Regulatory Toxicology and Pharmacology 119 (2021) 104834

Contents lists available at ScienceDirect

Regulatory Toxicology and Pharmacology

journal homepage: www.elsevier.com/locate/yrtph

Evaluating chemical similarity as a measure to identify potential substances of very high concern

Pim N.H. Wassenaar [a,b,*], Emiel Rorije [a], Martina G. Vijver [b], Willie J.G.M. Peijnenburg [a,b]

[a] National Institute for Public Health and the Environment (RIVM), P.O. Box 1, 3720BA, Bilthoven, the Netherlands
[b] Institute of Environmental Sciences (CML), Leiden University, P. O. Box 9518, 2300RA, Leiden, the Netherlands

ARTICLE INFO

Keywords:
Substances of very high concern
Screening and prioritization
Chemical similarity
Expert elicitation
Chemical grouping

ABSTRACT

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

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

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

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

1. Introduction

Worldwide, more than 350,000 chemicals and chemical mixtures are registered for production and use (Wang et al., 2020). Due to this large amount of substances, screening and prioritization are essential in order to focus chemical evaluation on those chemicals of highest concern. Chemical regulations particularly aim to minimize exposures and emissions of chemicals with serious and irreversible effects on human health or the environment as much as possible. In Europe, this specifically includes substances that are carcinogenic, mutagenic or reprotoxic (CMR); persistent, bioaccumulative and toxic (PBT) or very persistent and very bioaccumulative (vPvB); or substances with an equivalent level of concern, like endocrine disrupting (ED) substances. Substances that meet specific criteria for these endpoints of concern are identified via a hazard-based approach as Substances of Very High Concern (SVHC) within the REACH regulation (Registration, Evaluation, Authorization and Restriction of Chemicals; EC/1907/2006). The ultimate aim is to substitute these substances by safer (non-regrettable)

* Corresponding author. National Institute for Public Health and the Environment (RIVM), P.O. Box 1, 3720BA, Bilthoven, the Netherlands.
E-mail address: pim.wassenaar@rivm.nl (P.N.H. Wassenaar).

https://doi.org/10.1016/j.yrtph.2020.104834
Received 12 August 2020; Received in revised form 15 October 2020; Accepted 17 November 2020
Available online 20 November 2020

alternatives.

To facilitate the identification of potential (new-)SVHC substances, we recently developed a chemical similarity methodology that assesses whether a new chemical is structurally similar to a known SVHC (Wassenaar et al., 2019). A high resemblance in chemical structure might be an indication of comparable effects ('similar property principle' (Johnson and Maggiora, 1990)), and therefore could be a trigger for further evaluation. The developed methodology is based on a Dutch list of SVHCs (RIVM, 2018), which covers a broader range of chemicals than the EU-SVHC list under REACH, but are identified based on the same hazard criteria (see Supplemental Material S.1 for more details). The final models suggested a good performance on a dataset with non-SVHC and SVHC substances with CMR, PBT/vPvB and ED properties during 'internal' validation (balanced accuracies $\geq$ 80%) and outperformed several well-known predictive models when applied to this dataset (Wassenaar et al., 2019). Accordingly, these results are promising for further application within screening and prioritization activities on potential SVHCs.

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

Within this study we applied the newly developed chemical similarity models to the list of all registered substances under REACH, for which we do not (yet) have specific knowledge on potential concerns. We compared the chemical similarity as computed by the models with expert judgement classifications in order to assess the developed similarity models. The use of expert opinions is not uncommon in the field of predictive toxicology. For instance, expert classifications have been used within the development of the widely applied biodegradation models Biowin3 and Biowin4 – where the entire training dataset is formed by expert elicitation (US EPA, 2012) – and expert classifications are also applied within specific machine learning algorithms (i.e. active learning approaches) (Settles, 2010; Yang et al., 2019). Furthermore, within this study, illustrations are given aimed to show the model's potential for screening purposes (including single-substances and groups of substances). Specific examples that are discussed include phenolic benzotriazoles and bisphenol analogues.

2. Methods

2.1. REACH dataset

To investigate and assess model applicability, a dataset consisting of all REACH registered substances was used. For these substances we did not evaluate specific knowledge on potential concerns, which might be available for a subset of these substances in their REACH registration dossiers. The REACH registered substances were extracted from the webpage of the European Chemicals Agency (ECHA) on registered substances ((ECHA, 2019); extracted on 17-05-2019). In total, this list consisted of 24,694 entries representing 22,180 unique substances with chemical names and CAS-numbers. Based on this information, SMILES were generated using the KNIME (v3.7) workflow as developed by Gadaleta et al. (2018), which connects SMILES from different data sources to CAS-numbers and/or chemical names. For cases where multiple CAS-numbers were available per substance, only the first CAS-number was used for the KNIME input (n = 439).

Following the first part of the KNIME workflow, all entries are divided in three groups: Maintained, Rejected and Manual check. A substance is maintained when the retrieved SMILES from different sources are consistent. Substances are rejected when the retrieved SMILES are highly discordant or information is totally missing, whereas a further assessment is necessary when some identical and different/missing SMILES are retrieved (i.e. a manual check). Upon manual check some concordant SMILES had to be retrieved from other datasets. The following data sources were considered consecutively: ECHA dissemination site (https://echa.europa.eu/nl/search-for-chemicals; primary source of the substances), ChemicalBook (https://www.chemicalbook.com/; suggested by Gadaleta et al. (2018)) and Molbase (http://www.molbase.com/; suggested by Gadaleta et al. (2018)). When no SMILES was retrieved via the above-mentioned sources, google searches were performed. In addition, substances that could not be represented by a single SMILES were removed during the manual check. This included substances with chemical names that describe mixtures, chemical substances of unknown or variable composition, complex reaction products and biological materials (UVCBs; including petroleum, extracts, fatty acids, glycerides, hydrocarbons, oil, residues, resins and rosins), reaction masses, reaction products (including products) or polymers (Gadaleta et al., 2018). Furthermore, ionic substances that have large (organic) counter ions were excluded as they cannot be represented by a single structure.

The information of the manual check is used in the second part of the KNIME workflow. This results in a list of maintained substances with corresponding CAS-numbers and SMILES and a list of rejected substances for which no reliable SMILES could be (automatically) retrieved. Subsequently, all substances that are on a Dutch list of Substances of Very High Concern ((RIVM, 2018); extracted on 01-03-2018) were excluded from the maintained substances, as those substances are in the training dataset of the structural similarity models (see section 2.2).

2.2. Structural similarity screening

Subsequently, the dataset was screened with the structural similarity models as described by Wassenaar et al. (2019). Within these models, the structure of a chemical is compared to known CMR, PBT/vPvB and ED substances included on a Dutch list of SVHCs ((RIVM, 2018); extracted on 01-03-2018; Table S.1). This list covers a broader range of chemicals than the EU-SVHC list under REACH, but are identified based on the same hazard criteria as the EU-SVHC substances (i.e. REACH article 57 (European Parliament, 2006)). The generation and composition of this list of substances is more elaborately described by Wassenaar et al. (2019) and in the Supplemental Material S.1. Throughout the text, these substances are referred to as SVHCs.

Within the structural similarity models, first a fingerprint is generated for a substance based on its chemical structure. Secondly, the similarity of the fingerprint to the fingerprints of all SVHCs is expressed using a similarity coefficient. This results in similarity values to all SVHC substances, ranging from 0 (i.e. structures are considered as totally different) to 1 (i.e. structures are considered as identical). Thirdly, the similarity values are compared to a similarity threshold (i.e. a specific value between 0 and 1). Above the threshold, the substance is considered to be sufficiently structurally similar to assume comparable toxicological effects/concerns. The type of fingerprint, the similarity coefficient and the threshold applied in the structural similarity models, were determined in an optimization process (Wassenaar et al., 2019), and differ for the various SVHC-subgroups (Table 1).

The results of the structural similarity models were visualized within chemical similarity networks using Gephi (v0.9.2) (Bastian et al., 2009) for the different subsets (CMR, PBT/vPvB and ED). Within the similarity networks, only chemical similarities above the model threshold values

Table 1
Overview of the characteristics of the structural similarity models (Wassenaar et al., 2019). CMR = carcinogenic, mutagenic or reprotoxic substances; PBT/vPvB = persistent, bioaccumulative and toxic/very persistent and very bioaccumulative substances; ED = endocrine disrupting substances. The MACCS and CDK Extended fingerprint are generated using PaDEL-Descriptor (Yap, 2014) and the FCFP4 fingerprint (i.e. Functional-Class Fingerprints with a diameter of 4) with RDkit using Morgan fingerprints (Landrum, 2019). Names of the coefficients are provided as in accordance to Todeschini et al. (2012): CT4 = Consonni-Todeschini 4; SM = Simple Matching; SS3 = Sokal-Sneath 3.

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

[a] A different similarity coefficient is used in the CMR similarity model for substances that have less than 85 fragments identified in the fingerprint (<85) and substances with 85 or more fragments identified in the fingerprint (≥85).

were included (see Table 1). In addition, Gephi was used to visually cluster the substances according to 'Modularity Class' following the algorithm of Blondel et al. (2008).

2.3. Expert elicitation

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

1) Toxicological similarity: Do you expect similar toxicological effects/concerns for the unknown chemical based on chemical similarity, when compared to the chemical of known toxicological concern? The scoring was a binary answer 'Yes' or 'No' executed by 30 participants.
2) Chemical similarity: To which extent do you consider the two substances as structurally similar? The scoring was based on a 5-point Likert scale (Likert, 1932) executed by 10 participants (a guide for scaling has been provided to the experts; see Supplemental Material S.2).

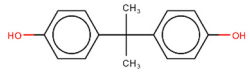
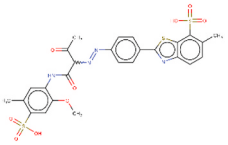
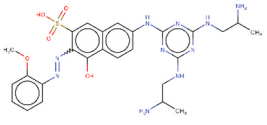
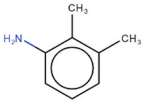
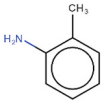
The results of the assessment phase were used to provide a statistical assessment of the ability of the structure similarity based computational models to reproduce the consensus expert elicitations regarding toxicological effects/concerns.

2.3.1. Assessment phase

The assessment phase consisted of 256 substance-pairs, consisting of a substance with unknown properties from the REACH dataset and its most similar SVHC according to the similarity models. In total, 96 substance-pairs represented the CMR model(s), 100 the PBT/vPvB model, and 60 the ED model (see Supplemental Material S.3 for more details on the selection). The inclusion of substances-pairs was according to stratified random sampling based on computer-generated similarity values. In other words, the REACH dataset was divided in bins of similarity scores from which one or multiple substance-pairs were randomly selected. Substance-pairs were selected in such a way to ensure balanced groups of similar and non-similar substance-pairs (i.e. above and below the similarity model thresholds, Table 1). The procedure used for selection of substance-pairs for the CMR-CT4 model differed slightly from the described procedure (see Supplemental Material S.3). The substance-pairs were provided in random order to the experts, without showing the computer calculated similarity values in order to avoid any influence on the expert opinions. In addition, one random substance-pair of the CMR, PBT/vPvB and ED subset was included three times in order to investigate the consistency in scoring.

Two groups of participants were requested to fill in the survey. One group consisted of toxicologists in training and were only requested to

Table 2
An example of the expert judgement exercise. Substance-pairs consisting of a substance with unknown properties and the most similar SVHC were provided to the experts, who had to answer a question on toxicological concern ('yes/no'-score) and a question on chemical similarity (5-point Likert scale (Likert, 1932): 1 = strongly disagree, 2 = disagree, 3 = neither agree/disagree, 4 = agree, 5 = strongly agree). During the assessment phase, both questions were asked separately.

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

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

2.3.2. Data analysis

The performance of the models is analyzed by comparing the predictions by the models to the predictions by a group of experts. In other words, we analyze whether the similarity models do highlight those chemicals that would also be selected as substances of potential concern by a group of experts. Within this study, the use of expert elicitation thus considers a direct expert response on chemical similarity and related concerns, and cannot be considered as an extensive (expert) evaluation of all available data on a specific chemical. The use of expert opinions is particularly suitable as this is exactly the intended purpose of the chemical similarity models: a quick, reproducible and automated screening tool that mimics the expert judgement that is frequently applied in various screening applications.

The results of the binary scoring on toxicological similarity were analyzed by using a confusion matrix (i.e. expert judgement vs similarity model prediction), in which the expert judgement scores were considered as the 'true'-effects. Following expert judgement scoring, a substance with unknown SVHC-effects was considered as potential SVHC based on majority voting (i.e. >50% 'yes'-score). Majority voting on chemical similarity is regularly applied in several setting, e.g. at the European Medicines Agency (Franco et al., 2014). Within this analysis, the predictions of assessors from both groups were combined. The results of the scoring of chemical similarity using the 5-point Likert scale, were used to analyze the relation between toxicological similarity (i.e. the results from the first question) and chemical similarity. In addition, the chemical similarity scores of the participants were compared to the computer-generated similarity values. When an expert did not answer the toxicological or chemical similarity question for a specific substance-pair, the observation was excluded from analysis (n = 14 and n = 3, respectively). All data was analyzed in R (v3.6) (R Core Team, 2019).

2.4. Illustrative cases

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

3. Results

3.1. REACH dataset

A reliable unique SMILES could be assigned to 9593 chemicals out of the 22,180 REACH registered substances. In Table 3 an overview of the physicochemical and structure properties of those substances is provided, as well as for the SVHC subsets as used in the similarity models.

3.2. Structural similarity screening

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

3.3. Expert elicitation

3.3.1. Assessment phase

The results for the binary question (toxicological concern), indicated that 102 of the 256 substances are considered potential SVHC following majority voting, with an average 'yes-voting' of 74% (±14% standard deviation) (Table S.2). In total, 154 substances were not considered potential SVHC based on expert elicitations with an average 'yes-voting' of 24% (±15% standard deviation) (Table S.2). Results for the second question (chemical similarity), indicate that the moderate values (i.e. 2-4) are selected more frequently – in 76% of the cases – compared to the extremes (i.e. 1 and 5). In addition, the average spread (i.e. standard error) around the mean of the chemical similarity score is lower for the extremes (Figure S.1), indicating slightly less variation among the experts. The individual scores as provided by the experts are shown in Figure S.2 and S.3 and summarized in Table S2. On average, the experts scored their own expertise to toxicity assessment and/or knowledge of molecular structures at 7.1 (n = 10 extended survey) and 5.5 (n = 20 short survey) out of a max of 10, respectively. No relation was observed between the expertise of the participants and their provided scores, and also not between the participants of the different groups (see Figure S.2 and S.3).

The relation between toxicological similarity (n = 30 experts) and chemical similarity (n = 10 experts) as assessed by the experts is shown in Fig. 1. For all subsets – CMR, PBT/vPvB and ED – there is a clear relationship observed (R^2 ranging from 0.84 to 0.89), and indicates the importance of chemical similarity for toxicological concern. Based on this relationship, 50% of the experts expect a comparable toxicological concern for a substance-pair with an average chemical similarity of around 3 in the Likert scale used for chemical similarity.

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

Table 3
Physicochemical and structural properties of the substances in the REACH dataset and the expert elicitation dataset. In addition, the properties of the SVHCs as included within the structural similarity models are provided for the different models (i.e. CMR, PBT/vPvB and ED). The ranges represent the 2.5th and 97.5th percentiles of the properties. Log K_{ow} was predicted according to EpiSuite (US EPA, 2012).

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

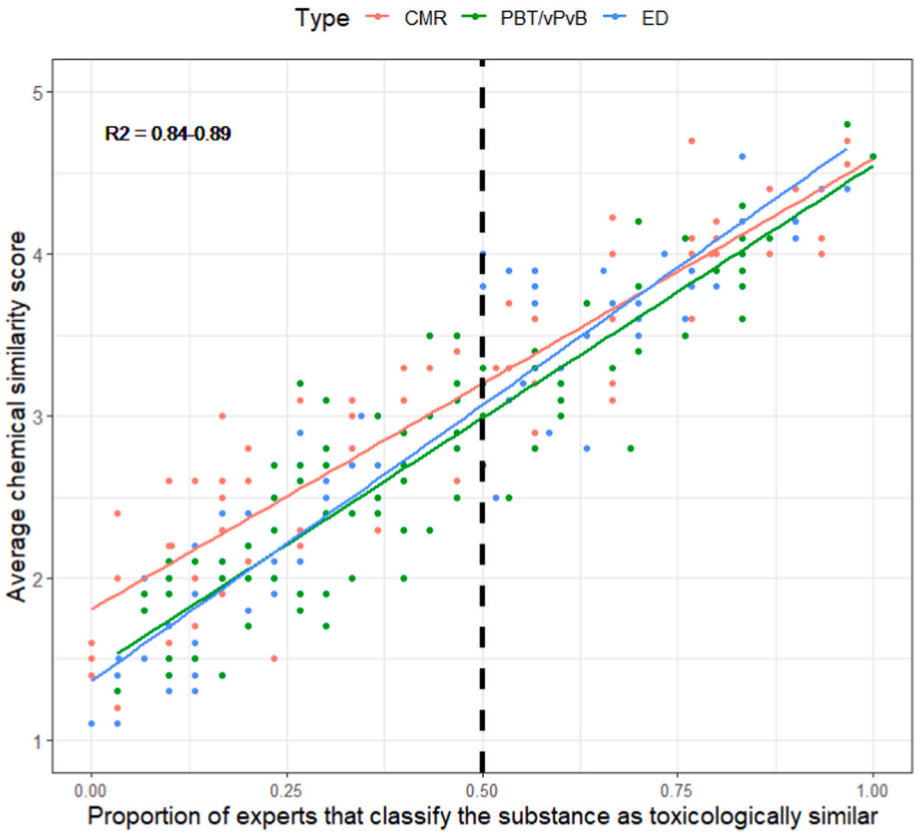


Fig. 1. Relation between toxicological similarity (n = 30 experts) and chemical similarity (n = 10 experts) as assessed by the experts for the 256 substance-pairs. R^2 quantifies the goodness of fit.

respect to toxicological similarity. With respect to chemical similarity, the experts provided the same answer in 59% of the cases. In cases they provided a different score, they varied on average with a score of 1.16. Furthermore, most variation was observed around substance-pairs with higher uncertainty (i.e. substance-pairs with around 50% 'yes'-scores for toxicological similarity, or an average chemical similarity around 3).

3.3.2. *Computer model performance compared to expert elicitation*

Table 4 shows the expert scores for the binary question (toxicological concern) in comparison to the computer predictions (also visualized in Figure S.4). The scores from the experts are taken as the 'true'-values based on majority voting (i.e. > 50% 'yes'-scores). It can be observed that the similarity models follow the expert opinions with a balanced accuracy between 0.69 and 0.87. The performance of the CMR model – when compared to the expert judgement exercise – is comparable to the performance obtained from an internal ('training') dataset (see Table 4) (Wassenaar et al., 2019). The predictive performance of the ED model is slightly below the performance observed during internal validation, whereas the model application performance for the PBT/vPvB subset is much lower when predicting expert opinions. Specific examples of substances that are differently classified by the similarity models, when compared to the expert judgement scores, are discussed in section 3.4.

3.4. *Illustrative cases*

The results of the structural similarity models, after application to the REACH dataset, were visualized by using chemical similarity networks (see Fig. 2 and Figure S.5-S.6). Based on these networks, groups of substances can be identified that are all predicted to be sufficiently structurally similar to one or multiple existing SVHCs, and therefore can be considered as potential SVHC. Within this section, several specific groups are highlighted to indicate varying applications of the models. In addition, general and substance-specific model limitations – as apparent from the REACH dataset screening and expert elicitations – are discussed.

Table 4
Cooper statistics for the computer similarity model classifications of potentially SVHC or non-SVHC when compared to the majority opinion of a group of human experts. The reference balanced accuracy represents the model performance following internal validation as analyzed by Wassenaar et al. (2019). TP = true positives, FP = false positives, TN = true negatives, FN = false negatives.

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

3.4.1. Group screening

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

All phenolic benzotriazoles that are considered structurally similar to a known SVHC, meet the P-screening criteria, and are close to or above the B-screening criteria (ECHA, 2017a; US EPA, 2012) (Table 6). Follow-up PBT/vPvB analyses are already being conducted within the REACH framework for three of these five substances (Table 6). Considering the chemical – and potentially the biological – similarity between these substances, this could be a trigger for further evaluation or assessment of this specific group of substances.

Furthermore, besides PBT/vPvB concern, the ED structural similarity model identifies substance nr.7 (in Tables 5–6) as structurally similar to an ED substance (i.e. 4-(1,1,3,3-tetramethylbutyl)phenol). This type of additional triggers, obtained via the similarity models, could lead to new hypotheses on chemical effects/concerns that could be further investigated.

3.4.2. Dissimilarity screening

The similarity models can also be used to evaluate dissimilarity. More specifically, it could be tested when a chemical is not considered to be sufficiently structurally similar to a known SVHC. Such evaluations are of particular interest to prioritize potential alternatives to known SVHCs (e.g. within safe-by-design development processes). Although the similarity models classify substances as potentially-SVHC versus non-SVHC, the difference between the similarity value and the established model thresholds (Table 1) could be an indication of the certainty for the classifications.

To illustrate the use for non-similarity screening, we here provide an example of bisphenol A (BPA) analogues. BPA is acknowledged as being a reprotoxic chemical with ED properties (nr.1 Fig. 3) (ECHA, 2020a). In Fig. 3, a sequence of BPA analogues is shown with slight changes in the BPA structure from chemical nr.2 up to larger changes in structure nr.8. For all those structures the chemical similarity to BPA is analyzed using both the CMR and ED structural similarity models. According to these models, substance nr.7 (tetramethyl bisphenol F; TM-BPF) and substance nr.8 (tetra(tertbutyl) bisphenol F; TTB-BPF) are considered to be structurally too dissimilar to BPA to be classified as potential SVHC, and could potentially be given higher priority within a safe-by-design development process. TM-BPF seems to show lower estrogenic activities compared to BPA (Soto et al., 2017), though further data analysis on TM-BPF is still ongoing (ECHA, 2020b), in which also other modes of action need to be considered. For instance, BPA also shows anti-androgenic effects (ECHA, 2017b), which are also predicted for TM-BPF (Mansouri et al., 2020). Furthermore, the ED properties of TTB-BPF are currently under investigation via a substance evaluation within REACH (ECHA, 2020c).

The absence of chemical similarity to BPA does not by definition mean no concerns. For instance, bisphenol S (BPS) is not considered to be structurally similar to BPA by the models (i.e. substance nr.9 in Fig. 3). However, biological analysis indicates that BPS – although via different pathways – has the potential to interfere with the endocrine system (NTP, 2017). Therefore, prioritization of alternatives ideally consists of a combination of chemical similarity with biological similarity, as for instance conducted by the NTP for several bisphenols (NTP, 2017). Within such evaluations, also chemical similarities to other SVHCs should be considered in order to prevent regrettable substitution. For instance, when BPS (substance nr.9 in Fig. 3) will be identified as ED-SVHC in future, substance nr.10 will be considered as a potential SVHC by the model (when applying current threshold values).

3.4.3. Interpretation of model results

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

For instance, chemical similarity does not mean that there is always a hazard concern. It is possible that a chemical has a high similarity with a SVHC, but that the specific functional group causing the concern is missing (see Table 7, Chemical nr.1 – the aromatic amine is the reason for the carcinogenic effects). On the other hand, absence of similarity does also not guarantee absence of toxicological concern. A chemical could exert specific effects via different mechanisms than the currently known SVHCs, or the structural overlap might just be too low according to the model (see Table 7, Chemical nr.2).

Nevertheless, our earlier work showed that the structural similarity model for identifying carcinogenic/mutagenic SVHCs performs better than a well-known structural-alert screening model applied to the same dataset (Wassenaar et al., 2019). This indicates the relevance of full chemical overlap for toxicity prediction, and might be explained by a closer relationship with partitioning properties of substances (as also illustrated for the phenolic benzotriazole backbone in the previous section). In addition, other fragments present in the substance may function as a (steric) shield, adjusting the stability or reactivity of specific fragments/substances (Karaman, 2013).

Specific model limitations were identified upon application to the REACH dataset, and upon comparison to the expert judgement scores. Some substances that were classified differently by the structural similarity models and the expert pool are shown in Table 7. In cases where the computer model predicts non-SVHC, whereas the experts see a toxicological concern, the substances generally have several functional fragments in common with the SVHC substance, that – according to the expert – are potentially responsible for the effects. However, the models do not regard them as similar, as their total structural overlap is considered insufficient. This is for instance due to differences in the linkage of atoms or the presence of other functional groups (e.g. Chemical nr.3–5, Table 7).

The cases for which the model predicts SVHC and the experts see no concern, differ per model. For the CMR-CT4 and ED model, these

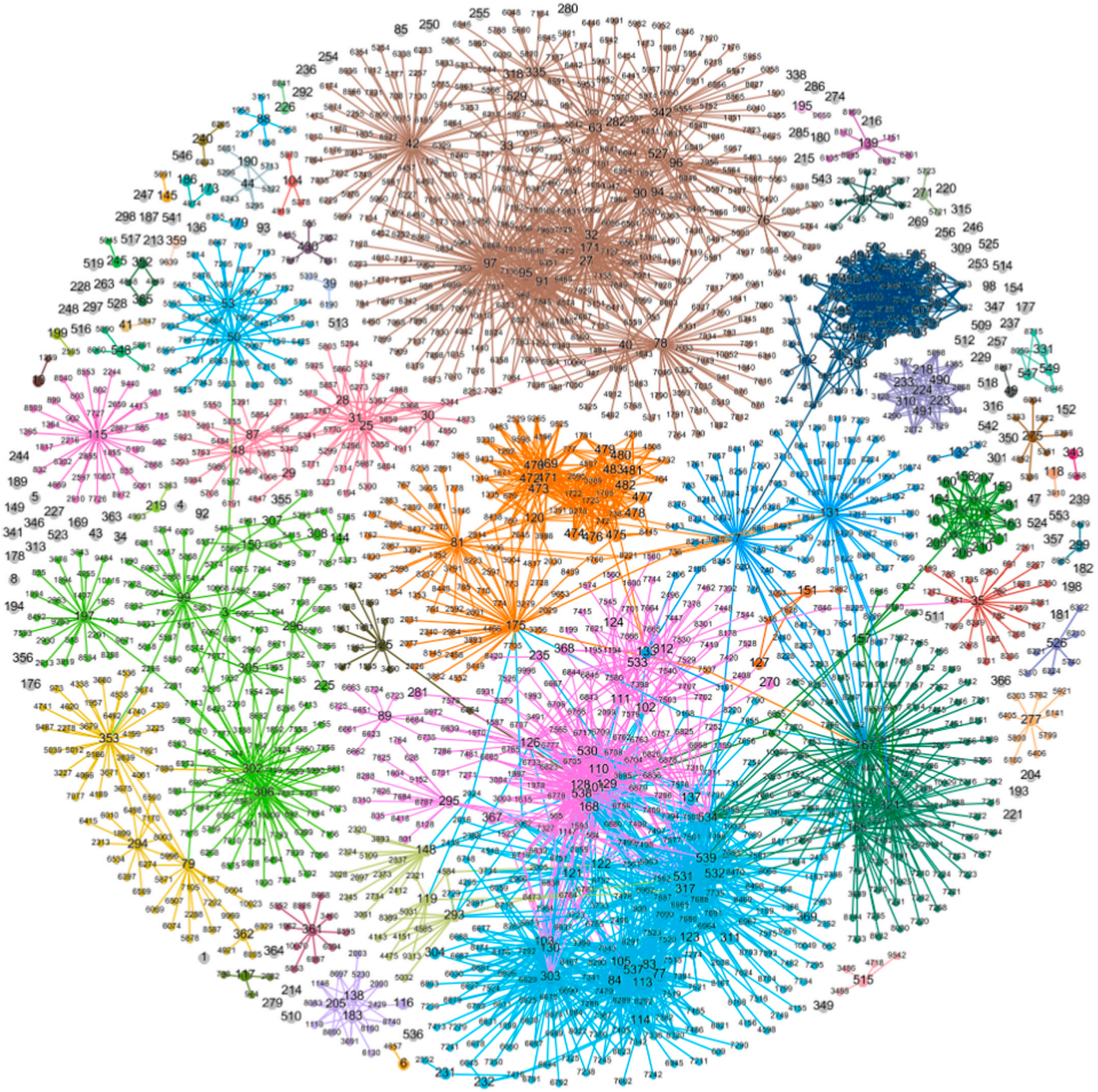


Fig. 2. Chemical similarity network of the REACH dataset substances and CMR-SVHC substances. The chemical similarity network was generated using Gephi by using Fruchtmann-Reingold layout and the similarity thresholds of the structural similarity model (see Table 1). Each node represents a chemical with corresponding ID-number. The numbers with a large font size represent the ID-numbers of CMR-SVHC substances (Table S.1) and the numbers with smaller font size represent the ID-numbers of REACH dataset substances. The lines represent a chemical similarity (i.e. similarity value above the model threshold) between the REACH dataset substance and a CMR-SVHC substance as predicted by the structural similarity model. The length of the lines (i.e. the distance between two connected nodes) does not represent the extent of similarity (i.e. the height of the similarity values as predicted by the similarity model). The colors represent clusters of substances that are considered to be structurally similar to the same SVHC substance(s). The clusters are predicted by Gephi using 'Modularity Class'. Some examples of classes: Brown – small organic oxygen compounds (e.g. nr. 32); Orange – phenols (e.g. nr.175), Blue – aromatic amines and nitro-aromatic compounds (e.g. nr. 303 and 317); Green – polycyclic aromatic hydrocarbons (e.g. nr. 158); Dark blue and purple – phthalates (e.g. nr. 224 and 501); Light pink – small chlorinated and brominated organic compounds (e.g. nr. 25); Light green – diphenyl methane-backbones (e.g. nr. 293). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

substances generally miss a specific fragment that is considered important for the concern by the experts (e.g. Chemical nr.6–7, Table 7). For the CMR-SM model, and partially also for the CMR-CT4 model, also other differently classified substances are identified, that are related to the presence and absence of ring-structures. Due to the use of the CDK extended fingerprint in the CMR model – which is a path-based fingerprint – not many additional fragments are identified for substances with a straight-chain of (carbon) atoms or when these atoms are structured in

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

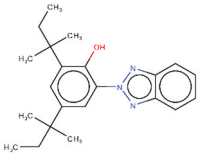
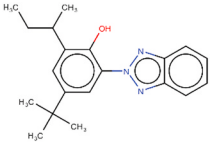
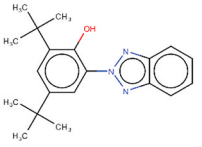
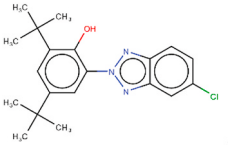
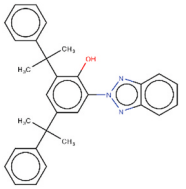
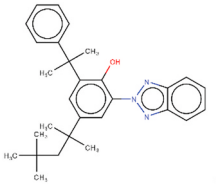
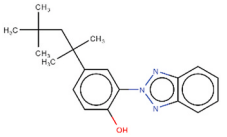
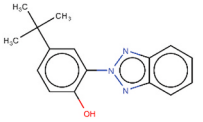
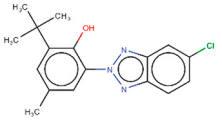
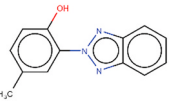
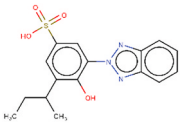
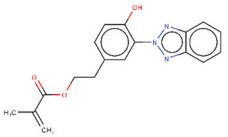
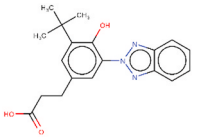
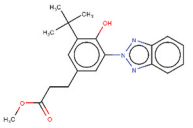
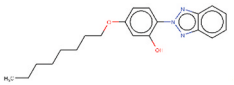
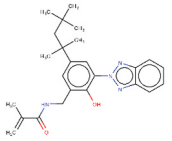
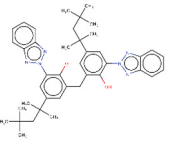
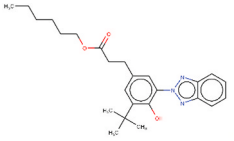
SVHC	1) 	2) 	3) 	4) 	
REACH substance *Structurally similar*	5) 	6) 	7) 	8) 	9) 
REACH substance *Not structurally similar*	10) 	11) 	12) 	13) 	14) 
	15) 	16) 	17) 	18) 	

Table 6
Additional substance information for the group of phenolic benzotriazoles as depicted in Table 5 (linked based on ID number). REACH-ID represents the number of the substances as depicted in Figure S.5. Substance properties were predicted with EpiSuite (US EPA, 2012). Values in bold meet the P/B-screening criteria (ECHA, 2017a).

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

[a] Undergoing PBT assessment.
[b] Regulatory management option analysis (RMOA) on persistence and ED.
[c] RMOA on CMR.
[d] RMOA on persistence, human health and reprotoxicity.
[e] Substance evaluation on sensitization (concluded: no follow-up).

a ring. Consequently, substances with this kind of variation (i.e. linear versus ring) could be considered as similar by the model, whereas this is not perceived as similar by the experts (e.g. Chemical nr.8, Table 7). For the PBT/vPvB model on the other hand, additional differences in classifications are related to the disregard of counts of specific fragments (e.g. counts of halogen substituents – multiple halogens make a substance more PBT-like; counts of aromatic structures – polyaromatic hydrocarbons are considered PBT/vPvB, where monoaromatic hydrocarbons are normally not PBT/vPvB, etc.; e.g. Chemical nr.9–10, Table 7). Furthermore, the type of halogen (i.e. F, Br, Cl, I) is not considered within the fragments as defined by the MACCS fingerprint. The type of halogen is regularly considered as important for the PBT/vPvB properties by the experts (e.g. Chemical nr. 11, Table 7), but is not always decisive (e.g. Chemical nr.12, Table 7). As a consequence of the underlying methodology (i.e. the applied fingerprint), a lower balanced accuracy is observed for the PBT/vPvB model when predicting the expert elicitation results. The abovementioned classification errors were also observed in the different clusters of REACH dataset substances visualized in Fig. 2 and Figure S.5-S.6.

4. Discussion

The goal of this study was to investigate the application performance of the newly developed structural similarity models (Wassenaar et al., 2019) on the broader universe of chemicals. As currently no external validation set could be developed based upon toxicological studies and regulatory decisions, we used expert judgement scores regarding the toxicological similarity between known SVHCs and chemicals with unknown SVHC properties, to derive a pseudo-external validation set. The use of expert opinion was particularly suitable as the ultimate goal of the

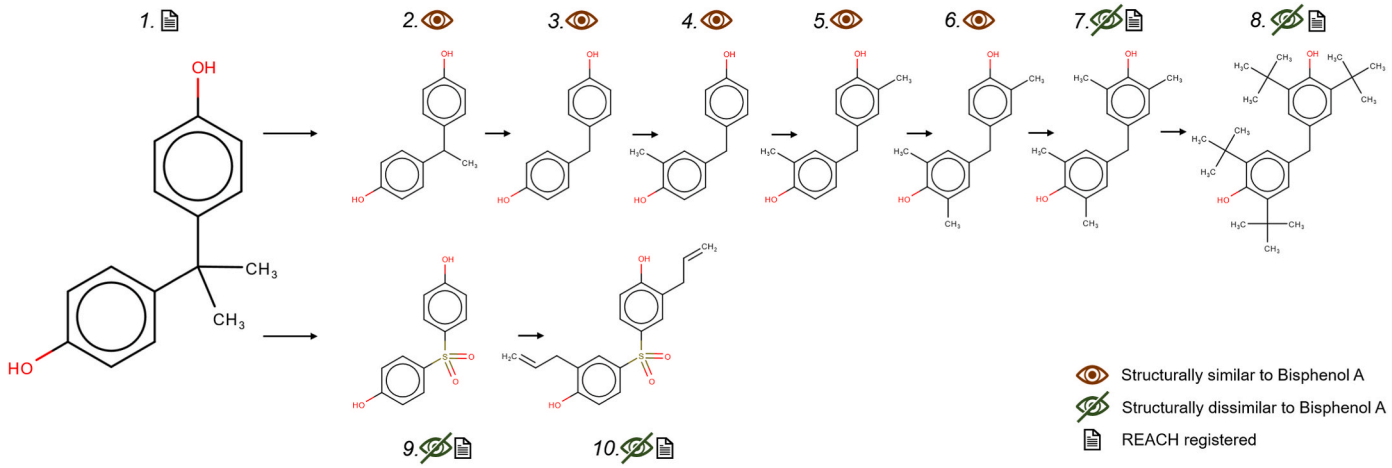
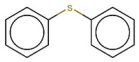
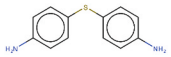
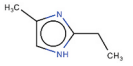
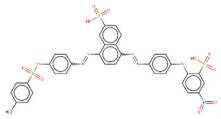
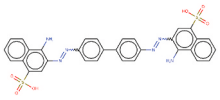
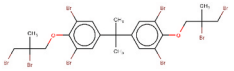
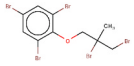
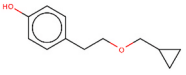
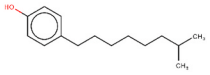
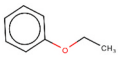
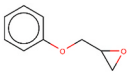
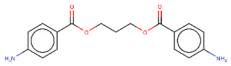
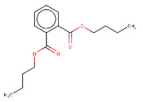
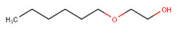
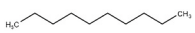
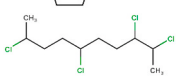
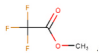
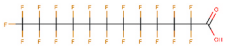
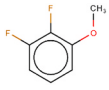
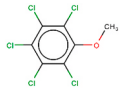
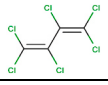


Fig. 3. Structural similarity of several bisphenol A (BPA) analogues to BPA according to the CMR and/or ED models. Model thresholds are 0.944 and 0.866 for the CMR and ED model, respectively (Table 1). 1 = BPA (CAS: 80-05-7); 2 = bisphenol E (CMR model 0.99, ED model 0.89); 3 = bisphenol F (BPF; CMR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

Table 7
Examples of substance-pairs that are differently classified by the computer similarity models when compared to the expert elicitation results. The first two substance-pairs are illustrative cases, which were not part of the expert elicitation.

ID	REACH substance	Most similar known SVHC	SVHC subset	Model prediction (sim. Value)	Expert prediction (proportion [a])
1			CMR (CT4)	SVHC (0.905)	- [b]
2			CMR (SM)	Non-SVHC (0.925)	- [c]
3			CMR (SM)	Non-SVHC (0.794)	SVHC (0.933)
4			PBT/vPvB	Non-SVHC (0.964)	SVHC (0.759)
5			ED	Non-SVHC (0.847)	SVHC (0.633)
6			CMR (CT4)	SVHC (0.906)	Non-SVHC (0.433)
7			ED	SVHC (0.873)	Non-SVHC (0.2)
8			CMR (SM)	SVHC (0.964)	Non-SVHC (0.4)
9			PBT/vPvB	SVHC (0.97)	Non-SVHC (0.133)
10			PBT/vPvB	SVHC (0.976)	Non-SVHC (0.467)
11			PBT/vPvB	SVHC (0.988)	Non-SVHC (0.5)
12			PBT/vPvB	SVHC (0.988)	SVHC (0.567)

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

computer similarity models is to provide an automated, fast and reproducible alternative to expert opinion, as expert consultation requires much more time, manpower and therefore money. Based on our analyses, comparable performance statistics were observed for the CMR model (balanced accuracy of 0.81 reproducing the expert elicitation), when compared to the predictive performance previously determined during internal validation (balanced accuracy of 0.80) (Wassenaar et al., 2019). For the ED model a relative high balanced accuracy was observed (0.87) reproducing the expert elicitation (compared to a balanced accuracy of 0.99 during internal validation), whereas a moderate balanced accuracy was observed for the PBT/vPvB model (0.69, as compared to a balanced accuracy of 0.95 during internal validation). In addition, we provided several examples for application and result interpretation of the models.

4.1. Application performance

4.1.1. Variation in expert elicitation

The expert scores showed a clear relationship between chemical similarity and toxicological similarity (Fig. 1). This indicates that, in general, a high chemical similarity is expected to be related to comparable toxic concerns. Accordingly, the key assumption of the structural similarity models seems valid, and is reproducing the assumptions used by toxicological experts. In addition, the relation between the expert SVHC predictions and the data as available within REACH is illustrated in Supplemental Material S.4 for several substances in the dataset.

When looking in more detail to the expert scores, it can be observed that there is some variation in scores across experts (Figure S2-S3). Variation between expert similarity scores has been observed and described in earlier studies, and is suggested to be related to intuition, perception and experience of the assessor (Franco et al., 2014; Lajiness et al., 2004; Lester et al., 2018; Maggiora et al., 2014). Intuitively and unconsciously, assessors reduce the complexity of chemical structures and score chemical similarity based on only a few structural features or patterns that are perceived as most essential (Kutchukian et al., 2012; Maggiora et al., 2014). The essentiality of the structural features or patterns may differ per individual based on their scientific experience (Franco et al., 2014; Maggiora et al., 2014). In addition, it has been suggested that the alignment of chemicals as provided to the experts (e.g. in which rotation/angle the non-SVHC and SVHC structures were shown) may influence the perception of similarity across experts to a different extent (Franco et al., 2014). Also, the similarity scale may not be interpreted in a uniform manner by the different assessors (Maggiora et al., 2014). Although measures have been taken to provide a guide for scaling, slight differences in applied scales were also observed in this study (Figure S3; i.e. some assessors only provided scores in the range of

2–4). Furthermore, it has been suggested that similarity scorings are context-dependent (i.e. dependent on the order of substances in the survey) (Maggiora et al., 2014). Therefore, we provided the experts with different (random) orders of the substances and subsets. Despite the variation among experts, the results indicate that averages are clearly related to chemical similarity. Therefore, the group averages can be considered as much more valuable than single expert scores, as in line with previous conclusions (the wisdom-of-crowds principle) (Hack et al., 2011; Lajiness et al., 2004), and were therefore applied in this research.

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

4.1.2. Performance considerations

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

Differences in substance classifications – between the models and the experts – were particularly related to the absence or presence of a specific functional fragment, as several fragments are being related to a specific effect (Benigni et al., 2008; Lombardo et al., 2014). Although the absence or presence of a single fragment could influence the toxicological concern, total chemical similarity may not be affected significantly, and therefore could result in different classifications between the experts and the models (Mellor et al., 2019). Vice versa, the absence or presence of a functional fragment does not necessarily mean that a specific effect will occur. The advantages of using full chemical overlap over structural alerts, is the closer relation to partitioning properties of substances and therefore also potentially to toxicokinetic and toxicodynamic processes. This is also reflected in the predictive performance of the models, according to internal and external validation, and suggests that equal treatment of mechanistically relevant and irrelevant fragments by the similarity models may not be a huge problem in practice. Acknowledging the limitations, these models show great potential to be applied in screening and prioritization approaches.

4.2. Advances in screening and prioritization

Within the risk assessment of chemicals, there is a general transition from substance-by-substance assessments to group assessment approaches, in which assessments for some individual substances could be made based upon their similarity to other tested chemicals within the group (read-across) or based on simple trends observed within the group (ECHA, 2020d; OECD, 2014). As illustrated in Fig. 2 and section 3.4, the structural similarity models could be used to identify relevant groups of chemicals that are structurally similar to one or more SVHCs. Such groups could be selected for further evaluation, in which then also biological similarity needs to be considered, including bioavailability, degradation, bioaccumulation, physicochemical properties and toxicity (ECHA, 2017c). In addition, the structural similarity models could also be used to fine tune read-across in groups that are predefined based on their biological mechanism, as proposed by Mellor et al. (2019) and illustrated in Fig. 3.

Grouping of chemicals is of particular interest in terms of effective use of available information, thereby aiming to reduce animal testing and potentially speeding up risk assessment and management, and, ultimately, increasing the level of protection for human health and the environment (ECHA, 2020d; OECD, 2014). In addition, group regulations could prevent regrettable substitution to a close structural analogue with comparable technical functioning and toxic properties (KEMI, 2018).

Currently, several group prioritizations and evaluations are already being conducted by ECHA, based on their in-house screening methodology (ECHA, 2020d). Examples of concluded, ongoing or test-cases for group evaluations include non-branched aliphatic fatty acids (ECHA, 2020d), per- and polyfluoroalkyl substances (PFAS) (ECHA, 2020e), organotin compounds and polyol acrylates (ECHA, 2018). Within their recent report, ECHA highlights the importance of further optimization of the screening of groups of substances (ECHA, 2020d). The developed similarity models could potentially contribute to such advancements. In addition, to further contribute to current ongoing activities on the identification of groups of chemicals of high concern, the screening of emission and monitoring data is highly encouraged. Specifically, because this kind of data provides different insights in substances of potential concern, as not all substances that are emitted to the environment are registered within specific legislations.

In addition, the developed structural similarity models could help define follow-up directions. As the models analyze chemical similarity to known SVHCs, the specific concern of the most similar SVHC(s) provides a relevant trigger and direction for follow-up analysis. Information on the specific concern could be combined with integrated testing and assessment strategies (ITS) – as included in several regulatory guidelines (ECHA, 2017a) – to define directions for further analysis. For example, when the most similar SVHC is considered a mutagen based on point-mutations, results of relevant Ames tests could be evaluated first. When a SVHC is considered as PBT with specific toxicity to algae, aquatic algae tests could be given higher priority compared to invertebrate or fish toxicity data.

4.3. Notes on application and future recommendations

The methodology as analyzed and evaluated in this research is made available in the form of a web-based tool at https://rvszoeksysteem.rivm.nl/ZzsSimilarityTool or in the form of a R-script in the supplemental material of Wassenaar et al. (2019). The tool should be applied as a first screening model and could help in prioritization and grouping of substances. It should be noted that similarity to the broader list of Dutch SVHC is investigated (extracted on 01-03-2018), rather than the smaller list of EU-SVHC under REACH (see Wassenaar et al. (2019) and Supplemental Material S.1 for more details). Furthermore, it should be highlighted that several SVHCs that are classified as such based on an 'equivalent level of concern', are not yet included in the models. This includes substances that are considered SVHC based on sensitizing properties, SVHCs with specific target organ toxicity after repeated exposure (STOT-RE), and SVHCs with persistent, mobile and toxic (PMT) properties. Additionally, the number and variation in ED substances that are currently classified as SVHC – and thus included in the model – is limited (n = 52). Therefore, substances with a steroid backbone will currently not yet be identified as similar to an existing ED substance, although they can be expected to have endocrine disrupting properties. Furthermore, we advise to not only apply the predictive model to the parent substance, but also to the breakdown products, as this may give different similarity outcomes. In addition, we noted that

the SMILES standardization step, as included in the workflow of the models, does not consistently apply to phthalates in the CMR model (i.e. some phthalate SMILES are adjusted, whereas others were not adjusted). This is likely related to a bug in the SMILES-standardization step of the PaDEL-Descriptor, in line with earlier reported bugs (Moriwaki et al., 2018; Yap, 2014). Exclusion of this step from the model workflow will not result in significant changes and conclusions (model conclusions of less than 1% of the substances in the REACH dataset change), but for individual (phthalate) substances it can make a difference. In addition, exclusion of the SMILES standardization step does not necessarily result in more reliable predictions for substances with an altered conclusion. For the PBT/vPvB model on the other hand, which is also based on a fingerprint derived from PaDEL-Descriptor, only marginal differences were observed (model conclusions of less than 0.1% of the substances in the REACH dataset change, n = 4).

Additionally, several future adjustments could be made to further improve the performance of the models. Based on the conducted analysis, the models seem to incorrectly classify substances in the direction of false positives (i.e. higher sensitivity then specificity, Table 4), though most false classified substances have a similarity to a SVHC close to the model's threshold (Figure S.4). Especially for the PBT/vPvB model – where many false positives were close to the threshold – adjustment of the threshold could be considered, depending on the application purpose of the model. For instance, adjustment of the threshold to 0.971 results in a balanced accuracy of 0.74, with markedly less false positives. Nevertheless, for screening applications false positives might be preferred in a regulatory context over false negatives. In addition, for the PBT/vPvB model specifically, future adjustments could consider an update of the underlying fingerprint. Inclusion of counts and types of halogen containing fragments will potentially improve the performance. Furthermore, the models in general could potentially be improved by expressing the results more quantitatively, instead of 'yes/no'-scores. For instance, by providing a probability score with the 'yes/no'-classification. This may improve result interpretation, as it helps to separate and identify borderline-cases from clear-cases. In addition, overall screening performance might be improved by combining the results of multiple screening models. Generally, an improved performance is observed when a consensus model is applied (Ballabio et al., 2017; Fernández et al., 2012), as underlying methods are generally based on varying types of information (e.g. structural features and physicochemical properties).

5. Conclusions

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

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

Funding source

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

The funding source had no involvement in the study and writing of the manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

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

Appendix A. Supplementary data

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

References

Ballabio, D., Biganzoli, F., Todeschini, R., Consonni, V., 2017. Qualitative consensus of QSAR ready biodegradability predictions. Toxicol. Environ. Chem. 99 (7–8), 1193–1216. https://doi.org/10.1080/02772248.2016.1260133.

Bastian, M., Heymann, S., Jacomy, M., 2009. Gephi: an open source software for exploring and manipulating networks. In: Third Int. AAAI Conf. Weblogs Soc. Media. https://doi.org/10.1136/qshc.2004.010033.

Benigni, R., Bossa, C., Jeliazkova, N., Netzeva, T., Worth, A., 2008. The Benigni/Bossa Rulebase for Mutagenicity and Carcinogenicity – a Module of Toxtree. EUR 23241 EN - 2008.

Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E., 2008. Fast unfolding of communities in large networks. J. Stat. Mech. Theor. Exp. 2008 (10) https://doi.org/10.1088/1742-5468/2008/10/P10008.

ECHA - European Chemicals Agency, 2020a. Substance infocard - 4,4'-isopropylidenediphenol [WWW Document]. URL. https://echa.europa.eu/substance-information/-/substanceinfo/100.001.133. accessed 5.27.20.

ECHA - European Chemicals Agency, 2020b. Substance infocard - 4,4'-methylenedi-2,6-xylenol [WWW Document]. URL. https://echa.europa.eu/substance-information/-/substanceinfo/100.023.980. accessed 5.27.20.

ECHA - European Chemicals Agency, 2020c. Substance infocard - 2,2',6,6'-tetra-tert-butyl-4,4'-methylenediphenol [WWW Document]. URL. https://echa.europa.eu/substance-information/-/substanceinfo/100.003.891. accessed 5.27.20.

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

ECHA - European Chemicals Agency, 2020e. Five European states call for evidence on broad PFAS restriction. ECHA/NR/20/13.

ECHA - European Chemicals Agency, 2019. Registered substances [WWW Document]. URL. https://echa.europa.eu/information-on-chemicals/registered-substances. accessed 5.17.19.

ECHA - European Chemicals Agency, 2018. Collaborative Approach Pilot Projects - Final Report. https://doi.org/10.2823/224234.

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

ECHA - European Chemicals Agency, 2017b. SVHC support document - 4,4'-isopropylidenediphenol (bisphenol A) [WWW Document]. URL. https://echa.europa.eu/documents/10162/908badc9-e65d-3bae-933a-3512a9262e59. accessed 7.17.20.

ECHA - European Chemicals Agency, 2017c. Read-across assessment framework (RAAF). https://doi.org/10.2823/619212.

European Parliament, 2006. REACH Regulation EC/1907/2006.

Fernández, A., Lombardo, A., Rallo, R., Roncaglioni, A., Giralt, F., Benfenati, E., 2012. Quantitative consensus of bioaccumulation models for integrated testing strategies. Environ. Int. 45, 51–58. https://doi.org/10.1016/j.envint.2012.03.004.

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

Gadaleta, D., Lombardo, A., Toma, C., Benfenati, E., 2018. A new semi-automated workflow for chemical data retrieval and quality checking for modeling applications. J. Cheminf. 10 https://doi.org/10.1186/s13321-018-0315-6.

Hack, M.D., Rassokhin, D.N., Buyck, C., Seierstad, M., Skalkin, A., Ten Holte, P., Jones, T. K., Mirzadegan, T., Agrafiotis, D.K., 2011. Library enhancement through the wisdom of crowds. J. Chem. Inf. Model. 51, 3275–3286. https://doi.org/10.1021/ci200446y.

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

Karaman, R., 2013. Prodrugs design based on inter- and intramolecular chemical processes. Chem. Biol. Drug Des. 82, 643–668. https://doi.org/10.1111/cbdd.12224.

KEMI - Swedish Chemicals Agency, 2018. Grouping of Chemical Substances in the REACH and CLP Regulations - PM 2/18.

Kutchukian, P.S., Vasilyeva, N.Y., Xu, J., Lindvall, M.K., Dillon, M.P., Glick, M., Coley, J. D., Brooijmans, N., 2012. Inside the mind of a medicinal chemist: the role of human bias in compound prioritization during drug discovery. PloS One 7 (11), e48476. https://doi.org/10.1371/journal.pone.0048476.

Lajiness, M.S., Maggiora, G.M., Shanmugasundaram, V., 2004. Assessment of the consistency of medicinal chemists in reviewing sets of compounds. J. Med. Chem. 47, 4891–4896. https://doi.org/10.1021/jm049740z.

Landrum, G., 2019. RDKit: open-source cheminformatics and machine-learning [WWW Document]. http://www.Rdkit.org/. https://doi.org/10.2307/3592822.

Lester, C., Reis, A., Laufersweiler, M., Wu, S., Blackburn, K., 2018. Structure activity relationship (SAR) toxicological assessments: the role of expert judgment. Regul. Toxicol. Pharmacol. 92, 390–406. https://doi.org/10.1016/j.yrtph.2017.12.026.

Likert, R., 1932. A technique for the measurement of attitudes. Arch. Psychol. 22.

Lombardo, A., Pizzo, F., Benfenati, E., Manganaro, A., Ferrari, T., Gini, G., 2014. A new in silico classification model for ready biodegradability, based on molecular fragments. Chemosphere 108, 10–16. https://doi.org/10.1016/j.chemosphere.2014.02.073.

Maggiora, G., Vogt, M., Stumpfe, D., Bajorath, J., 2014. Molecular similarity in medicinal chemistry. J. Med. Chem. 57 (8), 3186–3204. https://doi.org/10.1021/jm401411z.

Mansouri, K., Kleinstreuer, N., Abdelaziz, A.M., Alberga, D., Alves, V.M., Andersson, P.L., Andrade, C.H., Bai, F., Balabin, I., Ballabio, D., Benfenati, E., Bhhatarai, B., Boyer, S., Chen, J., Consonni, V., Farag, S., Fourches, D., García-Sosa, A.T., Gramatica, P., Grisoni, F., Grulke, C.M., Hong, H., Horvath, D., Hu, X., Huang, R., Jeliazkova, N., Li, J., Li, X., Liu, H., Manganelli, S., Mangiatordi, G.F., Maran, U., Marcou, G., Martin, T., Muratov, E., Nguyen, D.T., Nicolotti, O., Nikolov, N.G., Norinder, U., Papa, E., Petitjean, M., Piir, G., Pogodin, P., Poroikov, V., Qiao, X., Richard, A.M., Roncaglioni, A., Ruiz, P., Rupakheti, C., Sakkiah, S., Sangion, A., Schramm, K.W., Selvaraj, C., Shah, I., Sild, S., Sun, L., Taboureau, O., Tang, Y., Tetko, I.V., Todeschini, R., Tong, W., Trisciuzzi, D., Tropsha, A., Van Den Driessche, G., Varnek, A., Wang, Z., Wedebye, E.B., Williams, A.J., Xie, H., Zakharov, A.V., Zheng, Z., Judson, R.S., 2020. Compara: collaborative modeling project for androgen receptor activity. Environ. Health Perspect. 128 (2) https://doi.org/10.1289/EHP5580.

Mellor, C.L., Marchese Robinson, R.L., Benigni, R., Ebbrell, D., Enoch, S.J., Firman, J.W., Madden, J.C., Pawar, G., Yang, C., Cronin, M.T.D., 2019. Molecular fingerprint-derived similarity measures for toxicological read-across: recommendations for optimal use. Regul. Toxicol. Pharmacol. 101, 121–134. https://doi.org/10.1016/j.yrtph.2018.11.002.

Moriwaki, H., Tian, Y.S., Kawashita, N., Takagi, T., 2018. Mordred: a molecular descriptor calculator. J. Cheminf. 10 https://doi.org/10.1186/s13321-018-0258-y.

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

OECD - Organisation for Economic Co-operation and Development, 2014. Guidance on Grouping of Chemicals. ENV/JM/MONO(2014)4.

R Core Team, 2019. R: A Language and Environment for Statistical Computing. Vienna, Austria.

RIVM - National Institue for Public Health and the Environment, 2018. List of Dutch substances of very high concern [in Dutch] [WWW Document]. URL. https://rvszoeksysteem.rivm.nl/ZZSlijst/Index. accessed 3.1.18.

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

Soto, A.M., Schaeberle, C., Maier, M.S., Sonnenschein, C., Maffini, M.V., 2017. Evidence of absence: estrogenicity assessment of a new food-contact coating and the bisphenol used in its synthesis. Environ. Sci. Technol. 51, 1718–1726. https://doi.org/10.1021/acs.est.6b04704.

Todeschini, R., Consonni, V., Xiang, H., Holliday, J., Buscema, M., Willett, P., 2012. Similarity coefficients for binary chemoinformatics data: overview and extended comparison using simulated and real data sets. J. Chem. Inf. Model. 52 (11), 2884–2901. https://doi.org/10.1021/ci300261r.

US EPA - US Environmental Protection Agency, 2012. Estimation Programs Interface Suite for Microsoft Window, 4.1.

Wang, Z., Walker, G.W., Muir, D.C.G., Nagatani-Yoshida, K., 2020. Toward a global understanding of chemical pollution: a first comprehensive analysis of national and regional chemical inventories. Environ. Sci. Technol. 54, 2575–2584. https://doi.org/10.1021/acs.est.9b06379.

Wassenaar, P.N.H., Rorije, E., Janssen, N.M.H., Peijnenburg, W.J.G.M., Vijver, M.G., 2019. Chemical similarity to identify potential Substances of Very High Concern – an effective screening method. Comput. Toxicol. 12 https://doi.org/10.1016/j.comtox.2019.100110.

Yang, X., Wang, Y., Byrne, R., Schneider, G., Yang, S., 2019. Concepts of artificial intelligence for computer-assisted drug discovery. Chem. Rev. 119 (18), 10520–10594. https://doi.org/10.1021/acs.chemrev.8b00728.

Yap, C.W., 2014. PaDEL-Descriptor web page [WWW Document]. URL. http://www.yapcwsoft.com/dd/padeldescriptor/. accessed 5.27.20.