Joint Research Programme
BTO 2023.086 | December 2023

A deeper understanding of (PMOC) toxicity

Joint Research Programme

KWR

Bridging Science to Practice

Colophon

A deeper understanding of PMOC toxicity

BTO 2023.086| December 2023

This research is part of the Joint Research Programme of KWR, the water utilities and Vewin.

Project number
402045/352

Project manager
Dr. P.S. Bäuerlein

Client
BTO - Thematical research - Chemical safety

Author(s)
Dr. R.P.J. Hoondert, Dr. M. de Baat, Dr. M. Yanagihara, Dr. T. ter Laak

Quality Assurance
Dr.ir. T. Pronk, Dr. M. M. L. Dingemans

Sent to
This report is distributed to BTO-participants.

A year after publication it is public.

Keywords
PMOCs; Random Forest

Year of publishing
2023

More information
Dr. R.P.J. Hoondert
T
E

PO Box 1072
3430 BB Nieuwegein
The Netherlands

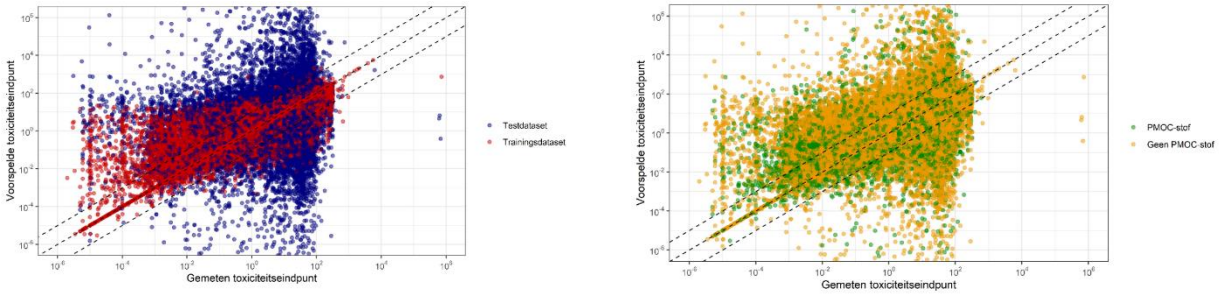
T +31 (0)30 60 69 511
E info@kwrwater.nl
I www.kwrwater.nl

Managementsamenvatting

Meer inzicht in manieren om de toxiciteit van persistente, mobiele organische stoffen (PMOC's) te voorspellen

Auteurs: Renske Hoondert, Miina Yanagihara, Thomas ter Laak

Onderzoek heeft meer inzicht gebracht in manieren om toxiciteit van stoffen te voorspellen. Een *random forest analyse* (een *machine learning* model) blijkt geen verband te laten zien tussen toxiciteit en stofstructuur, maar wel tussen stofkenmerken als mobiliteit en persistentie en de toxiciteit van een stof. Ook regressieanalyses lieten een significante correlatie zien tussen toxiciteit en de stofeigenschappen die de mobiliteit bepalen, maar met een lager voorspellend vermogen dan de *random forest analyse*. Het voorspellend vermogen van regressiemodellen op basis van stofstructuur was juist hoger dan het voorspellend vermogen van *random forest analyses* op basis van stofstructuur. Dit betekent dat de relatie tussen stofkenmerken en toxiciteit waarschijnlijk niet lineair (evenredig) is. Clustering van de toxiciteitstesten in de modellen resulteerde niet in betere voorspellingen. Stofstructuur en stofeigenschappen als voorspellende variabelen in modellen van een subset van toxiciteitstesten leverden wel voldoende informatie op om toxiciteitsklassen in plaats van toxiciteit te voorspellen (d.w.z. 'lage', 'medium', en 'hoge toxiciteit'). Dit kan de basis vormen voor een toekomstig hulpmiddel om toxiciteitsklassen te voorspellen in plaats van exacte toxiciteitswaarden.



Voorspelde toxiciteitswaarden (y-as) uitgezet tegenover gemeten toxiciteitswaarden (x-as) voor PMOC-stoffen en andere stoffen en voor de trainingsdataset (de data waarop het model gebaseerd is) en de testdataset (de data waarop het model NIET gebaseerd is). De modellen die zijn afgeleid in deze studie zijn gebaseerd op structuren van meer dan 5000 stoffen en meer dan 600 individuele toxiciteitstesten.

Belang: de relatie tussen stofstructuur, -eigenschappen en toxiciteit ophelderen

In een eerder project (zie BTO 2023.060 – Zijn PMOC's minder giftig?) is gekeken of persistente mobiele organische stoffen minder giftig zijn dan stoffen die dat niet zijn. Hieruit bleek dat meer mobiliteit van een stof vaak overeenkomt met een lagere toxiciteit. Maar er zijn andere factoren die mogelijk een rol kunnen spelen in de biologische activiteit van stoffen, zoals de structuur en de aan-/afwezigheid van bepaalde groepen. Ook rees de vraag of toxiciteitstesten konden worden geclusterd om de dataset te vergroten en mechanismen tussen specifieke stofeigenschappen (structurele elementen) en toxiciteit beter te begrijpen en te kunnen voorspellen.

Aanpak: de relatie tussen structuur, eigenschappen en toxiciteit van een stof

De relatie tussen de stofeigenschappen die stoffen persistent en mobiel maken en hun toxiciteit is met statistische technieken onderzocht. Met *random forest* analyses is bestudeerd welke stof-eigenschappen en groepen van atomen (structuurelementen) correleren met gemeten effectconcentraties in toxiciteitstesten. Deze

effectconcentraties komen uit de ToxCast dataset die 603 toxiciteitstesten en 5,114 stoffen omvat. Vervolgens is de relatie tussen deze stofeigenschappen en effectconcentraties geanalyseerd met *lineaire regressiemodellen*. Daarna zijn de toxiciteitstesten geclusterd op basis van diverse categorieën (bijvoorbeeld doeltype van de test of organisme-weefselcombinatie) en zijn deze technieken herhaald en geëvalueerd.

Resultaten: mobiele stoffen zijn gemiddeld minder giftig, persistentie zegt daarover niets

De *random forest analyse* in het huidige onderzoek toonde geen verband aan tussen toxiciteit en stofstructuur, terwijl wel een verband werd aangetoond tussen stofkenmerken (bijvoorbeeld mobiliteit en persistentie) en toxiciteit. De daaropvolgende *regressieanalyses* lieten opnieuw een significante correlatie zien tussen de toxiciteit en de stofeigenschappen die de mobiliteit bepalen, maar het voorspellend vermogen voor het *regressiemodel* was wel lager dan voor het *random forest* model, terwijl het voorspellende vermogen voor *regressiemodellen* op basis van stofstructuur juist hoger was dan voor *random forestmodellen* op basis van stofstructuur. Dit betekent dat de relatie tussen stofkenmerken en toxiciteit waarschijnlijk niet lineair (evenredig) is. De voorspellingen werden ook niet beter op het moment dat de toxiciteitstesten werden geclusterd.

Toepassing: beperkingen van de modellen belemmeren toepasbaarheid in (drink)watersector

Deze studie heeft meer inzicht gegeven in (het voorspellen van) de toxiciteit van stoffen met behulp van de ToxCast-database. De gebruikte modellen hebben echter beperkingen (zoals de gelimiteerde datasets per toxiciteitstest en hun onderlinge correlaties) die hun toepasbaarheid in de (drink)watersector belemmeren. De in de database opgenomen eindpunten van de toxiciteitstesten variëren aanzienlijk in bijvoorbeeld doeltype en testontwerp. Stofstructuur op zichzelf kan de verschillen in toxiciteit in deze testen niet verklaren en mogelijk zijn nog steeds onvoldoende data beschikbaar om betrouwbare correlaties af te leiden (zie figuur). Vaak leverden stofeigenschappen (vooral die gerelateerd aan persistentie en mobiliteit) als verklarende variabelen in veel gevallen (toxiciteitstesten) wél voldoende betrouwbare voorspellingen voor de activiteit in de test. Net als in eerder onderzoek (zie rapport BTO 2023.060) zijn mobielere, persistentere verbindingen doorgaans minder giftig. De betrouwbaarheid van de voorspellende modellen nam echter af wanneer testeindpunten werden geclusterd op basis van een van de categorieën. Stofstructuur en stofeigenschappen als voorspellende variabelen in modellen van een subset van toxiciteitstesten leverden voldoende informatie op om toxiciteitsklassen in plaats van toxiciteit te voorspellen (d.w.z. ‘laag’, ‘medium’, ‘hoog’). Om deze reden voorzien we in toekomstig onderzoek de ontwikkeling van een hulpmiddel om toxiciteitsklassen te voorspellen, in plaats van exacte toxiciteitswaarden voor deze specifieke subset van eindpunten. Daarnaast rijst de vraag of de toxiciteitstesten die zijn opgenomen in ToxCast wel de juiste testen zijn om daadwerkelijke effecten op organismeniveau te bepalen. Mogelijk leiden deze tot te hoge of juiste te lage schattingen van het effect van stoffen.

Rapport

Dit onderzoek is beschreven in het rapport *A deeper understanding of PMOC toxicity* (BTO 2023.086).

Contents

1 Introduction

Over the years, persistent and mobile (organic) compounds (PM(O)Cs) have been receiving increased attention in the drinking water sector. A compound will be labeled as a PMOC if it meets three criteria (partly based on criteria set by Neumann and Schliebner (2019)): i) the compound is organic; ii) the lowest organic carbon-water coefficient log Koc over the pH range of 4-9 is less than 4.0; and iii) the degradation half-life in fresh or estuarine water at 12 °C is higher than 40 days. PMOCs may pose threats to human health and the environment as the high mobility in water (as a result of their hydrophobicity) and persistence of these compounds lead to their occurrence and accumulation in surface water and drinking water sources. Additionally, some of these substances also tend to accumulate in the food chain, due to their bioaccumulative potential (Ghisi et al., 2019). The water sector is increasingly confronted with these substances. Because the high hydrophobicity of these chemicals makes it challenging to remove them by conventional water treatments, it is becoming increasingly important to estimate risks associated with PMOC emissions. For many of these chemicals, little is known about their toxicity as 'mobility' has historically been neglected as a prioritization criterium for ecotoxicological assessment. As it is time consuming and expensive to conduct *in vivo* and *in vitro* toxicity experiments, *in silico* approaches are useful in estimating hazards associated with such substances. The advantage is that they can be relatively easily conducted and do not require compound availability to conduct experiments. A disadvantage, however, is the lack of suitable, large datasets which may serve as basis e.g. for *in silico* predictive machine learning models (Hemmerich et al., 2020).

In a previous study, we focused on associations between toxicity of chemicals and their physicochemical properties that determine persistence and mobility in the environment. Random Forest analyses and multiple linear regression analyses indeed showed that properties related to polarity (hydrophilicity and mobility), particularly K_{OW} and K_{OC}, are inversely related to concentrations that elicit responses in bioassays ('effect concentrations' – AC_{50}), confirming that, in general, more polar chemicals are less toxic (BTO 2023.060: "Are PMOCs less toxic?"). The associations presented in the previous study indicate that PMOCs interact less with tissues, cell membranes, and receptors than similar but more hydrophobic chemicals, leading to lower intrinsic toxicity. However, the study also indicated that it may be difficult to predict (human) toxicity based on these physicochemical properties alone, as the processed ToxCast dataset of bioassay effect concentrations (at that time based on a list of water relevant PMOCs) covers not only a very diverse set of chemicals with different toxic modes of action, but also includes a large variety of assays with different toxicological endpoints. Furthermore, studies have shown that correlations between physicochemical descriptors (i.e. log K_{ow} and log K_{oc}) may not always be linear (Calleja et al., 1994a; Calleja et al., 1994b; Mackay et al., 2009). At the moment, a mechanistic understanding of what makes PMOCs more or less toxic is yet to be developed.

In the present study, we aim to deepen the understanding of PMOC toxicity by building upon the dataset and knowledge developed in this previous project and by producing QSARs (Quantitative Structure Activity Relationships) based on both linear regression analysis and random forest analysis (machine learning) to predict toxicity of PMOCs and non-PMOCs. Instead of focusing on a limited dataset of ~3000 chemicals, we now include the complete ToxCast database, containing tens of thousands of chemicals, and thousands of *in vitro* assays (Feshuk et al., 2023). Next to exploring non-linear relationships between toxicity and general physicochemical descriptors (e.g. K_{ow}, K_{oc}), we also look into structural properties and functional groups (taken from the OECD QSAR Toolbox (Schultz et al., 2018)) as explanatory variables in our models, and into grouping *in vitro* assays based on 'target' and 'study design' information from the individual experiments themselves. This information is based on annotations recommended by Phuong et al. (2014) on ToxCast assay characteristics including the intended target type, technological target type, assay design type, and signal direction.

In addition to the development of QSARs based on ToxCast data from *in vitro* assays, the identification of adverse outcome pathways (AOPs) may be useful to provide information on systemic toxicity, by linking chemical exposure to a series of events leading to an adverse health effect in humans. Several tools have been developed to provide information on adverse outcome pathways of chemicals, including the AOP wiki (Society for the Advancement of Adverse Outcome Pathways (SAAOP), 2023) and the AOP-helpFinder (Jaylet et al., 2023; Université Paris Cité, 2023). The latter tool highlights features to facilitate to search and interpret AOPs more easily (Jaylet et al. 2023). This tool is based on natural language processing (text mining) to search keywords in scientific literature stored in PubMed database, by screening abstracts. The search result is provided with a score to support the weight of evidence approach (Hardy et al., 2017). In the present study, we explored AOPs related to PMOCs that were listed based on the methodology described in paragraph 2.1. The AOP-helpFinder was employed in the analysis to search for possible AOPs from a wide range of previous studies in PubMed (Appendix II). A better understanding of PMOC toxicity will allow the identification of new and potentially hazardous PMOCs based on their chemical structures (e.g. specific structural alerts and features) and physicochemical properties that drive persistence and mobility. As such, the knowledge gap on the toxicity of PMOCs is narrowed, and PMT (persistent, mobile and toxic) chemicals can be identified from the larger pool of PMOCs for targeted risk mitigation.

2 Methods

2.1 Data acquisition, quality and formatting

To gain insight into the toxicity of PMOCs, the complete ToxCast dataset was downloaded, consisting of 21 databases, encompassing over 3.7 million toxicity data records (U.S. EPA, 2015). When only including active substances (substances triggering a toxicological response, an active hit call), the resulting dataset consists of 1677 individual *in vitro* assays endpoints (based on a smaller number of unique assays) from 20 sources (separate databases), with 357,010 data entries for 8,119 unique substances (based on CAS number). In ToxCast, the active concentration at which 50% of the effect is observed (AC_{50} in µM) is calculated using experimental concentration response series for a wide range of *in vitro* bioassays and three model types; a constant (two-parameter) model, a Hill (three parameter) S-model, and a gain-loss model, which is the product of two (three parameter) Hill models.

Log AC_{50}s (the active concentration at which 50% of the effect is observed) for the best predictive model (based on lowest Akaike Information Criterion (AIC) (Feshuk et al., 2023)) are calculated automatically. In ToxCast, concentration response series only get an active hit call (and high quality rating) when they meet three criteria (Filer et al., 2014):

1. The Hill model (S-curve model, see Figure 1) should emerge as the model with the best fit (based on lowest AIC – Akaike Information Coefficient (Feshuk et al., 2023))
2. The top of the modeled curve must be above the efficiency threshold (efficacy cutoff, the maximal experimental value on which the model was based)
3. For at least one concentration, the median response must be above the efficacy threshold

The complete dataset was cleaned up for analysis based on three criteria, based on:

- the number of active hit-calls (described above, especially criterion 1);
- the solubility of the chemicals (which should not be lower than the AC_{50} of the concentration-response curve);
- the sample size of the sub datasets.

In order to build a suitable database that can act as a training dataset in deriving our models, in the modelling exercise, we are only interested in experiments that meet aforementioned criteria, especially the first criterion. Below, a histogram depicting the percentage of the number of individual experiments in which the Hill model is the model with the best fit, for all individual *in vitro* assays, is shown (Figure 1). The figure shows that – in general – for most chemicals tested for the in vitro assay endpoints, the Hill model did NOT appear as the best fitted model (0-50%), while for a small subset of in vitro assay endpoints for all chemicals (100%) the Hill model appeared as the best fitted model.

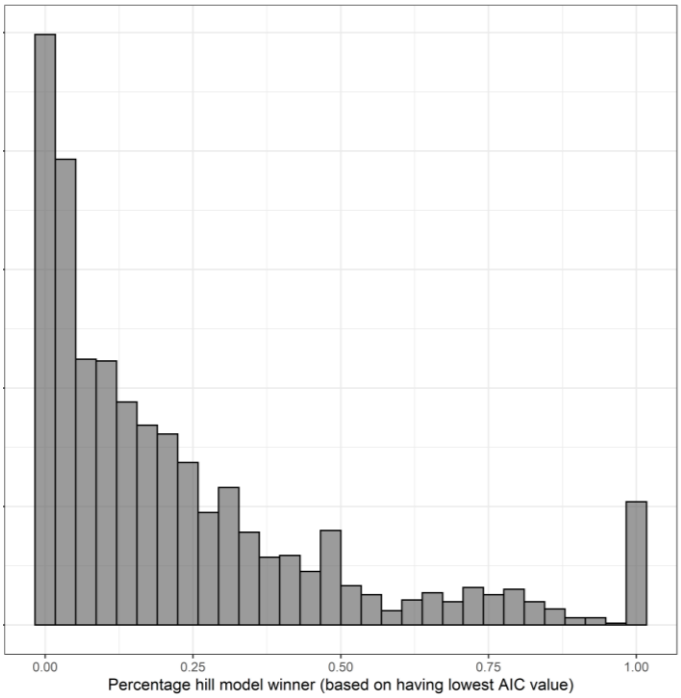


Figure 1: Histogram depicting the percentage of individual concentration-response experiments in which the Hill model is the best fitting model, per individual in vitro assay. X-axis: percentage hill model 'wins', y-axis: percentage of all experiments included in the dataset.

1) For 69 *in vitro* assays, for 100% of the individual experiments the Hill models appeared to have the best fit. However, sample sizes for these *in vitro* assays are extremely small (< 4 chemicals analyzed per *in vitro* assay endpoint; complete concentration-response curves – including replicates), and in total only 132 data records out of 357,010 data records (= single experiments) in the complete dataset are based on bioassays with a 100% hill model "winning" percentage. These data records were later removed from the dataset as the *in vitro* assays did not fulfill the sample size requirement (see criterion 3). For 160 *in vitro* assays for 0% of the individual experiments the Hill models appeared to have the best fit. These 160 individual *in vitro* assays covered a total of 9152 data records in the initial dataset. These *in vitro* assays were removed from the dataset.

The resulting data were combined with data on physicochemical parameters that drive mobility and persistence (K_{ow}, K_{oc}, molecular weight, degradation half-life (in days), and vapor pressure (in mmHg at 25 °C)) from EPISuite, resulting in a dataset of 281,369 data records, covering 6054 chemicals and 1627 *in vitro* assays.

2) Since low solubility of a compound frequently affects the actual exposure in a toxicity test (generally leading to underestimation of its effect) (Groothuis et al., 2015), poorly soluble chemicals (i.e. with a solubility in µM below the corresponding AC_{50}) were removed from the data set (Jonker & Van der Heijden, 2007). Solubility (Log_{10} M (molar)) was estimated using the WSKOWWIN v1.42 model. Substances with such low solubility often also have a high hydrophobicity and are for that reason not mobile and thus most likely not meeting the PMOC criteria (See introduction; log Koc < 4 (pH = 4.9), degradation rate in fresh water (at 12°C) over 40 days).

Truncating the data based on solubility resulted in a dataset of 5426 chemicals for 1588 *in vitro* assays, covering 162,743 data records or individual concentration-response curves. This selection of data based on defined criteria resulted in 43% reduction of the data records.

3) All *in vitro* toxicity tests for which less than 50 data entries (tested chemicals) were available were disregarded, to ensure an unbiased modelling practice. This cut-off of 50 data entries was based on the fact that 30% of the data were used as a test dataset, and a minimum of 30 data entries are required as

assumptions about the population distribution are not useful if the sample size does not exceed 30, since the sampling distribution approximates the standard normal distribution (Kwak & Kim, 2017).

Truncating the data based on this criterion resulted in a dataset of 5381 chemicals for 603 assay endpoints covering a smaller number of *in vitro* toxicity tests, covering 148,271 data records. This selection of data based on defined criteria resulted in 9% reduction of the data records.

The resulting data were combined with data on nested functional groups (structural fragments), extracted from the OECD QSAR Toolbox. Data were collected based on CAS number of chemicals. The Organic Functional Groups (OFG) system is designed to introduce some classification and systematisation of the various structural fragments in organic chemicals from a large database, and identify structurally similar chemicals. In total, 498 organic functional groups or structural fragments can be identified (European Chemicals Agency, 2014). For 7,771 chemicals, 396 structural fragments were identified in the formatted dataset based on ToxCast data, including hundreds of different dummy-variables (0-1). Combining the three datasets (i.e. the toxicity dataset, the dataset with physicochemical parameters and the dataset containing functional groups) and applying the three criteria (based on the Hill model being the 'winning' model, chemical solubility, and sample size) resulted in a final dataset of 5,114 chemicals, covering over 139 thousand individual AC_{50} values for 603 endpoints for *in vitro* assays. 4,780 of these chemicals are organic, of which 4,622 are mobile (Log K_{oc} below 4) and 1119 are persistent (half-life more than 40 days, BIOWIN3 score below 2.5) and mobile (PMOCs). Combining and formatting the datasets in the end resulted in a dataset with a broad coverage of chemicals (in terms of functional groups and physicochemical descriptors) from which approximately 20% consists of PMOCs. More information on the applicability domain of the models can be found in paragraph 2.7.

2.2 Data quality

Due to the diverse assay technologies and study designs deployed in the ToxCast database, a highly generalized and robust (median and median absolute deviation vs mean and standard deviation) set of calculations were performed to obtain robust AC_{50} values (U.S. Environmental Protection Agency (EPA), 2014). However, the ToxCast program has acknowledged that false positive and negative hit calls are possible using the automated methods, and has thus added a processing step to assign "flags" or "warnings" to the data (Ryan & Becker, 2017) related to a series of quality criteria such as 'noisy data', 'less than 50% efficacy', and 'borderline active result'. Ryan and Becker (2017) describe possible flags in the ToxCast dataset that may be considered when analyzing a list of possible results. However, they also note that their assignment is automated, and prone to error. Therefore, it may not be the best practice to set hard filters based on these flags. Because of this, although multiple flags have been found in the data (Figure 2), these were not used as criteria in formatting the data. In the figure below, the total number of registered flags within the formatted (training) dataset are shown.

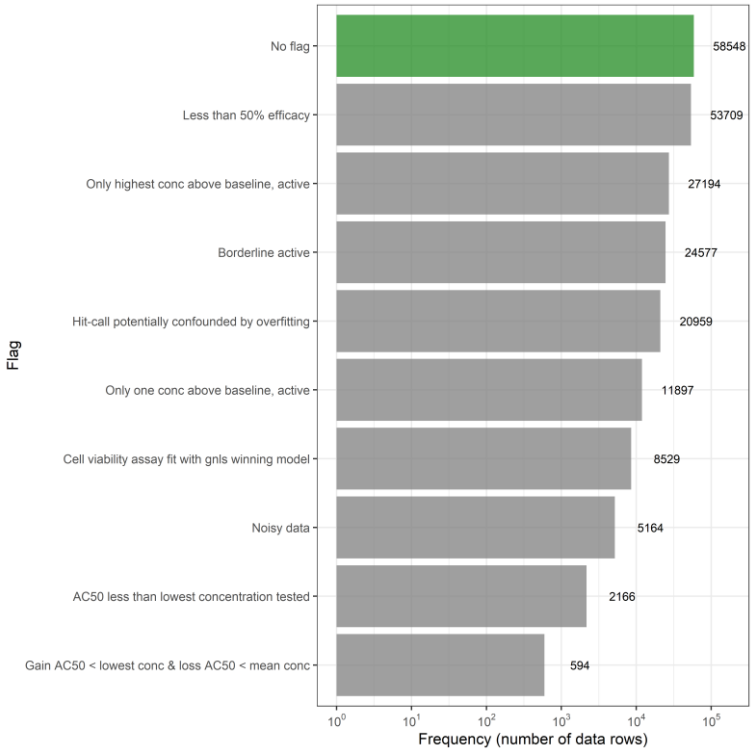


Figure 2: Frequency (number of data records out of 139,364) of flags included in the formatted dataset.

In total, for 58,548 (40.1%) individual data records out of 139,364 data records, no flags were found. For 53,709 data records (36.8%) from individual biochemical experiments the efficacy was below 50%, for 27,194 data records (18.6%) only the highest concentration was above the baseline, for 24,577 (16.8%) data records were borderline active, and for 20959 data records (14.4%) the hit-call was potentially confounded by overfitting. The sum of the classes exceed the total number of records as records can be flagged for multiple criteria. Overall, although a large sum of flags have been reported in the data, these flags are equally distributed over non-PMOCs and PMOCs, so it is safe to assume that the models produced in the research presented here are based on equally 'bad' data for both chemical groups.

For all individual concentration-response curves the width of the confidence interval of the AC_{50} (in log units) is reported in the ToxCast database. These confidence intervals are divided by the corresponding (mean) AC_{50} to be able to compare these values across all endpoint for *in vitro* assays and chemicals. Below, a histogram depicting these ratios for all individual experiments in all individual *in vitro* assays in which the Hill model is the model with the best fit is shown (Figure 3). On average, the confidence interval of the data is 5% of the Log AC_{50}. 95% of all confidence intervals (based on 139,364 individual data records) lies within 1.5 times the log AC_{50}. Although this may sound like a large average deviation, this deviation may be negligible when prioritizing chemicals based on a classification using AC_{50} values or based on threshold levels, which may deviate multiple magnitudes.

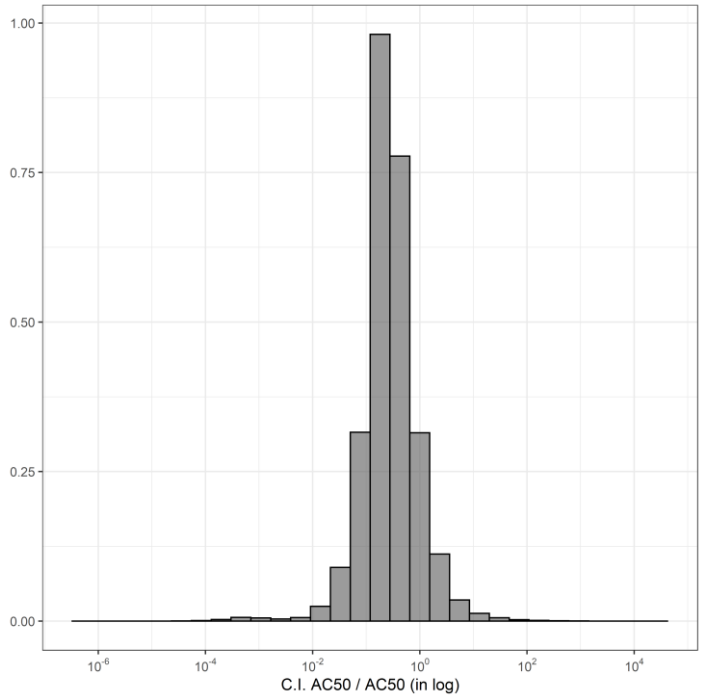


Figure 3: histogram depicting the ratio between the AC_{50} and its confidence interval.

2.3 Grouping of *in vitro* assays

Each data record in the ToxCast database has a distinct set of annotations, i.e., descriptive features that capture a particular aspect of the assay endpoint and *in vitro* assay used. Most of the 38 annotations are related to at least one other annotation (Figure 4). Since the formatted dataset used in the present research comprised a large number of *in vitro* assays, these were classified based on type and characteristics, such as intended target family, technological target type, assay design type, signal direction, target organism, and target tissue. The type and characteristics of the *in vitro* assays in the ToxCast dataset are described below (Phuong et al., 2014). These are equal to the annotations in the formatted database.

In vitro assays capture the effects of chemicals on different types of targets related to biological processes (Figure 4). The *intended target family* captures the objective (qualitative) form of the intended target (the representative genetic family or biological process of the target (e.g., cell cycle, neurodevelopment or DNA binding)), while the *technological target type* provides the measured (quantitative) form of the target used in the experimental methods (e.g., embryonic development, electrical activity, RNA production or molecular messenger) (Phuong et al., 2014). The *assay design type* of an *in vitro* assay is related to the technology used to translate a biological or physical process to a detectable signal (e.g. enzyme reporter or growth reporter), and the *signal direction* corresponds to the expected direction of the detected signal in relation to the negative control (either gain or loss) (Phuong et al., 2014) (Figure 4).

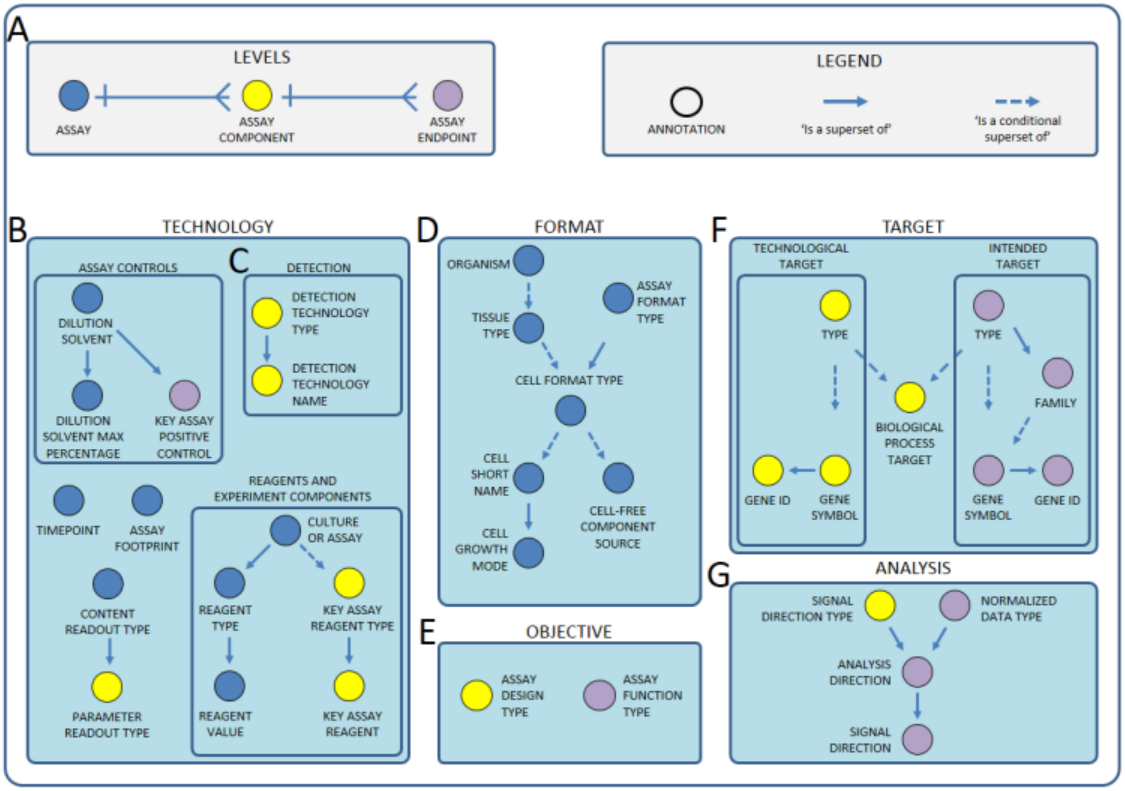


Figure 4: The assay annotation structure. The 38 annotations can be grouped into (among other things) (A) assay information, (B): technology information, (C): detection information, (D): format information, I design information, (F) target information, and (G) analysis information (Phuong et al., 2014).

Below, pie charts (Figure 5) show the relative frequencies of individual targets within set categories, including all intended target families, technological target types, assay design types, signal direction, and organism-tissue type. *In vitro* assay endpoints included in the formatted dataset are equally spread out over 49 intended target families (Figure 4), while most *in vitro* assays in the formatted dataset have a protein (41%) or RNA (40.3%) technological target type. The majority of *in vitro* assays have an assay design type related to either inducible reporters (48%), or binding reporters (23.1%). A slight majority of all *in vitro* assays have a loss signal direction (54.1%), and most *in vitro* assays are based on human (86.7 %) cells and mammalian liver cells (47.6 %). Human liver cells account for 47.3% of all data records. Ideally, an equal composition of annotations is included in the training dataset in the modelling process, as sub setting data based on an unequal distribution of annotations leads to bigger and smaller datasets, and model performance based on smaller datasets may be lower when applying the model on chemicals outside the training dataset.

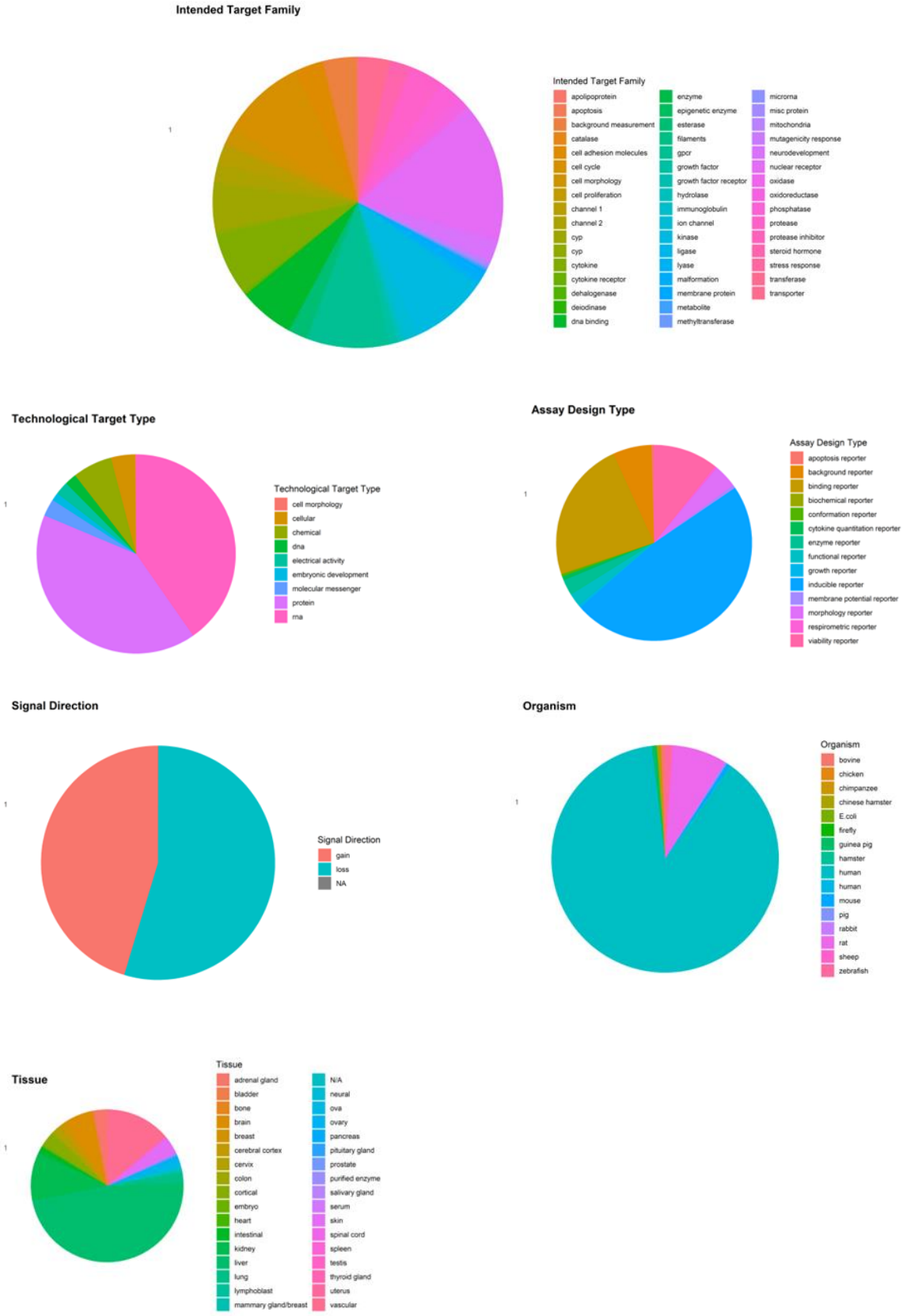


Figure 5: Pie charts showing the relative frequencies of targets withing categories of in vitro assays, based on intended target family, technological target type, assay design type, signal direction, organism and tissue.

2.4 Random Forest model

Random forest is a supervised learning algorithm using an ensemble of decision trees, capable of performing both regression and classification tasks. The algorithm continually randomly selects a subset of physicochemical descriptors or structural fragments/functional groups and subdivides the data based on these descriptors until a full tree is developed and analyzed for predictive power using these physicochemical descriptors or structural fragments/functional groups. The algorithm arrives at the best explanatory properties by always prioritizing the decision trees with the properties that perform best to explain toxicity. The randomization process reduces bias and decreases variance between and within trees. Random forest is, aside from its ability to build accurate classifiers, an often used objective method for feature importance assessment and selection. To get more insight into toxicity of chemicals (including PMOCs), random forest analyses were performed based on physicochemical characteristics and functional groups, taking toxicological endpoints (log_{10}-transformed AC_{50}) as a response variable for each toxicity test individually.

A fixed number of 5000 decision trees was used in the random forest analyses and the top physicochemical descriptors or top 10 functional groups explaining the most variance in AC_{50} were reported. Being a non-parametric method, Random forest analysis does not require the response variable and/or the predictors to be normally distributed.

The predictive power of variables within a Random Forest analysis is determined by calculating the %IncMSE (increase in mean-squared error of the predicted values). This is a measure for the importance of the feature; if the values of the feature are randomized in the same trees, what would be the drop in accuracy. This is the most robust and informative value within the analysis. This value is calculated by comparing the MSE when dropping explanatory variable (j) to overall (initial) MSE_0:

$$\%IncMSE = \frac{\left(MSE_j - MSE_0\right)}{MSE_0} * 100\% \quad [1]$$

where a higher number indicates a better prediction.

A disadvantage of any machine learning model, including random forest, is that it is so complicated that it can only be applied as a computer model. It is not intuitively easy to interpret. Additionally, the random forest output does not include quantitative regression coefficients and therefore does not provide insight into the magnitude or direction of the observed effect in an *in vitro* assay. Additional multiple linear regression analyses were thus performed to provide insight into the magnitude and direction of the relationship between toxicity (biological activity) and continuous variables (i.e. physicochemical descriptors or functional groups).

2.5 Multiple linear regression model

Additional QSARs for toxicity (next to the random forest models) were derived by fitting the following conceptual model to the formatted data:

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_n x_n \quad [2]$$

in which β_1 to β_n represent the regression coefficients associated with the 1st to nth X_1 to X_n chemical property (either a physicochemical descriptor or a functional group), and y represents the toxic potency on an endpoint in an *in vitro* assay (log_{10}-transformed AC_{50}). To enable comparison of results between *in vitro* assays and *in vitro* assay groups, prior to the derivation of the multiple linear regression model, the response variable (AC_{50}) was standardized using the z-score:

$$z_i = \frac{x_i - \bar{x}_i}{s_i} \quad [3]$$

which transforms the overall toxicity mean ($\bar{x}_i$) to 0 and the corresponding standard deviation (s_i) to 1 (Eriksson et al., 2003). Usually the values in a z-score are within -3 to 3. We derived multiple regression models separately for each *in vitro* assay and *in vitro* assay group, that incorporated the full set of physicochemical parameters and included at least 50 data records (see 2.1), using the *lm* function in R, Ver. 4.1.1 (Team, 2021). Uncomplicated models allow for easier interpretation and are for that reason more suitable for screening-level impact assessments. Therefore no interactions or quadratic functions were included in model derivation. Afterwards, the most influential predictors (physicochemical parameters) of toxicity (AC_{50}) were identified using the Relaimpo package R statistics, Ver. 4.1.1.

2.6 Model evaluation

Both the results from the Random Forest analysis and the multiple linear regression analysis were evaluated by plotting predicted effect concentrations against the observed effect concentrations, taking the aforementioned functional groups as explanatory variables, for all individual *in vitro* assay endpoints and *in vitro* assay groups, separately. 70% of the data were used in a training dataset and the remaining 30% served as a test dataset. These data were randomly selected. The models were evaluated using the coefficient of determination (R^2) for the training set(see Equation 4):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad [4]$$

where the 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 statistics explains the variance in the response variable 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 this study, R^2 values of 0.75, 0.50, or 0.25 for response variables will be described as substantial, moderate or weak, respectively, according to Hair et al. (2013) and Sarstedt et al (2021) (Hair et al., 2013; Sarstedt et al., 2021). The predictive power of the model is evaluated by calculating the Q^2 for the test dataset:

$$Q^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_{ext})^2}{\sum_{i=1}^{n}(y_{ext} - \bar{y}_{ext})^2} \quad [5]$$

where the Q^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 Q^2 statistic reflects predictive relevance, and measures whether a model has predictive relevance or not. Q^2 values above zero indicate that your values are well reconstructed and that the model has predictive relevance.

2.7 Applicability domain of the models

The domain of applicability is an important concept in QSARs. It allows to estimate the uncertainty of the prediction of a particular molecule based on how similar it is to chemicals used to build the model (Weaver & Gleeson, 2008). In this case, the applicability domain of the developed QSAR is the range of physicochemical properties (related to PMOC-properties; mobility and persistence), and the structural information (based on structural fragments/functional groups) on which the Random Forest model and the multiple linear regression model have

been developed. This defines the properties of any new chemicals for which the model is applicable to make predictions (Table x). Any predictions on new chemicals that have deviating properties, can be incorrect. The applicability domain of the three most important physicochemical descriptors related to persistence and mobility are described in the paragraphs below.

Table 1: Source of physicochemical properties included in the modelling process.

Property	Source
Octanol-water partitioning coefficient (K_{ow})	EPI Suite™ (experimentally based and estimated through
Soil sorption coefficient (K_{oc})	EPI Suite™ (experimentally based or based on MCI-method) and OPERA
Molecular weight	EPI Suite™
Biodegradation rate (half life in days)	OPERA (estimated half life in days based on PaDel descriptors)
Vapor pressure	EPI Suite™
Functional groups	Organic functional groups via OECD QSAR Toolbox

2.7.1 Soil sorption coefficient (mobility)

The soil sorption coefficient (K_{oc}) of chemicals was included as an explanatory variable in both the Random Forest model and the multiple linear regression model. In the formatted dataset (training dataset), we have included log K_{oc} values taken from two separate sources; experimental and predicted values from EPI Suite™ (KOWWIN v1.68) (EPISKOC_EXP and EPISKOC_MCI, respectively) (US EPA, 2022), and predicted values from OPERA (OPERAKOC) (Mansouri & Williams, 2017). While EPI Suite predicts log K_{oc} values based on the Randiç Molecular Connectivity Index (Randić, 2001), the OPERA model predicts log K_{oc} values based on PaDEL descriptors (1D, 2D, 3D descriptors and fingerprints) (Yap, 2011). If no experimental data were available for a compound (this was the case for 72125 chemicals (49.4%)), the log K_{oc} based on the molecular connectivity index was chosen. If no data were available on log $K_{oc\text{-}MCI}$, the log K_{oc} from OPERA was taken instead. Figure 6 shows the range of log K_{OC}s of chemicals included in the formatted dataset, separated by data source and method. The average log K_{oc} found in the formatted dataset is 2.66 with 95% of the log K_{oc} falling within the 0.82 – 4.82 range, implying that the majority of the chemicals in the formatted dataset is mobile (log K_{oc} < 4; (Neumann & Schliebner, 2019)).

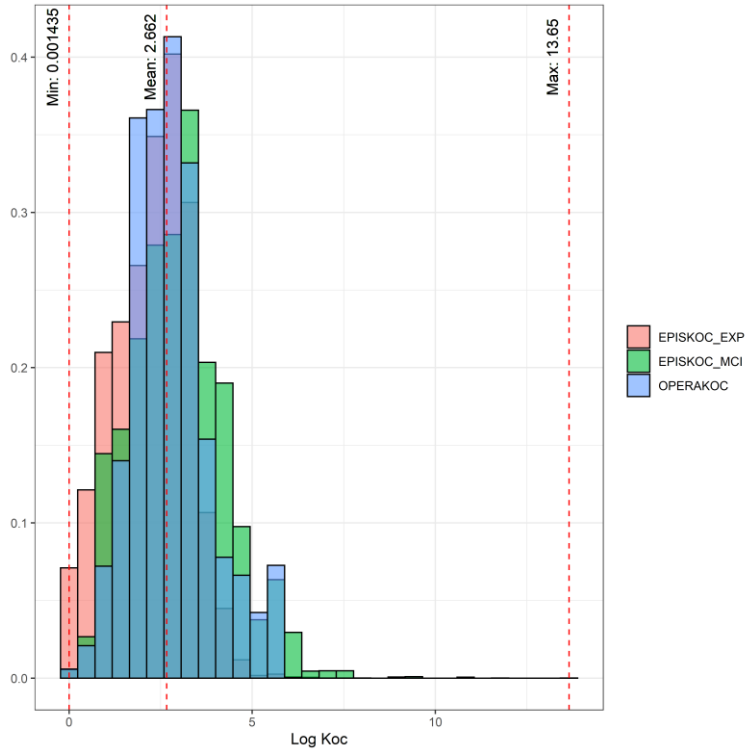


Figure 6: Histogram depicting the range/ applicability domain of log K_{oc} within the formatted dataset.

2.7.2 Octanol-water partitioning coefficient (mobility/bioaccumulation/bioavailability)

The octanol-water partitioning coefficient (K_{ow}) of chemicals was included as an explanatory variable in both the Random Forest model and the multiple linear regression model. In the formatted dataset (training dataset), we have included log experimental and predicted K_{ow} values from EPI Suite™ (KOWWIN v1.68) (EPISKOW_EXP and EPISKOW_Pred, respectively) (US EPA, 2022). EPI Suite uses a "fragment constant" method to predict K_{ow}. In the "fragment constant" method, a molecule is divided into fragments (atoms or larger structural fragments/functional groups) and the assigned coefficient values for each fragment are added to give the K_{ow} estimate, which is reported as a log. If no experimental data were available for a compound (which was the case for 3044 out of 5114 individual chemicals), the estimated K_{ow} was taken as a substitute. Figure 7 shows a histogram depicting the range of log K_{ow} of chemicals included in the formatted dataset, separated by data source and method. The average log K_{ow} found in the formatted dataset is 2.16, with 95% of the log K_{ow}s falling within the -0.43 – 4.6 range. Although there has not been a scientific consensus on threshold values for hydrophobicity/hydrophilicity, like with K_{oc}, The compounds used in the formatted dataset can be considered relatively hydrophilic (Log K_{ow} < 5).

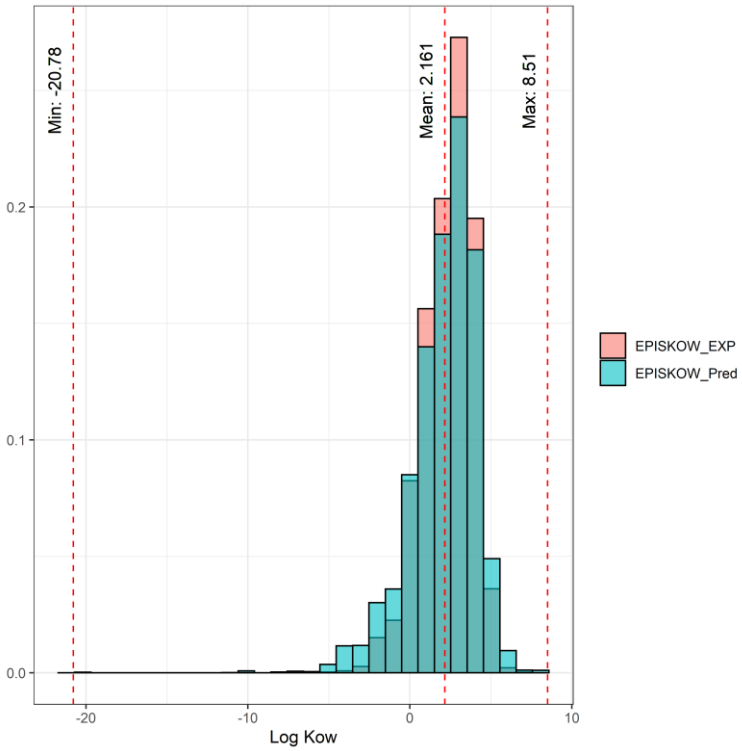


Figure 7: Histogram depicting the range/ applicability domain of log K_{ow}s within the formatted dataset.

2.7.3 Biodegradation rate (persistence)

The biodegradation rate (reported in half-life in days) of chemicals was included as an explanatory variable in both the Random Forest model and the multiple linear regression model. In the formatted dataset (training dataset), we have included biodegradation (half-life in days), estimated through OPERA. The OPERA model predicts biodegradation rates based on PaDEL descriptors (1D, 2D, 3D descriptors and fingerprints) (Yap, 2011). Figure 8 shows a histogram depicting the range of half-lives of chemicals included in the formatted dataset. The average half-life found in the formatted dataset is 0.92 = 8.3 days, with 95% of the half-lives falling within the 0.52 – 1.99 (3.34 – 98 days) range, implying that the chemicals in the formatted dataset are equally distributed with respect to their biodegradability (persistent compounds have a degradation half-life in fresh or estuarine water at 12 °C that is higher than 40 days (Neumann & Schliebner, 2019)).

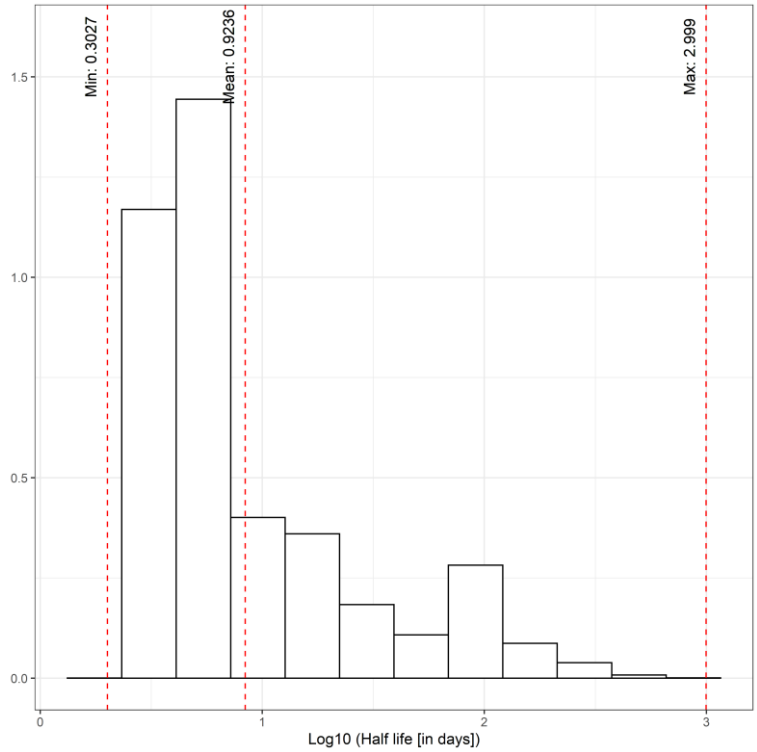


Figure 8: histogram depicting the range/ applicability domain of half lives for chemicals within the formatted dataset.

2.7.4 Vapor pressure and molecular weight

Additional to the PM-parameters included above, the vapor pressure and molecular weight of chemicals were included as an explanatory variable in both the Random Forest model and the multiple linear regression model. In the formatted dataset (training dataset), we have included vapor pressure taken from EPI Suite™ (MPBPWIN) and molecular weight from EPI Suite. The MPBPWIN model predicts vapor pressure (in mmHg at 25°C) based on molecular fragments. Figure 9A (left) shows a histogram depicting the range of the vapor pressure of chemicals included in the formatted dataset. The average log-transformed vapor pressure found in the formatted dataset is -6.74, with 95% of the vapor pressures falling within the -14.82 – -0.69 range. Figure 9B (right) shows a histogram depicting the range of the molecular weight of chemicals included in the formatted dataset. The average log-transformed molecular weight found in the formatted dataset is 2.36 (261.25 grams per mole), with 95% of the vapor pressures falling within the 2.09 – 2.63 (123.16 – 424.39 grams per mole) range.

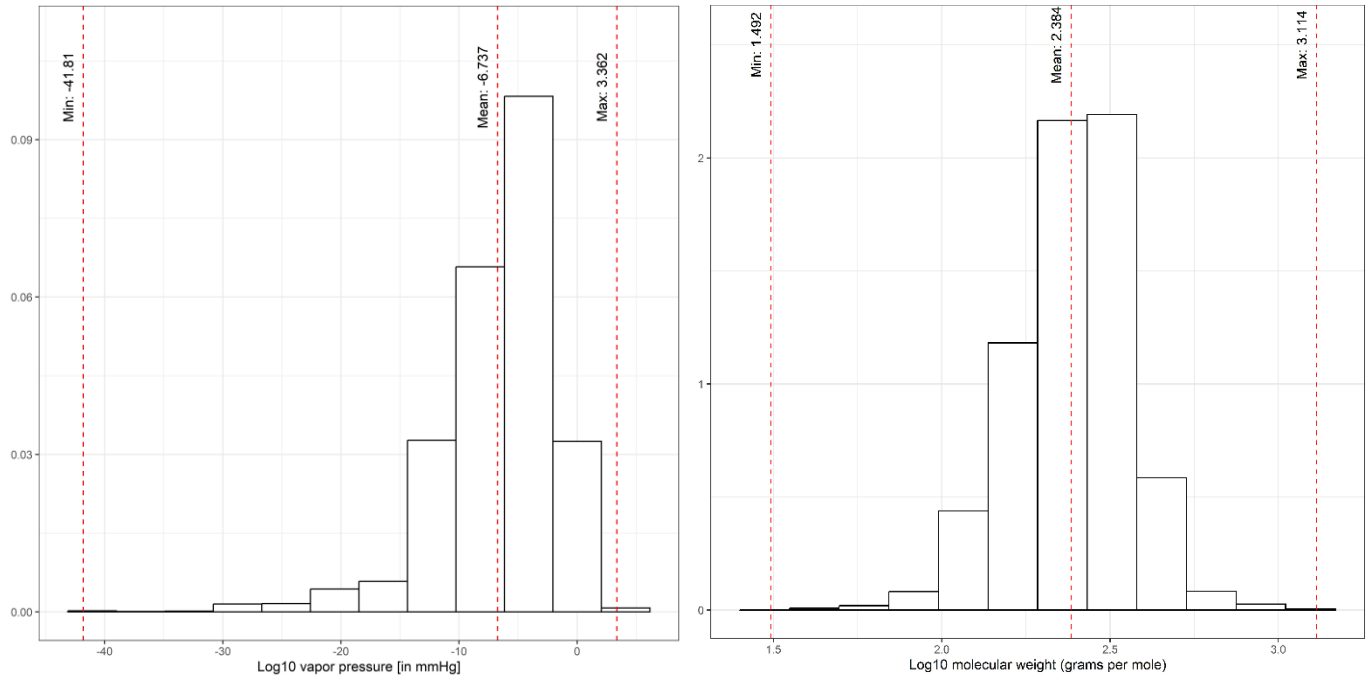


Figure 9: histogram depicting the range/ applicability domain of vapor pressure (left) and molecular weight (right) for chemicals within the formatted dataset.

3 Results

3.1 Toxicity

In Figure 10, the distribution of all toxicity values (AC_{50}s) across all *in vitro* assays and *in vitro* assay types is shown, for both chemicals labeled as PMOCs as well as other chemicals. The average AC_{50} – covering all *in vitro* assays and *in vitro* assay types – for PMOCs was 1.73 $\log_{10}$ µM (= µmol per liter) (median: 1.33, S.D.: 1.51), or 0.937 $\log_{10}$ mg/L (median: 0.67, S.D.: 0.906), while the average AC_{50} for non-PMOCs was 1. 46 $\log_{10}$ µM (median: 1.32, S.D.: 3.6), or 0.85 $\log_{10}$ mg/L (median: 0.69, S.D.: 2.48). Overall, AC_{50}s (in both µM and mg/L) associated with PMOCs were higher than AC_{50}s associated with non-PMOCs ($p < 0.05$, one-sided (upper-bound) t-test), giving a first indication of lower toxicity of PMOCs, compared to non-PMOCs, as was concluded in the previous BTO report on PMOC toxicity (BTO 2023.60). However, the PMOC-group consisted of a relatively small number of individual chemicals (n = 1116), while the non-PMOC-group consisted of a large, diverse group of chemicals (n = 3995), possibly eliciting a large variety of effects. Additionally, no differentiation was made between *in vitro* assay types, covering a vast amount of different effects.

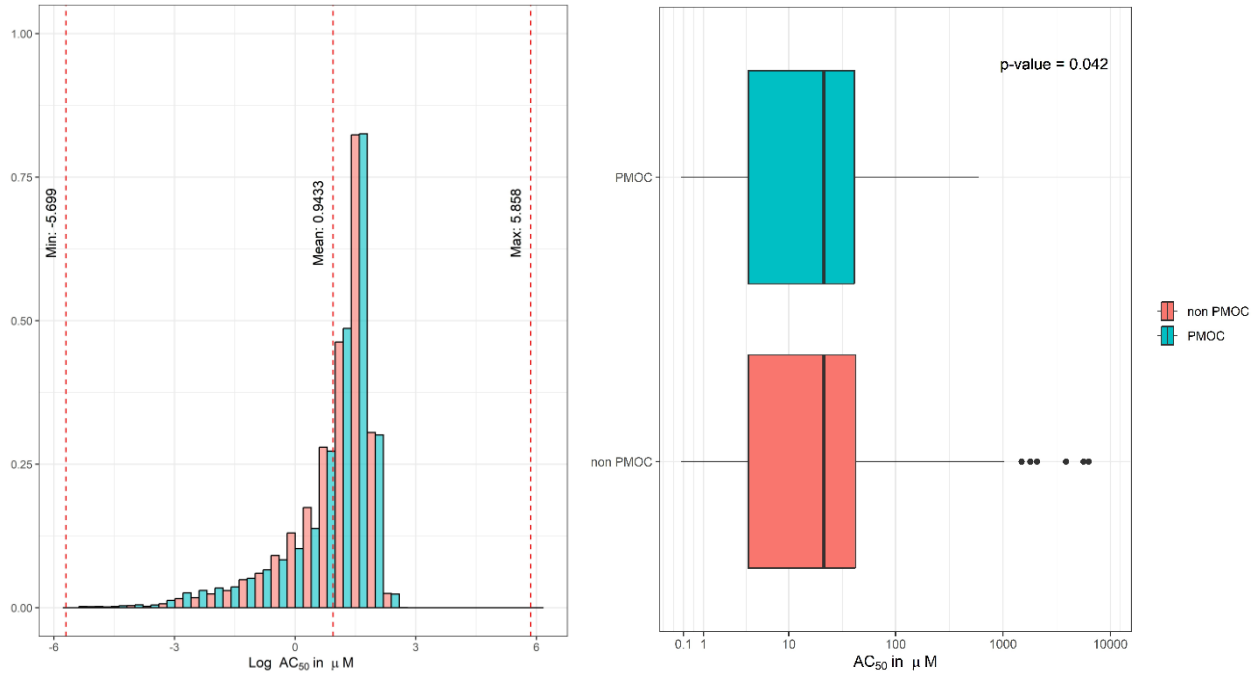


Figure 10: distribution of log-transformed endpoints (AC_{50}s in µM) for PMOCs, and for chemicals not labeled as PMOCs as a histogram (left) and boxplot (right).

In the figure and analysis above, no distinction was made between *in vitro* assays and – therefore – endpoint types. To enable comparison of results between *in vitro* assays and *in vitro* assay groups, prior to the derivation of the multiple linear regression model, the response variable (AC_{50}) was standardized to show relative toxicity and the relative position of the compound within the distribution of toxicities for each *in vitro* assay (see Equation 3). In Figure 11, the distribution of all z-transformed toxicity values (AC_{50}s) across all *in vitro* assays and *in vitro* assay types is shown, for chemicals labeled as PMOCs (see paragraph 2.1), and other chemicals (non PMOCs). A z value represents the deviation of AC_{50}s from the mean/average of all toxicity data per individual *in vitro* assay endpoint, expressed in

number of standard deviation units. Although the median of z values for the PMOC and non PMOC group may differ, the geometric average of all z values should be 0 (equal to the true average of all data). The average z value – standardized based on all individual *in vitro* assays – for PMOCs was 0.031 (median: -0.23, S.D.: 0.98), which implies that PMOCs have a slightly higher average AC_{50} value compared to all toxicity data for all individual *in vitro* assay endpoints. However, please note that z values for PMOCs may differ across individual in vitro assay endpoints, i.e. PMOCs may be less toxic when looking at – for instance – neurodevelopment, but may appear more toxic when looking at cytotoxicity. The average AC_{50} for non-PMOCs was -0.0054 (median: -0.24, S.D.: 1.00), implying that non-PMOCs may – when including all endpoints and assay types – be slightly more toxic than PMOCs in most *in vitro* assays. This was confirmed by a one-sided (upper-bound) t-test ($p < 0.05$).

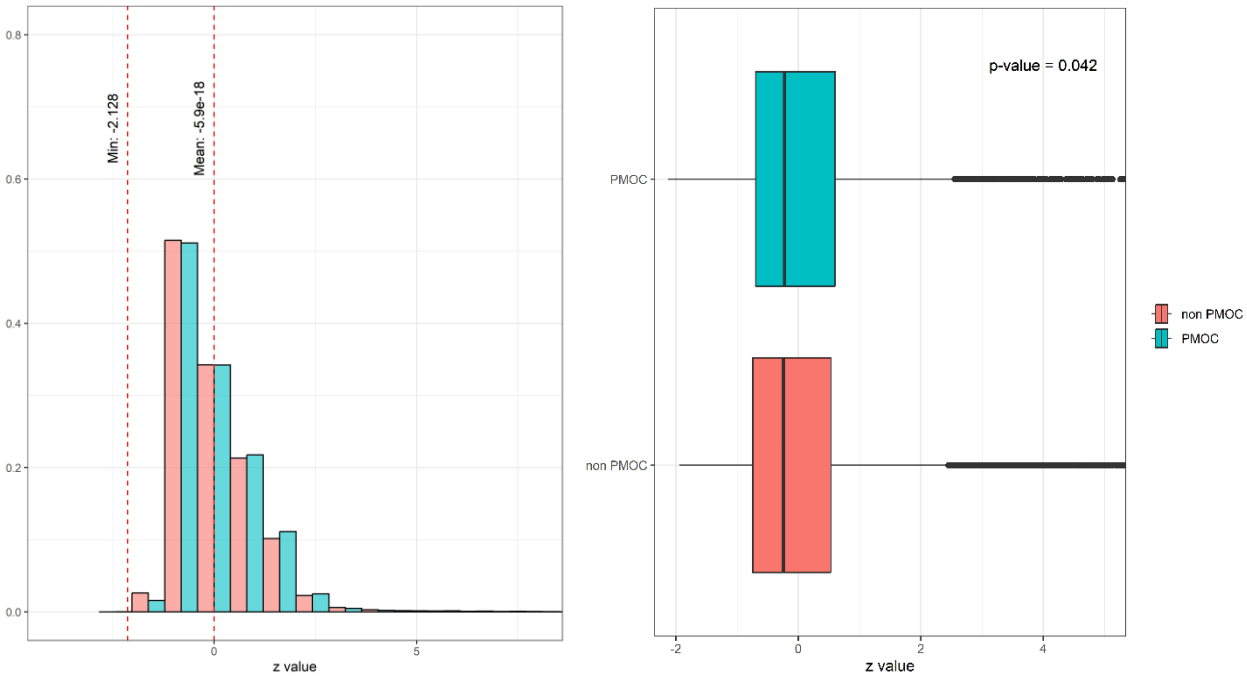


Figure 11: distribution of log-transformed endpoints (z values) for PMOCs, and for chemicals not labeled as PMOCs as a histogram (left) and boxplot (right)

In the random forest and linear regression analyses the functional groups/structural elements and physicochemical descriptors taken as explanatory variables were analyzed separately, as the structural elements and topological features of compounds in itself may be strongly correlated with physicochemical descriptors (Cocchi et al., 1999).

3.2 Structural fragments/functional groups

3.2.1 Random Forest

In general, the Random Forest analysis for all 603 *in vitro* assays separately, including only the functional groups as explanatory variables, explained on average -0.96% (median: -0.95%, S.E.: 0.004) of all variance in the toxicity data for both PMOCs and non-PMOCs (AC_{50}s). The highest percentage of variances explained were determined for TOX21_PXR_viability[1] (9.55%), while the lowest percentage variance explained by the Random Forest model were found for NVS_GPCR_rAdra2_NonSelective[2] (-7.97%). For almost 12% (11.9% - 72 assay endpoints) of all *in vitro* assay endpoints organotin was identified as the most important structural fragment for the prediction of toxicity in the Random Forest model, followed by steroids (8.5% - 52 assay endpoints), and acetals (3.1% - 19 assay endpoints). Chemicals including an organotin fragment included covered only 1.9% of the complete formatted database, including only 15 (out of 5114) compounds, which may have been very toxic, and covering only *in vitro* assay endpoint for which a small dataset was available. A total of 114 chemicals included a steroidal structural fragment (covering 4.7% of the total dataset) and a total of 55 chemicals included an acetal (covering only 0.8% of the complete formatted dataset). This shows that the distribution of chemicals within each individual dataset for each assay endpoint may differ considerably and may disproportionally steer the final conclusion. However, when weighing assay endpoints based on sample size of their respective datasets, organotin was again identified as the most structural fragment, with assay endpoint datasets for which organotin was identified as the most predictive structural fragment for toxicity covering over 16% of all data. Steroids again followed as the second most important structural fragment for the prediction of toxicity (assay activity) with a total coverage of 9.2% of all formatted data, while dithiocarbamates (and not acetals) were identified as the third most important predictor of assay activity, covering 6.2% of the complete formatted dataset.

Figure 12 shows the predicted effect concentrations (log AC_{50}s) – predicted by both Random Forest and multiple linear regression - plotted against the observed effect concentrations. taking the aforementioned functional groups as explanatory variables. In total, 65.6% of all individual predicted AC_{50}s were within a factor of 5 of the observed AC_{50}s; 17% of the predicted datapoints were more than a factor five below the observed datapoints (underestimated), and 17.3% were more than a factor five above the observed data (overestimated), which can be considered high. None of the predicted datapoints were a perfect fit, which indicates that no overfitting of the model occurs. This occurs when the model is too complex, when there are an overly large number of parameters compared to the number of observations. In that case, the model will perform well on training data, but poorly on test data. Although no overfitting in the model takes place, the accuracy of the random forest model, when including functional groups as explanatory variables, is very low.

3.2.2 Multiple linear regression analysis

In general, the multiple linear regression analysis for all 603 *in vitro* assays, separately, including only the structural fragments/functional groups as explanatory variables, explained on average 51.44% (median: 54.48%, S.E.: 0.042%) of all variance in the toxicity data (AC_{50}s), based on the adjusted R^2. The highest % of variances explained were determined for a specific *in vitro* assay focusing on cytotoxicity: BSK_Sag_PBMCCytotoxicity_up[3] (100%). However, as fitting of the multiple linear regression model is based on only 33 data entries, this high predictability is likely due to overfitting of the model. The lowest % variance explained by the multiple linear regression model was found for Tanguay_ZF_120hpf_PE_up[4] (an *in vitro* assay focusing on embryonic vascular disruption) (-2.4%). Figure 12 shows

[1] https://comptox.epa.gov/dashboard/assay-endpoints/TOX21_PXR_viability

[2] https://comptox.epa.gov/dashboard/assay-endpoints/NVS_GPCR_rAdra2_NonSelective

[3] https://comptox.epa.gov/dashboard/assay-endpoints/BSK_SAg_PBMCCytotoxicity_up

[4] https://comptox.epa.gov/dashboard/assay-endpoints/Tanguay_ZF_120hpf_PE_up

the predicted effect concentrations (Log_{10} AC_{50}s) plotted against the observed effect concentrations, based on the multiple linear regression model, taking the functional groups as explanatory variables (Equation 2). In total, 86.1% of all individual predicted AC_{50}s lied within a factor 5 of the observed AC_{50}s; 7.1% of the predicted datapoints were more than a factor five below the observed datapoints (underestimated), while 6.8% were more than a factor five 5 above the observed data (overestimated). 0.85% of the predicted datapoints were a perfect fit, which may indicate overfitting of the model.

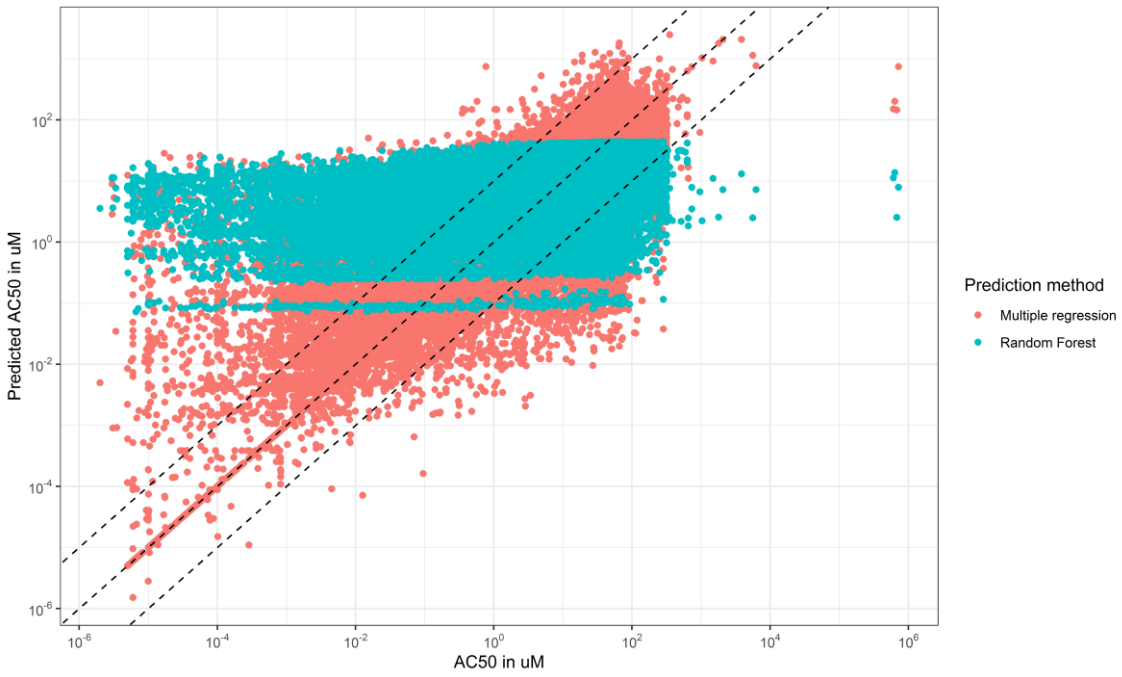


Figure 12: Predicted toxicity (AC_{50} in µM) in the training dataset by the multiple linear regression model and the Random Forest model versus observed toxicity, based on structural fragments/functional groups, clustered per individual in vitro assay, for the training dataset only. The middle dashed line represents the 1:1 ratio. The outer dashed lines represent the 1:5 and 5:1 ratio.

The predictive power of the individual multiple linear regression models for both the training dataset and the test dataset were evaluated by comparing predicted AC_{50}s with observed (experimental) values (Figure 13). Here, we see that 82.22% of all predicted data points were within a factor of five of the experimental AC_{50}s. 6.11% of the data points were overfitted (û-u=0; the difference between the predicted value and the observation is 0), 9.14% underestimated (u/5 > û; ; the predicted value is more than five times lower than the observation), and 9.05% overestimated (u*5 < û; the predicted value is more than five times higher than the observation). When we solely look at the test dataset, we see that 61.6% of all data records were within a factor five of the experimental AC_{50}s (training: 87%), 19.7% of the data records in the test dataset were underestimated (training: 7.06%), 18.75% of the data records in the test dataset were overestimated (training: 6.8%), and 0.025% of the data records in the test dataset were overfitted (training: 7.04%). Overall, 71.6% of the variation in AC_{50} values in the training dataset is explained by the linear regression model (R^2), when excluding interaction terms between the explanatory variables (Figure 13). Unfortunately, it was not possible to include interaction terms in the models, as the datasets are too limited to cover all possible combinations of functional groups/structural elements. When the model is applied to the test data set, a negative Q^2 was calculated, implying that functional groups/structural fragments – including all multiple linear regression models for all individual *in vitro* assays – on average did not perform well in predicting toxicity for chemicals outside the training dataset. However, large differences exist in the predictive power of the models across *in vitro* assays.

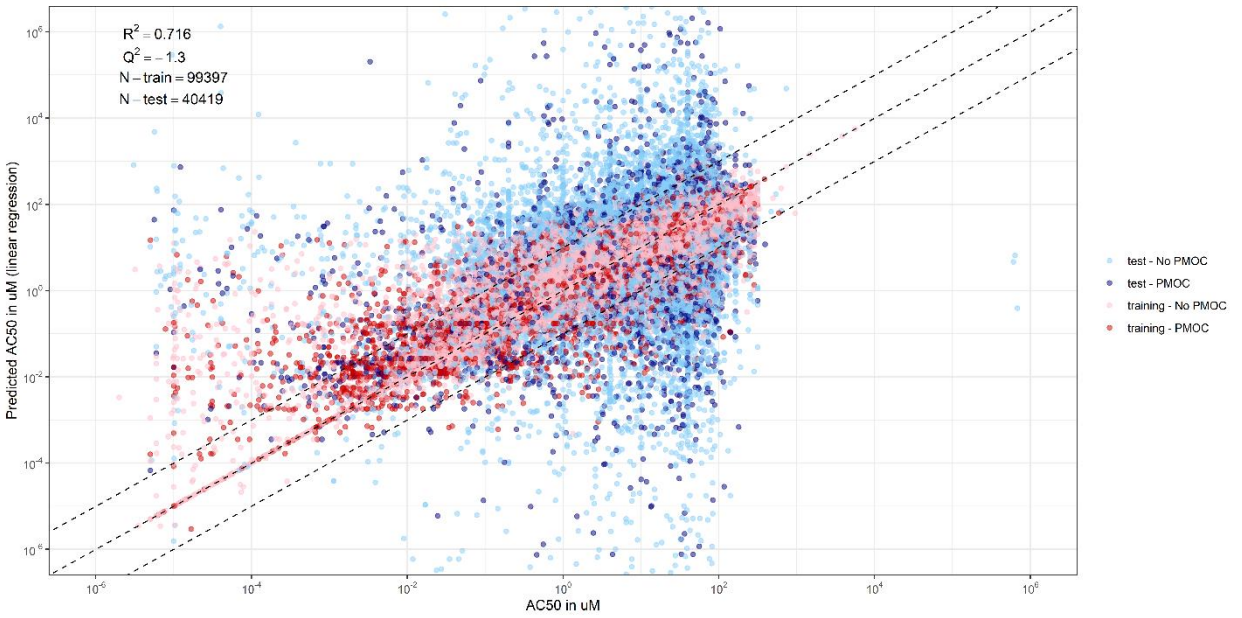


Figure 13: Predicted toxicity (AC_{50} in µM) by the multiple linear regression model versus observed (experimental) toxicity, based on structural fragments/functional groups and topological parameters, for both the training dataset and test dataset. The middle dashed line represents the 1:1 ratio. The outer dashed lines represent the 1:5 and 5:1 ratio.

In the figure below, the adjusted R^2, from the multiple linear regression model, based on structural fragments/functional groups (Figure 14), is plotted against the variance explained (%) by the Random Forest model modelled for all *in vitro* assays separately. The overall correlation between the R^2 from the multiple linear regression analysis and the variance explained by the Random Forest model is very low, which implies that the correlation between functional groups and toxicity can be better described by linear regression (taking into account all functional groups and chemicals) than by Random Forest, possibly omitting any correlation between structural fragments/functional groups and chemicals.

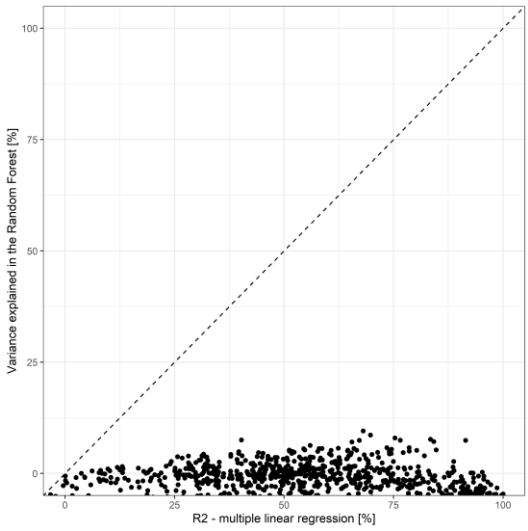


Figure 14: The adjusted R^2, from the multiple linear regression model, based on organic structural fragments/functional groups, plotted against the variance explained (%) by the Random Forest model, based on aforementioned descriptors, modelled for all in vitro assays separately. The dashed line represents the 1:1 ratio.

3.3 Physicochemical descriptors

3.3.1 Principal component analysis

Principal component analysis (PCA) was conducted to explore the characteristics of the physicochemical descriptors (log K_{OC}, log K_{OW}, biodegradation rate, vapor pressure, and molecular weight) prior to two analyses (Random Forest Analysis and multiple linear regression analysis), as these descriptors were continuous (in contrast to the binary structural fragment descriptors). The PCA was performed based on the scaled values of the six variables, i.e., the five physicochemical descriptors, and the log-transformed AC_{50} values.

Prominent principal components (PCs) emerged, with PC1 explaining 30.98%, PC2 explaining 24.77%, and PC3 explaining 16.69%, as shown in Figure 15. The score plots did not indicate any grouping among the data sets or differences between PMOCs and non-PMOCs groups, suggesting the six variables did not have enough information to be categorized into multiple groups. The biplot with PC1 and PC2 (Figure 15A) indicates that the molecular weight was positively associated with PC2. The other five factors had negative associations with PC2; the log K_{OC} and log K_{OW} were negatively related to PC1, and the log-transformed AC_{50} values were positively associated with PC1, together with the vapor pressure and the biodegradation rate. The results of PCA revealed how the six variables influenced each principal component and that a part of the variables had similar information. Further analysis would be required to investigate more precise relationships between the variables.

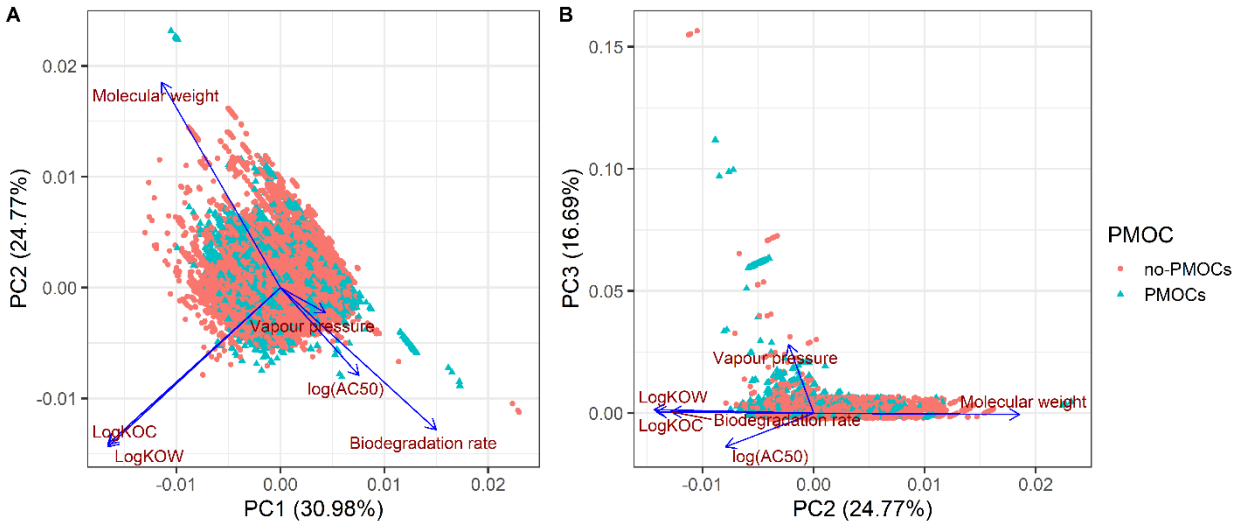


Figure 15: Principal component analysis (PCA) biplots based on PC1 and PC2 (A) and PC2 and PC3 (B). Each data point corresponds to a data record consisting of the six types of data. The colors of data points represent the chemical groups (PMOCs or no-PMOCs), according to the classification described in section 2.1.

3.3.2 Random Forest

In general, the Random Forest analysis for all 603 *in vitro* assays separately, including the five most important predictive physicochemical descriptors (log K_{oc}, log K_{ow}, biodegradation rate, vapor pressure and molecular weight) as explanatory variables resulted in explaining 15.78% (median: 14.37, S.E.: 0.03) of all variance in the toxicity data (AC_{50}s). The highest percentage of variances explained were determined for OT_ER_ERaERa_0480[5] (nuclear receptor type) (83.98%), while the lowest percentage variance explained by the Random Forest model were found for NVS_ENZ_hEphA1_Activator[6] (-52.78). Figure 16 shows the predicted effect concentrations (log_{10} AC_{50}s) – predicted by both Random Forest and multiple linear regression - plotted against the observed effect concentrations, based on Random Forest analysis, taking the aforementioned five physicochemical parameters as explanatory variables, for both signal directions separately. In total, 93% of all individual predicted AC_{50}s were within a factor 5 (which equals to approximately 5% of the complete toxicity data range) of the observed AC_{50}s; 2.4% of the predicted datapoints were more than a factor five below the observed datapoints (underestimated), while 4.6% were more than a factor five above the observed data (overestimated). 0.05% of the predicted datapoints were a perfect fit, which may indicate overfitting of the model.

3.3.3 Multiple linear regression analysis

In general, the multiple linear regression analysis, including the five most important predictive physicochemical descriptors (log K_{oc}, log K_{ow}, biodegradation rate (half-life in days), vapor pressure and molecular weight) as explanatory variables resulted in explaining 14.5% (median: 12%, S.E.: 0.01%) of all variance in the toxicity data (AC_{50}s), based on the adjusted R^2. The highest % of variances explained were determined for an *in vitro* assay focusing on a background reporter gene: TOX21_GR_BLA_Agonist_ch1 (62.6%), while the lowest % variance explained by the

[5] CompTox Chemicals Dashboard (epa.gov)
[6] CompTox Chemicals Dashboard (epa.gov)

multiple linear regression model was found for another *in vitro* assay focusing on a background reporter gene: ATG_M_32_CIS_dn (-15%). Figure 16 shows the predicted effect concentrations (log_{10} AC_{50}s) plotted against the observed effect concentrations, based on the multiple linear regression model, taking the aforementioned five physicochemical parameters as explanatory variables (Equation 2). In total, 69% of all individual predicted AC_{50}s lied within a factor 5 of the observed AC_{50}s; 15.1% of the predicted datapoints were more than a factor five below the observed datapoints (underestimated), while 15.9% were more than a factor five above the observed data (overestimated). None of the predicted datapoints were a perfect fit.

Based on the total dataset, linear regression analysis (R^2s) had a lower predicting power compared to the random forest analyses for the same toxicity tests. This indicates that there probably is no linear relationship between physicochemical descriptors and the response variable as the modelling exercise was based on single linear responses only, disregarding any interaction between parameters or non-linear responses. The linear regression coefficients for the 603 analyzed *in vitro* assays, taking standardized AC_{50} values as response variables, were negative for log K_{oc}, log K_{ow}, and molecular weight, with median regression coefficients of -0.04, -0.03, and -0.002 respectively, based on the absolute values (Figure 17). The median regression coefficients for log K_{ow}, log K_{oc}, and biodegradation rate all significantly differed from zero (p-value < 0.05; One sample t-test). Only the median value of regression coefficients for biodegradation rate (half-life in days), and vapor pressure were positive (0.44 and 0.02, respectively), however median values for both vapor pressure and molecular weight did not differ significantly from zero (p > 0.05; One sample t-test). Hence, in general, the majority of the investigated physicochemical properties were inversely related to toxicity (expressed as AC_{50}, with a higher AC_{50} indicating a lower toxicity and a lower K_{oc}/K_{ow} indicating a higher mobility), albeit to varying degrees. However, biodegradation rate (half-life in days) in general was proportionally related to AC_{50} values, implying that more persistent chemicals (higher half-life) tend to be less toxic (higher AC_{50}). These observations with respect to mobility and persistence are in line with conclusions drawn in the previous BTO-report on PMOC toxicity.

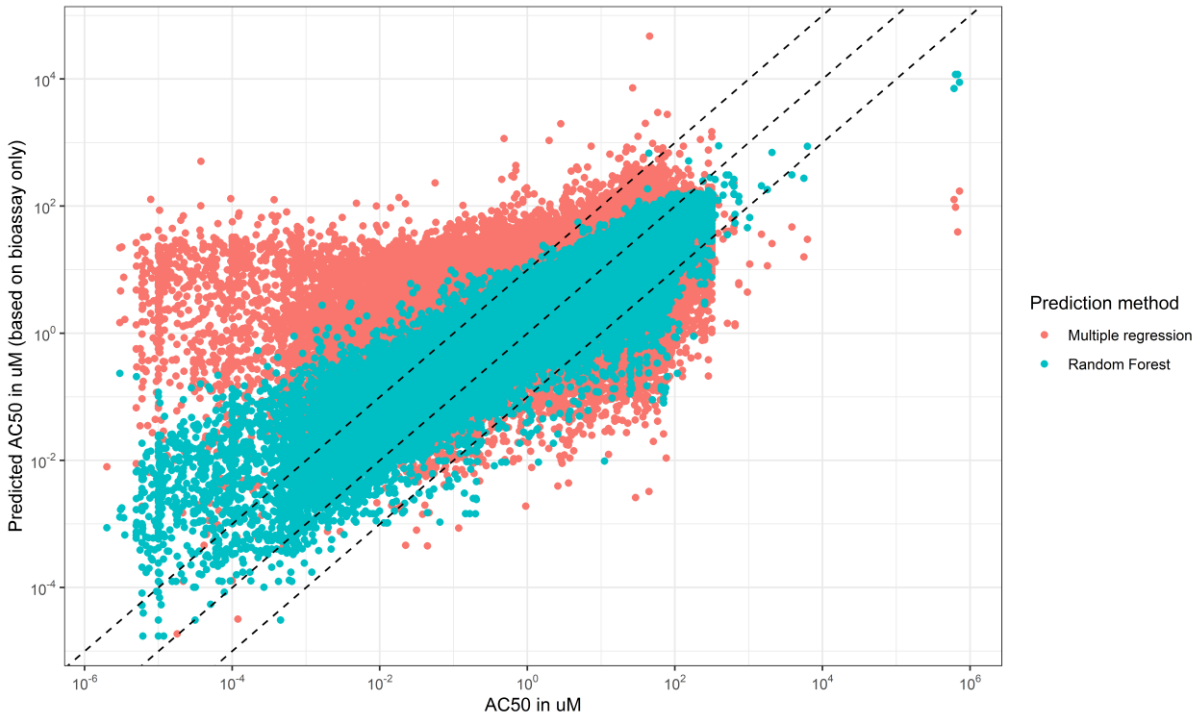


Figure 16: Predicted toxicity (AC_{50} in µM) by the multiple linear regression model and the Random Forest model versus observed toxicity, based on five physicochemical parameters (Log K_{ow}, Log K_{oc}, biodegradation rate, vapor pressure, and molecular weight), clustered per individual in vitro *assay. The dashed lines represent the 1:5 line.*

When standardizing the physicochemical descriptors by subtracting the values by the mean value per toxicity test and dividing the result by the standard deviation, rescaling the data to have a mean of zero and a standard deviation of one, we see a similar pattern, albeit more spread out than the unstandardized, absolute values . The influence of physicochemical descriptors that normally cover a large range of values (such as molecular weight, boiling point and biodegradation rate) become more apparent, as the values on the y-axis now inform us about the change in response (toxicity) when increasing the physicochemical descriptor by one standard deviation (i.e., a relative increase in descriptor value, rather than an absolute increase in descriptor value).

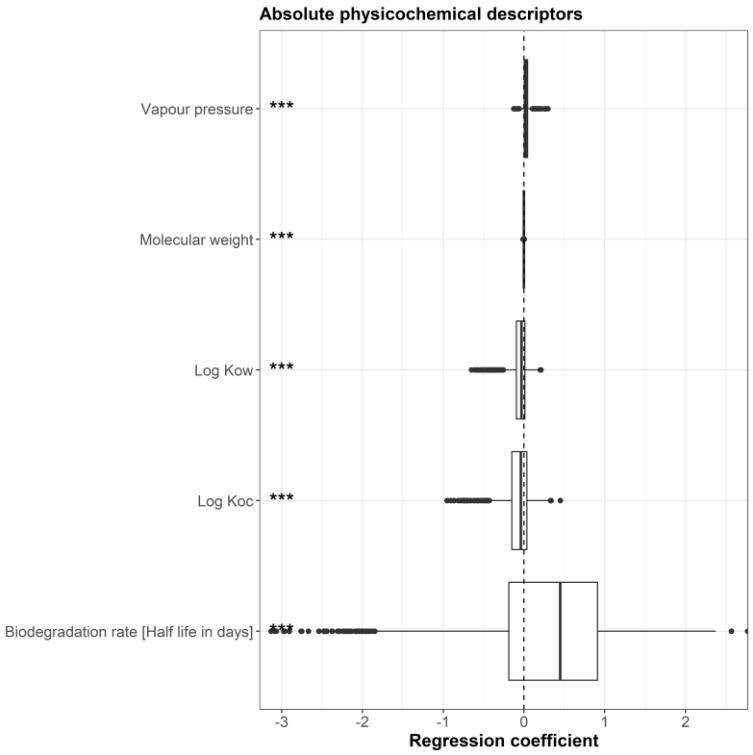


*Figure 17: Boxplots depicting the distribution of regression coefficients (minimum value, 25th percentile, median value, 75th percentile and maximum value) for vapor pressure (in mmHg), molecular weight (in g/mol), log octanol-water partition coefficient (log K_{ow}), log sorption coefficient to organic carbon (log Koc), boiling point (in °C), and biodegradation rate (half-life in days) resulting from the multiple linear regression analysis for 603 in vitro assays. ***; $\mu \neq 0$, $p < 0.001$, **; $\mu \neq 0$, $p < 0.01$, *; $\mu \neq 0$, $p < 0.05$, -; $\mu = 0$, $p > 0.05$. Parameters with significantly similar distributions of regression coefficients were assigned a similar letter.*

In the figure below, the adjusted R^2, from the multiple linear regression model, based on the five physicochemical descriptors (log K_{oc}, log K_{ow}, vapor pressure, molecular weight, and biodegradation (half-life in days)) (Figure 18), is plotted against the variance explained (%) by the Random Forest model modelled for all *in vitro* assays separately.

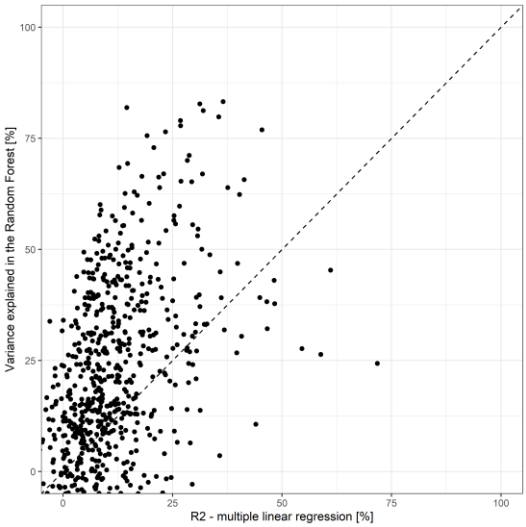


Figure 18: The adjusted R^2, from the multiple linear regression model, based on five physicochemical descriptors (log K_{oc}, log K_{ow}, vapor pressure, molecular weight, and biodegradation (half-life in days)), plotted against the variance explained (%) by the Random Forest model, based on aforementioned descriptors, modelled for all in vitro assays separately.

3.4 *In vitro* assay types

After grouping *in vitro* assays based on five category types (intended target family, technological target type, assay design type, signal direction, and organism-tissue combination), Random Forest analysis and the multiple linear regression analysis was based on the five physicochemical descriptors, as these showed to be the most consistent explanatory parameters, compared to taking structural fragments/functional groups as explanatory variables. Table 2 shows an overview of all results concerning both the Random Forest analysis and the multiple linear regression analysis. Overall, the best fits for both the Random Forest analysis and multiple linear regression analysis – when including physicochemical properties as explanatory variables – were obtained when categorizing the *in vitro* assays based on technological target type, which may be explained by an overlap in specific toxic modes of actions and specific target types used in the categorization of *in vitro* assays. Random Forest analysis in this case explained on average 55.27% of all variance in toxicity (AC_{50}) data, while multiple linear regression analysis on average explained 13.3% of all variance in toxicity (AC_{50}) data. However, when looking specifically at the percentage of predictions within a factor of five of experimental observations, the best Random Forest fit was obtained when categorizing assays based on organism-tissue combination, implying that sometimes interspecies differences may be greater and more important than inter-effect differences within a species. More details with respect to the results for each category can be found in Appendix I.I.

Table 2: Summary showing all results concerning both the Random Forest model and the multiple linear regression model, based on five physicochemical descriptors, when grouping in vitro assays based on five in vitro assay category types. Summary statistics include average % of variance explained, median % of variance explained (± S.E.), highest variance explained, lowest variance explained, % of data points overfitted, % of data points over- or underestimated and % of observations within a factor of five of the predicted values. The third table shows the number of data point overestimated, underestimated, overfitted, and predictions within a factor of five of the observed data when clustering the data based on assay categories for the test and training dataset when applying the linear regression model.

	Random Forest model						Linear regression model					
Category	**Mean variance explained (%)**	**Median variance explained (%)**	**Lowest variance explained (%)**	**Highest variance explained (%)**	**Target with lowest variance**	**Target with highest variance explained**	**Mean variance explained (%)**	**Median variance explained (%)**	**Lowest variance explained (%)**	**Highest variance explained (%)**	**Target with lowest variance**	**Target with highest variance explained**
Intended target family	28,01% (±0,06%)	31,22%	-26,07%	84,70%	Membrane protein	Neurodevelopment	9,74% (±0,02%)	9,01%	-3,98%	32,04%	Membrane protein	Mitochondria
Technological target type	55,27% (±0,04%)	54,85%	27,31%	84,70%	Cellular	Electrical activity	13,30% (±0,02%)	10,19%	0,95%	31,19%	DNA	Molecular messenger
Assay design type	42,73% (±0,06%)	38,25%	15,15%	84,70%	Enzyme reporter	Functional reporter	12,69% (±0,02%)	11,30%	1,52%	32,04%	Biochemical reporter	Respirometric reporter
Signal direction	41,71% (±0,00%)	41,71%	41,63%	41,78%	Gain	Loss	6,36% (±0,01%)	6,36%	4,29%	8,43%	Gain	Loss
Organism tissue	37,05% (±0,07%)	41,36%	-25,56%	83,24%	Human brain	Rat cortical	10,06% (±0,02%)	8,99%	-10,06%	31,81%	Guinea pig spleen	Rat kidney

	Random Forest model				Linear regression model			
Category	% Overestimated (> 5x)	% Underestimated (>5x)	% Overfitted	% Observations within a factor 5 of predicted values	% Overestimated (> 5x)	% Underestimated (>5x)	% Overfitted	% Observations within a factor 5 of predicted values
Intended target family	8,76%	6,95%	0,07%	84,28%	15,91%	21,25%	0,00%	62,48%
Technological target type	9,25%	7,40%	0,00%	83,35%	16,14%	19,85%	0,00%	64,01%
Assay design type	9,81%	7,77%	0,00%	82,42%	16,14%	21,75%	0,00%	62,11%
Signal direction	11,50%	9,35%	0,00%	79,15%	16,41%	21,82%	0,00%	61,78%
Organism-tissue combination	8,09%	6,37%	0,07%	85,54%	15,89%	21,01%	0,00%	63,11%

	Test				**Training**			
Category	Overestimated	Underestimated	Overfitted	Between	Overestimated	Underestimated	Overfitted	Between
Intended target family	15,7% (±4,0%)	12,6% (±7,5%)	0,0% (±0,00%)	71,7% (±11,0%)	17,2% (±6,2%)	16,0% (±9,6%)	0,00% (±0,00%)	66,8% (±12,9%)
Technological target type	18,4% (±5,9%)	19,2% (±12,9%)	0,00% (±0,00%)	62,4% (±18,6%)	19,1% (±6,0%)	19,8% (±12,8%)	0,00% (±0,00%)	61,1% (±18,1%)
Assay design type	17,0% (±4,4%)	18,4% (±7,4%)	0,00% (±0,00%)	64,6% (±11,3%)	16,8% (±5,4%)	19,7% (±6,7%)	0,00% (±0,00%)	63,6% (±11,6%)
Signal direction	17,8% (±2,2%)	23,5% (±12,8%)	0,00% (±0,00%)	58,7% (±15,0%)	17,8% (±2,0%)	23,6% (±12,6%)	0,00% (±0,00%)	58,6% (±14,6%)
Organism-tissue combination	16,0% (±4,8%)	16,2% (±10,6%)	0,00% (±0,00%)	67,9% (±14,9%)	17,4% (±6,7%)	16,2% (±9,9%)	0,00% (±0,00%)	66,4% (±15,3%)

4 Overall discussion

The present study was a continuation of the previous project "BTO 2023.60 - *Zijn persistente mobiele stoffen minder giftig*?", in which correlations between physicochemical descriptors (i.e. log K_{ow}, log K_{oc}, biodegradation rate, vapor pressure, and molecular weight) were explored. A significant inverse correlation between mobility (log K_{oc}) and toxicity was observed, indicating that a higher mobility results in a lower toxicity of the compound. However, in this previous study, all *in vitro* assay endpoints were analyzed individually and no special attention was given to differences and similarities with respect to toxicity when clustering *in vitro* assays based on *in vitro* assay type and toxicological mechanism. Additionally, only a relatively small subset of water relevant PMOCs were used as input in the modelling exercises.

The study presented in the current report aimed to gain a deeper understanding of PMOC toxicity, allowing the signaling of new and potentially hazardous PMOCs that may emerge in the aquatic environment, based on their chemical structures and physicochemical properties, by looking into the predictive power of both Random Forest and multiple linear regression models for a large set of toxicity data (including PMOCs). Additionally, to allow for sufficiently large datasets for model training and *in vitro* data were clustered based on assay type, based on annotations as described in paragraph 2.3 (intended target family, technological target type, assay design type, signal direction and organism/tissue combination). As molecular structure (i.e. structural alerts or functional groups) has been associated with bioactivity, in the present study structural fragments and structural fragments/functional groups were taken as predictors in both the multiple linear regression analyses, as well as the Random Forest models. Random Forest performed relatively well in predicting AC_{50} values from ToxCast, when taking physicochemical descriptors as explanatory variables for some individual assays (mean variance explained: 15.8%, max: 84%, when looking at individual assay types), compared to multiple linear regression analysis (mean R^2: 14.5% , max R^2: 62.9%), implying that any correlations between five physicochemical parameters and toxicity are not linear. Although R^2s were higher when taking structural fragments as explanatory variables in both models, Q^2s remained low, implying that multiple linear regression analysis also did not perform well when using the models to predict toxicity for compounds outside the training dataset, based on structural fragments. In general, the Random Forest analysis performed poorly, compared to the multiple linear regression analysis. This poor performance may have multiple explanations. Firstly, individual datasets may have been too small, although a cut-off point of 50 data rows was used as a criterion. Bigger datasets did provide a better model fit (i.e. a higher percentage of the variance explained by the Random Forest model), when analyzing all *in vitro* assays separately. However, the highest variance explained by the Random Forest model was only 6%, still indicating a poor fit of the model when including structural fragments and functional groups as explanatory variables. Secondly, the models may contain a large proportion of irrelevant features (functional groups), in which case the model struggles to learn the underlying patterns in the data. As the initial dataset in the present study contains a maximum of 396 features (structural properties and functional groups), the model may have difficulties in identifying specific functional groups that may be useful in classifying the data based on AC_{50}. Finally, the AC_{50} data itself is expected to be noisy. These toxicity data have been collected by multiple laboratories, scientists and analysts throughout the years, using different protocols. Although multiple flags (warning assigned by ToxCast – See paragraph 2.2), have been found in the data, these were in this research not used as criteria in truncation if the data, as these warnings covered over 50% of the complete dataset. The noise in the data may be reduced by standardizing the data (equation 3). However, standardizing the data makes the data less easy to interpret, as it adds complexity to the data; retransforming the data to interpret the results requires an extra step. Furthermore, the analyses performed in the present study may be less sensitive to noise in the data when using a

categorical (low toxicity, medium toxicity, and high toxicity) response variable rather than a continuous response variable.

Although a bit better than the Random Forest model, the multiple linear regression analysis also did not perform great when including functional groups/structural fragments as explanatory variables. Overfitting occurs more often, especially in cases where the number of data entries is limited, as we included hundreds of different dummy-variables (0-1) (See 2.1), instead of five continuous variables. Predictions from such a rank-deficient fit (i.e. a fit on a dataset for which not enough observations are available per factor level) may be underestimating or overestimating. Additionally, in order to provide a reliable toxicity prediction, large amounts of data are needed to cover a wide variety of structural properties and functional groups, which was not always the case for all *in vitro* assays. Although, especially for smaller datasets, no good fits were obtained for all *in vitro* assays, more reliable results for regression models including functional groups as explanatory variables may be obtained when dividing AC_{50} in two or three toxicity classes, as is typically done in commercial read-across and QSAR software (Chakravarti et al., 2012; Ciallella et al., 2022; Krewski et al., 2020; Russo et al., 2019). Additionally, non-linear models or the inclusion of interaction terms (combinations of structural fragments, which may have a synergistic toxic effect) may also increase the fit of linear regression models. According to work by Calleja et al. (1994b) non-linear models taking molecular structure as explanatory variables appear to have a better fit than linear models. In the current report, models were based on functional groups, represented as dummy variables, representing the absence or presence of a certain functional group/structural fragment (See paragraph 2.1). Regression coefficients associated with this qualitative information tells us the average increase (or decrease) of (log-transformed) toxicity in case the functional group is present. In the case of a significant decrease in toxicity associated with a dummy variable, the functional variable can be considered de-activating.

In the current report and modelling exercise AC_{50} data from ToxCast were used as response variables in both the Random Forest analysis, as well as the multiple linear regression analysis. Although these data provide information on bioactivity and potential mechanistic pathways that they act on, these data do not indicate hazard or an adverse effect *in vivo* (Huang et al., 2016). These data are typically used to prioritize chemicals based on expected bioactivity when *in vivo* toxicity data are lacking. To evaluate the potential hazard of a compound for which toxicity data are lacking, ToxCast (activity) data may be linked to biological events through adverse outcome pathways (AOPs). An AOP is a construct describing a sequential chain of causally linked biological events at different levels that lead to adverse effects. ToxCast data (on *in vitro* assays) may be clustered based on assay data corresponding to molecular initiating events (MIEs) in an AOP framework for a certain adverse effect as was done for thyroid disease by Nelms et al. (2018). There, ToxCast data were combined with chemical structure data (structural alerts) from OECD QSAR Toolbox and clustered corresponding to a set of MIEs within the AOP for hepatic steatosis.

5 Conclusions

In the present study, we used multiple linear regression modelling and Random Forest analysis to explore whether toxicity (AC_{50} values) can be predicted based on physicochemical characteristics and structural properties of chemicals, and to evaluate if sub-setting *in vitro* assay data based on assay characteristics aid in predicting toxicity (bioactivity) . In general, multiple linear regression explained more variance in toxicity when using functional groups/structural fragments (median over all individual *in vitro* assay endpoints: 54.5%) rather than physicochemical descriptors (median over all individual *in vitro* assay endpoints: 14.5%).

In addition to exploring the predictive power of Random Forest models and multiple linear regression models based on physicochemical characteristics, in the current study, *in vitro* assays were clustered based on assay type (intended target family, technological target type, assay design type, signal direction and organism-tissue combination), to investigate the impact of assay annotations on the predictability of AC_{50}s. Grouping based on technological target type resulted in the highest predictability of toxicity in both models, with an average of 55.27% of variance explained in the Random Forest model, and an average of 13.30.% explained in the multiple linear regression model). Although grouping of *in vitro* assays considerably increased the number of data rows in the training datasets used in the modelling exercises, the percentages of variances explained by both models when taking the physicochemical descriptors as explanatory variables, decreased significantly compared to grouping based on individual *in vitro* assays. Additionally, no considerable differences in the average percentages explained by the Random Forest model and multiple linear regression model could be observed when grouping the *in vitro* assays based on the five categories, based on assay annotations (intended target family, technological target type, assay design type, signal direction, organism-tissue combination). This was likely due to intercorrelation between *in vitro* assays within the different categories (e.g., the gross majority of *in vitro* assays in the neurodevelopment intended target family also tend to be in *electrical activity* technological target type). This implies that none of the individual categories included in this study to cluster the *in vitro* assays were suitable for the prediction of toxicity of chemicals. Nevertheless, for the large majority of individual *in vitro* assays as well as for groups of *in vitro* assays based on aforementioned grouping criteria, more than 80% of all predicted data points (by Random Forest based on physicochemical characteristics and by multiple linear regression based on functional groups) fell within a factor of five of the experimental data points, implying that the physicochemical descriptors and functional groups of compounds may still provide enough information to categorize toxicity into toxicity classes (based on AC_{50}s).

Although in the present study we gained a deeper understanding of (PMOC) toxicity using the ToxCast database, the aforementioned limitations (i.e. data limitation, rank deficiency and intercorrelation) of the models used hamper their applicability in the (drinking) water sector. *In vitro* assay endpoints included in the ToxCast database vary considerably in e.g. target type, tissue tested and assay design type and structural elements in itself may not solely explain differences in activity in these assays or there are still insufficient data available to derive reliable correlations. In contrast, physicochemical descriptors (especially the ones related to persistence and mobility) as explanatory variables in many cases (assay endpoints) provided sufficiently reliable predictions for activity in the assay. Just like in the previous study (See report BTO 2023.060), more mobile, more persistent compound tend to be less toxic. Note, however, that the reliability of the predictors decreased when assay endpoints were clustered based on one of the five aforementioned categories, implying that variation in AC_{50}s between *in vitro* assay endpoint are greater than the variation between chemicals within assay endpoints. However, structural elements and physicochemical descriptors of chemicals for models of a subset of *in vitro* assay endpoint did provide sufficient information to predict AC_{50}s and AC_{50} classes (i.e. 'low', 'medium', 'high'). For this reason, in future research we foresee the development of a tool to predict toxicity classes, rather than exact toxicity (AC_{50}) values for this particular subset of endpoints.

6 Bibliography

Calleja, M., Geladi, P., & Persoone, G. (1994a). Modelling of human acute toxicity from physicochemical properties and non-vertebrate acute toxicity of the 38 organic chemicals of the MEIC priority list by PLS regression and neural network. *Food and Chemical Toxicology, 32*(10), 923-941.

Calleja, M., Geladi, P., & Persoone, G. (1994b). QSAR models for predicting the acute toxicity of selected organic chemicals with diverse structures to aquatic non-vertebrates and humans. *SAR and QSAR in Environmental Research, 2*(3), 193-234.

Chakravarti, S. K., Saiakhov, R. D., & Klopman, G. (2012). Optimizing predictive performance of CASE Ultra expert system models using the applicability domains of individual toxicity alerts. *Journal of chemical information and modeling, 52*(10), 2609-2618. doi:10.1021/ci300111r

Ciallella, H. L., Russo, D. P., Sharma, S., Li, Y., Sloter, E., Sweet, L., Huang, H., & Zhu, H. (2022). Predicting prenatal developmental toxicity based on the combination of chemical structures and biological data. *Environmental science & technology, 56*(9), 5984-5998.

Cocchi, M., De Benedetti, P. G., Seeber, R., Tassi, L., & Ulrici, A. (1999). Development of Quantitative Structure–Property Relationships Using Calculated Descriptors for the Prediction of the Physicochemical Properties (n D, ρ, bp, ε, η) of a Series of Organic Solvents. *Journal of Chemical Information and Computer Sciences, 39*(6), 1190-1203.

Eriksson, L., Jaworska, J., Worth, A. P., Cronin, M. T., McDowell, R. M., & Gramatica, P. (2003). Methods for reliability and uncertainty assessment and for applicability evaluations of classification-and regression-based QSARs. *Environmental Health Perspectives, 111*(10), 1361-1375.

European Chemicals Agency. (2014). *Illustrative example with the OECD QSAR Toolbox workflow Part 1: Introductory note*. Retrieved from

Feshuk, M., Kolaczkowski, L., Dunham, K., Davidson-Fritz, S. E., Carstens, K. E., Brown, J., Judson, R. S., & Paul Friedman, K. (2023). The ToxCast pipeline: updates to curve-fitting approaches and database structure. *Front Toxicol, 5*, 1275980. doi:10.3389/ftox.2023.1275980

Filer, D. L., Kothiya, P., Setzer, W. R., Judson, R. S., & Martin, M. T. (2014). The ToxCast analysis pipeline: An R package for processing and modeling chemical screening data. *US Environmental Protection Agency: http://www. epa. gov/ncct/toxcast/files/MySQL% 20Database/Pipeline_Overview. pdf*.

Ghisi, R., Vamerali, T., & Manzetti, S. (2019). Accumulation of perfluorinated alkyl substances (PFAS) in agricultural plants: A review. *Environmental Research, 169*, 326-341. doi:https://doi.org/10.1016/j.envres.2018.10.023

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. doi:https://doi.org/10.1016/j.tox.2013.08.012

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*(1-2), 1-12.

Hardy, A., Benford, D., Halldorsson, T., Jeger, M. J., Knutsen, H. K., More, S., Naegeli, H., Noteborn, H., & Ockleford, C. (2017). Guidance on the use of the weight of evidence approach in scientific assessments. *EFSA Journal, 15*(8), e04971.

Hemmerich, J., Troger, F., Füzi, B., & G, F. E. (2020). Using Machine Learning Methods and Structural Alerts for Prediction of Mitochondrial Toxicity. *Mol Inform, 39*(5), e2000005. doi:10.1002/minf.202000005

Huang, R., Xia, M., Sakamuru, S., Zhao, J., Shahane, S. A., Attene-Ramos, M., Zhao, T., Austin, C. P., & Simeonov, A. (2016). Modelling the Tox21 10K chemical profiles for in vivo toxicity prediction and mechanism characterization. *Nature Communications, 7*(1), 10425. doi:10.1038/ncomms10425

Jaylet, T., Coustillet, T., Jornod, F., Margaritte-Jeannin, P., & Audouze, K. (2023). AOP-helpFinder 2.0: Integration of an event-event searches module. *Environment international, 177*, 108017. doi:https://doi.org/10.1016/j.envint.2023.108017

Jonker, M. T., & Van der Heijden, S. A. (2007). Bioconcentration factor hydrophobicity cutoff: An artificial phenomenon reconstructed. *Environmental science & technology, 41*(21), 7363-7369.

Krewski, D., Andersen, M. E., Tyshenko, M. G., Krishnan, K., Hartung, T., Boekelheide, K., Wambaugh, J. F., Jones, D., Whelan, M., & Thomas, R. (2020). Toxicity testing in the 21st century: progress in the past decade and future perspectives. *Archives of toxicology, 94*, 1-58.

Kwak, S. G., & Kim, J. H. (2017). Central limit theorem: the cornerstone of modern statistics. *Korean journal of anesthesiology, 70*(2), 144-156.

Mackay, D., Arnot, J., Petkova, E., Wallace, K., Call, D. J., Brooke, L., & Veith, G. (2009). The physicochemical basis of QSARs for baseline toxicity. *SAR and QSAR in Environmental Research, 20*(3-4), 393-414.

Mansouri, K., & Williams, A. (2017). *QMRF - Title: KOC model for the soil adsorption coefficient prediction from OPERA models*.

Nelms, M. D., Mellor, C. L., Enoch, S. J., Judson, R. S., Patlewicz, G., Richard, A. M., Madden, J. M., Cronin, M. T., & Edwards, S. W. (2018). A mechanistic framework for integrating chemical structure and high-throughput screening results to improve toxicity predictions. *Computational Toxicology, 8*, 1-12.

Neumann, M., & Schliebner, I. (2019). *Protecting the sources of our drinking water: The criteria for identifying persistent, mobile and toxic (PMT) substances and very persistent and very mobile (vPvM) substances under EU Regulation REACH (EC) No 1907/2006*. Retrieved from

Phuong, J., Sipes, N., Truong, L., Connors, K., Houck, K., & Martin, M. (2014). ToxCast assay annotation version 1.0 data user guide. *US Environmental Protection Agency*, 36.

Randić, M. (2001). The connectivity index 25 years after. *Journal of Molecular Graphics and Modelling, 20*(1), 19-35. doi:https://doi.org/10.1016/S1093-3263(01)00098-5

Russo, D. P., Strickland, J., Karmaus, A. L., Wang, W., Shende, S., Hartung, T., Aleksunes, L. M., & Zhu, H. (2019). Nonanimal models for acute toxicity evaluations: Applying data-driven profiling and read-across. *Environmental Health Perspectives, 127*(4), 047001.

Ryan, N., & Becker, R. (2017). A user's guide for accessing and interpreting ToxCast data: Bayer.

Sarstedt, M., Ringle, C. M., & Hair, J. F. (2021). Partial least squares structural equation modeling *Handbook of market research* (pp. 587-632): Springer.

Schultz, T. W., Diderich, R., Kuseva, C. D., & Mekenyan, O. G. (2018). The OECD QSAR toolbox starts its second decade. *Computational Toxicology: Methods and Protocols*, 55-77.

Society for the Advancement of Adverse Outcome Pathways (SAAOP). (2023). AOP Wiki. Retrieved from https://aopwiki.org/

Team, R. C. (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL

https://www.R-project.org/.

U.S. Environmental Protection Agency (EPA). (2014). *US EPA TOXCAST DATA RELEASE ASSAY QUALITY SUMMARY OCTOBER 2014*. Retrieved from

U.S. EPA. (2015). *ToxCast & Tox21 Summary Files from invitrodb_v3*. Retrieved from: https://www.epa.gov/chemical-research/toxicity-forecaster-toxcasttm-data

Université Paris Cité. (2023). AOP-helpFinder 2.0. Retrieved from https://aop-helpfinder.u-paris-sciences.fr/

US EPA. (2022). EPI Suite™-Estimation Program Interface. Retrieved from https://www.epa.gov/tsca-screening-tools/epi-suitetm-estimation-program-interface

Weaver, S., & Gleeson, M. P. (2008). The importance of the domain of applicability in QSAR modeling. *Journal of Molecular Graphics and Modelling, 26*(8), 1315-1326. doi:https://doi.org/10.1016/j.jmgm.2008.01.002

Yap, C. W. (2011). PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of computational chemistry, 32*(7), 1466-1474.

I Appendix : Individual Random Forest model and linear regression model by assay type.

I.I Intended target family

The intended target family attempts to represent the common targets across assay endpoints. These families pertain to gene families and include morphological and cell cycle concepts (U.S. EPA, 2015).

Random Forest model

In general, the Random Forest analysis, including the five most important predictive physicochemical descriptors (log K_{oc}, log K_{ow}, biodegradation rate, vapor pressure and molecular weight) as explanatory variables resulted in explaining 28.01% (median: 31.22%, S.E.: 0.06) of all variance in the toxicity data (AC_{50}s) when categorizing *in vitro* assays based on intended target family. Overall, when grouping *in vitro* assays based on intended target family, the highest percentage of variances explained by the Random Forest model were determined for *in vitro* assays related to neurodevelopment (84.74%), while the lowest % variance explained by the Random Forest model were found for *in vitro* assays related to membrane proteins (-26.07%) (Table 1). This implies that physicochemical descriptors included in the present study correlated strongly with neurodevelopmental activity AND that variation in AC_{50} values in the neurodevelopment dataset was proportional to or higher than the variation of physicochemical descriptors from chemicals in the dataset. Furthermore, this also implies that using a Random Forest model to predict toxicity for the subset of chemicals and in vitro assay endpoint focusing on membrane proteins result in a prediction that is worse than taking the average of all AC_{50} values. Figure 19 shows a heatmap visualizing to which extent the five physicochemical descriptors of interest correlate with toxicity for assays within one of the intended target families. The increase in MSE (%IncMSE) (Equation 1) corresponds to the extent to which the physicochemical parameter explains the variance in the Random Forest model.

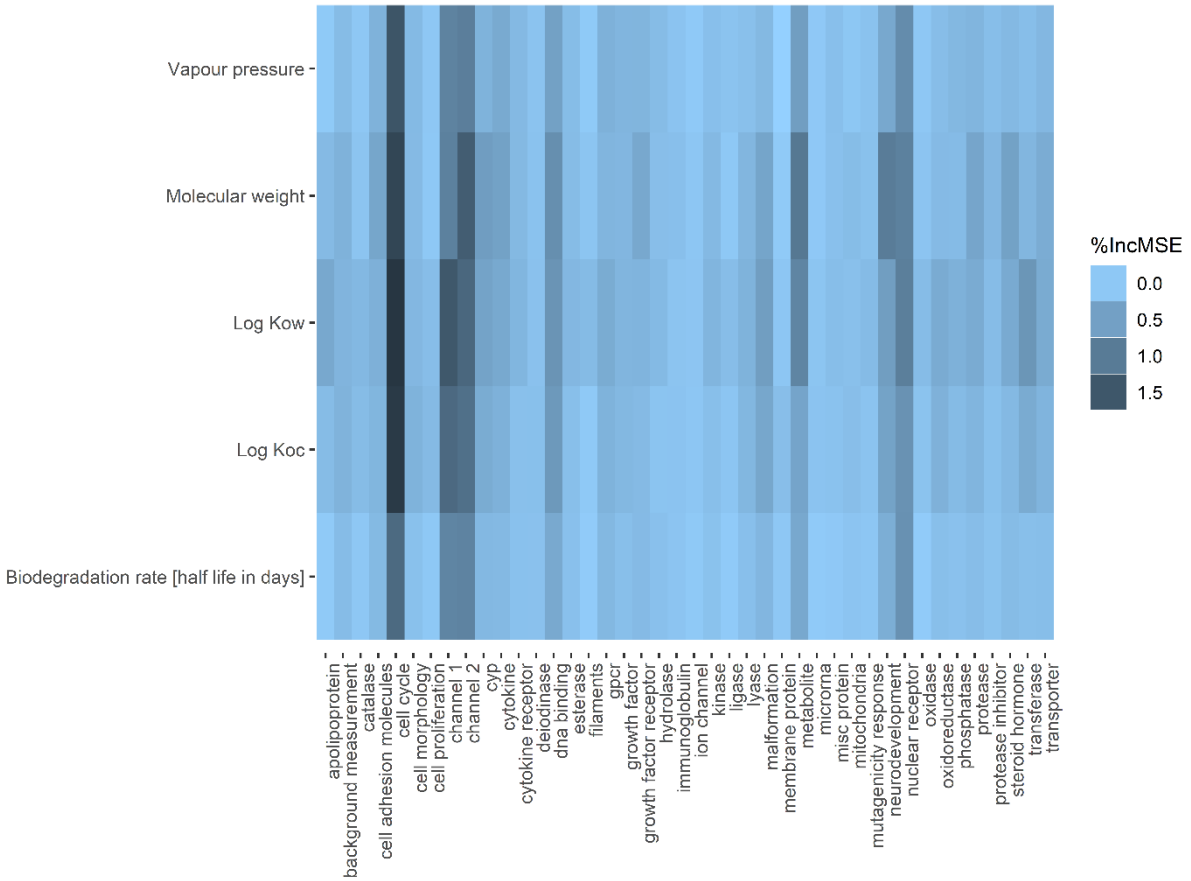


Figure 19: Heatmap visualizing the increase in MSE (= to which extent the variable explains the variance in toxicity in the Random Forest model) for the five individual physicochemical parameters (y-axis), when grouping the in vitro assays based on intended target family (x-axis).

Figure 20 shows the predicted effect concentrations (log_{10} AC_{50}s) plotted against the observed effect concentrations, based on Random Forest analysis, taking the aforementioned five physicochemical parameters as explanatory variables, for all individual intended target families separately. In total, 84.3% of all individual predicted AC_{50}s lied within a factor 5 of the observed AC_{50}s; 7% of the predicted datapoints were more than a factor five below the observed datapoints (underestimated), while 8.8% of all datapoints were more than a factor five above the observed data (overestimated). 0.07% of the predicted datapoints were a perfect fit, which may indicate overfitting of the model. Individual observed-predicted plots can be found below.

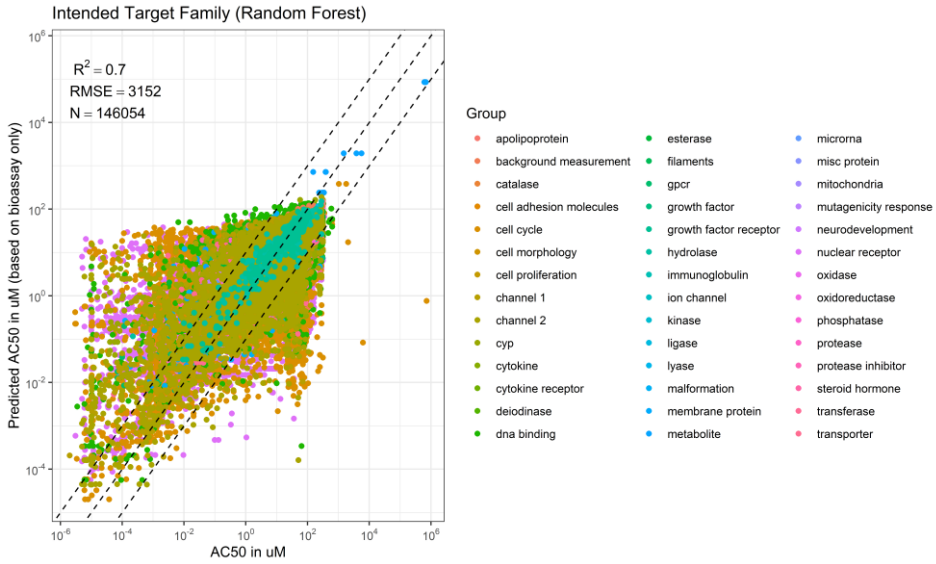


Figure 20: Predicted toxicity (AC_{50} in µM) by the Random Forest model versus observed toxicity, based on five physicochemical parameters (log K_{ow}, log K_{oc}, biodegradation rate, vapor pressure, and molecular weight), clustered per in vitro assay type (based on intended target family). The middle dashed line represents the 1:1 ratio. The outer dashed lines represent the 1:5 ratio and 5:1 ratio (the predicted values are 5 times higher/lower than the observed data).

Multiple linear regression model

In general, the multiple linear regression analysis, including the five most important predictive physicochemical descriptors (log K_{oc}, log K_{ow}, biodegradation rate (half-life in days), vapor pressure and molecular weight) as explanatory variables resulted in explaining 9.7% (median: 9.0%, S.E.: 0.02%) of all variance in the toxicity data (AC_{50}s) when categorizing *in vitro* assays based on intended target family, based on the adjusted R^2. Overall, when grouping bioassays based on intended target family, the highest % of variances explained were determined for *in vitro* assays related to mitochondrial target type (32%), while the lowest % variance explained by the multiple linear regression model were found for *in vitro* assays related to membrane proteins (-4%) (Table 1). Figure 21 shows the predicted effect concentrations (log_{10} AC_{50}s) plotted against the observed effect concentrations, based on the multiple linear regression model, taking the aforementioned five physicochemical parameters as explanatory variables (Equation 2), for all individual intended target families separately. In total, 62.84% of all individual predicted AC_{50}s lied within a factor 5 of the observed AC_{50}s; 21.25% of the predicted datapoints were more than a factor five below the observed datapoints (underestimated), while 15.91% of all datapoints were more than a factor five above the observed data (overestimated). None of the predicted datapoints were a perfect fit.

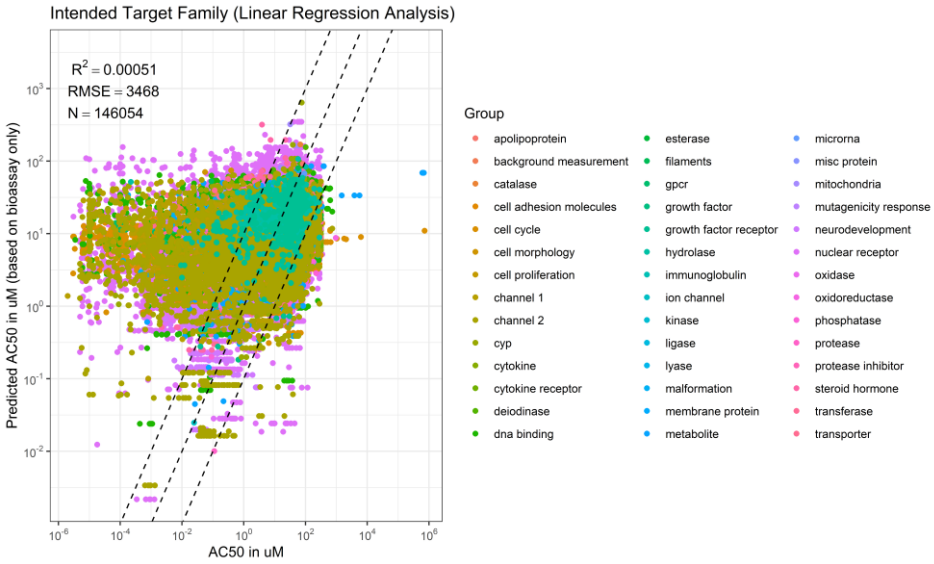


Figure 21: Predicted toxicity (AC_{50} in µM) by the multiple linear regression model versus observed toxicity, based on five physicochemical parameters (log K_{ow}, log K_{oc}, biodegradation rate, vapor pressure, and molecular weight), clustered per in vitro assay type (based on intended target family). The middle dashed line represents the 1:1 ratio. The outer dashed lines represent the 1:5 ratio and 5:1 ratio (the predicted values are 5 times higher/lower than the observed data).

Figure 22 shows the predicted effect concentrations ($\log_{10}$ AC_{50}s) plotted against the observed effect concentrations, based on both the multiple linear regression model and the Random Forest model, covering all individual intended target families. Overall, the Random Forest model had a higher predictive power (R^2 = 0.7, Figure 20) than the multiple linear regression model (R^2 = 0.00051, Figure 21), implying that the correlation between toxicity and the five physicochemical parameters of chemicals may be non-linear, when subdividing *in vitro* assays based on intended target family. Below, all individual predicted-observed plots when categorizing based on individual intended target families are shown (Figure 23).

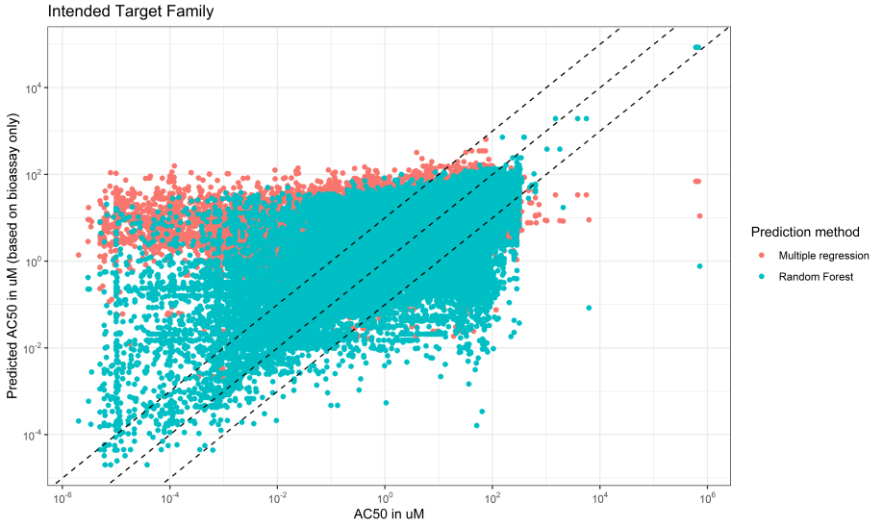


Figure 22: Predicted toxicity (AC_{50} in µM) by the multiple linear regression model and the Random Forest model versus observed toxicity, based on five physicochemical parameters (Log K_{ow}, Log K_{oc}, biodegradation rate, vapor pressure, and molecular weight), clustered per in vitro assay type (based on intended target family). The middle dashed line represents the 1:1 ratio. The outer dashed lines represent the 1:5 ratio and 5:1 ratio (the predicted values are 5 times higher/lower than the observed data).

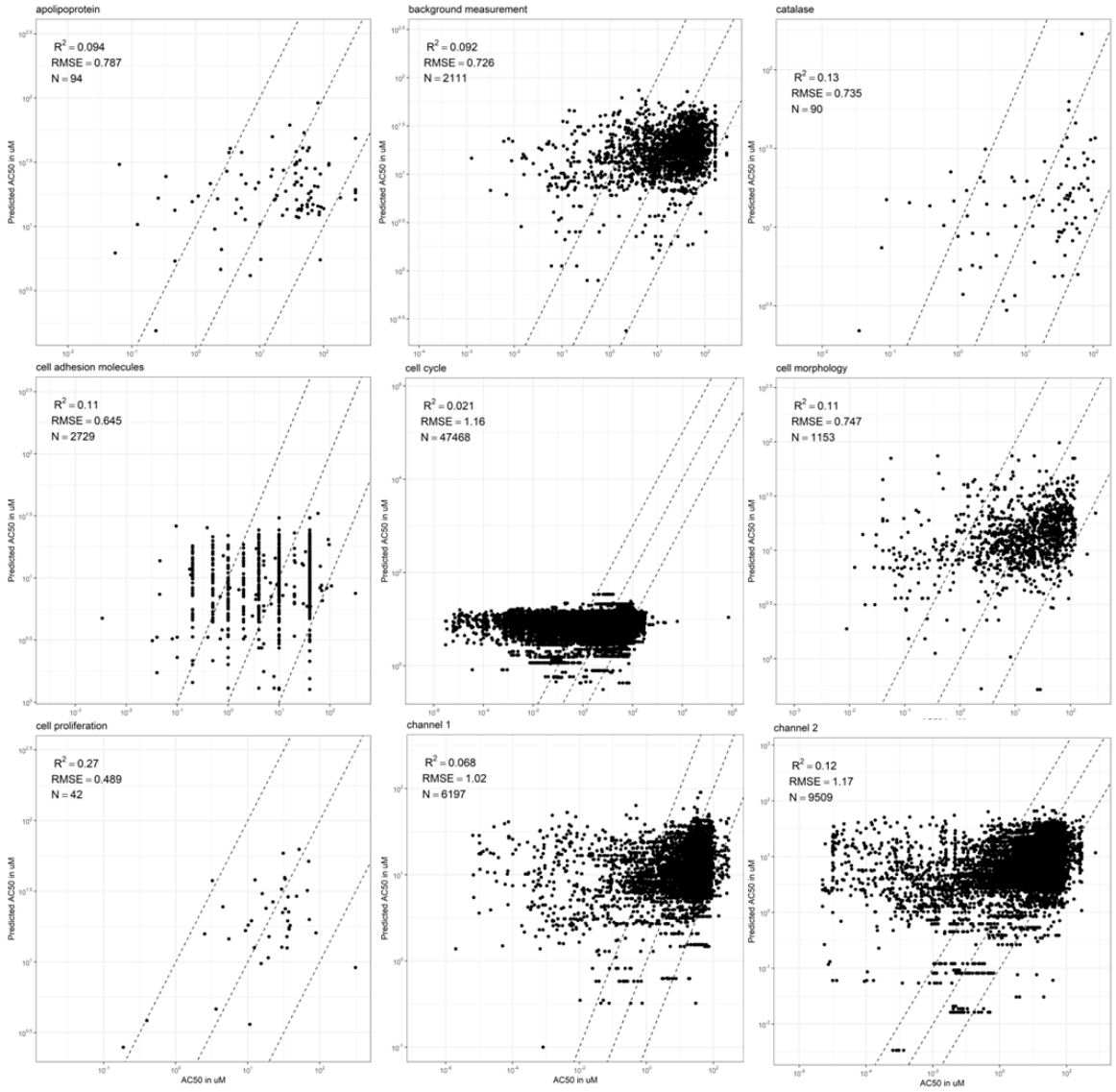


Figure 23: Predicted toxicity (AC_{50} in µM) by the multiple linear regression model versus observed toxicity, based on five physicochemical parameters (log K_{ow}, log K_{oc}, biodegradation rate, vapor pressure, and molecular weight), clustered per individual in vitro assay type (based on intended target family). The middle dashed line represents the 1:1 ratio. The outer dashed lines represent the 1:5 ratio and 5:1 ratio (the predicted values are 5 times higher/lower than the observed data).

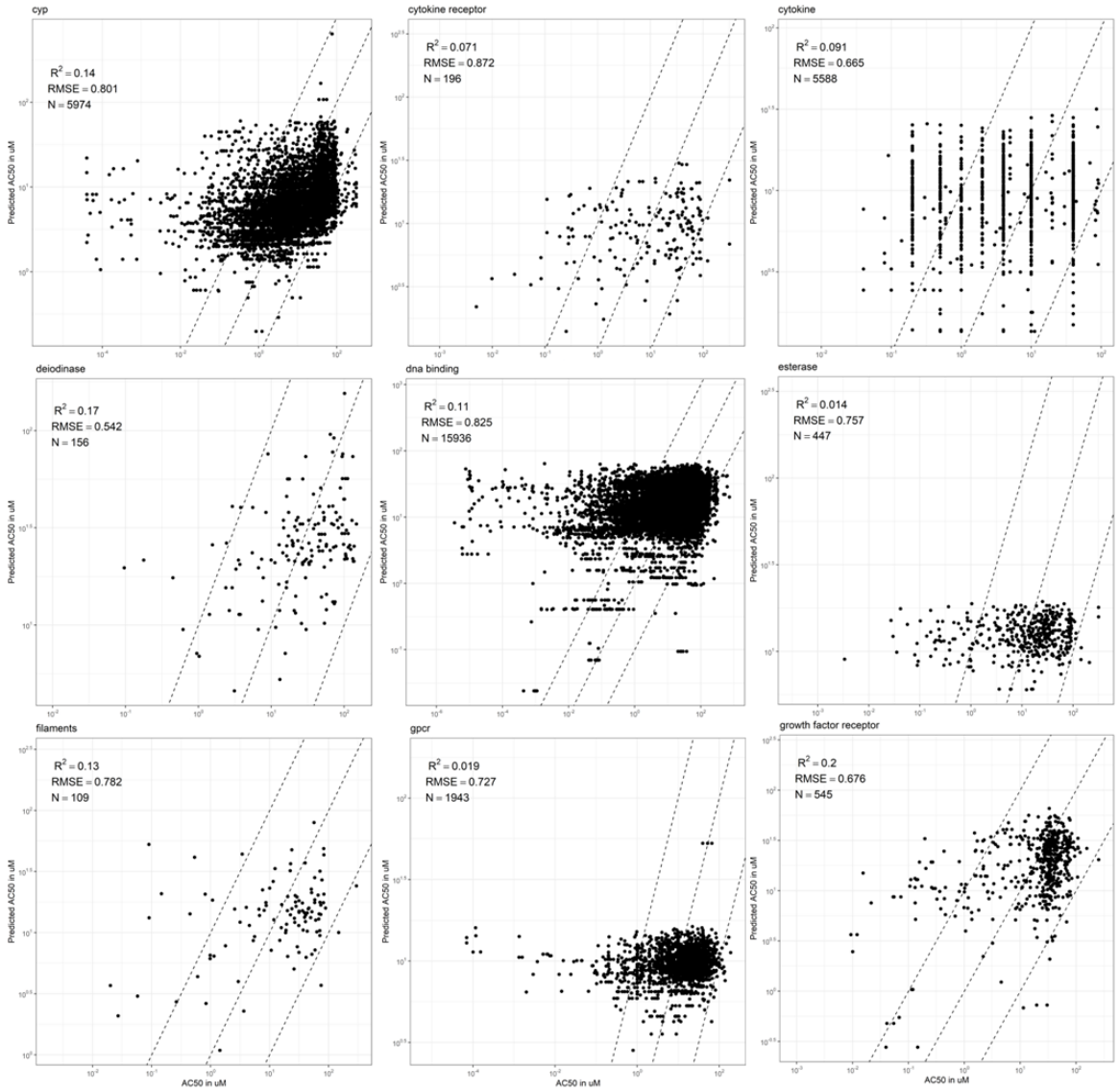


Figure 23 continued.

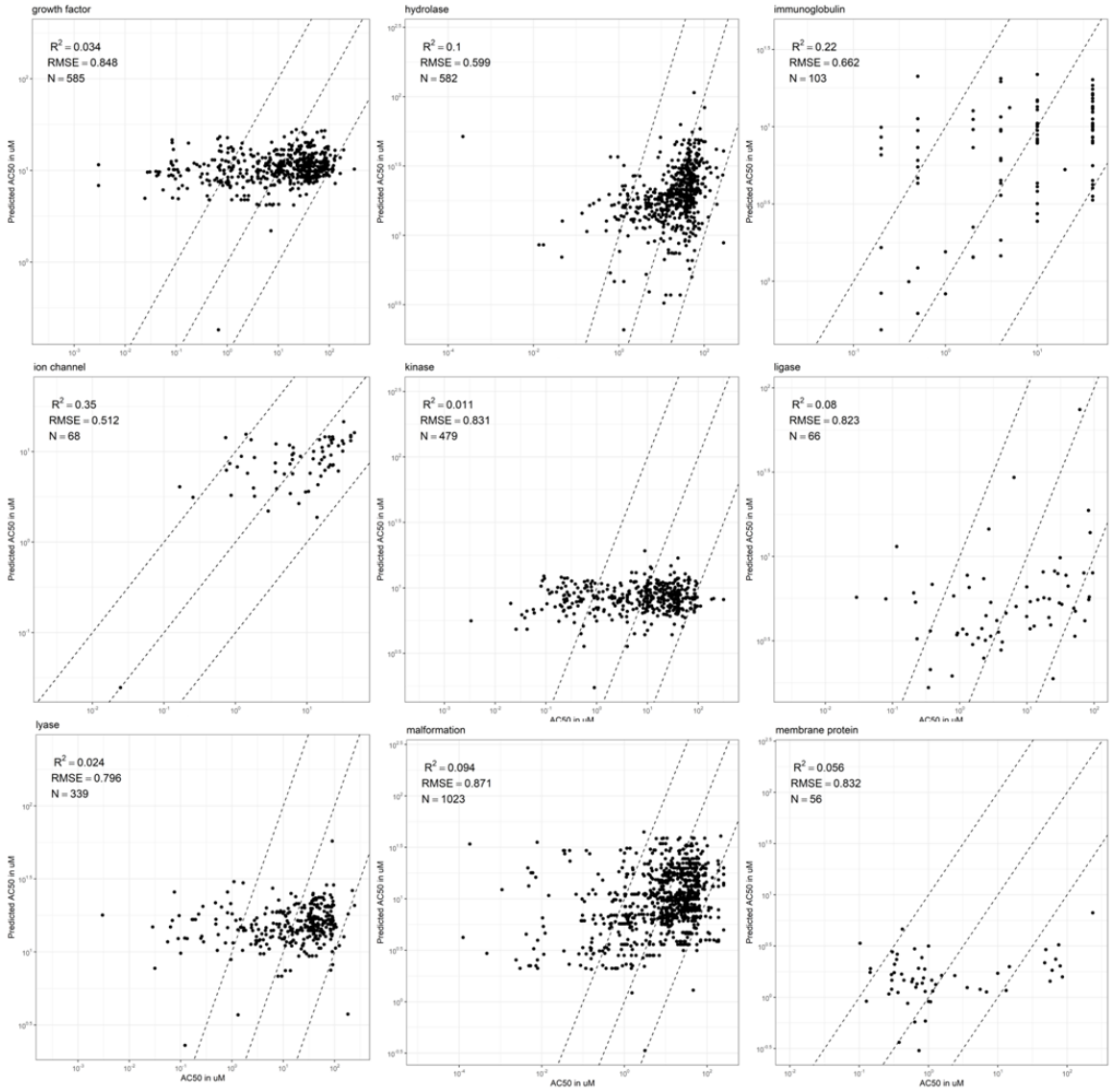


Figure 23 continued.

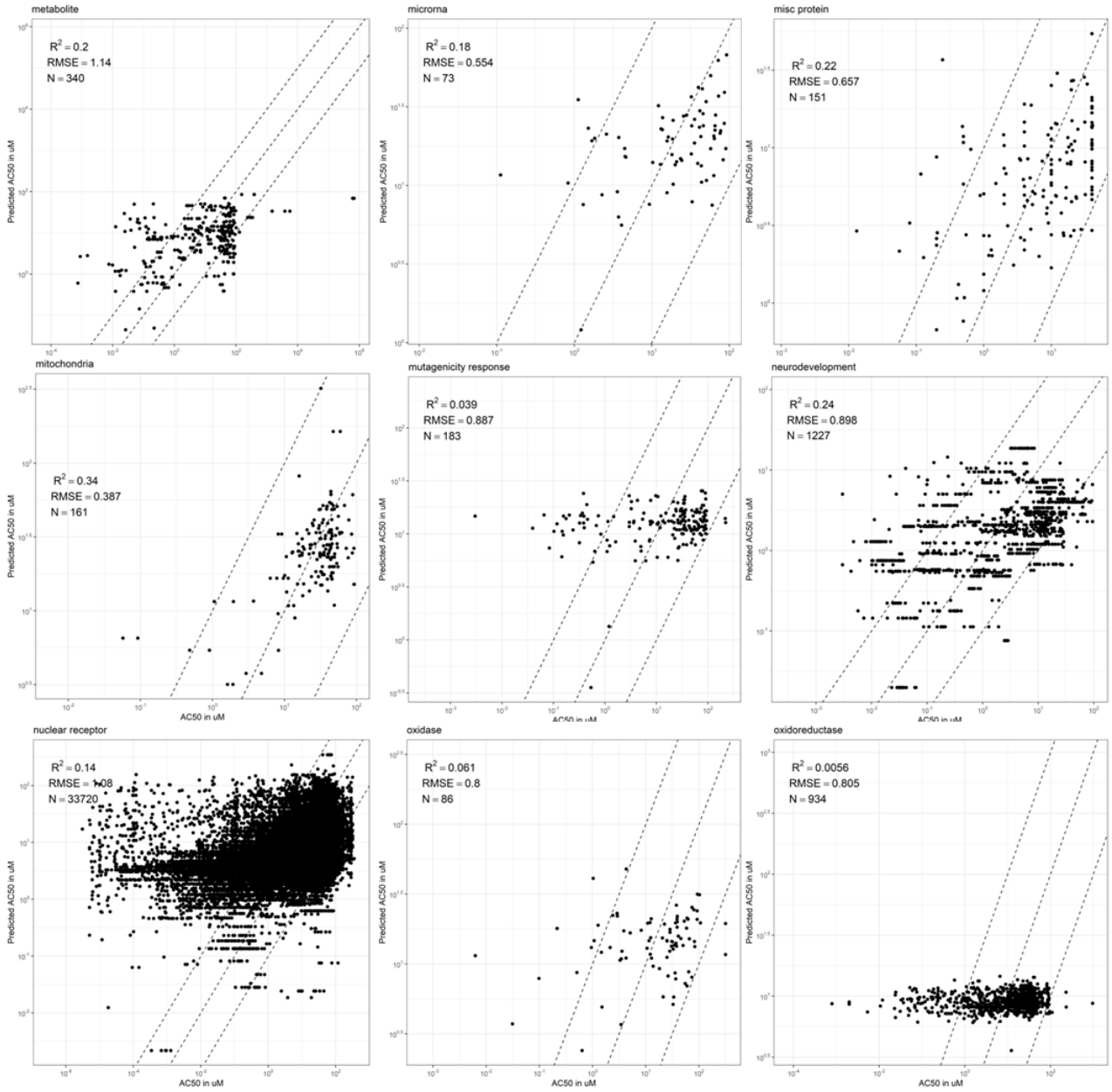


Figure 23 continued.

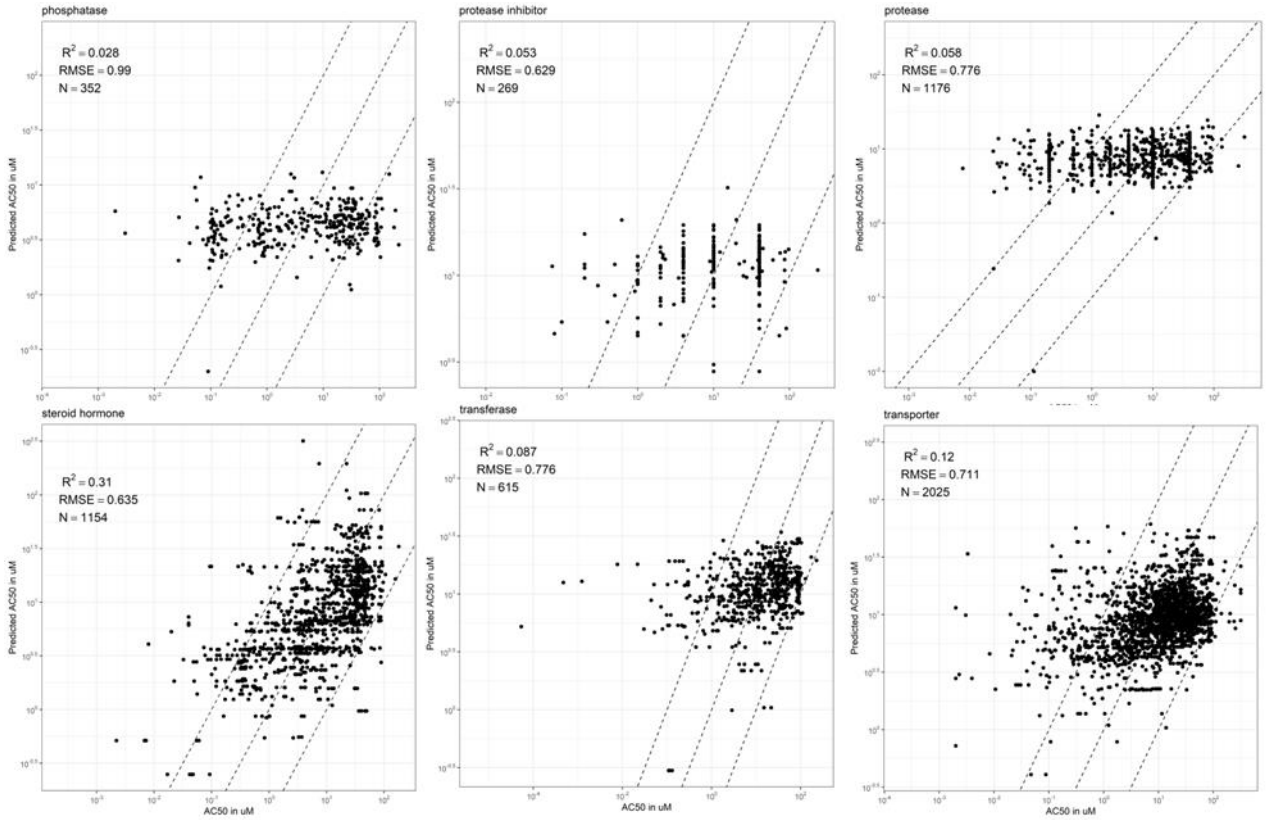


Figure 23 continued.

I.II Technological target type

The technological target type attempts to represent the individual targets across assay endpoints. These families pertain to gene families and include morphological and cell cycle concepts (U.S. EPA, 2015).

Random Forest model

In general, the Random Forest analysis, including the five most important predictive physicochemical descriptors (log K_{OC}, log K_{OW}, biodegradation rate, vapor pressure and molecular weight) as explanatory variables resulted in explaining 55.27% (median: 54.85%, S.E.: 0.04%) of all variance in the toxicity data (AC_{50}s) when categorizing *in vitro* assays based on technological target type. Overall, when grouping *in vitro* assays based on technological target type, the highest % of variances explained were determined for *in vitro* assays focusing on electrical activity (84.7%), while the lowest % variance explained by the Random Forest model was found for cellular *in vitro* assays (27.31%) (Table 1). This implies that physicochemical descriptors included in the present study correlated strongly with cellular *in vitro* assays AND that variation in AC_{50} values in the dataset with cellular assays was proportional to or higher than the variation of physicochemical descriptors from chemicals in the dataset. Figure 24 shows a heatmap visualizing to which extent the five physicochemical descriptors of interest correlate with toxicity for assays within one of the technological target types. The increase in MSE (%IncMSE) (Equation 1) corresponds to the extent to which the physicochemical parameter explains the variance in the Random Forest model.

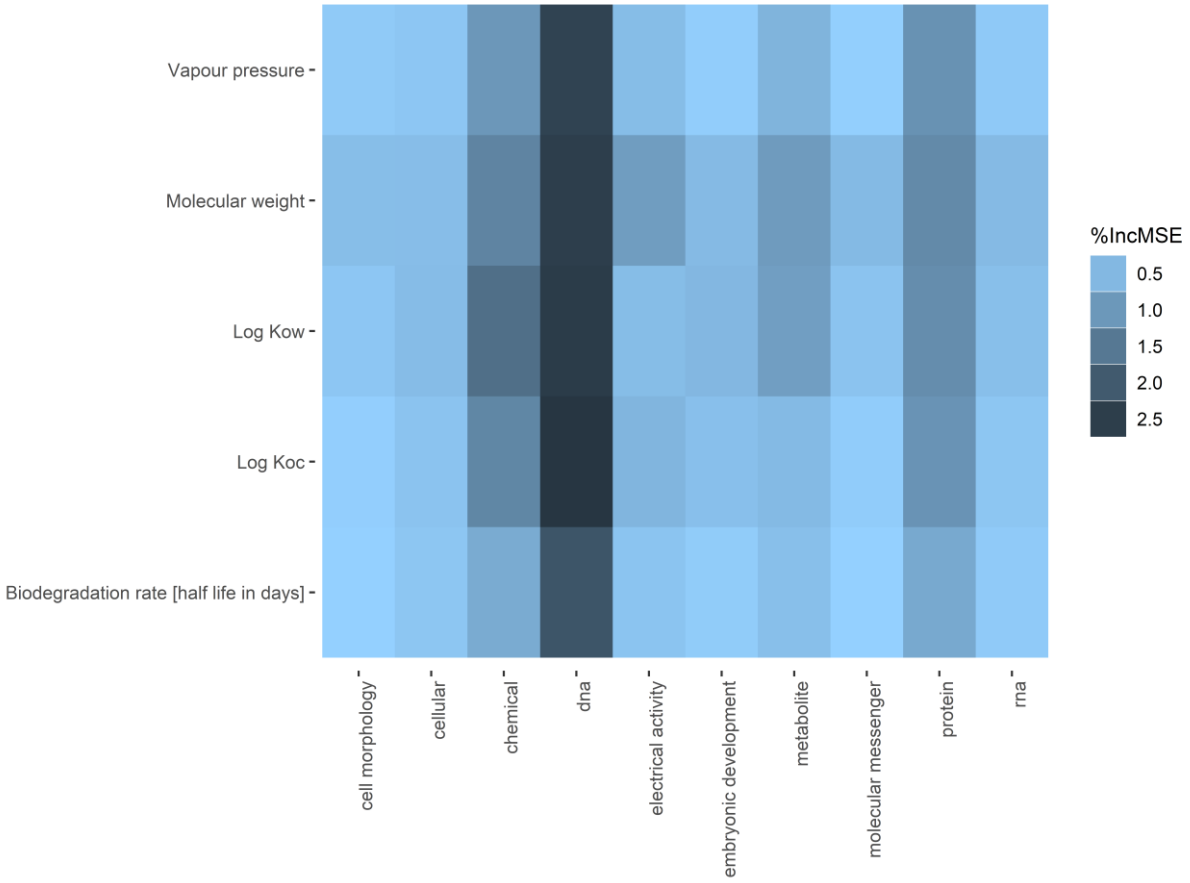


Figure 24:Heatmap visualizing the increase in MSE (= to which extent the variable explains the variance in toxicity in the Random Forest model) for the five individual physicochemical parameters (y-axis), when grouping the in vitro assays based on intended target family (x-axis).

Figure 25 shows the predicted effect concentrations (log AC_{50}s) plotted against the observed effect concentrations, based on Random Forest analysis, taking the aforementioned five physicochemical parameters as explanatory variables, for all technological target types separately. In total, 83.35% of all individual predicted AC_{50}s lied within a factor 5 of the observed AC_{50}s; 7.4% of the predicted datapoints were more than a factor five below the observed datapoints (underestimated), while 9.25% of all datapoints were more than a factor five above the observed data (overestimated). None of the predicted datapoints were a perfect fit. Individual observed-predicted plots can be found below.

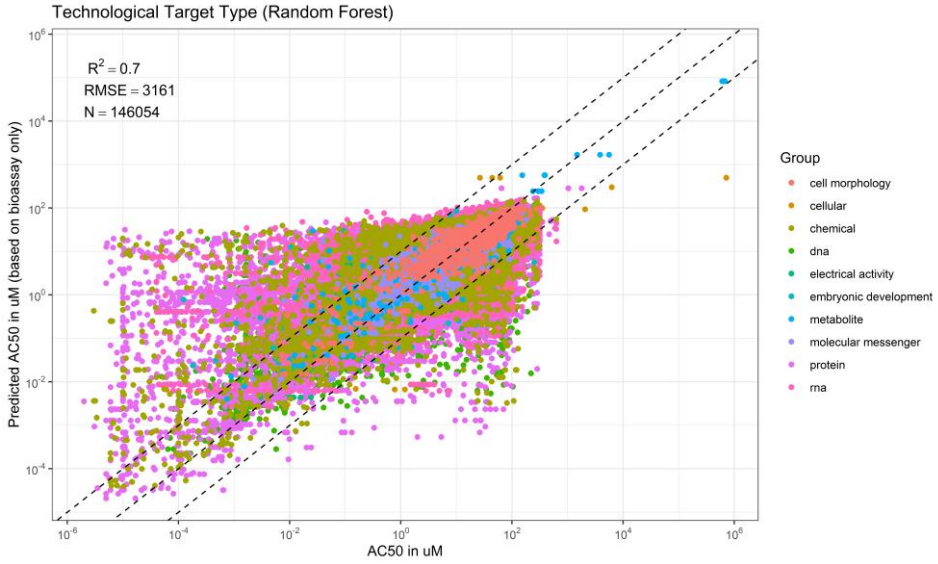


Figure 25: Predicted toxicity (AC_{50} in µM) by the Random Forest model versus observed toxicity, based on five physicochemical parameters (Log K_{ow}, Log K_{oc}, biodegradation rate, vapor pressure, and molecular weight), clustered per in vitro assay type (based on technological target type). The middle dashed line represents the 1:1 ratio. The outer dashed lines represent the 1:5 ratio and 5:1 ratio (the predicted values are 5 times higher/lower than the observed data).

Multiple linear regression model

In general, the multiple linear regression analysis, including the five most important predictive physicochemical descriptors (log K_{oc}, log K_{ow}, biodegradation rate (half-life in days), vapor pressure and molecular weight) as explanatory variables resulted in explaining 13.3% (median: 10.2%, S.E.: 0.02%) of all variance in the toxicity data (AC_{50}s) when categorizing *in vitro* assays based on technological target type, based on the adjusted R^2. Overall, when grouping *in vitro* assays based on technological target type, the highest % of variances explained were determined for *in vitro* assays related to molecular messaging (31.2%), while the lowest % variance explained by the multiple linear regression model were found for *in vitro* assays related to DNA (0.95%) (Table 1). Figure 26 shows the predicted effect concentrations (log_{10} AC_{50}s) plotted against the observed effect concentrations, based on the multiple linear regression model, taking the aforementioned five physicochemical parameters as explanatory variables (Equation 2), for all individual technological target types, separately. In total, 64% of all individual predicted AC_{50}s lied within a factor 5 of the observed AC_{50}s; 19.85% of the predicted datapoints were more than a factor five below the observed datapoints (underestimated), while 16.1% of all datapoints were more than a factor five above the observed data (overestimated). None of the predicted datapoints were a perfect fit (overfitted).

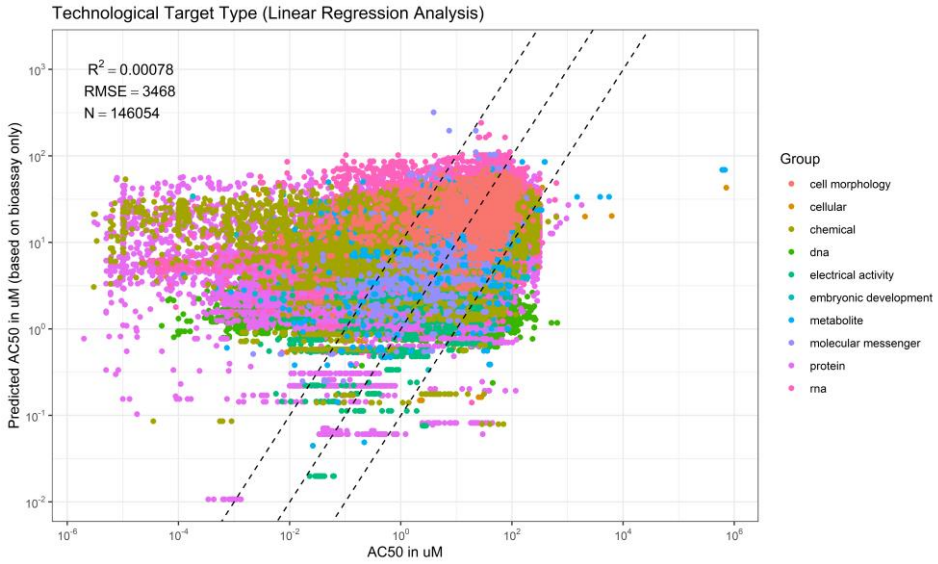


Figure 26: Predicted toxicity (AC_{50} in µM) by the multiple linear regression model versus observed toxicity, based on five physicochemical parameters (log K_{ow}, log K_{oc}, biodegradation rate, vapor pressure, and molecular weight), clustered per in vitro assay type (based on technological target type). The middle dashed line represents the 1:1 ratio. The outer dashed lines represent the 1:5 ratio and 5:1 ratio (the predicted values are 5 times higher/lower than the observed data).

Figure 27 shows the predicted effect concentrations ($\log_{10}$ AC_{50}s) plotted against the observed effect concentrations, based on both the multiple linear regression model and the Random Forest model, covering all individual intended target families. Overall, the Random Forest model had a higher predictive power (R^2 = 0.7, Figure 25) than the multiple linear regression model (R^2 = 0.00078, Figure 26), implying that the correlation between toxicity and the five physicochemical parameters of chemicals may be non-linear, when subdividing bioassays based on technological target type.

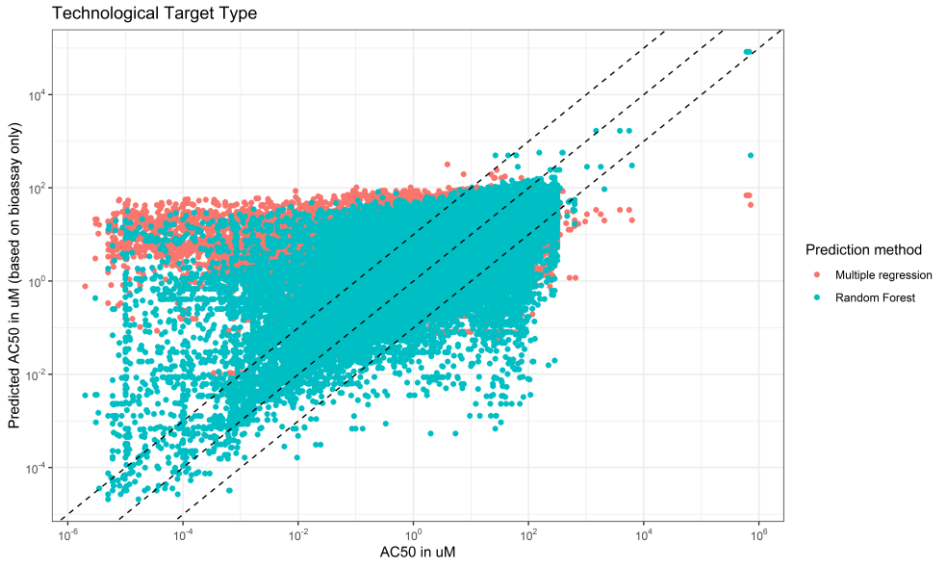


Figure 27: Predicted toxicity (AC_{50} in µM) by the multiple linear regression model and the Random Forest model versus observed toxicity, based on five physicochemical parameters (log K_{ow}, log K_{oc}, biodegradation rate, vapor pressure, and molecular weight), clustered per in vitro assay type (based on technological target family). The middle dashed line represents the 1:1 ratio. The outer dashed lines represent the 1:5 ratio and 5:1 ratio (the predicted values are 5 times higher/lower than the observed data).

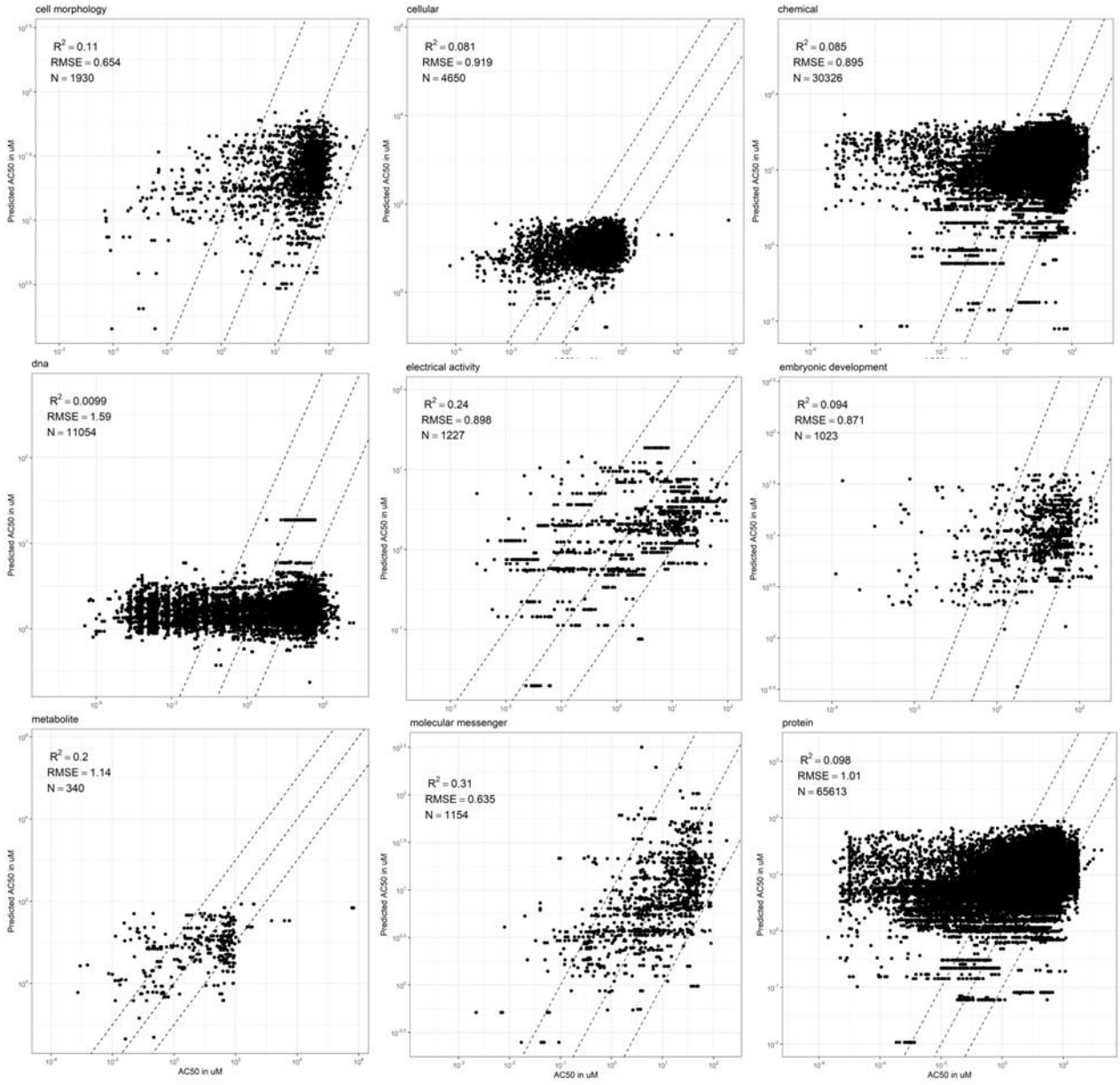


Figure 28: Predicted toxicity (AC_{50} in µM) by the multiple linear regression model versus observed toxicity, based on five physicochemical parameters (log K_{ow}, log K_{oc}, biodegradation rate, vapor pressure, and molecular weight), clustered per individual in vitro assay type (based on technological target type). The middle dashed line represents the 1:1 ratio. The outer dashed lines represent the 1:5 ratio and 5:1 ratio.

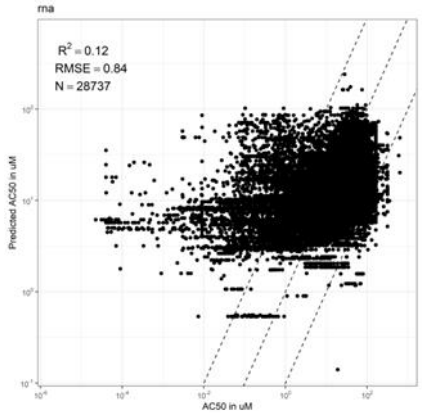


Figure 28 continued.

I.III Assay design type

The assay design type represents the method that a biological or physical process is translated into a detectable signal. (U.S. EPA, 2015). The assay design type annotation captures the method by which the technological target is measured.

Random Forest model

In general, the Random Forest analysis, including the five most important predictive physicochemical descriptors (log K_{oc}, log K_{ow}, biodegradation rate, vapor pressure and molecular weight) as explanatory variables resulted in explaining 42.73% (median: 38.25%, S.E.: 0.06%) of all variance in the toxicity data (AC_{50}s) when categorizing *in vitro* assays based on assay design type. Overall, when grouping *in vitro* assays based on assay design type, the highest % of variances explained were determined for *in vitro* assays characterized as functional reporters (84.65%), while the lowest % variance explained by the Random Forest model was found for bioassays characterized as enzyme reporters (15.15%) (Table 1). This implies that physicochemical descriptors included in the present study correlated strongly with *in vitro* assays characterized as functional reporters AND that variation in AC_{50} values in the dataset with functional reporter assays was proportional to or higher than the variation of physicochemical descriptors from chemicals in the dataset. Figure 29 shows a heatmap visualizing to which extent the five physicochemical descriptors of interest correlate with toxicity for assays within one of the assay design types. The increase in MSE (%IncMSE) (Equation 1) corresponds to the extent to which the physicochemical parameter explains the variance in the Random Forest model.

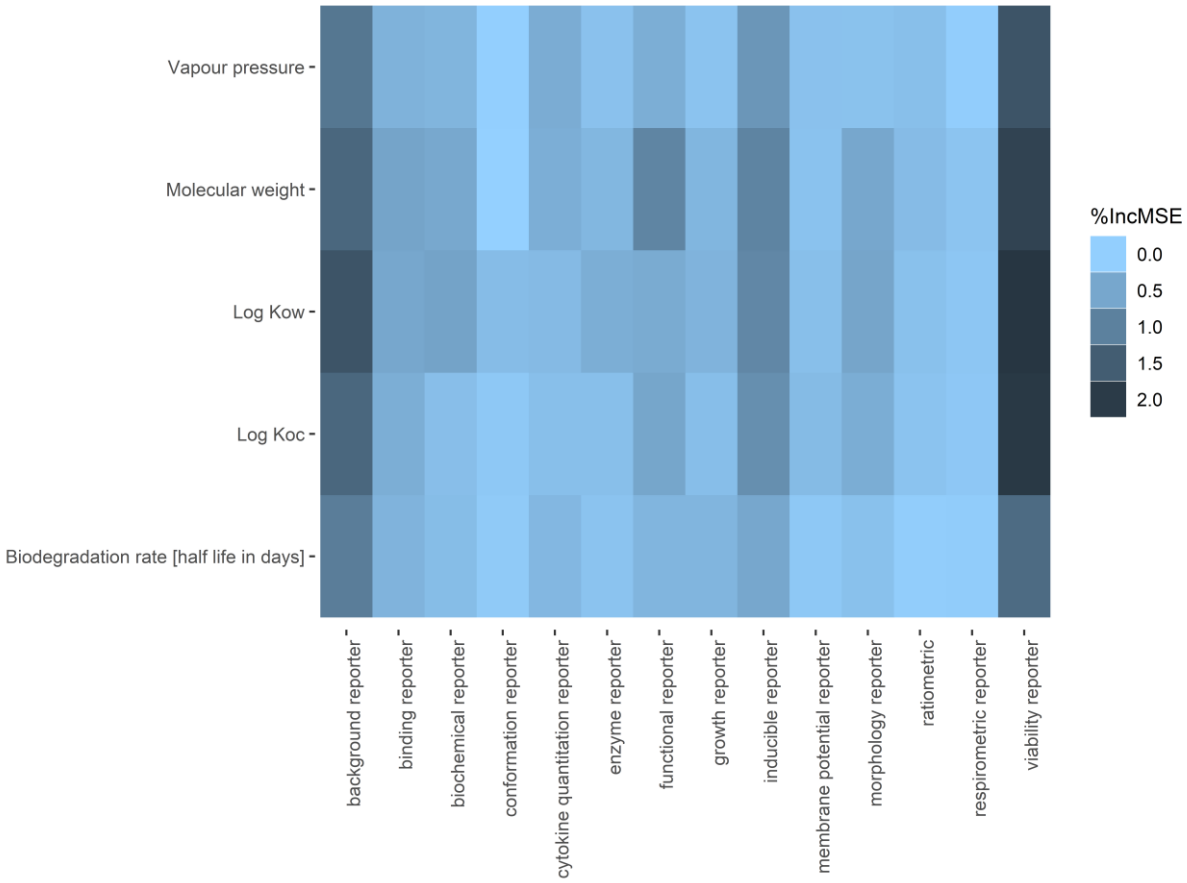


Figure 29: Heatmap visualizing the increase in MSE (= to which extent the variable explains the variance in toxicity in the Random Forest model) for the five individual physicochemical parameters (y-axis), when grouping the in vitro assays based on assay design type (x-axis).

Figure 30 shows the predicted effect concentrations (log_{10} AC_{50}s) plotted against the observed effect concentrations, based on Random Forest analysis, taking the aforementioned five physicochemical parameters as explanatory variables, for all assay design types separately. In total, 82.4% of all individual predicted AC_{50}s were within a factor 5 of the observed AC_{50}s; 7.8% of the predicted datapoints were more than a factor five below the observed datapoints (underestimated), while 9.8% of all datapoints were more than a factor five above the observed data (overestimated). None of the predicted datapoints were a perfect fit. Individual observed-predicted plots can be found below.

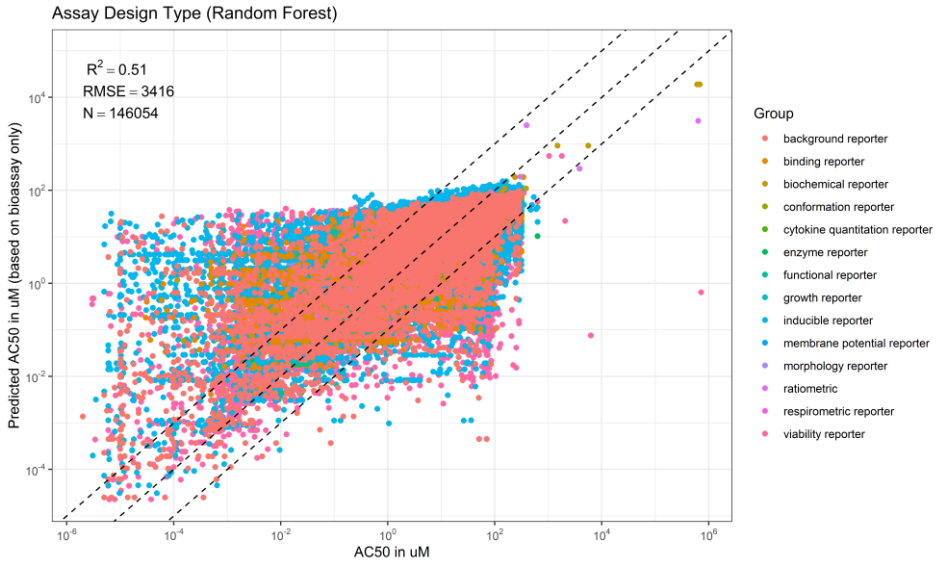


Figure 30: Predicted toxicity (AC_{50} in µM) by the Random Forest model versus observed toxicity, based on five physicochemical parameters (log K_{ow}, log K_{oc}, biodegradation rate, vapor pressure, and molecular weight), clustered per in vitro *assay type (based on assay design type). The middle dashed line represents the 1:1 ratio. The outer dashed lines represent the 1:5 ratio and 5:1 ratio (the predicted values are 5 times higher/lower than the observed data).*

Multiple linear regression model

In general, the multiple linear regression analysis, including the five most important predictive physicochemical descriptors (log K_{oc}, log K_{ow}, biodegradation rate (half-life in days), vapor pressure and molecular weight) as explanatory variables resulted in explaining 12.7% (median: 11.3%, S.E.: 0.02%) of all variance in the toxicity data (AC_{50}s) when categorizing *in vitro* assays based on assay design type, based on the adjusted R^2. Overall, when grouping *in vitro* assays based on assay design type, the highest % of variances explained were determined for *in vitro* assays characterized as respirometric reporters (32%), while the lowest % variance explained by the multiple linear regression were found for *in vitro* assays characterized as biochemical reporters (1.5%) (Table 1).

Figure 31 shows the predicted effect concentrations (log_{10} AC_{50}s) plotted against the observed effect concentrations, based on the multiple linear regression model, taking the aforementioned five physicochemical parameters as explanatory variables (Equation 2), for all individual assay design types, separately. In total, 62.1% of all individual predicted AC_{50}s were within a factor 5 of the observed AC_{50}s; 21.7% of the predicted datapoints were more than a factor five below the observed datapoints (underestimated), while 16.1% of all datapoints were more than a factor five above the observed data (overestimated). None of the predicted datapoints were a perfect fit.

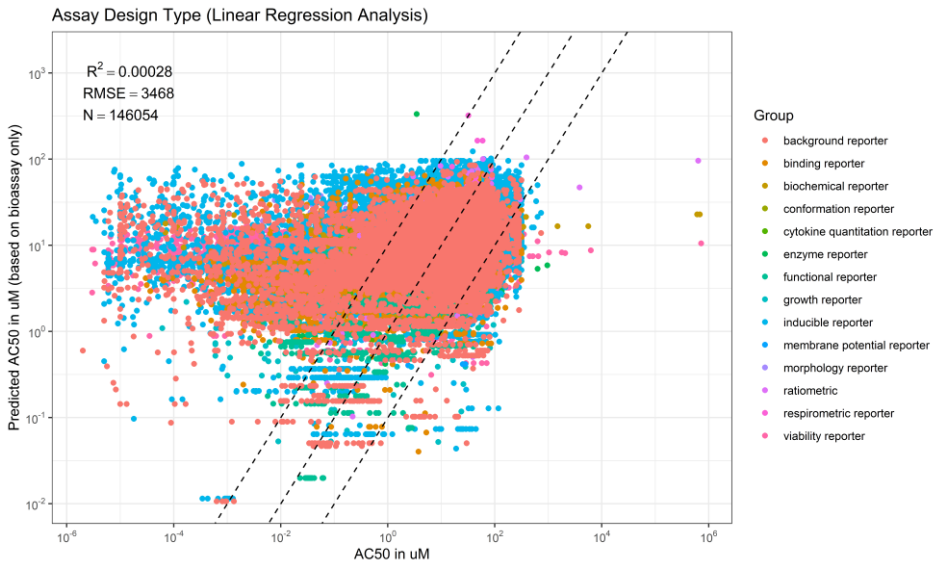


Figure 31: Predicted toxicity (AC_{50} in µM) by the multiple linear regression model versus observed toxicity, based on five physicochemical parameters (log K_{ow}, log K_{oc}, biodegradation rate, vapor pressure, and molecular weight), clustered per in vitro assay type (based on assay design type). The middle dashed line represents the 1:1 ratio. The outer dashed lines represent the 1:5 ratio and 5:1 ratio (the predicted values are 5 times higher/lower than the observed data).

Figure 32 shows the predicted effect concentrations ($\log_{10}$ AC_{50}s) plotted against the observed effect concentrations, based on both the multiple linear regression model and the Random Forest model, covering all individual intended target families. Overall, the Random Forest model had a higher predictive power (R^2 = 0.51, Figure 30) than the multiple linear regression model (R^2 = 0.00028, Figure 31), implying that the correlation between toxicity and the five physicochemical parameters of chemicals may be non-linear, when subdividing *in vitro* assays based on assay design type.

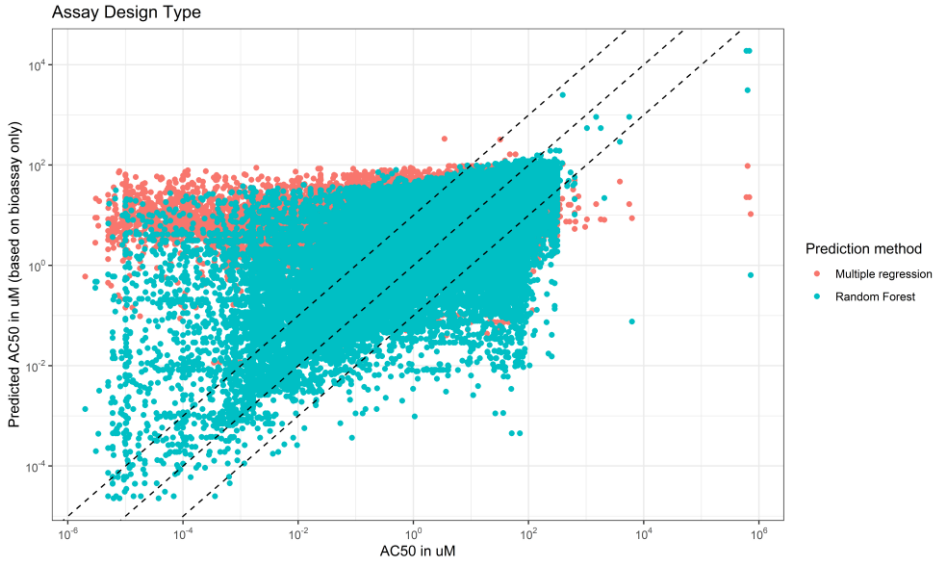


Figure 32: Predicted toxicity (AC_{50} in µM) by the multiple linear regression model versus observed toxicity, based on five physicochemical parameters (log K_{ow}, log K_{oc}, biodegradation rate, vapor pressure, and molecular weight), clustered per in vitro assay type (based on assay design type). The middle dashed line represents the 1:1 ratio. The outer dashed lines represent the 1:5 ratio and 5:1 ratio (the predicted values are 5 times higher/lower than the observed data).

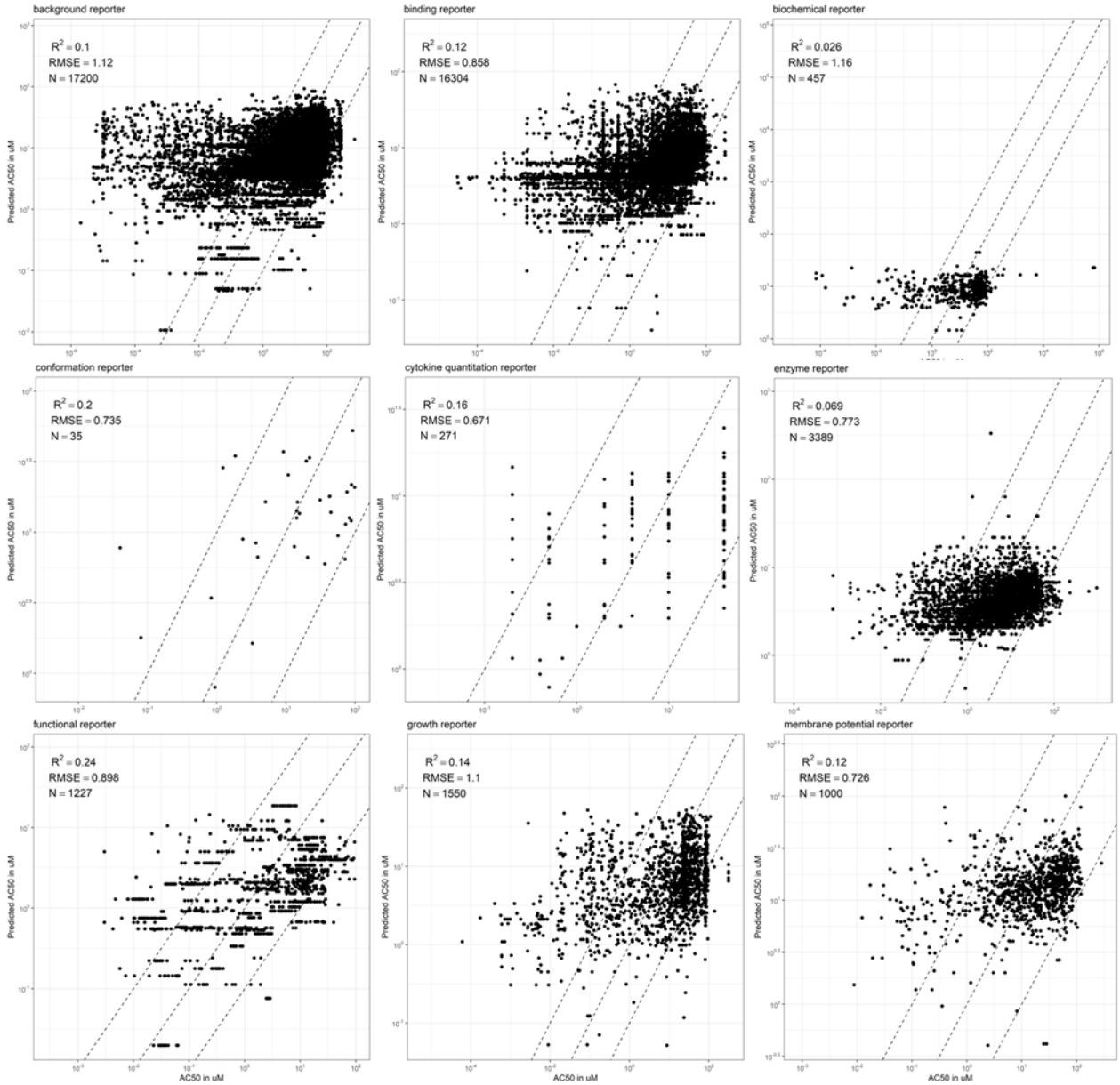


Figure 33: Predicted toxicity (AC_{50} in µM) by the multiple linear regression model versus observed toxicity, based on five physicochemical parameters (log K_{ow}, log K_{oc}, biodegradation rate, vapor pressure, and molecular weight), clustered per individual in vitro *assay type (based on assay design type). The middle dashed line represents the 1:1 ratio. The outer dashed lines represent the 1:5 ratio and 5:1 ratio (the predicted values are 5 times higher/lower than the observed data).*

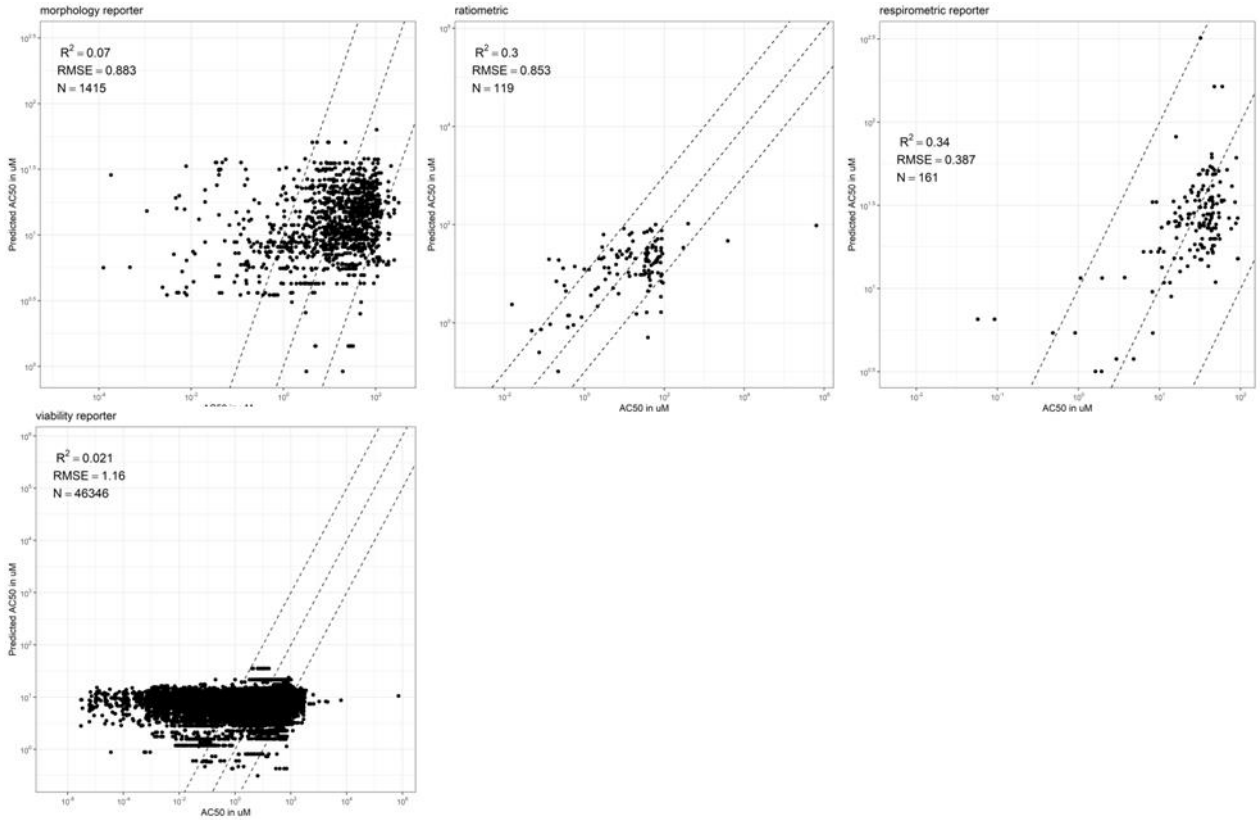


Figure 33 continued.

I.IV Signal direction

The signal direction (Figure 1G) indicates whether the *in vitro* assay endpoint provides 'gain' or 'loss' of signal data (U.S. EPA, 2015).

Random Forest model

In general, the Random Forest analysis, including the five most important predictive physicochemical descriptors (log K_{oc}, log K_{ow}, biodegradation rate, vapor pressure and molecular weight) as explanatory variables resulted in explaining 41.71% (median: 41.71, S.E.: 0.00024) of all variance in the toxicity data (AC_{50}s) when categorizing *in vitro* assays based on signal direction. Overall, when grouping *in vitro* assays based on signal direction, the highest % of variances explained were determined for *in vitro* assays with chemicals producing a hit (i.e., $AC_{50} \leq 1000$ µM) causing a loss of signal (41.78%), while the lowest % variance explained by the Random Forest model was found for bioassays with chemicals producing a hit causing a gain of signal (41.63%) (Table 1). *in vitro* assays characterized as functional reporters Figure 34 shows a heatmap visualizing to which extent the five physicochemical descriptors of interest correlate with toxicity for *in vitro* assays within one of the two signal directions. The increase in MSE (%IncMSE) (Equation 1) corresponds to the extent to which the physicochemical parameter explains the variance in the Random Forest model.

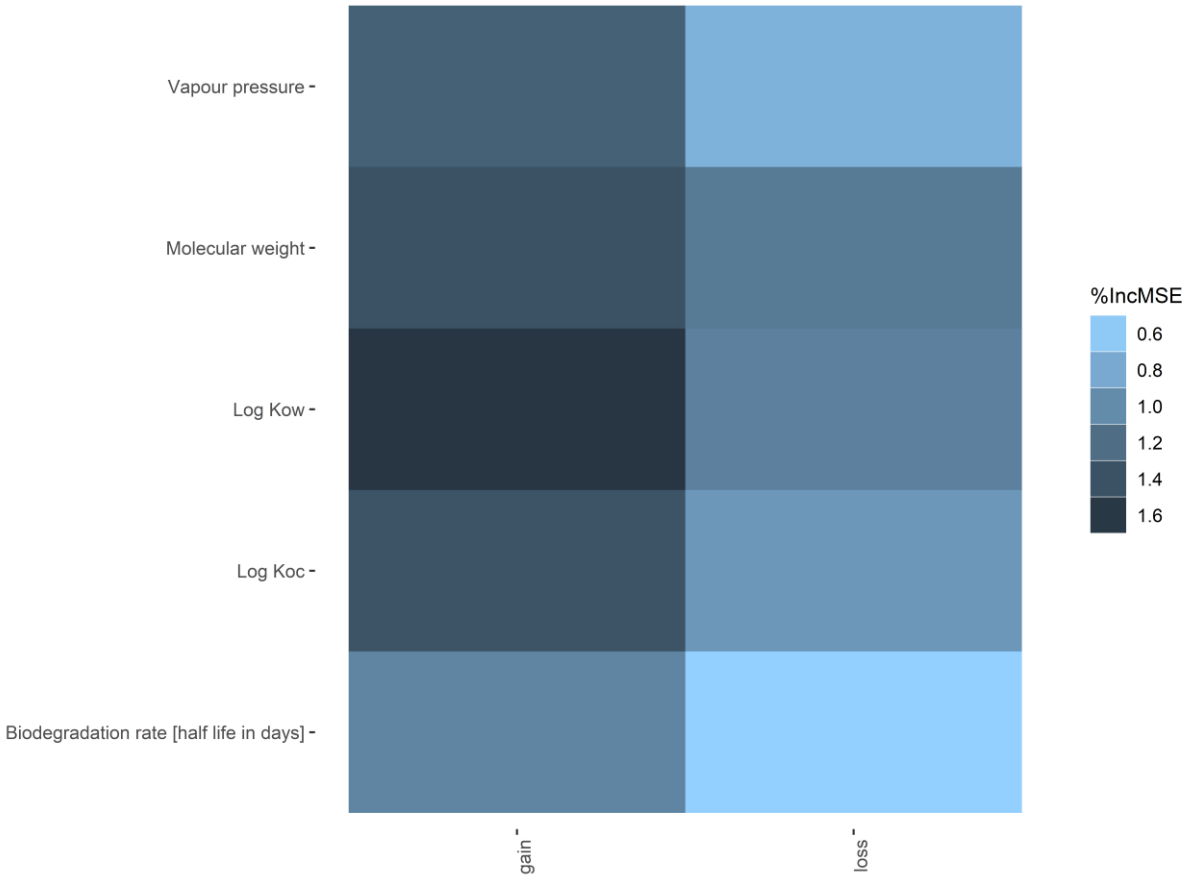


Figure 34: Heatmap visualizing the increase in MSE (= to which extent the variable explains the variance in toxicity in the Random Forest model) for the five individual physicochemical parameters (y-axis), when grouping the in vitro assays based on signal direction (x-axis).

Figure 35 shows the predicted effect concentrations (log_{10} AC_{50}s) plotted against the observed effect concentrations, based on Random Forest analysis, taking the aforementioned five physicochemical parameters as explanatory

variables, for both signal directions separately. In total, 79.15% of all individual predicted AC_{50}s lied within a factor 5 of the observed AC_{50}s; 9.35% of the predicted datapoints were more than a factor five below the observed datapoints (underestimated), while 11.5% of all datapoints were more than a factor five above the observed data (overestimated). 0.008% of the predicted datapoints were a perfect fit, which may indicate overfitting of the model. Individual observed-predicted plots can be found below.

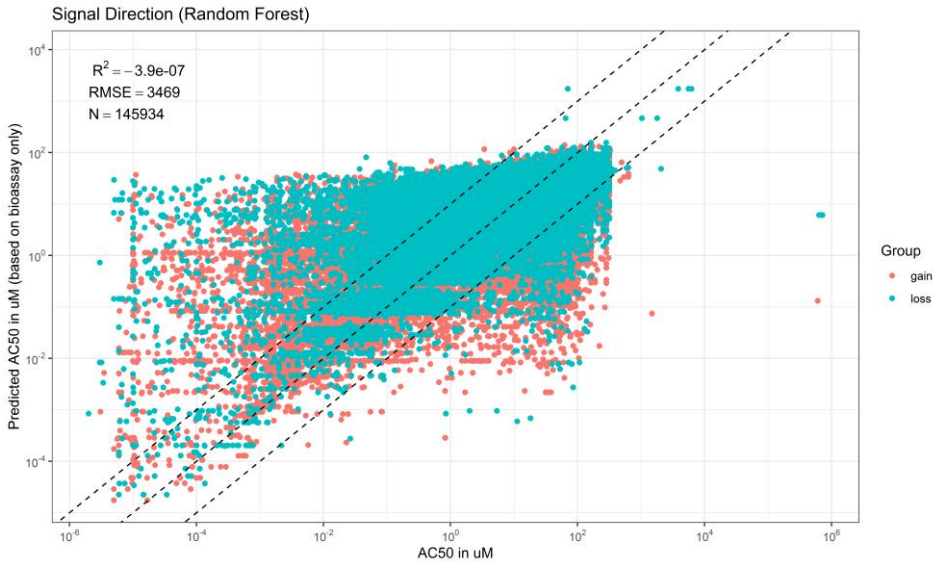


Figure 35: Predicted toxicity (AC_{50} in µM) by the Random Forest model versus observed toxicity, based on five physicochemical parameters (log K_{ow}, log K_{oc}, biodegradation rate, vapor pressure, and molecular weight), clustered per in vitro assay type (based on signal direction). The middle dashed line represents the 1:1 ratio. The outer dashed lines represent the 1:5 ratio and 5:1 ratio (the predicted values are 5 times higher/lower than the observed data).

Multiple linear regression model

In general, the multiple linear regression analysis, including the five most important predictive physicochemical descriptors (log K_{oc}, log K_{ow}, biodegradation rate (half-life in days), vapor pressure and molecular weight) as explanatory variables resulted in explaining 6.36% (median: 6.36% % , S.E.: 0.01%) of all variance in the toxicity data (AC_{50}s) when categorizing *in vitro* assays based on signal direction, based on the adjusted R^2. Overall, when grouping *in vitro* assays based on signal direction, the highest % of variances explained were determined for *in vitro* assays with chemicals producing a hit causing a loss of signal (8.4%), while the lowest % variance explained by the multiple linear regression model were found for *in vitro* assays with chemicals producing a hit causing a gain of signal (4.3%) (Table 1).

Figure 36 shows the predicted effect concentrations (log_{10} AC_{50}s) plotted against the observed effect concentrations, based on the multiple linear regression model, taking the aforementioned five physicochemical parameters as explanatory variables (Equation 2), for both signal directions, separately. In total, 61.78% of all individual predicted AC_{50}s lied within a factor 5 of the observed AC_{50}s; 21.8% of the predicted datapoints were more than a factor five below the observed datapoints (underestimated), while 16.4% of all datapoints were more than a factor five above the observed data (overestimated). None of the predicted datapoints were a perfect fit.

Figure 36: Predicted toxicity (AC_{50} in µM) by the multiple linear regression model versus observed toxicity, based on five physicochemical parameters (log K_{ow}, log K_{oc}, biodegradation rate, vapor pressure, and molecular weight), clustered per in vitro assay type (based on signal direction). The middle dashed line represents the 1:1 ratio. The outer dashed lines represent the 1:5 ratio and 5:1 ratio (the predicted values are 5 times higher/lower than the observed data).

Figure 37 shows the predicted effect concentrations (log_{10} AC_{50}s) plotted against the observed effect concentrations, based on both the multiple linear regression model and the Random Forest model, covering all individual intended target families. Overall, the Random Forest model had a slightly lower predictive power (R^2 = 3.9E-7, Figure 35) than the multiple linear regression model (R^2 = 0.00051, Figure 36), implying that the correlation between toxicity and the five physicochemical parameters of chemicals may be non-linear, when subdividing *in vitro* assays based on signal direction. However, both the R^2 and variance explained by the random forest model were very low, implying that sub setting *in vitro* assays based on signal direction does not lead to a better model fit in both cases.

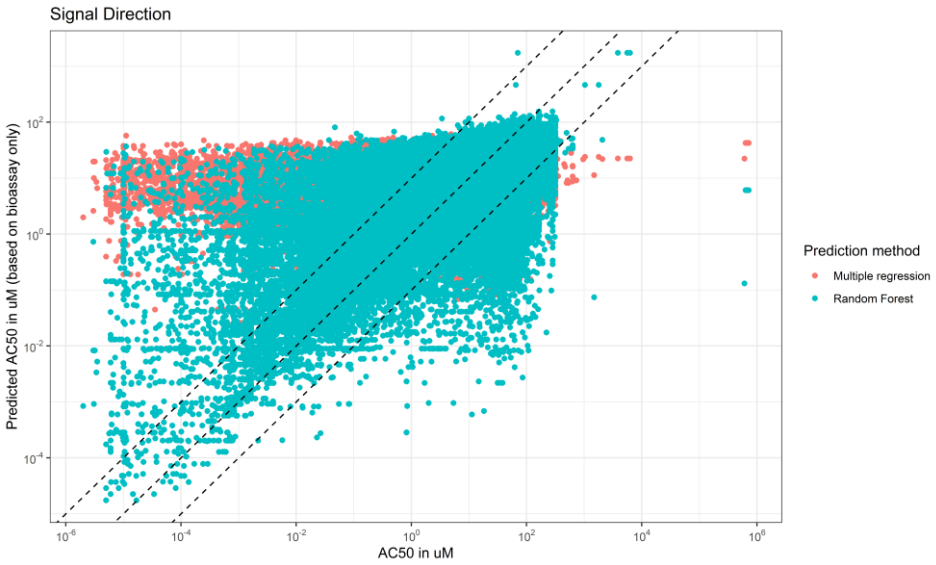


Figure 37: Predicted toxicity (AC_{50} in µM) by the multiple linear regression model versus observed toxicity, based on five physicochemical parameters (log K_{ow}, log K_{oc}, biodegradation rate, vapor pressure, and molecular weight), clustered per in vitro assay type (based on signal direction). The middle dashed line represents the 1:1 ratio. The outer dashed lines represent the 1:5 ratio and 5:1 ratio (the predicted values are 5 times higher/lower than the observed data).

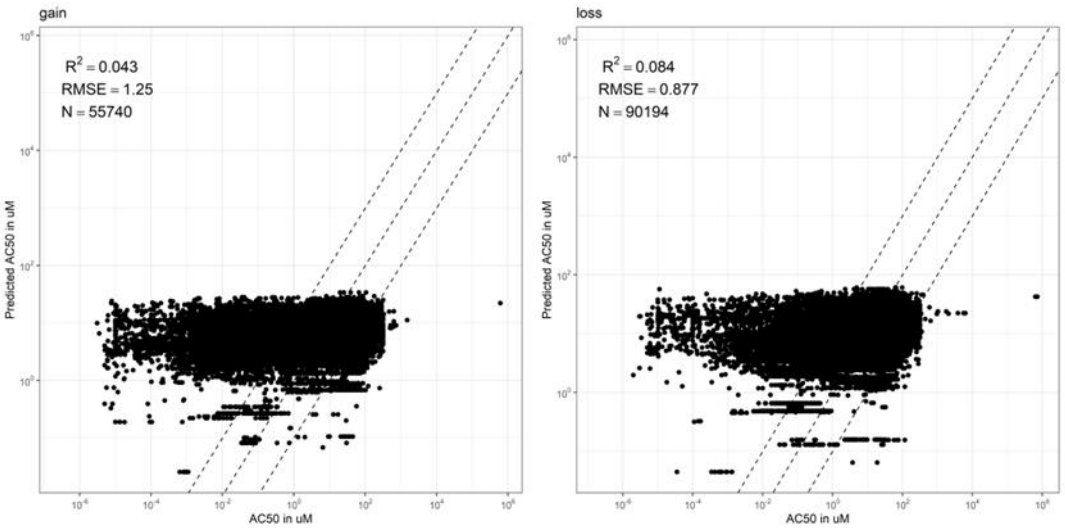


Figure 38: Predicted toxicity (AC_{50} in µM) by the multiple linear regression model versus observed toxicity, based on five physicochemical parameters (log K_{ow}, log K_{oc}, biodegradation rate, vapor pressure, and molecular weight), clustered per individual assay type (based signal direction). The middle dashed line represents the 1:1 ratio. The outer dashed lines represent the 1:5 ratio and 5:1 ratio (the predicted values are 5 times higher/lower than the observed data).

I.V Organism and tissue type

Random Forest model

In general, the Random Forest analysis, including the five most important predictive physicochemical descriptors (log K_{oc}, log K_{ow}, biodegradation rate, vapor pressure and molecular weight) as explanatory variables resulted in explaining 37.1% (median: 41.36%, S.E.: 0.07%) of all variance in the toxicity data (AC_{50}s) when categorizing *in vitro* assays based on organism-tissue combination. Overall, when grouping bioassays based on organism-tissue combination, the highest % of variances explained were determined for *in vitro* assays based on cortical rat cells (83.2%), while the lowest % variance explained by the Random Forest model was found for *in vitro* assays based on human brain cells (-25.56%) (Table 1). This implies that physicochemical descriptors included in the present study correlated strongly with assays using cortical rat cells AND that variation in AC_{50} values in the dataset including assays using corticol rat cells was proportional to or higher than the variation of physicochemical descriptors from chemicals in the dataset. Furthermore, this also implies that using a Random Forest model to predict toxicity for the subset of chemicals and in vitro assay endpoint using human brain cells result in a prediction that is worse than taking the average of all AC_{50} values, likely due to data limitations. Figure 39 shows a heatmap visualizing to which extent the five physicochemical descriptors of interest correlate with toxicity for *in vitro* assays within the organism-tissue combinations. The increase in MSE (%IncMSE) (Equation 1) corresponds to the extent to which the physicochemical parameter explains the variance in the Random Forest model.

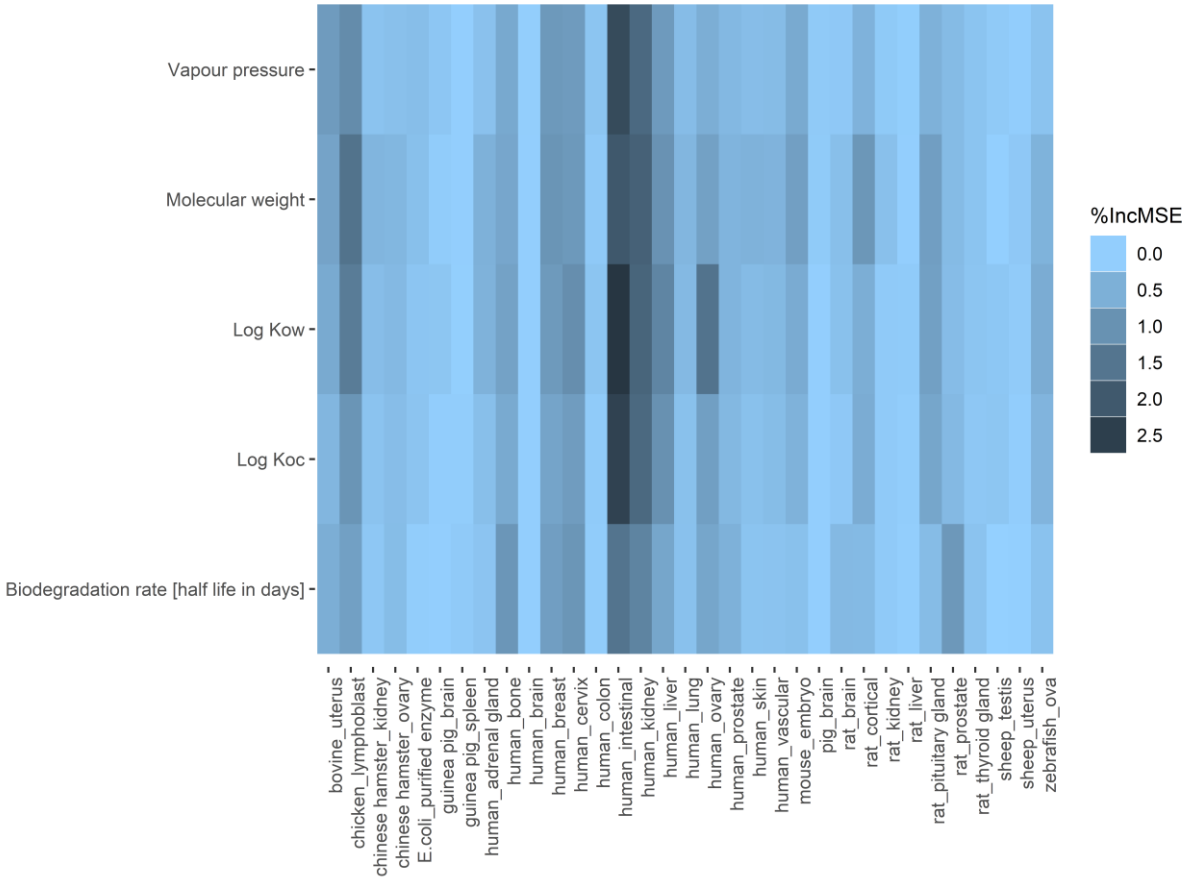


Figure 39: Heatmap visualizing the increase in MSE (= to which extent the variable explains the variance in toxicity in the Random Forest model) for the five individual physicochemical parameters (y-axis), when grouping the in vitro assays based on organism-tissue combination (x-axis).

Figure 40 shows the predicted effect concentrations (log_{10} AC_{50}s) plotted against the observed effect concentrations, based on Random Forest analysis, taking the aforementioned five physicochemical parameters as explanatory variables, for both signal directions separately. In total, 85.54% of all individual predicted AC_{50}s lied within a factor 5

of the observed AC_{50}s; 6.4% of the predicted datapoints were more than a factor five below the observed datapoints (underestimated), while 8.1% of all datapoints were more than a factor five above the observed data (overestimated). 0.07% of the predicted datapoints were a perfect fit, which may indicate overfitting of the model. Individual observed-predicted plots can be found below.

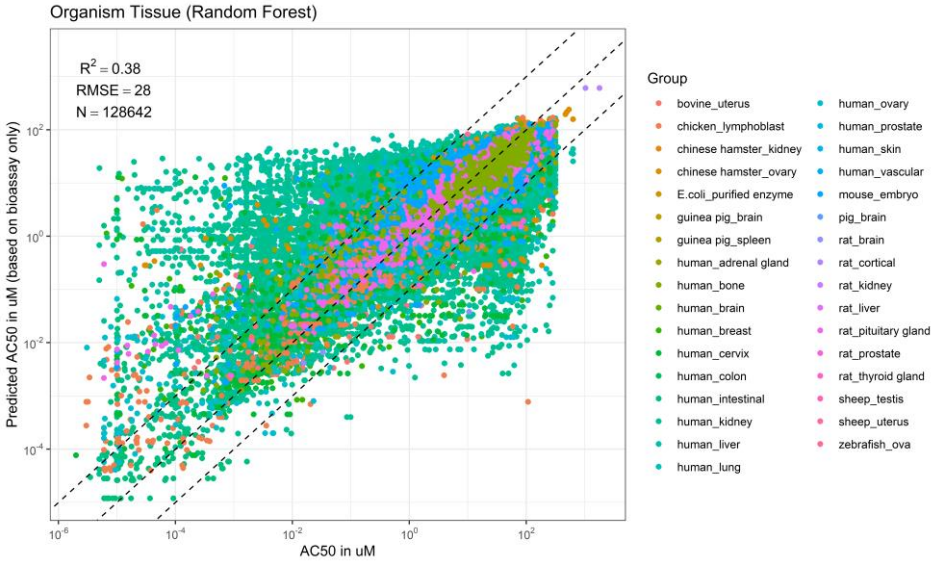


Figure 40: Predicted toxicity (AC_{50} in µM) by the Random Forest model versus observed toxicity, based on five physicochemical parameters (log K_{ow}, log K_{oc}, biodegradation rate, vapor pressure, and molecular weight), clustered per in vitro assay type (based on organism-tissue combination). The middle dashed line represents the 1:1 ratio. The outer dashed lines represent the 1:5 ratio and 5:1 ratio (the predicted values are 5 times higher/lower than the observed data).

Multiple linear regression model

In general, the multiple linear regression analysis, including the five most important predictive physicochemical descriptors (log K_{oc}, log K_{ow}, biodegradation rate (half-life in days), vapor pressure and molecular weight) as explanatory variables resulted in explaining 10% (median: 9% % , S.E.:0.02%) of all variance in the toxicity data (AC_{50}s) when categorizing *in vitro* assays based on organism-tissue combination, based on the adjusted R^2. Overall, when grouping *in vitro* assays based on organism-tissue combination, the highest % of variances explained were determined for *in vitro* assays based on rat kidney cells (31.8%), while the lowest % variance explained by the multiple linear regression model were found for *in vitro* assays based on guinea pig spleen cells (-10%) (Table 1).

Figure 41 shows the predicted effect concentrations (log_{10} AC_{50}s) plotted against the observed effect concentrations, based on the multiple linear regression model, taking the aforementioned five physicochemical parameters as explanatory variables (Equation 2), for all organism-tissue combinations, separately.

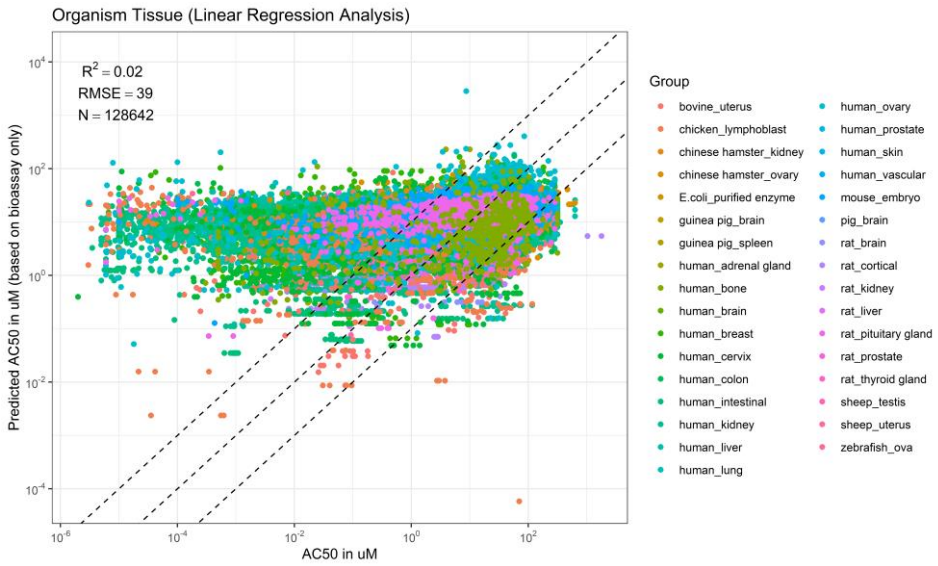


Figure 41: Predicted toxicity (AC_{50} in µM) by the multiple linear regression model versus observed toxicity, based on five physicochemical parameters (Log K_{ow}, Log K_{oc}, biodegradation rate, vapor pressure, and molecular weight), clustered per in vitro assay type (based on organism-tissue combination). The middle dashed line represents the 1:1 ratio. The outer dashed lines represent the 1:5 ratio and 5:1 ratio (the predicted values are 5 times higher/lower than the observed data).

Figure 42 shows the predicted effect concentrations ($\log_{10}$ AC_{50}s) plotted against the observed effect concentrations, based on both the multiple linear regression model and the Random Forest model, covering all individual intended target families. Overall, the Random Forest model had a higher predictive power (R^2 = 0.38, Figure 40) than the multiple linear regression model (R^2 = 0.02, Figure 41), implying that the correlation between toxicity and the five physicochemical parameters of chemicals may be non-linear, when subdividing *in vitro* assays based on organism-tissue combination. In total, 63.1% of all individual predicted AC_{50}s lied within a factor 5 of the observed AC_{50}s; 21% of the predicted datapoints were more than a factor five below the observed datapoints (underestimated), while 15.9% of all datapoints were more than a factor five above the observed data (overestimated). None of the predicted datapoints were a perfect fit.

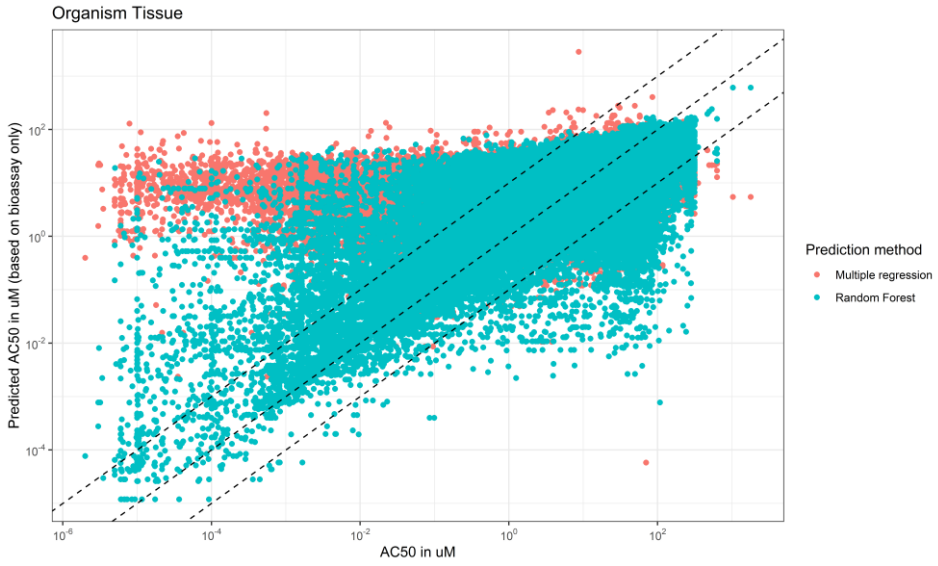


Figure 42: Predicted toxicity (AC_{50} in µM) by the multiple linear regression model versus observed toxicity, based on five physicochemical parameters (log K_{ow}, log K_{oc}, biodegradation rate, vapor pressure, and molecular weight), clustered per in vitro assay type (based on organism-tissue type). The middle dashed line represents the 1:1 ratio. The outer dashed lines represent the 1:5 ratio and 5:1 ratio (the predicted values are 5 times higher/lower than the observed data).

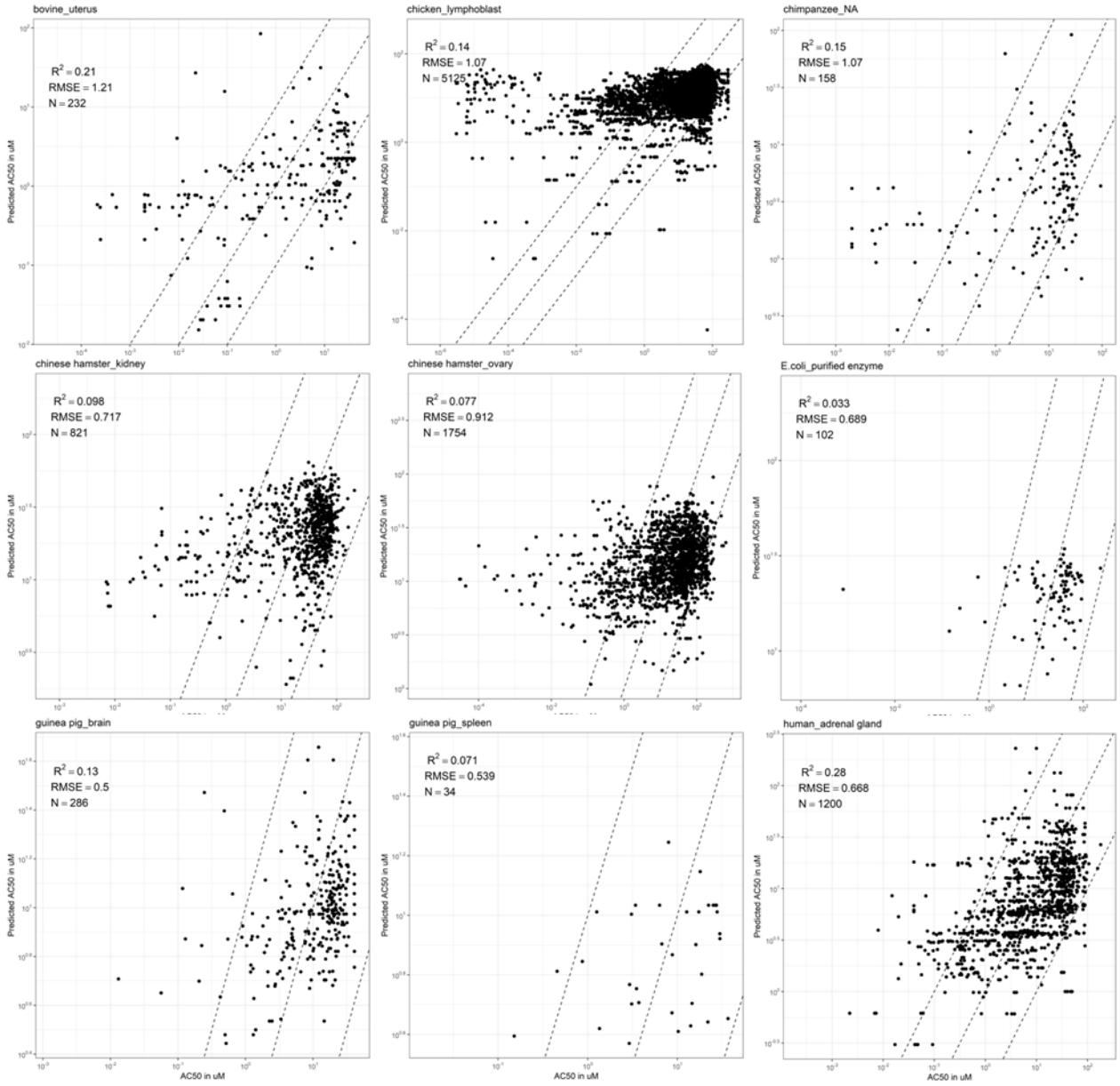


Figure 43: Predicted toxicity (AC_{50} in µM) by the multiple linear regression model versus observed toxicity, based on five physicochemical parameters (log K_{ow}, log K_{oc}, biodegradation rate, vapor pressure, and molecular weight), clustered per individual in vitro assay type (based on organism-tissue combination). The middle dashed line represents the 1:1 ratio. The outer dashed lines represent the 1:5 ratio and 5:1 ratio (the predicted values are 5 times higher/lower than the observed data).

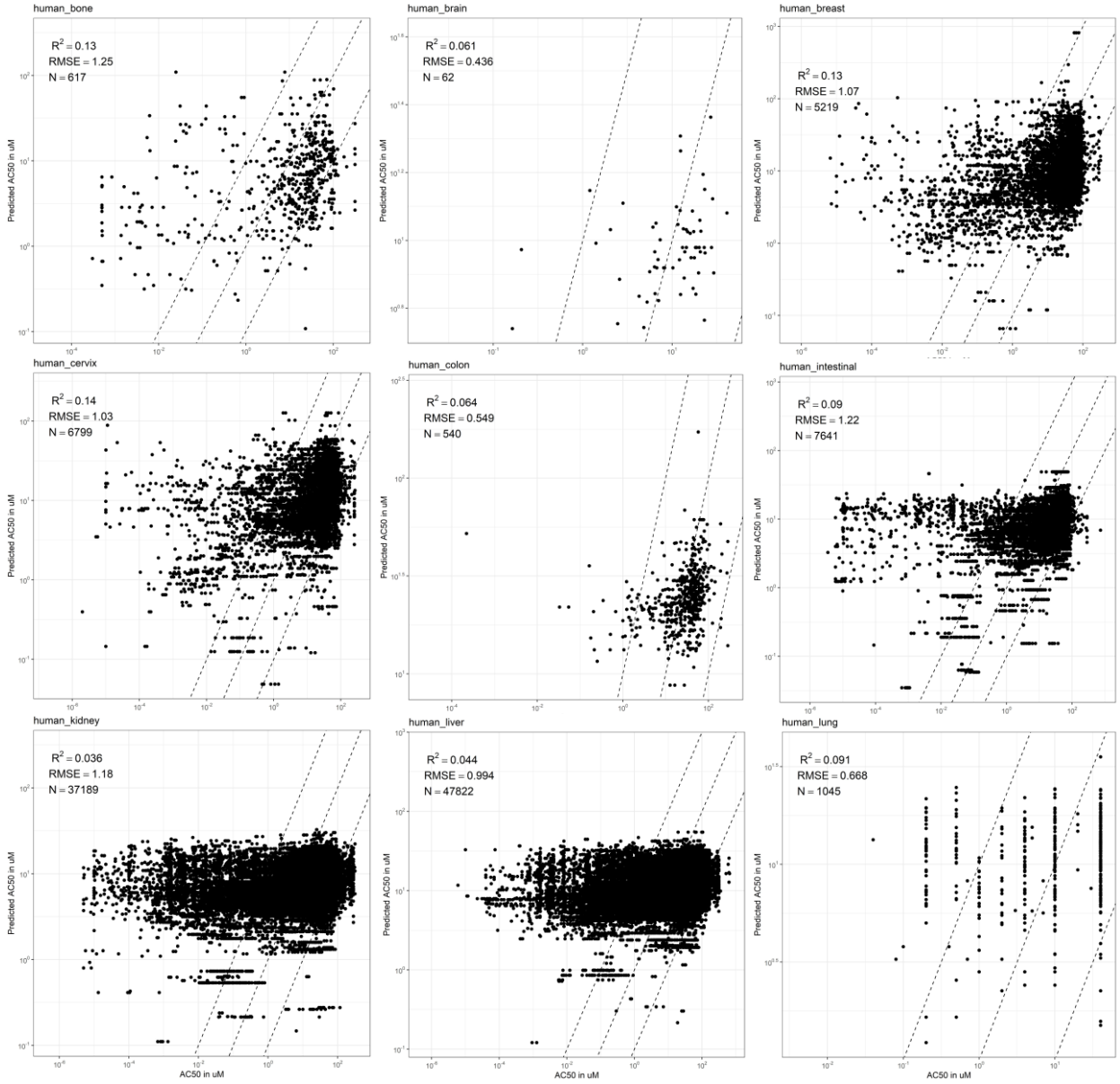


Figure 43 continued.

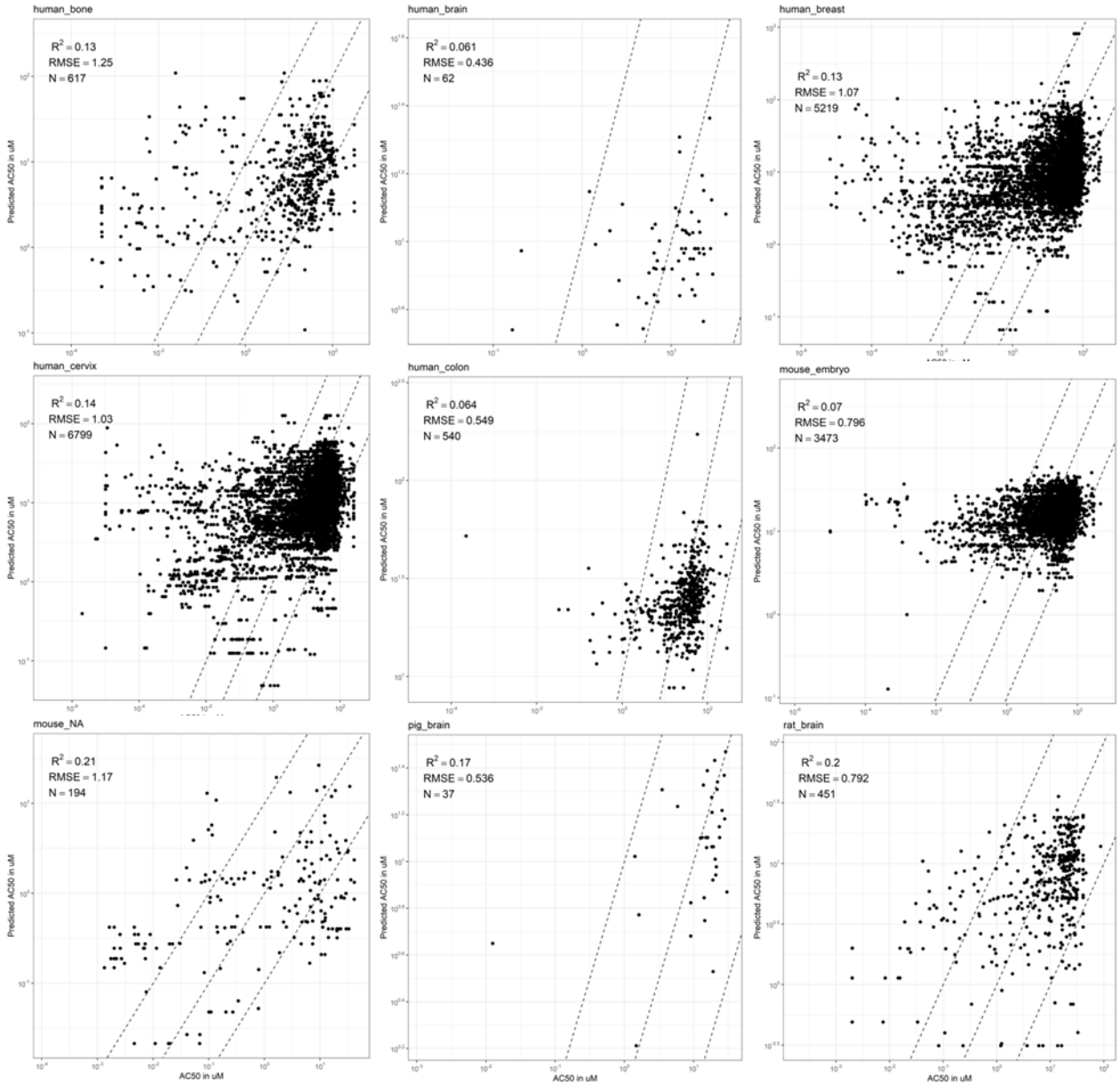


Figure 43 continued.

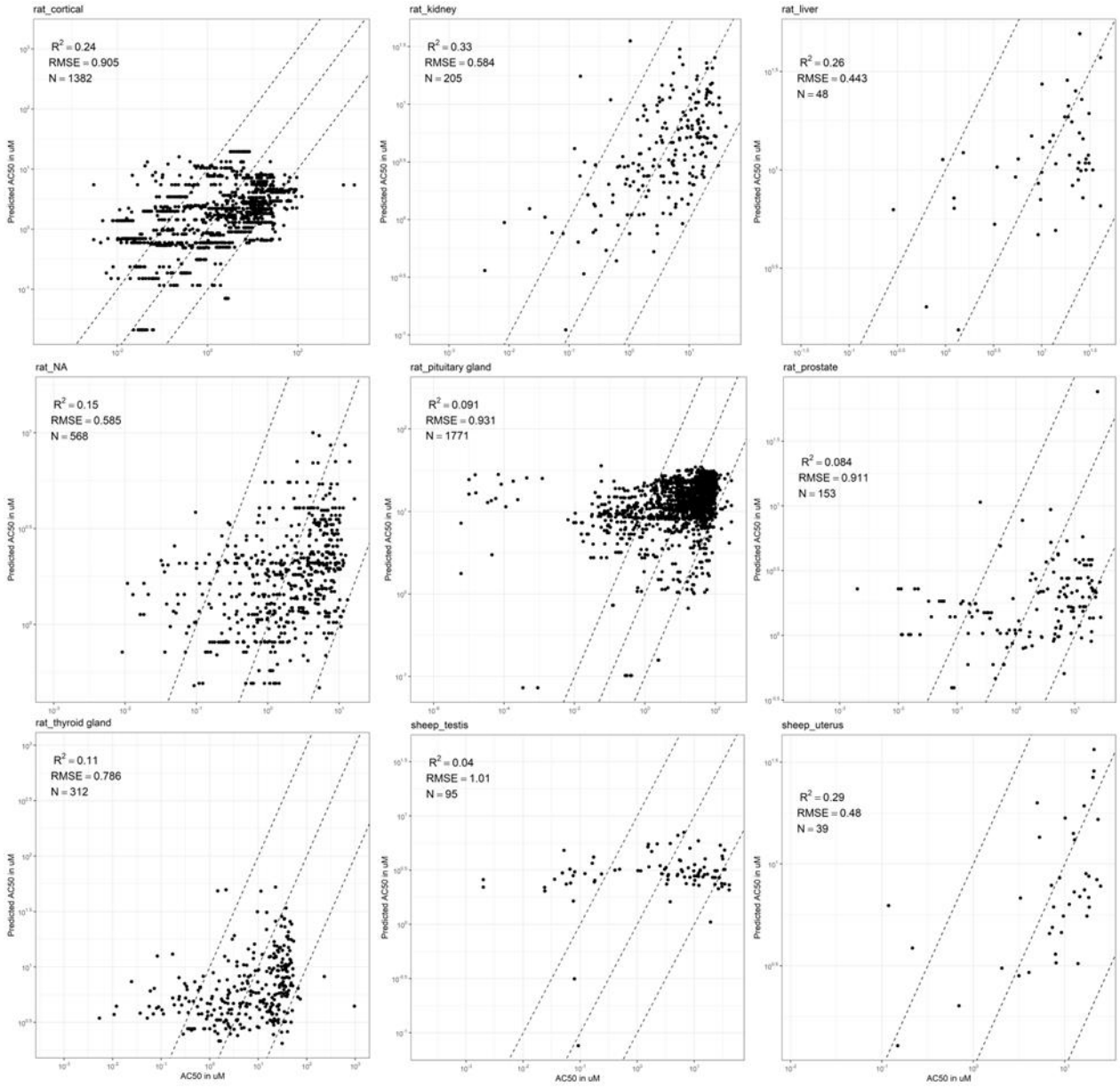


Figure 43 continued.

KWR

Memo

II Appendix: A preliminary Adverse Outcome Pathway analysis for PMOCs

II.I Introduction

Integrating knowledge of various biological reactions at molecular levels due to toxicants has attracted more attention in the field of risk assessment. Adverse outcome pathways (AOPs) were proposed as a conceptual framework to organize existing scientific knowledge by Ankley et al. 2010. These are models that identify the sequence of molecular and cellular events required to produce a toxic effect when an organism is exposed to a substance. AOPs consist of various key events (KEs) starting with a molecular initiating event (MIE) to lead to an adverse outcome (AO) that is relevant to a risk assessment context such as survival. A substantial effort has been made to enhance the AOPs for many chemicals, and identified AOPs are collected in the online database AOP-Wiki (https://aopwiki.org/), which is hosted by the Society for the Advancement of Adverse Outcome Pathways.
Additionally, more studies have been conducted to enhance AOPs search by exploring associations between stressors and KEs from scientific literature. The original method was applied to bisphenol A substituents and pesticides first (Carvaillo et al. 2019), after which it was developed into the web server (Jornod et al. 2022) and an updated version of a tool, AOP-helpFinder 2.0, which highlights features to facilitate to search and interpret AOPs more easily (Jaylet et al. 2023). This tool is based on natural language processing (text mining) to search keywords in scientific literature stored in PubMed database, by screening abstracts. The search result is provided with a score to support the weight of evidence approach (Hardy et al., 2017). The AOP-helpFinder has contributed already to several investigations of the mechanisms of exposure to per- and polyfluoroalkyl substances (PFAS) (Gundacker et al. 2022; Kaiser et al. 2022) Here, we explored AOPs related to PMOCs. The AOP-helpFinder was employed in the analysis to search for possible related AOPs from a wide range of previous studies in PubMed.

II.II Method

The AOP-helpFinder 2.0 (Jaylet et al. 2023) was used to explore the pathways that can be related to PMOCs, by matching a number of research articles stored in the PubMed database, using the webserver (Jornod et al. 2022). The stressor event analysis enables us to find AOPs that may have links to the target chemicals. The analysis requires two types of data: chemical names and event names. The chemicals are in this case the list of 1119 PMOCs as described in section 2.1 in the main report (BTO xxxx.xx – A deeper understanding of PMOC toxicity). The events indicate biological events related to AOPs, such as MIE, KE, and AO. For this study, the event names were taken from Kaiser et al. (2022), who conducted an AOP analysis to address associations between PFAS exposure and metabolic health outcomes. This list of events related to metabolism is shown in Table 3. The stressor event analysis was performed according to the default setting, i.e. the search was performed in the full abstract of research articles from PubMed, without a lemmatization process. By skipping the lemmatization process, the terms are kept in their natural forms without standardizing them to their root or base. Confidence scores were assigned to each combination of the stressors (1119 PMOCs) and the events based on the p-value derived from a Fisher's exact test. This metric was utilized to assess whether an occurrence demonstrates a higher frequency of association with a stressor (stressor-event) in contrast to another event (event-event). The scores were divided into five categories to facilitate the interpretation of the results: Low, Quite Low, Moderate, High, and Very High (Jaylet et al. 2023).

Table 3: List of the events used for the stressor event analysis (taken from Kaiser et al. 2022)

Insulin Resistance Syndrome	Syndrome X	Dysmetabolic Syndrome X
Type 2 Diabetes Mellitus	Insulin Sensitivity	Glucose Intolerances
Insulin Resistance	Metabolic syndrome	Dyslipidemias
Hyperlipidemias	High Blood Pressure	Hypertension
Central Obesity	Liver Diseases	Thyroid Diseases
Metabolic Cardiovascular Syndrome	Hyperglycemia	Dyslipoproteinemias
Abdominal Obesity		

II.III Results and discussion

Among the 1119 PMOCs, 479 chemicals were found in the stressor event analysis at one or more occurrences. The count of the occurrences, i.e., the number of links indicating associations between each chemical and stressor event , was about 1837 on average, with a range of 1–75561. The distribution of the events over the found stressors, i.e., the studied PMOCs (Figure 4),was based on 216,008 PubMed articles that provided one or more links. This indicates that abundant scientific literature was employed in this search. The event “Hypertension” had the largest number of links (Figure 45). The second largest number of links was found with the event “Type 2 Diabetes Mellitus”, and those two top links accounted for more than 50% of the total links. This means that the list of PMOCs was most commonly associated with these events in scientific literature. Figure 46 shows with which chemicals those events were often associated and the confidence score for each relationship. The confidence scores have five levels, and the results indicate that the Type 2 Diabetes Mellitus had the higher scores (“Very High”) in their links, compared with the links of the other events. Among 9101 combinations resulting from 479 stressors and 19 events, 3011, 30, 59, 32, and 250 stressor-event pairs were found to have confidence scores of Low, Quite Low, Moderate, High, and Very High, respectively (no links were found in the rest of the 5708 pairs). This indicates that the links between the stressors and the events were not statistically significant in most cases. To discuss the results more carefully, examining original literature would be essential; however, it should be noted that a systematic approach should be designed prior to the analysis of a multitude of studies. Overall, the current analysis suggests that the PMOCs we retrieved from several databases could be associated with various metabolic pathways. The usefulness of the AOP-helpFinder 2.0 was highlighted as a screening tool to search possible relevant metabolic pathways, which would be helpful for further risk assessment for these chemicals.

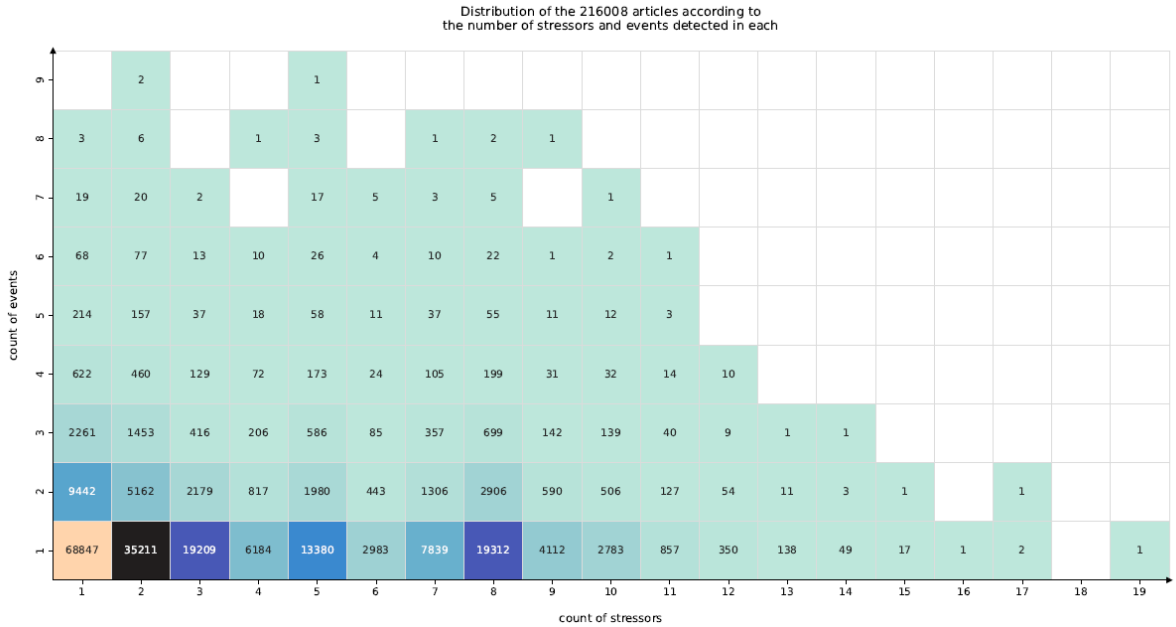


Figure 44: Distribution of the 216,008 articles according to the number of stressors and events detected in each.

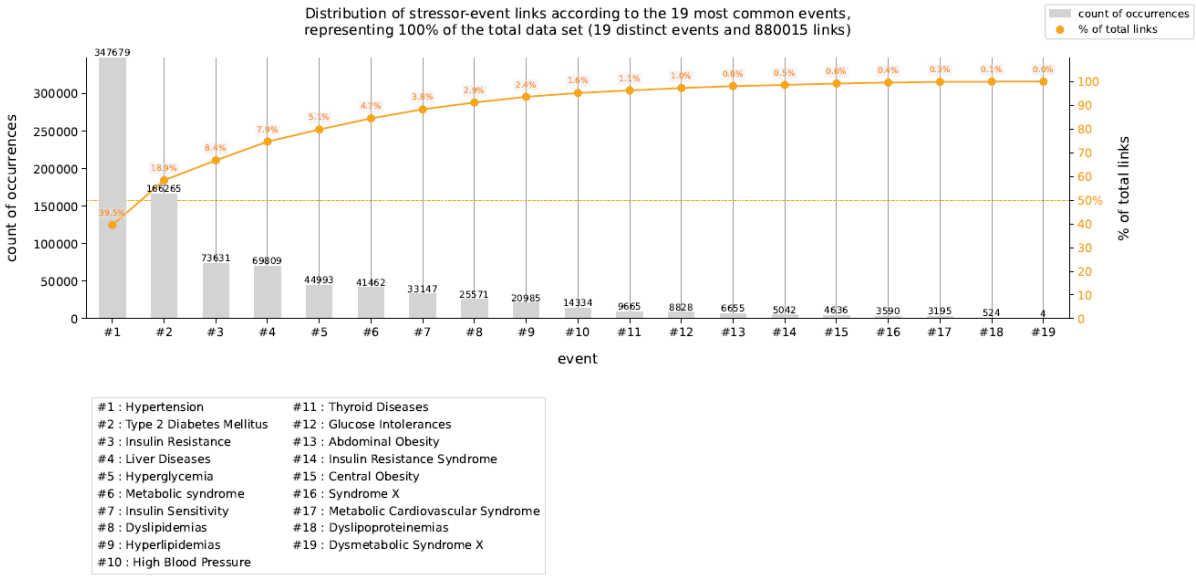


Figure 45: Distribution of stressor-event links according to the 19 most common events, representing 100% of the total data set (19 distinct events and 880,015 links).

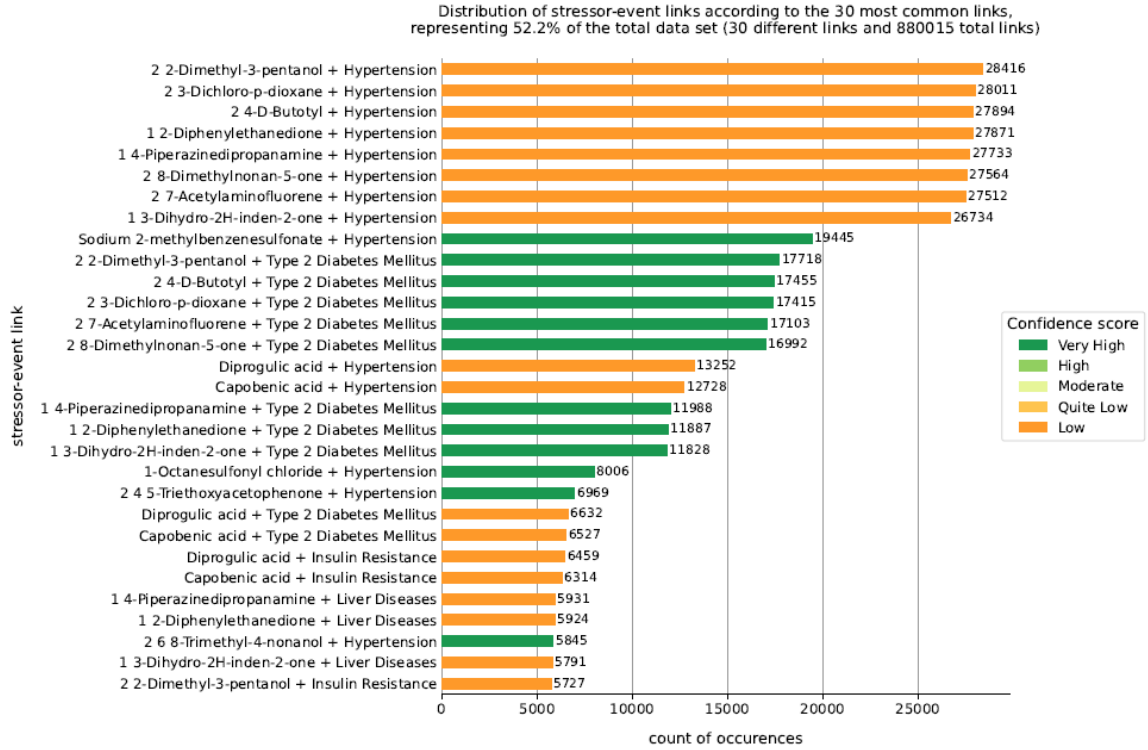


Figure 46: Distribution of stressor-event links according to the 30 most common links

Although Nelms et al. (2018) showed that AC_{50} results may be combined with AOPs in order to come to a more complete risk assessment, combining ToxCast data with AOP information was outside the scope of this present study. Future studies can explore possibilities to use experimental and predicted AC_{50} data from ToxCast for water relevant compounds in AOP pathway frameworks to relate *in vitro* toxicity (bioactivity) data to adverse effects *in vivo*.

II.IV References

Ankley GT, Bennett RS, Erickson RJ, Hoff DJ, Hornung MW, Johnson RD, Mount DR, Nichols JW, Russom CL, Schmieder PK. 2010. Adverse outcome pathways: a conceptual framework to support ecotoxicology research and risk assessment. Environ Toxicol Chem. 29(3):730–741.

Carvaillo J-C, Barouki R, Coumoul X, Audouze K. 2019. Linking bisphenol S to adverse outcome pathways using a combined text mining and systems biology approach. Environ Health Perspect. 127(4):47005.

Gundacker C, Audouze K, Widhalm R, Granitzer S, Forsthuber M, Jornod F, Wielsøe M, Long M, Halldórsson TI, Uhl M, et al. 2022. Reduced Birth Weight and Exposure to Per- and Polyfluoroalkyl Substances: A Review of Possible Underlying Mechanisms Using the AOP-HelpFinder. Toxics. 10(11). doi:10.3390/toxics10110684. http://dx.doi.org/10.3390/toxics10110684.

Jaylet T, Coustillet T, Jornod F, Margaritte-Jeannin P, Audouze K. 2023. AOP-helpFinder 2.0: Integration of an event-event searches module. Environ Int. 177:108017.

Jornod F, Jaylet T, Blaha L, Sarigiannis D, Tamisier L, Audouze K. 2022. AOP-helpFinder webserver: a tool for comprehensive analysis of the literature to support adverse outcome pathways development. Bioinformatics. 38(4):1173–1175.

Kaiser A-M, Zare Jeddi M, Uhl M, Jornod F, Fernandez MF, Audouze K. 2022. Characterization of Potential Adverse Outcome Pathways Related to Metabolic Outcomes and Exposure to Per- and Polyfluoroalkyl Substances Using Artificial Intelligence. Toxics. 10(8). doi:10.3390/toxics10080449. http://dx.doi.org/10.3390/toxics10080449.