Title

Assessing the Use of Quantitative Structure-Activity Relationship Models in Previously Unevaluated Analysis of Alternatives in the European Union Registration, Evaluation, Authorization and Restriction of Chemicals

Author

Chinen, Kazue

Publication Date

2019

UNIVERSITY OF CALIFORNIA

Los Angeles

Assessing the Use of Quantitative Structure-Activity Relationship Models in Previously Unevaluated Analysis of Alternatives in the European Union Registration, Evaluation, Authorization and Restriction of Chemicals

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Environmental Science

and Engineering

by

Kazue Kelly Chinen

2019

ABSTRACT OF THE DISSERTATION

Assessing the Use of Quantitative Structure-Activity Relationship Models in Previously Unevaluated in the European Union Registration, Evaluation, Authorization and Restriction of Chemicals Analysis of Alternatives

by

Kazue Kelly Chinen

Doctor of Environmental Science and Engineering

University of California, Los Angeles, 2019

Professor Michael K Stenstrom, Co-Chair

Professor Timothy Malloy, Co-Chair

Untested chemicals released into the market could have harmful effects on human health and the environment. Non-testing methods such as quantitative structure-activity relationship (QSAR) models may prevent these harmful consequences. However, without a meaningful evaluation of QSAR usage and proper documentation under the European Union's (EU) Registration, Evaluation, Authorization and Restriction of Chemicals (REACH) authorization process, the European Chemicals Agency (ECHA) will continue to make decisions as to whether to authorize Annex XIV chemicals are based on uncertain quality of these QSAR predictions.

The four major research questions of this study are: 1) To what extent are QSARs used in AoAs to support, eliminate or evaluate in a weight of evidence (WoE) context an alternative to an Annex XIV chemical?; 2) How did applicants document their QSAR use in AoAs?; 3) How was WoE using QSAR predictions in AoAs used to assess priority endpoints?; 4) How can battery ITS QSAR models further the evaluation of potentially harmful chemicals in AoAs? In order to conduct an analysis on QSARs, it was important to first have a firm understanding of how regulatory models work in the European Union. To become familiar with this process, two regulatory models were built at the Technical University of Denmark (DTU) for predicting agonism and antagonism of the Constitutive Androstane Receptor (CAR) for future upload to the Danish (Q)SAR Database. Applying this knowledge, data were collected from 189 AoAs through May 2017 to assess QSAR usage in AoAs, however, low numbers suggested that QSARs may not have been fully utilized. To explore possible reasons behind these statistics, an assessment of proper documentation of QSAR predictions in AoAs well as a review on the completeness of WoE using QSARs for higher-tier endpoints were performed. Results indicated that several completeness criteria were not met, including one of our priority criteria, structural analogues. In addition, only a limited number of AoAs used Woe with QSARs. A comparison of WoE using QSARs from the AoA sample with Danish EPA battery ITS QSAR predictions suggested that current use of single QSAR models continues to be limited.

The dissertation of Kazue Kelly Chinen is approved.

Jennifer A. Jay

Stephanie S. Pincetl

Joel Tickner

Timothy Malloy, Co-Chair

Michael K. Stenstrom, Co-Chair

University of California, Los Angeles

2019

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AD	Applicability domain
AhR	Arylhydrocarbon Receptor
AoA	Analysis of alternative
AOP	Adverse Outcome Pathway
ATBC	Acetyl tributyl citrate
BA	Balanced accuracies
C&L	Classification and Labelling
CAR	Constitutive androstane receptor
CAS	Chemical Abstracts Service number,
CLP	Classification, labelling and packaging of substances and mixtures
CMR	Carcinogenic, Mutagenic, Reproductive toxicity
CRS	Concentration-response series
CV	Cross-validation procedure
CYP	Cytochrome P450
Danish (EPA)	Danish Environmental Protection Agency
DBM	OASIS Database Manager
DBP	Dibutyl phthalate
DEHA	Bis(2-ethylhexyl) adipate
DIBE	Diisobutyl hexahydrophthalate
DOZ	Dioctyl azelate (
DPHP	Bis(2-propylheptyl) phthalate
DQD	Danish (Q)SAR Database
DTU Food	Technical University of Denmark National Food Institute

EC	European Community number
ECB	European Chemicals Bureau
ECHA	Environmental Chemicals Agency
ECOSAR	Ecological Structure Activity Relationships
EDC	Endocrine disrupting chemical
EFSA	European Food Safety Authority
EINECS	European Inventory of Existing Commercial Chemical Substances
ESR	Endpoint study record
EU	European Union
FN	False negative
GRADE	Grading of Recommendations Assessment, Development, and Evaluation
hAhR	Human aryl hydrocarbon receptor
hAR	Human androgen receptor
hCAR	Human constitutive androstane receptor
hER	Human estrogen receptor
hPXR	Human pregnane X receptor
HTS	High-throughput screening
IATA	Integrated approaches to testing and assessment
IRIS	Integrated Risk Information System, U.S. EPA
ITS	Integrated testing/information strategies
IUPAC	International Union of Pure and Applied Chemistry
LBD	Ligand binding domain
LOE	Lines of evidence
LPDM	Leadscope® Predictive Data Miner

MCC	Matthews correlations coefficient
MCDA	Multi-criteria decision analysis
MIE	Molecular initiating event
MOA	Mode of toxic action
NR	Nuclear receptor
NRC	National Resource Council
NUL	No Upper Limit threshold concentration
OECD	Organisation for Economic Co-operation and Development
PBT	Persistent, bioaccumulative and toxic
PLR	Partial logistic regression
PPV	Positive prediction value
qHTS	Quantitative high-throughput screening
(Q)SAR	Quantitative structure-activity relationship and structure-activity relationship
QMRF	QSAR Model Reporting Format
QPRF	QSAR Prediction Reporting Format
QSAR	Quantitative structure-activity relationship
RAC	Committee for Risk Assessment
REACH	Registration, Evaluation, Authorization and Restriction of Chemicals
RIVM	Netherlands National Institute for Public Health and the Environment
RSS	Robust study summaries
SD	Standard deviation
SDF	Structure data file
SEAC	Committee for Socio-Economic Analysis

SMILES	Simplified Molecular Input Line Entry System
SVHC	Substances of Very High Concern
TBC	Tributyl citrate
TCNES	Technical Committee on New and Existing Chemical Substances
TH	Thyroid hormone
TP	True positive
U.S. EPA	U.S. Environmental Protection Agency
U.S. Tox21	United States Toxicology Testing in the 21st Century
UGT	Uridine 5'- diphosphate-glucuronosyltransferase
vPvB	Very persistent and very bioaccumulative
WoE	Weight of evidence
VDR	Vitamin D receptor

Acknowledgements

I would like to thank my Co-Chairs and my committee members for their guidance and support.

I would also like to thank UCLA, the ESE department and my co-advisors, Dr. Michael Stentrom and Professor Timothy Malloy for the financial support that I have received over the course of my education. Specifically, I would like to acknowledge UCLA's Dissertation-Year-Fellowship (DYF), which allowed me to successfully complete my dissertation. In addition, I would like to thank Professor Donald Browne for giving me the opportunity to be a Teaching Assistant for his Engineering Ethics course in the last few months of my program.

I am grateful for my research assistant, Marian Chen, who helped to analyze data and develop the table of QSAR sources, which is presented as an appendix in Chapter 2 of this dissertation.

In addition, I would like to thank Eva Bay Wedebye, Senior Officer and my acting supervisor at the National Food Institute, Technical University of Denmark, Lyngby, Denmark, Nikolai Georgiev Nikolov, Senior Researcher at National Food Institute, Technical University of Denmark, Lyngby, Denmark, and Kyrylo Oleksandrovyc Klimenko, Postdoctoral Researcher at Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa. While I completed my program's internship component in Denmark, we had the opportunity to collaborate on the development of two QSAR models, which is presented in this dissertation. Chapter 2 is a version of our recently submitted work (Chinen K. K., Klimenko K., Nikolov, N.G., & Wedebye, E.B. (2019). QSAR modeling of different minimum potency levels for *in vitro* human CAR activation and inhibition and screening of 80,086 REACH and 54,971 U.S. substances, *Computational*

Toxicology. Approval has been obtained from Ms. Wedebye, Mr. Nikolov, and Dr. Klimenko to include this paper in my dissertation. My acting supervisor, Dr. Klimenko, Mr. Nikolov and I developed an optimized approach to QSAR-targeted processing of Tox21 data. Ms. Wedebye and Mr. Nikolov expanded findings from this approach by investigating possible statistical associations between hCAR and other endpoints.

I would also like to extend my gratitude to my Co-Advisor Professor Timothy Malloy, Professor of Law and Faculty Director of the UCLA Sustainable Technology and Policy Program. We collaborated on two policy papers, which are presented as two chapters in this dissertation. Chapter 3 is a version of a recently submitted work (Chinen, K. and Malloy, T.) (2019) QSAR Use in REACH Analyses of Alternatives to Predict Human Health and Environmental Toxicity of Alternative Chemical Substances. *Integrated Environmental Assessment and Management*. Chapter 4 is in preparation for publication (Chinen, K. and Malloy, T.) (2019) Exploring QPRF, WoE and ITS aspects of QSAR use in REACH AoAs. Approval has been obtained from Professor Malloy to include these papers in my dissertation.

I would also like to thank my ESE cohort member Lucy Lin for her support and advice throughout the years.

Vita

2009 M.S. Environmental Health Sciences, University of California, Los Angeles. Report title: Permeation of Tricresyl Phosphate Through Nitrile Glove Material.

2007 MPH Community Health Sciences, University of California, Los Angeles. Thesis title: UCLArts and Healing 2-Day Training Program for LAUSD Primary Intervention and Elementary School Counselors October 16-17, 2008.

Awards

2017 Dissertation Year Fellow

Publications

Chinen, K. & Malloy, T. (Submitted) QSAR Use in REACH Analyses of Alternatives to Predict Human Health and Environmental Toxicity of Alternative Chemical Substances.

Chinen, K., Wedebye, E. B., & Nikolov, N. G. & Klimenko, K. O. (Submitted) QSAR modeling of different minimum potency levels for *in vitro* human CAR activation and inhibition and screening of 80,086 REACH and 54,971 US substances.

Chinen, K. & Malloy, T. (In preparation) Analyzing QSARs in REACH AoAs by QPRF, WoE and ITS.

Chinen, K., Lau, S. L., Nonezyan, M., McElroy, E., Wolfe, B., Suffet, I. H., & Stenstrom, M. K. (2016) Predicting runoff induced mass loads in urban watersheds: Linking land use and pyrethroid contamination. *Water research*, 102, 607-618.

Ho, P., Chinen, K. K., Streja, L., Kreitzer, M. J., & Sierpina, V. (2011) Teaching group drumming to mental health professionals. *Explore: The Journal of Science and Healing*, 7(3), 200-202.

Professional experience during the doctoral program

2018 PhD employee at the Technical University of Denmark, Division of Diet, Disease Prevention and Toxicology. Gained proficiency in Leadscope® and Database Manager. Build two computational models that predict chemical toxicity for the Constitutive Androstane Receptor. and support policy for the Danish EPA. Presented a poster on the Development of a QSAR Model for Constitutive Androstane Receptor (CAR) Inhibition and Activation at the 18th International Conference on QSAR (Slovenia 2018)

2017 Analyst for the Government Accountability Office. Collaborated on 4 separate teams to perform regulatory audits contributing to three issued GAO reports . Assisted on a variety of project that included Boeing and SpaceX Commercial Crew Program risk assessments for NASA certification; analyzing contaminants from the Department of Defense (DoD) list of emerging contaminants that could be found in water; reviewing federal agencies' research funding, tools and methodology development and support of sustainable chemistry technologies; participating in conducting site visits and attending meetings on Capitol Hill.

1. Chapter 1: Introduction to Assessing the Use of Quantitative Structure-Activity Relationship Models in Previously Unevaluated Analyses of Alternatives in the European Union Registration, Evaluation, Authorization and Restriction of Chemicals

Every day thousands of untested industrial and synthetic chemicals put humans and wildlife at risk for long-term, serious human health and environmental effects. Traditional testing of chemicals is expensive and oftentimes employs the use of animal-testing, which is subject to both ethical and quality considerations (Akhtar 2015). Over the last decade, the search for cost-effective testing has shifted towards advanced technologies, such as quantitative structure-activity relationship (QSAR) models and human cell testing (Knudsen et al. 2013; NRC 2007; NRC 2014b). As computational models, QSARs predict the potential toxicity of untested chemicals based on the assumption that chemicals that are similar in structure have similar toxic endpoints (Malloy et al. 2017).

Under the Registration, Evaluation, Authorization and Restriction of Chemicals (REACH) authorization process, higher-tier endpoint substances known as Substances of Very High Concern (SVHC), are prioritized though limited testing data can impede their identification. According to ECHA, SVHCs are substances that have: *a*) carcinogenic, mutagenic, reprotoxic (CMR); or *b*) Persistent, bioaccumulative toxic chemical (PBT)/ very persistent or very bioaccumulative or toxic chemical (vPvB) high-tier endpoint criteria; or *c*) other toxicological properties of concern such as endocrine disrupting properties or PBT/vPvB properties, which do not fulfill the criteria of Annex XIII (ECHA 2019g). If placed on the Candidate list, SVHCs may be subject to further restriction (ECHA 2019e). Once an SVHC is included on the final Annex XIV list, however, companies must formally apply for authorization if they want continued use of their priority substance.

As part of the information requirements for authorization, applicants must submit health and environmental effects data to the Environmental Chemicals Agency (ECHA) in an analysis of alternatives (AoAs) application. Under authorization, ECHA's Committees for Risk Assessment (RAC) and Socio-Economic Analysis (SEAC) assess risks as well as the availability and feasibility of possible alternatives in the analysis of alternatives (AoA) report for Annex XIV substances (ECHA 2019a). Yet, for many untested alternative chemicals, data gaps exist (OECD 2014; Tickner and Jacobs 2016), for which QSARs can play a critical role in helping to fill in data gaps.

Despite ECHA's recommended use of non-testing methods, particularly in cases of data gaps (ECHA 2011b), there is limited insight into the extent to which AoA applicants use QSARs. Recent evaluations performed by ECHA on QSAR use have focused exclusively on registration dossiers, for which ECHA noted "poor justifications for using alternatives to vertebrate animal testing" (ECHA 2017a). However, with QSARs receiving so little attention under authorization, there is little knowledge available on the variations in quality of QSAR predictions and supporting information in REACH AoAs. This scenario is troubling, especially when QSARs generate vital and sometimes the only available data on a chemical.

Because of the advances in QSAR modeling and software platforms, quantitative structure-activity relationship (QSAR) models have been recognized for their ability to aid in certain SVHC detection (Jacobs 2004). QSARs can be used to screen both large inventories of chemicals and individual compounds to predict the toxicity of untested chemicals for a variety of endpoints. Currently, there are a limited number of *in silico* tools and QSAR models designed to predict toxicity for CMR/PBT or vPvB endpoints that modulate the constitutive androstane receptor (CAR), which is implicated in human thyroid regulation (Maglich et al. 2004; Qatanani

et al. 2005; Yan and Xie 2016), energy metabolism (Gao and Xie 2012; Wahlang et al. 2014), and fetal development (Qatanani and Moore 2005; Wyde et al. 2005). Küblbeck et al. (2008) selected CAR agonists based on ligand-binding by conducting a virtual screening study that included a preliminary screening of the Tripos LeadQuest® database of approximately 85,000 compounds to identify potentially active compounds. In a later study, Küblbeck et al. (2011) used his molecular docking technique along with other approaches to identify CAR inverse agonists. Lee et al. (2017) developed a predictive model to classify compounds as either activators or inactivators. A variety of CAR QSAR models have also captured 3-D modeling with precise predictions (Dring et al. 2010; Jyrkkärinne et al. 2008; Kato et al. 2017).

While these valuable studies predicted compounds that could either activate CAR and possibly trigger harmful metabolites, or, deactivate CAR, thereby interfering with vital metabolic actions, none of these models were developed on a robust qHTS dataset based on a wide range of concentrations tested on human cell lines. Moreover, 3-D QSAR modeling is limited by its inability to replicate ligand binding in real-time and its initial structural dataset (Jacobs 2004). Thus, having a large sample of high-quality data is important for a robust QSAR model, because it is a prerequisite for the development of a model having a large applicability domain and a high predictive accuracy.

The purpose of this paper is to understand the extent to which QSARs are being used in REACH AoAs. This research is important because no one has evaluated QSARs under authorization. In addition, all REACH QSAR evaluations have been performed at the level of registration. Thus, for chapters three and four, assessments are provided on how QSARs are used in AoAs as well as if required documentation that supports these predictions is provided. In addition, evaluations are made on the completeness of WoE using QSARs and the benefits of

ITS battery QSAR models. However, in order to assess QSARs in AoAs, a practical understanding of how QSARs work needed to be established. Chapter two therefore details the importance and development of two QSAR models for CAR at the Technical University of Denmark (DTU) based on a new data optimization approach to QSAR development for large inventory screening.

This study, which identifies QSAR model usage and target endpoint trends will allow ECHA regulators and stakeholders to gain insight into how applicants are applying QSAR guidance typically reserved for registrants. This research can help strengthen ECHA's risk assessment programs by providing additional information for regulatory decision-making, more specifically, on how how higher-tier endpoints can be better supported with QSAR predictions. Finally, implementing new guidance to support ITS battery QSAR testing is another forward-looking way to adapt to more integrated chemicals assessments. While this is the first step in providing information on QSAR usage in AoAs, this information may support ECHA's efforts in monitoring AoA applicants. Additionally, data generated from this study may facilitate the selection of safer alternatives by industry when seeking authorization, thereby reducing the number of opportunities to maintain the status quo. Finally, expanding the Danish (Q)SAR Database with more freely accessible QSARs and QSAR predictions may help to improve the quality of AoAs.

2. Chapter 2: QSAR modeling of different minimum potency levels for *in vitro* human CAR activation and inhibition and screening of 80,086 REACH and 54,971 US substances

Abstract:

Along with the Pregnane X Receptor (PXR), the Constitutive Androstane Receptor (CAR) is a key regulator of the metabolism and excretion of xenobiotics and endogenous compounds. Currently, tens-of-thousands of untested industrial compounds are released into the environment potentially exposing large parts of the population to chemical substances that have properties that may inhibit or activate crucial receptors, such as CAR. Inhibition or activation of CAR by xenobiotics can alter protein expression, leading to decreased or enhanced turnover of both xenobiotics and endogenous substances. Impacts from these alterations can potentially disturb physiological homeostasis and cause adverse effects. In the present study, the U.S. Tox21 high-throughput *in vitro* assay results for human CAR (hCAR) inhibition and activation are optimized in a comprehensive in-house process to derive training sets for different potency cut-offs and develop suites of quantitative structure-activity relationship (QSAR) models with binary outputs. Final expanded models, which include substances from the external validation sets, are developed for select minimum potency models. Rigorous cross- and external validations are used to demonstrate good predictive accuracies for the models. The final expanded models were applied to screen 80,086 European Union (EU) and 54,971 United States (U.S.) substances, and the models predicted around 60% of the substances within their respective applicability domains (AD). Finally, statistical comparisons of hCAR predictions and QSAR predictions for a number of other endpoints related to Pregnane X, aryl hydrocarbon, estrogen and androgen receptors, as well as mutagenicity, sensitization, cancer and teratogenicity from the Danish (Q)SAR database were made to investigate the possible implications of hCAR antagonists and agonists. The final

models and predictions made with these models for 650,000 substances will be made available on the free Danish (Q)SAR Database, which can aid in priority setting, read-across cases and weight-of-evidence assessments of chemicals.

Introduction

The constitutive androstane receptor (CAR) belongs to the human nuclear receptor (NR) superfamily, a 48-member group (Honkakoski et al. 2003; Maglich et al. 2001) of "orphan" and "adopted-orphan" NRs (di Masi et al. 2009; Kachaylo et al. 2011; Sonoda et al. 2008). In humans, the CAR protein is encoded by the NR1I3 gene from the NR subfamily 1, group I, member 3. The NR subfamily 1 group I also includes the Vitamin D Receptor (VDR) and the Pregnane X Receptor (PXR) (Alexander et al. 2015; Molnár et al. 2013; Wang et al. 2011). CAR is known for its 'constitutive' state. In the absence of a ligand, CAR has activity (Kretschmer and Baldwin 2005; Moore et al. 2003). Many known CAR agonists are also species-specific (Gong and Xie 2008; Qatanani and Moore 2005). CAR is expressed mainly in the liver and small intestine (di Masi et al. 2009; Honkakoski et al. 2003; Lu and Xie 2017) and mediates the induction of metabolizing enzymes, such as cytochrome P450 3A (CYP3A) isoenzymes, conjugation enzymes such as UDP glucuronosyltransferase family 1 member A1, and transporters such as P-glycoprotein (J.G. DeKeyser and C.J. Omiecinski 2010; Tabb and Blumberg 2006; Xu et al. 2005; Yan and Xie 2016). Along with the NR PXR, CAR is a principal regulator of the metabolism of xenobiotic compounds (Hakkola et al. 2018; Poso and Honkakoski 2006; Tabb and Blumberg 2006) . Both PXR and CAR cross-regulate their target genes cytochrome P450 (CYP) CYP2B and CYP3A (Francis et al. 2003) . CAR also plays an important role in the metabolism of a number of endogenous substances such as thyroid and

steroid hormones, cholesterol, bile acids, bilirubin, glucose, and lipids (Hakkola et al. 2018; Tabb and Blumberg 2006).

In some cases, the CAR upregulation of xenobiotic metabolism may lead to increased turnover of hormone and other endogenous substances, and subsequent decreased hormone levels in the body (Qatanani et al. 2005). Such interference in the regulation of endogenous hormones may have negative consequences in thyroid regulation (Miller et al. 2009), which is reflected in the adverse outcome pathway (AOP): 8 (under development) (Friedman et al. 2016). According to this AOP, activation of CAR or other NRs like PXR and the arylhydrocarbon receptor (AhR) can cause upregulated thyroid hormone (TH) catabolism, and lead to reduced TH levels, which may result in adverse neurodevelopmental outcomes in mammals (Friedman et al. 2016).

CAR is also involved in a number of other health outcomes. According to AOP: 107 (under review) (Peffer et al. 2018) CAR activation is the molecular initiating event that can lead to hepatocellular adenomas and carcinomas in the mouse and the rat. When mice were exposed to xenobiotics, CAR activation was found to be an important factor for tumor development (Huang et al. 2005; Kretschmer and Baldwin 2005; Yamamoto et al. 2004). Yet, CAR activation has been found to ameliorate diabetes (Dong et al. 2009). Alternatively, CAR inhibition may have negative consequences, namely, decreased metabolizing potential in the body, which leads to decreased turnover of endogenous hormones as well as decreased detoxification and excretion of xenobiotics (NIH 2019e). Furthermore, according to AOP: 58 (under development), CAR suppression may lead to hepatic steatosis (Angrish and Chorley 2018).

In an effort to reduce animal testing and increase the toxicity-related information level on chemical substances, Organisation for Economic Co-operation and Development (OECD) developments and European Union (EU) Registration, Evaluation, Authorization and Restriction of Chemicals (REACH) regulation have established guidance and policy to increase regulatory use of quantitative structure-activity relationship (QSAR) models (Benfenati et al. 2018; ECHA 2018; OECD 2018). QSARs are mathematical models that predict properties, (e.g. biological activities), based on chemical structure (Benfenati 2012; ECHA 2008; OECD 2007). Because QSAR predictions can be generated for large inventories of substances in a short amount of time, their use is well-suited for screening and priority setting (Rosenberg et al. 2017b). In some cases, QSAR predictions may also be used for a 1:1 replacement of experimental tests (ECHA 2008). For higher-tier health endpoints, QSAR predictions may contribute to Integrated approaches to testing and assessment (IATA) weight-of-evidence (WoE) assessments and read-across cases.

The primary objective of this study was to develop global binary QSAR models that can be used for screening purposes and single-compound identification of possible hCAR antagonists or agonists. A secondary interest in this study was to process the experimental training set data specifically for the development of QSAR models for prediction of minimum potency. We used high-throughput *in vitro* data sets from the U.S. Tox21 Program's qHTS assay for hCAR agonism and for hCAR antagonism (NIH 2019b), and the results were used to train and validate a number of QSAR models for hCAR inhibition and activation.

For priority setting purposes, the U.S. Tox21 Program applies quantitative high-throughput (qHTS) screening with the aim of identifying substances that may adversely affect human health. To date, the Tox21 chemical library holds approximately 10,000 diverse chemical substances, such as commercial chemicals, pesticides, food additives/contaminants, and medical

compounds (NTP 2019). The Tox21 CAR agonism data has previously been used as the basis for QSAR modeling (Matsuzaka and Uesawa 2019) , and others have also modeled CAR agonism using other smaller data sources (Jyrkkärinne et al. 2008; Kato et al. 2017; Lee et al. 2017).

In this study, we developed a set of criteria to process the data for our QSAR model development, including setting a minimum absolute effect, which should be observed at a maximum concentration threshold and occur at a non-cytotoxic concentration. We also filtered out luciferase inhibitors as likely false positive agonists and false negative antagonists. The processed data was ultimately used to build four final "expanded" models. For the validation procedures, all initial models underwent a DTU in-house two times 5-fold cross-validation (CV) as well as external validations with unused actives and inactives. The final models underwent the same CV procedure and external validation for specificity. These models were used to screen 80,086 structurally diverse pre-registered and/or registered substances under the EU and REACH regulation (DTU Food 2018a; DTU Food 2019), and 54,971 unique chemical structures from the U.S. Environmental Protection Agency (U.S. EPA) CoMPARA inventory (U.S. EPA 2019d). Generated QSAR predictions have the potential uses for: 1) priority setting; 2) single substance IATA WoE assessments; and 3) read-across support, (e.g. identifying quality source analogs and contributing to the hypothesis justification). To explore the possible roles of the hCAR receptor in relation to other biological activities, hCAR predictions were then statistically correlated with predictions from other QSAR models from the free online Danish (Q)SAR database (DTU Food 2019) including PXR binding/activation (Rosenberg et al. 2017b), AhR activation (Klimenko KO et al. 2019), thyroperoxidase inhibition (Rosenberg et al. 2017a), estrogen receptor (ER) activation, androgen receptor (AR) antagonism (Vinggaard et al. 2008), genotoxicity, cancer and teratogenicity.

The QSAR predictions from these screenings will be published on the free online Danish (Q)SAR Database. In addition, all final models will be published on the free, online Danish (Q)SAR models website for real-time prediction of user-submitted structures and download of detailed results in the QSAR Model Reporting Format (QMRF).

Materials and methods

Experimental datasets, definition of endpoints and developed QSAR-targeted data processing

We used results from the U.S. Tox21 Program available from the Tox21 Data Browser (NIH 2019e) and structures for the Tox21 substances from PubChem (NIH 2018). As part of the U.S. Tox21 Program, the U.S. NIH screened a total of 9,667 chemical substances for hCAR agonism and antagonism assays and for cell viability (NIH 2019c; NIH 2019d). Substances in the chemical library were not specifically selected to target hCAR agonism and/or hCAR antagonism nor were they previously suspected of affecting the hCAR receptor.

Previous publications have described the chemical structures, hCAR assays and Tox21 data analysis in more detail (Huang 2016; Lynch et al. 2018). For the cell culture, Tox21 qHTS testing used human hepatoma (HepG2) cells transfected with a double-stable human CAR and CYP2B6-2.2kb, in both agonist and antagonist mode (NIH 2019d; NIH 2019f). In addition, both assays screened 16 different response concentrations with varying concentration ranges among the different substances (NIH 2019e). Screening statistics of the agonist assay generated a Z' factor, which reflects the assay signal dynamic range and data variation associated with signal measurements, of 0.687 (Zhang et al. 1999), and a coefficient of variance close to 6.04% ± 1.56 (Lynch et al. 2018). Z' factors. Thus, an indicator of good performance is a Z' factor above 0.5 (Lynch et al. 2018; Zhang et al. 1999). In addition to these datasets, computer-readable structure data files (SDF) on the tested chemicals substances structures from PubChem: *a*) AID 1224893

on small molecule antagonists of the hCAR signaling pathway, and *b*) AID 1224892 on small molecule agonists of the hCAR signaling pathway) were used as a basis for our study. Assay results were provided by the U.S. Tox21 Program.

The US Tox21 activity profiling is primarily aimed at identifying potential mechanisms of action to prioritize substances for further in-depth toxicological evaluation. As part of the U.S. Tox21 data processing, concentration-response series (CRS), typically three per substance per assay, are fit to four-parameter Hill equations, and the outcomes are ranked into qHTS curve classes specific to this program, accounting for efficacy, p-value, asymptotes and inflection (Huang 2016). Furthermore, concentrations of half-maximal relative (AC50) rather than absolute activity (AC50) are calculated for activity and cell viability. For instance, if the maximum activity of a substance is 30% inhibition, Tox21's dose-response modeling Hill curve will give an AC50, (i.e. the concentration that causes half-maximal activity) of 15% inhibition. In the end, half-maximal AC50 values for activity and cell viability were applied to make activity outcome summary calls specific to this program, which in some cases also integrated results from additional counterscreens.

Rather than use the Tox21 summary calls (NIH 2019c; NIH 2019d), in this study, we undertook further QSAR-targeted processing of the Tox21 hCAR data by setting criteria for absolute activity for actives and setting criteria to only select the most robust inactives (Figure 2-1). Because we did not find any information from regulations, AOPs, scientific literature or other sources that would tell us which potency cut-offs we should apply to identify most relevant agonists/antagonists from a health impact perspective, or, if certain potency cut-offs formed a better basis for QSAR modeling than others, we decided to apply different potency cut-offs with a 25% effect (Vinggaard et al. 2008) occurring at or below six different thresholds: 10 µM, 20

μM, 30 μM, 40 μM 50 μM, and No Upper Limit (NUL) to construct our endpoint definitions. We then used each threshold to construct a data set. A lower cut-off equaled to a higher minimum potency, which could potentially form the basis for stronger alerts in a model, given a training set that contains a sufficient number of observations. On the other hand, NUL implied that we did not impose a concentration cut-off.

For each substance, our QSAR-targeted process led to the assignment of one of the following outcomes: “active”, “inactive”, or “inconclusive,” for which only actives and inactives were used for QSAR development and validation. For the data processing, we filtered each test CRS through in-house tools, specifically developed for the purpose of determining active responses with non-cytotoxic concentrations showing at least 25% effect (in absolute value), and accepted only the best Tox21 Hill curve classes. For inactives, we also required an initial Tox21 classification of curve class 4 (i.e. inactive) that exhibited no cytotoxicity up to 10 μM concentration (Vinggaard et al. 2008). In the end, the complete refinement procedure for each substance for both hCAR agonism and antagonism for each of the selected concentration thresholds (10 μM, 20 μM, 30 μM, 40 μM, 50 μM, and NUL) fell into five main data steps:

1. For agonism activity, only Tox21 curve classes 1.1, 2.1, 1.2 and 2.2, (i.e. all complete and incomplete curves with inflection, p-value < 0.05 and efficacy > 3 standard deviations (SD) of control) were accepted.
2. For antagonism activity, only curve classes -1.1, -1.2, -2.1 and -2.2, (i.e. all complete and incomplete curves with inflection, p-value < 0.05 and efficacy > 3SD of control) were accepted.

3. The highest concentration with non-cytotoxicity was calculated as the median for all cell viability CRSs for the analyzed substance, using the highest concentration with at least 80% viable cells for each CRS.
4. For agonist "actives", two-thirds of all hCAR activity CRS for the substance were required to fulfill the requirement of step 1 and have at least 25% effect at a non-cytotoxic concentration (as defined in "3") at or below the selected concentration threshold. For antagonist "actives", two-thirds of all hCAR activity CRS for the substance were required to fulfill the requirement of step 2 and have at least 25% effect at a non-cytotoxic concentration (as defined in "3") at or below the selected concentration threshold.
5. Substances for which all hCAR activity CRS were curve class 4 with no cytotoxicity up to at least 10 μM were assigned "inactive".

A small number of substances for both agonism and antagonism had at least 50% but not two-thirds of all hCAR activity CRS fulfilling the requirement of step 4 and were considered to be 'active' by manual expert judgment.

Because both hCAR agonism and antagonism assays were luciferase-based, we removed substances, which were luciferase inhibitors. Under certain circumstances, luciferase inhibitors may stabilize the enzyme, giving significant increases in luciferase levels in cells relative to untreated wells within the typical assay incubation time, which can lead to increases in luminescent signal by luciferase inhibitors in cell-based assays (Auld and Inglese 2018). As cell-based luciferase counterscreens for the hCAR agonism and antagonism assays were unavailable, we applied an *in chemico* luciferase inhibition screen for all Tox21 substances (NIH 2019a).

Structure curation

All data set structures underwent structure curation after the QSAR-targeted data processing (Figure 2-2). The curation was performed in OASIS Database Manager (DBM) 1.7.3 (Nikolov et al. 2006), which included additional in-house developed algorithms. First, we identified compounds with acceptable structures: Only structures exclusively containing atoms from the following list were kept: H, B, C, N, O, F, Na, Mg, Si, P, S, Cl, K, Ca, Br, and I. Records with structure errors identified by OASIS DBM were removed from the dataset. We then conducted a dissociation simulation by breaking ionic bonds and "neutralizing" the remaining structures. After this, we removed substances containing two or more organic components, (i.e. "mixtures"), and structures with less than two carbon atoms from the dataset. Furthermore, to assure that every chemical structure was only represented once in the data set, identical structures, (i.e. duplicates), were identified and removed according to the procedure described in Figure 2-2.

Training and external validation sets preparation

For both antagonism and agonism, we randomly split each of the six concentration threshold data sets into a training and a validation set (Figure 2-1). For each data set, we randomly selected 20% of the active structures for the validation set. The remaining 80% active structures were then assigned to the training set. Afterwards, we randomly selected ten times as many training set inactives to make a training set with a 1:10 distribution as this is the maximum ratio that the applied QSAR modeling software, Leadscope® Predictive Data Miner (LPDM) 3.5.3-5 can efficiently model. Any remaining inactives were used for the validation set. This meant that for external validation sets, inactives greatly out-numbered the actives. After the models were fully developed, we applied an independent external validation.

Lastly, we combined the training set and the external validation set for each of a number of selected concentration threshold to create four final 20μM and 50μM expanded models for the hCAR antagonism and hCAR agonism (Figure 2-1). An aim of the expanded models was to possibly improve model accuracy, robustness and/or applicability domain of the single training set-based models. Due to the 1:10 limitation, some negatives were randomly left out of the expanded models. These negatives were set aside to make independent external validations for specificity for the expanded models.

QSAR modeling and selection

To build all models for both antagonism and agonism, we used the commercial software Leadscope® Predictive Data Miner (LPDM), a component of Leadscope® Enterprise Server version 3.5.3-5 (Leadscope® 2019b). Structures were first imported into LPDM. Nine continuous molecular descriptors (AlogP, Hydrogen Bond Acceptors and Donors, Lipinski Score, Molecular Weight, Parent Atom Number, Parent Molecular Weight, Polar Surface Area, Rotatable Bonds) were calculated for each structure. Imported structures also underwent LPDM's systematic substructure analysis for indexing to facilitate faster data retrieval according to the 27,000 pre-defined fragment descriptors (Leadscope® 2002; Roberts et al. 2000). LPDM allows the user to generate additional training set-dependent fragment descriptors called "scaffolds," which may or may not coincide with the original library. From the entire descriptor set, which includes structural features, scaffolds and molecular descriptors, LPDM automatically selected the top 30% descriptors using the Yates X^2-test. For specific models, we applied this setting. LPDM models binary response variables using partial logistic regression (PLR). According to Valerio et al., the PLR method minimizes autocorrelation (Valerio Jr. et al. 2010),

PLR is used for a binary response variable and extracts factors by PLS using the responses as continuous data followed by logistic regression for classifications; this process is repeated until the criteria for optimum number of factors and features are reached. The binary classification model results are given as outcome probabilities from the logistic regression.

Training sets skewed towards a greater number of negatives, however, can often lead to models with a higher specificity, (i.e. true negatives), at the expense of the sensitivity, (i.e. true positives) (Valerio Jr. et al. 2010). LPDM, therefore, offers the option of building composite models, a method bearing some resemblance to bagging (Breiman 1996) though with full resampling of the smaller class and without replacement of the larger class. With this option, the modeler can set the desired ratio between the two activity classes and include up to ten sub-models with a 1:1 ratio, resampling the smaller class. In our experience, such models have close-to-equal sensitivity and specificity. Previous research has applied a "cocktail" model approach where the sub-models of composite models are aggregated with a model on the full skewed training set ("single model") (Rosenberg et al. 2017a). In earlier work, this approach has been shown to both increase specificity with only a small penalty on sensitivity compared to the composite models, as well as increase the balanced accuracy compared to either composite or single models (Klimenko KO et al. 2019).

With the purpose of selecting the specific modeling approach for further hCAR model development, we used the following approaches for all six agonism and six antagonism concentration threshold training sets in the initial model development:

1) single model, i.e. a non-composite model drawing on the full training set

2) composite model, i.e. 10 sub-models

3) composite 'cocktail' model, i.e. single model combined with the 10 composite sub-models

For all models, scaffolds were generated in LPDM from the training set structures and used along with the continuous descriptors and features. All models underwent a two times five-fold CV (i.e. removing 20% and making models on the remaining 80% structures to predict the 20% leave-out) by the LPDM algorithm. Currently, LPDM's variable selection algorithm, transfers knowledge of the selected descriptor set from the parent model when building the sub-models. Subsequently, LPDM's CV may give overly optimistic results. Thus, LPDM's CV was only used to assess the relative performance of the initial models for modeling approach selection. In the end, we selected a number of concentration-thresholds for antagonism and agonism, for which we used the expanded training set (i.e. combined initial training set and validation set for that concentration threshold) to build new, expanded models (Figure 2-1).

Applicability domain definition

We defined the applicability domain (AD) of our models as a combination of the following three components: 1) model-independent structure requirements; 2) LPMD's definition of a structural domain; 3) DTU Food's in-house definition of class probability refinement on the LPDM's output. We considered a test structure to be in AD if:

1. The test structure exclusively contained H, B, C, N, O, F, Na, Mg, Si, P, S, Cl, K, Ca, Br, and/or I, it is mono-constituent after de-salting, and it contains at least two carbon atoms.
2. The test structure met the following LPDM criteria:

a) LPDM's algorithms can calculate all molecular descriptors for the structure
b) the structure of the compound contains at least one structural feature used in the model
c) the structure of the compound has at least 30% similarity using the Jaccard (Jaccard 1901) (also known as Tanimoto) coefficient (Valerio Jr. et al. 2010) with a training set substance (based on Leadscope's built-in fragment library).

3. The test structure met the following criteria based on the positive prediction probability p between 0 and 1 calculated by LPDM as part of the prediction, with actives having a $p \geq 0.5$ and inactives having a $p < 0.5$ (Valerio Jr. et al. 2010): $p \geq 0.7$ is required for an active prediction call and a $p \leq 0.3$ for an inactive prediction call. Predictions closer to the cutoff ($p = 0.5$) are excluded, as they are likely to be less reliable.

Validation of the models

After using LPDM to guide the selection of the modeling approach, we applied DTU Food's in-house two times five-fold cross-validation procedure to measure the robustness and performance of the initial antagonism and agonism models (Figure 2-1). In this procedure, all CV sub-models were developed in isolation from the parent model as completely new models in Leadscope. Unlike LPDM's cross-validation procedure, the DTU Food's in-house CV conditions prevent any transfer of knowledge from the parent model to the CV sub-models. The five-fold approach was chosen as a robust leave-many-out cross-validation approach, and because removing any more actives might cause too large of a perturbation in the training set for the smaller active class.

To execute the in-house five-fold cross-validations, we first randomly divided the training set into five portions, each constituting 20% of the training set structures, while preserving the ratio of inactives to actives (10:1). For each of these five portions, the following steps were taken:

1) the 20% portion was removed from the full training set to create a sub-model's training set of 80%.
2) a CV sub-model was built from the reduced training set by applying the same development approach as for the parent model, but without transferring any variable selection information.
3) the 20% left-out portion was predicted by the 80% sub-model "external validation sets".

The whole procedure was performed twice, resulting in 10 CV prediction sets per threshold concentration. For all in-AD predictions in each threshold concentration's 10 prediction sets, we also calculated overall sensitivity, specificity and balanced accuracy as well as standard deviations (SD) between prediction results in the ten sub-models (Cooper II et al. 1979). In our study, we used Cooper et al.'s definitions: *a*) sensitivity is defined as the percentage of experimental actives predicted accurately; *b*) specificity is the percentage of experimental inactives predicted accurately; *c*) balanced accuracy (BA) as the average of specificity and sensitivity (Cooper II et al. 1979). To determine the percentage of substances with predictions within the AD of the DTU in-house CV models, we calculated the total coverage, i.e. "the proportion of the full set predicted within the AD of the model" of each threshold concentration's total 10 CV models (Rosenberg et al. 2017a).

To further evaluate the predictive performance of the initial DTU in-house CV threshold-concentration models for antagonism and agonism, models were subjected to an external validation using the 20% removed validation sets (Figure 2-1). Predictions, which were within AD, were then compared with the experimental results. Sensitivity, specificity, balanced accuracy and coverage were calculated for each model. Likewise, we applied the DTU Food's in-house cross-validation procedure to the expanded antagonism and agonism models for the selected concentration thresholds (Figure 2-1). Since all expanded models contained all of the actives from the initial training and validation sets, an external validation could only be performed for specificity using the unused inactives from the expanded models.

Screening large chemical inventories

To identify possible hCAR activators and inhibitors among current industrial chemicals, we applied the expanded QSAR models to two, large regulatory chemical libraries: the REACH pre-registered and/or registered substances compiled for the Danish (Q)SAR Database (DTU Food 2018a; DTU Food 2019), and a U.S. EPA substance list compiled for the U.S. EPA CoMPARA project (Mansouri et al. 2017) (Figure 2-1). Both the REACH substances and U.S. EPA set already underwent a similar structure preparation as described in the "Structure preparation" section. For our study, 80,086 QSAR-ready REACH structures, and 54,971 QSAR-ready U.S. EPA inventory structures were screened by the expanded QSAR models. As part of the predictions, we calculated the proportion of QSAR-predicted U.S. EPA and REACH-PRS substances within the AD, and of these, how many were predicted as active or inactive.

Statistical correlations of CAR predictions with QSAR predictions for other endpoints

To investigate possible associations between PXR, AhR, ER, AR, mutagenicity, sensitization, cancer and teratogenicity large-inventory prediction results, we correlated

screening results from the REACH set with Leadscope® QSAR predictions contained in the DTU-developed free online Danish (Q)SAR Database. Detailed information on Leadscope® QSAR models can be found in their QSAR Model Reporting Format (QMRF), which is freely downloadable from the Danish (Q)SAR Database (DTU FOOD et al. 2019). To gauge the strength of our correlations, specifically for predictions found in the common domain between CAR and the individual models, we calculated a Matthews correlations coefficient (MCC), a chi square (χ^2) test statistic, and sensitivity for CAR active antagonists and antagonists. Test statistics demonstrated the ability of the hCAR models at 'catching' actives from other models as well as how often hCAR models give positive predictions when the other models predict a negative outcome. Moreover, test statstics showed how good other models were at 'catching' actives from the hCAR models in addition to how often the other models give positive predictions when the hCAR models predict a negative response. For instance, when you compare actives and inactives from two different models, you can calculate sensitivity and specificity both ways, namely how good Model 1 is at 'catching' actives and inactives from Models 2, and how good Model 2 is at 'catching' actives and inactives from Model 1.

Results and discussion

In this study we developed QSAR models for hCAR antagonism and hCAR agonism for a number of different effect concentration thresholds and used final expanded models (for 25% effect at 20 μM and 50 μM) to screen 80,086 REACH substances and 54,971 U.S. EPA substances for hCAR antagonism and agonism.

The training and validation sets

We started with Tox21 experimental results for 9,667 substances for both hCAR antagonism and hCAR agonism from the Tox21 Data Browser and structures from PubChem.

After subjecting the initial data to our rigorous refinement process, each threshold concentration training set was reduced to approximately one tenth of the original dataset size. Detailed results from the QSAR-targeted data processing, structure curation, and training and validation sets can be viewed in Table 2-1.

QSAR modeling and selection

QSAR models were developed in LPDM based on all initial training sets using the three different modeling approaches. In all but one case (agonism 10 μM), the cocktail ("3") models had the highest balanced accuracies, due to increased specificities and only slightly decreased sensitivities compared to the composite ("2") models. We therefore chose to continue with the cocktail models.

We next chose the concentration thresholds to advance the modeling process and expand the training sets with substances from the validation sets. For both the antagonism and the agonism models, the differences in performance of the cocktail model across all concentration thresholds were small (Table 2-2). In other words, we did not discern some concentration thresholds as being more or less "modelable" than others. We also chose the 20 μM antagonism and agonism models as 'expansion models' for predicting higher potency substances, and the 50 μM antagonism and agonism models as 'screening models' of possible hCAR substances (Table 2-2). One reason we selected the 20 μM models for higher potency prediction as opposed to the 10 μM models, more specifically, the agonism model, was that the 10 μM agonism model had a rather small number of actives in the training set, making the training set more unstable. In addition, the 10 μM agonism model had a smaller AD compared to the 20 μM models. For the 50 μM models, we considered this threshold as still 'interpretable' for minimum potency, as

opposed to the NUL models, which indicate zero potency. More specifically, if one predicts an active with the 50 µM model, this means that the model predicts that the chemical will have minimum 25% effect at maximum 50 micro-M concentration, i.e. a minimum potency. The 20 µM and 50 µM expanded models for both hCAR antagonism and hCAR agonism were modelled using the expanded training sets presented in Table 2-1 and by the chosen cocktail approach in LPDM.

Predictive performance of all the initial QSAR models

All initial QSAR models underwent a two times five-fold DTU Food in-house cross-validation (CV) procedure as well as external validation with both the left out 20% active and all remaining inactives. Results are given in (Table 2-3). For antagonism, CV sensitivities ranged from 54.3% and 74.7%. However, with rather high SD values (i.e. 11.0% and 16.1%, for Ant-10 µM-QSAR and Ant-30 µM-QSAR, respectively), the CV sensitivities were not significantly different from each other. The high standard deviations (SD) of sensitivities for the CV results of both endpoints are likely due to the 20% removal of the relatively small sets of actives, which removes valuable information for classes not highly represented in the sets (Table 2-3). In contrast, CV specificities ranged between 92.4% and 97.2%. These high specificities had a smaller SD range of values, (i.e. 1.2% to 2.9%), which reflected the bigger inactive classes in the training sets. Balanced accuracies stayed between 75.7% and 84.0% (Table 2-3).

For the agonism models, CV sensitivity was much lower for the Ag-10 µM-QSAR, (i.e. 28.2%) and and ranged between 61.6% and 71.7% for the remaining models (Table 2-3). SD values had a higher range, (i.e. 12.2% and 32.3%) relative to the antagonism values (Table 2-3). Higher SD values reflected the rather small size of the training set active classes. For example,

the Ag-10 µM-QSAR initial model had only 26 actives compared to the Ant-10 µM-QSAR initial model with 86 actives. On the other hand, the CV-derived specificities for the initial agonism models ranged between 90.0% and 97.2%. Similar to the antagonism models, these high specificities were attributed to larger inactive classes in the training sets. SD values ranged between 2.2% for the Ag-No-Upper-Limit initial model and 6.3% for the Ag-10 µM-QSAR initial model. Balanced accuracies had a slightly lower range between 55.0% and 81.4%.

Unlike the CV sensitivities, external validation sensitivities had a wider range for both models, though specificities remained relatively high and within a narrow range. For antagonism, the external validations with the 20% leave-out actives and the remaining inactives gave sensitivities that ranged between 55.0% and 83.3% (Table 2-3). Specificities, however, were higher, (i.e. 91.7% for the and 94.2%). The range for balanced accuracy was also slightly higher (i.e. 73.8% to 88.4%) when compared to the CV balanced accuracies.

External validations for agonism with the 20% leave-out actives and the remaining inactives demonstrated a wider range of sensitivities. Sensitivities ranged between 37.5% for the Ag-30 µM-QSAR initial model and 100% for the Ag-10 µM-QSAR initial model (Table 2-3). Specificities, however, were higher than antagonism, ranging between 89.4% for the Ag-No-Upper-Limit initial model and 95.3% for the Ag-10 µM-QSAR initial model. Balanced accuracies also had a wider range. BAs ranged between 65.0% and 97.6% for both the Ag-10 µM and the Ag-30 µM-QSAR initial models, respectively.

For both antagonism and agonism, sensitivities lacked a clear trend, and varied much more than specificities. Variance was most likely due to the small number of actives in AD in the validation sets, as reflected in the true positive (TP) and false negative (FN) numbers (Table 2-

3). On the other hand, specificities were much more stable due to the high number of inactives in the validation sets. In this case, high specificities are especially desirable because the models do not generate a high percentage of false positives, which is important for endpoints such as hCAR antagonism and agonism. If substances in a given chemical universe are rarely 'positive', (i.e. the balance is very shifted towards negatives), then having a high specificity is even more important in order to not to overly "pollute" the true positives with false positives, thereby giving low positive prediction value (PPV), or low trust in positive predictions.

A possible trend for both antagonism and agonism models, however, could be the slight decrease in specificity with increasing concentration thresholds. This slight trend could possibly indicate that including actives with lower potency in the training set leads to decreasing quality of the positive alerts in the models. Because of the uncertainties of the sensitivities, comparisons between CV results and external validation results are most relevant for specificity. As shown in Table 2-3, the specificities from the external validations are close to the specificities from the CVs, with the latter being a few percent higher in some cases (SD taken into consideration).

Predictive performance of the 20 µM and 50 µM expanded QSAR models

All expanded QSAR models underwent a two times five-fold DTU Food in-house CV procedure as well as external validation for specificity using the inactives not included in the expanded training sets (Table 2-3). According to the CV results, the expanded antagonism 20 µM and 50 µM models had sensitivities of 58.4% and 72.4%, specificities of 97.1% and 82.0%, and BAs of 77.8% and 82.0%. For the expanded agonism 20 µM and 50 µM models, sensitivities were 72.2% and 78.4%, specificities were 93.5% and 91.4%, and BAs were 82.0% and 84.9%. In all cases the results for the expanded antagonism and agonism models were not dissimilar from the CV results of the corresponding initial models, taking SD into account. With the exception of

sensitivity for the 20 µM antagonism model, for both antagonism and agonism, the SDs for sensitivity and specificity were in all cases lower for the expanded models than the corresponding initial models, indicating more stability.

According to the external specificity validation results, the expanded antagonism 20 µM and 50 µM models had specificities of 93.2% and 92.0%. For the expanded agonism 20 µM and 50 µM models, external validation specificities were 91.5% and 90.6%. Notably, the specificity results for all antagonism and agonism expanded models had a difference of less than 1% from the external validation specificity results for the corresponding initial models. Although smaller than the 50 µM models, the 20 µM antagonism and agonism expanded models had slightly higher external validation specificities on these relatively large inactive validation sets. We speculated that some less discriminating positive alerts entered into the 50 µM model compared to the 20 µM models, especially for antagonism. Thus, when we included the "extra" actives for the higher thresholds, in this case, the positives in the 50uM model) these actives also included substances that are weaker than the ones in the 20uM model.

Screening results

We screened the U.S. EPA and REACH-PRS inventories using the expanded models for hCAR 20 µM and 50 µM antagonism and agonism models (Table 2-4). The 20 µM and 50 µM antagonism models had coverages between 54.3% and 63.1%. The two antagonism models predicted between 15.8% and 17.0% of the active substances within their respective ADs from the two inventories. The two agonism models predicted between 16.9% and 20.9% of the substances in AD from the two inventories. The prevalences of actives, or the percent of actives out of the total number of substances, were between 2.5% and 4%, after we conducted the QSAR-targeted data processing and depending on the concentration threshold. (Prevalences can

be derived from Table 2-1). It should be noted that the experimentally tested Tox21 library of substances was not selected based on suspicion of hCAR antagonism or agonism. At the same time, it is not known how well the library reflects the true prevalence of hCAR antagonists and agonists of the U.S. EPA and REACH inventories.

According to the rather large external validations, both the applied antagonism and the agonism models showed high specificities (90.7% - 93.2%) indicating that they do not give very many false positive predictions. Nonetheless, the models may have still have false positive rates, (which can be derived from the external validation specificity results), between 9.3% and 6.8%. These false positive rates may explain some though not all of the high percents, (i.e. 15.8% - 20.9%), of active predictions for hCAR antagonism and agonism in the two large inventories, thereby indicating a possible high number of hCAR antagonists and agonists in the U.S. EPA and REACH inventories.

Statistical correlations of CAR predictions with QSAR predictions for other endpoints

REACH screening results for both the expanded CAR 50μM antagonism and agonism models were statistically correlated with QSAR predictions within the same REACH set for PXR, AhR, ER, AR, mutagenicity, sensitization, cancer and teratogenicity endpoints. Predictions from these other QSAR models are stored in the free online Danish (Q)SAR Database (DTU FOOD et al. 2019), which is developed and maintained by the DTU co-authors. This analysis was performed to explore possible biological pathways and toxicities affected by CAR antagonists and agonists. When performing the correlations, only molecules in the common QSAR applicability domain of the relevant models were used.

Each of the two lists in Table 2-5 is sorted by the percent of hCAR positives overlapping with predictions from the respective other model (denoted by M1). As can be seen from Table 5, there are endpoints related to AhR, AR antagonism, PXR/CYP3A4 induction, cancer in rats and TPO inhibition on top of both lists for correlations with hCAR antagonism and agonism predictions. In other words, hCAR positive predictions are found at a much higher percentage among positive than among negative predictions from models for these endpoints. At the bottom of both lists are CYP2C9, CYP2D6 and skin irritation, where CAR positives are not found at very high percentages either among positives or among negatives from the other models. These low percentages may be regarded as a random overlap.

Because CAR, PXR and AhR share some ligands, we also wanted to see how well the three endpoints correlated to each other (Li et al. 2015). As expected, both CAR antagonism and agonism correlated strongly with PXR. However, more surprisingly, CAR antagonism and agonism correlated even more strongly with AhR (Table 2-6).

The tabulated counts in Table 2-6 show that 23,004 substances out of 80,086 REACH substances in the Danish (Q)SAR Database are in the common domain of the three models. Of these 23,004 substances, the vast majority were predicted negative by all three models, (i.e. 20,164 substances). hCAR had 561 positive predictions for which the other two models predicted negative (corresponding to 2% of the common domain); hPXR had 738 (3%) positive predictions for which the two other models predicted negative; and hAhR had 21 (0.1%) positive predictions for which the two other models predicted negative. However, all three models gave positive predictions for 243 substances. hCAR and hPXR shared the biggest number of common positive predictions, namely, the 243 substances, which were positive in all three models plus the 1,233

substances only positive in the hCAR and hPXR models. In addition, the AhR model rarely gave a positive prediction when the hCAR model gave a negative prediction in a combined 24 cases.

Conclusions

Our study presents two main results: 1) We developed an in-house QSAR-targeted data processing approach to extract Tox21 experimental results for QSAR development of different absolute minimum potency classes; 2) We developed, validated, and applied global, binary QSAR models for hCAR antagonism and agonism *in* vitro. All initial models were based on 10 μM, 20 μM, 30 μM, 40 μM, 50 μM and No Upper Limits threshold concentrations for both agonism and antagonism. All initial models underwent CV and external validations and showed high specificities of around 90-95% and good BAs. For both antagonism and agonism, new expanded models were developed for the 20 μM and 50 μM thresholds by incorporating the external validation set actives and ten times as many inactives into the training sets. These expanded models also underwent cross-validation, and external validation but only for specificity as there were no additional actives for a full external validation. In all but one case, the results for the expanded antagonism and agonism models were slightly better compared to the correspoinding initial models. However, results were not dissimilar from the CV results of the corresponding initial models when taking SD into account. External validations of specificity showed similar performance between initial and the corresponding expanded models.

Our four expanded models were used to screen two large chemical inventories from the U.S. and EU. Of the substances predicted within the ADs of the expanded models, the 20 μM agonism model predicted 8,265 (16.9%) REACH substances and 5,731 (17.4%) U.S. EPA substances to be positive; the 20 μM antagonism model predicted 8,058 (16.0%) REACH substances and 5,175 (15.8%) U.S. EPA substances to be positive. For antagonism, the 50 μM

expanded models predicted a slightly higher number of positives than the 20 μM expanded model.

Finally, we explored if a number of biological pathways and toxicity properties correlated statistically, (i.e. not investigated for causality) with predicted hCAR antagonists and agonists. This was done by correlating QSAR predictions from the expanded hCAR 50μM antagonism and agonism models with QSAR predictions for endpoints related to PXR, AhR, ER, AR, mutagenicity, sensitization, cancer and teratogenicity endpoints for 80,086 REACH substances contained in the Danish (Q)SAR Database.

Our study aims to be forward looking. Results from the developed QSAR-targeted data processing of Tox21 may contribute to future QSAR modeling studies. In addition, the developed hCAR antagonism and agonism QSAR models may be utilized in the future for support of screening, read-across or IATA WoE assessments. Predictions for 650,000 substances from the expanded antagonism and agonism 20 μM models have long-term implications for regulatory hazard assessments. These predictions will be published in the free online Danish (Q)SAR Database, and the models themselves will be made available in the free online Danish (Q)SAR Models website (DTU FOOD et al. 2019).

Conflict of Interest Statement

The authors declare that they have no conflict of interest in relation with this paper.

Acknowledgements

We would like to thank Informatics leader, Senior scientist Ruili Huang from the U.S. NIH and the Danish Environmental Protection Agency for their support in this project.

Tables

Table 2-1. Number of processed structures through QSAR-targeted data processing and structure curation, and resulting unique structures in the training and validation sets with the distribution of active and inactive.

	hCAR Antagonism			hCAR Agonism		
	All	Active	Inactive	All	Active	Inactive
Original data set	9667			9667		
10 μM						
After QSAR-targeted data processing	5868	160	5708	6977	108	6869
Acceptable structures	5430	136	5294	6422	106	6316
After luciferase inhibitors removal	5298	136	5162	6356	40	6316
After duplicates removal	4259	107	4152	5098	33	5065
Validation set (20% random for actives)	3313	21	3292	4812	7	4805
Training set (inactives = 10 * actives)	946	86	860	286	26	260
20 μM						
After QSAR-targeted data processing	5897	189	5708	7098	229	6869
Acceptable structures	5459	165	5294	6609	227	6382
After luciferase inhibitors removal	5327	165	5162	6492	110	6382
After duplicate removal	4277	128	4149	5147	84	5063
Validation set (20% random for actives)	3155	26	3129	4410	17	4393
Training Set (inactives = 10 * actives)	1122	102	1020	737	67	670
Expanded training set	1408	128	1280	924	84	840
Reduced validation set (only inactives)			2869			4223
30 μM						
After QSAR-targeted data processing	5920	212	5708	7126	257	6869
Acceptable structures	5481	187	5294	6636	254	6382
After luciferase inhibitors removal	5349	187	5162	6507	125	6382
After duplicate removal	4292	144	4148	5159	96	5063
Validation set (20% random for actives)	3027	29	2998	4312	19	4293
Training Set (inactives = 10 * actives)	1265	115	1150	847	77	770

(Table 2-1 continued)

	hCAR Antagonism			hCAR Agonism		
	All	Active	Inactive	All	Active	Inactive
Original data set	9667			9667		
40 μM						
After QSAR-targeted data processing	5925	217	5708	7129	260	6869
Acceptable structures	5485	191	5294	6639	257	6382
After luciferase inhibitors removal	5353	191	5162	6509	127	6382
After duplicate removal	4292	145	4147	5161	98	5063
Validation set (20% random for actives)	3016	29	2987	4303	20	4283
Training Set (inactives = 10 * actives)	1276	116	1160	858	78	780
50 μM						
After QSAR-targeted data processing	5956	248	5708	7267	398	6869
Acceptable structures	5515	221	5294	6774	392	6382
After luciferase inhibitors removal	5383	221	5162	6592	210	6382
After duplicate removal	4314	170	4144	5234	173	5061
Validation set (20% random for actives)	2818	34	2784	3716	35	3681
Training Set (inactives = 10 * actives)	1496	136	1360	1518	138	1380
Expanded training set	1870	170	1700	1903	173	1730
Reduced validation set (only inactives)			2444			3331
NUL[+]						
After QSAR-targeted data processing	5961	253	5708	7341	472	6869
Acceptable structures	5519	225	5294	6848	466	6382
After luciferase inhibitors removal	5387	225	5162	6641	259	6382
After duplicate removal	4315	172	4143	5271	212	5059
Validation set (20% random for actives)	2797	34	2763	3401	42	3359
Training Set (inactives = 10 * actives)	1518	138	1380	1870	170	1700

[+] For No Upper Limit (NUL) a concentration threshold cut-off was not set.

Table 2-2. Results from the LPDM two times five-fold cross-validation of all initial models by three approaches: 1) single model, 2) composite model, 3) composite 'cocktail' model.

Cocktail models (initial)	Approach	LPDMs two times five-fold cross validation results		
		Sensitivity (%)	Specificity (%)	Balanced accuracy (%)
Antagonism				
10 µM training set	1	54.1	99.0	76.6
	2	91.2	89.9	90.6
	3	86.6	97.4	92.0
20 µM training set	1	49.3	98.6	74
	2	87.8	88.3	88.1
	3	84.5	97.8	91.2
30 µM training set	1	52.9	98.5	75.7
	2	87.2	89.1	88.2
	3	84.9	97.2	91.1
40 µM training set	1	47.3	98.9	73.1
	2	87.8	87.3	87.6
	3	85.9	97.4	91.7
50 µM training set	1	46.0	99.0	72.5
	2	84.1	87.2	85.7
	3	83.5	97.9	90.7
No Upper Limit trainin	1	38.5	99	68.8
	2	84.6	86.3	85.5
	3	82.4	97.1	89.8

(Table 2-2. continued)

Cocktail models (initial)	Approach	LPDMs two times five-fold cross validation results		
		Sensitivity (%)	Specificity (%)	Balanced accuracy (%)
Agonism				
10 μM training set	1	23.5	99.4	61.5
	2	94.7	88.2	91.5
	3	81.3	96.8	89.1
20 μM training set	1	35.8	98.9	67.4
	2	96.6	89.0	92.8
	3	92.5	96.9	94.7
30 μM training set	1	21.1	99.5	60.3
	2	91.4	88.8	90.1
	3	87.7	98	92.9
40 μM training set	1	41.8	98.2	70.0
	2	92.2	88.2	90.2
	3	90.2	96.5	93.4
50 μM training set	1	40.4	98.6	69.5
	2	92.0	87.6	89.8
	3	90.3	97.9	94.1
No Upper Limit trainin	1	24.5	98.8	61.7
	2	91.0	87.2	89.1
	3	89.1	97.7	93.4

* TP: true positives, FP: false positives, TN: true negatives, FN: false negatives. The numbers are averages of the ten iterations as given by LPDM.

Table 2-3. Predictive performances for hCAR antagonism and agonism QSARs.

	DTU Food Cross-Validation† %							External Validation %							
Model (Cocktail Scaffolds)	Sensitivity (SD,%)	Specificity (SD,%)	Balanced accuracy (SD,%)	TP*	FP*	TN*	FN*	Sensitivity	Specificity	Balanced accuracy	Coverage	TP	FP	TN	FN
Ant-10 **μ**M-QSAR (initial)	69.9 (11.0)	93.1 (2.9)	81.5 (5.8)	75	80	1071	32	57.1	94.2	75.7	69.8	8	133	2164	6
Ant-20 **μ**M-QSAR (initial)	54.3 (11.5)	97.2 (1.2)	75.7 (5.7)	55	41	1423	46	83.3	93.5	88.4	62.1	15	126	1814	3
Ant-20 **μ**M-QSAR (expanded)	58.4 (14.2)	97.1 (0.7)	77.8 (7.2)	93	55	1877	65	-	93.4	-	61.3	-	116	1644	-
Ant-30 **μ**M-QSAR (initial)	74.7 (16.1)	93.3 (2.1)	84.0 (8.6)	101	101	1406	35	76.9	92.1	84.5	67.4	10	160	1866	3
Ant-40 **μ**M-QSAR (initial)	74.0 (12.6)	94.1 (2.4)	84.0 (5.9)	90	89	1429	32	56.3	93.5	74.9	66	9	129	1845	7
Ant-50 **μ**M-QSAR (initial)	72.4 (14.3)	92.6 (2.4)	82.5 (7.9)	112	129	1599	44	55	92.5	73.8	59.9	11	125	1544	9
Ant-50 **μ**M-QSAR (expanded)	72.4 (10.2)	91.6 (1.5)	82.0 (5.0)	130	173	1896	52	-	92.4	-	59.8	-	111	1350	-
Ant-No-Upper-Limit (initial)	70.9 (12.3)	92.4 (2.8)	81.6 (5.5)	111	129	1568	47	58.8	91.7	75.3	57.9	10	133	1469	7
Ag-10 **μ**M-QSAR (initial)	28.2 (32.3)	91.0 (6.3)	55.0 (24.5)	7	25	246	17	100	95.3	97.6	61.7	3	140	2825	0
Ag-20 **μ**M-QSAR (initial)	61.6 (22.3)	93.0 (3.2)	77.3 (12.1)	53	58	784	33	87.5	91	89.3	66.2	7	261	2652	1
Ag-20 **μ**M-QSAR (expanded)	72.2 (13.6)	93.5 (1.2)	82.8 (7.0)	72	71	1013	26	-	91.5	-	66.4	-	238	2564	-
Ag-30 **μ**M-QSAR (initial)	71.7 (17.5)	91.0 (4.8)	81.4 (8.8)	57	84	850	25	37.5	92.4	65	65	3	212	2582	5
Ag-40 **μ**M-QSAR (initial)	67.9 (14.0)	91.6 (3.2)	80.3 (7.1)	62	73	922	30	71.4	92.6	82	65.7	10	209	2604	4
Ag-50 **μ**M-QSAR (initial)	70.9 (12.2)	91.5 (3.3)	81.2 (6.6)	118	159	1708	48	77.3	90.9	84.1	67.4	17	226	2258	5
Ag-50 **μ**M-QSAR (expanded)	78.4 (10.6)	91.4 (1.9)	84.9 (5.6)	147	199	2119	42	-	90.6	-	67.7	-	211	2045	-
Agt-No-Upper-Limit (initial)	70.8 (12.5)	90.0 (2.2)	80.4 (6.0)	133	215	1953	53	84	89.4	86.7	65.2	21	232	1962	4

† A two times five-fold cross-validation.

* TP: true positives, FP: false positives, TN: true negatives, FN: false negatives. The numbers given are the totals over all ten CV sub-models.

Table 2-4. The coverage (AD) and the number of active/inactive predictions of the U.S. EPA and REACH-PRS inventories in the expanded 20 μM and 50 μM hCAR antagonism models and the expanded 20 μM and 50 μM hCAR agonism models.

Datasets	Total Predictions	Ant-20 μM-QSAR (expanded)			Ant-50 μM-QSAR (expanded)		
		In AD	Active (%)	Inactive (%)	In AD	Active (%)	Inactive (%)
REACH[†]	80,086	63.1%	8,058 (16.0)	42,441 (84.0)	57.0%	7,680 (16.8)	37,931 (83.2)
U.S. EPA[††]	54,971	59.6%	5,175 (15.8)	27,577 (84.2)	54.3%	5,062 (17.0)	24,767 (83.0)

Datasets	Total Predictions	Ag-20 μM-QSAR (expanded)			Ag-50 μM-QSAR (expanded)		
		In AD (%)	Active (%)	Inactive (%)	In AD (%)	Active (%)	Inactive (%)
REACH[†]	80,086	61.1%	8,265 (16.9)	40,631 (83.1)	63.7%	10,289 (20.2)	40,711 (79.8)
U.S. EPA[††]	54,971	59.8%	5,731 (17.4)	27,121 (82.6)	63.0%	7,254 (20.9)	27,389 (79.1)

[†] U.S. EPA QSAR-ready structures from a U.S. EPA inventory of synthetic chemical structures to which humans are potentially exposed.

[††] REACH QSAR-ready structures from the REACH pre-registered substances (PRS) list and/or REACH registrered substances.

Table 2-5. Correlations between prediction for REACH substances from the hCAR Ant-50 µM (expanded) model and hCAR Ag-50 µM (expanded) model and predictions from other DTU-developed and commercial LPDM QSAR models from the free online Danish (Q)SAR Database (only predictions in the domains of the pairwise correlated models were counted).

hCAR Ant-50 µM (expanded) Model 1 (M1)	Pos in both	Pos in M1, Neg in CAR	Neg in M1, Pos in CAR	Neg in both	MCC	Chi2	CAR positives of M1 positives (%)	CAR positives of M1 negatives (%)	M1 positives of CAR positives (%)	M1 positives of CAR negatives (%)
AhR agonism (rational)	408	84	3435	27383	0.27	2317	83	11	11	0
AhR agonism (random)	1064	340	2587	24347	0.43	5207	76	10	29	1
AR antagonism	836	275	4006	29277	0.32	3551	75	12	17	1
ER alpha agonism	731	344	2080	14135	0.36	2254	68	13	26	2
CYP3A4 induction	1247	623	2414	29655	0.44	6426	67	8	34	2
PXR agonism rat in vitro	459	314	4736	32621	0.19	1404	59	13	9	1
FDA cancer rat	2903	2207	2013	15506	0.46	4778	57	11	59	12
PXR agonism human in vitro	1391	1114	2063	22372	0.41	4507	56	8	40	5
PXR binding human in vitro (new model)	3376	2791	1676	22222	0.52	7988	55	7	67	11
TPO inhibition rat in vitro	2354	1990	1670	24713	0.49	7507	54	6	58	7
Bacterial reverse mutation test (Ames)	1646	1468	4125	29112	0.31	3488	53	12	29	5
FDA cancer rodent	2288	2113	1736	15879	0.44	4185	52	10	57	12
Teratogenic potential in humans	1042	1624	1984	13607	0.25	1144	39	13	34	11
Respiratory sensitisation in humans	2607	4127	290	9966	0.47	3701	39	3	90	29
PXR binding humans in vitro (old model)	2882	4757	1241	14890	0.37	3262	38	8	70	24
FDA cancer mouse	904	1536	4668	21357	0.13	518	37	18	16	7
Ashby structural alerts	2000	4761	3618	21600	0.16	854	30	14	36	18
Allergic contact dermatitis	2402	8589	759	17099	0.27	2161	22	4	76	33
CYP2C9 substrates in humans	158	636	3428	16488	0.01	4	20	17	4	4
CYP2D6 substrates in humans	137	791	3516	18539	-0.01	1	15	16	4	4
Severe skin irritation in rabbit	923	7288	3263	14722	-0.09	200	11	18	22	33

(Table 2-5 continued)

hCAR Ag-50 µM (expanded)	Pos in both	Pos in M1, Neg in CAR	Neg in M1, Pos in CAR	Neg in both	MCC	Chi2	CAR positives of M1 positives (%)	CAR positives of M1 negatives (%)	M1 positives of CAR positives (%)	M1 positives of CAR negatives (%)
Model 1 (M1)										
AhR agonism (rational)	671	64	4568	29793	0.31	3447	91	13	13	0
AhR agonism (random)	1487	298	3458	26897	0.46	6697	83	11	30	1
AR antagonism	1375	532	5019	30648	0.34	4317	72	14	22	2
PXR agonism human in vitro	2279	987	1390	26525	0.62	11826	70	5	62	4
CYP3A4 induction	1783	852	2793	32380	0.47	8220	68	8	39	3
FDA cancer rat	3361	1728	2551	18773	0.51	6917	66	12	57	8
Bacterial reverse mutation test (Ames)	2227	1180	4928	32148	0.38	5815	65	13	31	4
ER alpha agonism	1150	678	4213	13035	0.25	1211	63	24	21	5
FDA cancer rodent	2770	1695	2487	17741	0.47	5401	62	12	53	9
PXR binding human in vitro (new model)	3477	2168	2109	25485	0.54	9756	62	8	62	8
TPO inhibition rat in vitro	3040	2396	2858	25365	0.44	6615	56	10	52	9
PXR binding humans in vitro (old model)	3568	4412	947	17665	0.48	6221	45	5	79	20
Ashby structural alerts	2983	4603	4359	23536	0.24	2041	39	16	41	16
PXR agonism rat in vitro	466	905	6256	34692	0.09	348	34	15	7	3
FDA cancer mouse	796	2043	5901	22677	0.05	84	28	21	12	8
Respiratory sensitisation in humans	2040	5459	887	8470	0.23	911	27	9	70	39
Teratogenic potential in humans	972	2608	2064	14787	0.16	518	27	12	32	15
Allergic contact dermatitis	3253	10610	1690	15867	0.19	1119	23	10	66	40
Severe skin irritation in rabbit	1329	8712	4994	12972	-0.17	781	13	28	21	40
CYP2D6 substrates in humans	152	1064	4330	20729	-0.03	19	13	17	3	5
CYP2C9 substrates in humans	108	775	3984	18128	-0.03	19	12	18	3	4

* Sens1: Percent of hCAR actives also active in Model 1, Sens2: Percent of Model 1 actives also active in hCAR.

Table 2-6. Correlations between prediction for REACH substances from the hCAR Ant-50 µM (expanded) model and hPXR agonism and hAhR agonism (rational model) QSAR predictions from the free online Danish (Q)SAR Database (only predictions in the domains of the pairwise correlated models were counted).

hCAR agonism	hPXR agonism	hAhR agonism (rational)	Count
Negative	Negative	Negative	20,164
Negative	Negative	Positive	21
Negative	Positive	Negative	738
Negative	Positive	Positive	3
Positive	Negative	Negative	561
Positive	Negative	Positive	41
Positive	Positive	Negative	1,233
Positive	Positive	Positive	243

Figures

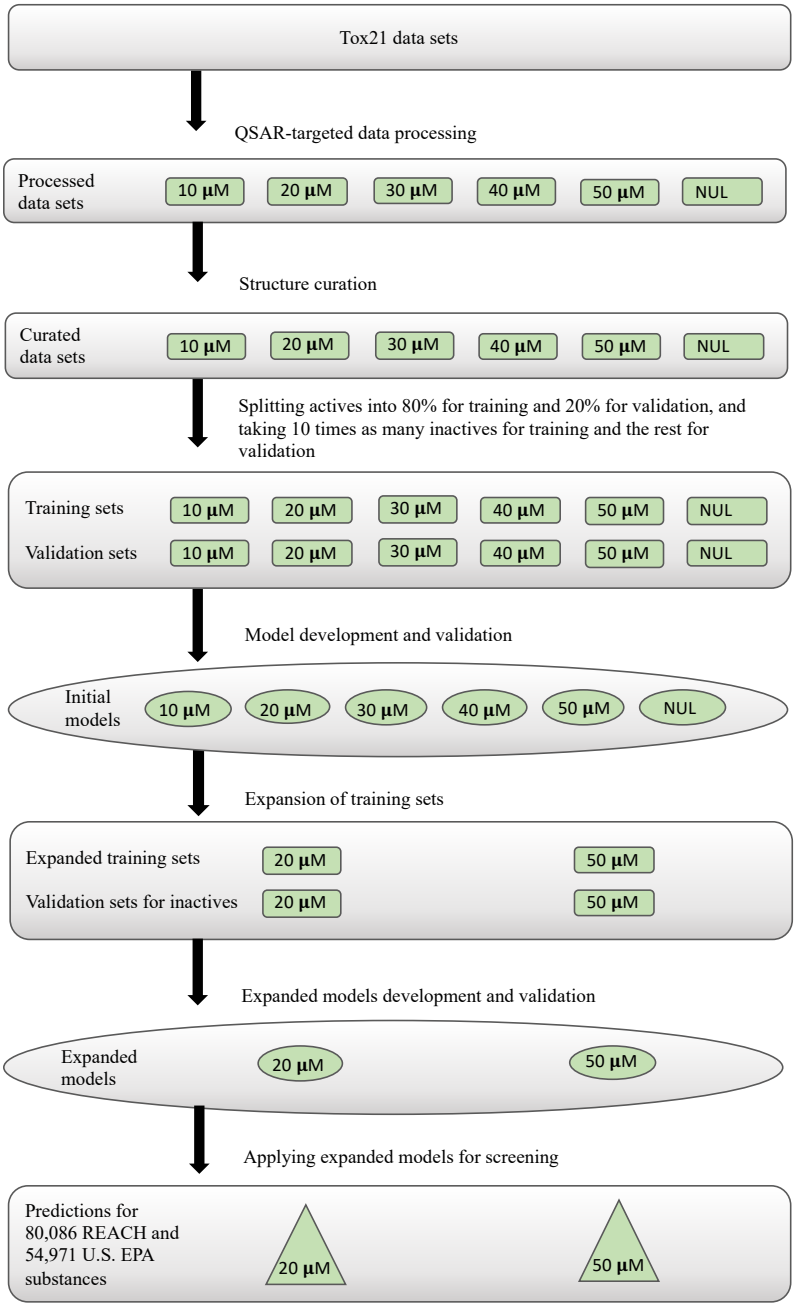


Figure 2-1. Overview of the process† of making training and validation sets, modeling, and predictions for hCAR antagonism and agonism activity.

†The rectangular boxes are for experimental data sets, the ovals are for models and the triangles are for the screening sets. The process was performed for both the agonism and antagonism sets.

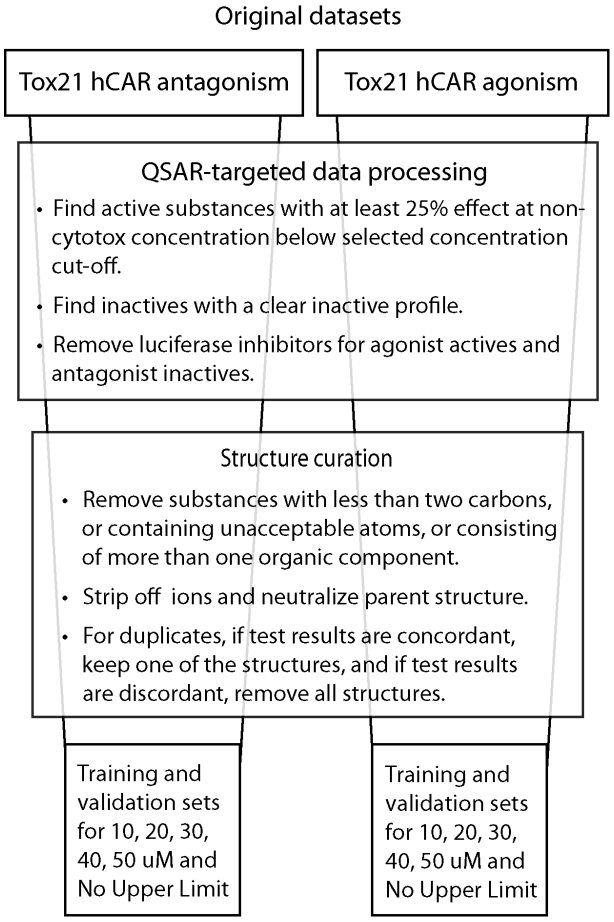


Figure 2-2. Data processing and structure curation.

3. Chapter 3: QSAR Use in REACH Analyses of Alternatives to Predict Human Health and Environmental Toxicity of Alternative Chemical Substances

Abstract:

In 2007, the European Union (EU) enacted the Registration, Evaluation, Authorization, and Restriction of Chemicals (REACH) to address growing concerns of hazardous chemicals in the EU market. Under REACH, companies seeking authorization to use priority substances identified as substances of very high concern (SVHCs) and included in the Authorization list, must apply and submit health and environmental effects data in analyses of alternatives (AoAs) to the Environmental Chemicals Agency (ECHA). To assess safer alternatives, especially in AoA hazard assessment cases where vital information could be missing or insufficient, quantitative structure activity relationship (QSAR) non-testing methods have gained increasing acceptance and importance. This article assesses AoA applicants' use of QSAR sources and QSAR documentation while looking for meaningful trends. In this assessment, usage and frequency of QSAR sources were evaluated in 189 analyses of alternatives for 15 physico-chemical properties and 19 human health and environmental endpoints to determine the scope of purpose of QSAR use in AoAs. We found that only 24 out of 189 applications cited QSAR sources to rank or evaluate the safety of their alternative substances relative to the Annex XIV chemical. For human health and environmental hazard endpoints, the Danish (Q)SAR Database (n=63) and unidentified QSARs (n = 33) were the most frequently cited QSAR sources by applicants. For QSAR usage, 7.9% per maximum opportunity (MOP) of alternatives with hazard endpoint QSAR predictions and 12% per MOP of physico-chemical QSAR predictions were used to report background information on an alternative. 3.0 % per MOP of hazard endpoint QSAR predictions supported the safety of the alternative while 0.7% per MOP of physico-chemical

QSAR predictions gave mixed support for their alternative's safety. Documentation regarding QSARs was absent in all 24 AoAs that used QSARs. Limited QSAR use and missing documentation may be the result of several factors including inconsistent regulatory guidance.

KEY WORDS: QSAR, analyses of alternatives, REACH, non-testing methods, predictive toxicology

Introduction

Authorization decisions under the European Union's (EU) Registration, Evaluation, Authorization, and Restriction of Chemicals (REACH) aim to replace substances of very high concern (SVHCs) with safer alternatives where economically and technically feasible (ECHA 2019). Under authorization, the European Chemicals Agency's (ECHA) Committees for Risk Assessment (RAC) and Socio-Economic Analysis (SEAC) assess risks as well as the availability and feasibility of possible alternatives in the analysis of alternatives (AoA) report for Annex XIV substances (ECHA 2019a). Guidance on the preparation of an application for authorization recommends the use of non-testing methods such as quantitative structure-activity relationships (QSARs), particularly in cases of data gaps (ECHA 2011b). QSARs, which are computational models that predict biological activity from compounds through statistical means, have become increasingly accepted in EU chemical regulation due to their cost-effectiveness and independence from animal testing. Yet, there is limited insight into the extent to which QSAR are used in AoAs. Recent evaluations performed by ECHA on alternative methods, (e.g. QSARs and read-across), focused exclusively on registration dossiers, for which "poor justifications for using alternatives to vertebrate animal testing" (ECHA 2016c), and several "deficiencies in the use of alternative testing methods," including the lack of proper documentation, supporting test data, and carefully developed chemical profiles were noted (ECHA 2017a)

This paper evaluates how QSARs are used to assess the safety of alternative chemicals in AoAs under REACH. Specifically, this study aims to answer three primary questions: *a)* What QSARs were used by AoA applicants?; *b)* How did applicants document their QSAR use in AoAs?; *c)* To what extent are QSARs used in AoAs to support, eliminate or evaluate in a weight of evidence (WoE) context an alternative to an Annex XIV chemical? This assessment also looks at the physico-chemical and human health and environmental hazard profiles of the alternatives

in order to look for trends in QSAR usage, support, and the role of WoE with respect to QSARs. These findings report on the nature and scope of QSAR use in AoAs as well as the applicants' use of the (Q)SAR modeling reporting format (QMRF) and the (Q)SAR prediction reporting format (QPRF) documentation. This paper concludes with a discussion on the likely need for additional regulatory guidance granted specifically for QSAR use under REACH authorization.

Background

Under REACH, the burden of establishing safety shifts from regulators to industry. In order to register a chemical, companies must compile data on their substance if manufactured or imported at greater than one ton per year (ECHA 2019d). All chemical properties, use and, when relevant, hazard and risk information requirements on the chemical are compiled in a registration dossier. Under Annex IV through X of the REACH regulation, a registrant may submit a proposal for chemical testing to fulfill its information requirements (ECHA 2019f). ECHA, the regulatory arm of REACH, and the Member States, assess dossier compliance and evaluate the proposal for testing. For each testing proposal, the company must detail its use or consideration of non-testing methods, including *in silico* modeling such as QSAR models, which give quantitative property and activity information in lieu of vertebrate testing. For the purpose of this study, we will use QSAR to indicate computational modeling, and (Q)SAR to signify both computational modeling and the structure-activity relationship (SAR), (i.e. grouping approach) (Benfenati 2012).

In the event that a chemical is identified and placed on a candidate list as an SVHC during the prioritization process, the chemical becomes eligible for authorization under Annex XIV. Assuming that the REACH Committee, with input from ECHA on high priority inclusions, then decides to include a candidate list SVHC on the Annex XIV list, the substance/use becomes

subject to a "sunset period," a period after which a substance/process can no longer be used unless authorization is obtained from the European Commission (EC) (EC 2019). However, to obtain authorization, an applicant must demonstrate, at a minimum, an absence of suitable alternatives. To assess if there are suitable replacements for an SVHC, an AoA is required (Article 62(4)(e). As part of the authorization process, an AoA must review all possible alternatives to Annex XIV substances based on the reduction of overall risks, and economic and technical feasibility (ECHA 2011b). This assessment includes the identification of potential alternatives, the screening process to select viable candidates, and an analysis of the suitability and availability of final selected alternatives. Applicants must also create a property and hazards profile for their alternative chemical(s), although, oftentimes, only limited toxicological information is available (NRC 2014b). Data gaps and limited information could present a significant obstacle to the identification of alternatives (Malloy et al. 2017). Still, QSARs can play a critical role in helping to fill these data gaps. (For the full REACH process, see Appendix S1).

QSAR Non-Testing Method Under REACH

Under REACH, QSARs are used to predict potential chemical toxicity of data-poor chemicals. QSARs are considered a non-testing tool because of their ability to predict data without further experimental testing (Worth 2010b). In general, QSARs are mathematical models that relate structural properties in the form of molecular descriptors, e.g. $\log K_{ow}$, to the biological activity of a chemical (Benfenati 2012). As a computational model, QSARs use algorithms, such as regression, to predict toxicity. Additionally, QSARs must be statistically validated to ensure model performance and prediction accuracy. Because of the uniqueness of each model, QSARs are accepted on a case by case basis under REACH (Benfenati et al. 2013).

Companies that use QSARs within a REACH process must first meet a number of conditions, which are defined under Annex XI as four criteria (European Parliament and the Council of the European Union 2006):

a) only scientifically validated models can be used

b) models have a defined domain of applicability;

c) models have appropriate measures of goodness-of-fit, robustness and predictivity as a prerequisite for use within "classification and labelling and/or risk assessment"

d) proper QSAR documentation must accompany the use of any QSAR

In addition to requiring valid QSARs, ECHA requires that companies provide reliable QSAR predictions. In practical terms, this means that QSAR data should be integrated with other *in vivo* and *in vitro* evidence. This is especially important when the experimental data are missing or insufficient. In these cases, companies should consider multiple data sources for a property or endpoint in a "totality of evidence" or weight of evidence (WoE) approach (ECHA 2016a). While ECHA has issued guidance that QSARs can be used to replace experimental data as long as QSAR results are considered "relevant, reliable and adequate for the purpose" and documented, in general, ECHA strongly advises a WoE approach that is based on experimental data considerations (ECHA 2008).

Because ECHA needs to verify QSAR results for regulatory decision-making, companies must formally document their QSAR predictions. Annex XI of the REACH regulation stipulates that adequate and reliable documentation, which outlines the reliability and quality of the QSAR model and prediction(s), must be provided by registrants when QSAR data is used (ECHA 2008). Currently, the main QSAR documents are the QMRF and the QPRF. QMRFs are typically prepared by the model developer, and provide information on the "source, type, development, validation, and possible applications of the model," which are aligned with the

OECD principles (ECHA 2008). QPRFs, on the other hand, detail the applicant's rationale for the predictions. QPRFs should include information on the environmental and human health endpoints, physico-chemical properties, and an assessment on whether the predicted toxicity of the alternative compared to the model's experimental data set is within acceptable limits (ECHA 2008). According to ECHA, both reporting formats must be submitted together with their company's registration dossier or authorization application (ECHA, personal communication).

Materials and Methods

Data Collection

To prepare for the review, we identified: *a)* the relevant endpoints; *b)* QSAR models, platforms, and databases; and *c)* AoAs. We started with a literature review of commonly targeted, regulatory properties and environmental and human health endpoints along with QSARs developed by or accepted under U.S. and E.U. regulatory regimes (Figure 3-1 Box 1). From this literature review, we created a list of 15 physico-chemical properties and 19 hazard endpoints based on three well-known QSAR guidance documents: *Guidance from ECHA's Practical Guide: How to Use and Report (Q)SARS* (ECHA 2016b), OECD's *Guidance Document on the Validation of (Quantitative) Structure Relationships ([Q]SAR] Models* (OECD 2007), and Cronin's (2010) *Prediction of Harmful Human Health Effects of Chemicals from Structure* (Figure 3-1 Box 2). Since ECHA does not endorse specific QSARs, we developed a standard list of QSARs (Appendix S2) to help identify QSARs tools and predictions used by AoA applicants (Figure 3-1 Box 3). For our baseline list, we selected an initial set of QSARs based on recommendations for QSARs applied under REACH (Benfenati 2012; ECHA 2016b; Worth et al. 2014). Our final or "default" list totaled 25 QSAR sources (Appendix S2). For the purpose of this study, we did not include a specific analyses of the OECD QSAR Toolbox, with

the exception of general descriptive statistics, as we collected this data under different QSAR codes. (For details on captured data for the OECD QSAR Toolbox, see Appendix 3).

Once we established our QSAR list, we reviewed 189 publicly available AoAs, located online at "Adopted opinions and previous consultations on applications for authorization" through May 2017 (ECHA 2019h) (Figure 3-1 Box 4). To collect data, we used a three-fold strategy: *a)* data-mining the 189 AoAs for QSAR models, guided by our rolling "default" list as a simple query; *b)* performing a visual inspection for QSAR usage; and *c)* recording any QMRFs and/or QPRFs on the ECHA website for Public Consultations. During our visual inspection of AoAs for QSAR usage, we targeted:

a) the citation and use of QSARs from our baseline QSAR list including any mention of applicability domain (AD) or the limits of the structural and information space of training sets or the experimental data used to make model predictions;

b) information on QSAR usage, which included the role of QSARs in supporting alternatives, and QSAR results in the chemical profiles, hazard assessments, reduction of overall risks, and conclusions of safety for the alternative chemical; and

c) QSAR results as part of WoE or as standalone evidence.

All data and findings were recorded in LibreOffice software Version 5.2.2.2.

In addition to data-mining AoAs for QSAR usage and models, we recorded all QSAR models/sources and predictions by endpoint. Qualitative values, (e.g. "negative for skin sensitization"), and quantitative values, (e.g. "LC_{50} = 3.83 mg/L for acute toxicity for fish") (Dow Italia Srl and Rohm and Haas France S.A.S. 2016) immediately qualified as QSAR predictions if a data point directly cited a QSAR source or if an AoA cited QSAR prediction(s) generated by a QSAR model/platform. If an applicant did not identify the QSAR source, but did specify a value as a QSAR prediction, we recorded the prediction as "unknown QSAR." We also

considered any value that reported its AD to be a QSAR prediction and used expert judgment for any remaining values that might be viewed as QSAR predictions. Furthermore, we organized all data, which we based on several assumptions when developing our approach, into two major units of analysis, (i.e. consultation by alternative, and alternative by QSAR). (For full details on units of analysis, see Appendix S4).

Lastly, we defined the "purpose" of the applicant's use of QSAR predictions. "Purpose" characterized the QSAR prediction's role in either supporting or eliminating the alternative with or without a WoE context (Table 3-1). All data were recorded for later use in our content analysis.

During the course of this data collection, we looked for QMRF and QPRF attachments on ECHA's consultations and opinions website (ECHA2019h), and recorded any instances where we found documentation. We forwarded our questions on authorization documentation requirements to ECHA regulatory officials and used these correspondences as informal interviews.

Descriptive Statistics

We used descriptive statistics to examine QSAR use (Figure 3-1 Box 5). We totaled counts across two dimensions: *a*) AoAs with cited QSAR use; *b*) QSAR predictions grouped by QSAR source (i.e., database, platform or model used) (Figure 3-1 Box 6). We based this tally on total QSAR prediction counts. Next, we tabulated the frequency of QSAR use. Data were assigned to either one of four categories: *a)* "QSARs applied," which meant an applicant cited a QSAR from our list for at least one value for specific endpoint or property; *b)* "QSARs not applied" or a QSAR from our list was not used for a specific endpoint or property; *c)* "No data available," which meant that the applicant reported "no data" in the value field, or "no

information available" or "not relevant"; *d)* "Unidentified QSAR tests," which meant that the applicant provided a QSAR value but did not indicate the source. We then "normalized" these values to reflect QSAR opportunities rather than strict counts. "Normalization" involved tallying the actual count of QSARs by the unit of consultation number by alternatives, and then dividing this number by the total alternative population count, which we defined as the maximum times a QSAR could have been used for either the 15 physico-chemical properties, or, the 19 hazard endpoints. Afterwards, we calculated their percentages. In the end, our analysis told us the maximum opportunities (MOP) a QSAR could be used in a hazard endpoint or property category. (For more information on MOP, see Appendix S5). By adjusting counts to percentages, we were able to align and scale the raw data to simple percentages for comparison. Finally, we analyzed the purpose(s) for which a QSAR was used (Figure 3-1 Box 7). We analyzed by elimination, support, and/or WoE by tabulating the count for each respective use and inspected trends. If an endpoint or property lacked QSAR predictions, or, we could not determine if its value was generated by a QSAR, the value received a non-applicability designation. (Full information on non-applicable designations can be found in Appendix S6).

Results

Frequency of QSAR Use

Applicants used QSAR data to rank potential alternatives or evaluate the safety of suitable alternative chemicals relative to the Annex XIV substance in approximately 13% of the AoAs or 24 out of 189 AoAs. Among the 24 AoAs, QSARs were used to assess 54 overlapping and unique alternatives. In addition, 11 of the 25 QSARs on our list were used by applicants. Out of the 24 AoAs, 19 of the AoAs cited QSAR predictions for eight of the 15 physico-chemical properties. The physico-chemical properties with the greatest amount of QSAR predictions

included water solubility (n = 37), surface tension (n = 24), and vapor pressure (n = 23). Physico-chemical properties with the least number of predictions were flash point (n = 8), density (n = 7), and boiling point (n = 11) (Table 3-2).

In the assessment of the toxicity of alternatives for human health and environmental hazard endpoints, six of the 24 AoAs cited QSAR predictions for 16 out of the 19 hazard endpoints. Endpoints with the most QSAR predictions included reproductive toxicity (n = 50), short-term toxicity to fish (n = 18), short-term toxicity to aquatic invertebrates (n = 15), and skin irritation/skin corrosion (n = 15). QSAR predictions were cited least for long-term toxicity to fish (n = 2), short-term toxicity to terrestrial invertebrates (n = 2) and hydrolysis (n = 1). We did not find QSAR predictions for acute toxicity, repeated dose toxicity or adsorption/desorption screening (n = 0) (Table 3-3).

The types of QSAR sources that applicants cited when assessing human health and environmental hazard endpoints consisted of a limited number of available QSARs though applicants more generally referenced the OECD QSAR Toolbox 13 times. As illustrated in Figure 3-2, applicants used the Danish (Q)SAR Database for nine out of the total 19 endpoints. The other major QSARs sources included ECOSAR, OASIS, and the FDA EDKB CoMFA QSAR method. The most popular MultiCase models were endocrine disruption models for estrogen receptor binding. A MultiCase model was also used to predict short-term toxicity for aquatic invertebrates. In addition, TOPKAT® was used mainly for the ecotoxicological endpoints. However, for physico-chemical properties, only MultiCASE, ACD/PhysChem and EPI Suite™ QSAR sources were used or cited (Figure 3-3). (Full details on the major QSAR sources can be found in Appendix S7).

Regarding the frequency of QSAR usage in hazard endpoints, only 11% of the hazard endpoints per MOP used a QSAR tool (Figure 3-4). In order of highest to lowest frequency of counts of QSARs per MOP, reproductive toxicity (1.2%) had the most QSAR predictions out of the 54 alternatives that could have potentially cited a QSAR prediction, while long-term toxicity to fish (0.2%), short-term toxicity to terrestrial invertebrates (0.2%), and hydrolysis (0.1%) had the least amount of QSAR use. No predictions were found for acute toxicity, repeated dose toxicity or adsorption/desorption screening endpoints.

Purpose of QSAR Use

Meaningful trends for both hazard endpoints and physico-chemical properties emerged in our assessment of QSAR use for elimination, support, and WoE. None of the QSARs were used to overtly eliminate alternatives with or without WoE (Figure 3-5). On the other hand, smaller trends emerged, which indicated a variety of QSARs uses. Specifically, 7.9% per MOP of hazard endpoint and 12% per MOP of physico-chemical and QSAR predictions were used to report background information for an alternative. For example, an LC_{50} QSAR prediction by ECOSAR for acute toxicity to aquatic species, (i.e. 0.023 mg/L) was included in a table of ecological data supporting Environment Canada decisions on methyl centralite for AoA consultation number 0005-02. Although this prediction, which according to ECOSAR guidelines (U.S. EPA 2012c) qualifies as acutely toxic, the applicant never discussed this prediction in the context of a possibly toxic estimate in either the environmental fate and behavior and ecotoxicology discussion or the reduction of overall risks section (DEZA A.S. n.d.-a).

In addition, “other” reasons, which included QSARs used to rank potential alternatives, emerged as a small trend. For “other” reasons, we observed a 2.9% per MOP frequency for

hazard endpoints and 0.4% per MOP frequency for physico-chemical properties. For example, in consultation numbers 0013-01 and 0013-02, applicants reported persistent, bioaccumulative and toxic (PBT) QSAR estimates for the initial assessment of the polymeric flame retardant alternative (INEOS Styrenics Netherlands BV et al. 2013a, b). In addition, consultation number 0077-01, the applicant used QSAR predictions to select reference values for the potential alternative dichloromethane (Eli Lilly S.A. Irish Branch n.d.). Under "other" reasons, QSAR predictions could also be used to compare endpoints for an alternative. For AoA consultation number 0005-01, some QSAR predictions suggested a more benign mammalian hazard profile for the alternative diisobutyl hexahydrophthalate (DIBE) relative to its Annex XVI chemical dibutyl phthalate (DBP), while others, more specifically, predictions for reproductive toxicity, gave warrant for concern (DEZA A.S. n.d.-b).

Trends changed slightly with respect to QSAR use for support of an alternative (Figure 3-6). 3.0% per MOP of hazard endpoint QSAR predictions supported an alternative. In AoA consultation number 0005-01, the applicant used QSAR outputs to support a more benign assessment of DIBE relative to the Annex XIV chemical,

> However, there is reason to believe that DIBE may have a more benign mammalian hazard profile than DBP, given the lack of a REACH Registration dossier (so far) and some aspects of the OECD QSAR outputs" (DEZA A.S. n.d.-b).

Notably, we identified several QSAR predictions that did not support an alternative. For example, in consultation 0006-01, a $\log K_{ow}$ prediction for DIBE suggested a low chronic aquatic toxicity category 2 classification (Sasol-Huntsman GmbH & Co. KG n.d.).

Additionally, “other” reasons ranked high in terms of frequency of QSAR use for hazard endpoints and physico-chemical properties, (i.e. 6.3% and 13% per MOP, respectively). Under QSAR use for supporting an alternative, “other” reasons included QSAR predictions used for a “background” purpose or to assess a potential alternative. For instance, consultation numbers 0037-01, 0038-01, 0039-01, 0040-01, 0041-01, and 0041-02, for the alternative Chromium III (chloride), cited QSAR predictions for water solubility, but then never drew from these predictions to indicate a direction of support for the alternative (AkzoNobel Pulp and Performance Chemicals 2015; Caffaro Brescia S.r.l. 2015; Ercros S.A. 2015; Kemira Chemicals Oy 2015; Solvay Portugal – Produtos Quimicos SA 2015).

Finally, we observed a small trend for “mixed” support for hazard endpoints and physico-chemical properties, (1.2% and 0.7% per MOP, respectively). In “mixed” support, an applicant could use QSAR predictions to show tentative support for an alternative as having a more benign profile than the Annex XIV chemical, but then also include evidence indicating a potential hazard(s). Consultation number 0005-02 for the alternative methyl centralite demonstrated this version of support for the endpoint short-term toxicity to aquatic invertebrates. In their AoA, the applicant referenced irritation and aquatic toxicity as areas of concern for methyl centralite in the Reduction of overall risks, for which a positive QSAR prediction for skin irritation and possibly toxic aquatic QSAR predictions were given by the Danish (Q)SAR Database) (DEZA A.S. n.d.-a). However, for aquatic toxicity, the results were, in fact, both supportive and unsupportive of methyl centralite. The applicant referenced several aquatic toxicity results from the Danish (Q)SAR Database as a possible concern (DEZA A.S. n.d.-a),

> QSAR modelling of the ecotoxic profile of methyl centralite indicated possible concern with regard to its aquatic toxicity, with predictions of LC50 or EC50 values < 1 mg/L in invertebrate, algal and bacterial species.

Yet, in the Canada List QSAR predictions, all aquatic QSAR predictions, with the exception of an EC_{50} or LC_{50} aquatic toxicity prediction for 0.023 mg/L reported, were greater than one mg/L although the applicant did not reference this fact in their concluding argument (DEZA A.S. n.d.-a).

Documentation of QSARs

None of the 24 AoAs using QSARs included or referred to documentation for the relevant QSARs. As specified by ECHA guidance, there are two ways one can provide documentation: *a)* QMRF and *b)* QPRF. However, neither was provided. Due to the specificity of QMRF documentation for model development, we could not identify any QMFRs associated with QSAR models in our study in either the JRC database (JRC 2019) or the Danish (Q)SAR database (DTU FOOD et al. 2019).

Discussion

In this study, we assessed the extent to which QSARs were used in REACH AoAs through March 2017. Our analysis demonstrated limited QSAR use in AoAs. Our study also revealed that supporting documentation for QSAR predictions was missing for all 24 AoAs with QSAR predictions. Lack of documentation is a significant potential obstacle for regulators and stakeholders trying to determine the quality of QSAR tools and data in AoAs. Additionally, existing regulatory guidance on QSAR documentation can be viewed as inconsistent, which may have factored into the missing QMRFs and QPRFs. Specific guidance or enforceable standards

for QSAR use and documentation in AoAs could help strengthen applicants' understanding of ECHA QSAR guidance for future AoAs.

Limited use of QSARs in AoAs

In our analysis, only 24 of the 189 AoAs used QSARs. In addition, while we identified 25 potential QSAR sources from literature, only 11 of these QSAR sources were actually used. The limited usage of QSARs in AoAs is similar to results reported in the 2014 ECHA report regarding QSAR use in registration (ECHA 2014). We believe this to be a significant trend: When companies do not exploit the gamut of QSAR sources, vital QSAR predictions could be omitted from regulatory decision making (Cronin 2010). These observations ultimately raise the important issue of whether QSARs are underused in AoAs.

More consistent use of appropriate QSARs in AoAs, however, will provide a better picture of the degree of hazard posed by the SVHC(s) relative to the alternative(s). One reason that QSAR predictions help create a more robust toxicological understanding is that animal testing is often flawed and can be inconsistent (Akhtar 2015). In these cases, QSARs would help to provide more accurate results (Lillicrap et al. 2016) as well as support uncertain test results (Lahl and Gundert-Remy 2008). Second, in the face of diverse and uncertain information, it is almost always better to have as much relevant information as possible. WoE perspective encourages the use of everything available, including QSARs, to get the best answer (ECHA 2010).

At the same time, ECHA has also been known to promote the use of standalone QSAR predictions for regulatory purposes, if only under ideal circumstances, though this allowance could have the opposite effect and limit QSAR use (ECHA 2008). In their recent 2017 *Endpoint*

specific guidance Version 6.0, ECHA addressed the opportunity for standalone non-testing methods "where possible" in addition to their WoE recommendation (ECHA 2017d). AoA consultation numbers 0006-01 and 0005-02 reflect this standalone QSAR approach. The applicants provided non-testing data from the OECD QSAR Toolbox, more specifically, FDA Teratogen Information System (TERIS) QSAR predictions from the Danish (Q)SAR Database, for alternatives such as Akardite II to fill in the experimental data gap for the reproductive toxicity and endocrine disruption potential of the alternative DIBE (DEZA A.S. n.d.-a; Sasol-Huntsman GmbH & Co. KG n.d.). Applicants' justifications for standalone QSAR use, such as "no [experimental] information is available on the potential reproductive toxicity of this substance" underscored the importance of QSAR predictions in their AoAs (DEZA A.S. n.d.-a).

Yet, we cannot ignore a potential alternative response of applicants: companies may decline to provide any data when their only source is QSAR predictions. While this is solely speculative, we have seen evidence of applicants rejecting QSAR data when it is the only available information. In AoA consultation numbers 0078-01 and 0077-01, treatment of standalone QSAR predictions reflect this viewpoint. Applicants dismissed a QSAR-based Predicted No-Effect Concentration **(**PNEC) value (i.e. 130 µg/L (100)) when developing a $PNEC_{freshwater}$ value for methylene chloride (DCM), because it was based solely on QSAR predictions (Dow Italia Srl and Rohm and Haas France S.A.S. 2016; Eli Lilly S.A. Irish Branch n.d.),

> All data regarding the individual submission by ECHA-CHEM are based on QSAR, only. Because of this, the resulting PNEC is regarded as being not reliable and consequently will not be discussed any further.

Interestingly, the $PNEC_{freshwater}$ value was set at more than twice the concentration of the QSAR-based PNEC value. Of course, without knowing the predicted or actual environmental concentration, we cannot ascertain if the $PNEC_{freshwater}$ would actually amount to a higher or lower acceptable risk (Syberg et al. 2009).

Because of the risk of not properly using QSARs brings to making a thorough assessment of AoAs, we strongly believe that ECHA should: 1) more clearly require the use of valid QSARs in an alternative's assessment particularly in the absence of other information; and 2) consistently communicate equal support and acceptance for both standalone predictions under strict conditions, and for WoE QSAR evidence. Absent such direction, some applicants may follow the practice of ignoring QSAR data as "unreliable" without discussion.

QMRF and QPRF documentation in AoAs

Under REACH, QSAR documentation serves an important role in establishing the reliability of QSAR predictions and the robustness of QSAR models used in AoAs. *Chapter R.6-QSARs and Grouping of Chemicals* explains that each document provides critical information about the QSAR model and prediction, which aids in the regulatory decision-making process. However, despite ECHA's general guidance, QSAR documentation was missing for all 24 AoAs that used QSARs. According to the ECHA Regulatory Advice Team, missing documentation could be attributed to one of two factors: either the QMRF and QPRF were claimed as confidential, or they were not included with the AoAs (ECHA, personal communication). Based on our observations, it appears that QMRFs and QPRFs were missing for reasons other than issues of confidentiality. We examined 24 non-confidential AoAs that applied QSAR predictions in their assessments of their alternative(s); none of them had associated QMRFs or QPRFs.

One reason for the missing documentation in AoAs could be the imprecise language in regulatory guidance describing the requirements for QMRFs and QPRFs. The language appears to alternate between strict and loose requirements, which may have even been interpreted by applicants as creating voluntary options. For example, in the 2016 Practical Guide – How to use and report (Q)SARs 3.1, ECHA stressed that a QPRF "should be prepared by the registrant," and its corresponding QMRF is "normally provided by the model developer" though it is "recommended" that both be attached to the registration dossier (ECHA 2016b). On the other hand, in *Guidance on Information Requirements and Chemical Safety Assessment, Chapters R.7a and R.7b: Endpoint specific guidance*, ECHA discusses standard documentation and justification of QSARs under Annex XI under terms such as "need" and "necessary" (ECHA 2017b; ECHA 2017d). Indeed, in personal communications ECHA's Regulatory Advice Team used somewhat inconsistent language regarding whether applicants must- as opposed to should- attach a QMRF and QPRF to the AoA when addressing (Q)SAR documentation (ECHA, personal communication). However, in a separate email to ECHA's Risk Management Implementation Unit, ECHA insisted that documentation was required for AoAs using QSARs (ECHA, personal communication). In short, companies submitting AoAs may have looked to registration-targeted guidelines for instructions on what to do when using a QSAR under authorization and concluded that documentation for authorization was optional.

At the same time, the current QMRF and QPRF format may not be a "one size fits all" framework for every type of QSAR model or QSAR source. For example, QSAR models may be built as part of an expert system, (i.e. Toxtree). In these instances, an applicant may need to modify some of the information such as supplementing decision tree-making rules in lieu of algorithms. In more extreme cases, an applicant may need to give a more detailed explanation

pertaining to which exact rules took place and how that impacts the quality of the prediction (Benfenati 2012). However, as to the exact format that ECHA would accept this information under authorization remains to be determined.

Ultimately, without QSAR model and prediction information, ECHA cannot determine modelling components or how an applicant understands the relevance of endpoints, though, the degree to which developers make their QSAR information available on their QSAR model(s) varies. For registration, ECHA has enforced this guidance. In their 2013 evaluation, ECHA reported an improvement of proper QMRF and QPRF documentation for registration dossiers (ECHA 2013) from previous years. Although, in prior evaluations, some registration dossiers lacked the required documentation for structure activity relationships (Schulte et al. 2012). In our study, AoA applicants generally gave information on the applicability domain of their QSAR prediction (For more information on applicability domain data, see Appendix S8). Yet, applicants did not always provide an explanation as to whether their QSAR model covered their alternative chemical. With a QPRF, ECHA would have been able to discern whether an applicant used an out of domain prediction particularly if the applicant did not initially report this information in their AoA. Perhaps in the future, an applicant will even be required to provide all evidence of their due diligence even if only to show that their chemical(s) could not be covered by the QSAR model due to an unacceptable structural feature such as counter ions, toxic metals or complex molecules (Klimenko KO et al. 2019). Nonetheless, a developer's proprietary rights over their QSAR model may have also factored into the missing documentation. While some QSARs are open access such as the JRC QSAR Model Database, Toxtree, Dart (Benfenati 2012) and Danish (Q)SAR Database, commercial software such as CASE Ultra (MultiCase) and OASIS CATALOGIC (LMC) and OASIS TIMES (LMC) does not lend the same level of QSAR

model transparency, although some QMRFs for commercial software QSARs (though none could be found for our study) can be found in the JRC QSAR Model Database (EURL ECVAM 2019). We believe that enforcement needs to extend to authorization if there are to be similar improvements in AoAs.

The importance of including QMRFs and QPRFs under REACH clearly extends beyond a formal compliance check list, yet, there is almost no regulatory guidance on QSAR documentation for authorization, with the exception of a small excerpt in *Chapter 3: Planning for Substitution: Guidance on Analysis of Alternatives*. In this excerpt, ECHA broadly states that in situations where important information is missing, surrogate information from tools such as QSARs may be used (ECHA 2011b). In short, almost all legislative language and guidelines describing QSAR requirements are reserved for registration. For example, REACH's Annex VI of Title VII, which applies to "registration, evaluation and the duty of care," states that registrants must consider all available data on their substances, including alternative sources such as QSARs (European Parliament and the Council of the European Union 2006).

Study limitations

Several factors may have contributed to the limitations of our study. First, although our sample of AoAs appears to be representative of AoAs as a whole, our sample had a significant number of AoAs with redacted language, which could have impacted our analysis. Second, unless the applicant made explicit reference to their QSAR models or sources, we only accepted data on a conditional basis. Third, supplier confidentiality protocols prohibited a complete hazard analysis on their product(s). Thus, for these AoAs, it was not possible to know the complete spectrum of hazard analysis and potential QSAR usage.

In addition, depending on how the applicant cited their QSAR software package and how we coded QSAR predictions, we may not have been able to identify the exact QSAR tool. In addition, because we coded QSAR predictions based on a set list of regulatory endpoints and QSAR sources, we may not have captured every QSAR in our sample of AoAs, particularly, if an endpoint or source fell outside of this list. Finally, while we were successful in generating descriptive statistics to illustrate rank, order, and frequency, these statistics were limited to descriptions of prioritization and missing data and did not capture the capacity and appropriateness of each QSAR in precise detail. For example, the frequency counts of certain QSARs may have depended on the tool's design for a specific hazard endpoint(s), which our statistics did not capture, rather than its popularity. Furthermore, applicants may have chosen their QSAR tool based on accessibility to the tool, (i.e. availability and user cost), ease of use, and perhaps confidence in the model's algorithm or platform's navigability, though more research is needed to confirm predisposing factors such as confidence or motivation. (For more information on study limitations, see Appendix S9).

Conclusions and Recommendations

The REACH regulation's goal to protect human health and the environment by using alternative non-testing methods in a competitive market has been partially met through the use of QSARs. However, we identified potential obstacles in this path to achieving this goal. To our knowledge, there is no other review of how QSARs are used in REACH analyses of alternatives. At the time of our study, only 24 out of 189 AoAs in our sample used QSARs, and only 11 of the 25 QSARs on our list were used by applicants. These low numbers suggest that QSARs may not have been fully utilized. We believe that our results point to three larger policy issues:

- a possible breach in policy whereby companies may not be exhausting all toxicity data sources, including QSARs, for their chemical(s), particularly in cases of insufficient or missing information;
- the necessity for clearer regulatory language and guidance on the use of QSARs in concert with other data, and where appropriate, on a standalone basis;
- a critical need for more explicit language requiring proper QSAR documentation under authorization.

While we conducted a preliminary examination of QSAR use within a limited number of AoAs, future studies are needed to understand the extent to which ECHA accepts AoAs, which do not fulfill ECHA QSAR requirements. In our sample, we identified 22 AoAs that used QSARs for final, selected alternatives and two that used QSARs to assess potential alternatives. Of these 24 AoAs, the European Commission authorized 16 AoAs, which did not attach required QSAR documentation, for continued use of an XIV chemical. However, to understand how ECHA factors in varying trends of QSAR usage into accepted applications, future studies that draw on a much larger AoA sample and include more recent AoAs are needed. In addition,

applicants who exist outside of industry may have better training and follow ECHA's QSAR guidance more closely, which could impact the consideration of their AoAs. Even if the European Committee authorizes an Annex XIV chemical where QSAR predictions are used to assess the safety of the alternatives, it is important to garner information on patterns of QSAR usage, and to target applications that do not fulfill QSAR guidance.

The REACH regulation highlights the need for applying new non-testing methods to bridge data gaps, promoting their acceptance, and delineating data requirements to ensure data quality in chemical risk assessment. As a non-testing method, QSARs play a critical role in the evaluation of chemicals on the Annex XIV list. Although ECHA has applied a cautious approach, QSARs are being used both in WoE and as standalone in AoAs. Focusing on the quality of QSAR data in AoAs and building on previous QSAR guidance to develop consistent regulatory language will likely increase the quality and volume of QSAR data. Thus, stronger industry regulatory requirements to provide information on alternative chemicals (Jacobs et al. 2016) as well as greater access and availability to comprehensive QSAR warehouses of predictions and models with demonstrated success, may be the key to reducing data gaps in AoAs. However, ECHA must ensure that reliable information is evident on all levels of QSAR documentation, and that high quality and user-friendly QSARs are accessible to industry. Along with the continued evaluation of the effectiveness of regulatory tools for AoAs, particularly for use by smaller companies (Tickner and Jacobs 2016), evaluation of data quality and effectiveness of guidance will become increasingly important in regulatory agendas worldwide, which accepts the predictive ability of QSARs to provide toxicological information on unknown chemical effects.

Acknowledgements

The authors would like to thank Marina Chen, who assisted with the data analysis and without which we would not have been able to develop the “default” QSAR list.

Tables

Table 3-1. Definitions and classifications of purpose for QSAR predictions used in AoAs.

Purpose	*Definition*	*Classifications*
Eliminate	QSAR prediction(s) eliminated an alternative based on a more toxic profile than the Annex XIV chemical	a) eliminate the alternative; b) report background information for the alternative without eliminating the alternative; c) other reason e.g. QSAR prediction not discussed in the final assessment of the alternative's reduction of overall risk or safety d) non-applicable e.g. no QSAR prediction
Support	QSAR prediction(s) supported an alternative as having an equally benign or more benign profile than the Annex XIV chemical	a) show an alternative as less benign; b) support an alternative; c) mixed support; d) other reason e.g. QSAR prediction not discussed in the final assessment of the alternative's reduction of overall risk or safety e) non-applicable QSAR data and/or no QSAR use
Weight of Evidence (WoE)	QSAR prediction(s) were used with a combination of soures to describe the alternative's a hazard profile	a) elimination used in a WoE context; b) elimination not used in a WoE context c) a combination of independent QSAR usage and QSARs used in WoE context to eliminate alternative

Table 3-2. Physico-chemical property non-testing methods regarding QSAR methods (i.e., QSARs applied, QSARs not applied, No data available or Unknown QSAR tests).

Physico-chemical property	QSAR counts[†]			QSAR sources		
	QSAR(s) used[a]	QSAR(s) not used[b]	No data available[c]	ACD/PhysChem	EPI Suite™	CASE Ultra (MultiCase)
Melting/freezing point	8	46	13	-	8	-
Boiling point	11	43	13	5	6	-
Density	7	47	13	7	-	-
Vapor pressure	23	23	21	6	17	-
Surface tension	24	22	21	24	-	-
Water solubility	37	24	6	1	36	-
Partition coefficient n- octanol/water	13	33	21	-	11	2
Flash point	8	38	21	8	-	-
Disassociation constant	-	-	67	-	-	-
Viscosity	-	-	67	-	-	-
Flammability	-	9	58	-	-	-
Explosive properties	-	9	58	-	-	-
Self-ignition temperature/auto-flammability	-	17	50	-	-	-
Oxidizing properties	-	5	62	-	-	-
Granulometry	-	11	56	-	-	-

(Table 3-2 continued)

† QSAR count unit of analysis is by type of QSAR model for the 54 alternatives (n = 67).

[a] QSAR prediction was reported for that specific physico-chemical property.

[b] A QSAR was not applied to that physico-chemical property, which may be for other reasons such as that specific property not being relevant to that QSAR source or no information is available.

[c] For 'no data available,' an applicant stated ”no data” in the value field for that physico-chemical property, or “no information available” or “not relevant" in the summary information for that alternative.

Table 3-3. Human health and environmental hazard endpoint non-testing methods regarding QSAR methods (i.e., QSARs applied, QSARs not applied, No information).

Human Health and Environmental Hazard Endoint	QSAR counts[†]		
	QSARs Used[a]	QSARs not applied[b]	No Information[c]
Acute Toxicity	-	-	118
Skin Irritation or skin corrosion	15	-	103
Eye Irritation	9	-	109
Skin Sensitization	7	1	110
Repeated dose toxicity	-	-	118
Mutagenicity	14	-	104
In vitro gene mutation in bacteria (Ames test)	8	-	110
Carcinogenicity	10	-	108
Reproductive toxicity	50	1	67
Short-term toxicity to fish	18	-	100
Long-term toxicity to fish	2	-	116
Short-term toxicity to aquatic invertebrates	15	-	103
Long-term toxicity to aquatic invertebrates	7	-	111
Toxicity to aquatic plants (algae)	10	-	108
Short-term toxicity to terrestrial invertebrates	2	-	116
Hydrolysis	1	-	117
Ready biodegradability	7	-	111
Bioaccumulation in aquatic species	12	-	106
Adsorption/desorption screening	-	-	118

(Table 3-3. continued)

Human Health and Environmental Hazard Endpoint	QSAR sources									
	Danish (Q)SAR Database	ECHA CHEM	ECOSAR	EPI Suite™	FDA EKDB CoMFA	M1-LC50	CASE Ultra (MultiCase)	Oasis	TOPKAT	Unidentified††
Acute Toxicity	-	-	-	-	-	-	-	-	-	-
Skin Irritation or skin corrosion	6	-	-	-	-	-	-	-	-	9
Eye Irritation	-	-	-	-	-	-	-	-	-	9
Skin Sensitization	5	-	-	-	-	-	-	1	1	-
Repeated dose toxicity	-	-	-	-	-	-	-	-	-	-
Mutagenicity	10	-	-	-	-	-	-	-	-	4
In vitro gene mutation in bacteria (Ames test)	8	-	-	-	-	-	-	-	-	-
Carcinogenicity	10	-	-	-	-	-	-	-	-	-
Reproductive toxicity	10	-	-	-	12	-	10	8	-	10
Short-term toxicity to fish	4	-	5	-	-	4	-	2	3	-
Long-term toxicity to fish	-	-	2	-	-	-	-	-	-	-
Short-term toxicity to aquatic invertebrates	6	-	5	-	-	-	1	-	3	-
Long-term toxicity to aquatic invertebrates	-	2	5	-	-	-	-	-	-	-
Toxicity to aquatic plants (algae)	4	-	5	-	-	-	-	-	1	-
Short-term toxicity to terrestrial invertebrates	-	-	2	-	-	-	-	-	-	-
Hydrolysis	-	-	-	1	-	-	-	-	-	-
Ready biodegradability	-	-	-	1	-	-	-	-	4	2
Bioaccumulation in aquatic species	-	-	-	6	-	-	-	4	-	2
Adsorption/desorption screening	-	-	-	-	-	-	-	-	-	-

† QSAR count unit of analysis is by type of QSAR model for the 54 alternatives (n = 118).

†† Unidentified means that the identify of the QSAR model was not reported in the AoA even if the QSAR result was provided.

[a] A specific QSAR prediction has been cited for that hazard endpoint.

[b] QSARs not applied to that hazard endpoint. Reasons may include that specific property not being relevant to that QSAR source or no information is available.

[c] "No Information" includes AoAs where QSAR predictions are not applied to an alternative's assessment and where an applicant reported "no information" (for which it is unknown if QSAR was run) in their non-testing summary table for an endpoint.

Figures

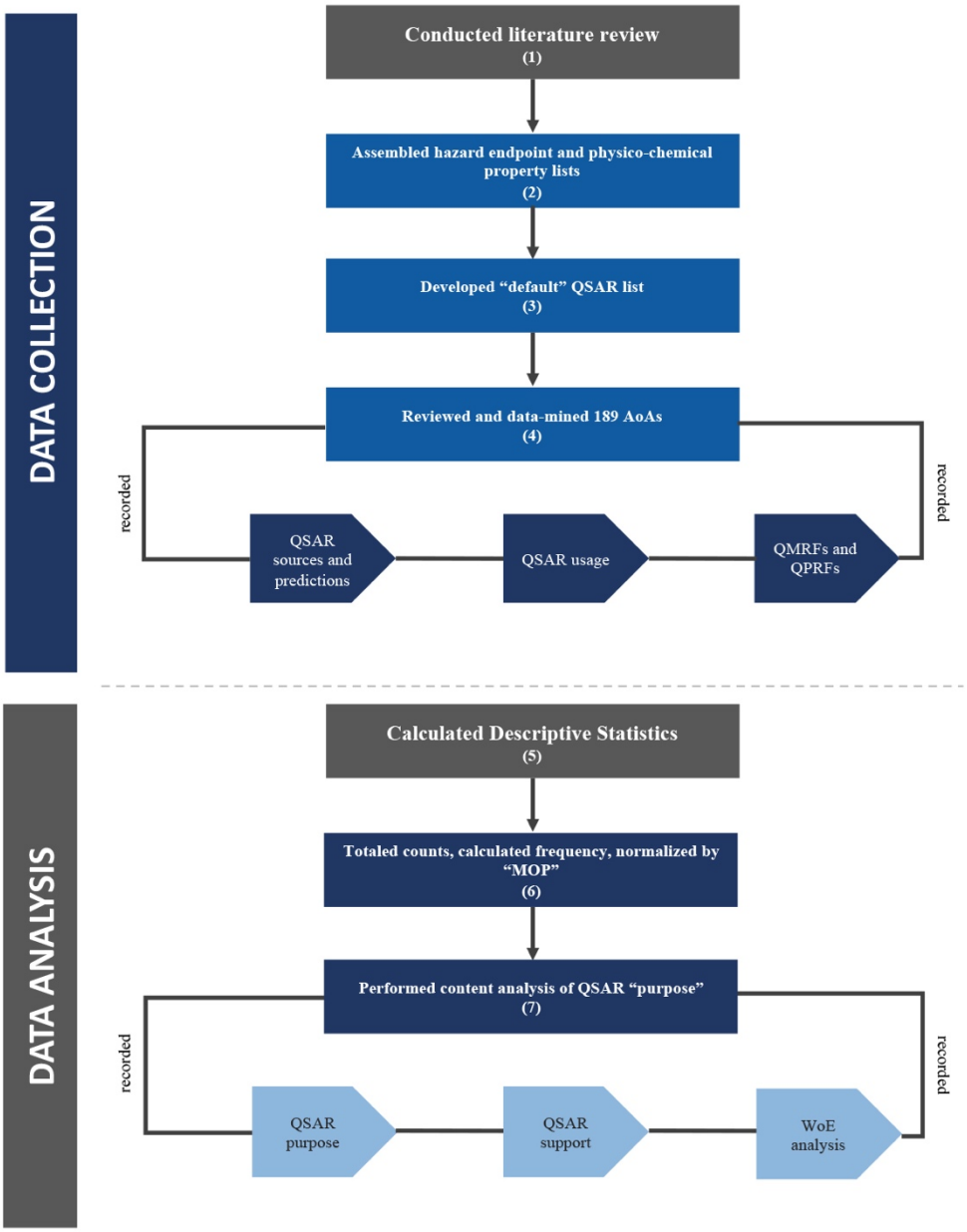


Figure 3-1. Flow chart for data collection and descriptive statistics.

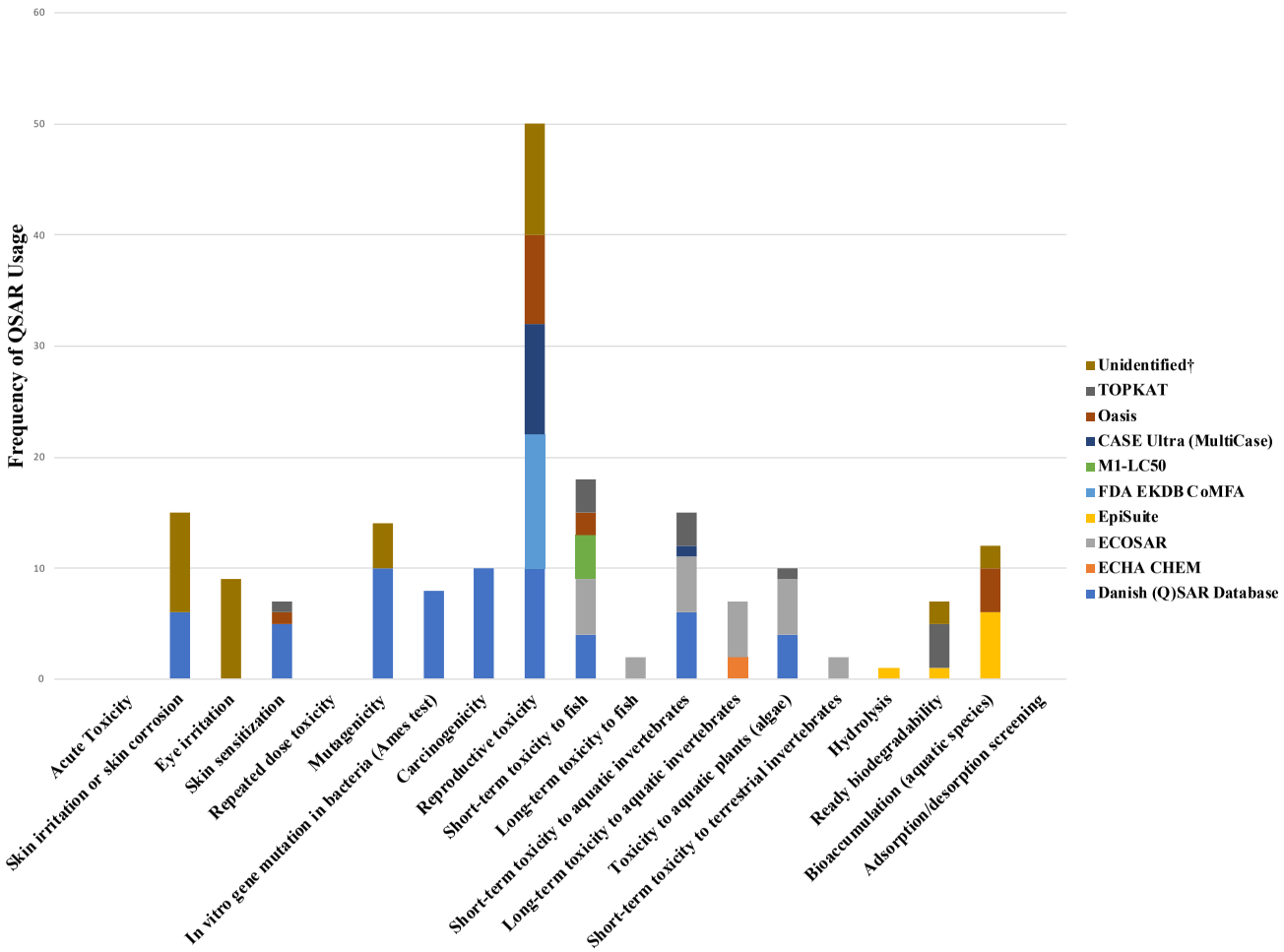


Figure 3-2. Frequency of QSAR usage for human health and environmental hazard endpoints.

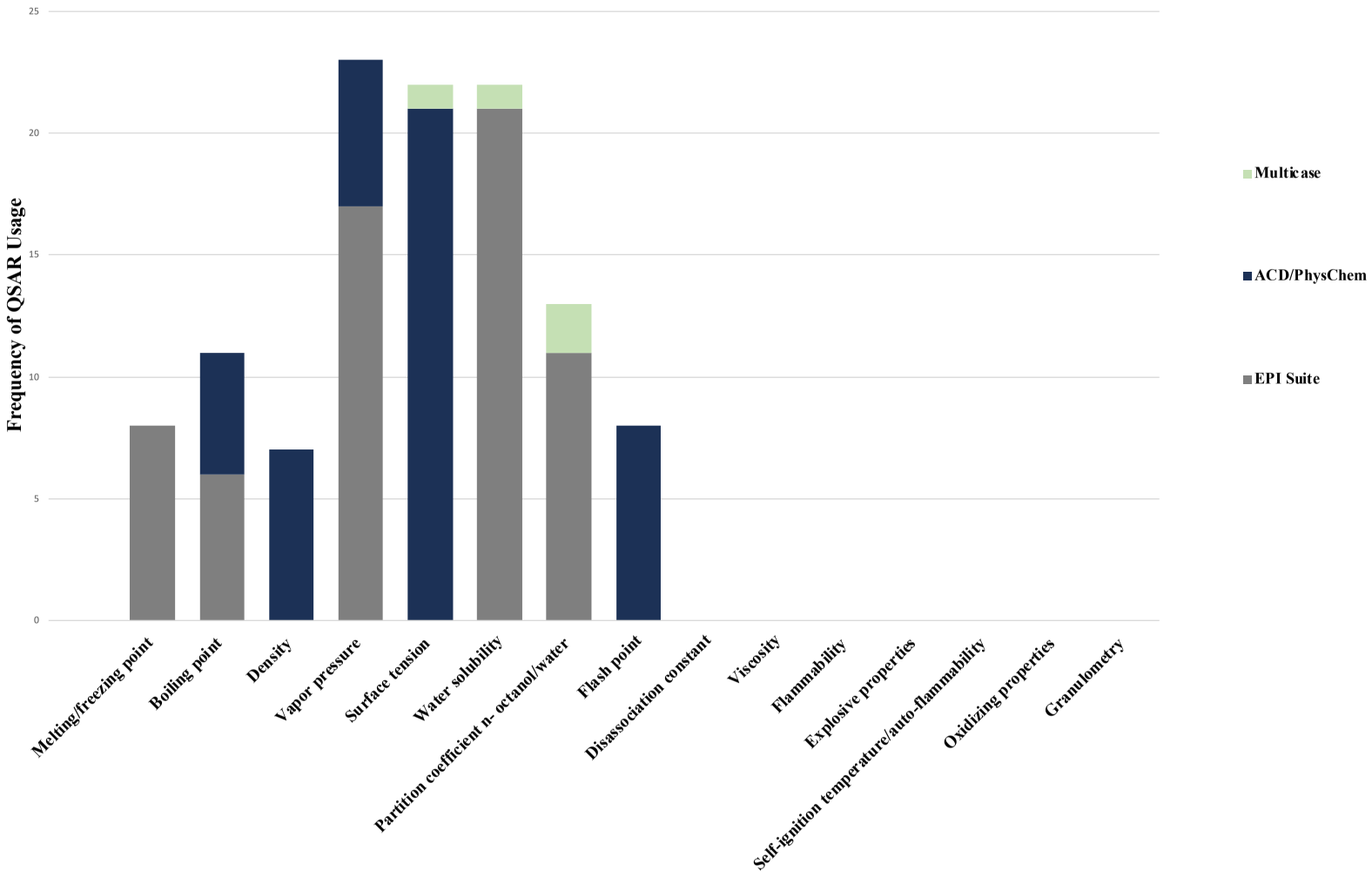


Figure 3-3. Frequency of QSAR usage for physico-chemical properties.

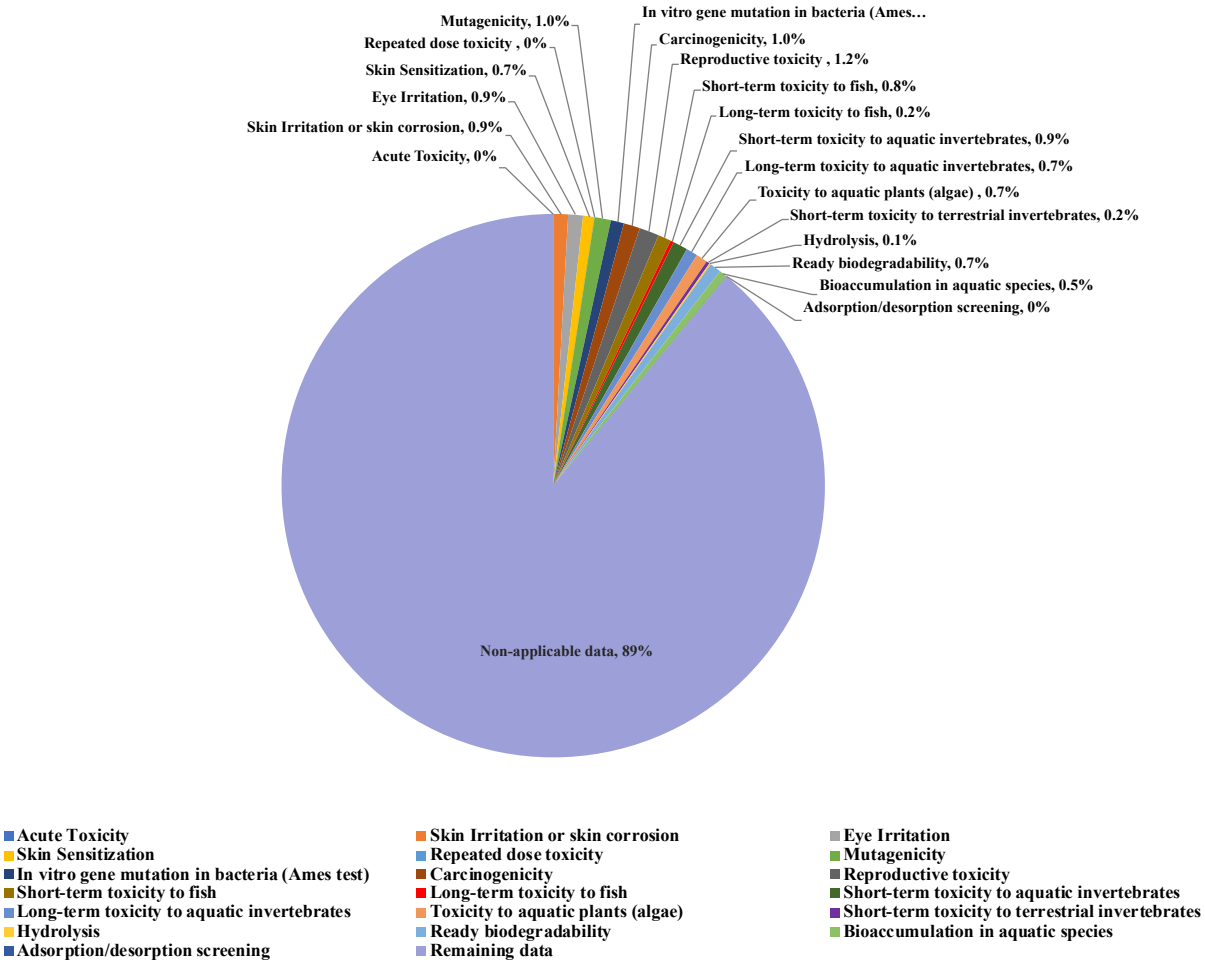
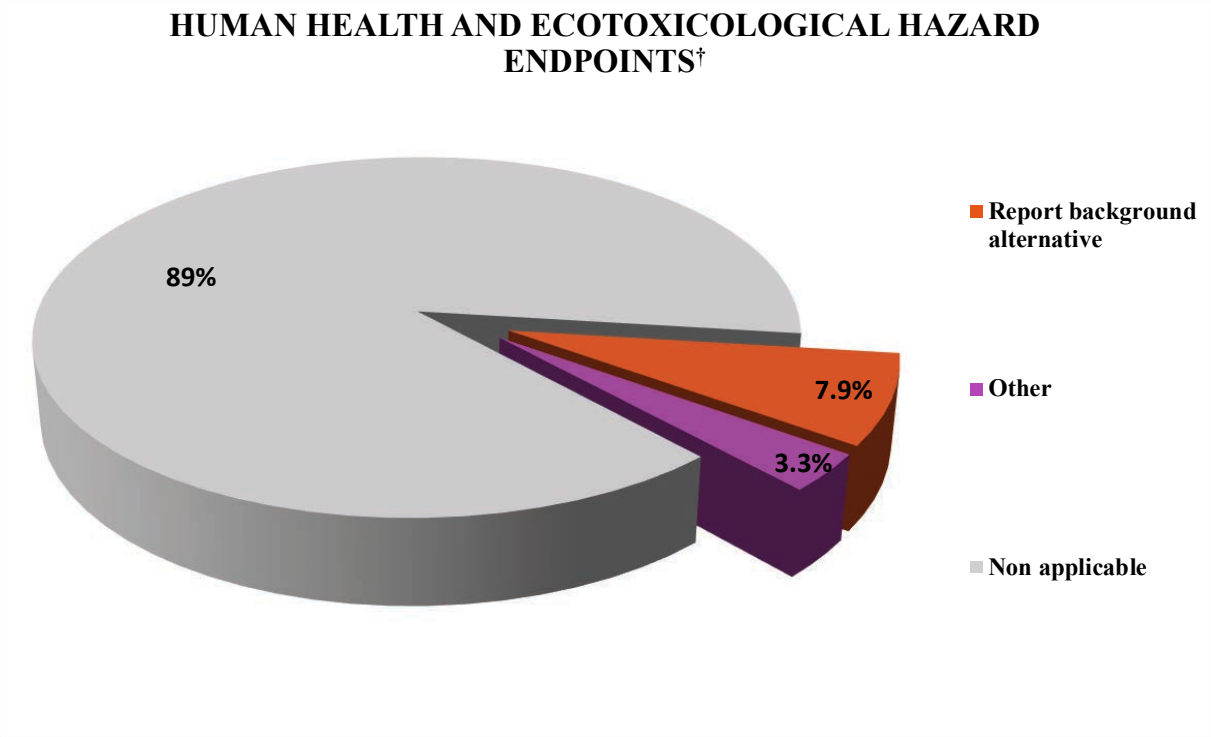


Figure 3-4. QSAR predictions for hazard endpoints by number of alternatives.



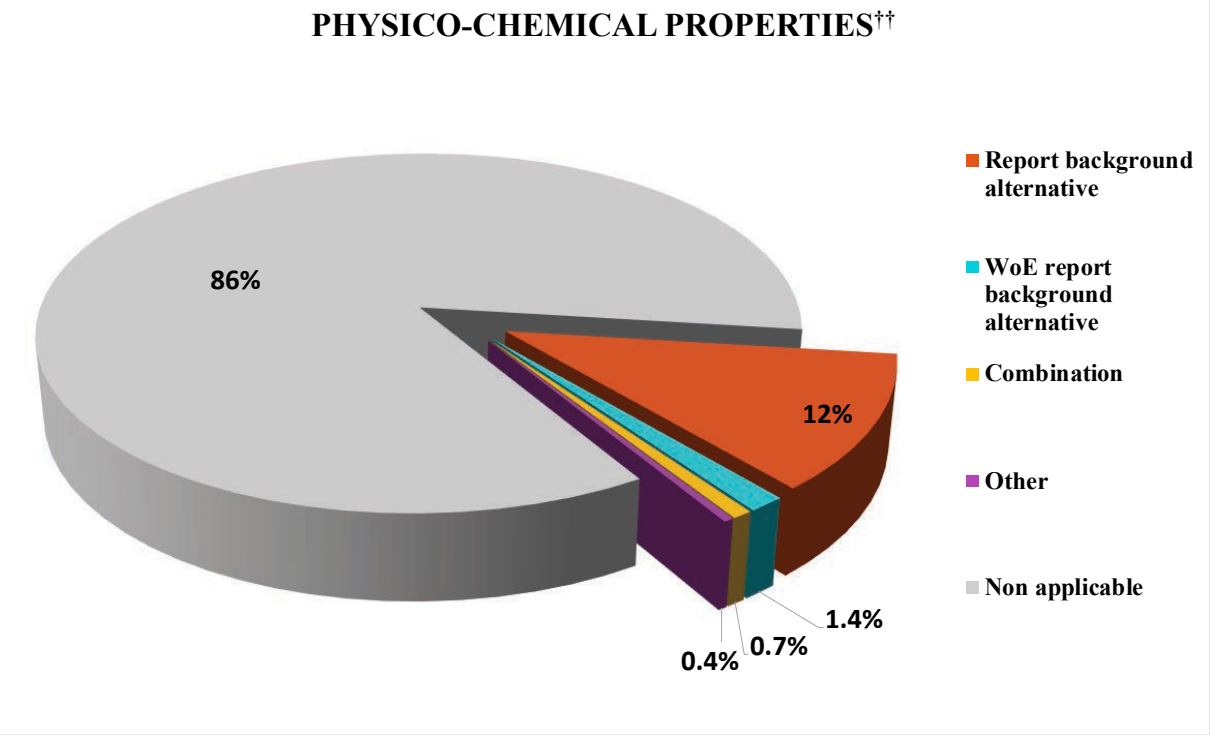
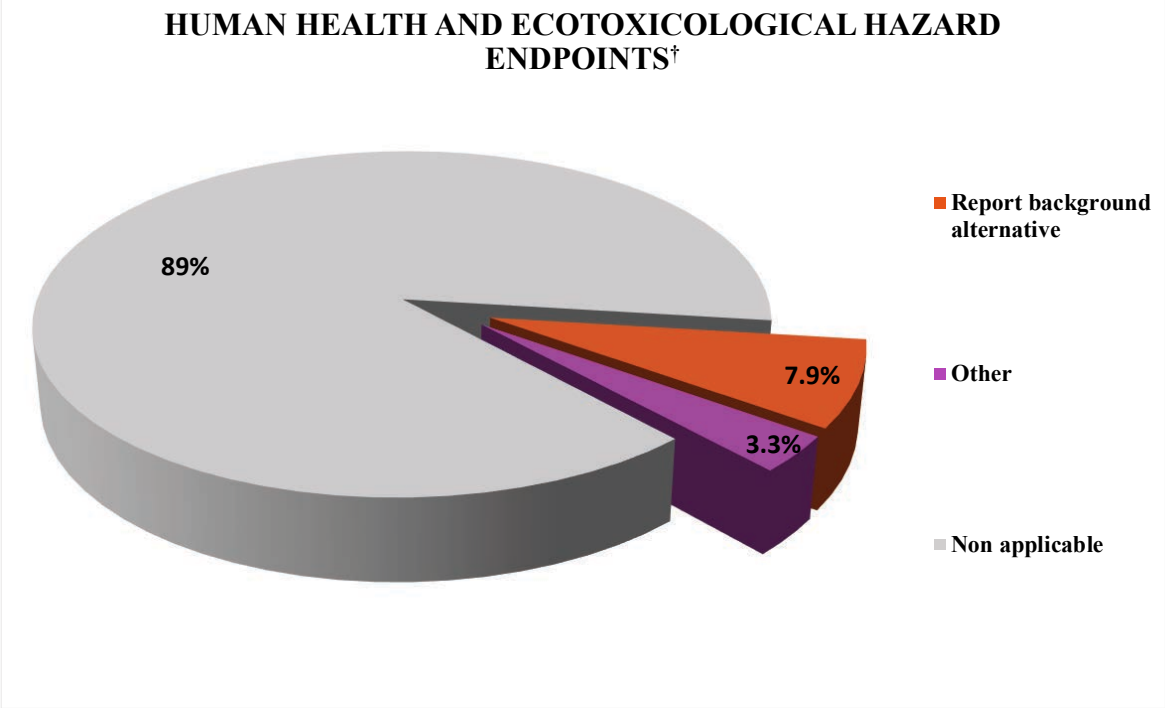


Figure 3-5. Percentages of AoA QSAR usage for the purpose of eliminating an alternative based on 19 human health and ecotoxicological hazard endpoints, and 15 physico-chemical properties.



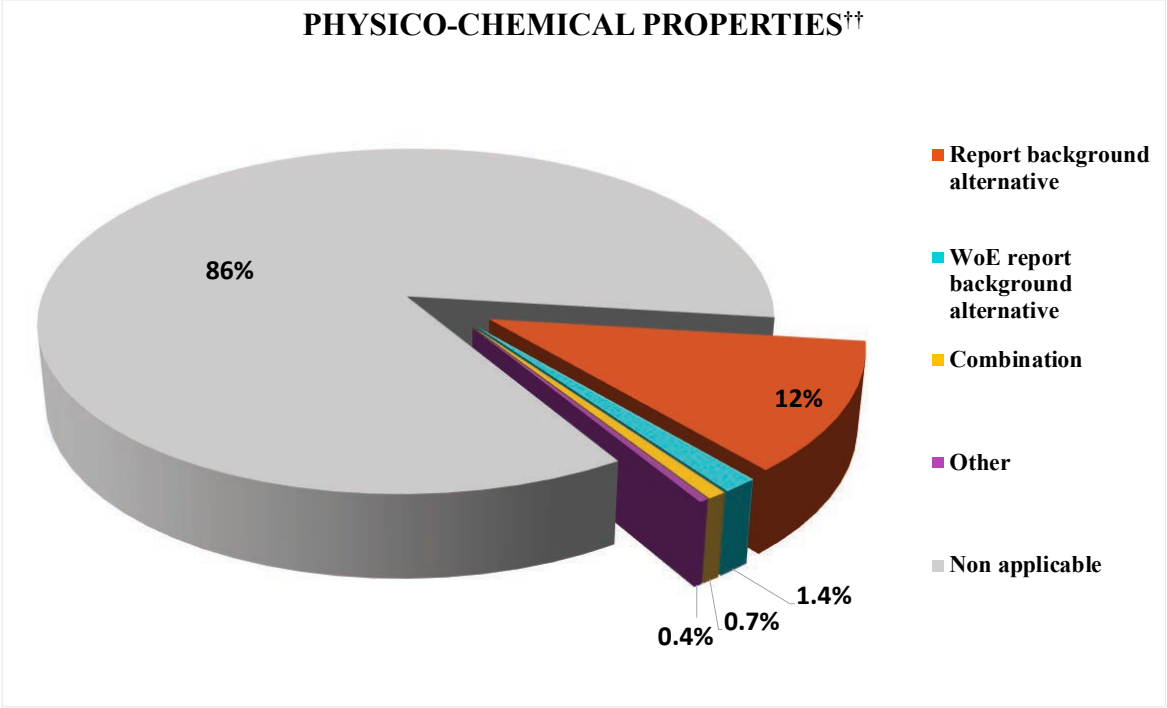


Figure 3-6. Percentages of AoA QSAR usage for the purpose of supporting an alternative as more benign than the Annex XIV chemical based on 19 human health and ecotoxicological hazard endpoints, and 15 physico-chemical properties.

Supporting information

Additional information on the REACH process (Appendix S1), the QSAR "default" list (Appendix S2), data collection methods and assumptions (Appendix S3), units of analysis (Appendix S4), background on MOP and units of analysis (Appendix S5), non-applicable category definition (Appendix S6), QSAR model/source descriptions (Appendix S7), applicability domain data (Appendix S8), and study limitations (Appendix S9).

Appendix S1

REACH process

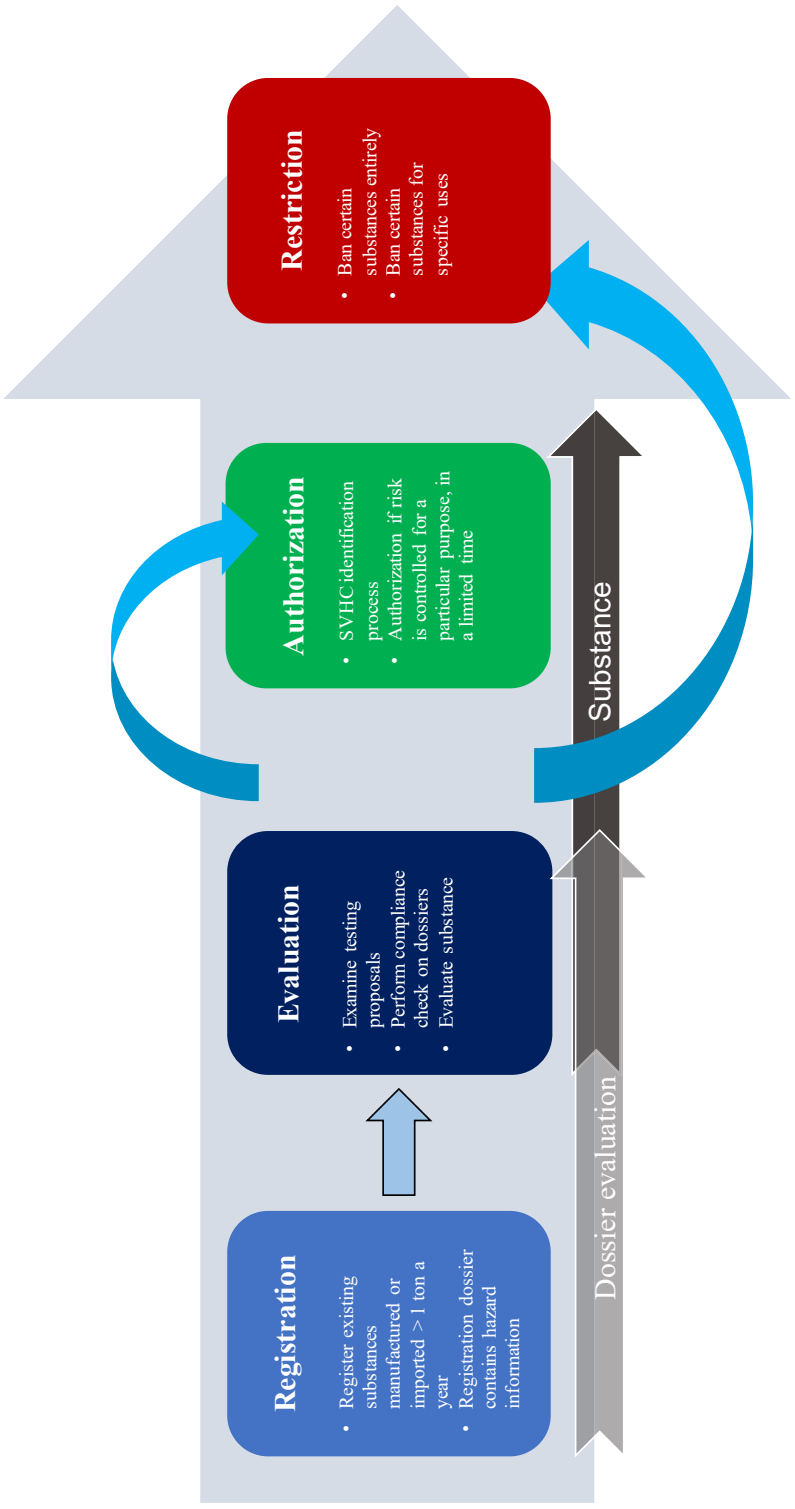


Figure 1S. Overview of REACH process (ECHA 2019d).

Appendix S2

QSAR “default” list

Table 1S. QSAR “default” list

[a] The only QSAR-related endpoint for Sarah Nexus is in vitro mutagenicity. However, the QMRF for In vitro mutagencity stated that the QSAR was built under Derek Nexus. Thus, we put the In vitro mutagenicity model under Derek Nexus.
[b] Leadscope® does not advertise its capability of estimating REACH endpoint "Repeated Dose Toxicity" on its official website but it is mentioned in 2016 ECHA QSAR report.
[c] This series of QSAR models for REACH endpoint "Short-term Aquatic Toxicity to Fish" were housed in the OECD QSAR Toolbox. We could only find information regarding developer and its function in the toolbox.
[d] The capability of OECD QSAR Toolbox was simplified because 1) it is a repeat of many listed QSAR applications, and 2) the OECD QSAR Toolbox does not categorize all QSAR sources under tab "QSAR," which limited our ability to confirm and pair all QSAR sources with REACH physico-properties and hazard endpoints.
[e] Because TOPS-MODE is an approach and not a QSAR platform, we listed the developers of the two TOPS-MODE QSAR QMRFs in the JRC database regarding Ames Test Mutagenicity.

[1] (ACD/Labs 2019)
[2] (EC JRC EURL ECVAM 2019)
[3] (Simulations Plus 2019)
[4] (DTU FOOD 2019)
[5] (DTU FOOD 2018a)
[6] (Lhasa Limited 2019a)
[7] (Lhasa Limited 2019b)
[8] (U.S. EPA 2012c)
[9] (U.S. EPA2019b)
[10] (U.S. EPA 2019e)
[11] (U.S. EPA 2012a)
[12] (U.S. FDA 2018)
[13] (Shi et al. 2001)
[14] (Hong et al. 2003)
[15] (KREATiS 2019)
[16] (ChemAxon 2019)
[17] (ChemAxon 2014)
[18] (Viswanadhan et al. 1989)
[19] (Hou et al. 2004)
[20] (Maunz et al. 2013a)
[21] (Maunz et al. 2013b)
[22] (Leadscope® 2019a)
[23] (Leadscope® 2019b)
[24] (MultiCase 2019)
[25] (Molcode 2019a)
[26] (Molcode 2019b)
[27] (OECD 2019)
[28] (2019)
[29] (U.S. EPA 2012b)
[30] (Pudenz and Frère 2017)
[31] (Dassault Systemes Biovia 2019)
[32] (U.S. EPA 2019c)

[33] (U.S. EPA 2016)
[34] (TerraBase Inc. 2019)
[35] (Estrada 2008)
[36] (IdeaConsult Ltd. 2018)
[37] (IRCCS 2019)

Appendix S3

Data Collection Methods and Assumptions

For this study, we used two approaches: 1) a consultation number by alternative approach, on which we based the maximum opportunities (MOP) approach, and 2) an alternative by QSAR (ABQ) approach that captures the complete number of QSARs in the 189 AoAs. More precisely, for the ABQ approach, we collected data on QSAR sources per consultation number per alternative, and then grouped the sources by physico-chemical property or hazard endpoint. For example, for every AoA, we recorded each alternative by line item, which included the QSAR source. If more than one QSAR source was used to make a prediction for an individual alternative, we included multiple line items of QSAR sources for the alternative. Unlike other regulatory approaches, such as ECHA's endpoint study records (ESR) approach for registration dossiers (ECHA 2014, 2017a), our unique approach tracks key information from individual applications as well as individual QSARs and how they are used. However, the 54 alternatives in our sample are not unique. In fact, we recorded the same alternative(s) for multiple consultation numbers.

Though our unique approach aims to examine how applicants used QSARs to fulfill regulation requirements, part of the process involved making key decisions and assumptions. We first excluded alternative processes that did not include a chemical substitution, even if the process used chemicals. We assumed that QSAR predictions would not be generated in the absence of a chemical substitution. In cases involving inconsistent reporting of QSAR predictions from applicants or reporting of 'no information,' we assumed that QSARs were applied. In some cases where applicants simply reported that results were within a QSAR's applicability domain, we assumed that a QSAR source was used to generate this prediction, and

therefore, included these QSAR predictions in our QSAR test totals. However, if "no information" was listed under a hazard endpoint, (i.e. reproductive toxicity) and "no information" listed under a sub-category, (i.e. developmental/teratogenic toxicity), we only counted a QSAR prediction once. Additionally, if an applicant reported that a prediction was the product of multiple QSARs, but only identified one QSAR source, we counted the identified QSAR source. Due to the variance among AoAs, we also accepted several different phrasings for applicability domain including: "within QSAR domain," "outside of QSAR domain," "undefined with regard to domain applicability," "operational limits," and "model's rules".

It is worth mentioning that some QSAR models are packaged within integrated software platforms incorporating a range of modeling and expert systems applications. When feasible, we singled out QSAR model(s) under these platforms. For example, we coded Food and Drug Administration (FDA) Endocrine Disruptor Knowledge Database (EDKB) Comparative Molecular Field Analysis (CoMFA) QSAR predictions as FDA EKDB CoMFA to specify the QSAR model within the platform (Tong et al. 2002). We also reclassified EPI Suite™ software and expressions of QSAR predictions to fit our study. Because EPI Suite™ software had several different user interfaces such as KOWIN™ or WSKOWWIN™, we grouped all interfaces under the EPI Suite™ software (EPA2019a). In addition, because EPI Suite™ employed the Modified Grain method (MGM) to estimate vapor pressure (Barley and McFiggans 2010), we assigned all vapor pressure QSAR predictions that cited MGM, but did not indicate a QSAR source, to EPI Suite™. For all QSAR software output, we used expressions of QSAR estimation (Cronin et al. 2003) and QSAR prediction interchangeably.

In order to include only relevant QSAR information , we applied limits to our AoA content analysis. All QSAR data for physico-chemical or hazard endpoint data in the AoA's

Appendix counted toward our data collection. If AoAs lacked alternatives, we excluded them from our in-depth QSAR analysis. Any AoAs that were assigned 'withdrawn' at the time of our study, were coded as "unavailable," and were excluded from our QSAR analysis. However, in order to obtain a robust sample and comprehensive understanding of practices used in REACH applications for authorizations as of May 2017, we based our total sample size on the 189 original applications.

Finally, to prioritize the identification of specific models over larger platforms and to control for falsely inflated frequency values, we did not include descriptive statistics specifically labeled as OECD QSAR Toolbox in our analysis. Our first justification for doing this was because an applicant oftentimes cited both the OECD QSAR Toolbox and the specific QSAR model for the same QSAR prediction. We thus decided to only include the tallies for the specific QSAR model even though we kept track of both specific models and OECD QSAR Toolbox counts. Alternatively, an applicant may have identified the OECD QSAR Toolbox as their QSAR source but then failed to cite the exact model. Upon locating the descriptor "prediction" within the applicant's explanation, which we used to justify that the value was actually a QSAR prediction, we then coded these types of values as "unidentified QSAR." However, we also coded for OECD QSAR Toolbox references in the event that we needed this information though we excluded OECD QSAR Toolbox data that did not cite a specific QSAR to be conservative. As the QSAR Toolbox hosts a variety of platforms, models, and information sources, we could not always confirm whether the value was a QSAR prediction or not.

Appendix S4

Details on the Consultation by Alternatives Unit of Analysis

For the total count of alternatives unit of analysis, we consolidated data that we initially collected by consultation number per alternative per QSAR source because we could only include one count per alternative. A simple decision-tree analysis approach was used to reduce the unit of analysis to consultation number per alternative.

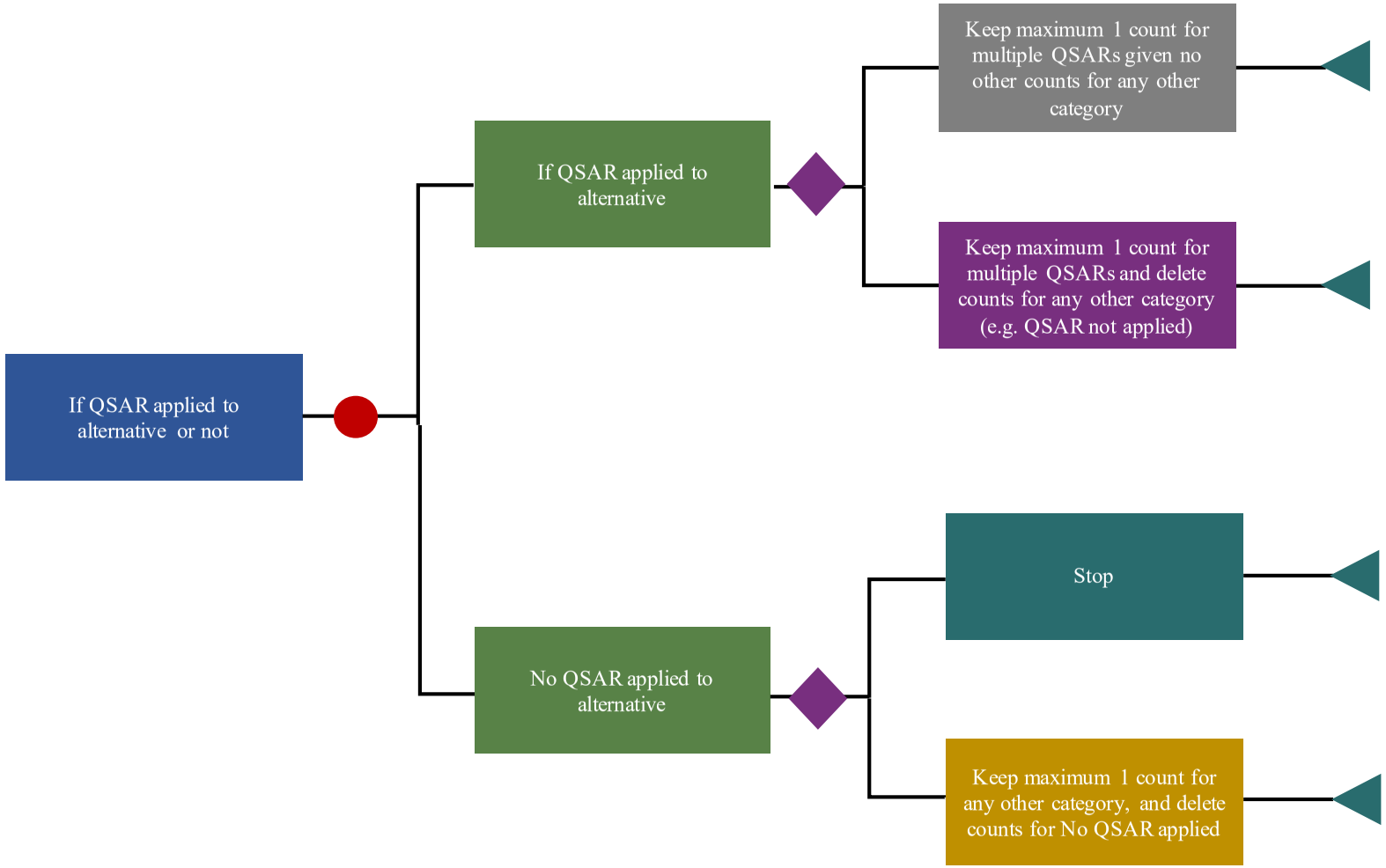


Figure 2S. Decision tree analysis for reducing unit of analysis to one count per consultation number per alternative.

Appendix S5

MOP

Total QSAR prediction counts included predictions that listed their QSAR models or sources, no information outputs with unspecified sources, and QSAR predictions, which could only be identified through references to their applicability domain.

To compare data, we "normalized" counts using the MOP approach to reduce bias for otherwise potentially misleading tallies, such as low counts. For instance, a low count could have been the result of a model's limited endpoint range, which, in turn, may have limited its total counts. As an example of an MOP calculation, we multiplied the total number of alternatives that cited QSARs (n = 54) within each AoA, and then multiple this amount by the total number of hazard endpoints (n = 19) or physico-chemical properties (n = 15),

Equation 1.

$$Maximum\ opportunities\ (MOP) = (number\ of\ endpoints\ or\ properties) \times\ (total\ alternatives)$$

Equation 2.

$$1{,}026 = (19) \times\ (54)$$

To calculate the percentage of MOP, we then divided the total amount of QSAR predictions per endpoint, for example, by its related MOP and then multiplied by 100,

Equation 3.

$$\left(\frac{8}{1{,}026}\right) \times 100\ =\ 0.8\%$$

In addition, for all calculations, we used either one of two basic units of analysis: *a)* alternatives by QSAR tool/model (sample number varied per AoA) served as the unit of analysis

when analyzing trends in hazard endpoints or physico-chemical properties by QSAR tool/model use or *b)* total count of alternatives (n=54) when examining data trends in QSAR use and support, and in certain cases, for endpoints and properties, which did not pertain to QSAR tools/models.

Appendix S6

Non-applicable Category

In general, most data points fit under the "non-applicable" category, (i.e. QSAR prediction data did not exist for that endpoint or we could not confirm if a certain value was in fact a QSAR prediction). For example, under QSAR usage, we calculated non-applicable data points for 86% per MOP and 89% per MOP for both physico-chemical properties and hazard endpoints, respectively).

Appendix S7

QSAR Model/Source Descriptions

The Danish (Q)SAR Database (n=63), a repository of over 600,000 estimates from more than 200 QSAR models that generates "battery" predictions from three QSAR systems (DTU 2018b), and unidentified QSARs, or, QSARs for which the applicant did not specify the QSAR model or tool, ranked second in frequency (n=36) were the most cited QSAR sources for human health and environmental hazard endpoints. ECOSAR (n=24), which is a library of 711 QSARs and which uses decision-tree analysis to predict aquatic toxicity (Mayo-Bean et al. 2012) was the third most cited QSAR source. In descending order, OASIS, EKDB CoMFA, TOPKAT and CASE Ultra (Multicase) were the next four most cited QSARs. OASIS or the Laboratory of Mathematical Chemistry OASIS (n=15), includes software suites and models for environmental fate and ecotoxicity endpoints (CATALOGIC) and human health endpoints (TIMES) (OASIS 2019). The FDA's National Center for Toxicological Research (NCTR) project's three-dimensional EKDB CoMFA QSAR method (n=12) predicts receptor binding affinity (Tong et al. 2002). TOPKAT (n=12) or (Toxicity Prediction by Komputer Assisted Technology) relies upon 2-D descriptor QSAR models available in the BIOVIA Discovery Studio predictive science application (Pudenz and Frère 2017). Finally, CASE Ultra (MultiCase) (n=11) includes statistical and expert rule-based systems with both alerts and statistics (MultiCase 2019).

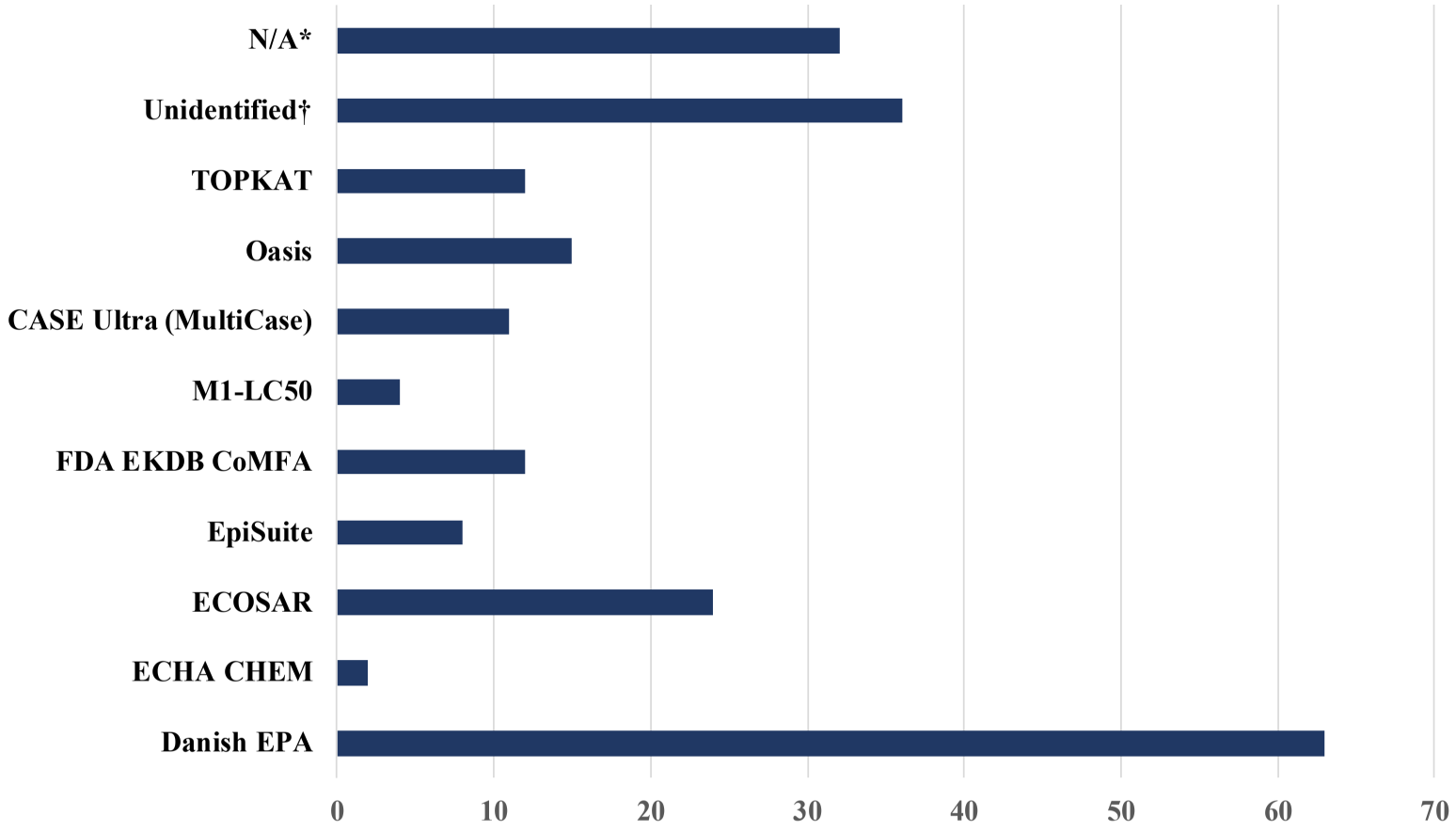


Figure 3S. Comparison of the frequency of QSAR sources used.

† Unidentified means that the identify of the QSAR model was not reported in the AoA even if the QSAR result was provided.

Table S2. Applicability domain of QSAR predictions for 54 alternatives in 24 AoAs.

Human Health and Environmental Hazard Endoint	QSARs applied[+]	Applicability Domain		
		AD information provided	No AD information provided	Combination of information[a]
Acute Toxicity	-	-	-	-
Skin Irritation or skin corrosion	15	15	-	-
Eye Irritation	9	9	-	-
Skin Sensitization	7	5	2	-
Repeated dose toxicity	-	-	-	-
Mutagenicity	14	14	-	-
In vitro gene mutation in bacteria (Ames test)	8	8	-	-
Carcinogenicity	10	10	-	-
Reproductive toxicity	50	35	10	5
Short-term toxicity to fish	18	8	10	-
Long-term toxicity to fish	2	-	2	-
Short-term toxicity to aquatic invertebrates	15	7	8	-
Long-term toxicity to aquatic invertebrates	7	-	7	-
Toxicity to aquatic plants (algae)	10	4	6	-
Short-term toxicity to terrestrial invertebrates	2	-	2	-
Hydrolysis	1	-	1	-
Ready biodegradability	7	-	7	-
Bioaccumulation in aquatic species	12	-	12	-
Adsorption/desorption screening	-	-	-	-

(Table S2. Continued)

[a] 'Combination of information' means that we "combined" AD information when an applicant provided multiple AD information for a single alternative in their AoA. For instance, in AoA 0005-02, the applicant DEZA a.s. generated multiple reproductive toxicity QSAR predictions with different AD interpretations for the following alternatives: Akardite II, Akardite III, DEHA, ATBC and IDP. Within each "combined" data point, at least one of the AD's met our AD criteria by including the wording: "within QSAR domain," "outside of QSAR domain," "undefined with regard to domain applicability," "operational limits," and "model's rules".

Appendix S9

Study limitations

Example of an AoA with redacted language: AoA consultation number 0004-02 (DEZA A.S. 2013b).

Examples of AoAs that did not directly cite after their reported values: AoA consultation numbers 0062-02 and 0066-02 .

Examples of AoAs with confidentiality protocols that prohibited a complete hazard analysis on their product(s): AoA consultation numbers 0063-02 and 0066-02 (Gentrochema BV n.d.; MTU Aero Engines AG n.d.).

Example of an AoA that listed EPI Suite™ as their software package: AoA consultation number 0080-01 (H&R Ölwerke Schindler GmbH 2016; OECD 1997).

Examples of QSARs not captured in our sample of AoAs: AoA consultation number 0005-02 (DEZA A.S. n.d.-a).

- MITI (for the alternatives methyl centralite, ethyl centralite Akardite I, isodecyl pelargonate (IDP): Because applicants did not provide enough information to confirm whether the reported values identified as 'Biodegradation' (by MITI) came from a MITI QSAR model or from the MITI-I screening test, which measures biological oxygen demand (BOD) (Pavan and Worth 2006), these values could not be counted towards our study's QSAR prediction counts. Moreover, because the 'MITI' QSAR model exists as part of BIOWIN5™ and BIOWIN6™ (Posthumus et al. 2005), we would have grouped 'MITI' under EPI Suite™ regardless if we knew MITI specified the QSAR model.
- GOBAS (for the alternatives methyl centralite, ethyl centralite Akardite I, isodecyl pelargonate (IDP): In addition, we did not capture biodegradation predictions

identified as ‘GOBAS’ under GOBAS QSAR model, but instead, incorporated these predictions under EPI Suite™. GOBAS BCF and BAF models are presently part of the EPI Suite™ BCFBAF™ model (Garg and Smith 2014). However, we could not find out exactly when GOBAS became part of BCFBAF™, thus, for consistency across all AoAs, we grouped GOBAS QSAR predictions under EPI Suite™.

Example of a QSAR designed for a specific hazard endpoints: FDA EDKB CoMFA. The FDA EDKB CoMFA, which solely predicts receptor binding affinity, ranked as one of the least cited QSAR tools in our study. Because our descriptive statistics were based on frequency of use, we were unable to completely control for this design limitation, however, we attempted to reduce biased interpretation of results by using percentages based on MOP.

4. Chapter 4: Exploring QPRF, WoE and ITS aspects of QSAR use in REACH AoAs

Introduction

Under the Registration, Evaluation, Authorization and Restriction of Chemicals (REACH) analysis of alternative's (AoA) process, quantitative structure-activity relationship (QSAR) models play an important role in information gathering and organizing frameworks. Already recognized as an alternative to testing under registration (ECHA 2014, 2016c, 2017a, c), QSARs have become increasing relevant tool in bridging data gaps and supporting weight of evidence (WoE) when assessing alternative substances (ECHA 2016b). Additionally, QSARs are growing in importance in integrated testing strategies (ITS) (Bassan and Worth 2008; Luechtefeld et al. 2018). For example, the REACH ITS framework for specific endpoints directs registrants to consider non-testing results, including QSAR predictions, when deciding if further animal testing is needed (ECHA 2017d). Despite the rising profile of QSARs within these frameworks, a gap exists in the evaluation of QSAR use and QSAR documentation under authorization (pending publication Chinen and Malloy 2019). Thus, an assessment of the different channels (e.g. WoE and ITS) by which QSAR predictions play a role in evidence gathering and organizing remains unaddressed for AoAs. For this study, we approached this disparity in information for QSAR predictions by conducting a substantive review of 24 AoAs through May 2017, which contained higher-tier endpoints under REACH. Understanding the manner in which applicants manage QSAR prediction information in AoAs and assess their potential within ITS will be valuable in promoting regulatory use of QSARs, and building out future platforms in the face of rapidly evolving technology.

Background

Quantitative structure activity-relationship (QSAR) models provide vital information for untested substances that lack or have limited experimental data due to the resource intensive costs associated with traditional animal testing (Benfenati et al. 2011) and act as an increasingly relevant tool under Registration, Evaluation, Authorization and Restriction of Chemicals (REACH) to bridge data gaps and support weight of evidence (WoE) when assessing alternative substances (ECHA 2016a). Moreover, QSARs have been gaining importance in integrated testing strategies (ITS) (Bassan and Worth 2008; Luechtefeld et al. 2018). QSARs are computational models that mathematically relate chemical structure to biological activity (Benfenati 2012). Yet, as more 21st century toxicological methods, such as QSARs, are integrated into information gathering and organizing frameworks for progressive regulations such as REACH, limited evaluations exist that evaluate how QSARs are used within these frameworks for the authorization process (pending publication Chinen and Malloy 2019). Proper supporting information and appropriate testing, which includes alternative methods such as QSARs, are especially critical for chemicals prioritized as hazards of highest concern (ECHA 2017c; ECHA 2019b). Without rigorous monitoring of these frameworks, particularly those which draw on QSAR predictions for authorization reports such as analyses of alternatives (AoAs), these highly toxic chemicals may circulate in the market without proper labeling (EPA 2009; EPA 2013) or safer substitution.

AoAs, which cite QSAR predictions and lack the required QSAR documentation, are one such example of the risks to proper assessment of alternative substances under authorization. AoAs are applications that companies submit under REACH for continued use of an priority Annex XIV substance (ECHA 2011b). Part of the purpose of AoAs is to examine whether safer alternatives exist relative to the Annex XIV substance. Yet, without background information on

the QSAR prediction(s), traditionally found in the QSAR Prediction Report Format (QPRF) to justify the reliability or adequacy of the prediction for a specific chemical by a specific model (ECHA 2008; OECD 2007; Rorije et al. 2008; Worth et al. 2011), an applicant's claims that a QSAR prediction is reliable or not cannot be verified (ECHA 2008). These risks associated with inconsistent monitoring of proper QSAR reporting and presentation in information frameworks under authorization has serious potential implications similarly found in other data organizing frameworks, which utilize QSARs. Two major information organizing frameworks that draw on combined results, but may include QSARs, and which can either be found in AoAs or have the potential to significantly impact AoAs are weight of evidence (WoE) and integrated testing strategies (ITS). Under REACH, WoE is an important approach for drawing reliable conclusions on a substance based on a collection of toxicity information with emphasis on avoiding unnnecessary animal testing. According to ECHA, WoE is a process that combines multiple lines of evidence while weighing the relative "strengths and weaknesses" in order to reach a conclusion on a substance's property (ECHA 2016a). For example, results from predictive non-testing methods, such as QSAR and read-across, could be combined to draw inferences on the hazard of a substance. (For an example of a formal WoE, see Appendix S1). WoE has also been used to develop ITS (ECHA 2016a). ITS is a method of collecting and combining results under a scientific approach (Hartung et al. 2013). Advancements in *in silico* models, such as QSARs, have enhanced the accuracy of predictions to even exceed animal tests (Luechtefeld et al. 2018). As a result, many tests are being combined in ITS as the next step forward in regulatory toxicologist assessment. Though ITS does not have an inherent regulatory purpose, ITS is used for regulatory decision-making, oftentimes in a WoE context (Worth 2010a). As an alternative method, QSARs have expanded under WoE and ITS methods, though not without potential challenges. Without clearer documentation and a more systematic WoE framework (Knudsen et

al. 2015; Malloy et al. 2017; Rotella 2011), information contained within WoE and ITS, such as QSAR predictions, may not be perceived as trustworthy information.

As a regulatory tool, WoE can be implemented according to a variety of approaches. These approaches include a range of methods from narrative to more systematic and quantitative types (Linkov et al. 2009; Martin et al. 2018). However, the degree of reproducibility and transparency differ significantly between the evaluation forms even though the different types of evaluations are equally considered under WoE (Linkov et al. 2009). For QSAR predictions in an AoAs, the high variability in WoE approaches may weaken the strength of supporting evidence, particularly if the supporting evidence is not well documented for QSARs. WoE is limited though; this framework may not always give the best outcome. In some cases, even if an applicant's WoE presents an articulate, well documented argument with a non-integrated battery of QSAR predictions, an applicant may still select the wrong QSAR model or approach, or, a non-integrated combination of predictions, which could include false positives. An integrated QSAR screening approach, on the other hand, which combines predictions based on algorithms, produces more powerful results. However, for AoAs, the potential benefits of using QSARs especially for priority higher-tier chemicals in an integrated testing strategies (ITS) framework, which combines predictions based on algorithms to produce more powerful results, remains unknown.

Substances with carcinogenic, mutagenic, reproductive toxicity (CMR), persistent, bioaccumulative and toxic/ very persistent and very bioaccumulative (PBT/vPvB) and certain endocrine disrupting properties, which are considered of higher concern under REACH (ECHA 2019c), are a likely target of WoE and ITS strategies especially if these substances are deficient in experimental data. Under REACH, QSARs within a WoE context have been increasingly used

to identify potential CMR and PBT/vPvB substances (ECHA 2016c). Some information requirements for higher-tier endpoints under REACH even point towards the use of QSARs to identify targeted substances (European Parliament and the Council of the European Union 2006; Lahl and Gundert-Remy 2008). The REACH ITS framework for specific endpoints, which includes CMR, also directs registrants to consider non-testing results, (e.g. QSAR predictions) when deciding if further animal testing is needed (ECHA 2017d). Despite their critical role in information gathering and testing assessments for CMR and PBT/vPvB substances, QSARs in AoAs have yet to be evaluated for these purposes, which could eventually become an issue for regulators attempting to deter companies from using harmful chemicals in their processes such as the European Chemicals Agency (ECHA), the regulatory arm of REACH.

Prompted by findings in several AoAs in a previous study, this study was used to address three main questions pertaining to QSAR information is reported in QPRFs, WoE and ITS: 1) To what degree do our sample AoAs contain information from the QPRF document? 2) How well do applicants meet regulatory and best practices WoE criteria when using QSARs in AoAs? 3) What are the differences in conclusions on toxicity between ITS battery QSAR model predictions and QSAR predictions generated from either individual models or multiple models, which do not employ ITS battery testing. We first data-mined 24 AoAs to see if applicants used an equivalent method to QPRFs by providing at least partial information. We prioritized this information because our research indicates this information could reveal the applicant's reasoning for using QSAR predictions. However, because we were also curious how WoE using QSAR predictions in AoAs were used to assess priority CMR/ PBT vPvB endpoints, we developed a checklist of criteria to assess the rigor of WoE analyses, or, what we call a 'completeness review' for our AoA sample (n=24). As the guidance document and the relevant

research were individually inadequate for our purposes, we combined four approaches to create a composite set of criteria. This checklist draws on ECHA WoE guidance as well as recent WoE literature and other best practices. We then assessed the degree to which an applicant met our WoE criteria in a completeness review of AoAs using QSAR predictions for higher-tier endpoints.

In cases where two or more AoAs assessed the same chemical, we adopted a consistency check review to see if applicants reached inconsistent conclusions. If the AoAs using QSAR predictions on the same chemical arrived at different conclusions, we examined the substantive aspects of the respective WoE analyses to understand why the analyses came out differently. Because of the importance placed on higher-tier endpoints, we examined WoE only in the context of CMR and PBT/vPvB endpoints. Our research underscores what we believe a regulator would need to make a sound decision on an alternative's relative safety in an AoA. Finally, to be forward looking, we added an integrated approach to QSAR use, when appropriate, in our paper. We compared the conclusions reached by applicants on the safety of the alternative relative to the Annex XIV chemical with results from the Danish EPA advisory list (EPA 2018b). We argue that the Danish EPA's use of ITS for further evaluation is a potentially valuable tool to confirm the identity of potentially harmful CMR chemicals.

We report findings regarding the breadth and context of QSAR usage in 24 REACH AoAs in addition to the role of QSARs in 21st toxicology strategies to increase our knowledge on how QSARs are used in information gathering and organizing frameworks. We conclude with a discussion on the need for enforcement of QSAR documentation under REACH authorization, guidance specifying the WoE criteria that should be met when using QSARs, and the future

direction of QSAR ITS modeling strategies within REACH. (For study limitations, see Appendix S2).

Methods

QPRF criteria

We collected data on 17 criteria (Table 4-1) from AoAs to investigate whether applicants provided information normally required under a QPRF. Since we were looking for results that had the most relevance in regulatory discussions, we focused on QSAR predictions for hazard endpoints. For substance information, we surveyed chemical identifiers, coding for Chemical Abstracts Service (CAS) number, European Community (EC) number, chemical name, structural formula, and structure codes, (i.e. text representations of a chemical's structure). For prediction information, we selected model identifiers such as the model's hazard endpoint and dependent variable (ECHA 2008), model name and version, predicted value including prediction cut-off values, input for the model used to generate the prediction such as the specific structure codes, and descriptor values, which codes molecular value into numerical value (Todeschini and Consonni 2008). We also focused on six QPRF "priority" criteria: *a)* applicability domain; *b)* structural analogues; *c)* predicted value; *d)* model endpoint; *e)* uncertainty; *f)* chemical and biological mechanisms to conduct a more in-depth analysis. (Table 4-1). For our study, we define QPRF "priority" criteria as foundational criteria, which provides the most fundamental information a regulator needs to draw a general conclusion on a QSAR prediction. Because we were already aware that all QMRFs were missing from previous research, we did not code for QMRF references in the QPRF. Lastly, we coded for applicability domain, (i.e. the limits used in a model to make reliable estimations) (Hanser et al. 2016), the inclusion and identification of structural analogues found in a model's training and test sets, prediction uncertainty, and if

appropriate, chemical and biological mechanisms, which according to Yuan et al. (2007) is the biological response underlying the mode of toxic action (MOA). To streamline our data, we excluded alternatives that did not employ a QSAR for a hazard endpoint.

During our analysis, we made several assumptions. First, we accepted multiple definitions of applicability domain (see Chapter 2 Methods: Data collection). Regarding uncertainty of QSAR predictions, we also accepted several versions including "acceptable," "limited similarity and no conclusion could be drawn," "doubtful reliability," "uncertain reliability," "robustness of prediction," "considered reliable," and "no conclusion should be drawn." However, we did not accept "no indication that model was operating outside of its operational limits," as this interpretation did not answer the degree to which an applicant considered the prediction trustworthy. For QSAR source/model names, we excluded "OECD QSAR" as a source because we could not confirm whether the applicant ran a QSAR model or if the QSAR prediction was cited from within the OECD (Q)SAR Toolbox, which would then mean the prediction could have been generated from any number of QSAR models. Finally, we accepted partial endpoint definitions.

WoE Completeness review

Our WoE completeness review began with an evaluation of REACH's approach to WoE criteria set out in the 2016 *Practical guide: How to use alternatives to animal testing to fulfil your information requirements for REACH registration*. For the purpose of our study, we define completeness review as a critical evaluation of the process and steps applicants took to formulate their WoE analysis for CMR and PBT/vPvB higher-tier endpoints when including QSAR predictions. Criteria included assembling information that factored in relevance, reliability adequacy and quantity, discrepancies in studies, proper documentation, expert judgement, and

"robust" summaries (Figure 4-1) (ECHA 2016a). However, in developing sub-criteria for this approach to our study, we came across inconsistencies and gaps in ECHA's guidance. In addition, the criterion to "pool" information did not define "pooling," nor were steps given on how to weigh the evidence or what to consider during data integration. Thus, we drew on elements from the 2018 National Resource Council's (NRC) evaluation on the U.S. EPA IRIS system, and best practices from Rhomberg et al.'s 2013 review *A survey of frameworks for best practices in weight-of-evidence analyses*, Martin et al.'s 2018 review *Weight of Evidence for Hazard Identification: A Critical Review of the Literature*, Suter et al.'s 2017 *A Weight of Evidence Framework for Environmental Assessments: Inferring Qualities*, and the European Safety Authority's (EFSA) *Guidance on weight of evidence*. Together, these frameworks and best practices formed a continuous framework that explicitly laid out a clear, transparent and structured WoE approach (EFSA Scientific Committee et al. 2017; Martin et al. 2018; NRC 2018; Suter et al. 2017). (For WoE criteria checklist, see Appendix S3). These supplemental regulatory guidances and best practices addressed three main deficiencies in ECHA's WoE guidance: *a*) creating a broad guidance to apply to other parts of REACH, particularly for study summaries, *b*) providing accepted metrics for "weighing" evidence; *c*) setting out specific steps for data integration. With this continuous framework, we developed a checklist to conduct a completeness review for how "well" AoAs articulated their WoE analysis. In this completeness review for WoE, the evaluation was limited to a procedural analysis and did not address substantive questions related to the quality of higher-tier WoE using QSARs. We also only considered WoE that used QSARs as opposed to all WoE used in an application (Benfenati 2012; ECHA 2016b).

Using this technical approach, we looked at all WoE higher-tier hazard endpoints from our original AoA sample citing QSARs (n = 24), which we collected through May 2017 (pending publication Chinen and Malloy 2019) . More specifically, our study focused on measuring the completeness across WoE using QSARs to gauge the degree to which applicants' efforts met our WoE criteria. (For sub-criteria coding for higher-tier endpoints used in WoE, see Appendix S4). We rated completeness on an increasing scale ranging from 0 = no criteria discussed to 5 = all criteria discussed (Figure 4-1). When endpoint data were not relevant to our WoE completeness review, such as potential alternatives, which do not typically provide sufficient hazard endpoint detail other than the type of data, or, the given information was not used in WoE context, we coded this information as '6' for non-applicable.

We recorded any WoE observations for each of the five main criteria in Libre Office Version: 6.2.4.2. Based on our descriptive statistics, we assessed the degree to which applicants met the criteria for a rigorous WoE analysis. For all descriptive statistics, we used "alternative per AoA consultation number" as the unit of analysis. In this study, there were a total number of 54 opportunities per alternative per consultation number to provide information for criteria.

ITS comparative analysis

Despite the lack of international regulatory consensus on ITS (Rovida 2010), several ITS frameworks, for which QSARs have been instrumental in assessing information, have already been developed. Comparison of the 24 AoAs with integrated CMR predictions from the 2018 Danish EPA advisory list for self-classification of hazardous substances began with a verification of Chemical Abstracts Service (CAS) numbers, which are unique numbers assigned to chemicals used in the science field (ACS 2019) (Table 4-2). (For the 2018 Danish EPA advisory list for self-classification of hazardous substances, see Appendix S5). To verify information, we inputted

the name of the alternative substance, as listed in the AoA's Table of Contents, into SciFinder®'s substance identifier search engine, which generated a profile. We then used the CAS number from the profile to verify if the given CAS number in the AoA correctly identified the alternative. If SciFinder® could not find a match based on the alternative's name, we queried the alternative's CAS number and used the molecular formula to confirm the correct identity of the alternative. If the query did not provide a molecular formula, we used the EC, IUPAC or one of the "other" names, as given in the AoA, to search for a matching CAS number; however, this was the least reliable method for confirming the correct identity of the alternative. Because we had repeating, non-unique alternatives, we verified each alternative by consultation number.

Once we confirmed the identity of each alternative, we compared AoA CAS numbers with CAS numbers in the 2018 Danish EPA advisory list using Excel (Version 16.28) (EPA 2018b). For any matches, we recorded the ITS battery QSAR prediction Advisory classification: Muta. 2 (Suspected of causing genetic defects); Carc. 2 (Suspected of causing cancer); or Repro. 2 (Suspected of damaging fertility or the unborn child) advisory classifications in an Excel spreadsheet.). CMR endpoints are defined under the Danish EPA's battery of model endpoints (Appendix S5).

For AoAs with matching CAS numbers, we visually inspected their applications for any CMR identifiers for the alternative. We focused primarily on the following sections: Mammalian hazard profile, Reduction of Overall Risk, Conclusion on suitability and availability, Comparison of hazards, and the Annex. Any supporting, conflicting or missing information was reported separately.

Results

QPRF criteria

Figure 4-2 gives a broad overview of the 16 criteria. Overall, criteria were unevenly distributed. No information was given in any AoA for the model's dependent variable, cut-off values for the prediction, model input for the prediction, structural analogues, and biological mechanisms. However, all QSAR predictions provided information on the structural formula, International Union of Pure and Applied Chemistry (IUPAC) name, and EC and CAS number criteria of their alternative. For chemical structure codes, (i.e. the remaining criterion), only Simplified Molecular Input Line Entry System (SMILES) codes were provided.

Results for four of the priority QPRF criteria were more evently distributed (Figure 4-2). For the applicability domain priority criteria, predictions were discussed as in domain 259 times while predictions were not discussed in the context of their applicability domains 109 times. While applicants commented on the uncertainty of predictions 266 times, they did not comment on a prediction's uncertainty 100 times. Though predicted values were cited the majority of time (n = 342), we identified qualitative or quantitative predicted values going unreported 26 times. The QSAR model endpoint was defined, at least partially, 285 times, though applicants did not report the model endpoint for 83 predictions. No information was given for either structural analogues or chemical and biological mechanisms.

AoAs that used QSARs in WoE for higher-tier endpoints

Of the 24 AoAs, only three used QSARs for at least one higher-tier endpoint (Table 4-3). These AoAs performed WoE for higher-tier endpoints on 11 unique alternatives. Because data were analyzed by the unit of alternative by consultation number, the same alternative may have

been assessed multiple times by different applicants. For instance, consultation numbers 0005-01 and 0006-1, both assessed the same alternative, diisobutyl hexahydrophthalate (DIBE).

One major issue that we encountered was that several PBT/vPvB QSAR predictions for the alternative Tributyl citrate (TBC) in consultation number 0005-02 had missing QSAR models/sources. When referencing information presented on the ECHA Dissemination Portal for Environmental fate and behavior and ecotoxicology, DEZA, a.s. reported, " [I]t was found to have a calculated bioconcentration factor (BCF) of 94.7 L/kg wet-wt." (DEZA A.S. n.d.-a). However, after rigorously reviewing content on TBC in AoA consultation number 0005-02, the value 94.7 L/kg wet-wt was the only BCF result for TBC, predicted or otherwise. Likewise, in consultation number 0006-01, the applicant Sasol-Huntsman GmbH & Co. KG reported a series of OECD Toolbox predictions for bioaccumulation for the alternative DIBE (DEZA A.S. n.d.-a). However, the applicant did not identify the source of these predictions, nor could we find any QSAR predictions in Table 4.2: Physico-chemical properties of DIBE or Table 4.5: Human health and environmental hazard profile for DIBE (DEZA A.S. n.d.-a).

Completeness review

QSARs used in WoE to assess CMR endpoints varied in quality of completeness for the five main criteria. Only one endpoint, reproductive toxicity, met all five WoE criteria for the alternative bis(2-ethylhexyl) adipate (DEHA) in consultation number 0005-02 (Table 4-4). In the same AoA, for the alternative Akardite II, the endpoint mutagenicity met four criteria: 1) "Pools" information; 2) Conflicting results; 3) Assesses reliability, relevance, adequacy, and quantity; 4) Assesses the overall WoE package. The endpoints that met the least amount of criteria were ready biodegradability, which indicates rapid breakdown of the substance in most environments (ECHA 2017b), and bioaccumulation or persistence (Pavan and Worth 2006) of the alternatives:

Akardite I, ethyl centralite and methyl centralite. WoE for dioctyl azelate (DOZ) did meet two criteria for bioaccumulation: 1) Assesses reliability, relevance, adequacy, and quantity; 2) Assesses overall WoE package. Figure 4-3 illustrates a high level view of these trends. In this figure CMR/PBT vPvB endpoint data are consolidated and organized by the number of criteria met.

Consistency check

Despite the similarities in AoAs, when applying a WoE consistency check, we observed less consistent results (Table 4-5). Only one pair of AoAs assessed the same alternative. For AoA consultation numbers 0005-01 and 0006-01, the alternative diisobutyl hexahydrophthalate (DIBE) was assessed using evidence for mutagenicity and carcinogenicity endpoints, and outcomes from *in vitro* gene mutation in bacteria (Ames test) assays. Although both AoAs acknowledged a concern for reproductive toxicity, both concluded that DIBE may be safer for human health relative to its Annex XIV chemical dibutyl phthalate (DBP). Specifically, positive QSAR predictions for teratogenicity and androgen receptor binding were suggestive of reproductive toxicity. In addition, both AoAs gave QSAR predictions for aquatic toxicity that suggested a low chronic toxicity (Dow Italia Srl and Rohm and Haas France S.A.S. 2016; Sasol-Huntsman GmbH & Co. KG n.d.).

For the endpoints mutagenicity and *in vitro* gene mutation in bacteria (Ames test), both applicants addressed the sub-criterion 'consistency' under the second main criterion (Figure 4-1). Applicants employed WoE using QSARs to point out a lack of alerts for the two endpoints. Furthermore, in the robust study summary (RSS) criterion, both applicant identified the single positive result among the two mutagenicity tests and Ames test (DEZA A.S. n.d.-b; Sasol-Huntsman GmbH & Co. KG n.d.). Each applicant also addressed the sub-criteria 'adequacy' or

use of the *in vitro* gene mutation in bacteria (Ames test) predictions when they pointed out the limited, peer-reviewed experimental data for DIBE (DEZA A.S. n.d.-b; Sasol-Huntsman GmbH & Co. KG n.d.).

Agreement between the two AoAs, however, ended under the two main criteria: *a*) 'Conflicting results' for mutagenicity; and *b*) 'Assess reliability, relevance, adequacy, quantity' for carcinogenicity. In consultation 0005-01, DEZA, a.s. included an informal 'scoring table,' which described the degree of certainty assigned to mutagenicity (DEZA A.S. n.d.-b). In contrast, a 'scoring table' was not included for for mutagenicity for consultation number 0006-01. Finally, for carcinogenicity, while DEZA, a.s. focused on justifying the 'adequacy' of QSAR and read-across predictions in the absence of experimental test results, Sasol-Huntsman GmbH & Co. KG in consultation 0006-01 separately addressed the sub-criteria 'quantity'. In addition, in the assessment of the overall WoE package for consultation 0005-01, DEZA, a.s. was the only applicant to address a concern for both reproductive toxicity and environmental toxicity.

ITS comparative analysis

The majority of AoA CAS numbers did not have a matching CAS number from the Danish EPA advisory list (Table 4-6). We, however, did find matches for four CAS numbers: 103-23-1 (DEHA), 53306-54-0 (DPHP), 7790-7 (ATBC) and 77-94-1 (TBC). All matching Danish EPA advisory list CAS numbers had a Repr. 2, Muta. 2 or Carc. 2 CLP classifications were not assigned to any of the matching chemicals.

Upon further inspection of the alternatives' AoAs, consultation numbers 0002-01, 0002-02, 0003-01, 0003-02, 0004-01, 0004-02 and 0005-02 identified reproductive toxicity for the alternative DEHA (Table 4-7). In each of the AoA's Reduction of overall risk assessments,

applicants reported CMR concerns in addition to DEHA's listing on ECHA's Community Rolling Action Plan (CoRAP) list, which frequently contains substances of PBT and CMR concern (ECHA 2019i). Teratogenicity was also mentioned in each of the AoA's Annexes. In addition, applicants reported a Repro. 2 notified classification for DEHA in the "Notified classification and labelling of DEHA according to CLP criteria" tables (ARKEMA 2013a, b; DEZA A.S. 2013a, b, n.d.-a; GRUPA AZOTY ZAKŁADY AZOTOWE KĘDZIERZYN S.A. 2013a, b). Moreover, consultation number 0005-02 cited uncertain reproductive toxicity in the Comparison of Hazards Table 4.53 (DEZA A.S. n.d.-a).

For the alternative bis(2-propylheptyl) phthalate (DPHP), no mention of reproductive toxicity was made in the Reduction of overall risk assessment for consultation numbers 0002-01, 0002-02, 0003-01, 0003-02, 0004-01, and 0004-02. Similarly, for the alternative acetyl tributyl citrate (ATBC), there is no mention of reproductive toxicity as a concern in the Reduction of overall risk assessment or the Comparison of Hazards tables ro4 consultation numbers 0002-01, 0002-02, 0003-01, 0003-02, 0004-01, 0004-02, and 0005-02. However, for consultation number 0005-02, in the Mammalian hazard profile, the applicant mentions a reproductive toxicity effect doses above 300 mg/kg bw/d (DEZA A.S. n.d.-a; SCENIHR 2015).

Finally, for the alternative tributyl citrate (TBC), the applicant stated a lack of "documented data" on TBC's reproductive toxic effects though a negative QSAR prediction originating from the TERIS database is given in Table 4.68: Human health and environmental hazard profile for TBC (DEZA A.S. n.d.-a).

Discussion

QPRF Equivalency

We began this study by exploring transparency in REACH QSAR documentation in AoAs. We collected data by prediction/QSAR source/alternative for hazard endpoints, and tabulated information supplied for 17 main QPRF criteria (Table 4-1), which included six priority criteria. We considered these criteria as fundamental information that a regulator would need to know to determine whether a QSAR's prediction is reliable or not. From our review, we found that several criteria lacked any information, including one of our priority criteria, structural analogues. The majority of applicants, however, provided prediction information for uncertainty, AD, predicted value and model endpoint, though, a range of predictions, (i.e. 5% - 24%) lacked any information. In particular, f the 'model endpoint' priority criterion, applicants provided some detailed information. Given the range of reported information on the model endpoint, some model endpoint descriptions were better detailed than others. For example, for consultation number 0005-02 for the alternative methyl centralite, the applicant DEZA, a.s. indicated both the specific assay used to develop the QSAR model in addition to the hazard endpoint: "genetic toxicity reported: *In vivo* – Mutagenicity, QSAR prediction for Rodent dominant lethal assay from the Danish (Q)SAR Database (DQD)" (DEZA A.S. n.d.-a). However, most endpoint descriptions did not contain the exact model endpoint or the experimental test. For the alternative ethyl centralite, also from consultation number 0005-02, information in Table 4.17: *Ecological data supporting decisions of Enviornment Canada* only gave the hazard endpoint, (i.e. EPI Suite (BCFWIN) Bioaccumulation potential: Log BCF (predicted by BCFWIN)) without giving any information about the assay or species used to develop the model (DEZA A.S. n.d.-a).

Our comparison of 24 AoAs suggests that current use of single QSAR models continues to be limited, and that ITS QSAR models could provide a huge benefit to the REACH AoA community. Our study also revealed that ITS QSAR predictions can significantly contribute to a broader understanding of a chemical in an alternative substance's assessment.

Despite the large amount of missing QPRF information, questions surrounded the degree to which a regulator would deem the existing information adequate if not fully sufficient. While an alternative stream of information was embedded in AoAs albeit informally and in incomplete form, including at least four of the priority criteria, this does not mean to say that the other criteria were inferior. In fact, several of these criteria, though not essential for a regulator to assess an AoA, if properly reported, could help a regulator more efficiently ascertain whether the QSAR prediction was reliable or not. For example, although we found that predicted values were reported 342 times, without cut-off values, especially for qualitative results, a regulator might not be able to judge an alternative's toxicity or safety without knowing the cut-off values offhand. In our study, missing cut-off values occurred for all QSAR predictions.

Similarly, the problem of missing model version information turned into a significant issue when we attempted to verify AD information. (For analysis by consultation numbers, see Appendix S6). For the alternative 'methyl centralite' in consultation number 0005-02 and the hazard endpoint 'Genetic toxicity: *In vivo* – Chromosomal effect' for the mouse bone marrow sister chromosome exchange assay, the applicant used the Danish (Q)SAR Database (DQD) to report an 'equivocal' result that was 'within QSAR domain' (DEZA A.S. n.d.-a). Yet, when we downloaded the same prediction on March 15, 2019, an inconclusive result was generated from battery QSAR models, which indicated the prediction was out of domain (DTU FOOD et al. 2019). Interestingly, the DQD reported a positive but out of domain result for each QSAR

prediction i.e. Leadscope, CASE Ultra, and SciQSAR. If we could verify that both results came from the same model version, perhaps we could identify a pattern where applicants cited QSAR results as in domain when they are, in fact, out of domain. On the other hand, it could simply be an isolated error. The DQD has been updated several times since the November 2015 launch (DTU Food et al. 2019). Arguably, these errors may not make a difference in the outcome of the relevant AoA. However, it is an indication of deficient practices that could matter in later AoAs.

We also looked at the quality of information that applicants provided. Based on our review, applicants commented on a prediction's uncertainty 266 times. However, these comments lacked depth, and fell short of communicating the importance owed to a discussion on a prediction's uncertainty. For example, many of the comments were single responses or short phrases such as "uncertain," or, "acceptable," or "doubtful reliability." Of course, leading regulators have established methodologies to fill out QPRFs, which provide well-substantiated reasoning. In the Netherlands National Institute for Public Health and the Environment (RIVM) Report 601779001/2007, several QPRF samples provided by the European Chemicals Bureau (ECB), Italy, were included. Specifically, these sample QPRFs demonstrate how a prediction could be explained in depth, which includes explaining a prediction's uncertainty. In the QPRF TOPKAT model prediction example for the substance cinnamaldehyde, prediction reliability (or certainty in this QPRF) was actually assigned a ranking to score the degree of prediction uncertainty/certainty. In addition, the training set's structural analogues and AD were used to justify their 'reliability' reasoning (Rorije et al. 2008),

> Cinnamic aldehyde is within the domain of the models and is also in the training set of both models. Structural analogues are weak and not thought to be particular similar since

the driving factor in the sensitisation behaviour of CAD is thought to be the unsaturated carbonyl system rather than the carbonyl group itself.

However, ECHA has yet to adopt or assign any of these compliance measures to REACH AoAs. While the information in a QPRF is vital in decoding the reasoning behind an applicant's conclusion on whether a prediction is acceptable under regulatory terms, the QPRF template is in no way perfect. Suggested changes to the QPRF format during the 2nd European Union (EU) Technical Committee on New and Existing Chemical Substances (TCNES) / (Q)SAR Working Group meeting (January 2006) included creating more defined headings such as 'other information regarding prediction reliability' to provide more useful information (Rorije et al. 2008). In addition, Walker et al. (Voyer and Heltshe 1984; Walker et al. 2003) suggested that predictions be accompanied by confidence intervals, especially, since descriptors are oftentimes generated by other QSARs, "thus increasing the potential for error propagation." Despite its vulnerabilities, the value of having a QPRF is evident, without which, cases of ambiguity and equivocal language such as stating four days for a chronic toxicity duration (H&R Ölwerke Schindler GmbH 2016) or defining the model endpoint but also writing the prediction was for an undefined endpoint (DEZA A.S. n.d.-a), cannot be resolved and could leave the regulator in a position of disregarding the QSAR prediction altogether.

AoAs that used QSARs in WoE for higher-tier endpoints and WoE completeness review

Findings from our completeness review of 24 AoAs, which assessed 54 non-unique alternatives, revealed that only a limited number of AoAs used Woe with QSARs. In addition, WoE completeness varied depending on the main criteria and hazard endpoint. (For WoE completeness by hazard endpoint, see Appendix S7). Notably, the three AoAs that used WoE involving QSAR predictions for reproductive toxicity, consistently met the majority of the five

main criteria to a higher degree than the other CMR/PBT/vPvB endpoints (Table 4-4). Ultimately, these results reveal the differences in completeness among the applicable sample AoAs. Moreover, these gaps in completeness provide insight into areas that need to be highlighted in future AoA WoE guidance.

The findings on completeness for WoE using QSARs for reproductive toxicity demonstrate the progress made in bridging data gaps for this endpoint. In a 2011 article on reproductive and developmental toxicity in REACH dossiers, the authors Rovida et al. (2011) recommended more support for the development of non-testing methods for reproductive toxicity testing. The article described how data gaps for endocrine disruptions also affected REACH information requirements. While REACH did not set out explicit guidance requesting information for this health endpoint, in our completeness review, applicants provided the most information for WoE using QSARs under reproductive toxicity. One possible explanation for this occurrence could be the increased access to the freely available, online QSAR models and QSAR predictions in the DQD. Alternatively, with REACH's increased focus on higher-tier endpoints, AoA applicants may simply have been more diligent in providing this information.

In contrast, some applicants failed to identify the sources of QSAR predictions in their WoE or did not address positive QSAR predictions in their WoE. Since OECD QSAR Toolbox houses a variety of sources and tools, which can generate different types of predictions, we could not categorize these as QSAR predictions with any confidence (OECD 2019). Not knowing what type of prediction was used in an AoA for a higher-tier endpoint has larger implications when considering the impact this might have on a regulator making a decision on imprecise information or disregarding important information because the analysis of alternatives is deficient.

Even so, DEZA, a.s. reported a positive biodegradation probability (i.e. Biodegradation = 0.0403) using an EPI Suite™ BIOWIN MITI QSAR model prediction in consultation number 0005-02, for the alternative methyl centralite. However, this prediction was not included in the discussion on environmental fate and pathways toxicity (DEZA A.S. n.d.-a),

> Available information, based largely on the outputs of various QSAR models, does not raise concern for either the persistence or bioaccumulative potential of the substance in the environment.

Even though there are other regulatory and best practice cut-off points for a substance to be considered not-readily biodegradable, methyl centralite's prediction of 0.0403 is far below any of these other biodegradable cut-off points, which should have been addressed under the conflicting results WoE criteria. According to Posthumus et al. (2005), a substance with a biodegradable score of 0.0403 is considered persistent. In addition, Posthumus et al. reported that for the EPI Suite™ BIOWIN MITI models, for a substance to be considered persistent, not only must the probability of a substance be < 0.5, but the substance must also meet two other criteria (Posthumus et al. 2005),

- the probability of the non-linear rapid BIODEG model is < 0.5 or
- the result of the ultimate survey model is < 2.2

Because the applicant was not transparent in how they weighted biodegradation, we could not conclusively say if DEZA, a.s. factored this positive prediction for persistence into their conclusions. Table 4.8 Ecological data supporting decisions of Environment Canada on methyl centralite listed several predictions and experimental results for which any of the

persistence and bioaccumulation data could have been part of the line of evidence for non-testing data before any weighting (DEZA A.S. n.d.-a).

Consistency check

As demonstrated in our consistency check on DIBE, different aspects of the analysis produced different outcomes (Table 4-5). For instance, while both applicants provided sub-criteria information for 'adequacy,' one focused on information for mutagenicity while the other gave evidence on reproductive and environmental toxicity. Needless to say, the lack of proper documentation may have put the regulator at a disadvantage. At the same time, expert judgement is viewed as an informal process, which could explain how an assessment ends up in different places. According to Suter et al. (2017), WoE is viewed as an "inferential process" where expert judgment is used to draw conclusions based on a variety of evidence. For our study, this means that in close cases, such as our consistency check for AoAs consultation numbers 0005-01 and 0006-01 for DIBE, an applicant's WoE, even on the same alternative, could come out differently. Furthermore, different WoE outcomes do not mean that more restrictive, prescriptive requirements for WoE are needed. Expert judgments are an acceptable approach in the eyes of major regulators (Martin et al. 2018). In fact, if a WoE approach is too formal, assessors may find the approach too cumbersome (Suter et al. 2017). However, from a regulator's point of view, companies are still required to explain how they assembled their WoE. A regulator could then make a more informed decision rather than judge an AoA based on incomplete WoE involving QSAR predictions for higher-tier endpoints.

ITS comparative analysis

Because ITS can be the precursor to evidence compiled in a WoE, we also compared 24 AoAs with the 2018 Danish EPA's advisory list to screen for potential CMR substances that

AoA applicants might have missed (Table 4-6). Findings from our comparison of 24 AoAs suggest that ITS QSAR battery models could provide significant benefit to the REACH AoA community. Our study revealed that ITS QSAR predictions can contribute to a broader understanding of a chemical in an alternative substance's assessment. For instance, the only QSAR prediction for reproductive toxicity was negative for potential teratogenicity for the alternative TBC in AoA consultation number 0005-02. However, the Danish EPA ITS battery QSAR model prediction predicted a positive response for reproductive toxicity. Having a more powerful result generated by a battery models such as the Danish EPA ITS battery QSAR models could impact an applicant's conclusion on the safety of TBC. Rather than drawing a conclusion of no concern for the hazard profile of an alternative based, in part, on a negative prediction for reproductive toxicity, an applicant could conclude that there is some degree of concern for reproductive toxicity. For example, in contrast to the absence of any QSAR predictions for reproductive toxicity for DEHA, DPHP, ATBC, the Danish EPA assigned a Repr. 2 advisory classification to the three alternatives. One possible explanation for this discrepancy could be that DPHP does not exhibit reproductive toxicity based on available studies of teratogenicity and reprotoxic effects at the highest doses (ARKEMA 2013a, b; DEZA A.S. 2013a, b; GRUPA AZOTY ZAKŁADY AZOTOWE KĘDZIERZYN S.A. 2013a, b). However, under Article 12(1) and Annex VI (ECHA 2011a), companies are still required to report non-testing methods when appropriate though these regulations specifically pertain to registration. At the same time, in the same consultation number for TCB (0005-02), the applicant noted an deficiency of information for reproductive toxicity (DEZA A.S. n.d.-a),

> There is also an absence of documented data on its reproductive effects, while no concerns have been raised with regard to either developmental or endocrine toxicity.

Ultimately, these findings show that ITS QSAR models for CMR endpoints are powerful tools that should be considered as part of the AoA process in identifying potentially toxic endpoints among alternatives. For example, even though the hazard class of Repr. 2 was similarly reflected in the AoAs for the alternative DEHA (ARKEMA 2013a, b; DEZA A.S. 2013a, b; GRUPA AZOTY ZAKŁADY AZOTOWE KĘDZIERZYN S.A. 2013a, b), in the end, the applicants did not factor reproductive toxicity into their conclusions. When noting the "slightly positive" response for the dominant lethal mouse assay, there is no mention of reproductive toxicity which, according to the Danish EPA, has the resulting effect of "early embryonic deaths" (EPA 2018a). We believe that an ITS framework that incorporates more advanced tools such as battery QSAR models not only provides better WoE information, but also serves to fill in data gaps. Ultimately, the advantage of using an ITS framework that incorporates QSAR models, in particular, the DQD, far exceeds the information gains from single use models in that both the applicant and regulator have even greater access to information, which allows a more complete evaluation of the safety of an alternative chemical.

Interestingly, the alternative ATBC, which does not have a harmonized classification and labeling (CLH), was reported by 12 companies to have Muta. 1B and Carc. 1B notified classification and labeling, which is a self-reported C&L. In addition, applicants reported DEHA as listed as group 3 of carcinogens (ARKEMA 2013a, b; DEZA A.S. 2013a, b, n.d.-a; GRUPA AZOTY ZAKŁADY AZOTOWE KĘDZIERZYN S.A. 2013a, b), which the Danish EPA ITS battery QSAR models for carcinogenicity or mutagenicity did not predict a positive response. One can speculate that the impurities in the formulation of ATBC could explain the Carc. 1B and Muta. 1B classification (ARKEMA 2013b). In addition, the IARC group 3 carcinogen classification was established based on "limited evidence in animals" (ARKEMA 2013b).

However, according to ECHA, conflicting evidence is not only acceptable but accounted for in their guidance on how to conduct a WoE assessment, and thus should have been included (ECHA 2016a).

Conclusions

Identifying trends under REACH AoAs in QPRF, WoE and ITS frameworks, which draw on QSAR predictions, is the first step towards understanding the degree to which QSAR predictions fulfill regulatory expectations as well as play a role in driving these frameworks forward. Our results suggest that without the enforcement of QPRF documentation, regulators may be at a disadvantage due to their limited access to a QSAR prediction's information. AoAs under consideration for this evaluation failed to provide any information at all for several criteria. Furthermore, results for meeting priority criteria showed that there is a baseline of information that one ought to know if submitting or assessing an AoA.

For AoAs that use WoE with QSAR predictions, findings from our consistency check provide important insight into the level of completeness. Gaps in consistency for meeting higher-tier WoE requirements under REACH could have significant implications for human health and the environment if positive predictions are not part of the final decision-making equation. One such application that could help ECHA in monitoring WoE use in AoAs and to make the process more standard and transparent could be the development of a platform for WoE in authorization similar to that of registration. Martin et al. (Martin et al. 2018) supports a more prescriptive approach though drawbacks to this approach include extra training and less flexibility in conducting WoE. Grading of Recommendations Assessment, Development, and Evaluation (GRADE) is another prescriptive approach, which uses a rating system to determine the "quality of evidence in systematic reviews and guidelines and grading strength of recommendations in

guidelines" (Guyatt et al. 2011). In fact, the NRC has already recommended this systematic approach to WoE for IRIS (NRC 2014a). Finally, incorporating elements from a multi-criteria decision analysis-based (MCDA) approach may help standardize the assigned weighting of information beyond the Klimisch scores used for reliability. According to Linkov et al. (2009), MCDA combines "value-based assessment" with expert decision-making and scientific judgment by weighting the individual lines of evidence. Ultimately, these recommendations have the potential to combine all steps into one unified process, integrating "social, political, and economic considerations" into the WoE framework as a whole (Linkov et al. 2009). In the end, there will no longer would be a need to separately analyze the Technical and Economic Feasibility portions of the AoA. Moreover, building from an existing WoE platform for alternative testing will likely increase the amounts of properly completed WoE. However, ECHA must first provide the necessary guidance for authorization users in order for this to happen.

The comparison of alternatives with the 2019 Danish EPA advisory list illustrates the narrow regulatory use of either single or ITS QSAR models. Limited consideration of these QSAR models as appropriate tools appears to be a repeating barrier to regulatory transparency, however, future studies will be needed to confirm this apparent trend. In conclusion, ECHA should provide guidance on ITS QSAR models for authorization and other areas under REACH that frequently encounter data gaps. For example, to encourage regulatory acceptance, ECHA should draft guidance that includes ITS QSAR models such as the Danish EPA to meet information requirements under authorization. Similarly, the Danish EPA and ECHA could partner as change agents to develop an international ITS framework within the AoA community.

Tables

Table 4-1. 17 QPRF criteria.

Category	QPRF criteria
Substance information	CAS number EC number Chemical name: IUPAC and CAS names Structural formula Structure codes (codes recognized by modeling software)
Prediction information	Endpoint* Dependent variable Model or submodel name Model version Predicted value (model result)* Predicted value (cut-off value) Input for prediction Descriptor values
Applicability domain	Domains* Structural analogues* Uncertainty of the prediction* Chemical and biological mechanisms*

* Priority criteria

Table 4-2. Alternatives from AoA sample (n = 24) used for comparison with 2018 Danish EPA advisory list for self-classification of hazardous substances.

Alternative Substance (listed in AoA)	Consultation number	CAS Index Name	Molecular formula	CAS number
1,2,4-Trifluorobenzene	0078-01	Benzene, 1,2,4-trifluoro-	C6 H3 F3	367-23-7
Acetyl Tributyl Citrate (ATBC)	0002-01	1,2,3-Propanetricarboxylic acid, 2-(acetyloxy)-, 1,2,3-tributyl ester	C20 H34 O8	77-90-7
Acetyl Tributyl Citrate (ATBC)	0002-02	1,2,3-Propanetricarboxylic acid, 2-(acetyloxy)-, 1,2,3-tributyl ester	C20 H34 O8	77-90-7
Acetyl Tributyl Citrate (ATBC)	0003-01	1,2,3-Propanetricarboxylic acid, 2-(acetyloxy)-, 1,2,3-tributyl ester	C20 H34 O8	77-90-7
Acetyl Tributyl Citrate (ATBC)	0003-02	1,2,3-Propanetricarboxylic acid, 2-(acetyloxy)-, 1,2,3-tributyl ester	C20 H34 O8	77-90-7
Acetyl Tributyl Citrate (ATBC)	0004-01	1,2,3-Propanetricarboxylic acid, 2-(acetyloxy)-, 1,2,3-tributyl ester	C20 H34 O8	77-90-7
Acetyl Tributyl Citrate (ATBC)	0004-02	1,2,3-Propanetricarboxylic acid, 2-(acetyloxy)-, 1,2,3-tributyl ester	C20 H34 O8	77-90-7
Acetyl Tributyl Citrate (ATBC)	0005-02	1,2,3-Propanetricarboxylic acid, 2-(acetyloxy)-, 1,2,3-tributyl ester	C20 H34 O8	77-90-7
Akardite I (1,2-Diphenyl Urea)	0005-02	Urea, N,N'-diphenyl-	C13 H12 N2 O	102-07-8
Akardite II (3-Methyl-1,1,-Diphenylurea)	0005-02	Urea, N'-methyl-N,N-diphenyl-	C14 H14 N2 O	13114-72-2
Akardite III (3-Ethyl-1,1,-Diphenyl Urea)	0005-02	Urea, N'-ethyl-N,N-diphenyl-	C15 H16 N2 O	18168-01-9
Benzene, Ethenyl-, Polymer With 1,3-butadiene, brominated (brominated co- polymer of styrene and butadiene) (Polymeric Flame Retardant (pFR)	0013-01	Benzene, ethenyl-, polymer with 1,3-butadiene, brominated polymer, manual registration, generic registration	N/A	1195978-93-8
Benzene, Ethenyl-, Polymer With 1,3-butadiene, brominated (brominated co- polymer of styrene and butadiene) (Polymeric Flame Retardant (pFR)	0013-02	Benzene, ethenyl-, polymer with 1,3-butadiene, brominated polymer, manual registration, generic registration	N/A	1195978-93-8
Chromium(III) chloride	0035-01	Chromium chloride (CrCl3)	Cl3 Cr	10025-73-7
Chromium(III) chloride	0036-01	Chromium chloride (CrCl3)	Cl3 Cr	10025-73-7
Chromium(III) chloride	0037-01	Chromium chloride (CrCl3)	Cl3 Cr	10025-73-7
Chromium(III) chloride	0038-01	Chromium chloride (CrCl3)	Cl3 Cr	10025-73-7
Chromium(III) chloride	0039-01	Chromium chloride (CrCl3)	Cl3 Cr	10025-73-7
Chromium(III) chloride	0040-01	Chromium chloride (CrCl3)	Cl3 Cr	10025-73-7
Chromium(III) chloride	0041-01	Chromium chloride (CrCl3)	Cl3 Cr	10025-73-7
Chromium(III) chloride	0041-02	Chromium chloride (CrCl3)	Cl3 Cr	10025-73-7
Di(2-ethylhexyl) adipate (DEHA)	0002-01	Hexanedioic acid, 1,6-bis(2-ethylhexyl) ester	C22 H42 O4	103-23-1
Di(2-ethylhexyl) adipate (DEHA)	0002-02	Hexanedioic acid, 1,6-bis(2-ethylhexyl) ester	C22 H42 O4	103-23-1
Di(2-ethylhexyl) adipate (DEHA)	0003-01	Hexanedioic acid, 1,6-bis(2-ethylhexyl) ester	C22 H42 O4	103-23-1
Di(2-ethylhexyl) adipate (DEHA)	0003-02	Hexanedioic acid, 1,6-bis(2-ethylhexyl) ester	C22 H42 O4	103-23-1
Di(2-ethylhexyl) adipate (DEHA)	0004-01	Hexanedioic acid, 1,6-bis(2-ethylhexyl) ester	C22 H42 O4	103-23-1
Di(2-ethylhexyl) adipate (DEHA)	0004-02	Hexanedioic acid, 1,6-bis(2-ethylhexyl) ester	C22 H42 O4	103-23-1

(Table 4-2 continued)

Alternative Substance (listed in AoA)	Consultation number	CAS Index Name	Molecular formula	CAS number
Bis(2-ethylhexyl) adipate (DEHA)	0005-02	Hexanedioic acid, 1,6-bis(2-ethylhexyl) ester	C22 H42 O4	103-23-1
Dichloromethane (methylene chloride) (DCM)	0077-01	-	C H2 Cl2	75-09-2
Dichloromethane (methylene chloride) (DCM)	0078-01	-	C H2 Cl2	75-09-2
Diethylene glycol dibutyl ether (possible substances alternatives)	0091-01	-	C12 H26 O3	112-73-2
Diisobutyl hexahydrophthalate (DIBE)	0005-01	Diisobutyl 1,2-cyclohexanedicarboxylate	C16H28O4	70969-58-3
Diisobutyl hexahydrophthalate (DIBE)	0006-01	Diisobutyl 1,2-cyclohexanedicarboxylate	C16H28O4	70969-58-3
Dioctyl azelate (DOZ)	0005-02	-	C25 H48 O4	103-24-2
Dioctylsebacate (Diethylhexylsebacate) (DEHS)	0002-01	Decanedioic acid, 1,10-bis(2-ethylhexyl) ester	C26H50O4	122-62-3
Dioctylsebacate (Diethylhexylsebacate) (DEHS)	0002-02	Decanedioic acid, 1,10-bis(2-ethylhexyl) ester	C26H50O4	122-62-3
Dioctylsebacate (Diethylhexylsebacate) (DEHS)	0003-01	Decanedioic acid, 1,10-bis(2-ethylhexyl) ester	C26H50O4	122-62-3
Dioctylsebacate (Diethylhexylsebacate) (DEHS)	0003-02	Decanedioic acid, 1,10-bis(2-ethylhexyl) ester	C26H50O4	122-62-3
Dioctylsebacate (Diethylhexylsebacate) (DEHS)	0004-01	Decanedioic acid, 1,10-bis(2-ethylhexyl) ester	C26H50O4	122-62-3
Dioctylsebacate (Diethylhexylsebacate) (DEHS)	0004-02	Decanedioic acid, 1,10-bis(2-ethylhexyl) ester	C26H50O4	122-62-3
DPHP Bis(2-Propylheptyl) Phthalate (DPHP)	0002-01	1,2-Benzenedicarboxylic acid, 1,2-bis(2-propylheptyl) ester	C28 H46 O4	53306-54-0
DPHP Bis(2-Propylheptyl) Phthalate (DPHP)	0002-02	1,2-Benzenedicarboxylic acid, 1,2-bis(2-propylheptyl) ester	C28 H46 O4	53306-54-0
DPHP Bis(2-Propylheptyl) Phthalate (DPHP)	0003-01	1,2-Benzenedicarboxylic acid, 1,2-bis(2-propylheptyl) ester	C28 H46 O4	53306-54-0
DPHP Bis(2-Propylheptyl) Phthalate (DPHP)	0003-02	1,2-Benzenedicarboxylic acid, 1,2-bis(2-propylheptyl) ester	C28 H46 O4	53306-54-0
DPHP Bis(2-Propylheptyl) Phthalate (DPHP)	0004-01	1,2-Benzenedicarboxylic acid, 1,2-bis(2-propylheptyl) ester	C28 H46 O4	53306-54-0
DPHP Bis(2-Propylheptyl) Phthalate (DPHP)	0004-02	1,2-Benzenedicarboxylic acid, 1,2-bis(2-propylheptyl) ester	C28 H46 O4	53306-54-0
Ethyl centralite	0005-02	Urea, N,N'-diethyl-N,N'-diphenyl-	C17H20N2O	85-98-3
Isodecyl pelargonate (IDP)	0005-02	Nonanoic acid, 8-methylnonyl ester	C19 H38 O2	109-32-0
Methyl centralite	0005-02	Urea, N,N'-dimethyl-N,N'-diphenyl-	C15 H16 N2 O	611-92-7
Methyl ethyl ketone (MEK)	0080-01	2-Butanone	C_4H_8O	78-93-3
Methyl ethyl ketone (MEK)	0081-01	2-Butanone	C_4H_8O	78-93-3
Ortho-xylene	0005-01	Benzene, 1,2-dimethyl-	C8H10	95-47-6
Ortho-Xylene	0006-01	Benzene, 1,2-dimethyl-	C8H10	95-47-6
Tributyl citrate (TBC)	0005-02	1,2,3-Propanetricarboxylic acid, 2-hydroxy-, 1,2,3-tributyl ester	C18 H32 O7	77-94-1

Table 4-3. Alternatives in AoAs that used QSARs[+] in WoE for higher-tier endpoints.

Consultation numbers	Applicants	Alternatives
0005-01	DEZA, a.s.	Diisobutyl hexahydrophthalate (DIBE)
0005-02	DEZA, a.s.	Methyl centralite Ethyl centralite Akardite I Akardite II Akardite III Bis(2-ethylhexyl) adipate (DEHA) Acetyl tributyl citrate (ATBC) Tributyl citrate (TBC) Dioctyl azelate (DOZ) Isodecyl pelargonate (IDP)
0006-01	Sasol-Huntsman GmbH & Co. KG	Diisobutyl hexahydrophthalate (DIBE)

Table 4-4. Number of criteria met for QSAR predictions used in WoE by CMR/PBT vPvB endpoints.

Consultation number	Alternative:	No. of criteria met	Criteria				
			RSS	Assesses reliability, relevance, adequacy, and quantity	"Pools" information	Conflicting results	Assesses overall package
MUTAGENICITY							
0005-02	Akardite I	1	-	-	x	-	-
0005-02	Akardite III	2	-	-	x	-	x
0006-01	Diisobutyl hexahydrophthalate (DIBE)	2	-	-	x	-	x
0005-02	Dioctyl azelate (DOZ)	2	-	-	x	-	x
0005-02	Ethyl centralite	2	-	-	x	-	x
0005-02	Isodecyl pelargonate (IDP)	2	-	-	x	-	x
0005-02	Methyl centralite	2	-	-	x	-	x
0005-02	Tributyl citrate (TBC)	2	-	-	x	-	x
0005-01	Diisobutyl hexahydrophthalate (DIBE)	3	-	-	x	x	x
0005-02	Akardite II	4	-	x	x	x	x
0005-02	Acetyl tributyl citrate (ATBC)	-	-	-	-	-	-
0005-02	Bis(2-ethylhexyl) adipate (DEHA)	-	-	-	-	-	-
***IN VITRO* GENE MUTATION IN BACTERIA (AMES TEST)**							
0005-02	Akardite III	2	-	-	x	-	x
0005-01	Diisobutyl hexahydrophthalate (DIBE)	2	-	-	x	-	x
0006-01	Diisobutyl hexahydrophthalate (DIBE)	2	-	-	x	-	x
0005-02	Ethyl centralite	2	-	-	x	-	x
0005-02	Isodecyl pelargonate (IDP)	2	-	-	x	-	x
0005-02	Methyl centralite	2	-	-	x	-	x
0005-02	Tributyl citrate (TBC)	2	-	-	x	-	x
0005-02	Akardite II	3	-	-	x	x	x
0005-02	Acetyl tributyl citrate (ATBC)	-	-	-	-	-	-
0005-02	Akardite I	-	-	-	-	-	-
0005-02	Bis(2-ethylhexyl) adipate (DEHA)	-	-	-	-	-	-
0005-02	Dioctyl azelate (DOZ)	-	-	-	-	-	-
CARCINOGENICITY							
0005-02	Akardite I	1	-	-	x	-	-
0005-02	Akardite III	2	-	-	x	-	x
0005-01	Diisobutyl hexahydrophthalate (DIBE)	2	-	-	-	x	x
0006-01	Diisobutyl hexahydrophthalate (DIBE)	2	-	-	-	x	x
0005-02	Dioctyl azelate (DOZ)	2	-	-	x	-	x
0005-02	Ethyl centralite	2	-	-	x	-	x
0005-02	Isodecyl pelargonate (IDP)	2	-	-	x	-	x
0005-02	Methyl centralite	2	-	-	x	-	x
0005-02	Akardite II	3	-	-	-	x	x
0005-02	Tributyl citrate (TBC)	3	-	-	-	x	x
0005-02	Acetyl tributyl citrate (ATBC)	-	-	-	-	-	-
0005-02	Bis(2-ethylhexyl) adipate (DEHA)	-	-	-	-	-	-

(Table 4-4 continued)

Consultation number	Alternative:	No. of criteria met	Criteria				
			RSS	Assesses reliability, relevance, adequacy, and quantity	"Pools" information	Conflicting results	Assesses overall package
REPRODUCTIVE TOXICITY							
0005-01	Diisobutyl hexahydrophthalate (DIBE)	1	-	-	-	-	x
0005-02	Methyl centralite	1	-	-	-	-	x
0005-02	Acetyl tributyl citrate (ATBC)	2	-	-	x	-	x
0005-02	Akardite I	2	-	-	x	x	-
0005-02	Akardite III	2	-	-	x	-	x
0006-01	Diisobutyl hexahydrophthalate (DIBE)	2	-	-	-	x	x
0005-02	Ethyl centralite	2	-	-	x	-	x
0005-02	Akardite II	3	-	x	x	-	x
0005-02	Dioctyl azelate (DOZ)	3	-	-	x	x	x
0005-02	Isodecyl pelargonate (IDP)	3	-	-	x	x	x
0005-02	Tributyl citrate (TBC)	3	-	-	x	x	x
0005-02	Bis(2-ethylhexyl) adipate (DEHA)	5	x	x	x	x	x
READY BIODEGRADABILITY							
0005-02	Akardite I	0	-	-	-	-	-
0005-02	Ethyl centralite	0	-	-	-	-	-
0005-02	Methyl centralite	0	-	-	-	-	-
0005-02	Isodecyl pelargonate (IDP)	1	-	-	-	-	x
0005-02	Acetyl tributyl citrate (ATBC)	-	-	-	-	-	-
0005-02	Akardite II	-	-	-	-	-	-
0005-02	Akardite III	-	-	-	-	-	-
0005-02	Bis(2-ethylhexyl) adipate (DEHA)	-	-				
0005-01	Diisobutyl hexahydrophthalate (DIBE)	-	-	-	-	-	-
0006-01	Diisobutyl hexahydrophthalate (DIBE)	-	-	-	-	-	-
0005-02	Dioctyl azelate (DOZ)	-	-	-	-	-	-
0005-02	Tributyl citrate (TBC)	-	-	-	-	-	-
BIOACCUMULATION							
0005-02	Akardite I	0	-	-	-	-	-
0005-02	Ethyl centralite	0	-	-	-	-	-
0005-02	Methyl centralite	0	-	-	-	-	-
0005-02	Isodecyl pelargonate (IDP)	1	-	-	-	-	x
0005-02	Dioctyl azelate (DOZ)	2	-	x	-	-	x
0005-02	Acetyl tributyl citrate (ATBC)	-	-	-	-	-	-
0005-02	Akardite II	-	-	-	-	-	-
0005-02	Akardite III	-	-	-	-	-	-
0005-02	Bis(2-ethylhexyl) adipate (DEHA)	-	-	-	-	-	-
0005-01	Diisobutyl hexahydrophthalate (DIBE)	-	-	-	-	-	-
0006-01	Diisobutyl hexahydrophthalate (DIBE)	-	-	-	-	-	-
0005-02	Tributyl citrate (TBC)	-	-	-	-	-	-

(Table 4-4 continued)

Notes:

(a) We only covered QSAR predictions in WoE, thus, hyphenated blank spaces meant that either there were no CMR/PBT vPvB QSAR predictions in WoE to analyze, which we coded as non-applicable.

(b) Numbers 1-5 corresponds to the number of criteria that was met on our checklist, where the total number of criteria was five.

(c) 0 indicates a QSAR prediction without a WoE context was cited for that endpoint.

(d) A hyphen '-' in the number of criteria column indicates that either the endpoint data not relevant to WoE e.g. evaluates potential alternative ,or, a QSAR prediction did not exist.

Table 4-5. Consistency check for AoA consultations 0005-01 and 0006-01 assessing the same alternative Diisobutyl hexahydrophthalate (DIBE) for the endpoints: mutagenicity, *in vitro* gene mutation in bacteria (Ames test), and carcinogenicity.

Alternative	**Mutagenicity: Conflicting results**			**In vitro gene mutation in bacteria (Ames test): Assessing reliability, relevance, adequacy, quantity**		**Carcinogenicity: Assessing reliability, relevance, adequacy, quantity**	
DIBE*	Addressed conflicting results to *in vitro* negative results	Scoring table	Non-applicable (i.e. sub-criteria not addressed)	Adequacy	Addressed consistency	Adequacy	Quantity
0005-01	X	X	-	X	X	X	-
0006-01	X	-	X	X	X	-	X

* Diisobutyl hexahydrophthalate (DIBE)

Table 4-6. Matching AoA and Danish EPA advisory list for self-classification of hazardous substances[§] CAS numbers for CMR endpoints.

CAS number	Alternative Substances (as listed in AoA)	C (QSAR)[+]	M (QSAR)[++]	R (QSAR)[+++]	CLP classification
10025-73-7	Chromium(III) chloride	No	No	No	
102-07-8	Akardite I (1,2-DIPHENYL UREA)	No	No	No	
103-23-1	Bis(2-ethylhexyl) adipate (DEHA)	No	No	**Yes**	**Repr. 2**
103-24-2	Dioctyl azelate (DOZ)	No	No	No	
109-32-0	Isodecyl pelargonate (IDP)	No	No	No	
112-73-2	Diethylene glycol dibutyl ether (possible substances alternatives)	No	No	No	
1195978-93-8	benzene, ethenyl-, polymer with 1,3-butadiene, brominated (brominated co- polymer of styrene and butadiene) (Polymeric Flame Retardant (pFR)	No	No	No	
122-62-3	DIOCTYLSEBACATE (DIETHYLHEXYLSEBACATE) (DEHS)	No	No	No	
13114-72-2	Akardite II (3-METHYL-1,1,-DIPHENYLUREA)	No	No	No	
18168-01-9	Akardite III (3-ETHYL-1,1,-DIPHENYL UREA)	No	No	No	
29063-28-3	octanol (mixed isomers) (possible substances alternatives)	No	No	No	
367-23-7	1,2,4-Trifluorobenzene	No	No	No	
53306-54-0	Bis(2-propylheptyl) Phthalate (DPHP)	No	No	**Yes**	**Repr. 2**
611-92-7	methyl centralite	No	No	No	
70969-58-3	Diisobutyl hexahydrophthalate (DIBE)	No	No	No	
75-09-2	Dichloromethane (methylene chloride) (DCM)	No	No	No	
77-90-7	ACETYL TRIBUTYL CITRATE (ATBC)	No	No	**Yes**	**Repr. 2**
77-94-1	Tributyl citrate (TBC)	No	No	**Yes**	**Repr. 2**
78-93-3	Methyl ethyl ketone (MEK)	No	No	No	
85-98-3	ethyl centralite	No	No	No	
95-47-6	Ortho-xylene	No	No	No	

+ C (QSAR) QSAR predictions for carcinogenicity

++ M (QSAR) predictions for mutagenicity

+++ R (QSAR) predictions for reproductive toxicity

§ All data comes from the Danish EPA's advisory list for self-classification of hazardous substances (EPA 2018b)

Table 4-7. Reproductive toxicity identifiers in AoAs with matching CAS numbers.

				Reproductive toxicity identifiers (AoAs)		
CAS #	**Alternative Substance (listed in AoA)**	**Danish EPA CLP classification**	**Consultation number(s)**	**Yes**	**No**	**Uncertain**
103-23-1	Di(2-ethylhexyl) adipate (DEHA)/ Bis(2-ethylhexyl) adipate (DEHA)	Repr. 2	0002-01	X		
			0002-02	X		
			0003-01	X		
			0003-02	X		
			0004-01	X		
			0004-02	X		
			0005-02			X
53306-54-0	Bis(2-Propylheptyl) Phthalate (DPHP)	Repr. 2	0002-01		X	
			0002-02		X	
			0003-01		X	
			0003-02		X	
			0004-01		X	
			0004-02		X	
77-90-7	Acetyl Tributyl Citrate (ATBC)	Repr. 2	0002-01		X	
			0002-02		X	
			0003-01		X	
			0003-02		X	
			0004-01		X	
			0004-02		X	
			0005-02		X	
77-94-1	Tributyl citrate (TBC)	Repr. 2	0005-02		X	

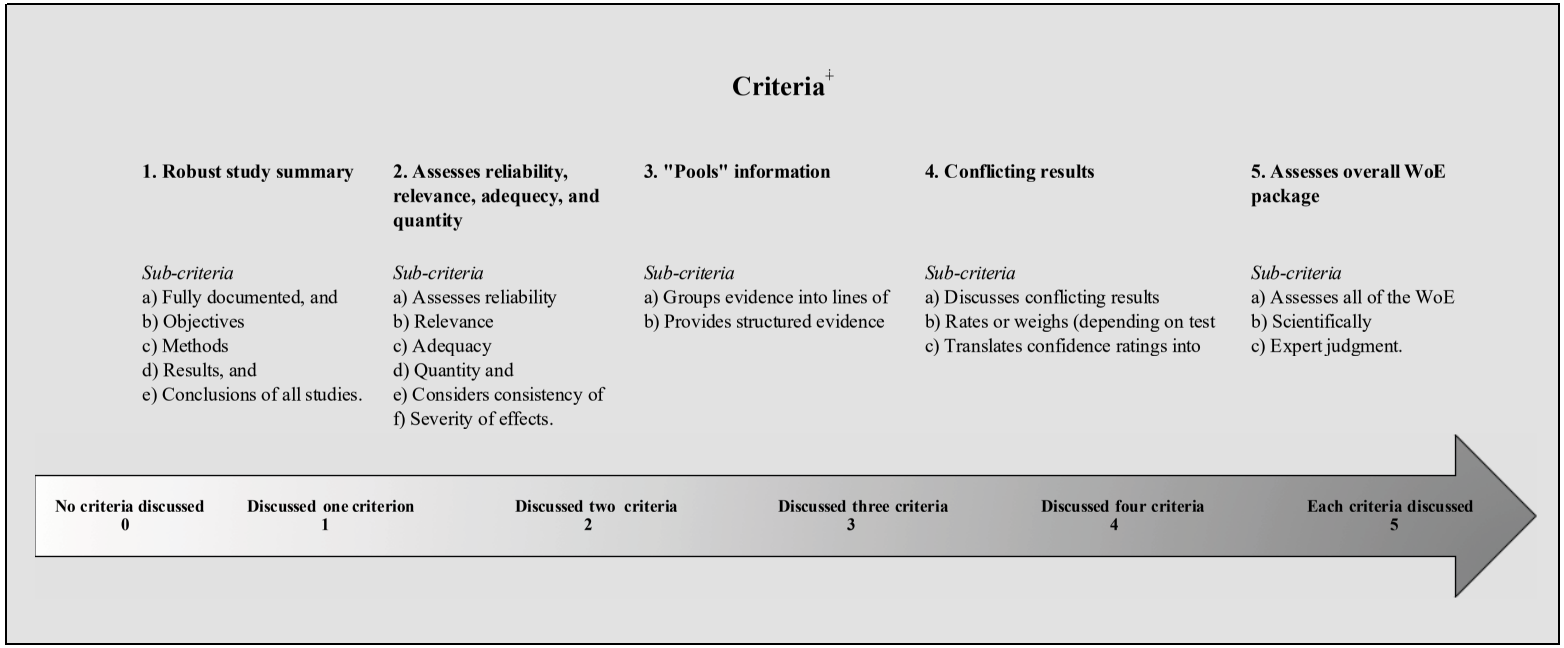


Figure 4-1. Checklist with quality scale for WoE completeness review.

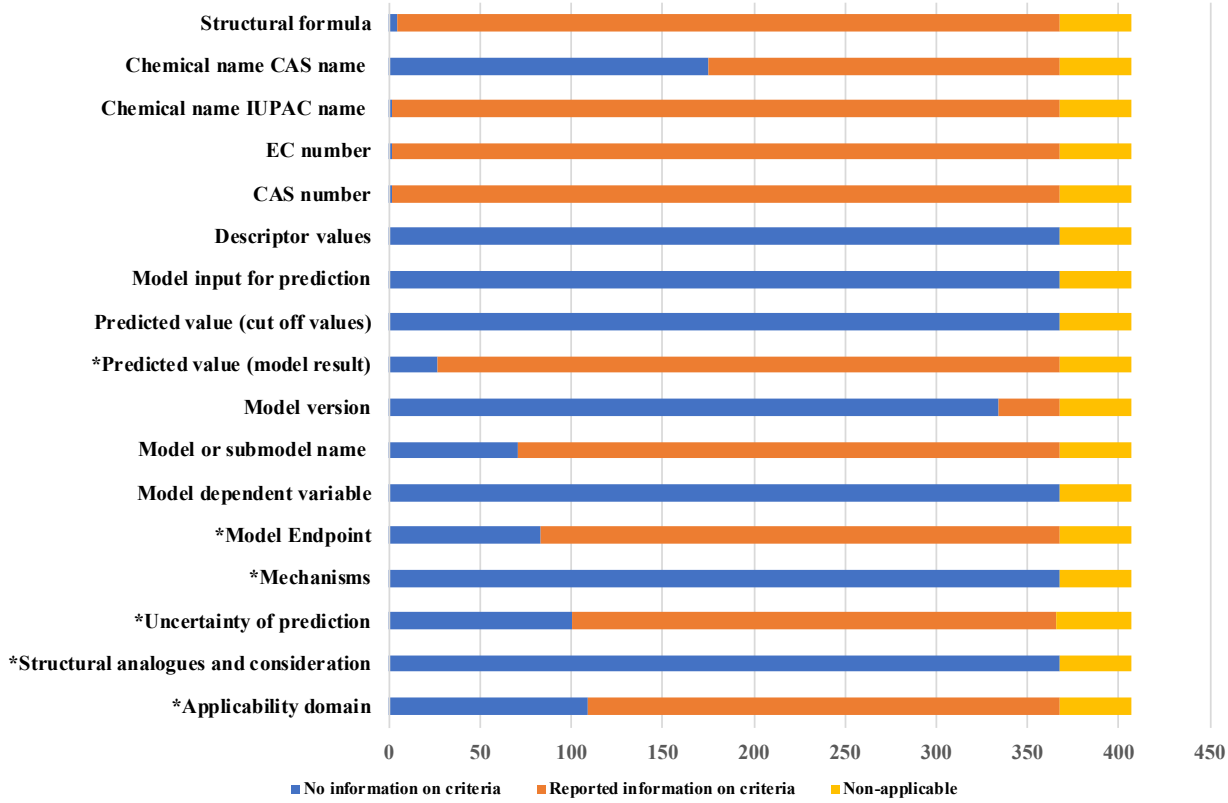


Figure 4-2. Descriptive statistics for 16 QPRF criteria in AoAs using QSAR predictions.

*Six priority criteria.

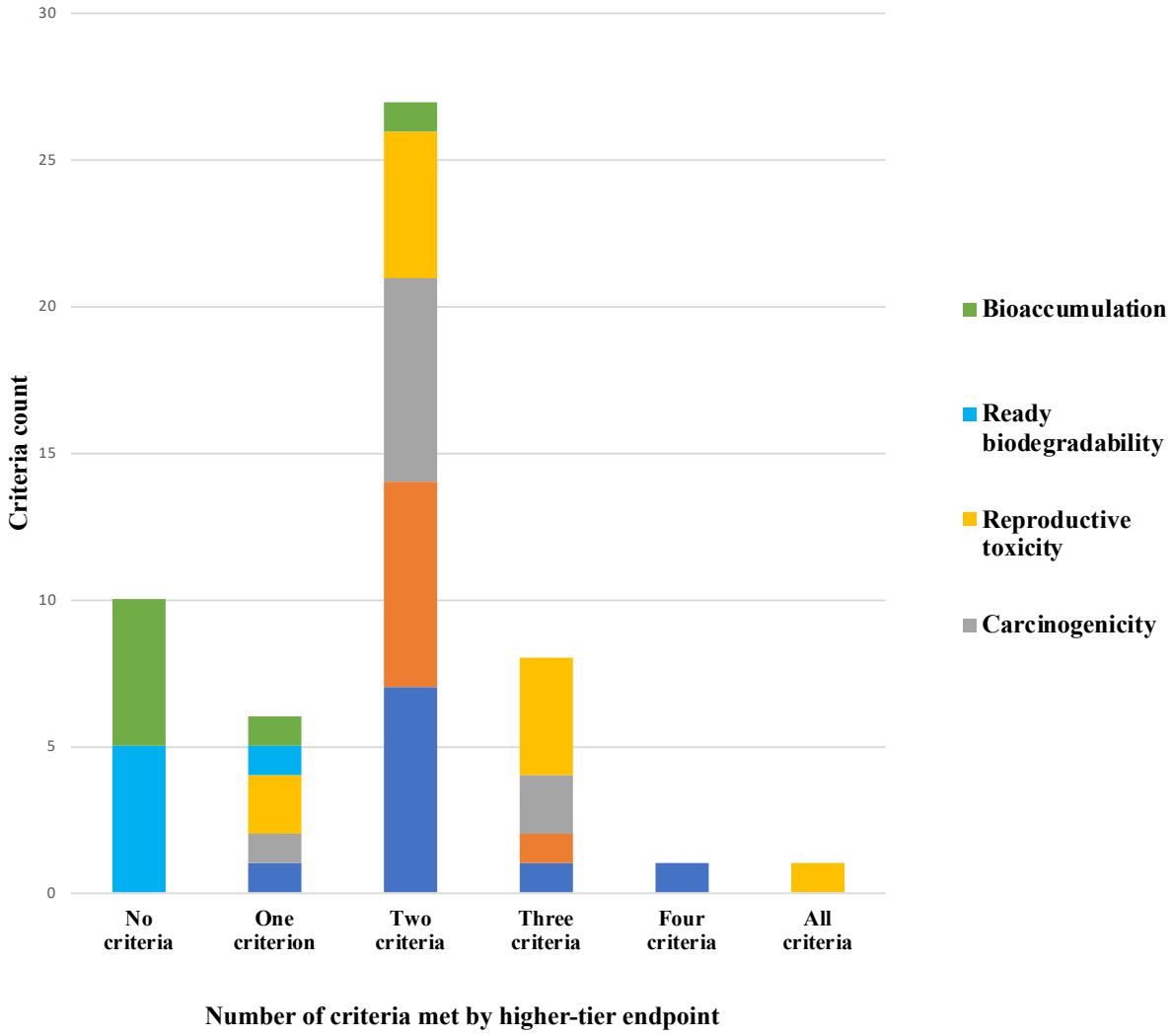


Figure 4-3. Number of criteria met in criteria checklist by QSARs used in WoE sub-divided by higher-tier endpoints.

Notes:

(a) Robust study summary: fully documented. Includes objectives, methods, results, conclusions of all studies.

(b) Assesses reliability, relevance, adequacy, and quantity. Considers consistency of results and severity of effects.

(c) "Pools" information by grouping evidence into lines of evidence and providing structured evidence tables.

(d) Conflicting results: Rates or weighs (depending on test method, data quality, endpoint) using scoring table, and translates confidence ratings into level of level of evidence for health effect.

(e) Assesses overall package. Scientifically justified/argued using expert judgment.

Supporting information

Additional information on sample WoE template (Appendix S1), study limitations (Appendix S2), WoE criteria checklist (Appendix S3), sub-criteria coding for the five main criteria (Appendix S4), Danish EPA advisory self-classifications (Appendix S5), consultation numbers (Appendix S6), WoE by CMR and PBT/vPvB endpoints (Appendix S7).

Appendix S1

Sample WoE template

Table A.1: Optional tabular format for summarising weight of evidence assessment for an emerging contaminant

Question: Hazard identification of an emerging contaminant		
Assemble the evidence	Select evidence	No toxicity data available: use read-across from already-tested similar compounds, *in silico* tools (QSAR) to predict toxicity
	Lines of Evidence	Identify lines of evidence for potential effect(s) from the presence of a structural alert or QSAR models, read-across from similar compounds
Weigh the evidence	Methods	Evaluate the reliability, relevance and consistency of the QSAR models. This can include weighing model results on a statistical basis (e.g. likelihood of a compound with a structural alert to express (a) toxic property(ies))
	Results	Toxicity value for each line of evidence, with associated assessment of reliability (e.g. through the applicability domain of the models used)
Integrate the evidence	Methods	If the estimates from the different models converge, the level of uncertainty regarding the toxic property(ies) can be evaluated (e.g. through the applicability domain of the models used). If the estimates do not converge, further modelling for the toxic property(ies) could be undertaken to evaluate whether the results can be improved
	Results	Integrated the toxicity value and uncertainty factor to derive a health based guidance value for the emerging contaminant: Summary Table

Table A.3: Optional tabular format for summarising weight of evidence assessment of an emerging contaminant

Question		***Hazard identification of an emerging contaminant***
Assemble the evidence	Select evidence	Nine QSAR models from two *in silico* platforms and a program for read–across were used to estimate mutagenicity potential (as assessed through bacterial reverse mutation test) of the target compound
	Lines of evidence	Except two, all estimates indicated the compound to be non-mutagenic. The exception was the QSAR model CAESAR within VEGA platform that predicted the compound as mutagenic, and the read-across programme ToxRead that showed one out of five similar compounds to be mutagenic
Weigh the evidence	Methods	VEGA provides a quantitative measurement of reliability and values higher than 0.8 ADI are considered more reliable. T.E.S.T. applies a filter to eliminate not reliable predictions. The results obtained from these platforms in this case are therefore reliable. ToxRead indicates the alerts associated with the effect and similar compounds. In case of chemicals with the toxicity value conflicting with the rule, the user should check if there are rules present only in the similar compound and not in the target, explaining the conflicting toxicity value. This is useful to evaluate the relevance of the lines of evidence, disregarding those that are not relevant
	Results	T.E.S.T. results consistently indicated non-mutagenicity. The VEGA models called SARpy and KNN showed higher indices for reliability, also predicted non-mutagenicity. The CAESAR and ISS models within the VEGA models showed relatively lower reliability. ToxRead results show that most of the compounds similar to the target compound were not-mutagenic. The only structural rule for mutagenicity found in one similar compound is not present in the target compound, and therefore is not relevant
Integrate the evidence	Methods	The *in silico* estimates have been integrated while considering the reliability and relevance of the individual values, together with the consistency of all the predicted values, to make an informed expert judgement about the probability that the target compound is not-mutagenic
	Results	The large majority of the *in silico* values are in concordance for non-mutagenicity of the target compound. One conflicting estimate is less reliable whereas the other is not relevant to the target compound. Considering all the evidence from this *in silico* assessment, it was concluded by informed expert judgement that the target compound is most likely (about 90% probability) to be non-mutagenic

ADI: Applicability Domain Index.

Figure 4-1S. Sample WoE using a structured evidence table (EFSA 2017)

Appendix S2

Study Limitations

Unless the description of the QSAR prediction was clearly ad-hoc and quantitative values were missing for predictions where there was already evidence that the QSAR source produces qualitative predictions, we assumed that qualitative and quantitative values reported in AoAs by applicants were the results generated by QSAR platforms. For example, in consultation number 0078-01 for the alternative 1,2,4-trichorobenzene by BIOWIN 3, the applicant did not provide an actual QSAR prediction and only said the biodegradation time frame predicted by BIOWIN 3 is "months and longer" for the fluoroisomer compared to "weeks to months" for the chlorinated benzene (p.105), However, in select instances, we could not go back and confirm the QSAR sources if they were not identified by model name and version. In AoA consultation number 0005-02, the applicant reported "unknown" for irritation: eye irritation for the alternative methyl centralite. Perhaps the applicant meant "inconclusive" for which we accepted "equivocal" in instances of DQD predictions. Model endpoint information may also have been embedded in an AoA, but this is information typically found in a QMRF. Without having this official document or access to the original QSAR platform, we could not verify if sporadic information in an AoA was what the developer actually described as the model endpoint. In cases where an applicated reported a QSAR prediction twice, for example, consultation number 0005-02 for the alternative ethyl centralite Irritation: Skin irritation/ corrosion (DEZA A.S. n.d.-a), we accepted it as two separate predictions because we did not have the QPRF to verify the prediction either way.

In addition, all endpoints were based on our curated list of endpoints from Chapter 2. While endpoints outside of this list may have excluded potential QSAR predictions from our analysis, the consistency of this endpoint classification allowed us to subject previous research to

a more in-depth analysis without further classification. However, if applicants did not report a specific endpoint, such as in a scoring table, this endpoint did not get coded. For instance, in consultation 0005-02, the applicant did not consider test results to be mutagenic, and subsequently left off mutagenicity from Table 4.31 (DEZA A.S.) thereby removing it from our coding.

Furthermore, our approach to data-mining and coding may have excluded some aspects of an applicant's WoE from our analysis. When analyzing each AoA, we maintained a narrow research path, and did not track criteria outside of our classifications. Specifically, for information to be considered under the final criteria "Assess overall WoE package," the applicant needed to make their scientific arguments within an AoA's "Reduction in overall risk." In consultation number 0005-02, scientific arguments were made for the alternative, Akardite I, however, the applicant made these arguments in the comparison of hazards sections. Therefore, any scientific argument written under the "Comparison of Hazards" section would not be coded under the "Assessment of Overall WoE" criteria but to another criteria such as criterion 3 'Pools information'. Similarly, for the same consultation number, biodegradability QSAR predictions for the alternative ethyl centralite are not discussed in either the RSS or Reduction of overall risk even though PBT/vPvB is discussed in the WoE for the alternative ethyl centralite (DEZA A.S. n.d.-a; Dow Italia Srl and Rohm and Haas France S.A.S. 2016). Because we are only considering CMR and PBT/vPvB endpoints in AoAs that use WoE with QSAR predictions, this information did not get coded. Furthermore, we did not always have the coding to explain when criteria were not met. Although some AoAs did not meet our five criteria, in other instances, the criteria simply did not exist in the AoA. For example, in consultation number 0005-02 for the alternative ethyl centralite, we coded the WoE for the reproductive toxicity endpoint as not providing any

conflicting results. However, we assigned this code because all of the information in the AoA for this endpoint was consistent, which would then have been coded separately for “consistence.”

If an applicant specified certain endpoints in their approach to WoE, only these endpoints could be considered when judging if the evidence met any of our five WoE criteria. For instance, in consultation 0005-02, the applicant specified Table 4.86 as the basis for “additional insight” into the alternative isodecyl pelargonate (IDP), thus, we could not factor any other endpoint that was not on this list into our completeness review, when considering “adequacy” or the usefulness of the information. More specifically, we could not code either biodegradation or bioaccumulation for “adequacy” because their QSAR predictions were listed on another table (DEZA A.S. n.d.-a).

Finally, because we used the 2018 Danish EPA advisory classifications list to screen for potential CMR substances, our results were subject to the factors that went into Danish EPA’s ITS QSAR model development, which has been subject to updates since 2001. For example, endocrine disrupting (ED) models, which were not included in the Danish EPA’s battery of QSAR models for reproductive toxicity, may be considered an important endpoint in reproductive toxicity for other model developers (Evans 2015). Even so, ED models have been used to identify mechanisms for reproductive toxicity (Jensen et al. 2008). Thus, while other ITS QSAR model developers may have taken a different approach in selecting endpoints for battery QSAR modelling as well as the selection of algorithms to integrate results, we based our results on the Danish EPA’s decision-making. In addition, our sample of AoAs was collected through May 2017; more current AoAs may have employed ITS QSAR modeling.

Appendix S3

WoE criteria checklist

Robust Study Summaries. Our checklist began with an assessment of robust study summaries (RSS) as the first criterion, which are provided as part of WoE under registration. According to ECHA, an RSS is (ECHA 2012),

> [A] detailed summary of the objectives, methods, results and conclusions of a full study report providing sufficient information to make an independent assessment of the study minimizing the need to consult the full study report (Article 3 (28) of REACH).

Under an endpoint study record (ESR), registrants must provide robust study summaries in the technical dossier for each key study used as part of the WoE (ECHA 2016a). ESRs, which record relevant endpoint information, are created for each study, such as an *in vitro* or *in vivo* study (ECHA 2016a). However, in order to meet ECHA's WoE criteria, a registrant needs to provide sufficient evidence, which means that multiple ESRs should be included as well as proper documentation (ECHA 2016a). For our checklist, we included the online ESR elements of "objectives, methods, and conclusions" for each supporting material. We also examined whether applicants provided full documentation for test study results (ECHA 2016a).

Reliability. The checklist's second criterion judged whether an applicant's WoE analysis established the reliability of the WoE studies. Under ECHA guidance, reliability is defined by whether the study is "relevant," "adequate," and "reliable." (ECHA 2016a). In their 2016 practical guide, ECHA defines these scoring terms by their level of appropriateness to the hazard endpoint of interest, using a Klimisch score to rate the reliability of the study (ECHA 2016a). Reliability and relevance were also identified as fundamental WoE components in the 2017

European Safety Authority (EFSA) Guidance on weight of evidence (EFSA Scientific Committee et al. 2017). Due to the importance placed on these WoE principles by the different regulatory agencies, we adopted these scoring terms as sub-criteria in our checklist, but excluded the Klimisch score, which we considered too prescriptive. ECHA also requires that companies gather as much information is available on the chemical (ECHA 2016a). We therefore added 'quantity' to our checklist as a sub-criterion to add rigor and transparency. In addition, because our completeness review covers higher-tier endpoints, such as CMRs, which have chronic and acute dose effects on human health and the environment, we counted consistency of results as well as severity and type of effects towards this checkpoint.

Lines of evidence. Because an organizational framework for assembling this information was missing in ECHA's WoE guidance, we turned to Rhomberg et al. (2013) and Martin et al.'s (2018) reviews on WoE framework and best practices. According to Martin et al. (2018), lines of evidence (LOE) are a useful grouping tool for similar information when assessing a substance's hazard. Furthermore, structured tables can help to present evidence. The National Resource Council (NRC) advises using structured tables so that different types of infromationc an be organized into "individual data streams" that connect to the areas of studies (NRC 2018). Both 'LOE' and 'structured tables' were thus added as sub-criteria. Since WoE draws on multiple LOEs for integration, we added these aspects to the third criteria of our checklist. However, we excluded mechanistic basis (MOA) into data assembly due to the limited number of existing MOAs (NRC 2018).

Conflicting results. When lines of evidence contain conflicting results, results need to be considered in a weighted manner (Rhomberg et al. 2013). Thus, we added 'conflicting results' as the fourth criterion in our checklist to consider the question of relative weight or strength of

evidence. However, aside from stating that high quality *in vivo* and *in vitro* results and studies should receive greater weight than QSAR results (ECHA 2016a), ECHA provided no further steps for rating and weighting these results. For a more explicit weight of evidence framework, we turned to Suter et al.'s (2017) assessment, which recommends the use of scoring tables. According to Suter et al., a scoring table, which is based on general criteria, such as "reliability" and "strength" of information, applies weighting with symbols such as "+, -, 0" to test the hypothesis on the chemical (Suter et al. 2017). Included in this step are addressing differences and inconsistencies in information. This step thus addresses risk factors that include "uncertainty" as well as data integration on points such as "bias," "rigor" and "cohesion" across studies (Rhomberg et al. 2013). For our checklist, we adopted a more flexible version of Suter et al.'s scoring table. We accepted any table that indicated and/or compared hazardous endpoints. For conflicting results, we accepted if an applicant explained any ambiguities and discrepancies (Suter et al. 2017). Because a regulator must be able to effectively infer the subject's impact, or in our case, the alternative chemical (Martin et al. 2018), we included the inference of any health effects from weighting to our checklist.

Final assessment. Finally, all evidence, which has been carefully classified and weighed, needs to be integrated into a final assessment based on expert judgment. ECHA notes that a WoE expert must have knowledge in the "relevant endpoints" and "study methods," and must be able to make scientific judgments (ECHA 2016a). Rhomberg et al. (2013) describes a WoE expert as someone who is a specialized in toxicology, epidemiology or methodology. Yet, ECHA does not lay out any expectations as to how this expert should be identified in a REACH AoA. While "prescriptive reporting templates" have been discussed as a way to systemize collective expert judgement, this type of large-scale regulatory change is beyond the scope of this paper (Martin et

al. 2018). For our study on AoAs, the “Conclusions” and “Reduction of Overall Risks” sections represent this final assessment, which we coded as “Assess overall WoE package.” To determine whether expert judgment was used, we looked for any detailed discussion in the form of a conclusion that considered the reliability, relevance and adequacy of WoE information, which has been integrated and compared, and assigned a weight to each piece of data ((ECHA 2016a). For this criterion, we examined the extent to which applicants drew conclusions on the safety of the alternative relative to the Annex XIV chemical.

Appendix S4

Sub-criteria coding for higher-tier endpoints

Table 4-1S. Sub-criteria coding for higher-tier endpoints used in Weight of Evidence.

Data codes	Sub-criteria codes
Higher-tier endpoint (0=No WoE, 1=yes WoE, 2=combination, 3=non-applicable)	0. QSAR but no WoE used;
	1. QSAR and yes WoE used;
	2. Combination of yes WoE and no WoE;
	3. No QSAR predictions made for this specific endpoint or property so WoE analysis is not relevant; or it is unknown if the prediction is from a QSAR.
Criteria[+] (0= no criteria, 1=one criteria, 2=two criteria, 3=3 criteria, 4=4 criteria, 5=all criteria, 6=non-applicable)	0. None of the 6 criteria was addressed in the WoE; this was not a "robust" summary but just a summary;
	1. At least one of the criterion was addressed in the WoE;
	2. At least two of the criteria were addressed in the WoE;
	3. At least three of the criteria were addressed in the WoE;
	4. Greater than or equal to four criteria were addressed in WoE;
	5. All criteria were addressed in WoE;
	6. Endpoint data not relevant to WoE e.g. evaluates potential alternative or info not used in WoE context or no WoE; the six criteria are: 1) Robust study summary, 2) fully documented, 3) objectives, 4) methods, 5) results.
Robust summary (1=objectives, 2=methods, 3=results, 4=conclusions, 5=documentation, 6=non-applicable)	1. Objectives of all test studies in RSS;
	2. Methods of all test studies in RSS;
	3. Results of all test studies in RSS;
	4. Conclusions of a full study report in RSS;
	5. Provides documentation (i.e. copies of the studies);
	6. Non-applicable endpoint data not relevant to WoE e.g. evaluates potential alternative or info not used in WoE context.
Assess reliability, relevance, adequacy, quantity (1=reliability, 2=relevance, 3=adequacy, 4=quantity, 5=consistency, 6=severity effects, 7=non-applicable)	1. Reliability: clarity and plausibility of the finding (ECHA 2016);
	2. Relevance: data and tests are appropriate for a particular hazard identification (ECHA 2016);
	3. Adequacy: usefulness of data for hazard/risk assessment purposes (ECHA 2016);
	4. Quantity: number of sources (ECHA 2016);
	5. Consistency of results especially within lines of evidence and categories of alternative data e.g. QSAR (ECHA 2016);
	6. Severity of the type of effects of concern (ECHA 2016);
	7. Non-applicable: endpoint data not relevant to WoE e.g. evaluates potential alternative or info not used in WoE context.

(Table 4-1S continued)

Data codes	Sub-criteria codes
Lines of evidence or structured evidence tables (1=lines of evidence, 2=tables, 3=non-applicable)	1. Lines of evidence (LOE) set of relevant items of information of similar type grouped to assess a hypothesis (Martin et al 2018);
	2. Tables displays individual pieces of evidence or categories of evidence for a hypothesis, such as the types shown here, scored with respect to properties or sub- properties and the overall weight (Suter 2017);
	3. Non-applicable; endpoint data not relevant to WoE e.g. evaluates potential alternative or info not used in WoE context.
Conflicting results (1=conflicting results, 2=scoring table, 3=health effect, 4=non-applicable)	1. Addresses conflicting results Note that high quality in vivo (read-across information) and in vitro data would generally carry more weight in the decision than a QSAR or an in-house in vitro method.;
	2. Scoring results: a weighting system e.g. + and - symbols to represent evidence that, respectively, supports, weakens, or has no effect on the credibility of a hypothesis (must include the endpoint in question) (Suter 2017);
	3. Health effects from endpoint be inferred from any weighting;
	4. Non-applicable; endpoint data not relevant to WoE e.g. evaluates potential alternative or info not used in WoE context.
Assesses overall package (1= assess, 2= scientifically argued, 3=expert judgment, 4=non-applicable)	1. Composed an assessment of overall hazard of alternative of interest factoring in this endpoint e.g. Conclusions or Reduction of Overall Risk;
	2. Scientifically argued: conclude whether the combined evidence is enough to draw a conclusion about the properties or the potential effects of the substance (ECHA 2016);
	3. Expert judgment (considers the reliability, relevance and adequacy, integrating and comparing different pieces of information and assigning a weight to each piece of data) (ECHA 2016);
	4. Non-applicable: endpoint data not relevant to WoE e.g. evaluates potential alternative or info not used in WoE context.

[+] Depending on the number of sub-criteria, not all of the criteria scale was applied.

Table 4-2S. Danish EPA advisory self-classification of hazardous substance by QSAR battery models.

Endpoint	Advisory classification	Description of CLP classification	Battery models
MUTAGENICITY			
Mutagenicity	Muta. 2	Suspected of causing genetic defects	Bacterial reverse mutation test (Ames test in S. typhimurium in vitro)
			Chromosome aberrations in CHO cells (in vitro) , commercial model from MultiCASE
			Chromosome aberrations in CHL cells (in vitro)
			Mutations in thymidine kinase locus in mouse lymphoma cells (in vitro)
			Mutations in HGPRT locus in CHO cells (in vitro)
			Micronucleus test in mouse erythrocytes (in vivo)
			Comet assay in mouse (in vivo)
CARCINOGENICITY			
Carcinogenicity	Carc. 2	Suspected of causing cancer	FDA RCA cancer male rat (*in vivo*), commercial models
			FDA RCA cancer female rat (*in vivo*) , commercial models
			FDA RCA cancer male mouse (*in vivo*) , commercial models
			FDA RCA cancer female mouse (*in vivo*) , commercial models
			Bacterial reverse mutation test (Ames test in S. typhimurium in vitro)
			Chromosome aberrations in CHO cells (*in vitro*)
			Chromosome aberrations in CHL cells (*in vitro*)
			Mutations in thymidine kinase locus in mouse lymphoma cells (in vitro)
			Mutations in HGPRT locus in CHO cells (in vitro)
REPRODUCTIVE TOXICITY			
Reproductive toxicity	Repr. 2	Suspected of damaging fertility or the unborn child	Teratogenic potential in Humans
			Dominant lethal mutations in rodents (*in vivo*)

[§]All data comes from the Danish EPA's advisory list for self-classification of hazardous substances (EPA 2018b).

Appendix S6

Consultation numbers

- 0005-02 methyl centralite for Genetic toxicity: *In vivo* – Chromosomal effect, the applicant wrote "equivocal" instead of inconclusive, and also reported that the results were in AD. But after querying the predictions myself in the DQD, we saw that results for that endpoint were out of domain (DTU FOOD et al. 2019). The applicant therefore appeared to incorrectly report the results.
- 0005-02 ethyl centralite for genotoxicity in vivo- For sister chromatid exchange in mouse bone marrow cells, it appears that the applicant reported the same QSAR prediction twice. Also the applicant wrote equivocal in domain, but the report says: "positive out of domain (battery)" (DTU FOOD et al. 2019).
- 0005-02 Akardite II: QSAR prediction for unscheduled DNA repair response based on a mouse bone marrow sister chromatid exchange assay, from the Danish (Q)SAR Database: Danish (Q)SAR Database battery result reported inconclusive out of domain but applicant stated equivocal in domain. And all 3 QSARs (Leadscope, Multicase SciQSAR) had "Pos out of domain" (DTU FOOD et al. 2019).
- 0005-02 Akardite III: Applicant reported equivocal instead of inconclusive and said in domain when the DQD report said out of domain. QSAR prediction for Chinese Hamster Ovary (CHO) cell assay for chromosome aberration test, from the Danish (Q)SAR Database (DTU FOOD et al. 2019).
- 0005-02 DOZ: QSAR prediction for chromosomal aberration in an *in vitro* COMET assay in mouse cells, from the DQD: we queried the predictions for DOZ and could not find a COMET assay for mouse cells in the DQD's *in vitro* Genotoxicity Endpoints

results. QSAR prediction for chromosome aberration in a Chinese Hamster Ovary (CHO) assay, from the DQD reported Negative, in domain for Chromosome Aberrations in Chinese Hamster Ovary (CHO) Cells (DTU FOOD et al. 2019).

- 0005-02 TBC: QSAR prediction for chromosome aberration in Chinese hamster ovary (CHO) cells from the DQD: applicant reported equivocal but still in domain. Results from the DQD reported negative in domain (DTU FOOD et al. 2019).

Appendix S7

WoE completeness by CMR and PBT/vPvB endpoints

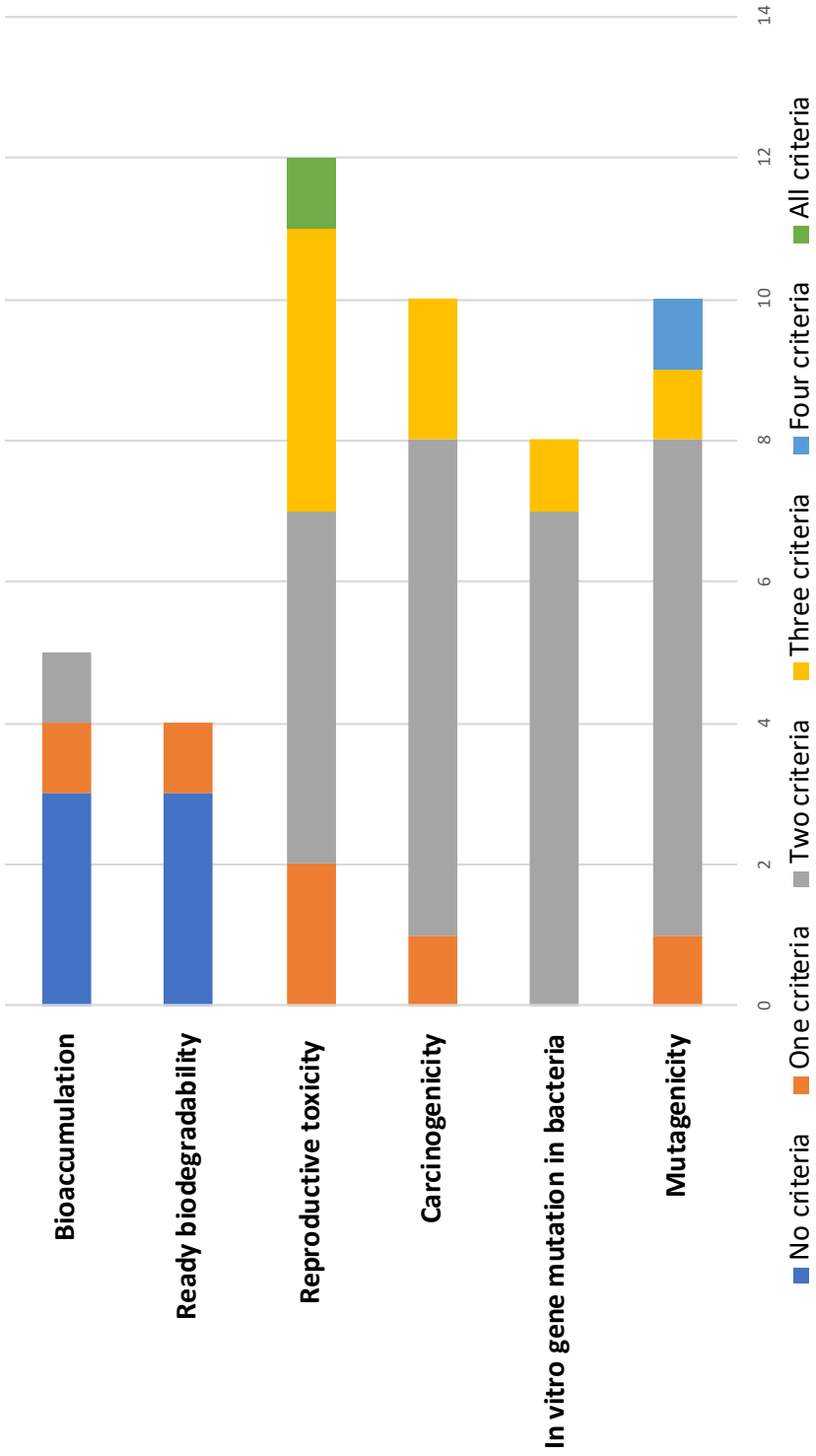


Figure 4-2S. Differences in number of criteria met in criteria checklist by QSARs used in WoE by CMR/PBT vPvB endpoints.

5. Chapter 5: Conclusions

Although QSARs have been assessed under registration, no one has evaluated QSARs used under authorization. More specifically, ECHA has yet to assess how AoA applicants use alternative methods such as QSARs, which are important for filling in data gaps, identifying positive alerts for higher-tier endpoints, and for strength of evidence systems such as WoE and ITS. This study, therefore, serves as an initial foundation to help strengthen future regulatory decisions within REACH authorization and more broadly for the future of alternatives assessment.

The research I conducted begins by setting up a context for QSAR use, which I then follow with a procedural examination of QSARs used in REACH AoAs. To advance the work related to the usage and purpose of QSARs in chapters three and four, I initiated a formal background training on how QSARs operate and are developed as regulatory models. Namely, I immersed myself in DTU's QSAR lab and learned the nature and use of QSARs through the development of four new QSARs for nuclear receptor CAR. Chapter three drew from this understanding of the proper use of QSARs and focused on profiling QSAR usage by applicants in REACH AoAs. Chapter four then provided a more in-depth analysis, revealing issues regarding transparency and WoE criteria fulfillment, in addition to the potential benefits of ITS battery QSAR predictions to identify potential CMR substances.

Future research needs

Limited use of QSARs

This analysis demonstrated limited QSAR use in AoAs. In this study, only 25 of the 189 AoAs utilized QSARs. One reason for limited QSAR use could be that in emphasizing the use of QSARs in a WoE context under ECHA's general approach to QSARs under REACH,

applicants may have shied away from information on their alternative for which QSAR predictions gave the only available data. In general, ECHA's guidance seems to accept QSARs mostly within the context of WoE, which requires all available historical and testing information on the substance, with a priority on experimental data. In this study, I found evidence of applicants prescribing to this guidance. In AoA consultation number 0078-01, the applicants *Dow Italia Srl* and *Rohm and Haas France S.A.S.* excluded the $PNEC_{freshwater}$ value for methylene chloride (DCM) from ECHA CHEM under individual submission because it was based solely on QSAR predictions (Dow Italia Srl and Rohm and Haas France S.A.S. 2016). The applicant Eli Lilly S.A. - Irish Branch for AoA consultation number 0077-01, who also assessed DCM, provided the same reasoning (Eli Lilly S.A. Irish Branch n.d.). At the same time, ECHA has issued guidance that supports the use of standalone QSAR predictions, at least under ideal conditions (ECHA 2008). In other AoAs, such as consultation numbers 0006-01 and 0005-02, QSAR data were the only information available (DEZA A.S. n.d.-a; Sasol-Huntsman GmbH & Co. KG n.d.). In these instances, standalone QSAR use may seemingly be interpreted as an acceptable approach to QSAR use in REACH though missing QSAR documentation hampers the ability of ECHA to corroborate whether the condition was "ideal" or not.

These conflicting regulatory points of views on whether QSARs can be used as standalone tools point to a larger regulatory debate: Does the greater good lie with QSARs being used with the most conservative approach? Or will pushing the boundaries of how QSAR predictions can be used ultimately prevent untested, toxic chemicals from escaping regulation? Who could help push this debate is industry. Results from an ORCHESTRA online questionnaire on QSARs allude to this point (Mays et al. 2012). If certain industries can successfully use certain QSAR tools for REACH then others may view this as an example of a successful

application of QSARs and in turn use QSARs themselves. Though not perfectly correlated, I observed a similar trend in the use of the Danish (Q)SAR Database. With its user-friendly interface, accurate predictions, and diverse database of over 600,000 substances (DTU FOOD et al. 2019), the Danish (Q)SAR Database was the most used out of all of the QSAR sources on the "default" list for most endpoints.

As part of this debate, definitive answers pertaining to the use of QSARs under REACH may change as QSARs become more advanced (Luechtefeld et al. 2018), and as regulations continue to evolve (Benigni et al. 2007; NRC 2014b; Worth 2010a). However, for now, unless more companies are encouraged to use QSARs, particularly under authorization, this debate may languish unnecessarily, and standalone QSAR use may continue to be the rare exception. But with the current awareness that a lack of data on alternative chemicals can lead to misinformed decision-making (Jacobs et al. 2016) as well as impact the consistency in the alternatives analysis process (Tickner et al. 2013), it is worth considering the implications of companies operating on an inconsistent understanding of QSAR requirements, and leaving out standalone QSAR data even if it is the only data that points to a safer alternative. I, therefore, recommend that ECHA consider producing more specific guidance that provides a "fit-for-purpose" approach to QSAR use for authorization and perhaps more generally for REACH to help applicants understand how to use QSAR data in AoAs. This guidance should establish the specific circumstances in which standalone use is required, those where is standalone is permitted, and those where standalone is disallowed.

QSAR documentation

This study revealed that for all 25 AoAs, applicants did not attach required supporting QMRF and QPRF documentation for QSAR predictions. Missing QMRF and QPRF

documentation poses a potential obstacle if regulators and stakeholders are trying to determine the quality of QSAR tools and data in AoAs. In addition, results from the QPRF information review showed that in most instances, an equivalent method of providing QPRF criteria in AoAs did not exist though applicants tended to provide the majority of priority criteria information as part of the AoA.

Given the missing documentation, one of my conclusions pointed towards the need for an explicit mandate to include QMRFs and QPRFs with AoAs if QSARs are used. Namely, because without setting clearer requirements under authorization, companies, which may already be disincentivized to find a safer alternative(s), particularly if their real motivation is the continued use of their incumbent chemical, may never inquire as to their obligation to include proper documentation, if by doing so, makes them more accountable. In fact, AoAs may already have been submitted that do not properly examine their QSAR models or provide suitable support in their WoE using QSARs. One of the most significant implications is that an Annex XIV chemical could get authorized based on inaccurate or incomplete information. While these points are hypothetical, I believe they lie on a possible trajectory of improperly documented AoAs that are submitted to ECHA.

WoE using QSARs

Only a limited number of AoAs used WoE with QSARs. Based on this study, WoE completeness depended on the main criteria and hazard endpoint. In addition, the completeness review showed that certain sub-criteria played a role in whether the main criteria were fulfilled or not, for example, expert judgment. For instance, as expert judgment comes into play, having a clearly explained WoE in an acceptable format becomes increasingly important. After all, regulators rely on the underlying logic of the author's qualitative conclusions to make sound

decisions (Rhomberg et al. 2013). In fact, according to the U.S. EPA, WoE hinges on expert decision-making (ECHA 2016a; U.S. EPA 2011). ECHA notes that a WoE expert must have knowledge of the "relevant endpoints" and "study methods," and be able to make scientific judgments (ECHA 2016a). Likewise, Rhomberg et al. (2013) describes a WoE expert as someone who is a specialized in "toxicology" or "epidemiology" or an equivalent field. Yet, REACH WoE guidance is missing any expectations as to how this expert should be identified in an AoA. Thus, including the names of the WoE experts along with contact information, and background could help shed light on their qualifications and possibly give a starting point for where applicants could improve their WoE completeness. At the same time, given the unique nature of WoE in regulatory submissions, which "inevitably requires subjective judgment," room needs to be made for the possibility that WoE conclusions may come out differently in similar cases (Suter et al. 2017).

ITS

In this study, I compared all 24 AoAs from the previous sample that used QSARs to assess final, selected alternatives (Chapter 2) with the 2019 Danish EPA QSAR Dangerous Properties list to screen for potential CMR substances that AoA applicants might have missed. Most AoAs did not have overlapping alternatives with chemicals on the Danish EPA advisory classification list. However, I found overlap for four chemicals between the Danish EPA advisory classification list and my CMR list, for which the Danish EPA assigned a Repr. 2 advisory classification for all four chemicals. More importantly, Danish EPA battery QSAR predictions did not consistently compare with the applicants' conclusions. While applicants did identify reproductive toxicity for the alternative bis(2-ethylhexyl) adipate (DEHA), no identification for reprotxicity was made by bis(2-propylheptyl) phthalate (DPHP), acetyl tributyl

citrate (ATBC), or tributyl citrate (TBC) for reproductive toxicity under normal exposure. Thus, the use of ITS not only has the potential to provide more information, but also, as demonstrated by my evaluation, to provide contrasting outcomes. Advancements in *in silico* models, such as QSARs, have enhanced the accuracy of predictions to exceed, at least under ideal conditions, animal tests (Luechtefeld et al. 2018), which in the past have been considered the "gold standard" for reliable results (Hartung et al. 2013). ITS is thus viewed as the next step forward in regulatory toxicologist assessment, which for future AoAs and even registration dossiers, could mean a greater breadth of information and perhaps a more complete hazard profile for alternatives being assessed. Though ITS does not have an inherent regulatory purpose, ITS is increasingly used for regulatory decision-making, oftentimes in a WoE context (Worth 2010a), which has the potential for far-reaching benefits in cases where experimental information is limited or missing.

Summary

This research highlights multiple areas of QSAR use in REACH AoAs, which contribute to the determination of safety of alternatives. Initially, four CAR models for antagonism and agonism were developed in Denmark at DTU to give a background and advance the analysis on QSAR usage in chapters three and four. In chapter three, I identified trends in model use and frequency upon data-mining 189 AoAs, which I further explored in chapter four. Results from this second study gave more substantive answers regarding the limited use of WoE contexts for QSARs as well as the small number of AoAs that met the five WoE criteria. Finally, a comparative analysis between the Danish EPA's advisory classification for overlapping chemicals in this study indicated that more information could be made available for CMR substances when using ITS battery QSAR models.

Even though this research was extensive, additional studies are needed to analyze AoAs from May 2017 to the present. More specifically, the small sample size of AoAs may not be representative of the work performed by a larger sample of applicants. For this study, most of the AoAs in my sample were conducted by companies or consultants who had little toxicological knowledge. As a result, their degree of training would have made it nearly impossible for the companies or consultants to know how to use a QSAR or to know to take advantage of available predictions. So perhaps the real issue is how to convert QSARs into useful tools like ECOSAR or the Danish (Q)SAR Database. Furthermore, with a larger sample size, additional comparisons could be drawn with the Danish EPA advisory classifications. As a possible solution, generating regulatory guidance for AoAs that defines rules to follow upon adopting QSAR strategies might help applicants understand how to maximize QSAR opportunities in their AoAs. Some of the methodologies employed in this study such as analyzing QSAR use by QSAR source could also be applied to QSAR use in registration dossiers. Currently, ECHA analyzes QSAR usage in registration by endpoint study records (ESR). However, tracking the types of QSAR sources could provide further background information for limited or frequent QSAR use for certain endpoints.

If monitoring the function of QSAR sources for filling in data gaps and generating high quality information through WoE, ITS and proper documentation help to provide more information in AoAs, they should be monitored with equal if not more stringency in other parts of REACH as well. Additionally, it is important to focus on how QSAR data contributes to assessing chemicals for higher-tier endpoints where vulnerable populations and wildlife are the most susceptible to harm. While more research is needed to assess QSAR usage under REACH authorization, ignoring the implications given in this study could subvert the purpose of AoAs

and influence the attitude of industry, which could impact towards other forms of alternatives assessment employed in other parts of the world.

References

[ECHA] European Chemicals Agency. 2008. Guidance on information requirements and chemical safety assessment. Chapter R.6: QSARs and grouping of chemicals. Available: https://echa.europa.eu/guidance-documents/guidance-on-information-requirements-and-chemical-safety-assessment [accessed January 22, 2019].

[ECHA] European Chemicals Agency. 2019. Authorisation. Available: https://echa.europa.eu/substances-of-very-high-concern-identification-explained [accessed November 20 2019].

ACD/Labs. 2019. ACD/Percepta Portal. Available: https://www.acdlabs.com/products/percepta/ [accessed November 11 2019].

Akhtar A. 2015. The flaws and human harms of animal experimentation. *Cambridge Quarterly of Healthcare Ethics* 24:407-419.

AkzoNobel Pulp and Performance Chemicals. 2015. Use 1: Use of sodium dichromate as an additive for suppressing parasitic reactions and oxygen evolution, pH buffering and cathode corrosion protection in the electrolytic manufacture of sodium chlorate, with or without subsequent production of chlorine dioxide; Use 2: Use of sodium dichromate as an additive for suppressing parasitic reactions and oxygen evolution, pH buffering and cathode corrosion protection in the electrolytic manufacture of potassium chlorate. Consultation number 0041-01. European Chemicals Agency. Available: https://echa.europa.eu/public-consultations [accessed November 12, 2019].

Alexander SP, Cidlowski JA, Kelly E, Marrion N, Peters JA, Benson HE, et al. 2015. The concise guide to PHARMACOLOGY 2015/16: Nuclear hormone receptors. *British journal of pharmacology* 172:5956, doi: https://doi.org/10.1111/bph.1335.

American Chemical Society (ACS). 2019. CAS REGISTRY and CAS Registry Number FAQs. Available: https://www.cas.org/support/documentation/chemical-substances/faqs [accessed August 30 2019].

Angrish M, Chorley B. 2018. AOP: 58: NR1I3 (CAR) suppression leading to hepatic steatosis. Available: https://aopwiki.org/aops/58 [accessed September 12 2019].

ARKEMA. 2013a. Analysis of Alternatives Non-Confidential Report: Industrial use in polymer processing by calendering, spread coating, extrusion, injection moulding to produce PVC articles [except erasers, sex toys, small household items (<10cm) that can be swallowed by children, clothing intended to be worn against the bare skin; also toys, cosmetics and food contact material (restricted under other EU regulation)]. 0002-02. European Chemicals Agency. Available: https://echa.europa.eu/public-consultations.

ARKEMA. 2013b. Analysis of Alternatives Non-Confidential Report: Formulation of DEHP in compounds, dry-blends and plastisol formulations. Industrial use in polymer processing by calendering, spread coating, extrusion, injection moulding to produce PVC articles [except erasers, sex toys, small household items (<10cm) that can be swallowed by children, clothing intended to be worn against the bare skin; also toys, cosmetics and food contact material (restricted under other EU regulation)]. 0002-01. European Chemicals Agency. Available: https://echa.europa.eu/public-consultations.

Auld DS, Inglese J. 2018. Interferences with luciferase reporter enzymes. In: Assay Guidance Manual [Internet], (Sittampalam GS, Grossman A, Brimacombe K, Coussens NP, Arkin M, Auld D, et al., eds). Bethesda (MD): Eli Lilly & Company and the National Center for Advancing Translational Sciences.

Barley M, McFiggans G. 2010. The critical assessment of vapour pressure estimation methods for use in modelling the formation of atmospheric organic aerosol. *Atmospheric Chemistry and Physics* 10:749-767.

Bassan A, Worth AP. 2008. The integrated use of models for the properties and effects of chemicals by means of a structured workflow. *QSAR & Combinatorial Science* 27:6-20, doi: https://doi.org/10.1002/qsar.200710119.

Benfenati E, Diaza RG, Cassano A, Pardoe S, Gini G, Mays C, et al. 2011. The acceptance of in silico models for REACH: Requirements, barriers, and perspectives. *Chemistry Central Journal* 5:58.

Benfenati E. 2012. Theory, guidance and applications on QSAR and REACH. First ed. Milan, Italy: Istituto di Ricerche Farmacologiche "Mario Negri". 1-175.

Benfenati E, Pardoe S, Martin T, Gonella Diaza R, Lombardo A, Manganaro A, et al. 2013. Using toxicological evidence from QSAR models in practice. *ALTEX - Alternatives to animal experimentation* 30:19-40, doi: https://doi.org/10.14573/altex.2013.1.019.

Benfenati E, Lombardo A, Roncaglioni A. 2018. Chapter 9: Computational toxicology and REACH. In: Computational Toxicology: Risk Assessment for Chemicals, (Ekins S, ed). Hoboken, NJ: Computational Toxicology: Risk Assessment for Chemicals, 245-268.

Benigni R, Netzeva TI, Benfenati E, Bossa C, Franke R, Helma C, et al. 2007. The expanding role of predictive toxicology: an update on the (Q)SAR models for mutagens and carcinogens. *Journal of Environmental Science and Health Part C* 25:53-97, doi: https://doi.org/10.1080/10590500701201828.

Breiman L. 1996. Bagging predictors. *Machine learning* 24:123-140, doi: https://doi.org/10.1007/BF00058655.

Caffaro Brescia S.r.l. 2015. Use of sodium dichromate as an additive for suppressing parasitic reactions and oxygen evolution, pH buffering and cathode corrosion protection in the electrolytic manufacture of sodium chlorite. Consultation number 0040-01. European Chemicals Agency. Available: https://echa.europa.eu/public-consultations [accessed November 12, 2019].

ChemAxon. 2014. ChemAxon Documentation: Calculator Plugins User's Guide Physico-chemical plugins. Available: https://docs.chemaxon.com/display/docs/Physico-chemical+plugins [accessed December 9 2019].

ChemAxon. 2019. Calculators and Predictors: High-quality physico-chemical calculations and predictions for drug discovery. Available: https://chemaxon.com/products/calculators-and-predictors [accessed December 9 2019].

Cooper II J, Saracci R, Cole P. 1979. Describing the validity of carcinogen screening tests. *British Journal of cancer* 39:87.

Cronin MT, Jaworska JS, Walker JD, Comber MH, Watts CD, Worth AP. 2003. Use of QSARs in international decision-making frameworks to predict health effects of chemical substances. *Environmental Health Perspectives* 111:1391, doi: 10.1289/ehp.5760.

Cronin MT. 2010. Prediction of harmful human health effects of chemicals from structure. In: Recent Advances in QSAR Studies Challenges and Advances in Computational Chemistry and Physics, Vol. 8, (Puzyn T, Leszczynski J, Cronin MT, eds): Springer, Dordrecht, 305-325.

Dassault Systemes Biovia. 2019. QSAR, ADMET AND Predictive Toxicology with BIOVIA Discovery Studio Datasheet. San Diego, CA: Dassault Systems. Available: http://www.3dsbiovia.com/products/datasheets/qsar-admet-and-predictive-toxicology-with-ds.pdf [accessed December 9, 2019].

DEZA A.S. 2013a. Analysis of Alternatives Non-Confidential Report: Use of DBP in propellants Formulation: Industrial use of DBP as a burning rate surface moderant, plasticiser and/or coolant in the formulation of nitrocellulose-based propellant grains. Use at industrial site: Industrial use of DBP-containing propellant grains in manufacture of ammunition for military and civilian uses, and pyrocartridges for aircraft ejection seat safety systems [includes propellants for police force ammunition and excludes propellants intended for manual, private reloading of ammunition cartridges by civilian users, i.e., licensed individual sports shooters and hunters. No direct consumer use of DBP or its mixtures is covered by this Use. . Consultation number 0004-01. European Chemicals Agency. Available: https://echa.europa.eu/applications-for-authorisation-consultation [accessed January 22, 2019].

DEZA A.S. 2013b. Analysis of Alternatives Non-Confidential Report: Formulation of DEHP in compounds, dry-blends and plastisol formulations. Industrial use in polymer processing by calendering, spread coating, extrusion, injection moulding to produce PVC articles [except erasers, sex toys, small household items (<10cm) that can be swallowed by children, clothing intended to be worn against the bare skin; also toys, cosmetics and food contact material (restricted under other EU regulation)]. Consultation number 0004-02. European Chemicals Agency. Available: https://echa.europa.eu/public-consultations [accessed January 22, 2019].

DEZA A.S. n.d.-a. Analysis of Alternatives Non-Confidential Report: Use in propellants: Sub-scenario 1: F-2: Industrial use as a burning rate surface moderant, plasticiser and/or coolant in the formulation of nitrocellulose-based propellant grains. Sub-scenario 2: IW-2: Industrial use of propellant grains in manufacture of ammunition for military and civilian uses, and pyrocartridges for aircraft ejection seat safety systems [excludes

propellants intended for manual reloading of ammunition cartridges by civilian users]. Consultation number 0005-02. European Chemicals Agency. Available: https://echa.europa.eu/public-consultations [accessed November 12, 2019].

DEZA A.S. n.d.-b. Analysis of Alternatives Non-Confidential Report: Use as an absorption solvent in a closed system in the manufacture of maleic anhydride (MA). Consultation number 0005-01. European Chemicals Agency. Available: https://echa.europa.eu/public-consultations [accessed January 22, 2019].

di Masi A, De Marinis E, Ascenzi P, Marino M. 2009. Nuclear receptors CAR and PXR: Molecular, functional, and biomedical aspects. *Molecular aspects of medicine* 30:297-343, doi: https://doi.org/10.1016/j.mam.2009.04.002.

Dong B, Saha PK, Huang W, Chen W, Abu-Elheiga LA, Wakil SJ, et al. 2009. Activation of nuclear receptor CAR ameliorates diabetes and fatty liver disease. *Proceedings of the National Academy of Sciences* 106:18831-18836, doi: https://doi.org/10.1073/pnas.0909731106.

Dow Italia Srl, Rohm and Haas France S.A.S. 2016. Analysis of Alternatives Public Version: Industrial use as a sulphonation swelling agent of polystyrene-divinylbenzene copolymer beads in the production of strong acid cation exchange resins. Consultation number 0078-01 European Chemicals Agency. Available: https://echa.europa.eu/public-consultations [accessed January 22, 2019].

Dring AM, Anderson LE, Qamar S, Stoner MA. 2010. Rational quantitative structure–activity relationship (RQSAR) screen for PXR and CAR isoform-specific nuclear receptor ligands. *Chemico-biological interactions* 188:512-525, doi: https://doi.org/10.1016/j.cbi.2010.09.018.

DTU Food (Technical University of Denmark National Food Institute). 2018a. User Manual for the Danish (Q)SAR Database. Available: http://qsarmodels.food.dtu.dk/DB_user_manual_21_12_2018.pdf [accessed November 18 2019].

DTU Food (Technical University of Denmark National Food Institute). 2018b. User Manual for the Danish (Q)SAR Database. Lyngby, Denmark: Technical University of Denmark. Available: http://qsarmodels.food.dtu.dk/DB_user_manual_21_12_2018.pdf [accessed January 3, 2019].

DTU Food (Technical University of Denmark National Food Institute), Danish EPA (Danish Environmental Protection Agency), Nordic Council of Ministers, ECHA (European Chemicals Agency). 2019. Danish (Q)SAR Database. Available: http://qsar.food.dtu.dk [accessed November 18 2019].

EC (European Commission). 2019. Growth: Internal Market, Industry, Entrepeneurship and SMEs: How REACH Works. Available: https://ec.europa.eu/growth/sectors/chemicals/reach/about_en [accessed June 8 2019].

EC JRC EURL ECVAM (European Commission Joint Research Centre European Union Reference Laboratory for Alternatives to Animal Testing). 2019. JRC QSAR Model Database. Available: https://qsardb.jrc.ec.europa.eu/qmrf [accessed November 17 2019].

ECHA (European Chemicals Agency). 2008. Guidance on information requirements and chemical safety assessment. Chapter R.6: QSARs and grouping of chemicals. Helsinki, Finland. Available: https://echa.europa.eu/documents/10162/13632/information_requirements_r6_en.pdf [accessed September 15, 2019].

ECHA (European Chemicals Agency). 2010. Practical guide 2: how to report weight of evidence. Helsinki, Finland.

ECHA (European Chemicals Agency). 2011a. Guidance on information requirements and chemical safety assessment Part B: Hazard Assessment. Helsinki, Finland. Available: https://echa.europa.eu/guidance-documents/guidance-on-information-requirements-and-chemical-safety-assessment [accessed September 1, 2019].

ECHA (European Chemicals Agency). 2011b. Guidance on the Preparation of an Application for Authorization. Hesinki, Finland. Available: https://echa.europa.eu/guidance-documents/guidance-on-reach [accessed January 22, 2019].

ECHA (European Chemicals Agency). 2012. How to report robust study summaries Practical Guide 3 Version 2.0 Helsinki, Finland. Available: https://echa.europa.eu/documents/10162/13643/pg_report_robust_study_summaries_en.pdf/1e8302c3-98b7-4a50-aa22-f6f02ca54352 [accessed August 3, 2019].

ECHA (European Chemicals Agency). 2013. Evaluation under REACH: Progress Report 2013. Helsinki, Finland. Available: https://echa.europa.eu/documents/10162/13628/evaluation_report_2013_en.pdf [accessed January 22, 2019].

ECHA (European Chemicals Agency). 2014. The Use of Alternatives to Testing on Animals for the REACH Regulation: Second report under Article 117(3) of the REACH Regulation. Helsinki, Finland. Available: https://echa.europa.eu/documents/10162/13639/alternatives_test_animals_2014_en.pdf [accessed January 22, 2019].

ECHA (European Chemicals Agency). 2016a. Practical guide: How to use alternatives to animal testing to fulfil your information requirements for REACH registration Version 2.0.

Helsinki, Finland. Available: https://echa.europa.eu/documents/10162/13655/practical_guide_how_to_use_alternatives_en.pdf [accessed January 22, 2019].

ECHA (European Chemicals Agency). 2016b. Practical guide: How to use and report (Q)SARs Version 3.1. Helsinki, Finland. Available: https://echa.europa.eu/documents/10162/13655/pg_report_qsars_en.pdf [accessed January 22, 2019].

ECHA (European Chemicals Agency). 2016c. Report on the Operation of REACH and CLP 2016. Helsinki, Finland. Available: https://echa.europa.eu/documents/10162/13634/operation_reach_clp_2016_en.pdf [accessed January 22, 2019].

ECHA (European Chemicals Agency). 2017a. The use of alternatives to testing on animals for the REACH Regulation: Third report under Article 117(3) of the REACH Regulation. Helsinki, Finland. Available: https://echa.europa.eu/documents/10162/13639/alternatives_test_animals_2017_en.pdf [accessed June 18, 2019].

ECHA (European Chemicals Agency). 2017b. Guidance on Information Requirements and Chemical Safety Assessment: Chapter R.7b: Endpoint specific guidance. Helsinki, Finland. Available: https://echa.europa.eu/web/guest/guidance-documents/guidance-on-information-requirements-and-chemical-safety-assessment [accessed January 22, 2019].

ECHA (European Chemicals Agency). 2017c. Evaluation under REACH: Progress Report 2017: 10 years of experience. Helsinki, Findland: European Chemicals Agency. Available: https://echa.europa.eu/documents/10162/13628/evaluation_under_reach_progress_en.pdf/24c24728-2543-640c-204e-c61c36401048 [accessed July 26, 2019].

ECHA (European Chemicals Agency). 2017d. Guidance on Information Requirements and Chemical Safety Assessment: Chapter R.7a: Endpoint specific guidance. Helsinki, Finland. Available: https://echa.europa.eu/web/guest/guidance-documents/guidance-on-information-requirements-and-chemical-safety-assessment [accessed January 22, 2019].

ECHA (European Chemicals Agency). 2018. QSAR models. Available: https://echa.europa.eu/support/registration/how-to-avoid-unnecessary-testing-on-animals/qsar-models [accessed December 7 2019].

ECHA (European Chemicals Agency). 2019a. About Us. Available: https://echa.europa.eu/about-us [accessed November 20 2019].

ECHA (European Chemicals Agency). 2019b. Understanding CLP. Available: https://echa.europa.eu/regulations/clp/understanding-clp [accessed August 27 2019].

ECHA (European Chemicals Agency). 2019c. ECHA Newsletter: What happens with potential chemicals of concern? Available: https://newsletter.echa.europa.eu/home/-/newsletter/entry/what-happens-with-potential-chemicals-of-concern- [accessed December 8 2019].

ECHA (European Chemicals Agency). 2019d. Understanding REACH. Available: https://echa.europa.eu/regulations/reach/understanding-reach [accessed June 8 2019].

ECHA (European Chemicals Agency). 2019e. Authorization. Available: https://echa.europa.eu/substances-of-very-high-concern-identification-explained [accessed November 20 2019].

ECHA (European Chemicals Agency). 2019f. ECHA Regulations: REACH Registration Information Requirements. Available: https://echa.europa.eu/regulations/reach/registration/information-requirements [accessed January 22 2019].

ECHA (European Chemicals Agency). 2019g. Hot topics: Endocrine disruptors. Available: https://echa.europa.eu/hot-topics/endocrine-disruptors [accessed April 25 2019].

ECHA (European Chemicals Agency). 2019h. Adopted opinions and previous consultations on applications for authorisation. Available: https://echa.europa.eu/regulations/reach/authorisation/applications-for-authorisation [accessed November 20 2019].

ECHA (European Chemicals Agency). 2019i. Community Rolling Action Plan. Available: https://echa.europa.eu/information-on-chemicals/evaluation/community-rolling-action-plan [accessed August 31 2019].

EFSA Scientific Committee, Hardy Anthony, Benford Diane, Halldorsson Thorhallur, Jeger Michael John, Knutsen Helle Katrine, et al. 2017. Guidance on the use of the weight of evidence approach in scientific assessments. *EFSA Journal* 15:e04971.

Eli Lilly S.A. Irish Branch. n.d. Analysis of Alternatives Public Version: Industrial use as a reaction medium and a solvating agent in mediating subsequent chemical transformation reactions leading to the manufacture of an Active Pharmaceutical Ingredient, Raloxifene Hydrochloride. Consultation number 0077-01. European Chemicals Agency. Available: https://echa.europa.eu/public-consultations [accessed January 22, 2019].

EPA (Danish Environmental Protection Agency). 2009. The advisory list for self-classification of dangerous substances. Part Environmental Project No. 1322 2010, (Niemelä JR, Wedebye EB, Nikolov NG, Jensen GE, Ringsted T, Ingerslev F, et al., eds). Copenhagen, Denmark: Danish EPA.

EPA (Danish Environmental Protection Agency). 2013. Use of QSAR to identify potential CMR substances of relevance under the REACH regulation: Environmental Project No. 1503,

2013. (Wedebye EB, Niemelä JR, Nikolov NG, Dybdahl M, eds). Copenhagen, Denmark: Danish Environmental Protection Agency.

EPA (Danish Environmental Protection Agency). 2018a. Brief documentation on the (Q)SAR models and the algorithms used for the advisory list for self classification. Available: https://eng.mst.dk/chemicals/chemicals-in-products/assessment-of-chemicals/the-advisory-list-for-self-classification-of-hazardous-substances/ [accessed September 19 2019].

EPA (Danish Environmental Protection Agency). 2018b. The advisory list for self classification of hazardous substances. Available: https://eng.mst.dk/chemicals/chemicals-in-products/assessment-of-chemicals/the-advisory-list-for-self-classification-of-hazardous-substances/ [accessed September 19 2019].

Ercros S.A. 2015. Use of sodium dichromate as an additive for suppressing parasitic reactions and oxygen evolution, pH buffering and cathode corrosion protection in the electrolytic manufacture of sodium chlorate with or without subsequent production of chlorine dioxide or sodium chlorite. Consultation number 0037-01. European Chemicals Agency. Available: https://echa.europa.eu/public-consultations [accessed November 12, 2019].

Estrada E. 2008. How the parts organize in the whole? A top-down view of molecular descriptors and properties for QSAR and drug design. *Mini reviews in medicinal chemistry* 8:213-221, doi: https://doi.org/10.2174/138955708783744128.

European Parliament and the Council of the European Union. 2006. Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending Directive 1999/45/EC and repealing Council Regulation (EEC) No 793/93 and Commission Regulation (EC)

No 1488/94 as well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC. Available: http://eur-lex.europa.eu/eli/reg/2006/1907/oj/eng [accessed January 22, 2019].

Evans TJ. 2015. Reproductive toxicity and endocrine disruption of potential chemical warfare agents. In: Handbook of Toxicology of Chemical Warfare Agents: Elsevier, 599-613.

Francis GA, Fayard E, Picard F, Auwerx J. 2003. Nuclear receptors and the control of metabolism. *Annual review of physiology* 65:261-311, doi: https://doi.org/10.1146/annurev.physiol.65.092101.142528.

Friedman KP, Gilbert ME, Crofton KM. 2016. AOP: 8: Upregulation of Thyroid Hormone Catabolism via Activation of Hepatic Nuclear Receptors, and Subsequent Adverse Neurodevelopmental Outcomes in Mammals Available: https://aopwiki.org/aops/8 [accessed September 12 2019].

Gao J, Xie W. 2012. Targeting xenobiotic receptors PXR and CAR for metabolic diseases. *Trends in pharmacological sciences* 33:552-558.

Garg R, Smith CJ. 2014. Predicting the bioconcentration factor of highly hydrophobic organic chemicals. *Food and chemical toxicology* 69:252-259, doi: https://doi.org/10.1016/j.fct.2014.03.035.

Gentrochema BV. n.d. Analysis of Alternatives: Use of Sodium dichromate for surface treatment of metals such as aluminium, steel, zinc, magnesium, titanium, alloys, composites, sealings of anodic films. Consultation number 0063-02. Available: https://echa.europa.eu/public-consultations [accessed January 22, 2019].

Gong H, Xie W. 2008. Chapter 7: Animal Models of Xenobiotic Nuclear Receptors and their Utility in Drug Development. In: Nuclear receptors in drug metabolism, (Xie W, ed): Wiley Online Library, 185-210.

GRUPA AZOTY ZAKŁADY AZOTOWE KĘDZIERZYN S.A. 2013a. Analysis of Alternatives Non-Confidential Report: Formulation of DEHP in compounds, dry-blends and plastisol formulations. Industrial use in polymer processing by calendering, spread coating, extrusion, injection moulding to produce PVC articles [except erasers, sex toys, small household items (<10cm) that can be swallowed by children, clothing intended to be worn against the bare skin; also toys, cosmetics and food contact material (restricted under other EU regulation)]. 0003-01. European Chemicals Agency. Available: https://echa.europa.eu/public-consultations.

GRUPA AZOTY ZAKŁADY AZOTOWE KĘDZIERZYN S.A. 2013b. Analysis of Alternatives Non-Confidential Report: Formulation of DEHP in compounds, dry-blends and plastisol formulations: Industrial use in polymer processing by calendering, spread coating, extrusion, injection moulding to produce PVC articles [except erasers, sex toys, small household items (<10cm) that can be swallowed by children, clothing intended to be worn against the bare skin; also toys, cosmetics and food contact material (restricted under other EU regulation)]. 0003-02. European Chemicals Agency. Available: https://echa.europa.eu/public-consultations.

Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. 2011. GRADE guidelines: 1. Introduction—GRADE evidence profiles and summary of findings tables. *Journal of clinical epidemiology* 64:383-394, doi: https://doi.org/10.1016/j.jclinepi.2010.04.026.

H&R Ölwerke Schindler GmbH. 2016. Analysis of Alternatives: Public Version: Industrial use as a solvent and anti-solvent of the feedstock and intermediate product streams in the combined de-waxing and de-oiling of refining of petroleum vacuum distillates for the production of base oils and hard paraffin waxes. Consultation number 0080-01. Available: https://echa.europa.eu/public-consultations [accessed January 22, 2019].

Hakkola J, Bernasconi C, Coecke S, Richert L, Andersson T, Pelkonen O. 2018. Cytochrome P450 Induction and Xeno-Sensing Receptors Pregnane X Receptor, Constitutive Androstane Receptor, Aryl Hydrocarbon Receptor and Peroxisome Proliferator-Activated Receptor α at the Crossroads of Toxicokinetics and Toxicodynamics. *Basic & clinical pharmacology & toxicology* 123:42-50, doi: https://doi.org/10.1111/bcpt.13004.

Hanser T, Barber C, Marchaland J, Werner S. 2016. Applicability domain: towards a more formal definition. *SAR and QSAR in Environmental Research* 27:865-881.

Hartung T, Luechtefeld T, Maertens A, Kleensang A. 2013. Food for thought… integrated testing strategies for safety assessments. *Altex* 30:3, doi: 10.14573/altex.2013.1.003.

Hong H, Fang H, Xie Q, Perkins R, Sheehan DM, Tong W. 2003. Comparative molecular field analysis (CoMFA) model using a large diverse set of natural, synthetic and environmental chemicals for binding to the androgen receptor. *SAR and QSAR in Environmental Research* 14:373-388, doi: https://doi.org/10.1080/10629360310001623962.

Honkakoski P, Sueyoshi T, Negishi M. 2003. Drug-activated nuclear receptors CAR and PXR. *Annals of medicine* 35:172-182, doi: https://doi.org/10.1080/07853890310008224.

Hou T, Xia K, Zhang W, Xu X. 2004. ADME evaluation in drug discovery. 4. Prediction of aqueous solubility based on atom contribution approach. *Journal of chemical information and computer sciences* 44:266-275, doi: https://doi.org/10.1021/ci034184n.

Huang R. 2016. A quantitative high-throughput screening data analysis pipeline for activity profiling. In: High-Throughput Screening Assays in Toxicology Methods in Molecular Biology, Vol. 1473, (Zhu H., Xia M., eds). New York, NY: Human Press, 111-122.

Huang W, Zhang J, Washington M, Liu J, Parant JM, Lozano G, et al. 2005. Xenobiotic stress induces hepatomegaly and liver tumors via the nuclear receptor constitutive androstane

receptor. *Molecular endocrinology* 19:1646-1653, doi: https://doi.org/10.1210/me.2004-0520.

IdeaConsult Ltd. 2018. Toxtree- Toxic Hazard Estimation by decision tree approach. Available: http://toxtree.sourceforge.net [accessed December 9 2019].

INEOS Styrenics Netherlands BV, INEOS Styrenics Ribecourt SAS, INEOS Styrenics Wingles SAS, Synthos Dwory 7 spółka z ograniczoną odpowiedzialnością spółka komandytowo-akcyjna, Synthos Kralupy a.s., StyroChem Finland Oy, et al. 2013a. Formulation of flame retarded expanded polystyrene (EPS) to solid unexpanded pellets using hexabromocyclododecane as the flame retardant additive (for onward use in building applications). Consultation number 0013-01. Available: https://echa.europa.eu/public-consultations [accessed July 7, 2019].

INEOS Styrenics Netherlands BV, INEOS Styrenics Ribecourt SAS, INEOS Styrenics Wingles SAS, Synthos Dwory 7 spółka z ograniczoną odpowiedzialnością spółka komandytowo-akcyjna, Synthos Kralupy a.s., StyroChem Finland Oy, et al. 2013b. Manufacture of flame retarded expanded polystyrene (EPS) articles for use in building applications. Consultation number 0013-02. Available: https://echa.europa.eu/public-consultations [accessed July 7, 2019].

IRCCS (Istituto di Ricerche Farmacologiche Mario Negri), Kemijski inštitut, Sciensano, Politecnico di Milano 1863, U.S. EPA (U.S. Environmental Protection Agency), Fera (Food and Environment Research Agency), et al. 2019. VEGA QSAR. Available: https://www.vegahub.eu/portfolio-item/vega-qsar/ [accessed December 9 2019].

J.G. DeKeyser, C.J. Omiecinski. 2010. Constitutive Androstane Receptor. In: Comprehensive toxicology Second Edition, Vol. 2, Part 2nd, (McQueen C, ed). The Boulevard, Langford Lane, Kidlington OX5 1GB, United Kingdom: Elsevier, 169-181.

Jaccard P. 1901. Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines. *Bull Soc Vaudoise Sci Nat* 37:241-272.

Jacobs M. 2004. In silico tools to aid risk assessment of endocrine disrupting chemicals. *Toxicology* 205:43-53, doi: https://doi.org/10.1016/j.tox.2004.06.036.

Jacobs MM, Malloy TF, Tickner JA, Edwards S. 2016. Alternatives assessment frameworks: research needs for the informed substitution of hazardous chemicals. *Environmental health perspectives* 124:265–280, doi: https://doi.org/10.1289/ehp.1409581.

Jensen GE, Niemelä JR, Wedebye EB, Nikolov NG. 2008. QSAR models for reproductive toxicity and endocrine disruption in regulatory use–a preliminary investigation. *SAR and QSAR in Environmental Research* 19:631-641.

Jyrkkärinne J, Windshügel Br, Rönkkö T, Tervo AJ, Küblbeck J, Lahtela-Kakkonen M, et al. 2008. Insights into ligand-elicited activation of human constitutive androstane receptor based on novel agonists and three-dimensional quantitative structure− activity relationship. *Journal of medicinal chemistry* 51:7181-7192, doi: https://doi.org/10.1021/jm800731b.

Kachaylo E, Pustylnyak V, Lyakhovich V, Gulyaeva L. 2011. Constitutive androstane receptor (CAR) is a xenosensor and target for therapy. *Biochemistry (Moscow)* 76:1087, doi: https://doi.org/10.1134/S0006297911100026.

Kato H, Yamaotsu N, Iwazaki N, Okamura S, Kume T, Hirono S. 2017. Precise prediction of activators for the human constitutive androstane receptor using structure-based three-dimensional quantitative structure–activity relationship methods. *Drug metabolism and pharmacokinetics* 32:179-188, doi: https://doi.org/10.1016/j.dmpk.2017.02.001.

Kemira Chemicals Oy. 2015. Use of sodium dichromate as an additive for suppressing parasitic reactions and oxygen evolution, pH buffering and cathode corrosion protection in the

electrolytic manufacture of sodium chlorate with or without subsequent production of chlorine dioxide or sodium chlorite. Consultation number 0038-01. European Chemicals Agency. Available: https://echa.europa.eu/public-consultations [accessed November 12, 2019].

Klimenko KO, Rosenberg SA, Dybdahl M, Wedebye EB, Nikolov NG. 2019. QSAR modelling of a large imbalanced aryl hydrocarbon activation dataset by rational and random sampling and screening of 80,086 REACH pre-registered and/or registered substances. *PLOS ONE* 14, doi: https://doi.org/10.1371/journal.pone.0213848.

Knudsen T, Martin M, Chandler K, Kleinstreuer N, Judson R, Sipes N. 2013. Predictive models and computational toxicology. Totowa, NJ: Humana Press, doi: https://doi.org/10.1007/978-1-62703-131-8_26.

Knudsen TB, Keller DA, Sander M, Carney EW, Doerrer NG, Eaton DL, et al. 2015. FutureTox II: in vitro data and in silico models for predictive toxicology. *Toxicological Sciences* 143:256-267, doi: https://doi.org/10.1093/toxsci/kfu234.

KREATiS. 2019. iSafeRat® Available: https://www.kreatis.eu/en/qsars-products-services.php?endpoint=0 [accessed 2019 December 9].

Kretschmer XC, Baldwin WS. 2005. CAR and PXR: xenosensors of endocrine disrupters? *Chemico-biological interactions* 155:111-128.

Küblbeck J, Jyrkkärinne J, Poso A, Turpeinen M, Sippl W, Honkakoski P, et al. 2008. Discovery of substituted sulfonamides and thiazolidin-4-one derivatives as agonists of human constitutive androstane receptor. *Biochemical pharmacology* 76:1288-1297, doi: https://doi.org/10.1016/j.bcp.2008.08.014.

Küblbeck J, Jyrkkärinne J, Molnár F, Kuningas T, Patel J, Windshügel Br, et al. 2011. New in vitro tools to study human constitutive androstane receptor (CAR) biology: discovery and

comparison of human CAR inverse agonists. *Molecular pharmaceutics* 8:2424-2433, doi: 10.1021/mp2003658.

Lahl U, Gundert-Remy U. 2008. The use of (Q) SAR methods in the context of REACH. *Toxicology mechanisms and methods* 18:149-158.

Leadscope®. 2002. Leadscope: Enterprise User Manual Version 2.2. Columbus, OH: Leadscope, Inc.

Leadscope®. 2019a. QSAR Models. Available: http://www.leadscope.com/model_appliers/ [accessed December 9 2019].

Leadscope®. 2019b. Leadscope. Available: http://www.leadscope.com [accessed December 9 2019].

Lee K, You H, Choi J, No KT. 2017. Development of pharmacophore-based classification model for activators of constitutive androstane receptor. *Drug metabolism and pharmacokinetics* 32:172-178, doi: https://doi.org/10.1016/j.dmpk.2016.11.005.

Lhasa Limited. 2019a. Derek Nexus,. Available: https://www.lhasalimited.org/products/derek-nexus.htm [accessed November 17 2019].

Lhasa Limited. 2019b. Sarah Nexus. Available: https://www.lhasalimited.org/products/sarah-nexus.htm [accessed November 17 2019].

Li D, Mackowiak B, Brayman TG, Mitchell M, Zhang L, Huang S-M, et al. 2015. Genome-wide analysis of human constitutive androstane receptor (CAR) transcriptome in wild-type and CAR-knockout HepaRG cells. *Biochemical pharmacology* 98:190-202, doi: https://doi.org/10.1016/j.bcp.2015.08.087.

Lillicrap A, Belanger S, Burden N, Du Pasquier D, Embry MR, Halder M, et al. 2016. Alternative approaches to vertebrate ecotoxicity tests in the 21st century: A review of

developments over the last 2 decades and current status. *Environmental toxicology and chemistry* 35:2637-2646.

Linkov I, Loney D, Cormier S, Satterstrom FK, Bridges T. 2009. Weight-of-evidence evaluation in environmental assessment: review of qualitative and quantitative approaches. *Science of the Total Environment* 407:5199-5205, doi: https://doi.org/10.1016/j.scitotenv.2009.05.004.

LMC (Laboratory of Mathematical Chemistry): OASIS. 2019. Software: TIMES 2.27.16 and CATALOGIC 5.11.16. Available: http://oasis-lmc.org/products/software.aspx [accessed December 8 2019].

Lu P, Xie W. 2017. Xenobiotic receptors in the crosstalk between drug metabolism and energy metabolism. In: Drug Metabolism in Diseases: Elsevier, 257-278.

Luechtefeld T, Marsh D, Rowlands C, Hartung T. 2018. Machine learning of toxicological big data enables read-across structure activity relationships (RASAR) outperforming animal test reproducibility. *Toxicological Sciences* 165:198-212, doi: https://doi.org/10.1093/toxsci/kfy152.

Lynch C, Mackowiak B, Huang R, Li L, Heyward S, Sakamuru S, et al. 2018. Identification of Modulators that Activate the Constitutive Androstane Receptor from the Tox21 10K Compound Library. *Toxicological Sciences*.

Maglich JM, Sluder A, Guan X, Shi Y, McKee DD, Carrick K, et al. 2001. Comparison of complete nuclear receptor sets from the human, Caenorhabditis elegans and Drosophila genomes. *Genome biology* 2:research0029-0021, doi: https://doi.org/10.1186/gb-2001-2-8-research0029.

Maglich JM, Watson J, McMillen PJ, Goodwin B, Willson TM, Moore JT. 2004. The nuclear receptor CAR is a regulator of thyroid hormone metabolism during caloric restriction. *Journal of Biological Chemistry*.

Malloy T, Zaunbrecher V, Beryt E, Judson R, Tice R, Allard P, et al. 2017. Advancing alternatives analysis: The role of predictive toxicology in selecting safer chemical products and processes. *Integrated environmental assessment and management* 13:915-925, doi: https://doi.org/10.1002/ieam.1923

Mansouri K, Kleinstreuer N, Watt E, Harris J. 2017. CoMPARA: Collaborative Modeling Project for Androgen Receptor Activity. Available: https://www.researchgate.net/publication/316606155_CoMPARA_Collaborative_Modeling_Project_for_Androgen_Receptor_Activity.

Martin P, Bladier C, Meek B, Bruyere O, Feinblatt E, Touvier M, et al. 2018. Weight of evidence for hazard identification: A critical review of the literature. *Environmental health perspectives* 126:076001.

Matsuzaka Y, Uesawa Y. 2019. Optimization of a Deep-Learning Method Based on the Classification of Images Generated by Parameterized Deep Snap a Novel Molecular-Image-Input Technique for Quantitative Structure–Activity Relationship (QSAR) Analysis. *Frontiers in bioengineering and biotechnology* 7:65, doi: https://doi.org/10.3389/fbioe.2019.00065.

Maunz A, Gütlein M, Rautenberg M, Vorgrimmler D, Gebele D, Helma C. 2013a. Lazar: a modular predictive toxicology framework. *Frontiers in pharmacology* 4, doi: https://doi.org/10.3389/fphar.2013.00038.

Maunz A, Gütlein M, Rautenberg M, Vorgrimmler D, Gebele D, Helma C. 2013b. Lazar: a modular predictive toxicology framework. *Frontiers in pharmacology* 4:38, doi: https://doi.org/10.3389/fphar.2013.00038.

Mayo-Bean K, Moran K, Meylan B, Ranslow P, U.S. Environmental Protection Agency (USEPA), Syracuse Research Corporation, et al. 2012. Methodology Document for the Ecological Structure-Activity Relationship Model (Ecosar) Class Program: ESTIMATING TOXICITY OF INDUSTRIAL CHEMICALS TO AQUATIC ORGANISMS USING THE ECOSAR (ECOLOGICAL STRUCTURE ACTIVITY RELATIONSHIP) CLASS PROGRAM. MS-Windows Version 1.11. Available: https://www.epa.gov/sites/production/files/2015-09/documents/ecosartechfinal.pdf [accessed January 23, 2019].

Mays C, Benfenati E, Pardoe S. 2012. Use and perceived benefits and barriers of QSAR models for REACH: findings from a questionnaire to stakeholders. *Chemistry Central Journal* 6:159, doi: https://doi.org/10.1186/1752-153X-6-159.

Miller MD, Crofton KM, Rice DC, Zoeller RT. 2009. Thyroid-disrupting chemicals: interpreting upstream biomarkers of adverse outcomes. *Environmental health perspectives* 117:1033-1041, doi: https://doi.org/10.1289/ehp.0800247.

Molcode (Molcode distributed by CompuDrug). 2019a. MolCode Toolbox. Available: http://www.compudrug.com/molcode_toolbox [accessed December 8 2019].

Molcode (Molcode distributed by CompuDrug). 2019b. REACH QSAR. Available: http://reachqsar.com [accessed December 8 2019].

Molnár F, Küblbeck J, Jyrkkärinne J, Prantner V, Honkakoski P. 2013. An update on the constitutive androstane receptor (CAR). *Drug metabolism and drug interactions* 28:79-93, doi: https://doi.org/10.1515/dmdi-2013-0009.

Moore JT, Moore LB, Maglich JM, Kliewer SA. 2003. Functional and structural comparison of PXR and CAR. *Biochimica et Biophysica Acta (BBA)-General Subjects* 1619:235-238, doi: https://doi.org/10.1016/S0304-4165(02)00481-6.

MTU Aero Engines AG. n.d. Analysis of Alternatives: Non-confidential report: Surface treatment for aerospace applications for civil and military uses, comprising treatment of new components for aircraft engines as well as maintenance, repair and overhaul work on aircraft engine components, unrelated to functional chrome plating. Consultation number 0066-02. Available: https://echa.europa.eu/public-consultations [accessed January 23, 2019].

MultiCase. 2019. CASE Ultra Models: Statistical Models & Expert Rules. Available: http://www.multicase.com/case-ultra-models [accessed December 8 2019].

Nicot T. 2018. Re: QSARs used during Analysis of Alternatives (AoA) require documentation. (Message to Kazue Chinen, ed). 12 November 2018. Email.

NIH (National Institutes of Health). 2018. PubChem: Bioassay. Available: https://pubchem.ncbi.nlm.nih.gov/# [accessed December 8, 2018 2018].

NIH (National Institutes of Health). 2019a. PubChem AID 1224835: qHTS assay to identify small molecule inhibitors of firefly luciferase. Available: https://pubchem.ncbi.nlm.nih.gov/bioassay/1224835 [accessed September 20 2019].

NIH (National Institutes of Health). 2019b. PubChem AID 1224892: qHTS assay to identify small molecule agonists of the constitutive androstane receptor (CAR) signaling pathway: Summary. Available: https://pubchem.ncbi.nlm.nih.gov/bioassay/1224892 [accessed December 8 2019].

NIH (National Institutes of Health). 2019c. PubChem AID 1224892: qHTS assay to identify small molecule agonists of the constitutive androstane receptor (CAR) signaling pathway:

Summary. Available: https://pubchem.ncbi.nlm.nih.gov/bioassay/1224892 [accessed September 20 2019].

NIH (National Institutes of Health). 2019d. PubChem AID 1224893: qHTS assay to identify small molecule antagonists of the constitutive androstane receptor (CAR) signaling pathway: Summary. Available: https://pubchem.ncbi.nlm.nih.gov/bioassay/1224893 [accessed September 12 2019].

NIH (National Institutes of Health). 2019e. Tox21 Public Data. Available: https://tripod.nih.gov/tox21/assays/ [accessed November 26 2019].

NIH (National Institutes of Health). 2019f. PubChem AID 1224838: qHTS assay to identify small molecule antagonists of the constitutive androstane receptor (CAR) signaling pathway. Available: https://pubchem.ncbi.nlm.nih.gov/bioassay/1224838 [accessed September 12 2019].

Nikolov N, Grancharov V, Stoyanova G, Pavlov T, Mekenyan O. 2006. Representation of chemical information in OASIS centralized 3D database for existing chemicals. *Journal of chemical information and modeling* 46:2537-2551.

NRC (National Research Council). 2007. Toxicity testing in the 21st century: a vision and a strategy. 0309109922. Washington, DC: National Academies Press. Available: https://www.nap.edu/catalog/11970/toxicity-testing-in-the-21st-century-a-vision-and-a [accessed January 23, 2019].

NRC (National Research Council). 2014a. Review of EPA's Integrated Risk Information System (IRIS) Process (2014). Washington, D.C.: The National Academies of Science. Available: https://www.nap.edu/catalog/18764/review-of-epas-integrated-risk-information-system-iris-process [accessed July 28, 2019].

NRC (National Research Council). 2014b. A framework to guide selection of chemical alternatives. Washington, D.C. Available: https://www.nap.edu/catalog/18872/a-framework-to-guide-selection-of-chemical-alternatives [accessed November 20, 2019].

NRC (National Research Council). 2018. Progress Toward Transforming the Integrated Risk Information System (IRIS) Program: A 2018 Evaluation (2018). Washington, D.C.: The National Academies of Science. Available: http://dels.nas.edu/Report/Progress-Toward-Transforming-Integrated-Risk/25086 [accessed July 31, 2019].

NTP (National Toxicology Program). 2019. Tox21: Toxicology in the 21st Century (Tox21). Available: https://ntp.niehs.nih.gov/results/tox21/index.html [accessed November 26 2019].

OECD (The Organisation for Economic Co-operation and Development). 1997. SIDS Initial Assessment Profile for SIAM 6 (Paris, France, 9-11 June 1997). Methyl Ethyl Ketone (MEK). Available: http://webnet.oecd.org/Hpv/UI/SIDS_Details.aspx?id=31C513F8-2B0D-4DE8-9A14-8463CD709ADD [accessed January 23, 2019 2019].

OECD (The Organisation for Economic Co-operation and Development). 2007. Guidance Document on the Validation of (Quantitative)Structure-Activity Relationships [(Q)SAR] Models, OECD Series on Testing and Assessment, No. 69. ENV/JM/MONO(2007)2. Paris (FR): OECD. Available: https://read.oecd-ilibrary.org/environment/guidance-document-on-the-validation-of-quantitative-structure-activity-relationship-q-sar-models_9789264085442-en#page1 [accessed November 20, 2019].

OECD (The Organisation for Economic Co-operation and Development). 2014. Guidance on Grouping of Chemicals, Second Edition. Available: http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono%282014%294&doclanguage=en [accessed January 23, 2019 2019].

OECD (The Organisation for Economic Co-operation and Development). 2018. OECD Quantitative Structure-Activity Relationships Project [(Q)SARs] Available: http://www.oecd.org/chemicalsafety/risk-assessment/oecdquantitativestructure-activityrelationshipsprojectqsars.htm [accessed December 3, 2018 2018].

OECD (The Organisation for Economic Co-operation and Development). 2019. The OECD QSAR Toolbox. Available: https://www.oecd.org/chemicalsafety/risk-assessment/oecd-qsar-toolbox.htm [accessed November 20 2019].

Pavan M, Worth AP. 2006. Review of QSAR models for ready biodegradation.

Peffer RC, Bailey KA, Lake L-K, Kristin , Lake BG. 2018. AOP: 107: Constitutive androstane receptor activation leading to hepatocellular adenomas and carcinomas in the mouse and the rat. Available: https://aopwiki.org/aops/107 [accessed September 12 2019].

Poso A, Honkakoski P. 2006. Ligand recognition by drug-activated nuclear receptors PXR and CAR: structural, site-directed mutagenesis and molecular modeling studies. *Mini reviews in medicinal chemistry* 6:937-943, doi: https://doi.org/10.2174/138955706777935008.

Posthumus R, Traas T, Peijnenburg W, Hulzebos E. 2005. External validation of EPIWIN biodegradation models. *SAR and QSAR in Environmental Research* 16:135-148, doi: 10.1080/10629360412331319899.

Pudenz S, Frère AR. 2017. Quantitative structure activity-relationships for toxicity screening of pesticide impurities for regulatory submission–practical experiences and challenges. *Toxicological & Environmental Chemistry* 99:1071-1077, doi: https://doi.org/10.1080/02772248.2016.1265649.

Qatanani M, Moore D. 2005. CAR, the continuously advancing receptor, in drug metabolism and disease. *Current drug metabolism* 6:329-339.

Qatanani M, Zhang J, Moore DD. 2005. Role of the constitutive androstane receptor in xenobiotic-induced thyroid hormone metabolism. *Endocrinology* 146:995-1002.

Rhomberg LR, Goodman JE, Bailey LA, Prueitt RL, Beck NB, Bevan C, et al. 2013. A survey of frameworks for best practices in weight-of-evidence analyses. *Critical reviews in toxicology* 43:753-784, doi: https://doi.org/10.3109/10408444.2013.832727.

Roberts G, Myatt GJ, Johnson WP, Cross KP, Blower PE. 2000. LeadScope: software for exploring large sets of screening data. *Journal of chemical information and computer sciences* 40:1302-1314, doi: https://doi.org/10.1021/ci0000631.

Rorije E, Hulzebos E, Hakkert B. 2008. The EU (Q) SAR Experience Project: reporting formats. Templates for documenting (Q) SAR results under REACH. *RIVM report 601779001*.

Rosenberg SA, Watt ED, Judson RS, Simmons S, Friedman KP, Dybdahl M, et al. 2017a. QSAR models for thyroperoxidase inhibition and screening of US and EU chemical inventories. *Computational Toxicology* 4:11-21, doi: https://doi.org/10.1016/j.comtox.2017.07.006.

Rosenberg SA, Xia M, Huang R, Nikolov NG, Wedebye EB, Dybdahl M. 2017b. QSAR development and profiling of 72,524 REACH substances for PXR activation and CYP3A4 induction. *Computational Toxicology* 1:39-48, doi: https://doi.org/10.1016/j.comtox.2017.01.001.

Rotella D. 2011. New Horizons in Predictive Toxicology: Current Status and ApplicationRoyal Society of Chemistry.

Rovida C. 2010. Food for thought… why no new in vitro tests will be done for REACH by registrants. *Altex* 27:175-183.

Rovida C, Longo F, Rabbit RR. 2011. How are reproductive toxicity and developmental toxicity addressed in REACH dossiers? *Altex* 28:273-294.

Sasol-Huntsman GmbH & Co. KG. n.d. Use as an absorption solvent in a closed system in the manufacture of maleic anhydride (MA). Consultation number 0006-01. European Chemicals Agency. Available: https://echa.europa.eu/public-consultations [accessed January 23, 2019].

SCENIHR (Scientific Committee on Emerging and Newly-Identified Health Risks). 2015. The safety of medical devices containing DEHP- plasticized PVC or other plasticizers on neonates and other groups possibly at risk (2015 update). Luxembourg, Belgium: European Commission. Available: https://publications.europa.eu/en/publication-detail/-/publication/9c14b179-d3ae-11e5-a4b5-01aa75ed71a1/language-en [accessed August 31, 2019].

Schulte C, Tietjen L, Bambauer A, Fleischer A. 2012. Five years REACH–lessons learned and first experiences. I. an authorities' view. *Environmental Sciences Europe* 24:31, doi: 10.1186/2190-4715-24-31.

Shi LM, Fang H, Tong W, Wu J, Perkins R, Blair RM, et al. 2001. QSAR models using a large diverse set of estrogens. *Journal of Chemical Information and Computer Sciences* 41:186-195, doi: https://doi.org/10.1021/ci000066d.

Simulations Plus. 2019. ADMET PredictorTM: ADMET property prediction and QSAR model-building application. Available: http://www.simulations-plus.com/software/admetpredictor/ [accessed December 8 2019].

Solvay Portugal – Produtos Quimicos SA. 2015. Use of sodium dichromate as an additive for suppressing parasitic reactions and oxygen evolution, pH buffering and cathode corrosion protection in the electrolytic manufacture of sodium chlorate with or without subsequent production of chlorine dioxide or sodium chlorite. Consultation number 0039-01.

European Chemicals Agency. Available: https://echa.europa.eu/public-consultations [accessed November 12, 2019].

Sonoda J, Pei L, Evans RM. 2008. Nuclear receptors: decoding metabolic disease. *FEBS letters* 582:2-9, doi: https://doi.org/10.1016/j.febslet.2007.11.016.

Suter G, Cormier S, Barron M. 2017. A weight of evidence framework for environmental assessments: Inferring qualities. *Integrated environmental assessment and management* 13:1038-1044.

Syberg K, Jensen TS, Cedergreen N, Rank J. 2009. On the use of mixture toxicity assessment in REACH and the water framework directive: a review. *Human and Ecological Risk Assessment* 15:1257-1272.

Tabb MM, Blumberg B. 2006. New modes of action for endocrine-disrupting chemicals. *Molecular endocrinology* 20:475-482, doi: https://doi.org/10.1210/me.2004-0513.

TerraBase Inc. 2019. TerraBase Inc.'s TerraQSAR™ Computation Programs. Available: http://www.terrabase-inc.com [accessed December 9 2019].

Tickner J, Geiser K, Rudisill C, Schifano J. 2013. Alternatives assessment in regulatory policy: history and future directions. In: Chemical alternatives assessment, (Hester RE, Harrison RM, eds). London: Royal Society of Chemistry, 256-288.

Tickner J, Jacobs MM. 2016. Improving the Identification, Evaluation, Adoption and Development of Safer Alternatives: Needs and Opportunities to Enhance Substitution Efforts within the Context of REACH. Lowell, Massachusetts: University of Massachusetts Lowell, Lowell Center for Sustainable Production. Available: https://echa.europa.eu/documents/10162/13630/substitution_capacity_lcsp_en.pdf [accessed January 23, 2019].

Todeschini R, Consonni V. 2008. Handbook of molecular descriptorsJohn Wiley & Sons.

Tong W, Perkins R, Fang H, Hong H, Xie Q, Branham S, et al. 2002. Development of quantitative structure-activity relationships (QSARs) and their use for priority setting in the testing strategy of endocrine disruptors. *Regul Res Perspect* 1:1-16.

Tunnela O. 2017. Re: Enquiry to ECHA for REACH Regulatory obligations on QMRFs and QPRFs in Analyses of Alternatives (AoA). (Message to Kazue Chinen, ed). 7 December, 2017. Email.

Tunnela O. 2018. Re: Enquiry to ECHA on REACH Regulatory obligations in Authorisation. (Message to Kazue Chinen, ed). 17 January 2018. Email.

U.S. EPA (U.S. Environmental Protection Agency). 2011. Endocrine Disruptor Screening Program: Weight of Evidence: Evaluating results of EDSP Tier 1 screening to identify the need for Tier 2 testing. Washington, D.C. Available: https://www.regulations.gov/document?D=EPA-HQ-OPPT-2010-0877-0021 [accessed September 18, 2019].

U.S. EPA (U.S. Environmental Protection Agency). 2012a. Sustainable Futures / P2 Framework Manual 2012 4. Filling Data Gaps: Introduction to Predictive Models. EPA-748-B12-001. Washington, D.C. Available: https://www.epa.gov/sites/production/files/2015-05/documents/04.pdf [accessed December 8, 2019].

U.S. EPA (U.S. Environmental Protection Agency). 2012b. Sustainable Futures / P2 Framework Manual 2012 7. Estimating Persistence, Bioaccumulation, and Toxicity Using the PBT Profiler. EPA-748-B12-001. Washington D.C. Available: https://www.epa.gov/sites/production/files/2015-05/documents/07.pdf.

U.S. EPA (U.S. Environmental Protection Agency). 2012c. Sustainable Futures / P2 Framework Manual 2012 6. Estimating Aquatic Toxicity Using ECOSAR. EPA-748-B12-001.

Washington, D.C. Available: https://www.epa.gov/sites/production/files/2015-05/documents/06.pdf [accessed December 8, 2019].

U.S. EPA (U.S. Environmental Protection Agency). 2016. User's Guide for T.E.S.T. (version 4.2) (Toxicity Estimation Software Tool) A Program to Estimate Toxicity from Molecular Structure. EPA/600/R-16/058. Washington, DC. Available: https://www.epa.gov/sites/production/files/2016-05/documents/600r16058.pdf.

U.S. EPA (U.S. Environmental Protection Agency). 2019a. EPI Suite™-Estimation Program Interface. Available: https://www.epa.gov/tsca-screening-tools/epi-suitetm-estimation-program-interface [accessed June 9 2019].

U.S. EPA (U.S. Environmental Protection Agency). 2019b. Ecological Structure Activity Relationships (ECOSAR) Predictive Model. Available: https://www.epa.gov/tsca-screening-tools/ecological-structure-activity-relationships-ecosar-predictive-model [accessed December 9 2019].

U.S. EPA (U.S. Environmental Protection Agency). 2019c. Toxicity Estimation Software Tool (TEST). Available: https://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test# [accessed December 9 2019].

U.S. EPA (U.S. Environmental Protection Agency). 2019d. CoMPARA: Collaborative Modeling Project for Androgen Receptor Activity (SOT). Available: https://cfpub.epa.gov/si/si_public_record_report.cfm?Lab=NCCT&dirEntryId=339684 [accessed September 12 2019].

U.S. EPA (U.S. Environmental Protection Agency). 2019e. EPI Suite™-Estimation Program Interface. Available: https://www.epa.gov/tsca-screening-tools/epi-suitetm-estimation-program-interface [accessed November 18 2019].

U.S. FDA (U.S. Food & Drug Administration). 2018. Science & Research: EDKB Resources. Available: https://www.fda.gov/ScienceResearch/BioinformaticsTools/EndocrineDisruptorKnowledgebase/ucm084342.htm [accessed December 9 2019].

Valerio Jr. LG, Yang C, Arvidson KB, Kruhlak NL. 2010. A structural feature-based computational approach for toxicology predictions. *Expert opinion on drug metabolism & toxicology* 6:505-518, doi: https://doi.org/10.1517/17425250903499286.

Vinggaard AM, Niemelä J, Wedebye EB, Jensen GE. 2008. Screening of 397 Chemicals and Development of a Quantitative Structure− Activity Relationship Model for Androgen Receptor Antagonism. *Chemical research in toxicology* 21:813-823, doi: https://doi.org/10.1021/tx7002382.

Viswanadhan VN, Ghose AK, Revankar GR, Robins RK. 1989. Atomic physicochemical parameters for three dimensional structure directed quantitative structure-activity relationships. 4. Additional parameters for hydrophobic and dispersive interactions and their application for an automated superposition of certain naturally occurring nucleoside antibiotics. *Journal of chemical information and computer sciences* 29:163-172, doi: https://doi.org/10.1021/ci00063a006.

Voyer R, Heltshe J. 1984. Factor interactions and aquatic toxicity testing. *Water Research* 18:441-447.

Wahlang B, Falkner KC, Clair HB, Al-Eryani L, Prough RA, States JC, et al. 2014. Human receptor activation by aroclor 1260, a polychlorinated biphenyl mixture. *Toxicological Sciences* 140:283-297.

Walker JD, Carlsen L, Jaworska J. 2003. Improving opportunities for regulatory acceptance of QSARs: the importance of model domain, uncertainty, validity and predictability. *QSAR & Combinatorial Science* 22:346-350.

Wang P, Xiao X, Chou K-C. 2011. NR-2L: a two-level predictor for identifying nuclear receptor subfamilies based on sequence-derived features. *PloS one* 6:e23505, doi: https://doi.org/10.1371/journal.pone.0023505.

Worth A, Lapenna S, Lo Piparo E, Mostrag-Szlichtyng A, Serafimova R. 2011. A Framework for assessing in silico Toxicity Predictions: Case Studies with selected Pesticides. *JRC report EUR* 24705.

Worth A, Barroso J, Bremer S, Burton J, Casati S, Coecke S, et al. 2014. Alternative methods for regulatory toxicology–a state-of-the-art review. Report EUR 26797 EN. Available: https://www.researchgate.net/profile/Joao_Barroso3/publication/288256624_Alternative_methods_for_regulatory_toxicology_A_state-of-The-art_review/links/568a50cf08ae1e63f1fbbae7.pdf [accessed January 23, 2019].

Worth AP. 2010a. Chapter 13: The Role of QSAR Methodology in the Regulatory Assessment of Chemical. In: Recent advances in QSAR studies Challenges and Advances in Computational Chemistry and Physics, Vol. 8, (Puzyn T, Leszczynski J, Cronin MT, eds): Springer, Dordrecht, 367-382.

Worth AP. 2010b. The Role of QSAR Methodology in the Regulatory Assessment of Chemicals. In: Recent advances in QSAR studies Challenges and Advances in Computational Chemistry and Physics, Vol. 8, (Puzyn T, Leszczynski J, Cronin MT, eds): Springer, Dordrecht, 367-382.

Wyde ME, Kirwan SE, Zhang F, Laughter A, Hoffman HB, Bartolucci-Page E, et al. 2005. Di-n-butyl phthalate activates constitutive androstane receptor and pregnane X receptor and

enhances the expression of steroid-metabolizing enzymes in the liver of rat fetuses. *Toxicological Sciences* 86:281-290.

Xu C, Li CY-T, Kong A-NT. 2005. Induction of phase I, II and III drug metabolism/transport by xenobiotics. *Archives of pharmacal research* 28:249, doi: https://doi.org/10.1007/BF02977789.

Yamamoto Y, Moore R, Goldsworthy TL, Negishi M, Maronpot RR. 2004. The orphan nuclear receptor constitutive active/androstane receptor is essential for liver tumor promotion by phenobarbital in mice. *Cancer research* 64:7197-7200, doi: https://doi.org/10.1158/0008-5472.CAN-04-1459.

Yan J, Xie W. 2016. A brief history of the discovery of PXR and CAR as xenobiotic receptors. *Acta Pharmaceutica Sinica B* 6:450-452, doi: https://doi.org/10.1016/j.apsb.2016.06.011.

Yuan H, Wang Y-Y, Cheng Y-Y. 2007. Mode of action-based local QSAR modeling for the prediction of acute toxicity in the fathead minnow. *Journal of Molecular Graphics and Modelling* 26:327-335.

Zhang J-H, Chung TD, Oldenburg KR. 1999. A simple statistical parameter for use in evaluation and validation of high throughput screening assays. *Journal of biomolecular screening* 4:67-73, doi: https://doi.org/10.1177/108705719900400206.