Journal of Water & Health



Journal of Water and Health Vol 00 No 0, 1 doi: 10.2166/wh.2024.127

# Reducing sample size by clustering: A way to make risk assessment feasible for large groups of organic compounds?

Renske P. J. Hoondert [a], B. A. Wols [a,b] and Patrick Steven Bäuerlein [a,*]

[a] KWR Water Research, Groningenhaven 7, Nieuwegein 3433 PE, The Netherlands
[b] Wetsus, Oostergoweg 9, Leeuwarden 8911 MA, The Netherlands
*Corresponding author. E-mail: patrick.bauerlein@kwrwater.nl

RPJH, 0000-0001-6990-9851; BAW, 0000-0002-9264-3673; PSB, 0000-0002-1110-5997

## ABSTRACT

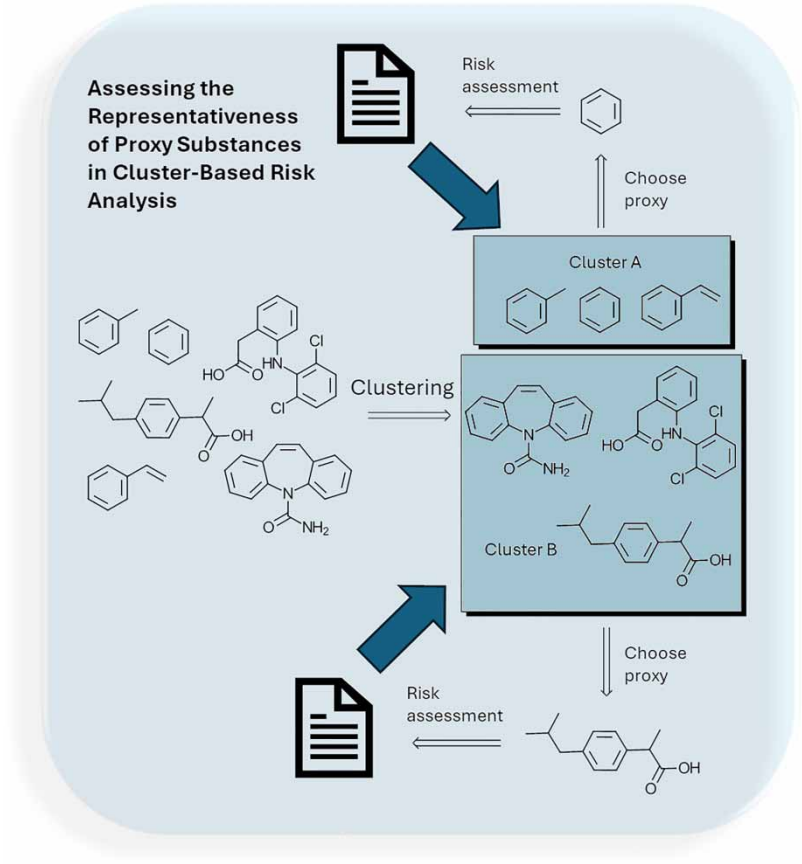
This research addresses the presence of substances of very high concern (SVHCs) confronting the drinking water sector. Responding adequately to the potential hazards by SVHCs, knowledge of emission pathways, toxicity, presence in drinking water sources, and removability during water production is crucial. As this information cannot be received for each compound individually, we employed a detailed clustering approach based on chemical properties and structures of SVHCs from lists with over 1,000 compounds. Through this process, 915 substances were divided into 51 clusters. We tested this clustering in risk assessment. To assess the risks, we developed toxicity prediction models utilizing random forests and multiple linear regression. These models were applied to make toxicity predictions for the list of compounds. This study shows that clustering is a viable approach to reducing sample size. In addition, the toxicity models provide insights into the potential human health risks. This research contributes to more informed decision-making and improved risk assessment in the drinking water sector, aiding in the protection of human health and the environment. This principle is generally applicable. If in a group a suitable representative is found, data from experiments with this compound can be used to gauge the behaviour of chemicals in this group.

**Key words:** clustering, drinking water, prediction, SVHC, toxicity

## HIGHLIGHTS

- More than 1000 mostly substance of very high concern (SVHC) substances were grouped into 51 clusters.
- Toxicity prediction with models based on functional groups was low for the random forest analysis.
- Linear regression models were better suited for toxicity predictions.
- Predictability of the toxicity of substances based on structural properties is still insufficient.
- Based on most of the models, unknown substances cannot be classified as SVHCs.

## GRAPHICAL ABSTRACT




## INTRODUCTION

Several international treaties and legal frameworks, including the REACH regulation (Schulte *et al.* 2012), the OSPAR convention (OSPAR 2017), the Water Framework Directive (European Union 2000), and the persistent organic pollutants regulation (European Union 2019), set rules for substances that pose risks to humans and the environment. These frameworks have different lists of substances whose use and/or emissions must be reduced. Companies are required by governments to prevent the discharge and emission of substances of very high concern (SVHCs) into the environment due to their hazardous nature. SVHCs meet the criteria set out in Article 57 of REACH, as they are potentially carcinogenic, reprotoxic (endocrine disruptors), or accumulate in the food chain (very persistent and very bioaccumulative (vPvB), persistent, bioaccumulative, and toxic). The industry faces new challenges with the emergence of SVHCs such as Per- and polyfluoroalkyl substances (PFAS) (Hale *et al.* 2020), waste pharmaceuticals, and pharmaceutical residues (Davey *et al.* 2022). The EU 'Directive 2000/60/EC' requires Member States to identify water bodies used for the abstraction of water intended for human consumption, to monitor them, and to take the necessary measures to avoid deterioration in their water quality to reduce the level of purification treatment required in the production of water that is fit for human consumption (Directive (EU) 2020/2184). Several approaches to meet these requirements include source control to prevent SVHCs from entering the environment, continuous improvement every 5 years to assess further reduction, and substitution with less hazardous substances or processes.

Given the extensive list of more than 1,600 compounds classified as SVHCs in the Netherlands, and additional candidates (potential SVHCs), it is time consuming to test new abatement techniques or to perform toxicological assessments for each compound individually. As for many chemicals (including SVHCs), toxicological information is lacking, *in silico* techniques (e.g., quantitative structure-activity relationships (QSARs) and read-across) may be used to predict the toxicity of chemicals and to categorize them as (potentially) toxic. The advantage of *in silico* approaches is that they can help in assessing health

risks associated with SVCHs without requiring extensive experiments. QSARs can be used to correlate the chemical structure of new, unknown substances with specific SVHC properties related to toxicity (Bettis 2019). These models are based on statistical linear regression models for molecular descriptors, including structural alerts and functional groups/structural fragments, which have shown to be important in determining the toxicity of these substances. Another method of dealing with these large numbers of substances is clustering. For example, clustering of chemicals has already been used to develop a risk-based monitoring strategy (Sjerps *et al.* 2021). In addition, structural similarity models have been developed to identify potential SVHCs based on their similarity to known SVHCs (Wassenaar *et al.* 2021; Wassenaar *et al.* 2022).

In this study, we present an approach to streamline toxicological testing. By creating clusters of chemically similar substances, a smaller representative group can be tested and the results extrapolated to the remaining chemicals in the same cluster. This approach aims to include as many chemicals as possible in each cluster without causing too much variation in terms of structure within the cluster, while keeping the number of clusters manageable. Subsequently, QSARs were developed to predict biological activity in *in vitro* assays, which were evaluated separately for each cluster and assay. This methodology aims to reduce the need for extensive testing of every chemical, thereby improving efficiency and cost-effectiveness in assessing potential environmental and health effects.

## METHODS

### Selection of chemicals of concern

In this study, the selected compounds are of potential relevance to Dutch drinking water production, as they are either identified on government lists or recognized as compounds of high concern in the drinking water sector, including some non-SVHC compounds (see Table 1 for the lists of compounds that were used and the final set of compounds after clustering is given in the SI). Solid matter such as microplastics, nanoplastics, or nanoparticles, although present in the environment (Bäuerlein *et al.* 2017; Bäuerlein *et al.* 2022; Bäuerlein *et al.* 2023), as well as inorganic compounds, are not included and were not considered for this study.

In the initial phase, compounds were cross-referenced with a chemical database PubChem (Kim *et al.* 2022) to obtain their chemical fingerprints and descriptors. Compounds not listed in the database were omitted. The search process involved finding each compound in PubChem using either the substance name or CAS number, from which the isomeric smile codes were extracted, followed by cross-referencing to eliminate duplicates. Table 1 presents these compounds, detailing the count from each original list, resulting in 1,135 compounds. Subsequent steps included the removal of very hydrophobic compounds (logp >8, logp was obtained from PubChem), resulting in 1,079 compounds. Boiling point and solubility predictions from the EPI suite (USEPA 2012) were calculated for these compounds, narrowing the focus to organic compounds due to their relevance in drinking water production. Compounds without boiling point or solubility information were removed beforehand. After these curation steps, the final list on which clustering was performed contained 931 organic compounds.

### Cluster analysis

Cluster analysis is performed using information from chemical fingerprints and descriptors. A chemical fingerprint is a coded representation of a given molecule where each bit in the code indicates whether a particular substructure is present in the

**Table 1** | Number of compounds on each list and after each step

| | SVHC (Rivm 2024) | pSVHC (Rivm 2024) | PMT (Holmberg *et al.* 2018) | Gd-MRI x-ray (Kools *et al.* 2013; Rijn 2023) | HPLC_UV | Water company PMT |
|---|---|---|---|---|---|---|
| Compounds | 1,566 | 261 | 157 | 9 | 153 | 121 |
| Compounds in PubChem | 795 | 217 | 156 | 9 | 149 | 121 |
| Compounds after removal of duplicates | 741 | 216 | 156 | 9 | 149 | 121 |
| Compounds after removal of inorganic compounds | 611 | 200 | 156 | 9 | 149 | 120 |
| Cumulative | 611 | 811 | 932 | 941 | 1,074 | 1,135 |

SVHC, substances of very high concern; pSVHC, potential substances of very high concern; PMT, persistent, mobile, toxic; Gd-MRI, list of Gadolinium containing compounds; HPLC_UV, list of compounds detectable with HPLC-UV, used in the Netherlands.

molecule. The following fingerprint types have been considered (the number of bits indicates the number of substructures on which the fingerprints were based):

- CACTVS fingerprint from PubChem (881 bits, retrieved using PubChemPy 1.0.4 (Kim *et al.* 2018))
- MACCS (Molecular ACCess System) fingerprint (167 bits, using rdkit 2022.9.1 (RDKit))
- Morgan fingerprint (2048 bits, using rdkit 2022.9.1 (RDKit)), this was ultimately not included because it was not distinctive enough.

Descriptors are numerical values of a particular property of the molecule that can be easily determined based on the structural formula using chemical informatics techniques. The following descriptors have been considered in the cluster analysis:

- From PubChem: charge, tpsa, rotatable_bond_count, molecular_weight, heavy_atom_count, h_bond_acceptor_count, h_bond_donor_count (retrieved using PubChemPy 1.0.4 (Kim *et al.* 2018))
- Via rdkit (2022.9.1, (RDKit)): xlogp. Via rdkit, many more descriptors can be calculated, but these are not known for all substances, and therefore – in that case – fewer substances can be included.

This resulted in 1,056 descriptors and fingerprints. Missing descriptor data were removed in several steps by removing descriptors with little data, descriptors without variation in the data, or substances with little data, in such a way that as little data as possible is 'lost'. This resulted in 825 descriptors and fingerprints that were used for the cluster analysis.

The clustering was done via K-means clustering (using the Python module scikit-learn 1.1.3, (Pedregosa *et al.* 2011)) based on the aforementioned fingerprint types and descriptors. The clustering algorithm was applied for different numbers of clusters, varying from 2 to 150 clusters, to determine the optimal number of clusters for the selected 931 compounds.

The clustering was evaluated to determine how well the substances fit within a cluster and how many clusters were needed to represent the lists. Two parameters were used to evaluate:

- Silhouette score. The silhouette score is a measure of how well a substance fits within a cluster, expressed as a value between $-1$ and 1. Greater than 0 means that the substance fits better within its own cluster than within other clusters, while for silhouette scores below 0, the opposite is true.
- Similarity index. Here, the MACCS fingerprints of two molecules are compared, and a similarity index is calculated between 0 and 1 (0 unalike and 1 alike). The Tanimoto similarity index is used (Bajusz *et al.* 2015). In a cluster, the average similarity of each substance was calculated concerning the other substances in the cluster. A similar approach has been tested previously (Wassenaar *et al.* 2022).

Once clustering was completed, the optimal number of clusters was determined through evaluation. Substances were removed if it was expected that these will not occur in the water sources. These substances include very volatile substances (a boiling point $< 20$ °C) and very poorly soluble substances (with a solubility $<0.1$ mg/L). In addition, clusters with only one or two substances have been removed to reduce the number of substances in the target substance analysis method. For the removed substances that were not expected to be relevant for water sources, the suitability of their inclusion in another cluster was also investigated.

## Toxicity prediction

### Data collection

To gain insight into the toxicity of (SVHC) substances, the complete ToxCast dataset was downloaded, consisting of 21 individual databases containing more than 3.5 million experimental toxicity data records (U.S. EPA 2012). Only dose–response series with an active *hit-call* (related to the reliability of the series) are included in the analyses, resulting in 357,000 data records. ToxCast calculates the active concentration at which 50% of the effect is observed ($AC_{50}$ in μM) using experimental dose–response series for a wide range of *in vitro* assays. Log $AC_{50}$s for the best predictive model are automatically calculated and are converted to milligrams per litre based on the substance's molecular weight. The resulting dataset covers 1,388 individual unique *in vitro* assay endpoints from 20 sources/laboratories, with 235,207 data lines for 7,566 unique substances (based on CAS number). However, these endpoints are not specifically related to SVHC properties and cover a wide range of toxicological mechanisms. The compound list used in the present study, based on chemicals after clustering and post-processing contains 915 substances (see Results and Discussion section), of which 517 substances (57% of the list, 45,142 data records) are included in the ToxCast database for 1,349 unique *in vitro* assays.

Toxicity data were combined with EPI Suite data on physicochemical descriptors and partition coefficients associated with bioaccumulation and mobility (Log $K_{oc}$ and Log $K_{ow}$). These parameters were chosen to be included as explanatory variables in the model development as they are associated with increased acute toxicity due to their relevance to bioconcentration in membranes (Lambert *et al.* 2022). The entire dataset was then cleaned for analysis based on several criteria: (i) poorly soluble compounds (i.e., with a solubility in μM (WSKOWWIN v1.42) below the corresponding $AC_{50}$) were removed from the dataset as this often affects the actual exposure in a toxicity test (typically leading to an underestimation of its effect) (Groothuis *et al.* 2015); (ii) all *in vitro* assay endpoints for which data for fewer than 50 chemicals were available were excluded from further analysis to ensure unbiased (unbiased) modelling. Applying these two criteria resulted in a toxicity dataset covering 5,101 chemicals (including 444 SVHCs), 603 *in vitro* assay endpoints, encompassing over 139,000 data entries. These data were combined with data on structural fragments (functional organic groups as defined by the US EPA), taken from the OECD QSAR Toolbox. Highly correlated structural fragments ($r > 0.97$) were removed from the dataset to avoid multicollinearity and prediction bias (Næs & Mevik 2001). Where values of two pairwise physicochemical descriptors were highly correlated, the variable with the highest mean absolute correlation with a multitude of other variables was removed.

## Model description

Toxicity predictions for chemicals (including SVHCs) were made using two types of models: random forest analysis and multiple linear regression analysis. Random forest analyses are performed using the *randomForest* package in R (v. 4.3.1), while the linear regression analyses are performed using the *lm()* function in the *stats* package in R (v 4.3.1). Five hundred decision trees in the random forest analyses were chosen based on a trade-off between processing time and prediction error reduction. Although model accuracy tends to increase with an increasing number of decision trees, at some point, the increase in processing time outweighed the increase in model accuracy. The top 10 structural elements explaining the most variance in $AC_{50}$s were reported and used as explanatory variables in the multiple regression analysis. The explanatory power of variables within a random forest analysis is determined by calculating the %IncMSE (increase in mean-squared error of predicted values), which is calculated through:

$$\%\text{IncMSE} = \frac{(\text{MSE}_j - \text{MSE}_0)}{\text{MSE}_0} \times 100\% \tag{1}$$

where MSE is defined as mean-squared error (the average squared difference between the predicted value and the actual value) for decision tree $j$ relative to the baseline (0). Since the %IncMSE in this case shows the extent to which the model performs less when a relevant variable is not included, a higher %IncMSE value means a better predicting variable.

## Model evaluation

After combining the experimental toxicity data with the structural elements, 70% of the data were used as the training dataset and the remaining 30% served as the test dataset. The data were randomly allocated to these two lists, using the R *sample()* function. The models were evaluated using coefficient of determination ($R^2$) and cross-validated redundancy ($Q^2$) as validation metrics, for the model's ability to predict the toxicity of compounds in the test dataset:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

where $R^2$ is calculated as 1 – residual sum of squares (RSS) and the total sum of squares (TSS), $y_i$ is the observed $AC_{50}$ for compound $i$, $\hat{y}_i$ is the predicted $AC_{50}$ for compound $i$, and $\bar{y}$ is the average $AC_{50}$ in the training set. The $R^2$ statistic explains the variance in the response variable that is explained by the explanatory variable(s). Over the years, there has been ample discussion on the $R^2$ threshold above which a model can be considered a good predictive model. In the present study, $R^2$ values of 0.75, 0.50, or 0.25 for response variables will be described as substantial, moderate, or weak, respectively (Hair *et al.* 2013;

Sarstedt *et al.* 2021). The predictive power of the model was assessed by calculating the $Q^2$ for the test dataset:

$$Q^2 = 1 - \frac{\sum_{i=1}^{n} (y_{ext} - \hat{y}_{ext})^2}{\sum_{i=1}^{n} (y_{ext} - \bar{y}_{ext})^2}$$

where the $Q^2$ is calculated as 1 – RSS and the TSS, $y_{ext}$ is the observed $AC_{50}$ for compound $i$ in the test set, $\hat{y}_{ext}$ is the predicted $AC_{50}$ for compound $i$ in the test set, and $\bar{y}_{ext}$ is the average $AC_{50}$ in the test set. The $Q^2$ statistic reflects predictive relevance and measures whether a model has predictive relevance. $Q^2$ values above zero indicate that your values are well reconstructed and that the model has predictive relevance. However, this does not say anything about the quality of the prediction, only that the model predicts better than taking the average of the observed values (Rigdon 2014).

In addition, model accuracy, specificity and sensitivity have been calculated. Because these evaluation criteria are based on nominal response variables (the substance is considered toxic or not toxic), the continuous $AC_{50}$ values per individual *in vitro* toxicity test endpoint have been converted into toxicity classes (low toxicity or high toxicity) based on the distribution of $AC_{50}$ values (per toxicity test), whether the $AC_{50}$ falls in the lowest 25% (high toxicity) or the highest 75% (low toxicity). The sensitivity, specificity, accuracy, negative predictive value (NPV), and positive predictive value (PPV) of the models to predict the correct bioassay response (high, low) were calculated based on these toxicity classes.

## RESULTS AND DISCUSSION

### Cluster analysis

#### Chemical clustering

The 'similarity' scores of each substance in a cluster are shown in Figure 1 for different numbers of clusters. Figure 2 shows the average 'similarity' and 'silhouette' scores as a function of the number of clusters. These scores increase with an increasing number of clusters. For the silhouette score, it also shows the number of clusters containing substances with a score below 0, implying that these substances do not fit well into that cluster. From about 50 clusters onwards, this proportion between the cluster scores and the cluster sizes remains almost constant (with an increasing number of clusters). In addition, the number of clusters including very few substances (fewer than five substances) logically increases with an increasing number of clusters. Based on this comparison of the number of clusters, 70 clusters were chosen as the optimal number, as with a greater number of clusters, the added value then becomes relatively small.

#### Post-processing and restructuring of clusters

Post-processing was done to reduce the number of clusters. Removing very volatile substances and very poorly soluble substances results in a reduction of the number of substances to 726 in 63 clusters. Finally, clusters with one or two substances were not included, reducing the number of substances to 710 in 51 clusters. As several substances and clusters were excluded from the analysis for various reasons, an additional step was undertaken to reincorporate these excluded substances or those within the excluded clusters into the remaining clusters. This integration was achieved by determining the 'similarity' of each substance in an excluded cluster with those in the remaining clusters. Each substance was then assigned to the cluster with which it shared the highest average 'similarity' score, ensuring a comprehensive and inclusive clustering process. Figure 3 illustrates an example where substances were added to various clusters following the described method. After reclustering, 915 compounds in 51 clusters were obtained for the toxicity predictions (a list of compounds is given in the Supporting Information).

### Toxicity predictions

#### Toxicity data

Boxplots showing the $AC_{50}$ (in mg/L) distributions for each cluster are shown in Figure 3 (left plot). These include substances that show clear activity in multiple *in vitro* assay endpoints as well as substances that specifically trigger responses in specific assays and thus may only be toxic when looking at certain assay endpoints (e.g., neural development). Although in the cluster analysis substances with fewer than three substances were removed, combining the data with toxicity data from ToxCast resulted in certain clusters that included fewer than three substances, as toxicity data were not available for all substances

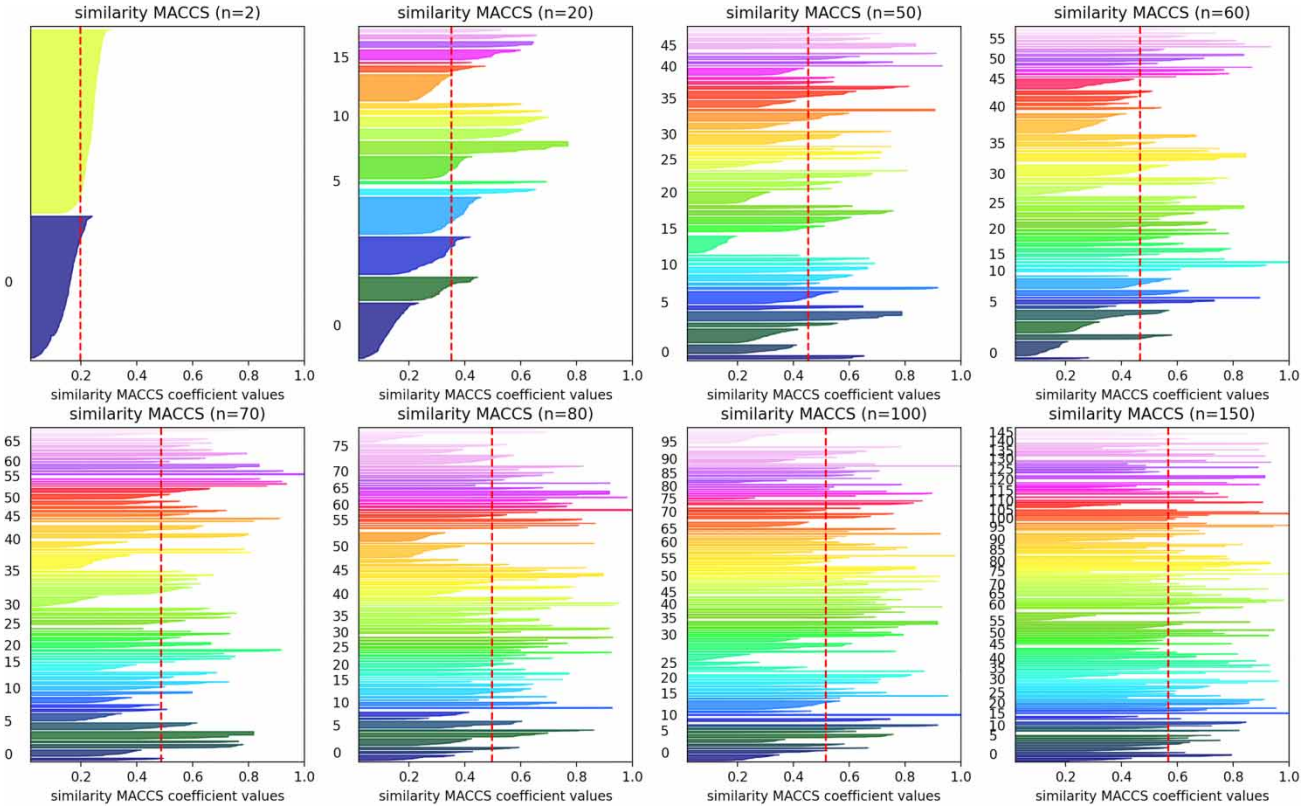



**Figure 1** | Evaluation of the 'similarity' scores of each substance in a cluster for various number of clusters (the subplots represent an increasing number of clusters from left to right). Substances in the same cluster have the same colour.

in a cluster. In total, toxicity data were available for 517 out of 915 substances. For two clusters (cluster 39 and cluster 42), toxicity data were available for fewer than three substances. Substances within cluster 18 had the lowest $AC_{50}$s – and thus the highest toxicity (mean = 2.03 μM – 0.02 mg/L, S.E. = 8.08, for all types of *in vitro* assay endpoints). However, the data within this cluster were not normally distributed, resulting in a median value (12.9 μM – 0.13 mg/L) that was significantly different from the mean value. The highest $AC_{50}$s (mean = 36.1 μM – 51 mg/L, median = 5.8 μM – 8 mg/L, S.E. = 8.64), and thus, the lowest toxicity was observed for compounds in cluster 61.

The $AC_{50}$ values presented earlier vary greatly between assay endpoint types. Moreover, a low number of substances within a cluster may bias the results. For example, cluster 18 consists of only six chemicals (radiocontrast agents), with 1,389 data entries (including different *in vitro* assay endpoints), and cluster 61 consists of seven substances (corrosion inhibitors), covering 293 data entries. Just from the $AC_{50}$ data, it is not possible to tell whether the chemicals within cluster 18 are more toxic than the chemicals within other clusters, as this depends heavily on the distribution of the substances in the subsets and the distribution of the corresponding $AC_{50}$ values. Therefore, for each chemical-$AC_{50}$ combination, the relative position (percentile within the distribution for each *in vitro* assay endpoint to standardize the data) of the $AC_{50}$s were plotted, assuming normally distributed $AC_{50}$ data per assay endpoint. The resulting boxplots based on percentiles (Figure 3, right plot) now indicate how the data within the cluster are compared to the $AC_{50}$s of other chemicals (both from within and outside its cluster). The percentiles of chemicals within the boxplots are generally uniformly distributed – from 0 (most toxic) to 1 (least toxic) – implying that while chemicals from certain clusters may exert the biggest response in one *in vitro* assay, this may not imply the highest toxicity in others. This means that SVHC from different clusters may each have a very specific toxic mechanism of action. However, results from a one-way analysis of variance (ANOVA) with least-significant difference test on the $AC_{50}$ data, including cluster number as an explanatory variable, indicate that some clusters (including cluster 18) tend to be more toxic (have lower $AC_{50}$s) in certain *in vitro* assay endpoints compared to other clusters, while other clusters (including cluster 61) tend to be significantly less toxic in the majority of the *in vitro* assay endpoints ($p < 0.001$). Results from a factorial ANOVA, including both cluster number and assay endpoint as independent (explanatory) variables, reveals that the variation

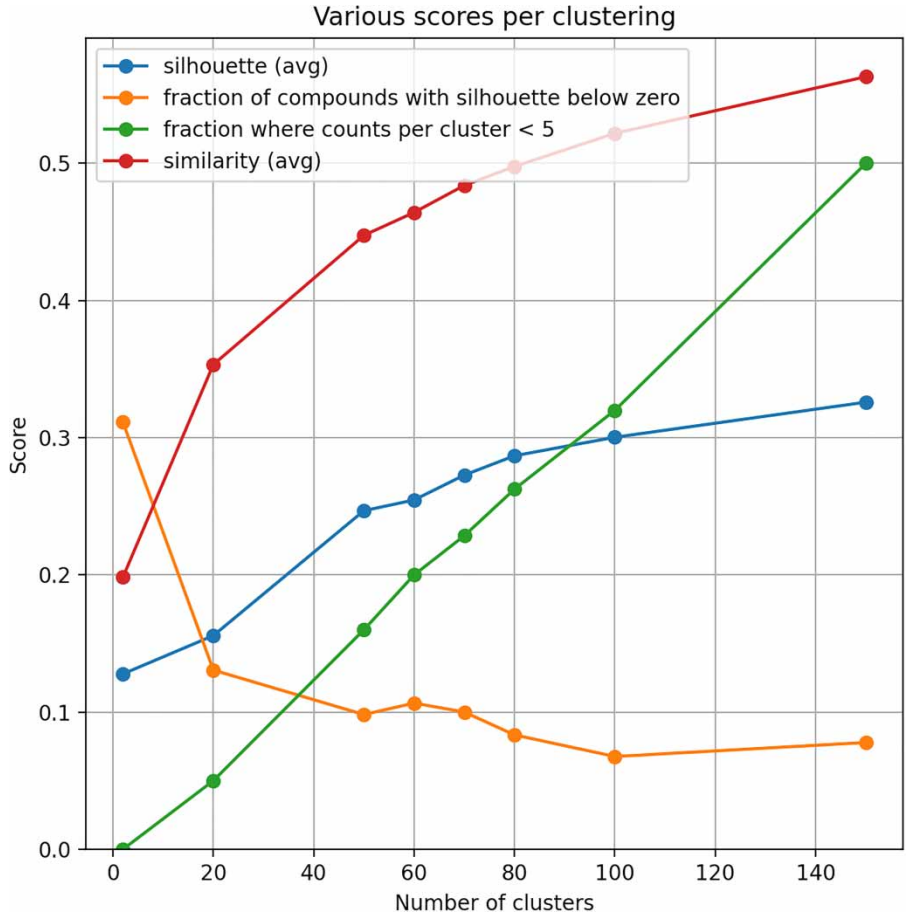



**Figure 2** | Average evaluation score of substances in a cluster at different numbers of clusters.

in $AC_{50}$ data within clusters is significantly fewer than between clusters when data are subset based on *in vitro* assay endpoint ($p < 0.001$). This implies that clustering SVHCs based on MACCS fingerprints leads to clusters of chemicals that likely have a relatively similar toxic mechanism of action.

### Random forest models

In 8.2% of all random forest analyses, the acetoxy group was identified as the most important structural element for predicting the toxicity of chemicals, followed by saturated carbocycles (8%), with the latter being identified as an important structural alert related to mutagenicity (Yang *et al.* 2020). However, only five compounds contain a functional organotin group, while only two and three compounds contain an acetoxy group and a saturated carbocycle group, respectively. In 26.1% of all random forest analyses (475 *in vitro* assay endpoints), organotin was identified as the most important structural element for predicting the toxicity ($AC_{50}$) of chemicals. However, this was based on a small dataset of chemicals that contain an organotin group. This makes these results uncertain and not (yet) useful for the risk assessment of substances for which no data are available.

### Multiple linear regression models

Predicted $AC_{50}$s were plotted against the measured $AC_{50}$s for all substances in the final formatted dataset based on the criteria described in the Methods section (over 5,000 chemicals; Figure 4, top). Models included in the present study are considered highly relevant if predicted values fall within a factor of five of the observed $AC_{50}$s, shown in Table 2. Overall, 74.2% of the variation in the $AC_{50}$ values in the training dataset is explained by the linear regression model ($R^2$). However, when the model is applied to the test dataset, only 27.3% of the variation is explained by the model ($Q^2$). $Q^2$ values above zero indicate that your values are well reconstructed and that the model has predictive relevance. However, this does not say anything about the

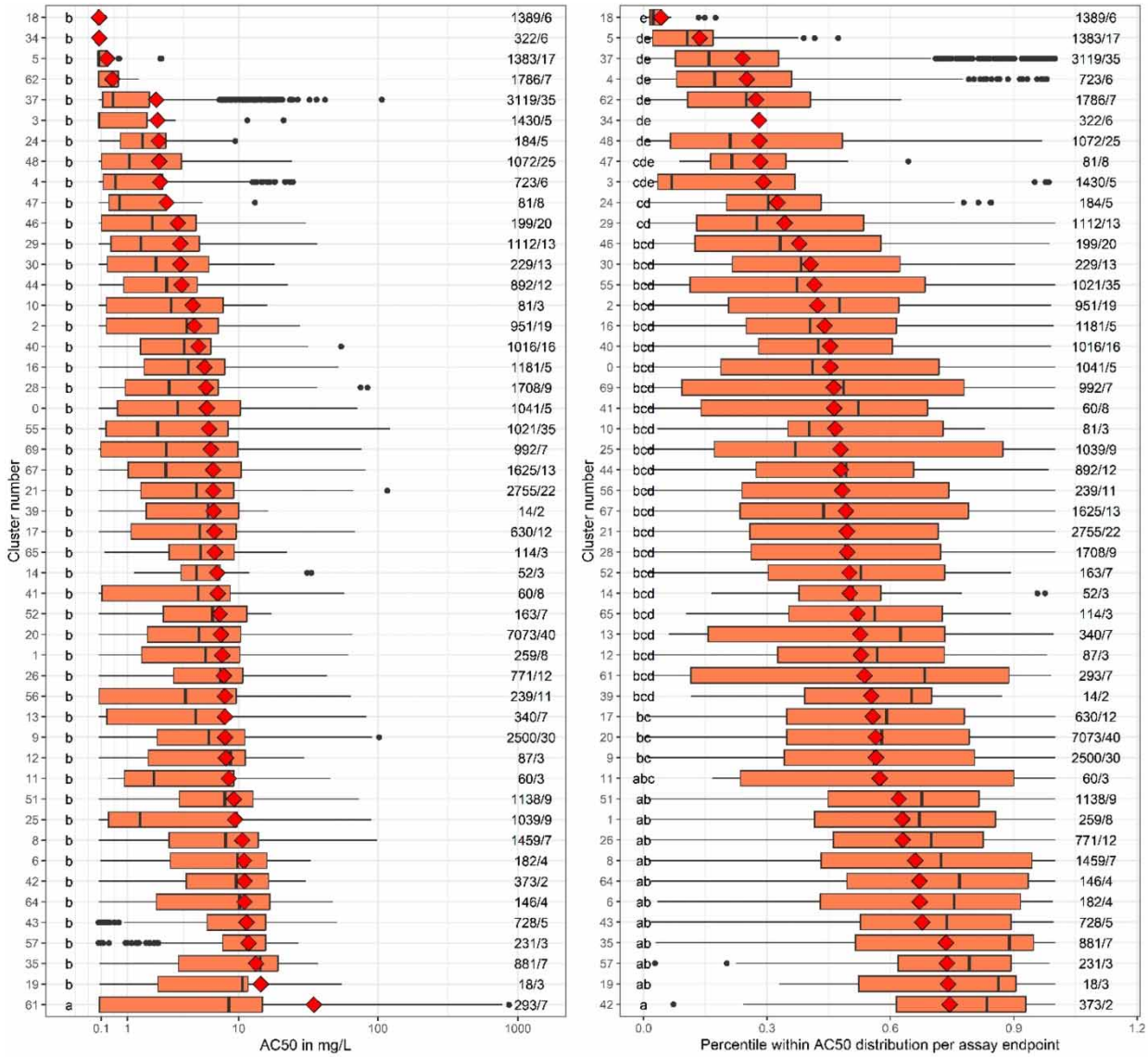



**Figure 3** | Boxplots showing toxicity values ($AC_{50}$; active concentrations at which 50% of the observed effect was observed) for the clusters included in this study (left) and boxplots showing percentiles of substances within the SVHC clusters relative to all $AC_{50}$ values within assay endpoints. Black letters (a,b,c,…) indicate significant differences between clusters (Fisher's Least Significant Difference (LSD)-test, $p < 0.05$). Equal letters indicate no significant difference between clusters. The red diamonds within the boxplots indicate the mean per cluster. The grey numbers on the right indicate the size of the dataset (number of data lines per cluster) and the number of individual chemicals per cluster.

quality of the prediction, only that the model predicts better than taking the average of the observed values (Rigdon 2014). Although a $Q^2$ value of 27.3% may prove the model's predictive relevance, its reliability in risk assessment for unknown substances remains low. However, this $Q^2$ value was based on results from a large variety of *in vitro* assay endpoints. $Q^2$ values, and thus predictive relevance, vary greatly among these assay endpoints, meaning that for some endpoints high predictive power was observed. Figure 4(a) and 4(b) show results from two *in vitro* assay endpoints, one for which the model has a high predictive relevance (**OT_AR_ARSRC1_0480**) and one for which the model has a low predictive relevance (**TOX21_PR_BLA_Antagonist_viability**).

Model performance varied widely between clusters (Figure S1, Supporting Information). Eleven of 44 clusters for which assay endpoint responses could be predicted had higher $R^2$ than the average $R^2$ of 74.2%. However, for four clusters, fewer than 30 data entries were included in the formatted dataset. Twenty-four of 44 clusters had $Q^2$s above 0, indicating

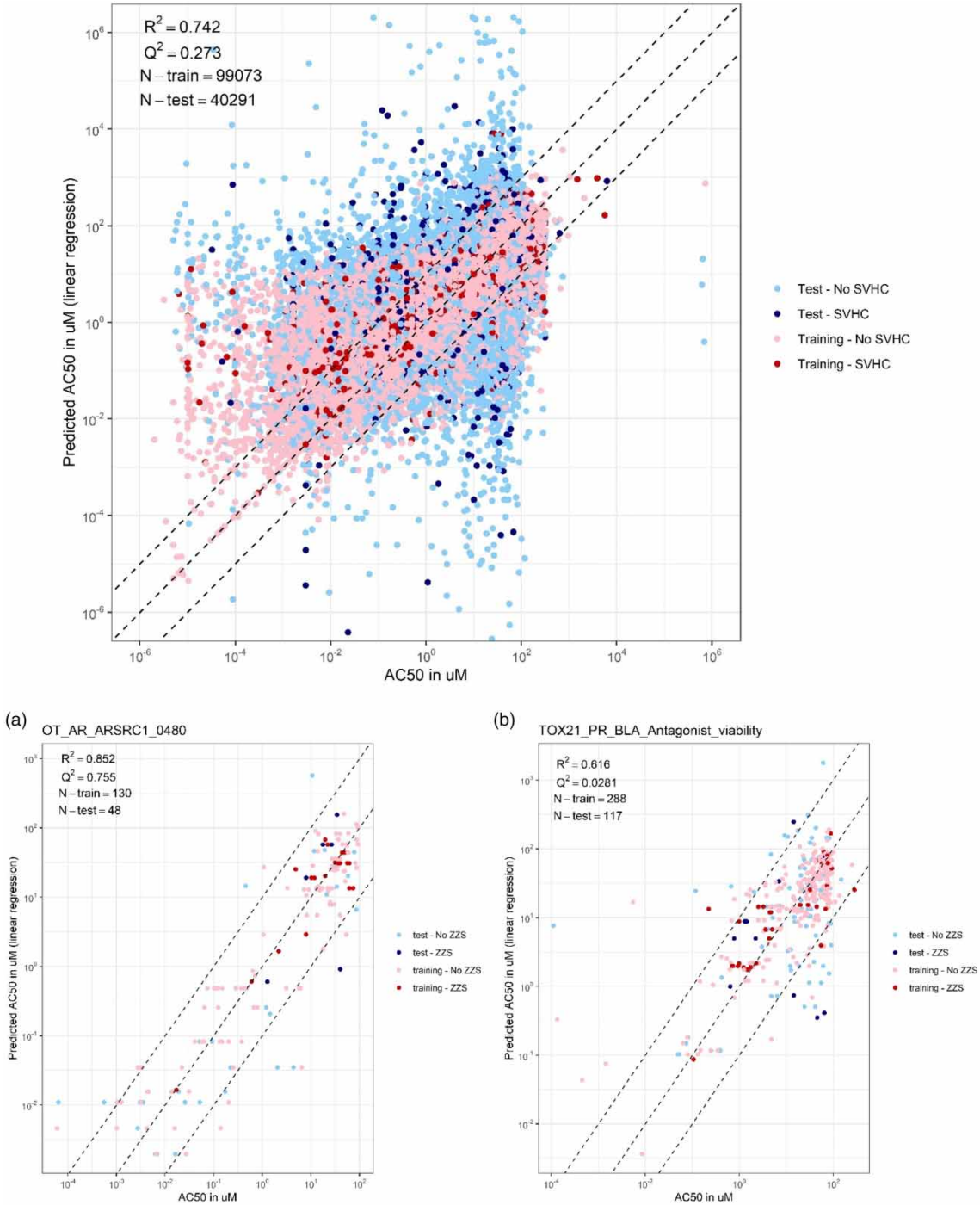



**Figure 4** | Predicted toxicity ($AC_{50}$ in μM) by the multiple linear regression model versus observed toxicity, based on structural elements, for both the training and test dataset, and chemicals classified as SVHC and chemicals not classified as SVHC. ($AC_{50}$; active concentrations at which 50% of the effect was observed.)

that the models have good predictive relevance for chemicals outside the training dataset. However, the predictive power of models trained on data from individual clusters when applying the models on test datasets was not necessarily higher than for the model trained on the complete non-SVHC dataset, probably due to the fact that a very diverse set of *in vitro* assays were included in these predictions.

**Table 2** | Percentages of predicted values falling within a factor of five of the observed values (between), above observed values (overestimated) or below observed values (underestimated)

| | | Between (%) | Overestimated (%) | Underestimated (%) | Overfitted (%) |
|---|---|---|---|---|---|
| All data | Overall | 75.5 | 11.5 | 10.9 | 2.1 |
| | Test | 67.4 | 16.8 | 15.8 | 0.0 |
| | Training | 83.6 | 8.0 | 5.3 | 3.1 |
| SVHCs | Overall-SVHC | 73.5 | 17.7 | 7.5 | 1.3 |
| | Test-SVHC | 70.1 | 21.0 | 8.9 | 0.0 |
| | Training-SVHC | 76.0 | 15.5 | 7.0 | 1.5 |

*Note*: The percentage of overfitted data points are included (overfitted).

### Sensitivity, specificity, and accuracy

In addition to regression analyses, the substances were classified on the basis of their $AC_{50}$ and categorized as 'low' or 'high' toxicity. Predicted values were also labelled based on the toxicity distributions described (see Methods section). The sensitivity of all individual models (the percentage of correctly predicted 'high toxicity' classes – true positives) was 65.6%. The specificity of the models (the percentage of correctly predicted 'low toxicity' classes – true negatives) was 81.4%. This means that overall the models are better at avoiding false negatives than false positives, which in turn implies that the models are conservative. The NPV for all models was 84.1%, which means that 84.1% of all predicted substances with a 'low toxicity' categorization actually were relatively non-toxic. The PPV for all models was 61.1%, implying that 61.1% of all substances for which 'high toxicity' was predicted actually were highly toxic. The overall accuracy of the models was 78.7%, implying that 78.7% of all toxicity categories were correctly predicted. When looking at differences among the clusters, we see that the accuracy, specificity, and sensitivity of the models vary greatly across clusters (Figure 5). However, bias may be introduced by the fact that clusters of SVHCs by definition are toxic. Note that the derivation of the models was not only based on data for SVHCs. Looking at model performance indicators, we see that the toxicity of substances within a number of clusters can be predicted better than the toxicity of substances not labelled as SVHCs. Examples of sufficiently large (>30 data entries) clusters for which the calculated accuracy, sensitivity, and specificity are higher than for non-SVHC data include cluster 69 (including seven chemicals used as UVB blockers) and cluster 48 (including 55 chemicals – fuel oil chemicals).

Although predictions for continuous data ($AC_{50}$s) were relatively unreliable, predictions for toxicity classes are much more accurate. However, there are large differences in the calculated sensitivity specificity and accuracy between *in vitro* assay endpoints and clusters of SVHCs. To designate a substance as a potential SVHC based on modelling with data from the ToxCast database, it is important to select the appropriate *in vitro* assay endpoints that most strongly correlate with the SVHC criteria (including carcinogenicity, mutagenicity, and reproductive toxicity). Within the clusters included in this study, chemicals may occur that differ from other chemicals in the cluster with regard to the biological activity of the substance in various assays.

## CONCLUSIONS

The lists of SVHC and potentially substances of very high concern together with high-concern compounds within the drinking water sector were clustered on the basis of molecular structures in order to end up with a list of SVHC substances covering the broad chemical space of SVHC. Initially, the more than 1,000 SVHC substances were classified into 70 clusters, which were then reduced to 51 clusters based on a number of criteria. The predictability of the toxicity of substances (SVHCs and non-SVHCs) based on functional groups and physicochemical parameters was examined. In general, the predictive power using linear regression models varied considerably among *in vitro* assay endpoints and clusters. There were no major differences between the predictability of SVHC and non-SVHC toxicity in either model, although the differences in predictability were greater between the training and test datasets for non-SVHCs. Large differences in toxicity predictability were observed between clusters of SVHCs which may be due to differences in sample size, the combination of *in vitro* assays, and combinations of functional groups in the dataset. The overall conclusion is that for the majority of *in vitro* test endpoints, current analyses are unlikely to be sufficient to predict the toxicity of substances based on structural properties. Unknown substances cannot be classified as SVHCs on the basis of most of the models, especially not on the basis of models combining a large number of test endpoints. Although the proposed models in the present study do not seem to be sufficiently reliable to

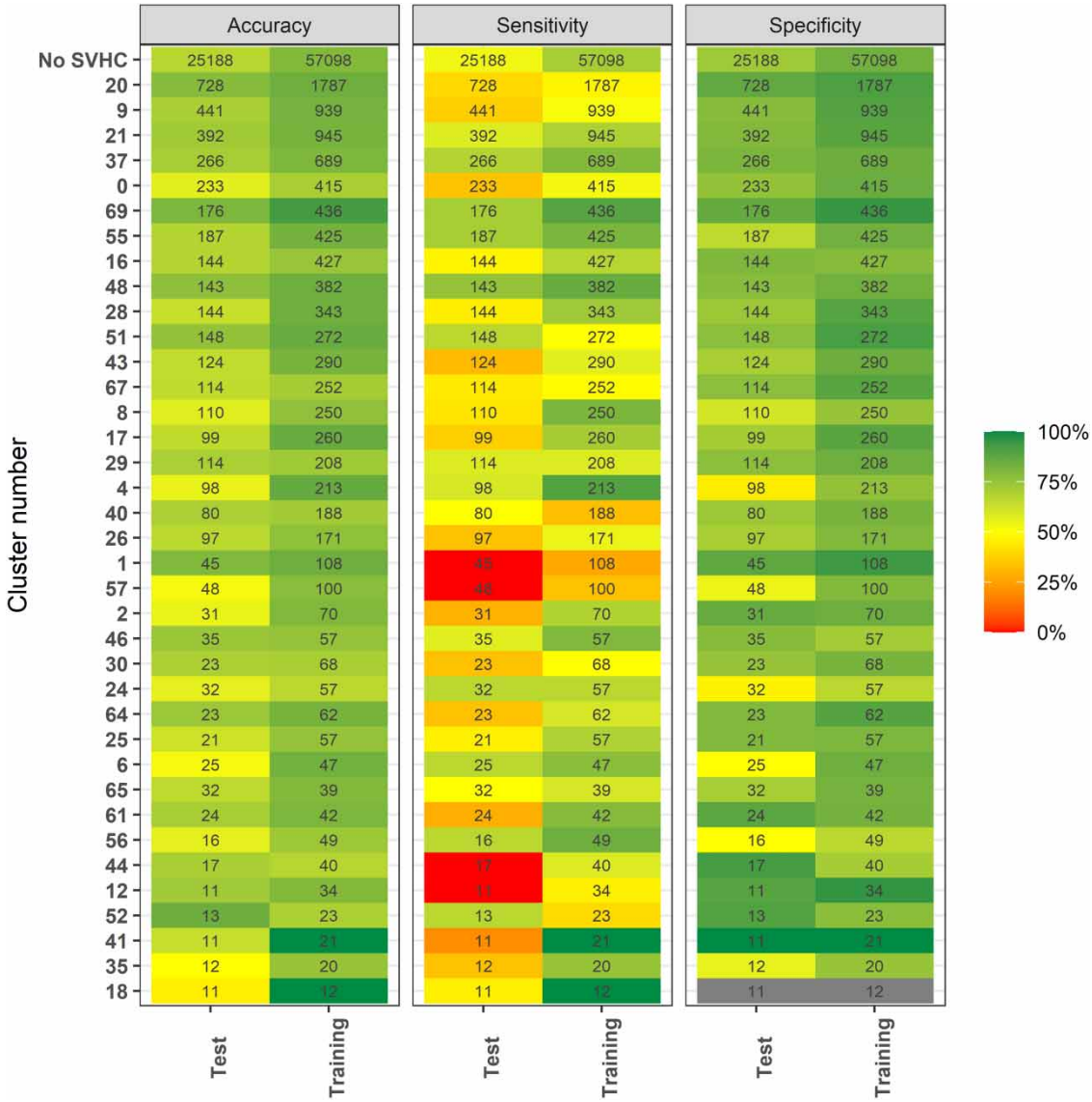



**Figure 5** | Model performance indicators (sensitivity, specificity, and accuracy) for the results by cluster for the test and training dataset (the dataset on which the models are not and are based, respectively). The models are trained per *in vitro* assay endpoint. The black numbers in the tiles indicate the sample size of the subset. The dataset is ordered by dataset size (largest dataset at the top). Clusters with fewer than 10 data entries for the test set were excluded from the figure.

be implemented in the risk assessment of substances, this study showed that clustering of similar SVHC based on MACCS fingerprints leads to clusters with chemicals exerting a similar toxicological mechanism of action. In addition, the toxicity predictions are expected to be reliable enough to make a statement about the toxicity class of a substance, as evidenced by the high calculated sensitivity and specificity of the models in the present study. In general, by clustering chemicals based on MACCS fingerprints, we can significantly reduce the number of samples. This approach allows us to compare manageable groups instead of an overwhelming number of individual chemicals, making the task feasible. Future research will further apply the concept of clustering to the field of drinking water purification. The pivotal question in this context is whether selecting a single representative compound from each cluster could reliably forecast the purification efficiency for all other compounds within the same cluster. The criterion for selecting this representative compound is its detectability, favouring the one that can be identified most easily. Of course, this principle can be applied to other types of experiments. If a suitable representative of a group is found, experiments can be carried out with that compound, and its behaviour can be taken as representative of the behaviour of all chemicals in the same group.

## DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

## CONFLICT OF INTEREST

The authors declare there is no conflict.

## REFERENCES


Bajusz, D., Rácz, A. & Héberger, K. 2015 Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics* **7** (1), 20.

Bäuerlein, P. S., Emke, E., Tromp, P., Hofman, J. A. M. H., Carboni, A., Schooneman, F., De Voogt, P. & Van Wezel, A. P. 2017 Is there evidence for man-made nanoparticles in the Dutch environment? *Science of the Total Environment* **576**, 273–283.

Bäuerlein, P. S., Hofman-Caris, R. C. H. M., Pieke, E. N. & Ter Laak, T. L. 2022 Fate of microplastics in the drinking water production. *Water Research* **221**, 118790.

Bäuerlein, P. S., Pieke, E. N., Oesterholt, F., Ter Laak, T. & Kools, S. A. E. 2023 Microplastic discharge from a wastewater treatment plant: Long term monitoring to compare two analytical techniques, LDIR and optical microscopy while also assessing the removal efficiency of a bubble curtain. *Water Science and Technology* **87**, 39–56.

Bettis, J. L. 2019 An introduction to (Q) SAR with respect to regulatory submissions. In: *Integrated Safety and Risk Assessment for Medical Devices and Combination Products* (In: Gad, S. C., ed.). Springer, Raleigh.

Davey, C. J. E., Kraak, M. H. S., Praetorius, A., Ter Laak, T. L. & Van Wezel, A. P. 2022 Occurrence, hazard, and risk of psychopharmaceuticals and illicit drugs in European surface waters. *Water Research* **222**, 118878.

European Union 2000 Water Framework Directive (2000/60/EC). European Union, Brussels.

European Union 2019 Regulation (EU) 2019/1021 of the European Parliament and of the Council of 20 June 2019 on persistent organic pollutants (recast). European Union, Brussels.

Groothuis, F. A., Heringa, M. B., Nicol, B., Hermens, J. L. M., Blaauboer, B. J. & Kramer, N. I. 2015 Dose metric considerations in in vitro assays to improve quantitative in vitro–in vivo dose extrapolations. *Toxicology* **332**, 30–40.

Hair, J. F., Ringle, C. M. & Sarstedt, M. 2013 Partial least squares structural equation modeling: Rigorous applications, better results and higher acceptance. *Long Range Planning* **46**, 1–12.

Hale, S. E., Arp, H. P. H., Schliebner, I. & Neumann, M. 2020 Persistent, mobile and toxic (PMT) and very persistent and very mobile (vPvM) substances pose an equivalent level of concern to persistent, bioaccumulative and toxic (PBT) and very persistent and very bioaccumulative (vPvB) substances under REACH. *Environmental Sciences Europe* **32**, 155.

Holmberg, R., Wedebye, E. B., Nikolov, N. G. & Tyle, K. H. 2018 *How Many vPvM/PMT Substances Have Been Registered Under REACH? – vPvM/PMT Screening by Using the Danish QSAR Database*. Danmarks Tekniske Universitet (DTU), Kongens Lyngby.

Kim, S., Thiessen, P. A., Cheng, T., Yu, B. & Bolton, E. E. 2018 An update on PUG-REST: RESTful interface for programmatic access to PubChem. *Nucleic Acids Research* **46**, W563–W570.

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., Zaslavsky, L., Zhang, J. & Bolton, E. E. 2022 Pubchem 2023 update. *Nucleic Acids Research* **51**, D1373–D1380.

Kools, S. A. E., Roskam, G. D., Verheul, M. R. A. & Pieters, B. J. 2013 MRI contrast media, Magnetic Resonance Imaging (MRI) contrast media in het aquatisch milieu, Vereniging van rivierwaterbedrijven. RIWA Rijn.

Lambert, F. N., Vivian, D. N., Raimondo, S., Tebes-Stevens, C. T. & Barron, M. G. 2022 Relationships between aquatic toxicity, chemical hydrophobicity, and mode of action: Log Kow revisited. *Archives of Environmental Contamination and Toxicology*. **83** (4), 326–338.

Næs, T. & Mevik, B. H. 2001 Understanding the collinearity problem in regression and discriminant analysis. *Journal of Chemometrics: A Journal of the Chemometrics Society* **15**, 413–426.

OSPAR 2017 *Assessment Document of Land-Based Inputs of Microplastics in the Marine Environment*. OSPAR, London.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R. & Dubourg, V. 2011 Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research* **12**, 2825–2830.

RDKIT. *RDKit: Open-Source Cheminformatics*. Available from: https://www.rdkit.org [Online]. [Accessed].

Rigdon, E. 2014 Rethinking partial least squares path modeling: Breaking chains and forging ahead. *Long Range Planning* **47**, 161–167.

Rijn, R. 2023 *Jaarrapport 2022* (In: Stroomberg, G., ed.). Riwa Rijn, Nieuwegein.

RIVM 2024 *Zeer Zorgwekkende Stoffen* [Online]. Rivm. Available from: https://rvs.rivm.nl/onderwerpen/zeer-zorgwekkende-stoffen (accessed 9 November 2023).

Sarstedt, M., Ringle, C. M. & Hair, J. F. 2021 Partial least squares structural equation modeling. In: *Handbook of Market Research* (Homburg, C., Klarmann, M. & Vomberg, A. eds), Springer, Cham.

Schulte, C., Tietjen, L., Bambauer, A. & Fleischer, A. 2012 Five years REACH – lessons learned and first experiences. I. An authorities' view. *Environmental Sciences Europe* **24**, 31.

Sjerps, R. M. A., Brunner, A. M., Fujita, Y., Bajema, B., De Jonge, M., Bäuerlein, P. S., De Munk, J., Schriks, M. & Van Wezel, A. 2021 Clustering and prioritization to design a risk-based monitoring program in groundwater sources for drinking water. *Environmental Sciences Europe* **33**, 32.

USEPA 2012 Estimation Programs Interface Suite™ for Microsoft® Windows, v 4.11 United States Environmental Protection Agency, Washington, DC.

Wassenaar, P. N. H., Rorije, E., Vijver, M. G. & Peijnenburg, W. J. G. M. 2021 Evaluating chemical similarity as a measure to identify potential substances of very high concern. *Regulatory Toxicology and Pharmacology* **119**, 104834.

Wassenaar, P. N. H., Rorije, E., Vijver, M. G. & Peijnenburg, W. J. G. M. 2022 ZZS similarity tool: The online tool for similarity screening to identify chemicals of potential concern. *Journal of Computational Chemistry* **43**, 1042–1052.

Yang, H., Lou, C., Li, W., Liu, G. & Tang, Y. 2020 Computational approaches to identify structural alerts and their applications in environmental toxicology and drug discovery. *Chemical Research in Toxicology* **33** (6), 1312–1322.

First received 2 April 2024; accepted in revised form 20 June 2024. Available online 4 July 2024