

SOFTWARE NOTE

ZZS similarity tool: The online tool for similarity screening to identify chemicals of potential concern

Pim N. H. Wassenaar^{1,2}  | Emiel Rorije¹ | Martina G. Vijver² | Willie J. G. M. Peijnenburg^{1,2}

¹National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands

²Institute of Environmental Sciences (CML), Leiden University, Leiden, The Netherlands

Correspondence

Pim N. H. Wassenaar, National Institute for Public Health and the Environment (RIVM), P.O. Box 1, 3720 BA Bilthoven, The Netherlands.

Email: pim.wassenaar@rivm.nl

Funding information

Ministry of Infrastructure and Water Management

Abstract

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

KEYWORDS

chemical similarity, classification model, screening and prioritization, substances of very high concern

1 | INTRODUCTION

Evaluation and regulation of chemical substances is crucial to ensure safe production and use of chemicals. For substances that are of concern, regulatory measures can be implemented that assure a minimization of emissions and exposure, and/or could stimulate the substitution by safer (non-regrettable) alternatives. Such actions contribute to the European ambitions of a toxic-free environment.¹ However, as available evaluation capacity is limited, it is essential to first evaluate (and subsequently regulate) those substances that are of highest concern. To facilitate the identification of substances of potential concern, we recently developed structural similarity models

that evaluate the structural similarity of substances with unknown hazard properties to known Substances of Very High Concern (SVHC).² Substances are identified as SVHC based on a regulatory decision process, in which available data is evaluated and compared to specific criteria (see Supplemental Material S1 for more details). When a substance is identified as SVHC, there are specific consequences for production/emission and use. This relates particularly to industrial chemicals. The developed models are based on structural similarity, which is considered an important descriptor in various research fields, including toxicology (e.g., for read-across³) and pharmacology (e.g., for virtual screening⁴⁻⁶), as a high resemblance in chemical structure could be an indication of comparable properties and effects ('similar

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2022 The Authors. *Journal of Computational Chemistry* published by Wiley Periodicals LLC.

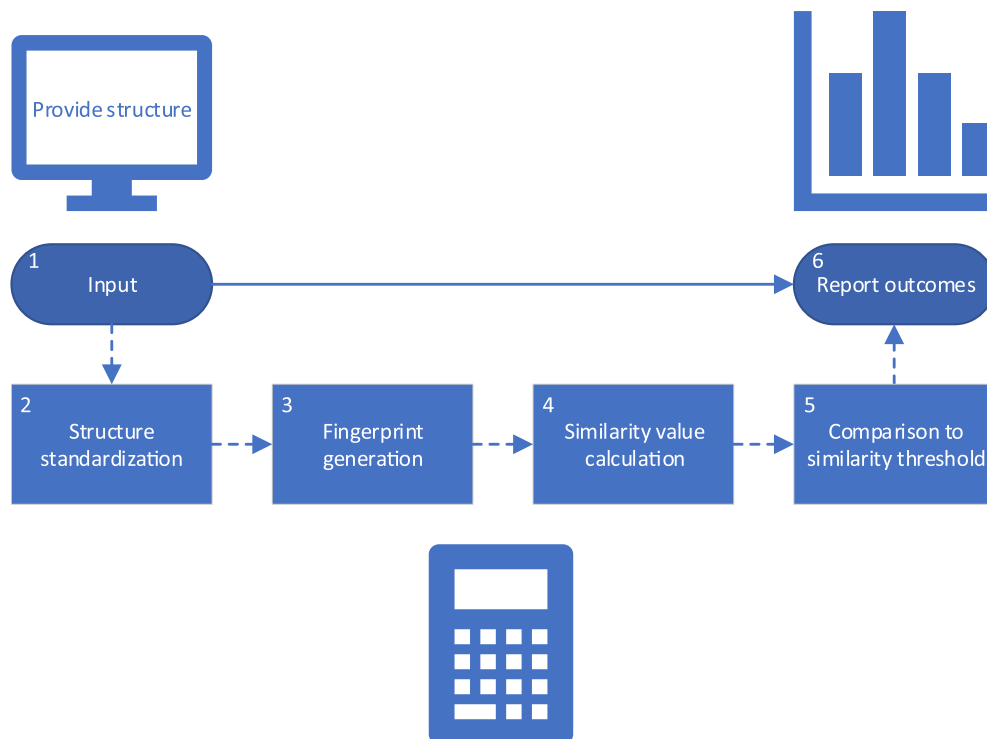


FIGURE 1 Illustration of the workflow of each of the separate similarity models that are incorporated in the ZS similarity tool (note that there are some variations for the specific sub-models, see section ‘3.2 models’). Step 1 and 6 consider the input and output as shown by the ZS similarity tool, and step 2–5 are used to calculate and predict the structural similarity. The exact specifications of step 3–5 differ per SVHC-category. An input structure can be provided as SMILES or CAS-number (step 1), which is converted to a standardized SMILES to ensure equal comparison to SVHC structures (step 2). The standardized SMILES is used to generate chemical fingerprints using PaDEL-descriptor¹³ (step 3). The fingerprint of the input structure is compared to the fingerprints of all SVHCs of a specific category to calculate similarity values by using a similarity coefficient (step 4). The calculated similarity values are compared to a similarity threshold to predict whether the input structure is considered sufficiently structurally similar to an SVHC (step 5), and the results are reported (step 6). For each SVHC-category a specific model was developed and optimized, that consists of a unique fingerprint, coefficient and threshold combination; and the outcomes are reported separately for each SVHC-category

property principle’).⁷ Therefore, substances that are structurally similar to known SVHCs might be selected for further evaluation.

The SVHC similarity models are based on chemical fingerprints and similarity coefficients^{8,9} and the workflow of the models is illustrated in Figure 1. Separate similarity models have been developed for three groups of SVHCs, including (1) SVHCs with carcinogenic (C), mutagenic (M) or reprotoxic (R) properties (i.e., CMR), (2) SVHCs with persistent, bioaccumulative and toxic (PBT) or very persistent and very bioaccumulative (vPvB) properties (i.e., PBT/vPvB), and (3) SVHCs with endocrine disrupting (ED) properties. These models showed promising performance statistics (with balanced accuracies of 0.80–0.99),² and showed a reasonable performance on a broader universe of chemicals as analyzed by a pseudo-external validation (with balanced accuracies of 0.69–0.87).¹⁰ In addition, the model predictions appear to be more robust than expert judgments.¹⁰ To enable the use of the similarity models by academia, industrial partners and risk assessors (including regulators that have to decide on emission permits), we have made the models publicly available via a freely accessible web-based tool, named the ZS similarity tool: <https://rvszoekstelsysteem.rivm.nl/ZzsSimilarityTool> (ZS = ‘Zeer Zorgwekkende Stoffen’ [in Dutch], which is literally translated as substances of very

high concern). Accordingly, this tool can be used to screen and prioritize chemicals for further action and evaluation within varying frameworks and safe-by-design trajectories, and is already applied in various screening activities.^{11,12}

Upon obtaining more experience with the application of the similarity models, we identified several methodological aspects that could be further optimized to improve the performance of the models.^{2,10} Particularly, the PBT/vPvB model misclassified various substances due to amongst others insufficient consideration of the type and number of halogenated fragments and aromatic structures. Moreover, the SVHC-categorization insufficiently reflected the current SVHC status, and the binary nature of the predictions limited the interpretation of the results. Therefore, the current study aims to improve the performance of the models by (1) separating the known SVHCs in more specific subgroups, (2) (re-)optimizing similarity models for the various SVHC-subgroups, and (3) improving interpretability of the predicted outcomes by providing a confidence score. In addition, the underlying reference dataset of SVHC substances was updated. The improvements as described in this study are directly incorporated in the ZS similarity tool, and enhance the applicability of the models.

TABLE 1 Aspects of the structural similarity models that are adjusted within the current study.

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

Abbreviations: SVHC—substances of very high concern; CMR—carcinogenic (C), mutagenic (M) or reprotoxic (R) properties; PBT/vPvB—very (v) persistent (P), bioaccumulative (B) and toxic (T) properties; CLP—classification, labelling and packaging of substances and mixtures; POP—persistent organic pollutants.

2 | METHODS

The methodological aspects of the similarity models that are adjusted in this study are shown in Table 1, and include an update of the underlying SVHC dataset, a re-categorization of the SVHCs into sub-groups, a (re-)optimization of the similarity models, and the addition of a quantitative outcome score.

2.1 | Dataset

The dataset of SVHCs was updated between 2018 to 2021 based on the substances that were included on a Dutch list of SVHCs (January 25, 2021¹⁴) following the same refinement procedure as previously described.² This list includes substances that are identified based on the same hazard criteria as the European SVHCs, but—besides the European SVHC list—are derived from various additional sources as well. Therefore, this Dutch list of SVHCs covers a slightly broader range of chemicals than the EU-SVHCs under REACH (see Supplemental Material S1 for a detailed description of the composition of the Dutch list of SVHCs).

The substances included on the refined/final SVHC list were categorized based on their hazard class (in which a chemical can belong to multiple hazard classes). Distinctions were made between the 'classical' SVHC hazard categories, including C, M, R, PBT and vPvB. Substances were added to these categories when they were considered to have such specific effects according to their inclusion on the European SVHC list, the European CLP list Annex VI, or on the list of Persistent Organic Pollutants (POPs). All POPs were considered as PBT and/or vPvB within the dataset. In addition, as within our

previous work, a specific ED category was used. Substances were added to this category when they were included on the European SVHC list based on ED effects. Substances on the Dutch list of SVHCs that do not belong to any of the above-mentioned categories were included in the 'Other'-category, like substances on the European SVHC list with persistent, mobile and toxic (PMT) properties, specific target organ toxicity after repeated exposure (STOT-RE) or sensitizing properties. In addition, substances were also included in the 'Other'-category when they were only included on the Dutch list of SVHCs based on other sources, like substances on the OSPAR list for priority action¹⁵ or priority hazardous substances according to the Water Framework Directive (see Supplemental Material S1 for more details about the Dutch list of SVHCs).

For modeling purposes, also a list of non-SVHCs was required. We used the same list as used by Wassenaar et al.,² but excluded the substances that were now included on the 'new' list of SVHCs (resulting in a total of 406 substances). This list consists of substances that are considered inherently safe (i.e., all substances on REACH Annex IV), and includes approved biocides and pesticides (which have been tested experimentally and are negative for all SVHC-endpoints).

As chemical similarity evaluations require unambiguous chemical structures as input information, we normalized and standardized all SMILES to QSAR ready structures with a Kekulé representation.¹⁶ This was done by extracting the QSAR ready structures from a CAS-SMILES list from the US-EPA¹⁷ or by generating QSAR ready structures with a KNIME workflow.¹⁸ In exceptional cases, where no QSAR ready SMILES could be generated, the most uniform representation was manually selected (e.g., from PubChem or ECHA dissemination site).

2.2 | Models

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

The selection of the best performing similarity models was specifically restricted to fingerprints that could be generated with PaDEL-Descriptor (as those are incorporated in the online ZZS similarity tool).¹³ This includes the Substructure, MACCS, E-State, PubChem, Klekota-Roth and CDK Extended fingerprint. Fingerprints were generated for all QSAR ready SMILES of the SVHC and non-SVHC substances with PaDEL-Descriptor, enabling PaDEL to remove salts, detect aromaticity, and standardize tautomers and nitro groups.¹³ These fingerprints were all tested in combination with the JT, HL, CT4, SS3, Coh, SM and Yu2 similarity coefficients.^{2,9} More details on the fingerprints and similarity coefficients are provided in Supplemental Material S2 (Tables S1 and S2).

We analyzed the predictive performance of the varying fingerprint-coefficient combinations for classifying the substances in the dataset as (potential) SVHC or non-SVHC per SVHC category (i.e., CM, R, PBT/vPvB, ED and Other). For each fingerprint-coefficient combination similarity values were calculated. Non-SVHC substances

were compared to all SVHCs, whereas SVHCs were compared to all other SVHCs (excluding itself), and per substance only the highest similarity value was retained. Next, the maximum balanced accuracy was determined (Equation (1)), by selecting the optimal threshold (i.e., a value between 0 and 1) to predict (potential) SVHC status versus non-SVHC status. Details are according to Wassenaar et al.² An overview of these various steps is provided in Supplemental Material S2 (Figure S1).

Selection of (or adjustments to) the best performing models focus on quantitative performance statistics (i.e., balanced accuracy), but also included qualitative selection criteria (which could vary between hazard classes), where necessary. For instance, in the case of a symmetric similarity coefficient (i.e., coefficients in which absence and presence of features that are in common between two structures contribute equally to the determined similarity), specific care was given to symmetric coefficient bias (i.e., the phenomenon where chemicals with less than a specific number of fragment features are always predicted to be structurally similar to an SVHC due to high overlap in absent features) (see Reference [2] for a more detailed description). Furthermore, for the PBT/vPvB-model (as well as the 'Other'-model) specific attention was given to the performance on the broader universe of chemicals, as these models had a relative low external performance in a previous evaluation.¹⁰

$$\text{Balanced accuracy (bAcc)} = \frac{\text{Sensitivity} + \text{Specificity}}{2} = \frac{\frac{\text{TP}}{\text{TP} + \text{FN}} + \frac{\text{TN}}{\text{TN} + \text{FP}}}{2} \quad (1)$$

2.3 | Outcomes

A quantitative confidence score was added to the binary model predictions (i.e., the yes or no prediction on sufficient structural similarity). The confidence scores represent the confidence in the structural similarity between a chemical and an SVHC, and are derived from the similarity values. The following stepwise procedure was followed for each similarity model. First, we iteratively assessed the performance for all distinguishing similarity values (i.e., threshold values) based on the subgroup specific SVHC and non-SVHC datasets, and derived balanced positive predictive values (bPPV) for each similarity threshold value (see Equation (2)). Second, the bPPVs were min-max normalized to confidence values ranging from 0 to 1 (i.e., 0%–100%), in which the model's optimal threshold value was set to a confidence value of 0.5 (i.e., 50%). Third, we fitted two functions through these normalized bPPVs. One function is fitted to the similarity values ranging from 0 to the model's optimal threshold (with confidence scores ranging from 0% to 50%), and the other function is fitted to the similarity values from the model's optimal threshold till 1 (with confidence scores ranging from 50% to 100%). Depending on the distribution of the bPPV for all similarity thresholds values, a corresponding function was selected (e.g., exponential or sigmoidal function). In cases where no clear distribution pattern was observed, a linear trend was used. The fitted functions at least had to cover the confidence ranges from 0.5% to 49.5% and 50.5% to 99.5%, and must sufficiently represent

the derived bPPV points, where possible. When necessary, the fit was manually optimized to meet these conditions, by for instance constraining the bottom or top of the curves at specific similarity values, or by providing additional weight to specific datapoints. A visual example of the fitting through a distribution of bPPV values as a function of similarity threshold values is given in the results section (Figure 2).

$$\text{Balanced positive predictive value (bPPV)} = \frac{\text{Sensitivity}}{\text{Sensitivity} + (1 - \text{Specificity})} = \frac{\frac{\text{TP}}{\text{TP} + \text{FN}}}{\frac{\text{TP}}{\text{TP} + \text{FN}} + \left(1 - \frac{\text{TN}}{\text{TN} + \text{FP}}\right)} \quad (2)$$

All analyses within this study were performed in R (unless otherwise specified)¹⁹ using *caret*, *ChemmineR*, *caTools*, and *ROCR*.^{20–23}

3 | RESULTS AND DISCUSSION

3.1 | Dataset

The new dataset consists of 621 substances, of which 80 structures were not yet included in the previous dataset. In addition, eight structures were removed (e.g., as they do not meet the SVHC criteria anymore), or were represented by newly included ($n = 3$) or already existing structures (see Supplemental Material Excel for more details). Furthermore, we re-categorized the substances across the hazard classes to better reflect the current SVHC status and thereby improve the interpretability (e.g., distinction between EU-based SVHCs versus SVHCs that are only identified as a Dutch SVHC; and distinction between CM- and R-concerns). The distribution of substances within this updated dataset across the different hazard categories is shown in Table 2, and the individual substances are included in Supplemental Material Excel.

TABLE 2 Overview of the new dataset and the distribution over hazard categories, in comparison to the previous dataset as included in Reference [2].

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

Note: 1—In the previous work, CM and R were combined as one class ($n = 306$). 2—In the previous work, no 'Other'-category was included. 3—The 'Other'-category consists of 10 substances that are identified as EU-SVHC based on PMT ($n = 3$) or respiratory sensitizing properties ($n = 7$). All others are not identified as EU-SVHC, EU-CLP or POP, but are included on the Dutch list of SVHCs based on specific concerns related to similar endpoints (C: $n = 3$, M: $n = 1$, R: $n = 14$, PBT: $n = 64$, PBT/vPvB: $n = 29$, ED: $n = 6$, PMT: $n = 2$, and others: $n = 2$) from other sources (e.g., OSPAR¹⁵; in which PBT/vPvB concerns are dominating).

3.2 | Models

Specifically the PBT/vPvB model required improvement according to the performance on the broader universe of chemicals.¹⁰ Despite the excellent performance on classifying substances in the original SVHC dataset, the PBT/vPvB model misclassified many substances when applied to a broader set of chemicals, due to amongst others insufficient consideration of the type and number of halogenated fragments and aromatic structures. In addition, the performance of the other similarity models (i.e., CMR and ED models) were reanalyzed as several adjustments have been made, including an update of the SVHC dataset and a new categorization of substances (i.e., CM, R, PBT/vPvB, ED and 'Other'-models).

Optimization of the CM- and R-models based on the new datasets indicated that the CDK Extended fingerprint with SM-coefficient was the best or second best performing fingerprint-coefficient combination for the CM- and R-dataset, respectively. For the R-dataset, the Extended-Coh combination scores best followed closely by the Extended-SM fingerprint-coefficient combination (with balanced accuracies of 0.814 and 0.808, respectively). These results are comparable to the results from our previous study, in which the Extended-SM fingerprint-coefficient combination outperformed all other combinations for the CMR-dataset with comparable optimal similarity thresholds (i.e., a threshold of 0.946 for the CM-dataset, 0.944 for the R-dataset and 0.944 for the combined dataset in the previous study).² We decided to additionally use an asymmetric similarity coefficient (i.e., JT or CT4 coefficient) for substances with a low number of fingerprint bits, as symmetric coefficient bias was observed. Statistical derivation of an optimal cut-off value (i.e., below which number of fingerprint bits the JT or CT4 coefficient should ideally be used) resulted in broad uncertainty ranges due to a limited number of substances in the subsets. As comparable best-performing models were derived for the new CM- and

R-dataset as previously determined, we decided to retain the CMR-model. The established optimal threshold and cut-off specifications are given in Table 3, and showed to be robust to minor changes in the dataset and do not specifically require an adjustment of the optimized parameters. Moreover, this decision was justified by the fact that besides an update of the dataset there was no specific incentive to improve the performance of the CMR-model based on the previous evaluations.

Revision of the PBT/vPvB model using the MACCS-SM fingerprint-coefficient combination was required considering its performance on the broader universe of chemicals.¹⁰ In addition, as many not (yet) EU-recognized PBT/vPvB chemicals were reallocated to the 'Other'-category, also specific attention was given to the optimization of the similarity models for this group. The Klekota-Roth, PubChem and CDK Extended fingerprint were identified as best performing fingerprints based on performance statistics for the PBT/vPvB-SVHCs and non-SVHCs. However, upon a more in-depth analysis of the predicted similarities (including false positives and false negatives) and its applicability on the broader universe of chemicals, it could be concluded that the Klekota-Roth fingerprint is not suitable to predict structural similarity amongst PBT/vPvB chemicals. The Klekota-Roth fingerprint provides a lot of emphasis to (small) linear chains of varying sizes and to relatively large fragments, but insufficiently weighs typical PBT-related fragments like aromatic-ring structures. In addition, for relatively many chemicals only a limited number of fragments are identified, and accordingly such chemicals are more easily (but often incorrectly) predicted as structurally similar to a PBT/vPvB-SVHC. The PubChem and Extended fingerprints have their own strengths and limitations. The PubChem fingerprint specifically weighs aromatic structures and halogens, but does not systematically cover the whole chemical structure. The Extended fingerprint specifically considers all fragments present within a chemical, but focusses specifically on path-based fragments which may insufficiently describe ring-structures.

TABLE 3 Overview of the final models—including performance statistics—to predict structural similarity to SVHCs.

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

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

As both fingerprints have their own unique flaws, they were combined to form the final PBT/vPvB model. The corresponding best performing coefficients included the JT, CT4 and SS3 coefficients. As the observed differences between these coefficients are in the details (see Supplemental Material S2) and are partially related to the determined optimal threshold values, we selected the fingerprint-coefficient combinations with the best performance on the expert judgment dataset (given preference to high bPPVs and few false positives, to ensure confidence in model predictions). The final PBT/vPvB model uses both the PubChem-JT and Extended-CT4 fingerprint-coefficient combinations, and only predicts that a chemical is structurally similar to a PBT/vPvB-SVHC when both models support this conclusion (see Table 3).

In the ED-dataset, four new ED-SVHCs were added and five SVHCs were allocated to the 'Other'-category. The results of the model optimization indicate that multiple fingerprint-coefficient combinations can be considered as the best-performing model, all with a balanced accuracy of 0.99 (including models based on the CDK Extended fingerprints and Klekota-Roth fingerprints). Although previously the RDKit based FCFP4 fingerprint with the SS3 coefficient was considered most optimal (with an equal performance statistics), we now pragmatically selected a PaDEL-based fingerprint as those were incorporated in the online ZZS similarity tool. We selected the CDK Extended fingerprint with JT-coefficient from the best performing fingerprint coefficient combinations. The Extended fingerprint was chosen above the Klekota-Roth fingerprint, as the Klekota-Roth fingerprint only considers a specific number of pre-specified fragments, and therefore, might be less specific when applied to a broader universe of chemicals. In addition, the JT-coefficient was selected as this coefficient uses an asymmetric function for which there is no risk of symmetric coefficient bias, and was preferred above the CT4 coefficient (see Supplemental Material S2).

The 'Other'-category consists of SVHC substances whose properties are different from the above-mentioned categories, or whose properties are not universally recognized as such. These substances were separated from the CM, R, PBT/vPvB and ED-dataset to better reflect the current SVHC status and thereby improve the interpretability. For the 'Other'-dataset very comparable observations and conclusions were made as compared to the PBT/vPvB-dataset, and this might not be a surprise considering the broad representation of PBT/vPvB related chemical concerns in the 'Other' category (>70%, see Table 2). The only differences between both models are the most optimal threshold values that are derived from the dataset (see Table 3).

Although we specifically assessed the performance of PaDEL-based fingerprints within this study, comparable or lower performances were observed for the RDKit related fingerprints that were previously tested as well.^{2,24}

3.3 | Outcomes

Within the previous models, only a dichotomous, qualitative, prediction of the concern was made for the structural similarity of a chemical to an SVHC. Based on the similarity score and model specific threshold, the models predicted whether or not a chemical is

sufficiently structurally similar to an SVHC (and thus predicted to be a potential SVHC). To support a better interpretation of the outcomes for prospective model users, a quantitative confidence score is added to this binary prediction. The developed quantitative scores describe the confidence in structural similarity between a chemical and an SVHC, with a higher confidence for higher structural similarity. This supports the intuitive interpretation that a substance that is more similar to an existing SVHC is also predicted with more certainty to have SVHC properties. The confidence score functions were derived separately for each model and are based on the normalized bPPV for substances in the SVHC and non-SVHC dataset. A similarity value equal to a model's optimized threshold was given 50% confidence, with a maximum confidence of 100% (in case of a similarity score of 1) and a minimum confidence of 0% (in case of a similarity score of 0). An example of such a function is shown in Figure 2, and a detailed overview of all derived confidence functions (including figures showing the bPPV as a function of the similarity value) is provided in Supplemental Material S3 (Table S3 and Figure S2). The functions do not aim to provide an exact confidence trigger, but are meant to provide additional (data-driven) information that could guide interpretation and follow-up evaluation. We specifically did not include a predictive score for non-similarity to an SVHC (for instance based on negative predictive values), as the models only make statements about the similarity and not the absence of similarity to SVHCs. This is related to the fact that it cannot be concluded that a substance is not a potential SVHC based on a lack of structural similarity, as a substance might exert effects through different (yet unknown) modes of action.

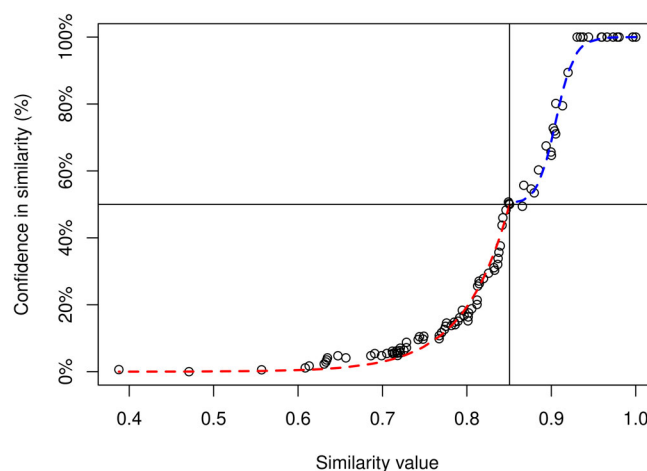


FIGURE 2 Relation between the structural similarity value and the confidence in the predicted structural similarity between a chemical and a Reprotoxic (R)-SVHC based on the CDK extended-CT4 fingerprint-coefficient combination. The fitted curves describe the normalized bPPV as a function of the similarity value used as a threshold value, and are derived from the R-SVHC and non-SVHC datasets (for substances with less than 85 fragment features, that is, bits in the CDK extended fingerprint). The vertical line represents the model's optimized threshold value (0.851) giving the best balanced accuracy, and the horizontal line represents the 50% confidence score. More details are presented in Supplemental Material S3

TABLE 4 Application of the newly optimized similarity models to a dataset of 9456 REACH registered substances.

Model	Similar substances	Similar substances by previous models	50%–75% confidence	75%–90% confidence	≥90% confidence
CM-combined ¹	1060	- ³	701	149	210
CM < 85	688		466	76	146
CM ≥ 85	372		235	73	64
R-combined ¹	936	- ³	729	98	109
R < 85	522		376	60	86
R ≥ 85	414		353	38	23
PBT/vPvB	53 ²	360 ⁴	38	13	2
ED	109	139 ⁵	86	13	10
Other	129 ²	554 ^{4,6}	32	46	51

Note: The confidence-bins represents the number of substances that are predicted to be structurally similar to an SVHC with a specific confidence in the structural similarity. The previously used similarity models are described by Wassenaar et al.² 1—Combination of two sub-models. 2—For two chemicals the PubChem fingerprint could not be generated (total = 9454). 3—The CM- and R-models were not adjusted. 4—For one chemical the MACCS fingerprint could not be generated (total = 9455). 5—For 82 chemicals the RDKit equivalent FCFP4-fingerprint could not be generated (total = 9374). 6—The previously derived PBT/vPvB model was applied to the ‘Other’-dataset (as the ‘Other’-SVHCs mainly consists of SVHCs previously included in the PBT/vPvB-SVHC dataset).

3.4 | Application to a broader universe of chemicals

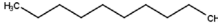
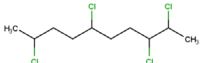
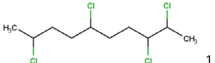
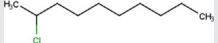
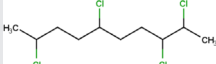
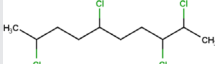
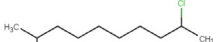
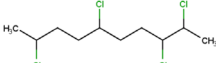
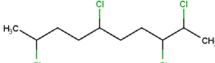
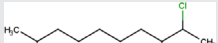
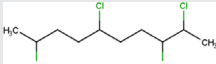
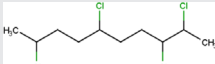
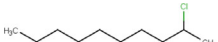
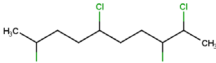
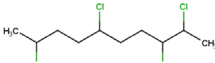
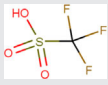
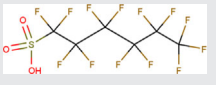
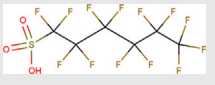
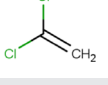
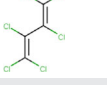
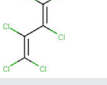

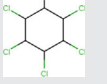

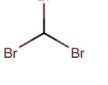


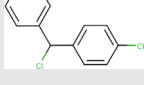
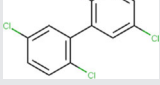
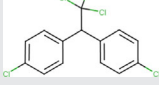
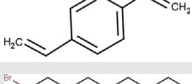
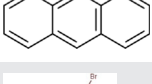
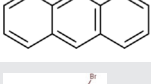
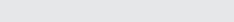

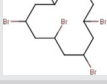
To illustrate the effects of the model adjustments, we applied the newly optimized similarity models to a dataset of REACH registered substances that was used by Wassenaar et al.¹⁰ Under the European chemicals legislation REACH, manufacturers and importers of substances are responsible for the registration of substances produced or imported in the EU above one ton per year via a registration dossier. This list of registered substances is managed by the European Chemicals Agency (<https://echa.europa.eu/information-on-chemicals/registered-substances>) and was previously extracted and prepared for analysis.¹⁰ Here, this dataset was slightly adjusted, by converting the SMILES to QSAR ready SMILES, similarly as performed for the SVHC dataset (see Section 2.1). This resulted in a dataset of in total 9456 REACH registered substances. The results of the screening are shown in Table 4, in which also the results of the previous similarity models are included (using the newly updated SVHC-dataset).

Based on the results as shown in Table 4, it can be observed that far more substances are identified as potential CM or R, compared to the other three categories. This difference can likely be explained by a larger diversity in SVHC structures within the CM- and R-categories. As previously shown, these categories have much more ‘single-point-of-knowledge’ structures, compared to PBT/vPvB and ED-SVHCs which can be divided into relatively few groups of chemicals.² Therefore, it cannot simply be concluded that the PBT/vPvB-, ED- and ‘Other’-models are more strict compared to the CM- and R-models. Nevertheless, the combination of P-, B- and T-properties (or vP- and vB-properties) might be a more stringent condition compared to C-, M- or R-properties. The addition of confidence scores to the similarity models, however, allows for a better interpretation of the predicted results (with a higher structural similarity resulting in a higher confidence in predicted results).

Furthermore, Table 4 and Supplemental Material S4 (Figure S3) indicate very comparable distributions in confidence scores across the varying categories. The results for the ‘Other’-SVHC category are the only exception, for which a relative high amount of structures are identified that are structurally very similar to an ‘Other’-SVHC. This is, at least partially, related to the steep increase in confidence scores in relation to the similarity values, which follows from the model's derived bPPV (see Supplemental Material S3). In addition, this might also be related to a coincidental high representation of structurally very similar substances in the screened dataset.

Predictions for the PBT/vPvB- (and ‘Other’-) model have much improved (with a much lower number of incorrect SVHC predictions) and can be better interpreted using the additional confidence scores (see Table 5) as compared to predictions reported in Wassenaar et al.¹⁰ These new models better consider the number of halogenated fragments (examples 1–7, Table 5), the type of halogenated fragments (example 8, Table 5), and the backbone/aromatic structures (examples 9–11, Table 5). These are important aspects that define the PBT/vPvB-properties of chemicals. Similar improvements are observed for the ‘Other’-model, see Supplemental Material S5 (Table S4). The added value of the confidence scores is particularly evident from examples 1–5 in Table 5, where an increased confidence in structural similarity is observed with an increase in halogenated fragments. In addition, these confidence scores are very useful when interpreting the similarity amongst groups of structurally similar substances, as illustrated in Supplemental Material S5 (Tables S5 and S6) where we present the confidence scores for previously discussed case-studies.¹⁰ Despite the many improvements, also a deficiency of the models can be observed which particularly relates to the use of the CDK Extended fingerprint (see example 12, Table 5). As this fingerprint considers a path-based fingerprint, not many additional fragments are identified for substances with a straight-chain of (carbon) atoms or when these atoms are structured in a ring.

TABLE 5 Specific examples of predictions by the PBT/vPvB-model, including confidence scores in structural similarity.

ID	Substance with 'unknown' properties	Previous most similar known SVHC	Previous model prediction	New most similar known SVHC	New model prediction	New model confidence in structural similarity
1			SVHC		Non-SVHC	43%
2			SVHC		SVHC	75%
3			SVHC		SVHC	83%
4			SVHC		SVHC	93%
5			SVHC		SVHC	96%
6			SVHC		Non-SVHC	23%
7			SVHC		Non-SVHC	7%
8			SVHC		Non-SVHC	1%
9			SVHC		Non-SVHC	0%
10			SVHC		SVHC	75%
11			SVHC		Non-SVHC	2%
12			Non-SVHC		SVHC	81%

Note: The examples illustrate the model's improved consideration of the number of halogenated fragments (examples 1–7), type of halogenated fragments (example 8), and backbone/aromatic structures (examples 9–11). A deficiency of the new model is illustrated with example 12. Similar improvements are observed for the 'other'-model, see Supplemental Material S5. 1—This SVHC considers the third most similar SVHC, with a comparable similarity value for the two other most similar SVHCs (i.e., all with 43% confidence in structural similarity). This structure is included in this table as illustrative example in relation to examples 2–5.

3.5 | ZZS similarity tool

The updated dataset and re-optimized similarity models were incorporated in the online ZZS similarity tool (<https://rvszoekstysteem.rivm.nl/ZzsSimilarityTool>). Also the user-interface was improved by adding the possibility to use a CAS-input as well as a batch-job possibility, besides the already existing SMILES-input option (see Figure 3). The CAS-search feature was included by adding a list of >700.000 CAS-

SMILES combinations, originating from the US-EPA.¹⁷ We refined this list by removing entries without a CAS-number or a QSAR ready SMILES, and removed any chirality description within the SMILES (as chirality has not been used in the similarity model optimization). Furthermore, we ensured that SMILES from substances in the final updated SVHC dataset (for which a CAS-number is available; see Section 2.1) were consistent or included. Some more details on the implementation of the similarity models and use of the ZZS similarity

ZZS similarity tool

Search hide

Enter the SMILES or the CAS-number of a substance to find structural similar ZZS substances. i

SMILES

CAS-number

Calculate

Batch search hide

You can enter multiple SMILES and/or CAS-numbers here. The results will be saved as a tab-separated-values .txt file. i

Batch Search

Download File

Model description show

FIGURE 3 The ZZS similarity tool main web-page with the input modes: Single search and batch search (using SMILES and/or CAS-numbers)

tool are provided in Supplemental Material S6 (Figures S4 and S5 and Table S7).

3.6 | Notes on application and future recommendations

The developed similarity tool has to be considered as a first screening methodology, which can be easily applied to identify and prioritize potential SVHCs. Although comparable approaches to the ZZS similarity tool have been developed before,^{25,26} these earlier methods are not available for use, are only marginally described, and/or do not provide an optimized and validated methodology (resulting in an unknown predictive performance). In contrast to these approaches, these aspects are specifically covered by the ZZS similarity tool. The user of the similarity tool should be aware of the fact that SVHC substances are identified as such based on a regulatory decision process, and that this status has particular regulatory consequences for

industrial chemicals. Accordingly, potential SVHC predictions are most valuable for industrial chemicals. The tool can also be applied to any other type of chemicals, but the results should always be considered as screening results and should be interpreted as follows:

1. Positive (i.e., SVHC) predictions indicate that a chemical is sufficiently structural similar to an existing SVHC to be marked as a potential SVHC. Such a prediction should not be interpreted as a conclusive outcome due to the potential presence of so-called ‘activity cliffs’ (i.e., two very similar chemicals which have an unexpectedly high difference in activity/toxicity),²⁷ and therefore require follow-up analyses. The tool helps to guide such a follow-up, as the specific concern of the most similar SVHC(s) provides a relevant direction for further evaluations.
2. Negative (i.e., non-SVHC) predictions indicate the absence of sufficient structural similarity to any of the SVHCs. Accordingly, related regulatory consequences may—at the moment—not be applicable for the new chemical. It should be noted, however, that a negative

prediction cannot guarantee absence of any toxicological concern. A chemical could for instance exert different non-SVHC types of effects or could exert specific SVHC effects via different mechanisms than the currently known SVHCs (which are mainly industrial chemicals). In addition, the structural overlap might just be too low according to the models to trigger a warning. To assist expert users, the tool always shows the most similar SVHCs (even when the highest structural similarity to an SVHC is below the model's threshold value), so that the structural similarity could always be evaluated manually as well.

The ZZS similarity tool can be applied to all organic chemicals, as the chemical similarity itself can be considered an applicability domain descriptor. If a chemical is sufficiently structurally similar to an existing SVHC, the chemical is clearly within the applicability domain of the model. The similarity models can be used for non-dissociating inorganic or metal-organic chemicals (e.g., organotin substances), to generate a first prediction, but results should be interpreted with care. Furthermore, inorganic chemicals with arsenic, beryllium, cadmium, chromium, lead, mercury, nickel and cobalt-metal derivatives will by definition be predicted as SVHC substances. For these chemicals, the metal atoms (or ions) are thought to be the cause of toxicological concern, irrespective of the (organic) groups or counter-ions present in the inorganic molecule. Additionally, we advise to apply the ZZS similarity tool to parent chemicals as well as breakdown products/metabolites when such information is available, as transformation products may give different similarity outcomes than the parent chemical.

Furthermore, it should be noted that the ED similarity model is limited by the number (and variation) of substances that are classified as ED-SVHC, and for instance does currently not include substances with a steroid backbone, that are very likely to be endocrine active. Accordingly, the user should be reminded that the model only identifies structural similarity to known ED-SVHCs, and that absence of similarity should thus not be interpreted as an absence of possible ED-effects. It is recommended to further develop the ED model when more substances are classified as ED-SVHC, or by including known endocrine disrupting substances like natural substrates.

Future improvements of the models could focus on the evaluation of more sophisticated fingerprints that possibly better define the structural aspects of chemicals, like count-based fingerprints or 3D-based fingerprints. Particularly, 3D-based fingerprints could be relevant as they present a group of important descriptors for determining binding affinity as well as several other properties.^{28,29} However, their use also contains challenges and uncertainties related to chemical conformations and alignments,^{8,30} and more advanced fingerprints do not necessarily outperform the predictive performance of 2D-binary fingerprints.³¹ Accordingly, the currently evaluated methodology, which shows very reasonable performance, is considered adequate, especially for the proposed screening activities. This is also confirmed by an earlier comparison of the performance of the similarity tool with the performance of existing screening methodologies, including Toxtree (i.e., Benigni/Bossa rulebase for mutagenicity and carcinogenicity), DART and the PB-score tool.³²⁻³⁴ The results indicated a higher performance for the similarity

models and indicate the added value and relevance of structural similarity for identifying potential SVHC substances.² Nevertheless, for extensive screening exercises overall screening performance might be improved by combining the results of multiple screening models. By combining models that are based on various types of information (e.g., also structural features and physicochemical properties), more reliable and consistent predictions could potentially be obtained.

4 | CONCLUSIONS

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

ACKNOWLEDGMENTS

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

DATA AVAILABILITY STATEMENT

The data that supports the findings of this study are available in the supplementary material of this article, and additional data are available from the corresponding author upon reasonable request.

ORCID

Pim N. H. Wassenaar  <https://orcid.org/0000-0001-8155-861X>

REFERENCES

- [1] European Commission, Chemicals Strategy for Sustainability—Towards a Toxic-Free Environment 2020.
- [2] P. N. H. Wassenaar, E. Rorije, N. M. H. Janssen, W. J. G. M. Peijnenburg, M. G. Vijver, *Comput. Toxicol.* **2019**, *12*, 100110. <https://doi.org/10.1016/j.comtox.2019.100110>
- [3] 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, *Regul. Toxicol. Pharmacol.* **2019**, *101*, 121. <https://doi.org/10.1016/j.yrtph.2018.11.002>
- [4] A. Cereto-Massagué, M. J. Ojeda, C. Valls, M. Mulero, S. Garcia-Vallvé, G. Pujadas, *Methods* **2015**, *71*, 58. <https://doi.org/10.1016/j.ymeth.2014.08.005>
- [5] T. G. Kristensen, J. Nielsen, C. N. S. Pedersen, *Comput. Struct. Biotechnol. J.* **2013**, *5*, e201302009. <https://doi.org/10.5936/csbj.201302009>

- [6] Y. Yang, J. Zhan, Y. Zhou, *J. Comput. Chem.* **2016**, *37*, 1734. <https://doi.org/10.1002/jcc.24380>
- [7] M. A. Johnson, G. M. Maggiora, *Concepts and Applications of Molecular Similarity*, Wiley, New York **1990**.
- [8] P. Willett, *Mol. Inf.* **2014**, *33*, 403. <https://doi.org/10.1002/minf.201400024>
- [9] R. Todeschini, V. Consonni, H. Xiang, J. Holliday, M. Buscema, P. Willett, *J. Chem. Inf. Model.* **2012**, *52*, 2884. <https://doi.org/10.1021/ci300261r>
- [10] P. N. H. Wassenaar, E. Rorije, M. G. Vijver, W. J. G. M. Peijnenburg, *Regul. Toxicol. Pharmacol.* **2021**, *119*, 104834. <https://doi.org/10.1016/j.yrtph.2020.104834>
- [11] C. E. Smit, P. N. H. Wassenaar, L. de Boer, N. M. H. Janssen, Research into substances of potential concern in Dutch surface water [In Dutch], <https://www.h2owaternetwerk.nl/vakartikelen/onderzoek-naar-mogelijk-zorgwekkende-stoffen-in-nederlands-oppervlaktewater> (accessed December 1, 2021).
- [12] J. Hartmann, E. M. J. Verbruggen, E. Rorije, M. van der Aa, P. N. H. Wassenaar, A. Bannink, Screening and prioritising PMT substances: development of a robust T-score. https://www.umweltbundesamt.de/sites/default/files/medien/362/dokumente/day_2_afternoon_02_julia_hartmann.pdf (accessed August 12, 2021).
- [13] C. W. Yap, *J. Comput. Chem.* **2011**, *32*, 1466. <https://doi.org/10.1002/jcc.21707>
- [14] RIVM—National Institute for Public Health and the Environment, List of Dutch Substances of Very High Concern [in Dutch]. <https://rvsoeksysteem.rivm.nl/ZZSlijst/Index> (accessed January 25, 2021).
- [15] OSPAR, chemicals for priority action. <https://www.ospar.org/work-areas/hasec/hazardous-substances/priority-action>. (accessed December 1, 2021).
- [16] K. Mansouri, C. Grulke, R. Judson, A. Richard, A. Williams, N. Kleinstreuer, *Open-source QSAR-ready chemical structure standardization workflow*, **2021**. <https://doi.org/10.23645/epacomptox.15070041.v1>
- [17] US EPA—United States Environmental Protection Agency, Chemistry Dashboard Data: DSSTox QSAR Ready File.
- [18] K. Mansouri, Standardization workflow for QSAR-ready chemical structures pretreatment. <https://github.com/kmansouri/QSAR-ready> (accessed March 10, 2021).
- [19] R Core Team, R: A Language and Environment for Statistical Computing **2021**.
- [20] M. Kuhn, R package, caret: Classification and Regression Training. Version 6.0-90, CRAN **2021**.
- [21] Y. Cao, A. Charisi, L. C. Cheng, T. Jiang, T. Girke, *Bioinformatics* **2008**, *24*, 1733 <https://doi.org/10.1093/bioinformatics/btn307>
- [22] J. Tuszynski, caTools: Tools: Moving Window Statistics, GIF, Base64, ROC AUC, etc. R package version 1.18.2, CRAN **2021**.
- [23] T. Sing, O. Sander, N. Beerenwinkel, T. Lengauer, *Bioinformatics* **2005**, *21*, 3940. <https://doi.org/10.1093/bioinformatics/bti623>
- [24] G. Landrum, RDKit: Open-source Cheminformatics and machine-learning.
- [25] ChemSec, Methodology for grouping the SIN List and development of the SINilarity tool. <https://chemsec.org/publication/sin-list/methodology-for-grouping-the-sin-list-and-development-of-the-sinilarity-tool/>.
- [26] ECHA—European Chemicals Agency, Screening Definition Document—Methodology for identifying (groups of) potential substances of concern for (further) regulatory action **2019**.
- [27] M. Cruz-Monteagudo, J. L. Medina-Franco, Y. Pérez-Castillo, O. Nicolotti, M. N. D. S. Cordeiro, F. Borges, *Drug Discov. Today* **2014**. <https://doi.org/10.1016/j.drudis.2014.02.003>
- [28] A. Nicholls, G. B. McGaughey, R. P. Sheridan, A. C. Good, G. Warren, M. Mathieu, S. W. Muchmore, S. P. Brown, J. A. Grant, J. A. Haigh, N. Nevins, A. N. Jain, B. Kelley, *J. Med. Chem.* **2010**, *53*, 3862.
- [29] P. W. Finn, G. M. Morris, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2013**, *3*, 226. <https://doi.org/10.1002/wcms.1128>
- [30] M. P. Seddon, D. A. Cosgrove, M. J. Packer, V. J. Gillet, *J. Chem. Inf. Model.* **2019**, *59*, 98. <https://doi.org/10.1021/acs.jcim.8b00676>
- [31] M. D. M. AbdulHameed, R. Liu, P. Schyman, D. Sachs, Z. Xu, V. Desai, A. Wallqvist, *Comput. Toxicol.* **2021**, *18*, 100162. <https://doi.org/10.1016/j.comtox.2021.100162>
- [32] R. Benigni, C. Bossa, N. Jeliakova, T. Netzeva, A. Worth, The Benigni/Bossa rulebase for mutagenicity and carcinogenicity—a module of Toxtree. EUR 23241 EN-2008, **2008**.
- [33] S. Wu, J. Fisher, J. Naciff, M. Laufersweiler, C. Lester, G. Daston, K. Blackburn, *Chem. Res. Toxicol.* **2013**, *26*, 1840. <https://doi.org/10.1021/tx400226u>
- [34] 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, **2011**.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: P. N. H. Wassenaar, E. Rorije, M. G. Vijver, W. J. G. M. Peijnenburg, *J. Comput. Chem.* **2022**, *43*(15), 1042. <https://doi.org/10.1002/jcc.26859>