinated biphenyls in Tokyo Bay elucidated by δ13C and δ15N isotope ratios as guides to trophic web structure, Marine Pollution Bulletin. 58 (2009) 663–671. https://doi.org/10.1016/j.marpolbul.2008.12.022.

[201] H.R. Harvey, K.A. Taylor, H. V. Pie, C.L. Mitchelmore, Polycyclic aromatic and aliphatic hydrocarbons in Chukchi Sea biota and sediments and their toxicological response in the Arctic cod, Boreogadus saida, Deep-Sea Research Part II: Topical Studies in Oceanography. 102 (2014) 32–55. https://doi.org/10.1016/j.dsr2.2013.07.013.

[202] K.L. Foster, G.A. Stern, J. Carrie, J.N.L. Bailey, P.M. Outridge, H. Sanei, R.W. Macdonald, Spatial, temporal, and source variations of hydrocarbons in marine sediments from Baffin Bay, Eastern Canadian Arctic, Science of the Total Environment. 506–507 (2015) 430–443. https://doi.org/10.1016/j.scitotenv.2014.11.002.

[203] A.W. Kim, C.H. Vane, V.L. Moss-Hayes, D.J. Beriro, C.P. Nathanail, F.M. Fordyce, P.A. Everett, Polycyclic aromatic hydrocarbons (PAHs) and polychlorinated biphenyls (PCBs) in urban soils of Glasgow, UK, Earth and Environmental Science Transactions of the Royal Society of Edinburgh. 108 (2017) 231–247. https://doi.org/10.1017/S1755691018000324.

[204] C.H. Vane, A.W. Kim, D.J. Beriro, M.R. Cave, K. Knights, V. Moss-Hayes, P.C. Nathanail, Polycyclic aromatic hydrocarbons (PAH) and polychlorinated biphenyls (PCB) in urban soils of Greater London, UK, Applied Geochemistry. 51 (2014) 303–314. https://doi.org/10.1016/j.apgeochem.2014.09.013.

[205] G.R. Southworth, J.J. Beauchamp, P.K. Schmieder, Bioaccumulation potential of polycyclic aromatic hydrocarbons in Daphnia pulex, Water Research. (1978). https://doi.org/10.1016/0043-1354(78)90080-5.

[206] S.D. Dimitrov, I.A. Dermen, N.H. Dimitrova, K.G. Vasilev, T.W. Schultz, O.G. Mekenyan, Mechanistic relationship between biodegradation and bioaccumulation. Practical outcomes, Regulatory Toxicology and Pharmacology. (2019). https://doi.org/10.1016/j.yrtph.2019.104411.

[207] E.A.J. Bleeker, E.M.J. Verbruggen, Bioaccumulation of polycyclic aromatic hydrocarbons in aquatic organisms - RIVM report 601779002/2009, 2009.

[208] S. Koenig, P. Fernández, M. Solé, Differences in cytochrome P450 enzyme activities between fish and crustacea: Relationship with the bioaccumulation patterns of polychlorobiphenyls (PCBs), Aquatic Toxicology. (2012). https://doi.org/10.1016/j.aquatox.2011.10.016.

[209] E. Nfon, I.T. Cousins, D. Broman, Biomagnification of organic pollutants in benthic and pelagic marine food chains from the Baltic Sea, Science of the Total Environment. 397 (2008) 190–204. https://doi.org/10.1016/j.scitotenv.2008.02.029.

[210] I. Waszak, K. Jonko-Sobuś, A. Ożarowska, G. Zaniewicz, Estimation of native and alkylated polycyclic aromatic hydrocarbons (PAHs) in seabirds from the south coast of the Baltic Sea, Environmental Science and Pollution Research. (2020). https://doi.org/10.1007/s11356-020-10653-y.

[211] F.A. De Lima Ribeiro, M.M.C. Ferreira, QSAR model of the phototoxicity of polycyclic aromatic hydrocarbons, Journal of Molecular Structure: THEOCHEM. (2005). https://doi.org/10.1016/j.theochem.2005.01.026.

[212] B.L. Boese, J.O. Lamberson, R.C. Swartz, R. Ozretich, F. Cole, Photoinduced toxicity of PAHs and alkylated PAHs to a marine infaunal amphipod (Rhepoxynius abronius),

Archives of Environmental Contamination and Toxicology. (1998). https://doi.org/10.1007/s002449900311.

[213] ECHA - European Chemicals Agency, C&L Inventory, (2021). https://echa.europa.eu/information-on-chemicals/cl-inventory-database (accessed January 20, 2021).

[214] D.L. Elmendorf, C.E. Haith, G.S. Douglas, R.C. Prince, Relative rates of biodegradation of substituted polycyclic aromatic hydrocarbons, Bioremediation Chlorinated Polycyclic Aromat. Hydrocarbon Compd. (1994).

[215] G. Jonsson, R.K. Bechmann, S.D. Bamber, T. Baussant, Bioconcentration, biotransformation, and elimination of polycyclic aromatic hydrocarbons in sheepshead minnows (Cyprinodon variegatus) exposed to contaminated seawater, Environmental Toxicology and Chemistry. (2004). https://doi.org/10.1897/03-173.

[216] NITE - National Institute of Technology and Evaluation, Biodegradation and Bioconcentration data under CSCL, (2020). https://www.nite.go.jp/en/chem/kasinn/cscl_data.html (accessed March 2, 2020).

[217] H. Lin, G.D. Morandi, R.S. Brown, V. Snieckus, T. Rantanen, K.B. Jørgensen, P. V. Hodson, Quantitative structure-activity relationships for chronic toxicity of alkyl-chrysenes and alkyl-benz[a]anthracenes to Japanese medaka embryos (Oryzias latipes), Aquatic Toxicology. (2015). https://doi.org/10.1016/j.aquatox.2014.11.027.

[218] W.H. Shin, X. Zhu, M.G. Bures, D. Kihara, Three-dimensional compound comparison methods and their application in drug discovery, Molecules. 20 (2015) 12841–12862. https://doi.org/10.3390/molecules200712841.

[219] A. Nicholls, G.B. McGaughey, R.P. Sheridan, A.C. Good, G. Warren, M. Mathieu, S.W. Muchmore, S.P. Brown, J.A. Grant, J.A. Haigh, N. Nevins, A.N. Jain, B. Kelley, Molecular Shape and Medicinal Chemistry: A Perspective, Journal of Medicinal Chemistry. 53 (2010) 3862–3886. https://doi.org/10.1021/jm900818s.

[220] P.W. Finn, G.M. Morris, Shape-based similarity searching in chemical databases, Wiley Interdisciplinary Reviews: Computational Molecular Science. 3 (2013) 226–241. https://doi.org/10.1002/wcms.1128.

[221] M. Cruz-Monteagudo, J.L. Medina-Franco, Y. Pérez-Castillo, O. Nicolotti, M.N.D.S. Cordeiro, F. Borges, Activity cliffs in drug discovery: Dr Jekyll or Mr Hyde?, Drug Discovery Today. 19 (2014) 1069–1080. https://doi.org/10.1016/j.drudis.2014.02.003.

[222] M.D.M. AbdulHameed, R. Liu, P. Schyman, D. Sachs, Z. Xu, V. Desai, A. Wallqvist, ToxProfiler: Toxicity-target profiler based on chemical similarity, Computational Toxicology. 18 (2021). https://doi.org/10.1016/j.comtox.2021.100162.

[223] M.P. Seddon, D.A. Cosgrove, M.J. Packer, V.J. Gillet, Alignment-Free Molecular Shape Comparison Using Spectral Geometry: The Framework, Journal of Chemical Information and Modeling. 59 (2019) 98–116. https://doi.org/10.1021/acs.jcim.8b00676.

[224] B. Aas, Thesis MSc (UU/RIVM) - Chemical Similarity Screening With Machine Learning and Active Learning Using Physical Chemical Properties, 2020.

[225] J. Baars, Thesis MSc (UvA/RIVM) - The effect of added chemical attributes on the classification of graph based molecular datasets, 2020.

[226] P. Hartog, Thesis MSc (LU/RIVM) - Predicting and Prioritizing Possible Endocrine Disruptors using Machine Learning, 2021.

[227] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence. 1 (2019) 206–215. https://doi.org/10.1038/s42256-019-0048-x.

[228] A.S. Rifaioglu, H. Atas, M.J. Martin, R. Cetin-Atalay, V. Atalay, T. Dogan, Recent applications of deep learning and machine intelligence on in silico drug discovery: Methods, tools and databases, Briefings in Bioinformatics. 20 (2019) 1878–1912. https://doi.org/10.1093/bib/bby061.

[229] RIVM - National Institute for Public Health and the Environment, Working method Substance Advice ZZS for licensing authorities. Version 2 [In Dutch], (2020). https://rvs.rivm.nl/sites/default/files/2020-04/Werkwijze%20NL%20Stofadviezen%20v2.0%20definitief_2.pdf (accessed September 14, 2021).

[230] R. van Herwijnen, Beoordeling van de gegevens ten behoeve van de ZZS status van 28 stoffen voor DCMR. RIVM. Rapport nr. 14529C00 [In Dutch], 2021.

[231] P. van Breemen, J.P.A. Lijzen, E. Rorije, C.E. Smit, E.M.J. Verbruggen, PREMISS Toxicological modules (in prep.), n.d.

[232] Ministry of Infrastructure and Water Management, Letter to the parliament about actions on environmental safety and risks. IENW/BSK-2020/95432 [In Dutch], 2020. https://www.rijksoverheid.nl/documenten/kamerstukken/2020/06/03/omgevingsveiligheid-en-milieurisico-s (accessed August 30, 2021).

[233] H.M. Braakhuis, P.T. Theunissen, W. Slob, E. Rorije, A.H. Piersma, Testing developmental toxicity in a second species: are the differences due to species or replication error?, Regulatory Toxicology and Pharmacology. 107 (2019). https://doi.org/10.1016/j.yrtph.2019.104410.

[234] H.-J. Klimisch, M. Andreae, U. Tillmann, A Systematic Approach for Evaluating the Quality of Experimental Toxicological and Ecotoxicological Data, Regulatory Toxicology and Pharmacology. 25 (1997) 1–5.

[235] C.T.A. Moermond, A. Beasley, R. Breton, M. Junghans, R. Laskowski, K. Solomon, H. Zahner, Assessing the reliability of ecotoxicological studies: An overview of current needs and approaches, Integrated Environmental Assessment and Management. 13 (2017) 640–651. https://doi.org/10.1002/ieam.1870.

[236] N.P. du Sert, V. Hurst, A. Ahluwalia, S. Alam, M.T. Avey, M. Baker, W.J. Browne, A. Clark, I.C. Cuthill, U. Dirnagl, M. Emerson, P. Garner, S.T. Holgate, D.W. Howells, N.A. Karp, S.E. Lazic, K. Lidster, C.J. MacCallum, M. Macleod, E.J. Pearl, O.H. Petersen, F. Rawle, P. Reynolds, K. Rooney, E.S. Sena, S.D. Silberberg, T. Steckler, H. Würbel, The arrive guidelines 2.0: Updated guidelines for reporting animal research, PLoS Biology. 18 (2020). https://doi.org/10.1371/journal.pbio.3000410.

[237] C.T.A. Moermond, R. Kase, M. Korkaric, M. Ågerstrand, CRED: Criteria for reporting and evaluating ecotoxicity data, Environmental Toxicology and Chemistry. 35 (2016) 1297–1309. https://doi.org/10.1002/etc.3259.

[238] A. Mangold-Döring, C. Grimard, D. Green, S. Petersen, J.W. Nichols, N. Hogan, L. Weber, H. Hollert, M. Hecker, M. Brinkmann, A Novel Multispecies Toxicokinetic Modeling Approach in Support of Chemical Risk Assessment, Environmental Science and Technology. 55 (2021) 9109–9118. https://doi.org/10.1021/acs.est.1c02055.

[239] M. Ågerstrand, A. Beronius, Weight of evidence evaluation and systematic review in EU chemical risk assessment: Foundation is laid but guidance is needed, Environment International. 92–93 (2016) 590–596. https://doi.org/10.1016/j.envint.2015.10.008.

[240] G. Suter, J. Nichols, E. Lavoie, S. Cormier, Systematic Review and Weight of Evidence Are Integral to Ecological and Human Health Assessments: They Need an Integrated Framework, Integrated Environmental Assessment and Management. 16 (2020) 718–728. https://doi.org/10.1002/ieam.4271.

[241] European Commission, Findings of the Fitness Check of the most relevant chemicals legislation (excluding REACH) and identified challenges, gaps and weaknesses, 2019.

[242] S.E. Escher, H. Kamp, S.H. Bennekou, A. Bitsch, C. Fisher, R. Graepel, J.G. Hengstler, M. Herzler, D. Knight, M. Leist, U. Norinder, G. Ouédraogo, M. Pastor, S. Stuard, A. White, B. Zdrazil, B. van de Water, D. Kroese, Towards grouping concepts based on new approach methodologies in chemical hazard assessment: the read-across approach of the EU-ToxRisk project, Archives of Toxicology. 93 (2019) 3643–3667. https://doi.org/10.1007/s00204-019-02591-7.

[243] ECHA - European Chemicals Agency, The Petroleum and Coal stream Substances (PetCo)

Working Group, (2021). https://echa.europa.eu/petco-working-group (accessed August 30, 2021).

[244] CEFIC LRI, ECO42: Fate-Directed Toxicity Testing and Risk Assessment of UVCBs, (2021). http://cefic-lri.org/projects/eco42-fate-directed-toxicity-testing-and-risk-assessment-of-uvcbs/ (accessed August 30, 2021).

[245] U.G. Sauer, R.A. Barter, R.A. Becker, E. Benfenati, E. Berggren, B. Hubesch, H.M. Hollnagel, K. Inawaka, A.M. Keene, P. Mayer, K. Plotzke, R. Skoglund, O. Albert, 21st Century Approaches for Evaluating Exposures, Biological Activity, and Risks of Complex Substances: Workshop highlights, Regulatory Toxicology and Pharmacology. 111 (2020). https://doi.org/10.1016/j.yrtph.2020.104583.

[246] ECHA - European Chemicals Agency, Perfluoroalkyl chemicals (PFAS), (2021). https://echa.europa.eu/hot-topics/perfluoroalkyl-chemicals-pfas (accessed August 30, 2021).

[247] K. Savolainen, U. Backman, D. Brouwer, B. Fadeel, T. Fernandes, T. Kuhlbusch, R. Landsiedel, I. Lynch, L. Pylkkänen, Nanosafety in Europe 2015-2025: Towards Safe and Sustainable Nanomaterials and Nanotechnology Innovations, 2013.

[248] M. Herzler, P. Marx-Stoelting, R. Pirow, C. Riebeling, A. Luch, T. Tralau, T. Schwerdtle, A. Hensel, The "EU chemicals strategy for sustainability" questions regulatory toxicology as we know it: is it all rooted in sound scientific evidence?, Archives of Toxicology. 95 (2021) 2589–2601. https://doi.org/10.1007/s00204-021-03091-3.

[249] Safe-by-Design NL, Safe-by-Design, (2021). https://www.safe-by-design-nl.nl/home+english/default.aspx (accessed September 14, 2021).

[250] NWO - Dutch Research Council, Research data management, (2020). https://www.nwo.nl/en/research-data-management (accessed August 30, 2021).

[251] European Commission, Horizon 2020 Online Manual: Data Management, (2020). https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm (accessed August 30, 2021).

[252] R.J. Kavlock, T. Bahadori, T.S. Barton-Maclaren, M.R. Gwinn, M. Rasenberg, R.S. Thomas, Accelerating the Pace of Chemical Risk Assessment, Chemical Research in Toxicology. 31 (2018) 287–290. https://doi.org/10.1021/acs.chemrestox.7b00339.

Summary

At the moment, over 350.000 chemicals are registered worldwide for production and use. Although the use of chemicals provides numerous benefits and our daily life is drowned with it, their application may harm human health and the environment. To manage the safety of chemicals, particular chemical legislations are in place in numerous countries and regions around the world, and make use of risk and hazard assessments. However, there are several challenges for current risk and hazard assessments, including i) a lack of (reliable) data, and ii) a relative slow and inefficient evaluation and regulation process. In this thesis, I investigate specifically whether more extensive and targeted use of chemical similarity within risk and hazard assessment has the potential to improve these aspects. Chemical similarity could be a valuable factor as similarities between two chemicals could be a sign of similar physicochemical and/or toxic properties (i.e. the similar property principle). The separate sections within this thesis specifically focus on chemical similarity in relation to screening, data generation and evaluation of substances, with specific emphasis on Substances of Very High Concern (SVHC). This includes substances with carcinogenic (C), mutagenic (M) or reprotoxic (R) properties, substances with persistent, bioaccumulative and toxic (PBT) or very persistent and very bioaccumulative (vPvB) properties, or substances with an equivalent level of concern, like endocrine disrupting (ED) substances.

In chapter 2, we have developed a methodology that predicts whether a chemical is a potential SVHC based on structural similarity to chemicals already identified as SVHC. We developed a dataset of 546 substances with identified CMR, PBT/vPvB and/or ED properties from a Dutch list of SVHCs, and 411 substances that lack these hazardous properties. Next, we systematically analyzed the performance of 112 similarity measures for predicting general SVHC properties as well as for three various SVHC-subgroups separately (i.e. CMR, PBT/vPvB and ED specific similarity models). The evaluated similarity measures were unique combinations of 16 binary fingerprints and 7 similarity coefficients. These fingerprint-coefficient combinations were used to classify the substances in the dataset as (potential) SVHC or non-SVHC based on structural similarity to an SVHC using an optimal similarity threshold value. The subgroup specific models showed to outperform one specific overall model. The best similarity measures showed a high predictive performance with a balanced accuracy of 80% (i.e. correct identifications) for CMR, 95% for PBT/vPvB and 99% for ED subgroups. Accordingly, this effective screening methodology consisting of chemical similarity models showed great potential for early-stage identification of potential SVHCs.

In chapter 3, we applied the similarity models as developed in chapter 2 to the broader universe of chemicals to evaluate the application performance of the models. We used a 'pseudo-external assessment' on a set of chemicals (n=60-100 for the varying SVHC-subgroups)

that were putatively assessed as SVHC or non-SVHC based upon consensus scoring using expert elicitations (n=30 experts). Expert scores were direct evaluations based on structural similarity to the most similar SVHCs according to the similarity models, and did not consider an extensive evaluation of available data. Subsequently, the expert scores were compared to the predictions of the similarity models to evaluate the performance of the models. The use of expert opinions was particularly suitable as this is exactly the intended purpose of the chemical similarity models: a quick, reproducible and automated screening tool that mimics the expert judgement that is frequently applied in various screening applications. These analyses indicated a good statistical performance for the CMR and ED models (balanced accuracies > 80%), whereas a moderate performance with a balanced accuracy of 69% was observed for the PBT/vPvB model when compared to expert opinions. For the PBT/vPvB model, particularly false positive substances were identified, indicating the necessity of outcome interpretation. In addition, within chapter 3, model predictions were analyzed via qualitative approaches and discussed via specific examples to identify the model's strengths and limitations. The models proved to be effective in identifying groups of substances of potential concern, to be valuable for dissimilarity screening in safe-by-design trajectories, and to provide clear follow-up directions for substances of potential concern.

In chapter 4, the developed similarity models were further optimized based on the results and conclusions from chapters 2 and 3. We specifically improved the models by i) separating known SVHCs in more specific subgroups (i.e. CM, R, PBT/vPvB, ED and Other), ii) (re-)optimizing similarity models for the various SVHC-subgroups, and iii) improving interpretability of the predicted outcomes by providing a confidence score. The improvements were directly incorporated in a freely accessible web-based tool, named the ZZS similarity tool: https://rvszoeksysteem.rivm.nl/ZzsSimilarityTool. Accordingly, this tool can be used by risk assessors, academia and industrial partners to screen and prioritize chemicals for further action and evaluation within varying frameworks, and could support the identification of tomorrow's substances of concern.

In chapter 5, I covered another aspect of similarity by specifically focusing on (biological) variability and uncertainty. Generally, two substances can be considered as similar when they are structurally similar (i.e. subject in chapters 2-4) and biologically similar (i.e. they have the same or a predictable trend in biological activity). However, in order to conclude that substances are biologically similar or dissimilar, we need to know the individual variation in biological activity. To gain additional insight in the potential impact of variability on similarity assessments, we analyzed and evaluated the variation in fish bioconcentration factors (BCF) for single substances. We developed a new database consisting of BCF values for single substances to investigate the relationship between BCF values and several test characteristics. We observed that BCF values for single substances could be significantly influenced by

i) simultaneous exposure to multiple chemicals, ii) the used BCF calculation methodology, and iii) the used type of test species. Furthermore, we observed a high variation in BCF values for single substances, even when we corrected for the impact of the abovementioned test characteristics. The results Indicate that a 95% confidence range for a BCF value of 2500 could range from 953 ('not-bioaccumulative') to 6561 ('very bioaccumulative'). The remaining (unexplained) variation is likely related to variation in laboratory practices and biological variation, and questions the robustness of a single BCF value. This chapter shows that the use of one single BCF value leads to high uncertainty in bioaccumulation assessments, and indicates that (biological) similarity between two substances cannot be evaluated independently from the potential high variability in derived effects of individual substances.

In chapter 6, we investigated the use of chemical similarity for evaluation purposes. We specifically assessed the PBT/vPvB properties of a group of 884 alkylated three-ring polycyclic aromatic hydrocarbons (PAHs) to which also the parent three-ring PAHs anthracene and phenanthrene belong to. For this evaluation the hydrocarbon block method was used. Within this method, groups of constituents with similar physicochemical properties and structure are treated as a single entity and are expected to have comparable environmental fate and hazard properties. We collected experimental data and model predictions for constituents that belong to this group, and specifically focused on properties that are relevant for the PBT-assessment. Subsequently, relative trend analyses were applied with this data, to investigate the PBT-properties of the various chemicals in this group. The results consistently indicate that alkylated three-ring PAHs are more persistent, bioaccumulative, and toxic than the parent three-ring PAHs. As the parent three-ring PAHs are currently identified within Europe as PBT/vPvB substances, the alkylated three-ring PAHs could also be considered as PBT/vPvB. Accordingly, chapter 6 illustrates that chemical similarity could serve as a basis for the evaluation of PBT-properties, despite the potential presence of variability in biological activity (chapter 5). Hence, this work confirms the underlying read-across hypothesis (for which chemical similarity forms the fundamental basis), and thereby also confirms the viability and validity of chemical similarity as a screening and prioritization feature for further chemical evaluation and regulation (chapters 2-4).

The results of this thesis indicate that chemical similarity could be used to identify and evaluate SVHC-properties of single and groups of chemicals. Accordingly, I promote more extensive use of chemical similarity within risk and hazard assessment as it has the ability to circumvent several issues related to a lack of data and evaluation efficiency. By further investing in the use of chemical similarity, in combination with other innovations, we will be able to make the transition to a more efficient and effective chemical regulatory system. And as such, the use of chemical similarity allows us to consciously diminish unforeseen and unintended impacts

of any of the substances that are part of the chemical universe, and thereby enables us to strengthen the protection of human health and the environment.

Samenvatting

Wereldwijd zijn er meer dan 350.000 chemicaliën geregistreerd voor productie en gebruik. Alhoewel het gebruik van chemicaliën tal van voordelen biedt en onmisbaar is in ons dagelijks leven, kan het gebruik ervan schadelijk zijn voor de volksgezondheid en het milieu. Om de veiligheid van chemische stoffen te beheren, is er specifieke stoffen wetgeving van kracht in tal van landen en regio's over de hele wereld, waarin gebruik wordt gemaakt van risico- en gevarenbeoordelingen. Er zijn echter verschillende uitdagingen voor de huidige risico- en gevarenbeoordelingen welke veroorzaakt worden door onder andere i) een gebrek aan (betrouwbare) gegevens, en ii) een relatief traag en inefficiënt evaluatie- en regulatieproces. In dit proefschrift onderzoek ik of het gebruik van chemische gelijkenis ('chemical similarity') binnen risico- en gevarenbeoordeling de potentie heeft om deze aspecten te verbeteren. Chemische gelijkenis zegt iets over de overlap in structuur tussen stoffen en kan een waardevol aspect zijn, aangezien overeenkomsten tussen twee chemicaliën een teken kunnen zijn van vergelijkbare fysisch-chemische en/of toxische eigenschappen (dit is ook wel bekend als het 'similar property principle'). De afzonderlijke secties in dit proefschrift richten zich op chemische gelijkenis in relatie tot screening, datageneratie en evaluatie van stoffen, met een specifieke nadruk op Zeer Zorgwekkende Stoffen (ZZS). Dit omvat stoffen met kankerverwekkende (C), mutagene (M) of reprotoxische (R) eigenschappen, stoffen met persistente, bioaccumulerende en toxische (PBT) of zeer persistente en zeer bioaccumulerende (vPvB) eigenschappen, of stoffen van gelijkwaardige zorg, zoals hormoonverstorende (ED) stoffen.

In hoofdstuk 2 hebben we een methodiek ontwikkeld die voorspelt of een chemische stof mogelijke ZZS-eigenschappen heeft op basis van structurele gelijkenis met reeds geïdentificeerde ZZS stoffen. Deze methodiek is ontwikkeld op basis van een dataset bestaande uit 546 stoffen met geïdentificeerde CMR-, PBT/vPvB- en/of ED-eigenschappen afkomstig van de Nederlandse ZZS-lijst, en 411 stoffen die deze gevaarlijke eigenschappen niet hebben. Vervolgens analyseerden we de prestaties van 112 verschillende gelijkenismodellen voor het voorspellen van ZZS-eigenschappen in het algemeen, als ook voor drie verschillende ZZS-categorieën afzonderlijk (d.w.z. CMR-, PBT/vPvB- en ED-specifieke gelijkenismodellen). De geëvalueerde gelijkenismodellen betroffen unieke combinaties van 16 verschillende zogenaamde 'molecular fingerprints' (moleculaire vingerafdrukken) en 7 verschillende gelijkeniscoëfficiënten. Deze vingerafdruk-coëfficiënt combinaties werden gebruikt om de stoffen in de dataset te classificeren als (mogelijke) ZZS of niet-ZZS. Deze classificatie is gebaseerd op de mate van structurele gelijkenis met een bestaande ZZS, en een bijbehorende model-specifieke optimale drempelwaarde. De modellen die geoptimaliseerd waren voor specifieke ZZS-categorieën bleken beter te presteren dan één algemeen model. De beste modellen toonde een hoge voorspellende waarde met een nauwkeurigheid van 80% (d.w.z.

correcte identificaties) voor CMR, 95% voor PBT/vPvB en 99% voor ED-subgroepen. Gezien deze prestaties blijkt dit een effectieve screeningmethodologie, welke een grote bijdrage kan leveren aan de vroege identificatie van mogelijke ZZS stoffen.

In hoofdstuk 3 hebben we de in hoofdstuk 2 ontwikkelde gelijkenismodellen toegepast op het bredere universum van chemische stoffen, om de bredere toepassing van de modellen te evalueren. In dit werk hebben we gebruik gemaakt van een 'pseudo-externe beoordeling' van een set chemicaliën (n=60-100 voor de verschillende ZZS-categorieën). Deze chemische stoffen, waarvan de daadwerkelijke eigenschappen onbekend zijn, werden geclassificeerd als ZZS of niet-ZZS op basis van een consensus beoordeling door een groep van 30 experts. De classificaties van de experts waren gebaseerd op een snelle beoordeling van de structurele gelijkenis tussen een stof met onbekende eigenschappen en de structureel meest vergelijkbare ZZS (volgens de gelijkenismodellen). In deze beoordelingen heeft dus geen evaluatie plaatsgevonden van mogelijk beschikbare (toxiciteits)studies. De beoordelingen van de experts werden vervolgens vergeleken met de beoordelingen van de gelijkenismodellen, om de model prestaties beter in kaart te brengen. Het gebruik van de expert-classificaties was bijzonder geschikt voor dit doeleinde, omdat dit precies het beoogde doel is van de ontwikkelde gelijkenismodellen: een snelle, reproduceerbare en geautomatiseerde screeningstool die het expertoordeel nabootst. De analyses toonde goede statistische prestaties voor de CMR- en ED-modellen (nauwkeurigheid > 80%) ten opzichte van de expert-classificaties, terwijl een matige prestatie met een nauwkeurigheid van 69% werd waargenomen voor het PBT/vPvB-model. De matige presentatie van het PBT/vPvB-model werd met name veroorzaakt door foutpositieve-classificaties. Dit benadrukt de noodzaak om de positieve voorspellingen van het PBT/PvB-model verder te interpreteren als vervolgstap. Daarnaast hebben we in hoofdstuk 3 de prestaties van de modellen geanalyseerd via een aantal specifieke voorbeelden om zowel de sterke als zwakkere punten van de modellen te identificeren. De modellen blijken effectief te zijn in het identificeren van groepen van mogelijk zorgwekkende stoffen, kunnen waardevol zijn in 'safe-by-design' trajecten, en geven een duidelijke richting voor vervolgonderzoek voor mogelijk zorgwekkende stoffen.

In hoofdstuk 4 zijn de ontwikkelde gelijkenismodellen verder geoptimaliseerd op basis van de resultaten en conclusies uit hoofdstuk 2 en 3. We hebben de modellen specifiek verbeterd door i) de ZZS stoffen verder op te splitsen in specifiekere categorieën (d.w.z. CM, R, PBT/vPvB, ED en Overig), ii) het (opnieuw) optimaliseren van de gelijkenismodellen voor de verschillende ZZS-categorieën, en iii) het verbeteren van de interpreteerbaarheid van de voorspelde uitkomsten door het toevoegen van een betrouwbaarheidsscore. De verbeteringen zijn direct verwerkt in een openbare tool, genaamd de 'ZZS similarity tool': https://rvszoeksysteem.rivm.nl/ZzsSimilarityTool. Door deze methodiek beschikbaar te maken in de similarity tool kan het laagdrempelig gebruikt worden door risicobeoordelaars, academici en partijen binnen

de industrie om chemicaliën te screenen en te prioriteren voor verdere actie en evaluatie in verschillende kaders. Als zodanig kan het gebruik van de similarity tool de identificatie van de zorgwekkende stoffen van morgen bevorderen.

In hoofdstuk 5 behandel ik een ander aspect rond 'similarity' (gelijkenis) door specifiek te focussen op (biologische) variabiliteit en onzekerheid. Over het algemeen kunnen twee stoffen als vergelijkbaar worden beschouwd als ze structureel vergelijkbaar zijn (zoals onderzocht in hoofdstukken 2-4) en biologisch vergelijkbaar (d.w.z. ze hebben dezelfde of een voorspelbare trend in biologische activiteit). Om echter te kunnen concluderen dat stoffen biologisch vergelijkbaar dan wel verschillend zijn, moeten we de individuele variatie in biologische activiteit kennen. Om meer inzicht te krijgen in de mogelijke impact van variabiliteit op gelijkenis beoordelingen, hebben we de variatie in vis bioconcentratie factoren (BCF) geanalyseerd en geëvalueerd voor individuele stoffen. We ontwikkelden een nieuwe database met BCF-waarden voor individuele stoffen om de relatie tussen BCF-waarden en verschillende testkenmerken te onderzoeken. De resultaten toonden aan dat BCF-waarden van een specifieke stof significant kunnen worden beïnvloed door i) gelijktijdige blootstelling aan meerdere chemicaliën, ii) de gebruikte BCF-berekeningsmethode, en iii) het gebruikte type organisme. Daarnaast zagen we een grote variatie in BCF-waarden voor individuele stoffen, zelfs wanneer we corrigeerden voor de impact van bovengenoemde testkenmerken. De resultaten illustreren dat een 95%-betrouwbaarheidsbereik voor een BCF-waarde van 2500 zou kunnen variëren van 953 ('niet-bioaccumulerend') tot 6561 ('zeer bioaccumulerend'). De resterende (onverklaarde) variatie houdt waarschijnlijk verband met uitvoeringsvariatie en biologische variatie, en zet vraagtekens bij de robuustheid van een BCF-waarde. Dit hoofdstuk laat hiermee zien dat het gebruik van één enkele BCF-waarde leidt tot grote onzekerheid bij de beoordeling van bioaccumulatie, en geeft aan dat (biologische) gelijkenis tussen twee stoffen niet onafhankelijk kan worden beoordeeld van de mogelijk grote variabiliteit in de effecten van individuele stoffen.

In hoofdstuk 6 hebben we het gebruik van chemische gelijkenis onderzocht voor evaluatiedoeleinden. In dit werk hebben we specifiek de PBT/vPvB-eigenschappen beoordeeld van een groep van 884 gealkyleerde drie-ring polycyclische aromatische koolwaterstoffen (PAKs). Tot deze groep behoren ook de 'parent' drie-ring PAKs, antraceen en fenantreen. Voor deze evaluatie hebben we de 'hydrocarbon block method' gebruikt. Binnen deze methode worden groepen constituenten met vergelijkbare fysisch-chemische eigenschappen en structuur als een enkele entiteit behandeld. Voor een dergelijke groep constituenten wordt verwacht dat ze een vergelijkbaar verspreidingspatroon en vergelijkbare gevaareigenschappen hebben. We hebben experimentele data en modelvoorspellingen verzameld voor de constituenten die tot deze groep behoren, waarbij we ons specifiek gericht hebben op eigenschappen die relevant zijn voor de PBT-beoordeling. Vervolgens

hebben we relatieve trendanalyses toegepast op deze data om de PBT-eigenschappen van de verschillende PAKs in deze groep te onderzoeken. De resultaten laten een consistent patroon zien, waarin gealkyleerde drie-ring PAKs persistenter, bioaccumulerender en toxischer zijn dan de 'parent' drie-ring PAKs. Aangezien de 'parent' drie-ring PAKs momenteel als PBT/vPvB zijn geïdentificeerd in Europa, zouden de gealkyleerde drie-ring PAKs ook als PBT/vPvB beschouwd kunnen worden. Hiermee illustreert hoofdstuk 6 dat chemische gelijkenis gebruikt kan worden als basis voor de evaluatie van PBT-eigenschappen, ondanks de mogelijke aanwezigheid van variabiliteit in biologische activiteit (hoofdstuk 5). Daarnaast bevestigt dit werk de onderliggende read-across hypothese (waarvoor chemische gelijkenis de fundamentele basis vormt), en bevestigt daarmee ook de validiteit van het gebruik van chemische gelijkenis voor screenings- en prioriteringsdoeleinde (hoofdstukken 2-4).

Dit proefschrift laat zien dat chemische gelijkenis kan worden gebruikt om ZZS-eigenschappen te identificeren en te evalueren van zowel individuele chemische stoffen als van groepen van stoffen. Daarnaast laat dit proefschrift zien dat het gebruik van chemische gelijkenis mogelijkheden biedt om verschillende problemen te omzeilen die verband houden met een gebrek aan gegevens en de efficiëntie van het evaluatieproces. Om deze redenen pleit ik voor een uitgebreider gebruik van chemische gelijkenis binnen risico- en gevarenbeoordeling. Door verder te investeren in het gebruik van chemische gelijkenis, in combinatie met andere innovaties, kunnen we de transitie maken naar een efficiënter en effectiever stoffenbeleid. Als zodanig stelt het gebruik van chemische gelijkenis ons in staat om bewust onvoorziene en onbedoelde effecten van eenieder van de stoffen die deel uitmaken van het universum van chemische stoffen te verminderen, en stelt ons daardoor in staat de bescherming van de volksgezondheid en het milieu te versterken.

Curriculum Vitae

Pim Nicolaas Hubertus Wassenaar was born on the 3rd of April in 1993 in Beverwijk, the Netherlands. After graduating with a VWO diploma from the Bonhoeffercollege in Castricum in 2011, he joined the Bachelor's program in Biomedical Sciences at the VU University in Amsterdam. He conducted his Bachelor's internship at the Institute for Environmental Studies (IVM, VU University, Amsterdam) under supervision of prof.dr.ir. J. Legler, where he studied the effects of endocrine disrupting chemicals on *in vitro* adipocyte differentiation. In 2014, he started the Master's program in Ecology with a specialization in Environmental Chemistry and Toxicology at the VU University. In this Master's program, he conducted a literature review and two Master's internship projects. During his literature review, he conducted a systematic review about the effects of early life exposure to endocrine disrupting chemicals on obesity development in rodents, under supervision of prof.dr.ir. J. Legler (IVM, VU University). His first Master's internship was conducted at the IVM under supervision of prof.dr. M.H. Lamoree. In this project, Pim investigated the toxicological effects of surface water extracts in snails (Lymnaea stagnalis) by an *in vitro* acetylcholinesterase inhibition assay and metabolomics. His second Master's internship was conducted at the Centre for Safety of Substances and Products (VSP) of the National Institute for Public Health and the Environment (RIVM, Bilthoven), under supervision of dr. F.A. van Broekhuizen, E. Rorije (RIVM) and prof.dr. P.E.G. Leonards (VU University). He developed a generic framework for the prioritization, evaluation and regulation of petroleum substances with persistent, bioaccumulative and toxic (PBT) properties under the SVHC Roadmap of REACH. After graduating *cum laude* for his Master's degree, he started working at the RIVM in 2016 as an environmental risk assessor of chemical substances. In 2018, next to his job at the RIVM, Pim started his PhD project in a part-time position at the Institute of Environmental Sciences (CML, Leiden University, Leiden), under supervision of prof.dr.ing. M.G. Vijver and prof. dr.ir. W.J.G.M. Peijnenburg. During his research, Pim completed the Postdoctoral Education in Toxicology to become a European Registered Toxicologist (ERT). As of February 2022, Pim continued to work full-time at the RIVM-VSP as scientific researcher.

List of Publications

Peer reviewed articles

Pim N.H. Wassenaar, Emiel Rorije, Martina G. Vijver & Willie J.G.M. Peijnenburg (Submitted). ZZS similarity tool: the online tool for similarity screening to identify chemicals of potential concern.

Emiel Rorije*, **Pim N.H. Wassenaar***, Jaap Slootweg, Lonneke C. van Leeuwen, Fleur A. van Broekhuizen & Leo Posthuma (2022). Characterization of ecotoxicological risks from unintentional mixture exposures calculated from European freshwater monitoring data: Forwarding prospective chemical risk management. *Science of the Total Environment*, 822, 153385.

Pim N.H. Wassenaar & Eric M.J. Verbruggen (2021). Persistence, bioaccumulation and toxicity-assessment of petroleum UVCBs: A case study on alkylated three-ring PAHs. *Chemosphere*, 276, 130113.

Pim N.H. Wassenaar, Emiel Rorije, Martina G. Vijver & Willie J.G.M. Peijnenburg (2021). Evaluating chemical similarity as a measure to identify potential substances of very high concern. *Regulatory Toxicology and Pharmacology*, 119, 104834.

Pim N.H. Wassenaar, Eric M.J. Verbruggen, Ellen Cieraad, Willie J.G.M. Peijnenburg & Martina G. Vijver (2020). Variability in fish bioconcentration factors: Influences of study design and consequences for regulation. *Chemosphere*, 239, 124731.

Pim N.H. Wassenaar, Emiel Rorije, Nicole M.H. Janssen, Willie J.G.M. Peijnenburg & Martina G. Vijver (2019). Chemical similarity to identify potential substances of very high concern – an effective screening method. *Computational Toxicology*, 12, 100110.

Cora F. Dacosta, **Pim N.H. Wassenaar**, Ivana Dencic, Ana Morao, Evelyn H.W. Heugens, Li Shen & Michiel C. Zijp (2019). Can we assess innovative bio-based chemicals in their early development stage? A comparison between early-stage and life cycle assessments. *Journal of Cleaner Production*, 230, 137-149.

Charles W.M. Bodar, Job Spijker, Johannes P.A. Lijzen, Susanne L. Waaijers-van der Loop, Richard J. Luit, Evelyn H.W. Heugens, Martien P.M. Janssen, **Pim N.H. Wassenaar** & Theo P. Traas (2018). Risk management of hazardous substances in a circular economy. *Journal of Environmental Management*, 212, 108-114.

Pim N.H. Wassenaar, Leonardo Trasande & Juliette Legler (2017). Systematic review and meta-analysis of early life exposure to bisphenol A and obesity related outcomes in rodents. *Environmental Health Perspectives*, 125(10), 106001.

Pim N.H. Wassenaar & Juliette Legler (2017). Systematic review and meta-analysis of early life exposure to di(2-ethylhexyl) phthalate and obesity related outcomes in rodents. *Chemosphere*, 188, 174-181.

Sara Tufi, **Pim N.H. Wassenaar**, Victoria Osorio, Jacob de Boer, Pim E. G. Leonards & Marja H. Lamoree (2016). Pesticide mixture toxicity in surface water extracts in snails (Lymnaea stagnalis) by an in vitro acetylcholinesterase inhibition assay and metabolomics. *Environmental Science & Technology*, 50(7), 3937-3944.

Others (institutional reports and book chapters)

Els C. Smit, **Pim N.H. Wassenaar**, L. de Boer & Nicole M.H. Janssen (2021). Onderzoek naar mogelijk zorgwekkende stoffen in Nederlands oppervlaktewater. H2O Waternetwerk-Online.

Pim N.H. Wassenaar (2021). 'Complex mixtures/UVCBs', in Cornelis A.M. van Gestel et al. (ed.) Environmental toxicology, an open online textbook. Amsterdam: Vrije Universiteit Amsterdam, Chapter 2.2.7.

Eric M.J. Verbruggen, Marino Marinković & **Pim N.H. Wassenaar** (2020). Ecotoxicological risk limits for PFOS in soil and groundwater. RIVM letter report 2020-0085.

Pim N.H. Wassenaar (2019). 'Hydrocarbons', in Cornelis A.M. van Gestel et al. (ed.) Environmental toxicology, an open online textbook. Amsterdam: Vrije Universiteit Amsterdam, Chapter 2.3.5.

Joris T.K. Quik, Johannes P.A. Lijzen, Job Spijker, Arjen van Drongelen, Esther van der Grinten, Anne Hollander, Rob de Jonge, Jeroen P.G. van Leuken Caroline T.A. Moermond, Leon R.M. de Poorter, Heike Schmitt, Bastiaan J. Venhuis, Peter L.A. van Vlaardingen, Susanne L. Waaijers-van der Loop, **Pim N.H. Wassenaar**, Michiel C. Zijp & Patrick G.P.C. Zweers (2019). Creating safe and sustainable material loops in a circular economy. RIVM letter report 2018-0173.

Patrick G.P.C. Zweers, Julie K. Verhoeven, **Pim N.H. Wassenaar**, Job Spijker, Richard J. Luit, Jurgen H. Ganzevles & Martien P.M. Janssen (2018). Advice for guidance with background document for performing a risk inventory for substances of very high concern in waste streams. RIVM letter report 2017-0168.

Eric M.J. Verbruggen, **Pim N.H. Wassenaar** & Els C. Smit (2017). Water quality standards for PFOA. RIVM letter report 2017-0044.

Pim N.H. Wassenaar, Richard J. Luit & Lonneke C. van Leeuwen (2017). Concentration limit value for substances of very high concern in waste streams. RIVM letter report 2017-0099.

Pim N.H. Wassenaar, Nicole Janssen, Leon R.M. de Poorter & Charles W.M. Bodar (2017). Substances of very high concern and the transition to a circular economy. RIVM letter report 2017-0071.

Johannes P.A. Lijzen, **Pim N.H. Wassenaar**, Els C. Smit, Connie J.A.M. Posthuma, Ellen Brand, Frank A. Swartjes, Eric M.J. Verbruggen & Ans J.F.M. Versteegh (2017). Risk limits for PFOA in soil and groundwater. RIVM letter report 2017-0092.

** These authors have contributed equally to this work and share first authorship.*

Dankwoord

Het promotietraject bestaande uit vier jaar onderzoek is een prachtige reis geweest. De tijd is voorbijgevlogen, helemaal wanneer ik me besef dat ik bijna twee jaar grotendeels vanuit huis heb gewerkt. Het resultaat is dit (naar eigen zeggen) prachtige proefschrift. Graag wil ik hier een aantal mensen in het bijzonder bedanken.

Ten eerste natuurlijk mijn promotoren, Willie Peijnenburg en Martina Vijver. Bedankt voor de kansen die ik van jullie heb gekregen, de vrijheid binnen het onderzoek en de adviezen die jullie mij gaven. Ik wil jullie bedanken voor jullie altijd snelle reactie als ik ergens tegen aan liep, jullie scherpe opmerkingen en voor het constant benadrukken van de impact en het bredere plaatje. Ook waardeer ik onze gesprekken en het vele lachen. Jullie dynamiek is fantastisch en jullie enthousiasme is zeer motiverend geweest!

Ook bedank ik graag mijn klankbordgroep (en tevens coauteurs van de verschillende artikelen in dit proefschrift) bestaande uit Charles Bodar, Ellen Cieraad, Emiel Rorije, Eric Verbruggen, Fleur van Broekhuizen, Nicole Janssen en mijn promotoren. Wij kwamen jaarlijks bij elkaar om de voortgang van mijn promotietraject te bespreken, maar ook heb ik met jullie een-op-een vele uren gediscussieerd. Ik wil jullie bedanken voor jullie kritische vragen, inzichten, enthousiasme en bijdrage aan dit proefschrift. Zonder jullie inbreng was het proefschrift niet geworden wat het nu is!

Mijn oprechte dank aan Charles Bodar en Jan Roels voor het vertrouwen en de mogelijkheden die jullie mij hebben gegeven om mijn promotie naast mijn werk bij het RIVM te doen. Dit waardeer ik enorm. Daarnaast wil ik jou, Charles, specifiek bedanken voor je betrokkenheid, je investering in mijn persoonlijke ontwikkeling, je begeleiding en adviezen. Zonder jou was ik niet waar ik nu ben.

Graag bedank ik mijn CML-collega's. In het bijzonder Tom Nederstigt, Bregje Brinkmann en Olivier Burggraaff, waarmee ik ongeveer tegelijk aan het promotie-avontuur ben begonnen, en mijn kamergenoten Di Dong en Franco Donati. Ik wil jullie bedanken voor de gezelligheid, de vele leuke gesprekken en onze inhoudelijke discussies. Ook wil ik specifiek Sammy Koning, Susanna van den Oever en Merve Findik bedanken voor alle ondersteuning. Verder bedank ik de vele collega's van de EB en IE afdelingen, en in het bijzonder de mensen uit de Ecotox groep. Alhoewel ik maar twee dagen in de week aanwezig was op het CML voelde ik me echt thuis en onderdeel van het team, heel veel dank daarvoor!

Daarnaast bedank ik graag mijn RIVM-collega's. Specifiek de collega's van de afdeling MSP, maar ook binnen het centrum VSP en de vele mensen daarbuiten. Ik wil jullie allen bedanken

voor de fijne samenwerkingen, de gezelligheid en jullie betrokkenheid! Het voelt als een voorrecht om met zo'n grote groep gepassioneerde experts te mogen werken, en de goede sfeer heeft absoluut een positieve uitwerking gehad op dit proefschrift.

Ook wil ik graag een moment nemen om de mensen te bedanken die mij aan het begin van mijn wetenschappelijke carrière begeleid en ondersteund hebben. Speciale dank gaat uit naar Juliette Legler. Jij hebt mij geïnspireerd om voor de toxicologie te kiezen. Ook bedank ik Marja Lamoree en Pim Leonards voor jullie begeleiding en steun tijdens mijn opleiding. Jullie hebben mij altijd inhoudelijk uitgedaagd en op waarde geschat. Het noemen van mijn toenmalig medestudenten Stan de Groot, Tim Jonkers en Jeroen Meijer is hier ook op zijn plaats. Bedankt voor de gezelligheid tijdens de opleiding. Ik vind het fantastisch om te zien dat we allen nog actief zijn in dit veld en elkaar met regelmaat spreken. Daarnaast bedank ik graag nogmaals Fleur van Broekhuizen en Emiel Rorije. Jullie hebben mij begeleid tijdens mijn stage op het RIVM, waarna ik direct doorstroomde in de functie van wetenschappelijk medewerker. Tot op de dag van vandaag hebben wij nog vele gesprekken en inhoudelijke discussies waar ik veel motivatie en plezier uit haal.

Naast pure wetenschap en werk, zijn er ook tal van mensen die ik wil bedanken voor sociale aspecten. Alhoewel jullie inhoudelijke input aan dit proefschrift minder groot is, is jullie bijdrage aan mijn plezier van onschatbare waarde. Ik bedank mijn vrienden voor onze vakanties en andere activiteiten, en mijn voetbalteam voor de ontspanning en de kampioenschappen. Ik bedank graag mijn familie, Rob, Astrid, Bram en Nienke, en iedereen die daar in de loop der jaren bij is gekomen, waaronder ook mijn schoon- en stief-familie. Het is fijn om een veilige haven te hebben waar ik altijd terecht kan voor een goed gesprek en niet-werk gerelateerde zaken, of simpelweg gewoon een spelletje. Bedankt voor de gezelligheid en al jullie steun, interesse en betrokkenheid!

Tot slot sluit ik graag af met het bedanken van Margot. Dank je wel voor je onvoorwaardelijke steun, het meedenken als ik weer eens ergens mee zit en alle lol die we samen hebben. Jouw bijdrage aan dit proefschrift is van onschatbare waarde geweest waarvoor ik je ontzettend dankbaar ben!