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

Pim N.H. Wassenaar[1,2 *], Emiel Rorije[1], Nicole M.H. Janssen[1], Willie J.G.M. Peijnenburg[1,2], Martina G. Vijver[2]

[1] National Institute for Public Health and the Environment (RIVM), Centre for Safety of Substances and Products, P.O. Box 1, 3720 BA Bilthoven, The Netherlands

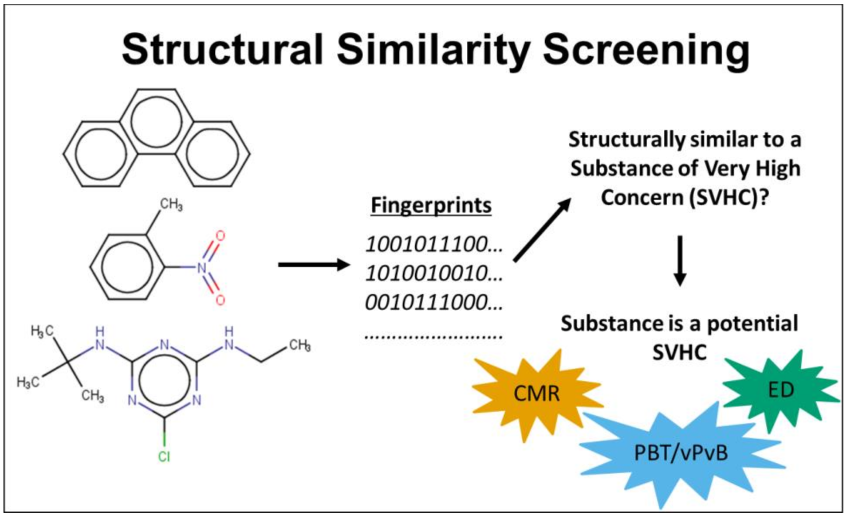
[2] Institute of Environmental Sciences (CML), Leiden University, P. O. Box 9518, 2300 RA Leiden, The Netherlands

* Corresponding author: Pim N.H. Wassenaar. Email: pim.wassenaar@RIVM.nl. Address: National Institute for Public Health and the Environment (RIVM), Centre for Safety of Substances and Products, P.O. Box 1, 3720 BA Bilthoven, The Netherlands.

Abbreviations[1]

[1] *SVHC = Substances of Very High Concern ; CMR = Carcinogenic, Mutagenic or Reprotoxic ; PBT = Persistent, Bioaccumulative and Toxic ; vPvB = very Persistent and very Bioaccumulative ; ED = Endocrine Disruption; SPOKs = Single Point of Knowledge structures ; K_{ow} = octanol/water partition coefficient ; UVCB = Substances of Unknown or Variable composition, Complex reaction products or Biological materials ; ECFP = Extended Connectivity Fingerprints ; FCFP = Functional-Class Fingerprints ; JT = Jaccard-Tanimoto coefficient ; HL = Harris-Lahey coefficient ; CT4 = Consonni-Todeschini 4 coefficient ; SS3 = Sokal-Sneath 3 coefficient ; Coh = Cohen coefficient ; SM = Simple Matching coefficient ; Yu2 = Yule 2 coefficient; TP = True Positives ; FP = False Positives ; FN = False Negatives ; TN = True Negatives.*

Graphical Abstract


Structural Similarity Screening
CH3
O
N
O
H3C
H
N
N
H
N
CH3
H3C
CH3
N
N
Cl
Fingerprints
1001011100…
1010010010…
0010111000…
……………………
Structurally similar to a Substance of Very High Concern (SVHC)?
Substance is a potential SVHC
CMR
ED
PBT/vPvB

Highlights

- Potential Substances of Very High Concern can be identified by chemical similarity.
- High balanced accuracies (≥0.8) were obtained for all SVHC-subgroup models.
- Improvement of the ED model by extending the database is considered necessary.
- The best performing similarity models can be used for screening and prioritization.

Abstract

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

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

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

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

Keywords: Substances of Very High Concern, Screening, Chemical similarity, Classification model.

1. Introduction

In recent decades, exposure to specific chemicals appeared of greater concern than previously anticipated, including concerns for polychlorinated biphenyls (PCBs), dichlorodiphenyltrichloroethane (DDT) and perfluorooctanesulfonic acid (PFOS) [1]. In many cases, when safety concerns are raised, widespread exposure has often already occurred, and typically the set of available toxicity data is inadequate to introduce risk management measures immediately. Consequently, chemicals of potential concern continue to be emitted, with the risk of significant effects on human and environmental health in the long-term. Therefore, it is important to signal emerging concerns and improve the early stage identification of hazardous chemicals before widespread exposure occurs. This endeavor is also acknowledged by the European Commission in their long-term vision towards a non-toxic environment [2,3]. In particular, high priority is given to so-called substances of very high concern (SVHC), which include substances with carcinogenic, mutagenic or reprotoxic (CMR) properties, substances with persistent, bioaccumulative and toxic (PBT) or very persistent and very bioaccumulative (vPvB) properties, or substances with endocrine disrupting (ED) properties [4]. Substances can be identified as SVHC following a regulatory decision process in which all available data is evaluated.

To improve the identification of potential SVHCs, it is essential to make efficient use of the limited amount of available (fate and toxicity) data. Several models have been described in the literature that predict hazard properties of chemicals from simple properties, like aquatic toxicity based on the octanol/water partition coefficient (K_{ow}) and/or structural alerts [5–7], or based on more complex algorithms [8–13]. Many of these models are (at least partially) based on the structural property principle, which assumes that (structurally) similar chemicals are likely to have similar properties [14]. Although these models are very useful to predict the effect of a chemical on a specific endpoint, their applicability to identify potential SVHC substances is limited. This is a consequence of the fact that the group of SVHC substances covers a broad range of different toxicological endpoints and mode of actions - and are only identified following a regulatory decision process. Within current models it is difficult to simulate such a regulatory weight-of-evidence approach. Potentially, total chemical similarity to known SVHC substances can be a useful way to estimate (potential) SVHC status, as such a method might be able to cover more information on SVHC identification properties.

To our knowledge, only two models, both with the aim of prioritization, attempt to identify potential SVHCs directly based on structural similarity to substances already identified as being SVHCs, including the SINimilarity tool developed by ChemSec [15], and screening scenarios as applied by the European Chemical Agency (ECHA) within the SVHC Roadmap program [16]. However, these methods do not provide optimized and cross-validated methodologies, resulting in an unknown predictive performance. If a high predictive accuracy could be achieved using only chemical similarity information, the lack of toxicity information can be bypassed, and those substances of potential SVHC concern, that are currently deemed “safe” in the absence of toxicity information, can be prioritized for further follow-up action. In addition, the chemical similarity information also provides a clear follow-up direction, as the potential concern is directly related to the concern of the most similar SVHC substance.

The aim of the present study was to evaluate the efficiency of a broad set of similarity measures for the identification of potential SVHCs, with a specific focus on separately identifying CMR, PBT/vPvB and ED concerns. We built upon the knowledge gained (see e.g. [17]) for calculating chemical similarity, that generally consists of two main elements: a descriptor (or representation) of the chemical structure and a similarity coefficient. First, descriptors are used to characterize the molecules that are compared by assigning numerical values to structures [17–19]. These values are in most methods related to the absence or presence of specific chemical substructures and are often encoded in fixed-length bit-strings (consisting of zeros and ones) [20]. These bit-strings are also known as fingerprints. Secondly, similarity coefficients are used to quantitatively express the similarity between two chemical descriptors [17,19,21]. For our purpose, the similarity between two fingerprints can be used to quantify the structural overlap between a chemical with unknown hazardous properties and known SVHCs. Many types of descriptors and similarity coefficients are available and there is no similarity measure that consistently is most effective (i.e. there is no single best “fingerprint - coefficient” combination for all applications)

[17,20,22]. Our study outcome provides the most optimal set of similarity measures as a first screening model to identify substances of potential SVHC concern.

2. Methods

The study approach consists of four general steps (Figure 1). First, a dataset of substances with and without CMR, PBT/vPvB and/or ED properties was constructed (paragraph 2.1). Secondly, binary fingerprints were generated for all substances in the datasets (paragraph 2.2). Thirdly, similarity values (i.e. quantitative values of chemical similarity) were calculated between substances by comparing the fingerprints with similarity coefficients (paragraph 2.3). Only the extent of similarity to substances with identified CMR, PBT/vPvB and/or ED properties leading to the SVHC status was investigated. Finally, we determined an optimal similarity threshold and the predictive performance of each "fingerprint - coefficient" combination (paragraph 2.4). Steps two to four were reiterated for multiple "fingerprint - coefficient" combinations, as well as for different SVHC subgroups (i.e. for CMR, PBT/vPvB and ED separately and together), in order to identify the optimal model(s) based on balanced accuracy. A more elaborate description of these steps is provided in the following paragraphs.

2.1 Dataset

In order to identify chemicals of (potential) concern based on structural similarity to known toxicants, a set of known CMR, PBT/vPvB and ED substances is required. For this purpose, a Dutch list of substances of very high concern was selected, as all substance on this list have CMR, PBT/vPvB and/or ED properties (see [23]; extracted on 01-03-2018). This list covers a broader range of chemicals than the EU-SVHC list under REACH, but are identified based on the same hazard criteria as the EU-SVHC substances (i.e. REACH article 57 [4]). The generation and composition of this list of substances is more elaborately described in Supplemental Material S.1.

In addition, for modelling purposes we also compiled a list of substances that are known not to have CMR, PBT/vPvB and/or ED properties. All substances on the REACH Annex IV - which lists chemicals that are considered to be inherently safe - were selected for this purpose, as well as all approved biocides and pesticides (see [24,25]; extracted on 23-05-2018). The list of biocides and pesticides is suited for our purpose as all substances approved for introduction on the European market have been tested experimentally and are negative for CMR, PBT/vPvB and ED endpoints, according to the SVHC criteria.

Several adjustments were made to the compiled substance lists, as chemical similarity searches require a specific and unambiguous chemical structure as input information. In cases that a group of substances was included in one of the above-mentioned lists (e.g. polychlorinated naphthalenes), representative chemical structures were generated and selected for inclusion in order to ensure that the structures represent the varying types of branching and/or substituents (e.g. tri- up till octachloro naphthalene, with two isomers per chlorine-atom count). When a substance is a mixture or a UVCB (Substances of Unknown or Variable composition, Complex reaction products or Biological materials), only the (representative) chemical structures of those components causing the concern were included (e.g. benzene in some of the UVCBs). When a substance is considered a non-SVHC substance, the main constituent(s) were included. Each unique chemical structure was included once in the final list. In addition, specific metal-complexes (i.e. based on arsenic, beryllium, cadmium, chromium, lead, mercury, nickel and cobalt) and fibers were excluded. For these metal-based complexes, it is generally the metal atom causing the concern, irrespective of the organic counterparts. In case of fibers, the toxicity is (also) determined by physical aspects other than their chemical structure (e.g. diameter, length and shape). In addition, all inorganic substances were removed from the list of non-SVHC substances.

In total, a dataset of 546 SVHC and 411 non-SVHC single chemical structures was compiled (see Supplemental Material Excel). Of the 546 SVHC substances, 306 are known to have CMR properties, 209 to have PBT/vPvB properties, and 52 are known to have ED properties. All chemical structures were represented by a (single) SMILES code [26] and all charged structures were converted to their neutral counterparts, where possible (Supplemental Material S.2). These SMILES codes were used for the analyses.

2.2 Fingerprints

We restricted this study to binary fingerprints based on 2D-fragments, as they tend to be more selective than whole molecule descriptors. Moreover, 2D-fragments descriptors are (computationally) easier to handle than 3D-fragment descriptors [17]. The fingerprints were selected in such a way to ensure maximum diversity and include dictionary-based, path-based, circular-based and pharmacophore-based fingerprints (Table 1) [27]. The fingerprints were generated using freely available resources, including the software packages RDkit and PaDEL-Descriptor (based on the Chemistry Development Kit (CDK) libraries) [28,29]. For all non-dictionary based fingerprints, a string length of 1024 bits was used. More details on the generation of the fingerprints are given in Supplemental Material S.3.

2.3 Similarity coefficients

The similarity between two 2D-binary fingerprints of known SVHCs and non-SVHC substances can be computed by using various formulas, the so-called similarity coefficients. When comparing two binary fingerprints, four different bit-combinations could be identified - denoted as *a*, *b*, *c* and *d*. *A*, *b*, *c* and *d* represent the counts that a feature is present in one structure and absent in the other (“x=1 and y=0”), absent in the first and present in the second structure (“x=0 and y=1”), present in both (“x=1 and y=1”) and absent in both (“x=0 and y=0”), respectively. These four numbers are combined in similarity coefficients to quantify chemical similarity. In total, 44 different similarity coefficients are available to calculate similarity values between binary fingerprints [21]. We selected seven coefficients for our analysis based on diversity and based on their performance as observed by Todeschini et al. (2012) and Floris et al. (2014) [21,30] (see Table 2). Similarity coefficients “SS1”, “Ja” and “Gle” all showed a high performance within Todeschini et al. 2012, but have an exactly similar performance as the JT-coefficient. Therefore, it has been decided to only include the JT-coefficient within this study. All included similarity coefficients were rescaled to provide similarity values between 0 and 1 using Equation 1, similar to Todeschini et al. (2012) [21].

$$s' = \frac{s + \alpha}{\beta} \qquad \textit{Equation } 1$$

Where s is the original similarity value (Table 2), s’ is the rescaled function in the range [0, 1], and α and β are numerical parameters whose values are reported in Table 2. When α = 0 and β = 1, this means that no transformation has been applied [21].

Table 1: Binary fingerprints included in this study.

Name	Number of bits	Type of fingerprint	Source
Substructure Fingerprints	307	Dictionary based fingerprints	PaDEL-Descriptor [29]
MACCS Fingerprints	166		
E-State Fingerprints	79		
PubChem Fingerprints	881		
Klekota-Roth Fingerprints	4860		
CDK Extended Fingerprints	1024	Topological or Path-based fingerprints	
Atom Pairs Fingerprints	1024		RDkit [28]
Topological Torsion Fingerprints	1024		
Extended Connectivity Fingerprints (diameter = 0) (ECFP0)	1024	Circular fingerprints *	
Extended Connectivity Fingerprints (diameter = 2) (ECFP2)	1024		
Extended Connectivity Fingerprints (diameter = 4) (ECFP4)	1024		
Extended Connectivity Fingerprints (diameter = 6) (ECFP6)	1024		
Functional-Class Fingerprints (diameter = 0) (FCFP0)	1024	Circular/pharmacophore fingerprints *	
Functional-Class Fingerprints (diameter = 2) (FCFP2)	1024		
Functional-Class Fingerprints (diameter = 4) (FCFP4)	1024		
Functional-Class Fingerprints (diameter = 6) (FCFP6)	1024		

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

Table 2: Similarity coefficients included in this study (obtained from [21]*).*

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

Simple Matching (SM)	$s = \frac{c + d}{c + a + b + d}$	0	1	S	None
Yule 2 (Yu2)	$s = \frac{\sqrt{cd} - \sqrt{ab}}{\sqrt{cd} + \sqrt{ab}}$	+1	2	Q	c=p, d=p or ab=0 → s=1

Names of the coefficients are provided as in accordance to Todeschini et al. 2012 [21], *though the definition of a and c are switched in Todeschini et al. 2012* [21]. *The column “Class” represents the type of coefficient: S = symmetric coefficient (counts a and d are considered equally); A = asymmetric coefficient (only count a is considered); Q = correlation based coefficients that are transformed to obtain a value between zero and one. The column “conditions” represents conditions that were assumed in order to avoid singularities. Den = denominator; p = a + b + c + d.*

2.4 Performance assessment

2.4.1 Performance statistics

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

For each fingerprint-coefficient combination, we determined the maximum balanced accuracy (Equation 2), by selecting the optimal threshold (i.e. a value between 0 and 1) to predict (potential) SVHC status versus non-SVHC status. Substances with a similarity value equal to or above this threshold are predicted to be structurally similar to a substance with CMR, PBT/vPvB or ED properties to such an extent that they are potential CMR, PBT/vPvB or ED themselves (and vice versa). When using a threshold value, the number of ‘True Positives (TP)’, ‘False Positives (FP)’, ‘False Negatives (FN)’ and ‘True Negatives (TN)’ predictions can be determined for a fingerprint-coefficient combination, as well as the balanced accuracy (Equation 2). By iteratively assessing the fingerprint-coefficient performance for all distinguishing threshold values (ranging from 0-1), the optimal threshold, with maximum balanced accuracy could be determined. The optimal threshold was selected for each specific fingerprint-coefficient combination to ensure equal model comparisons.

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

2.4.2 Best model selection

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

2.4.3 Best model evaluation

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

Furthermore, we analyzed the robustness of the best performing models by assessing the performance after two different robustness checks. Within the first robustness check, we extended the non-SVHC dataset by adding the substances of the “non-relevant” SVHC subgroup to the non-SVHC dataset. To illustrate, for the CMR-model, all PBT/vPvB and ED SVHC substances that do not have CMR properties were considered as not-CMR, and thus added to the non-SVHC set for this robustness check. This robustness check could not have been conducted on the overall model, as in this case all SVHC subgroups are relevant. Within a second robustness check, we reduced the number of representative structures of group entries that were included within the SVHC as well as within the non-SVHC set to generally two structures (see Supplemental Material Excel). In addition, some structurally similar substances are represented various times in the SVHC or non-SVHC datasets, including a large number of individual PCB isomers, chlorinated dibenzofurans, chlorinated dibenzodioxins and polybrominated diphenyl ethers on the PBT/vPvB dataset. To determine the robustness of the best performing models, such groups have also been reduced to a representation of generally two representative structures (see Supplemental Material Excel). The performance of the adjusted datasets within the different robustness checks was assessed similarly as described above, using the optimal threshold of the best-performing model.

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

The performance of the best predictive models was also compared to existing methodologies – using the SVHC dataset – including Toxtree (i.e. Benigni/Bossa rulebase for mutagenicity and carcinogenicity), DART and the PB-score tool [6,7,31]. For this analysis, the presence of a structural alert from Toxtree and/or DART was interpreted as a prediction of SVHC status based on CMR properties.

Besides performance evaluation, also applicability domain was analyzed by determining the 95th percentile of molecular weight, log K_{ow} [5], number of atoms, number rings and number of aromatic rings within the applied datasets.

All data was analyzed in R (version 3.5.1) [32], using *caret*, *ChemmineR*, *caTools*, *ROCR* and *rcdk* [33–37].

3. Results

3.1 Best model selection

3.1.1 Overall model performance

Table 3 shows the ten best performing models when all CMR, PBT/vPvB and ED substances are taken together in a single SVHC dataset. A wide variety of fingerprints was identified in the top ten models, including dictionary-based, path-based, circular-based and pharmacophore-based fingerprints. In contrast, one similarity coefficient, the Simple Matching (SM), is dominating the top ten models. Furthermore, it can be observed that relatively high optimal similarity thresholds are determined. The height of the threshold is highly related to the used similarity coefficient, and is specifically high for the SM coefficient (Figure S.1). This is a consequence of the fact that *c* and *d* variables are treated as similar in this coefficient (Table 2).

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

Table 3: Ten best performing fingerprint-coefficient combinations for the dataset with all CMR, PBT/vPvB and ED substances included. Also specific subgroup performances – in balanced accuracy - are provided based on the optimal overall threshold values. The numbers represent the number of SVHC substances, 411 non-SVHC substances were included. Highest balanced accuracies are given in italic bold. AUC is the area under the curve of ROC-plot.

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

3.1.2 Subgroup model performance

The best performing similarity models optimized for the separate CMR, PBT/vPvB and ED subgroups are shown in Table 4 (in row one till three, respectively). For the ED subgroup, 30 out of the 112 tested different similarity measures showed similar predictive performance, but the rank of the fingerprints and coefficients separately shows a highest rank for the FCFP4 fingerprint and the SS3 similarity coefficient. The best performing combination of fingerprint and similarity coefficient is different for the different subgroups, and a (slightly) higher balanced accuracy is obtained when compared to the best performing overall model (Table 3).

Table 4: Best performing fingerprint-coefficient combination for the CMR, PBT/vPvB and ED subgroups, including balanced accuracies after robustness checks (see section 3.2). The CMR model was improved by combining a symmetric and asymmetric coefficient in order to prevent symmetric coefficient bias (see section 3.2). In robustness check 1, the SVHC substances that did not belong to the subgroup of concern were added to the dataset as non-SVHCs. In robustness check 2, the number of representative structures for group entries and structurally similar substances were reduced to generally two structures in the SVHC and non-SVHC set. The numbers represent the number of SVHC substances. The number of non-SVHC substances varies between the full model assessment (n=411) and the robustness checks (see 3.2.2). '-' means that it is not possible to calculate a single AUC for a combination of two models. AUC is the area under the curve of ROC-plot.

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

3.2 Best model evaluation

3.2.1 Symmetric coefficient bias

By applying the “Extended fingerprint – SM coefficient” combination for the CMR dataset, with a 0.944 similarity threshold, all substances with less than 63 fingerprint bits were considered to be similar to CMR-SVHCs (Figure 2A). This coefficient bias is also observed upon visual inspection of the FP-substances, perceiving a better similarity assessment with increased number of fingerprint bits (e.g. ‘methyl octanoate’ and ‘3-propanolide’; or ‘Captan’ and ‘Captafol’; Figure 2B).

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

3.2.2 Robustness checks

The robustness of the best-performing subgroup models was investigated via two robustness checks (Table 4). Within the first robustness check, the SVHC substances that did not belong to the subgroup of concern were added to the dataset as non-SVHCs (i.e. ‘robustness check 1’). For the best performing CMR model, 651 non-SVHC substances were included, for the best PBT/vPvB model 748 non-SVHC substances and for the best ED model 905 non-SVHC substances. Within the second robustness check, we reduced the number of representative structures for group entries and structurally similar substances of the SVHC and non-SVHC set to generally two structures (i.e. ‘robustness check 2’). In total, 30 substances were excluded from the non-SVHC set, 35 from the CMR subset, 96 from the PBT/vPvB subset, and 34 from the ED subset.

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

3.2.3 Single-point-of-knowledge

The CMR and PBT/vPvB subgroup have a quite broad basis with 306 and 209 substances, respectively, whereas the ED subgroup only consists of 52 substances. Within the PBT/vPvB and ED subgroups, some groups of very similar structures can be identified, and only a few single-point-of-knowledge structures (SPOKs) are included (Figure 3). SPOKs are substances that are not comparable to any other substance in the subgroup and thus are single-point-of-knowledges within the dataset (i.e. the FN). Within the ED substances, four groups and one distinct substance are present; in the PBT/vPvB subgroup, 15 groups and 17 distinct substances were identified (giving 1 and 17 false negatives, respectively). On the contrary, the CMR-SVHC dataset is much more diverse in chemical structures and contains much more SPOKs, reflected in the high number of FN-classified substances (n=107). For the CMR subgroup, no unambiguous hierarchical clustering can be generated as the CT4-SM coefficient combination does not fulfill the mathematical conditions for all substances (i.e. similarity between substance x and y is not necessarily similar to the similarity between y and x). Nevertheless, some groups can be identified, including polycyclic aromatic hydrocarbons, haloalkanes, cyclic and acyclic ethers, alkyl phenols, phthalates, aromatic amines, nitroaromatics and chloroaromatics. As a consequence of the high structural diversity, the calculated balanced accuracy is also lower for the CMR subgroup compared to the

PBT/vPvB and ED groups. It should be noted that the SPOK false negatives will be included in the full dataset of SVHC substances when applying the model to a new substance.

3.2.4 Performance of existing models

The performance of a CMR model (i.e. the sum outcome from Toxtree and DART [7,31]) on the used SVHC-set was analyzed. Substances were considered as CMR by the model when a Toxtree or DART alert was identified. A balanced accuracy of 0.62 was determined, with a sensitivity of 0.78 and a specificity of 0.47. Furthermore, the performance of a PBT model was evaluated (i.e. PB-score tool [6]). For four substances no PB-score could be calculated as no log K_{aw} could be estimated. For the used dataset, a balanced accuracy of 0.73 was determined, with a sensitivity of 0.53 and a specificity of 0.93. No ED model was analyzed because of the limitations identified in the ED-similarity model (see discussion).

4. Discussion

As ever-increasing amounts of substances are produced, applied and emitted, it is important to focus attention on assessing the risks of those substances that are most likely to actually cause problems. Therefore, there is a need for efficient screening and prioritization methods to identify chemicals with a high potential of being hazardous. Within this study we evaluated the efficiency of a set of similarity measures for the identification of (potential) SVHCs. Based on our approach, we identified the three best performing models for CMR, PBT/vPvB and ED subgroups, that all show a promising balanced accuracy ($\geq$0.8) based on the used dataset.

4.1 Model performance

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

For the PBT/vPvB substances, the MACCS fingerprint performed best. The MACCS fingerprint contains only 166 predefined bits and was particularly developed to categorize substances in functional groups [38]. The PBT/vPvB dataset has a low structural diversity, with many substances sharing common structural features (Figure 3), including aromatic-rings and high levels of halogenation. In addition, small substances are often not considered PBT/vPvB, as in general a lower octanol-water-partitioning is observed for smaller substances, and this in turn is related to the bioaccumulation potential [39]. Apparently, the MACCS fingerprint is very effective in making a distinction between PBT/vPvB and non-PBT/vPvB substances based on these common features. Consequently, a high predictive performance is observed for this dataset (0.951).

The CMR substances are structurally much more diverse, with 107 SPOKs in the SVHC dataset. This diversity is also reflected in the most optimal fingerprint, the Extended Fingerprint. This path-based fingerprint, which is based on the well-known Daylight fingerprint [40], recognizes all paths within a structure consisting of 1-9 atoms (i.e. search depth of 8 bonds) and also includes some additional bits that describe ring features [29]. Compared to dictionary-based fingerprints, it is assumed that this method is more suitable to capture the broad diversity in CMR substances, as it characterizes all possible fragments within a structure.

As the balanced accuracy for the CMR subgroup was relatively low (compared to the PBT/vPvB and ED groups), we added an extra fingerprint that encodes for the presence of CMR-specific fragments identified in expert-models like Toxtree and DART [7,31]. Nonetheless, the inclusion of the mechanistically based substructures in the fingerprint did not lead to any improvement in the predictive performance (Supplemental Material S.5). Apparently, the size of the dataset and the fragments present in the optimal fingerprint already cover the specific structural features that have been linked to our collective knowledge of mechanisms of action leading to CMR effects. The additional fingerprint is therefore excluded again.

For ED substances, the FCFP-4 is identified as best performing fingerprint. FCFP-4 identifies fragments based on functional group patterns. It recognizes atoms as hydrogen donors, hydrogen acceptors, aromatics, halogens, basic-atoms and acidic-atoms, and it identifies fragments based on patterns between these atoms (e.g. hydrogen donor – hydrogen acceptor – hydrogen donor) [28]. Endocrine disruptors generally interact with specific hormone receptors or interact with proteins in the hormone pathway [41], and such (receptor) binding properties are potentially identified best by the features covered in the FCFP-fingerprint. Furthermore, the diameter of 4 (FCFP-4) scored slightly better for the similarity search than a diameter of 2 or 6, which is in line with earlier findings [42]. Rogers and Hahn (2010) [42] concluded that a diameter of four is typically sufficient for similarity searches whereas a diameter of six or eight is best for activity learning methods.

Despite the very high performance for the ED subgroup (0.990), prediction results from this model should be interpreted with caution. The currently used ED-SVHC dataset is limited as it only consists of a few number of substances that have a large structural overlap (Figure 3) and consequently results in higher uncertainty around the optimal threshold value compared to the other models (Figure S.3). In addition, there is only one substance on the ED-list with a hormone backbone (i.e. Diosgenin). The reason for the low number of identified ED-SVHC substances is partially related to the fact that only those substances are identified as ED for which SVHC-identification is of added regulatory value. In addition, only recently guidance and criteria are developed for the identification of ED substances [43]. It is recommended to further develop the ED model when more substances are classified as ED-SVHC, or by including known endocrine disrupting substances such as the natural substrates (and synthetic variants derived thereof) interacting with estrogen/androgen/thyroid and steroidogenic pathways. With a broader dataset, a more sophisticated screening model will be possible. Based on the current dataset the ED-SVHC similarity model is expected to miss many (potential) ED substances.

A higher performance is observed for the best-scoring CMR and PBT/vPvB similarity models compared to existing models [6,7,31], when using the SVHC dataset. This indicates the value and relevance of the structural property principle for identifying potential SVHC substances. For the ED model, no comparison was made with existing models because of the limitations as mentioned above.

4.2 Focus and restriction of the modelling

We limited our assessment to the performance of 2D-binary fingerprints, and the presence or absence of 2D-fragments. More sophisticated fingerprints are also available, including count-based fingerprints, taking into account how many times a fragment is present, or 3D-fingerprints that consider chemical conformation. Particularly, 3D-fingerprints could be relevant to identify potential ED substances, as receptor-binding properties are highly important for this group. In general, however, 2D-binary fingerprints are most popular as they are an acceptable trade-off between the wealth of (possible) information and simplicity, enabling an easy and quick comparison [17,30]. Especially for the proposed screening activities, the currently evaluated methodology is considered adequate.

In principle, all non-SVHC substances that have been used for modelling purposes within this study are tested on CMR, PBT/vPvB and ED properties. Nevertheless, it is possible that some substances are currently not identified as such, but will become a SVHC substance in future, when new information becomes available or when new evaluations are conducted. For instance, glyphosate is included in the non-SVHC list used in this study, although its carcinogenicity is currently extensively discussed [44,45]. Furthermore, as shown in Figure 2, Captafol is considered as CMR substance whereas its close structural analogue Captan is not (see Supplemental Material S.1). Captafol is classified as a carcinogen category 1B (leading to SVHC status), and Captan as a carcinogen category 2 [46]. Although the model identifies Captan as a false positive, the results could be very useful and may provide further arguments for (de)-classification of these substances. For instance, within European regulatory frameworks, a category 2 classification (for carcinogenicity but also for mutagenicity and reproductive toxicity) is often the highest classification that can be agreed upon when there are insufficient (experimental) data to support a category 1B classification [47].

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

4.3 Use and applicability domain of the model

The assumption, that structurally similar substances are likely to have similar properties, seems valid based on our analysis and model performances. The proposed similarity models focus on multiple endpoints (i.e. CMR, PBT/vPvB and ED) and could be applied as a first screening model, enabling to prioritize further follow-up analyses. The model directly highlights the most similar SVHC substance(s), which could provide additional information on the specific concerns. The absolute results should not be interpreted as a conclusive outcome. The methodology is framed to give systematic and transparent ways to identify relations that would not manually be identified. Based on the follow-up, it could be concluded that 1) the substance is likely to have similar effects, 2) that further data is required to substantiate the outcome, or 3) that the substance is not expected to have CMR, PBT/vPvB or ED properties.

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

A short guide on the application of the methodology is provided in Supplemental Material S.3. With respect to the applicability domain, an increase in reliability is observed with an increase in structure complexity for all three models, especially for the CMR model (i.e. number of atoms and different atom types). The structure similarity models are not applicable to arsenic, beryllium, cadmium, chromium, lead, mercury, nickel and cobalt-metal derivatives. For these chemicals, the metal atoms (or ions) are thought to be the cause of concern, irrespective of the (organic) groups present in the inorganic molecule.

These metal-based complexes are by definition predicted to be SVHC substances. However, the models can be used to generate a first prediction for non-dissociating metals (e.g. organotin substances). In principle, the chemical similarity itself is an applicability domain descriptor. If the new substance is sufficiently similar to an existing SVHC, the substance is clearly within the applicability domain of the model. Furthermore, physicochemical boundaries (i.e. 95th percentiles) have been calculated for the different models based on molecular weight, log K_{ow}, number of atoms, number of rings and the number of aromatic rings (Table S.2). The similarity methodology does not discriminate between pristine substances or environmental and/or metabolic breakdown products; this model is applicable to both. Risk assessors, we therefore advise not only to apply the predictive model to the parent substance, but also to the breakdown products as well as possible tautomers, as these may give different similarity outcomes.

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

5. Conclusions

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

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

6. Declaration of interests

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.

7. Acknowledgements

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

8. References

[1] S. Sauvé, M. Desrosiers, A review of what is an emerging contaminant, Chem. Cent. J. (2014). doi:10.1186/1752-153X-8-15.

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

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

[4] European Parliament, REACH Regulation EC/1907/2006, 2006.

[5] US Environmental Protection Agency (US EPA), Estimation Programs Interface Suite for Microsoft Window, v4.1, (2012).

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

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

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

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

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

ds.pdf.

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

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

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

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

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

[16] European Chemicals Agency (ECHA), Screening Definition Document - Methodology for identifying (groups of) potential substances of concern for (further) regulatory action, 2019.

[17] P. Willett, The calculation of molecular structural similarity: Principles and practice, Mol. Inform. (2014). doi:10.1002/minf.201400024.

[18] R.D. Brown, Descriptors for Diversity Analysis, Perspect. Drug Discov. Des. 7 (1997) 31–49.

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

[20] P. Willett, J.M. Barnard, G.M. Downs, Chemical similarity searching, J. Chem. Inf. Comput. Sci. (1998). doi:10.1021/ci9800211.

[21] R. Todeschini, V. Consonni, H. Xiang, J. Holliday, M. Buscema, P. Willett, Similarity coefficients for binary chemoinformatics data: Overview and extended comparison using simulated and real data sets, J. Chem. Inf. Model. (2012). doi:10.1021/ci300261r.

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

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

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

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

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

[27] S. Riniker, G.A. Landrum, Open-source platform to benchmark fingerprints for ligand-based virtual screening, J. Cheminform. (2013). doi:10.1186/1758-2946-5-26.

[28] G. Landrum, RDKit: Open-source Cheminformatics and machine-learning, Http://Www.Rdkit.Org/. (2019). doi:10.2307/3592822.

[29] C.W. Yap, PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints, J. Comput. Chem. (2011). doi:10.1002/jcc.21707.

[30] M. Floris, A. Manganaro, O. Nicolotti, R. Medda, G.F. Mangiatordi, E. Benfenati, A generalizable

definition of chemical similarity for read-across, J. Cheminform. (2014). doi:10.1186/s13321-014-0039-1.

[31] S. Wu, J. Fisher, J. Naciff, M. Laufersweiler, C. Lester, G. Daston, K. Blackburn, Framework for identifying chemicals with structural features associated with the potential to act as developmental or reproductive toxicants, Chem. Res. Toxicol. (2013). doi:10.1021/tx400226u.

[32] R Core Team, R: A Language and Environment for Statistical Computing, Vienna, Austria. (2018).

[33] M. Kuhn, R Package: caret, Ver. 6.0-81, CRAN. (2018).

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

[35] J. Tuszynski, caTools: Tools: moving window statistics, GIF, Base64, ROC AUC, etc. R package version 1.17.1.1, URL Http//CRAN. R-Project. Org/Package= CaTools. (2018).

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

[37] R. Guha, Chemical Informatics Functionality in R, J. Stat. Softw. (2007). doi:10.18637/jss.v018.i05.

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

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

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

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

[42] D. Rogers, M. Hahn, Extended-connectivity fingerprints, J. Chem. Inf. Model. (2010). doi:10.1021/ci100050t.

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

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

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

[46] European Chemicals Agency (ECHA), Search for Chemicals, (2019). https://echa.europa.eu/en/information-on-chemicals (accessed September 20, 2002).

[47] M. Woutersen, M. Beekman, M.E.J. Pronk, A. Muller, J.A. de Knecht, B.C. Hakkert, Does REACH provide sufficient information to regulate mutagenic and carcinogenic substances?, Hum. Ecol. Risk Assess. (2018). doi:10.1080/10807039.2018.1480351.

9. Figures

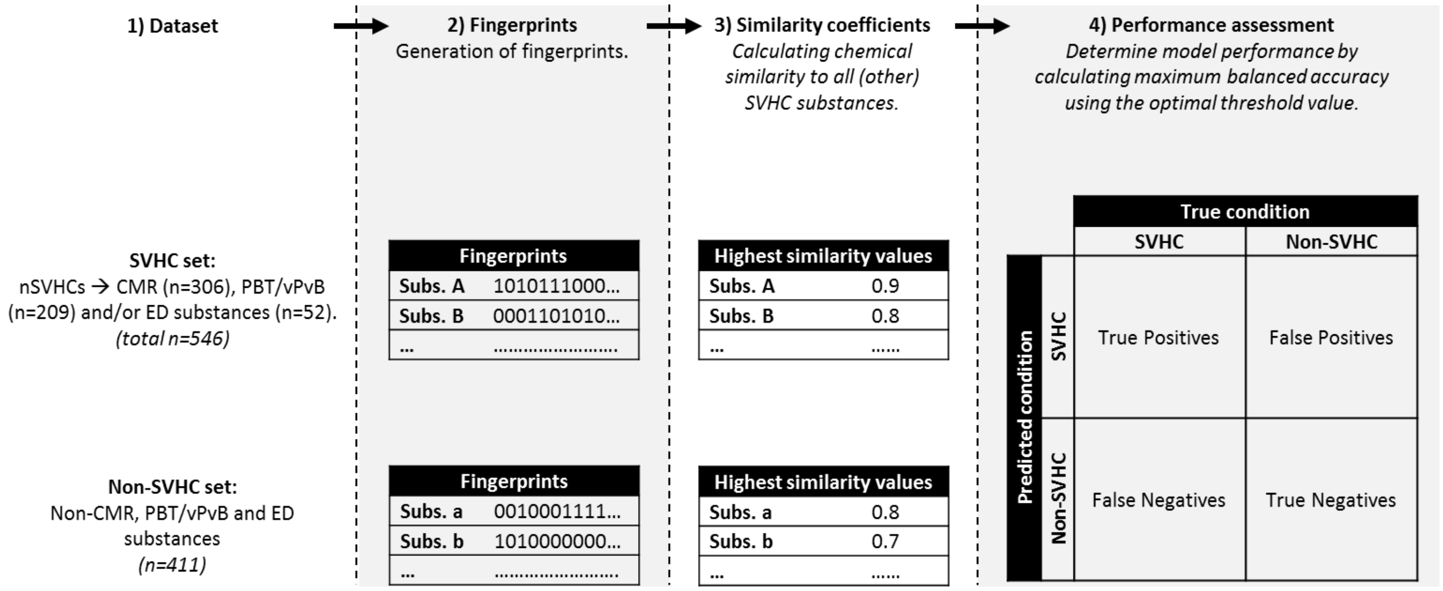
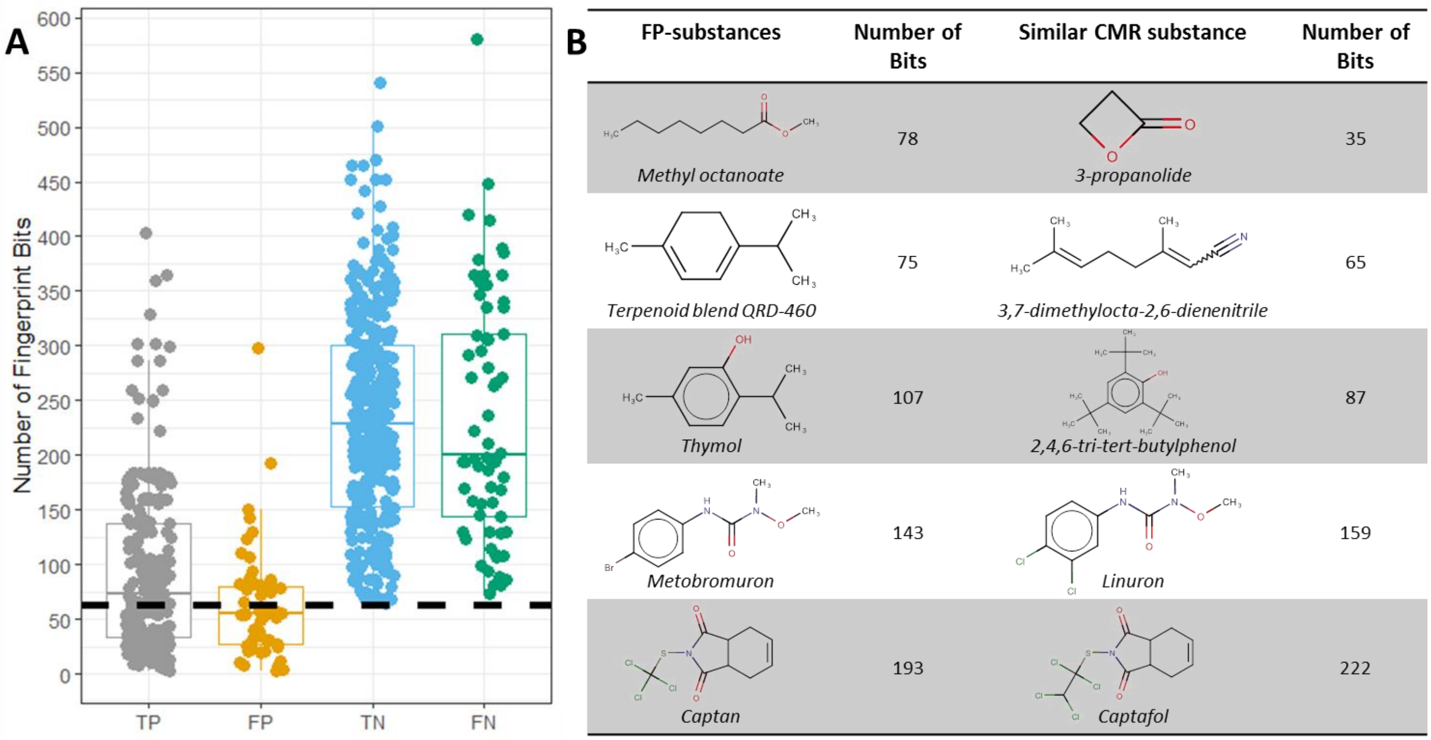

1) Dataset
2) Fingerprints
Generation of fingerprints.
3) Similarity coefficients
Calculating chemical similarity to all (other) SVHC substances.
4) Performance assessment
Determine model performance by calculating maximum balanced accuracy using the optimal threshold value.
SVHC set:
nSVHCs → CMR (n=306), PBT/vPvB (n=209) and/or ED substances (n=52).
(total n=546)
Non-SVHC set:
Non-CMR, PBT/vPvB and ED substances
(n=411)
Fingerprints
Subs. A 1010111000…
Subs. B 0001101010…
… ……………………
Fingerprints
Subs. a 0010001111…
Subs. b 1010000000…
… ……………………
Highest similarity values
Subs. A 0.9
Subs. B 0.8
… ……
Highest similarity values
Subs. a 0.8
Subs. b 0.7
… ……
True condition
SVHC
Non-SVHC
Predicted condition
SVHC
Non-SVHC
True Positives
False Positives
False Negatives
True Negatives

Figure 1. Overview of the methodology divided into four steps. Steps two to four were reiterated for multiple fingerprint-coefficient combinations.



FP-substances	Number of Bits	Similar CMR substance	Number of Bits
Methyl octanoate	78	3-propanolide	35
Terpenoid blend QRD-460	75	3,7-dimethylocta-2,6-dienenitrile	65
Thymol	107	2,4,6-tri-tert-butylphenol	87
Metobromuron	143	Linuron	159
Captan	193	Captafol	222

Figure 2: Classification of the CMR-SVHC and non-SVHC substances using the "Extended Fingerprint – SM coefficient" combination. A) Fingerprint bit count distributions across the different classifications: True Positive, False Positives, True Negatives and False Negatives. All substances with less than 63 fingerprint bits are classified as positive (dashed-line). B) Illustration of some False Positive classified substances and the most similar CMR substance. With an increase in the number of fingerprint bits, less ambiguous similarities are established.

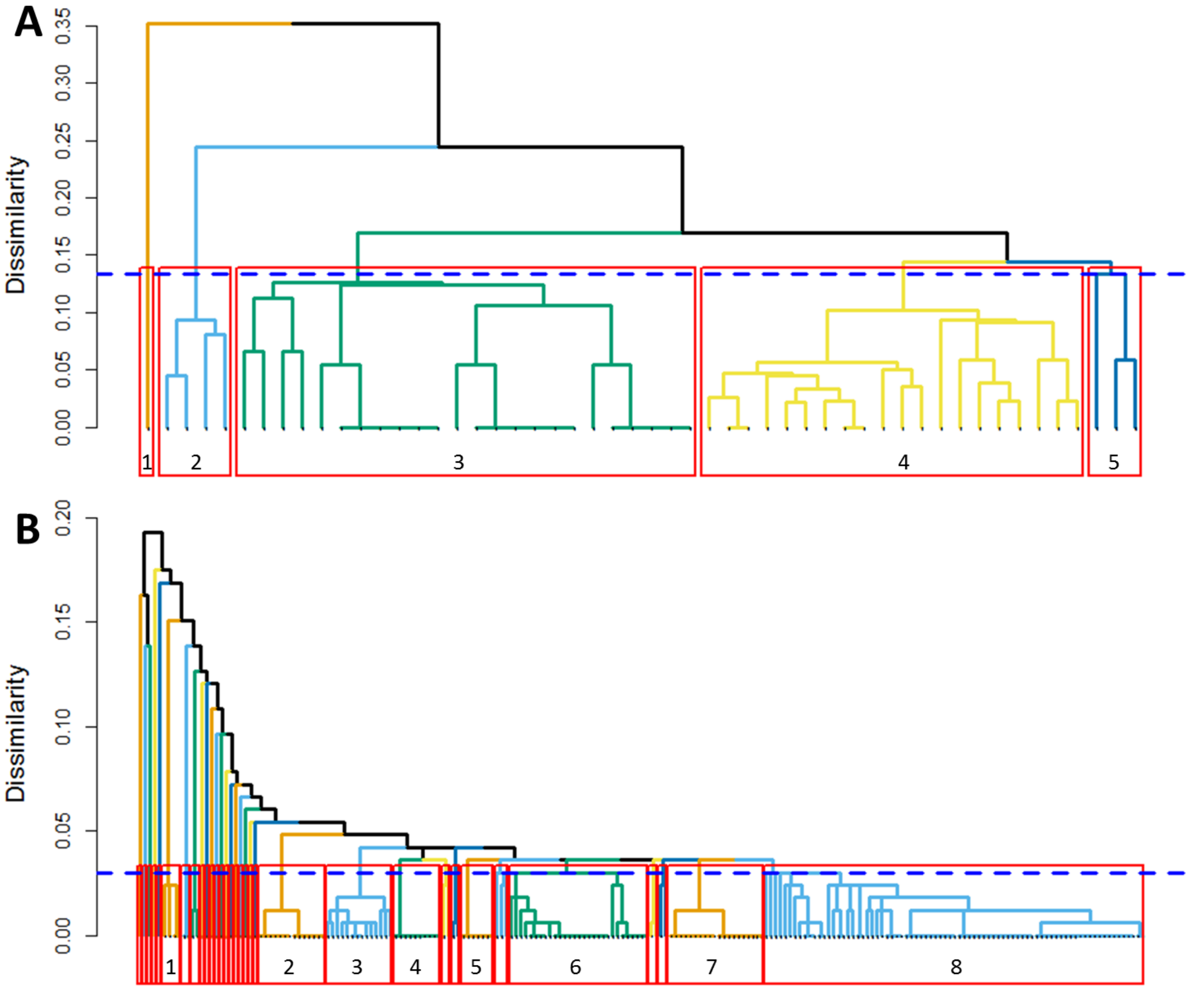


Figure 3: Hierarchical clustering for the ED and PBT/vPvB subgroups based on single linkage method. For ED, the FCFP4 fingerprint and SS3 coefficient are plotted, and for PBT/vPvB the MACCS fingerprint and SM coefficient. The y-axis describes the dissimilarity between the SVHC structures and is equal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