QSAR modeling of ToxCast assays relevant to the molecular initiating events of AOPs leading to hepatic steatosis

Domenico Gadaleta, Serena Manganelli, Alessandra Roncaglioni, Cosimo Toma, Emilio Benfenati, Enrico Mombelli

QSAR modeling of ToxCast assays Relevant to the Molecular Initiating Events of AOPs Leading to Hepatic Steatosis

Domenico Gadaleta[†], Serena Manganelli[†], Alessandra Roncaglioni[†], Cosimo Toma[†], Emilio Benfenati[†], Enrico Mombelli[‡,]*

†Laboratory of Environmental Chemistry and Toxicology, Department of Environmental Health Sciences, IRCCS - Istituto di Ricerche Farmacologiche Mario Negri, Via la Masa 19, 20156 Milano, Italy

‡Models for Ecotoxicology and Toxicology Unit (DRC/VIVA/METO), Institut National de l'Environnement Industriel et des Risques (INERIS), 60550 Verneuil en Halatte, France

* Corresponding author enrico.mombelli@ineris.fr

ABSTRACT: Non-alcoholic hepatic steatosis is a worldwide epidemiological concern since it is among the most prominent hepatic diseases. Indeed, research in toxicology and epidemiology has gathered evidence that exposure to endocrine disruptors can perturb cellular homeostasis and cause this disease. Therefore, assessing the likelihood of a chemical to trigger hepatic steatosis is a matter of the utmost importance. However, systematic *in vivo* testing of all the chemicals humans are exposed to is not feasible for ethical and economical reasons. In this context, predicting the molecular initiating events (MIE) leading to hepatic steatosis by QSAR modeling is an issue of practical relevance in modern toxicology.

In this article, we present (Q)SAR models based on random forest classifiers and DRAGON molecular descriptors for the prediction of *in vitro* assays that are relevant to MIEs leading to hepatic steatosis. These assays were provided by the ToxCast program and proved to be predictive for the detection of chemical-induced steatosis. During the modeling process, special attention was paid to chemical and toxicological data curation. We adopted two modeling strategies (undersampling and balanced random forests) to develop robust QSAR models from unbalanced datasets. The two modeling approaches gave similar results in terms of predictivity and most of the models satisfy a minimum percentage of correctly predicted chemicals equal to 75%. Finally, and most importantly, the developed models proved to be useful as an effective *in silico* screening test for hepatic steatosis.

INTRODUCTION

Hepatic steatosis, also known as fatty liver disease, is defined by an intrahepatic fat content to 5% or more of liver weight[1]. It is a widespread liver pathology and according to epidemiological investigations about 20-30% of people in Western countries are affected by nonalcoholic hepatic steatosis with an incidence that ranges between 2 and 1000 person-years[2].

Several degrees of severity have been identified according to the percentage of lipid accumulation and steatosis is considered as light (5% to 33%), moderate (33% to 66%) or severe (more than 66%)[2]. The disease represents a major health risk since it is the most frequent reason for altered enzymological activity in the liver and is associated with type 2 diabetes, dyslipidemia and obesity. In some cases nonalcoholic hepatic steatosis can also lead to serious diseases such as cirrhosis and hepatocellular carcinoma.[3]

From a toxicological point of view, the disease can be initiated by chemicals present in the environment that can induce adverse effects, such as the accumulation of fatty acids in the liver.[3, 4] This class of chemicals is part of the endocrine disrupting chemicals (EDC) and, according to some studies, an early-life exposure to EDC may increase the risk of developing hepatic steatosis in adulthood. [3]

EDC comprise several chemical classes (e.g. plasticizers, polychlorinated biphenyls) and include natural and industrial chemicals. Their biological activity is linked to the same mechanism of interference with biological homeostasis and this involves their binding to transcription factors (TF).[5, 6] This mechanism, means that the liver is one of the privileged targets for these chemicals since it expresses many TF involved in hepatic lipid metabolism.[3]

Nevertheless, it is very difficult to forecast the disrupting potency of chemicals of interest because of their structural heterogeneity and because of the wide range of intracellular targets and

pathways they can interact with. In this context, the possibility of predicting the potential of chemicals to interact with TF whose function is directly, or indirectly, linked to hepatic steatosis is of great practical importance. The objective can be achieved thanks to the development of (Quantitative) Structure-Activity Relationships ((Q)SAR) that describe the relationship between the structure of chemicals and their biological activity by means of a mathematical algorithm.[7]

However, it is often hard to develop predictive QSARs when different mechanisms of action contribute in inducing the apical adverse effect.[8] In the light of this toxicological complexity, several authors have recently highlighted the potential of QSARs in predicting the ability of chemicals to act on single molecular initiating events (MIE) upstream of a more complex apical endpoint, such as steatosis.[6, 9-11]

MIEs are an important piece of information associated with the conceptual scheme of Adverse Outcome Pathways (AOP). This is a sequence of events that starts from a MIE (e.g. activation of a receptor), proceeds through a sequential series of key events at cellular and sub-cellular levels (e.g. up-regulation of fatty acid translocase) and ends with an *in vivo* adverse outcome (e.g. hepatic steatosis).[12]

AOPs are of direct importance since, when an AOP is available for an *in vivo* endpoint, predictions based on structural analogies (read-across predictions) can be substantiated by generating data that characterize the MIE or other downstream key-events.[11] For instance, for well-established AOPs, integrating the read-across and *in vitro* assays was already proved effective by Strickland et al. for skin sensitization.[13]

For these reasons, AOPs are now central to several initiatives and this plays an important role in the framework of the EU-ToxRisk project[14] that prompted the present study. (Q)SAR models describing MIEs are particularly important since they define the chemical space that can elicit the

AOP they are associated with[15] and they model a biological phenomenon which is a major determinant of *in vivo* results.[6] In addition, because of the direct link between AOPs and Integrated Approaches to Testing and Assessment (IATA), (Q)SAR models predicting MIEs are also valuable for chemical safety assessment.[9]

The aim of the present work was to develop QSAR models as a function of DRAGON[16] molecular descriptors to predict the biological activities of a series of TF identified as MIEs of AOPs leading to hepatic steatosis. Experimental data for these MIEs (the independent variable of the QSAR models) were retrieved from a collection of high-throughput *in vitro* reporter gene assays conducted as part of the ToxCast program[17]. The results produced in this program offer the great advantage of making available data on over 700 high-throughput assays for several hundreds of chemicals.[17]

This wealth of information is very important for developing a comprehensive framework of *in silico* models. Previous results indicate that effective predictive (Q)SAR models can be implemented thanks to high-throughput screening (HTS) data in order, for instance, to assess and prioritize thousands of chemicals with regard to their ER-related activity[18].

In this study, we developed a collection of (Q)SAR models to predict these MIEs on the basis of these HTS data and as a function of DRAGON chemical descriptors.[16]

Since most of these data were extremely unbalanced, i.e. with a very low proportion of active chemicals, appropriate modeling techniques were employed to avoid the classification bias towards majority class examples.[19] In addition, the effect of feature selection on modeling performance was explored. Finally, a consensus approach was also investigated and it was found to be more suitable than single models to predict the activation of MIEs leading to hepatic steatosis.

For the sake of completeness it is also noted that three of the modeled targets (AhR, Nrf2 and PPARγ) were also modeled during the Tox21 challenge.[20] During which, however, data were not curated in terms of general cytotoxicity as we did during our work. In addition, as far as chemical purity is concerned, we also rejected chemicals that failed quality control including only those associated with purity greater than 90%.

At the end, the developed QSAR models were applied to screen chemicals for steatotic activity. In this regard, QSAR predictions and experimental ToxCast i*n vitro* HTS assay data showed a similar potential for screening purposes. In the light of these results, QSAR models here presented proved to be a valid support for the identification of potentially hazardous chemicals.

MATERIALS AND METHODS

Biological overview of the modeled MIEs

Information on MIE upstream of the adverse effect (i.e. hepatic steatosis) were collected from the AOPwiki[21] that provides details about several AOPs leading to hepatic steatosis (Fig. 1).

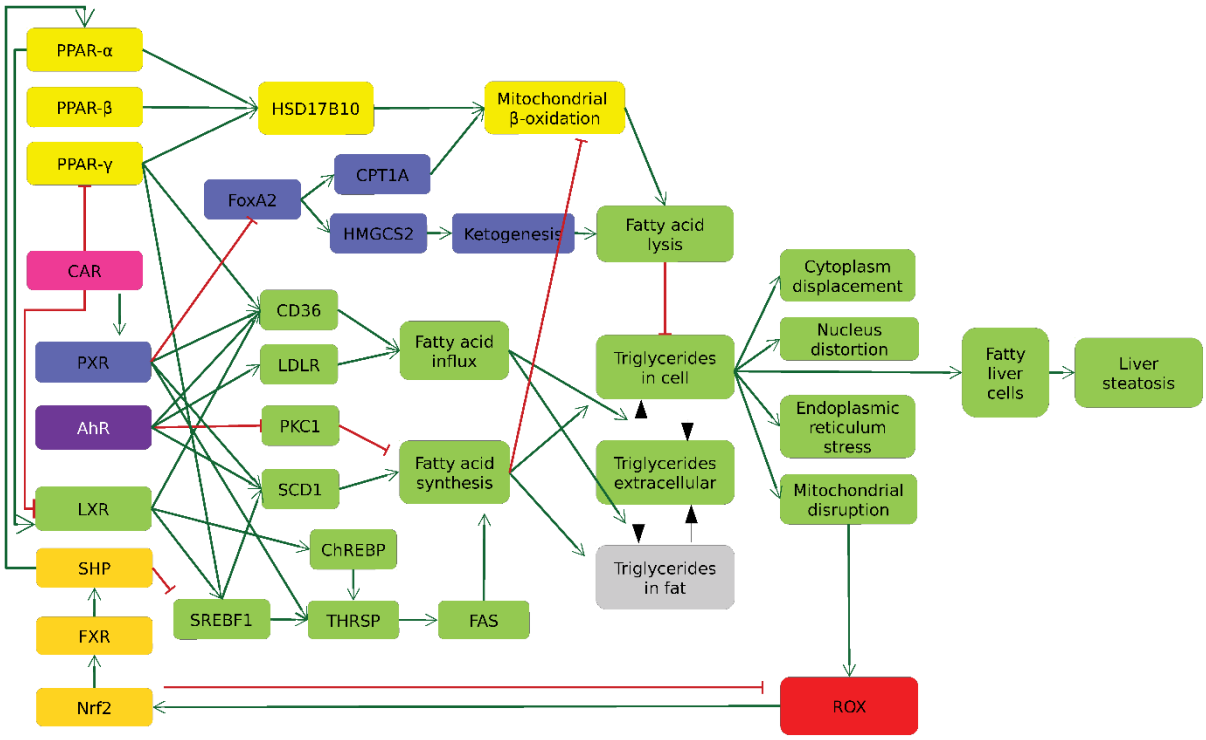

Figure 1. Schematic depiction of AOPs leading to hepatic steatosis. Green and red lines indicate activating and inhibitory effects respectively. The figure summarizes the AOP network obtained from the integration of AOPs 57, 34, 36, 60 and 61 from the AOPWiki.[21]

The following cellular targets were identified as potential MIE for our modeling analysis:

- the peroxisome proliferator-activated receptors (PPARα, PPARβ, PPARγ);
- the constitutive androstane receptor (CAR);
- the pregnane X receptor (PXR);
- the aryl hydrocarbon recetor (AhR);

- the liver X receptor (LXR);
- the Nuclear factor (erythroid-derived 2)-like 2 (Nrf2);
- the Farnesoid X receptor (FXR)

As reported in AOPwiki[21] disclaimer no definitive mechanistic understanding of the role in hepatic steatosis of the MIEs we modeled has yet been achieved. However, to provide the reader with a biological insight about the current state of knowledge of the modeled MIEs, we summarize here this role in hepatic steatosis.[21] Readers should refer to Figure 1 for a synthetic depiction of the intersecting AOPs leading to hepatic steatosis.

Some MIEs that lead to hepatic steatosis were not modeled because of a lack of data of sufficient quality (FXR, CAR, PPARβ, see the methodological section) or because there was not a sufficient weight of evidence supporting them.

AhR (AOP 57: AhR activation leading to hepatic steatosis)

The activation of this receptor directly leads to up-regulation of fatty acid translocase CD36 (FAT/CD36) which is a scavenger protein-mediating uptake and intracellular transport of long-chain fatty acids (FA). Indirectly, the activation of AhR also contributes to hepatic steatosis by a) inhibiting mitochondrial fatty acid beta-oxidation, b) mediating the induction of stearoyl-CoA desaturase (SCD1), c) decreasing the expression of phosphoenolpyruvate carboxykinase 1 (PKC1), (a control point for glycolysis/gluconeogenesis pathway) and d) increasing the uptake of low-density lipoprotein (LDL receptor).

LXR and PPARγ (AOP 34: LXR activation leading to hepatic steatosis)

This AOP describes the linkage between hepatic steatosis triggered by nuclear receptors activation (PPARγ and LXR). The liver X receptor (LXR) regulates the homeostasis of cholesterol, fatty acid, and glucose. Activation of this receptor leads to the induction of the following targets: fatty acid translocase CD36, the sterol regulatory element-binding protein (SREBP-F1c), carbohydrate-responsive element-binding protein (ChREBP), fatty acid synthase (FAS) and stearoyl-CoA desaturase (SCD1). The first event leads to an increased influx of fatty acids while the others lead to *de novo* synthesis of fatty acids. Similarly, activation of PPARγ directly leads to an up regulation of the fatty acid translocase CD36.

PPARα (AOP 36: Peroxisomal fatty acid beta-oxidation inhibition leading to steatosis)

PPARα activation increases the catabolism of fatty acids by inhibiting the accumulation of triglycerides. The AOPwiki indicates that fatty acid oxidation in liver tissue is controlled by PPARα signaling networks. The PPARα signaling network controls expression of the genes in metabolic pathways that catalyze fatty acid oxidation reactions. Down-regulation of PPARα hydroxysteroid (17β) dehydrogenase 10 (HSD17B10) inhibits mitochondrial β−oxidation.

PXR (AOP 60: NR1I2 (pregnane X receptor) activation leading to hepatic steatosis)

PXR activation directly leads to up regulation of CD36 and of stearoyl-CoA desaturase. These two events cause respectively hepatic accumulation of triglycerides and fatty acids. The activation of PXR also inhibits the forkhead box protein A2 (FoxA2) whose final effect is a reduction of fatty acid lysis.

Nrf2 (AOP 61: NFE2L2/FXR activation leading to hepatic steatosis)

The TF tor the erythroid 2-related factor 2 Nrf2 triggers the expression of genes that protect cells from oxidative and electrophilic stress. Nrf2 is also a negative regulator of genes that promote steatosis.[22]

The activation of Nrf2 directly leads to the induction of the Farnesoid X receptor which, in turns, results in activation of the Small Heterodimer Partner and PPARα. These two last events then protect against steatosis by inhibiting the sterol regulatory element-binding protein and increasing the catabolism of fatty acids.

ToxCast data

The models described here were obtained on the basis of the ToxCast data corresponding to the October 2015 release.[23] ToxCast data were downloaded from the EPA website (https://www.epa.gov/chemical-research/toxicity-forecaster-toxcasttm-data), containing a collection of files with information for more than 8,000 unique substances and DSSTox standard chemical fields (chemical name, CASRN, structure, etc.) for EPA ToxCast chemicals and the larger Tox21 chemical list. This information included also quality control grades for chemicals, details on the assays and results summarized by AC_{50} values.

Curation of toxicological data

Experimental data used in this work were isolated from a collection of 24 *in vitro* HTS assays from the ToxCast program, executed by Attagene Inc. (RTP, NC), under contract to the U.S. EPA (Contract Number EP-W-07-049). During this program, several experiments evaluated the impact

of more than 8,000 chemicals on the previously described TFs involved in the MIE of steatosis AOP.[24]

Assays were designed to combine libraries of so called CIS- and TRANS-regulated TF reporter constructs.

In 'CIS-assays' all members of the TF family recognize the same binding sequence in the promoter, so these assays evaluate the integral activity of the entire TF family (e.g., the entire family of PPARs) and do not distinguish specific receptor isotypes.[25]

Conversely, 'TRANS-assays' use a library of hybrid reporter constructs specific for each isoform of a given TF family so that individual members of a TF family can be distinguished (e.g., PPARα, PPARγ, and PPARβ).[25]

Both down and up regulation assays were considered separately for evaluating agonistic and antagonistic activity towards a given TF, respectively. Endpoints referring to the activation of the receptor are indicated by the suffix "up" at the end of the name of the modeled TF whereas endpoints referring to a deactivation are indicated by the suffix "dn" (e.g. PXR_up and PXR_dn when referring the pregnane x receptor).

In all cases, a micromolar concentration for each chemical–assay combination was reported as the negative logarithm of the half-maximal activity concentration (pAC_{50}). Zero values indicate inactive and tested chemical-assays combination. For classification purposes, chemicals with zero values for a given assay were considered as inactive, while chemicals with a continuous pAC_{50} value were considered active. Results from TRANS-assays were considered if specific isoforms of a TF listed among the AOPs for steatosis (e.g., PPARα, PPARγ), while CIS-assays or a combination of CIS- and TRANS-assays were used if TF isoforms were not specified. In the latter case, CIS- and TRANS- outputs were combined to obtain more balanced datasets by increasing

the proportion of active chemicals. A chemical was labeled as active if it was active in at least one assay and inactive if it was inactive in both types of assay.

Table S2 in Supporting Information summarizes the original list of selected assays, their classification as CIS- or TRANS-assays, and the number of active and inactive chemicals for each assay.

Chemical structure curation

The contents of ToxCast data were analyzed in depth. Analytical chemistry analysis over the course of the ToxCast project pertains to overall chemicals library management. As a result, quality control (QC) was the driving decision for the first screening of ToxCast data to be used for modeling. We retained only chemicals exceeding 90% purity and chemicals associated to lower purity, other anomalies (e.g. withdrawn chemicals) or not yet analyzed were not included.

Chemicals satisfying these criteria were extracted from the ToxCast files and assigned identifier numbers (IDs), SMILES/InChI, and experimental activities expressed as the negative logarithm of AC_{50} (of molar concentration) for the endpoints of interests.

The structures were checked by removing inorganic chemicals and mixtures, correcting inaccurate SMILES codes with the help of chemical databases, such as ChemSpider[26], ChemIDplus[27], and neutralizing salts.

An *in-house* software[28] was used to identify and remove duplicates. For a given set of duplicated structures (e.g., due to the presence of stereoisomers or salts), if their experimental activities were identical, then only one chemical was kept. If their experimental properties were different, both the chemicals were removed.

A Python script executing the MolVS standardizer (based on RDkit libraries) was written to obtain canonical tautomers. Canonical SMILES were coded using the istMolBase software[29], based on CDK libraries.

Z-score pruning

Judson et al. reported that for approximately half the chemicals tested during the ToxCast project, cytotoxicity was observed in the range of concentrations tested. Thus, a significant proportion of measured activities may represent a false positive response caused by assay interference process linked to a cytotoxicity-related 'burst' of activities.[30]

To isolate chemicals that can be considered as true positives with greater confidence, we applied the pruning strategy proposed by Judson.[30] We used the median log(AC_{50}) and the corresponding median absolute deviation (MAD) associated with positive cytotoxicity assays (between 2 and 33 assays) in combination with the AC50 of the assay of interest to compute a z-score (Eq. 1) that was assigned to each chemical.

$$Z(chemical, assay) = \frac{-logAC50(chemical, assay) - median[-logAC50(chemical, cytotoxicity)]}{global\ cytotoxicity\ MAD} \quad (1)$$

A high z-score for a given chemical-assay pair identifies a region where there is little or no superposition between cytotoxicity and observed effect; therefore this adds confidence to hit-calls that are very likely to be associated with specific biomolecular interactions (true positives). Conversely, chemicals associated with low z-scores are more likely to be false positives confounded by cytotoxicity. We took a z-score threshold of three was considered to select

chemicals that can be considered as specifically active. This cut-off was mentioned as a reasonable choice by Judson et al. for avoiding the risk of taking false positives into account.[30]

This procedure eliminated less reliable data from the original datasets at the cost of a 16-80% decrease in the number of active chemicals depending on the assay. Table S2 in Supporting Information gives a complete overview of the number of excluded chemicals for each considered assay. The superposition of data pruning for chemical purity and cytotoxicity in some cases left only a very small number of active chemicals (less than 50). In the end, nine assays relating to six TF were considered as endpoints for QSAR derivation. Table 1 summarizes the number of active and inactive chemicals for each maintainedassay.

Table 1. Summary of data relative to modeled assay responses

TF[a]	Assay type[b]		Inactive chemicals[c]	Active chemicals (z-score ≥3)[d]	% Active[e]	Excluded chemicals (z-score <3)[f]	TS_{FULL}[g]	VS[h]	TS_{US}[i]
PXR	TRANS+CIS[k]	up	529	640	55	225	934	235	934
PXR	TRANS+CIS[k]	dn	1269	83	6	42	1079	273	132
LXR	TRANS+CIS[k]	up	1296	68	5	30	1089	275	108
LXR	TRANS+CIS[k]	dn	990	141	12	263	904	227	224
AhR	CIS	up	1148	162	12	84	1045	265	258
AhR	CIS	dn	1301	50	4	43	1079	272	80
NRF2	CIS	up	723	346	32	325	853	216	552
PPARγ	TRANS	up	871	266	23	257	908	229	424
PPARα	TRANS	up	1259	64	5	71	1057	266	100

[a]Modeled TF. [b]Assay type (CIS- or TRANS-assay, up or down regulation). [c]Number of inactive chemicals. [d]Number of active chemicals (with information on the z-score). [e]Percentage of active chemicals. [f]Number of excluded chemicals whose positive result is probably caused by generalized cytotoxicity. [g]Size of the full training set. [h]Size of the validation set. [i]Size of the undersampled training set. [k]Chemicals showing activity for at least one assay were considered active for the combined assays; chemicals that were inactive in all single assays were considered inactive in the combined assay.

.

Statistical evaluation of the models

Performance in classification was evaluated using information retrieved from confusion matrices: the number of true positives (TP, i.e. chemicals that are correctly recognized as active), the number of true negatives (TN, i.e. chemicals that are correctly recognized as inactive), the number of false negatives (FN, i.e. misclassified active chemicals), the number of false positives (FP, i.e. misclassified inactive chemicals).

The Matthews correlation coefficient[31] (MCC) was used as a measure to assess the classification accuracy of the models (Eq. 2):

$$MCC = \frac{(TP \cdot TN) - (FP \cdot FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (2)$$

The MCC is a suitable statistical indicator in the presence of unbalanced datasets since it gives a particularly robust performance for different classifiers.[19] For this reason we adopted it as a quality criterion during model selection.

In addition, Cooper statistics were calculated.[32] According to these statisticsaccuracy (ACC), also called concordance, is the number of correctly predicted chemicals divided by the total number of chemicals (Eq. 3):

$$ACC = \frac{TP + TN}{TP + TN + FN + FP} \quad (3)$$

The sensitivity (SE) estimates the proportion of active chemicals that are correctly predicted (Eq. 4):

$$SE = \frac{TP}{TP + FN} \quad (4)$$

The specificity (SP) estimates the proportion of inactive chemicals that are correctly predicted (Eq. 5):

$$SP = \frac{TN}{TN + FP} \quad (5)$$

Balanced accuracy (BA) was used as a general measure of correct classification rate suitable for unbalanced datasets (Eq. 6):

$$BA = \frac{SN + SP}{2} \quad (6)$$

Finally, the Area Under a Receiver Operating Characteristic Curve (AUROC) was calculated. AUROC measures the probability that a predictor will rank a randomly chosen positive instance higher than a randomly chosen negative one. It is calculated from a Receiver Operating Characteristic (ROC) curve, that is created by plotting SE against (1-SP) at various threshold settings. The AUROC varies between 0 and 1 and under 0.5 the classifier is considered uninformative.[33]

Balanced accuracy (BA) and the AUROC enable a comparison with results obtained in the framework of the Tox21 Challenge.[20]

Internal validation procedures

Two procedures were implemented and applied in KNIME and R to test the derived QSAR model "internally": 10-fold-Cross-Validation (10-fold-CV) and Y-scrambling. Cross-Validation (CV) is a widely used statistical technique for internal validation, in which different proportions of chemicals (in our case ten segments) are iteratively held-out of the training set used for model development and "predicted" as new by the developed model in order to verify internal predictivity.[34]

Y-scrambling[35] was employed to demonstrate that the models were not the result of chance correlation. During this validation approach, the response variables were randomly shuffled *n* times (in our case 500), and the correlation between them and the descriptors was computed. In the absence of chance correlation, the performances of the scrambled models should decrease drastically.[35]

Model development

Classification models were derived for each of the nine assays (endpoints) listed in Table 1. The various steps for model development were outlined below andthe entire modeling workflow is summarized in Figure 2.

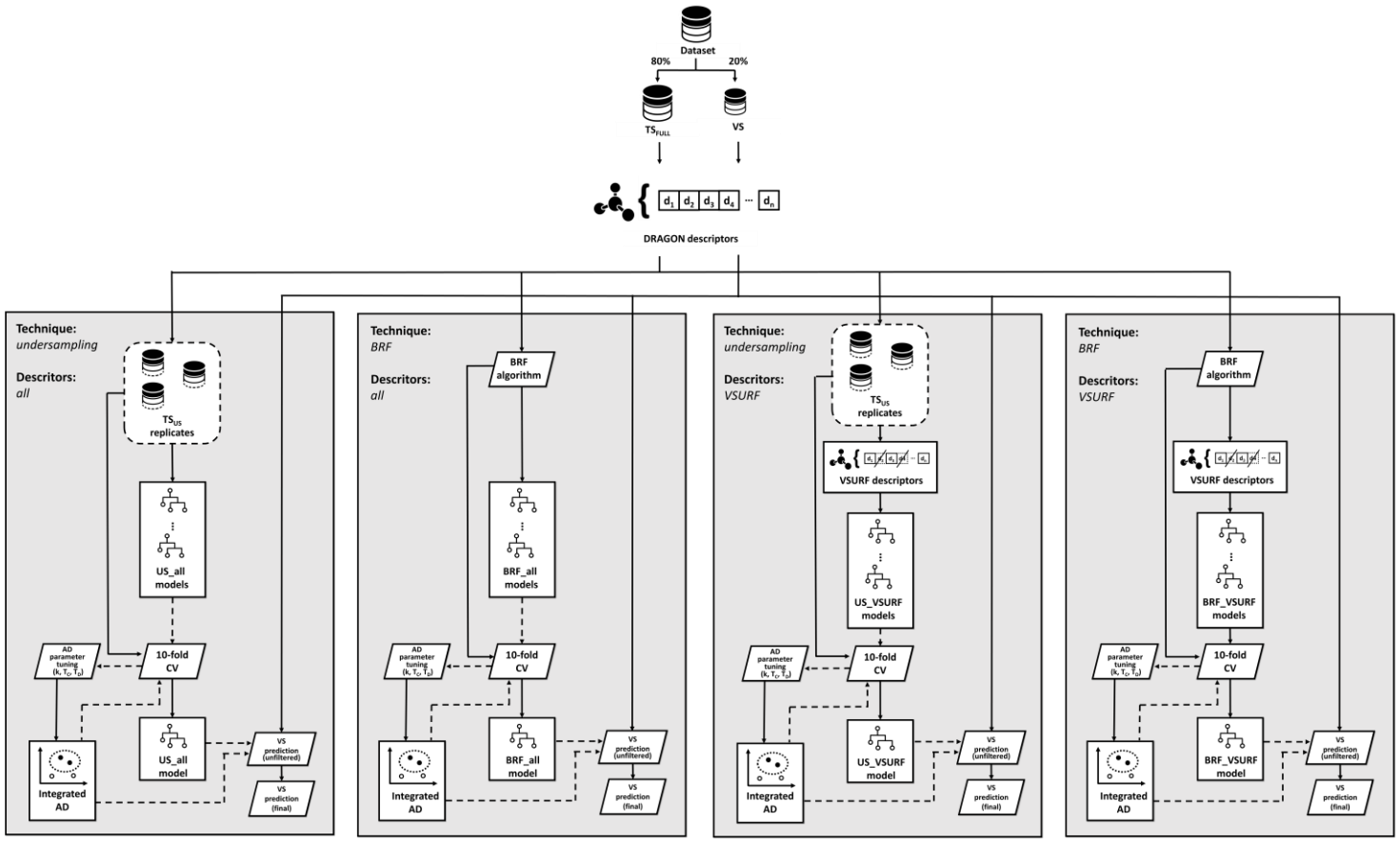

Figure 2. Modelling workflow. First, the entire dataset was split into a TS_{FULL} (80% of the initial number of chemicals) for model derivation and a VS (20% of the initial number of chemicals) for the external validation of the models. In the left-hand panels, undersampling was used to derive RF models from balanced training sets (TS_{US}, three replicates) that were subsequently validated, internally and externally. In the right-hand panels, balanced random forest (BRFs) were used to derive models from TS_{FULL}. In both cases, results from a 10-fold internal cross-validation were employed to fine-tune the parameters of the applicability domains (AD). Best models were selected considering the coverage and the final predictivity in internal and external validation.

Data split

Datasets for each endpoint were randomly divided into a training set (TS_{FULL}, 80% of the original dataset) and a validation set (VS, 20% of the original dataset) comprising the same proportion of active and inactive chemicals as the original dataset. Table 1 reports the number of chemicals in TS_{FULL} and VS for each of the modeled endpoint. A complete overview of the datasets used for model derivation in reported in Table S1 of Supporting Information.

Calculation and selection of descriptors

Molecular descriptors were calculated for each chemical with the DRAGON software.[16] Descriptors were pruned by constant and semi-constant vales (i.e. standard deviation < 0.01), then if a couple of descriptors was characterized by an absolute pair correlation greater than 90%, the descriptor with the highest pair correlation with all the other descriptors was removed.

Optimal subsets of descriptors for modeling were obtained with the R package VSURF.[36] The algorithm was applied exclusively to the training sets and it consists in a three step variable selection based on the logic underpinning the Random Forest algorithm (i.e. permutation importance and out-of-bag error). The first step eliminates irrelevant descriptors according to the permutation-based RF score of importance and a user-defined threshold. The second step finds important descriptors closely related to the response variable (interpretation step) and the third step (prediction step) identifies a sufficient parsimonious set of important descriptors leading to a good prediction of the response variables.[36] The VSURF selection procedure was carried out as a function of a number of trees ranging from 25 to 251.

Algorithm description

Preliminary analysis showed that straightforward linear approaches (i.e. linear discriminant analysis) yielded unsatisfactory results (data not shown). Therefore we decided to evaluate non-linear approaches such as RF.[37, 38]

The majority of datasets were highly unbalanced towards negative chemicals and attempts at developing predictive QSAR models as a function of classical RFs were ineffective (data not shown). Two different modeling approaches were therefore applied for each endpoint to model these highly unbalanced datasets.

- The first approach was based on undersampling[39], i.e. random deletion of the most represented class (i.e. negative chemicals) until both classes were equal in number. This approach generated a dataset (TS_{US}) more suitable for treatment with classical machine learning methods. Three replicates were generated from each dataset, so replicates for a given endpoint had the same active chemicals, but different inactive chemicals (Table 1). In the end, the model returning the best performance in 10-fold CV was retained (Table 1). RF implemented in KNIME[40] was used to derive undersampling based models.
- The second approach consisted of a Balanced Random Forest (BRF), which is a combination of under-sampling and the ensemble idea. This technique artificially alters the class distribution so that classes are represented equally in each tree.[41] The randomForest[42] R package (version 4.6-12) was used for the BRF approach.

These two modeling approaches (undersampling and BRF) were applied to the entire pool of descriptors and to the descriptors selected with the R package VSURF, for a total of four different sets of models for each endpoint. For models based on feature selection, the number of trees was selected according to indications given by VSURF. Indeed, the package enables the identification

of best descriptor subsets as a function of the number of trees. For models based on the entire pool of descriptors, the same combinations of trees probed during the optimization of VSURF output was used and the one returning the lowest error in prediction estimated by 10-fold CV was retained. In all the cases the final retained number of trees was located at the beginning of the plateau indicating a stabilization of the prediction error.

The *mtry* values used in undersampling and BRF models were those provided by default in KNIME node and R randomForest package respectively, calculated as the square root of the initial number of variables. These techniques for balancing the datasets were used for all endpoints except PXR_up regulation, which did not require their application, as its data distribution was well-balanced. For this endpoint, RF algorithms implemented in KNIME and R were used without altering the class distribution in the initial training set or in bagging samples.

Applicability Domain

For each model selected, the Applicability Domain (AD), i.e. «the response and chemical structure space in which the model makes predictions with a given reliability»[43] was defined. Predictions made by a model outside this space are the result of predictive extrapolation, and are potentially associated with greater predictive uncertainty.

Previous work showed that useful metrics for AD definition with RF are the variability of predictions among individual RF trees (a wide variation indicates less accurate predictions) and the distance to closest neighbors (long distances indicate that a chemical is not surrounded by similar structural analogs).[44] For this reason, in the present work both aspects were considered for the definition of models AD.

ADs were optimized for each model by fine-tuning the two metrics with respect to predictivity:

- For the first AD criterion we estimated the percentage of trees within the RF yielding the same prediction (i.e. confidence). A confidence threshold (T_c) was implemented in KNIME and gradually incremented by 0.05, from 0.55 to 0.75. Chemicals with confidence lower than this threshold were considered as outside the model AD.
- The second AD criterion took account of the structural domain of the model. This was achieved by evaluating the degree of structural similarity of a given chemical to those included within the TS. A distance matrix containing Euclidean distances for each pair of chemicals in the TS was calculated, then for each TS chemical the mean distance from its first k neighbors was calculated. TS chemicals were then sorted on the basis of these distances and the value corresponding to a given percentile of the distribution of distances was used as a threshold (T_D) beyond which chemicals were excluded from the AD.

For the external validations, the same procedure was repeated calculating the prediction confidence and the distances of each VS chemical from their neighbors within the TS, then T_D was used to identify chemicals outside of AD. For the present work, we adopted the Euclidean distance calculated on the scaled and centered descriptors used by the models as a similarity criterion; values assigned to k were 1 and 5; values assigned to T_D were those corresponding to the 100^{th}, the 97.5^{th}, the 95^{th} and the 90^{th} percentiles of the TS distance distributions.

Optimization of the applicability domain

For each model, the three parameters (T_C, k and T_D) were evaluated in each possible combination, for a total of 32 combinations. The best parameter combination was then selected to obtain the best compromise of MCC and AD coverage in 10-fold-CV.

- A threshold was set on the coverage in 10-fold-CV (T_{Cov_CV}), i.e. a minimum percentage of TS chemicals included within the AD during an internal 10-fold-CV. All the possible combinations of parameters (i.e. T_C, k and T_D) returning a coverage higher than T_{Cov_CV} were collected. The procedure was repeated four times, each time imposing a lower T_{Cov_CV} i.e. (from 40% to 70%, 10% step).
- The best combination of parameters was selected for each model by systematically evaluating MCC values both in internal (i.e. 10-fold cross-validation) and external validation. The evaluation started with combinations of parameters that fulfilled the highest T_{Cov_CV} (i.e. 70%). Parameters combinations returning MCCs in internal and external validation equal to a given threshold or higher (T_{MCC}) were selected. The first (highest) T_{MCC} threshold was initially set at 0.80. If more than one valid combination was found in this way, the one with the highest MCC in internal ten-fold CV was selected.
- If no models were valid for the first T_{MCC}, (i.e. 0.80), the selection was repeated by gradually reducing T_{MCC} in 0.10 steps down to 0.40.
- If no acceptable models associated with the highest T_{COV_CV} were identified, the search was repeated on models fulfilling lower a lower T_{Cov_CV} (60-40%).
- If no models were found even in this last case, the model with the highest mean MCC (considering 10-fold-CV, and VS) was retained.

Consensus modeling

Consensus modeling was also evaluated by integrating predictions of single models. Predictions from the four models were integrated with a majority vote approach.[45] A consensus prediction was produced for samples with at least three out of four concordant predictions (i.e. 75%). Conversely, chemicals with ambiguous predictions were discarded. AD of each model was also considered during the integration process. Indeed, if more than one single prediction out of four was outside the model AD, then the entire consensus prediction was considered out of AD. External validation performance were used to validate the application of the consensus strategy for each MIE.

Virtual screening of steatotic chemicals

QSAR models were evaluated for the virtual screening of steatotic chemicals. Experimental steatosis data were isolated from the *in vitro* cell morphology assay "APR_Hepat_Steatosis_48hr_up" from the ToxCast program, executed by Apredica (Watertown, Massachusetts), under contract to the U.S. EPA (Contract Number EP-D-13-013). This *in vitro* assay is a "morphology reporter" assay that enables the characterization of the regulation of steatosis in rat hepatocytes by means of fluorescent imaging of the probe LysoTracker red.[46] Indeed, LysoTracker Red is an acidophilic fluorescent dye that loads predominantly into lysosomes.[47] This specific behavior of the probe enables the monitoring of lysosomal permeabilization which is an upstream event in the cascade that leads to hepatocyte lipid overloading.[47]

Steatosis data overlapping with the datasets used for MIE modelling were considered for validation of the screening procedure. This overlapping resulted in a total of 213 chemicals with *in vitro* steatosis data (17 positives, 196 negatives) and also experimental data for the nine MIE

endpoints. This screening dataset (SS1) was first used to evaluate the correlation between steatosis and MIE experimental data. Furthermore, the analysis was extended to MIE predictions to evaluate the ability of QSAR models to identify steatotic chemicals.

A second validation of the screening procedure based on QSAR predictions was made on chemicals that were associated with experimental data on steatosis but that were not associated with experimental MIE data. In this case, chemical structures were curated with the same procedure already described in (see "Chemical structure curation" paragraph). The second screening dataset (SS2) included 90 chemicals with *in vitro* steatosis data (6 positives, 84 negatives) without experimental data for the MIE endpoints considered.

Models on MIE were evaluated for their ability to identify steatotic chemicals. Chemicals were ranked based on the percentage of activated (i.e. 1) MIEs. Chemicals activating a higher number of MIEs were considered more likely to activate at least one of the biological pathways summarized in Figure 1, resulting in a final steatosis outcome. Predictions from the consensus models described above were used for the analysis here described.

The virtual screening procedure was applied in three phases:

- Screening of SS1 based on ToxCast experimental data for assays relative to the nine modeled MIEs. This served to assess the existence of a correlation between the selected MIE assays and the steatosis experimental data.
- Screening of SS1 based on predictions returned by consensus models for the nine modeled MIEs. This served to confirm the capability of QSAR prediction to return screening results similar to those obtained with the use of experimental data.
- Screening of SS2 based on predictions returned by consensus models for the nine modeled MIEs. This served as final external validation of the virtual screening strategy

and to assess the real-life capability of QSAR models of prioritize steatotic chemicals with respect to decoys (i.e. inactive chemicals).

ROC curve

Receiver Operating Characteristic (ROC) curves are a widely used strategy to evaluate the results of virtual screening methods.[48, 49] A ROC curve is a plot of true-positive versus false-positive rates for all chemicals ranked by the virtual screening approach. The area under the ROC (AU-ROC) curve is the probability of active chemicals being ranked earlier than inactive chemicals. In this case, chemicals were ranked on the basis of the percentage of activated (i.e. positive) MIEs among the nine in this work.

Then, the AU-ROC was calculated to verify the correlation between steatosis data and MIE experimental and predicted data.

RESULTS

Information on the internal and external validation of the final models selected for each endpoint is reported in Table 2 (undersampling models) and Table 3 (BRF models). For each of the two methods (BRF and undersampling), the best models were reported among those based on VSURF descriptors and those based on all descriptors. Selection was based on the highest mean between MCC in 10-fold CV and in external validation. A complete overview of statistics is available in Tables S3 to S6 of Supporting Information.

Regardless of the modeling approach and of the endpoint modeled, the adopted definition of AD always resulted in an increase of predictivity even if the increase was coupled to a severe reduction of the coverage (> 50%) (Table 2 and 3).

The predictivity estimated on the external validation sets was comparable to or, sometimes even higher than its internal counterpart, indicating models that are not overfitted (Tables 2 and 3). Moreover, for all the models the mean internal MCC (i.e. 10-fold-CV) calculated after 500 scrambling iterations of the dependent variable was always lower than 0.01 (Table S6 in Supporting Information). This confirms that the QSAR models presented are not due to chance correlation.[35]

As regards the 10-fold-CV, the results, of the BRF when analyzed in terms of MCC, were often lower than with the undersampling approach. This was not unexpected since it is somehow an artefact due to the large variability for part of the dataset between the highly unbalanced data (TS_{FULL}) considered in the internal validation of BFR models and the perfectly balanced datasets (TS_{US}) in the case of undersampled models. See the discussion below for a detailed discussion of MCC behavior.

Pre-filtering the initial number of descriptors through the VSURF approach seemed effective for improving the modeling results. This was confirmed by two observations. First, when comparing MCC values for 10-fold CV models obtained with the undersampling procedure (Tables S2 and S3), they appeared much higher when descriptors selected with VSURF were used. Second, the MCC values associated with the VS (without considering AD) were frequently better, although not always satisfactory, for models derived with descriptors selected by VSURF regardless of the modeling approach (BFR or undersampling).

Table 2. Performance of the best RF models obtained by undersampling. For each MIE, the best model between the model based on all descriptors and the model based on a VSURF selection was selected by considering the mean MCC in internal and external validation within the AD.

		PXR_up		PXR_dn		LXR_up		LXR_dn		AhR_up		AhR_dn		NrF2_up		PPARγ_up		PPARα_up	
No. of descriptors		15*		8*		310		318		318		6*		9*		16*		282	
No. of trees		101		51		101		101		101		51		51		51		101	
T_D[a]		0.60		0.75		0.60		0.70		0.65		0.65		0.70		0.65		0.65	
T_C[b]		100th		95th		90th		90th		100th		97.5th		100th		100th		90th	
NN[c]		1		1		1		1		1		1		1		1		1	
AD[d]		all	inAD	all	inAD	all	inAD	all	inAD	all	inAD	all	inAD	all	inAD	all	inAD	all	inAD
10-fold CV	**#[e]**	934	747	132	76	108	54	224	100	258	134	80	53	552	255	424	301	100	40
	P[f]	512	407	66	37	54	27	112	45	129	69	40	24	276	121	212	159	50	27
	N[g]	422	340	66	39	54	27	112	55	129	65	40	29	276	134	212	142	50	13
	ACC	0.74	0.78	0.83	0.89	0.53	0.61	0.64	0.82	0.64	0.74	0.73	0.83	0.75	0.85	0.69	0.76	0.68	0.78
	SE[h]	0.79	0.85	0.83	0.92	0.52	0.63	0.69	0.87	0.63	0.77	0.75	0.88	0.67	0.79	0.75	0.82	0.66	0.70
	SP[i]	0.68	0.69	0.82	0.87	0.54	0.59	0.59	0.78	0.65	0.71	0.75	0.83	0.71	0.75	0.77	0.80	0.70	0.92
	MCC[j]	0.47	0.55	0.65	0.79	0.06	0.22	0.28	0.65	0.28	0.48	0.50	0.70	0.38	0.53	0.52	0.62	0.36	0.59
	BA[l]	0.73	0.77	0.83	0.90	0.53	0.61	0.64	0.82	0.64	0.74	0.73	0.83	0.75	0.85	0.69	0.77	0.68	0.81
	AUC[m]	0.80	0.82	0.86	0.87	0.56	0.62	0.74	0.83	0.72	0.81	0.80	0.87	0.80	0.82	0.75	0.80	0.72	0.80
	%[n]	1.00	0.80	1.00	0.58	1.00	0.50	1.00	0.45	1.00	0.52	1.00	0.66	1.00	0.46	1.00	0.71	1.00	0.40
VS	**#[e]**	235	177	273	132	275	123	227	101	265	131	272	175	216	101	229	168	266	85
	P[f]	128	102	17	7	14	6	29	16	33	19	10	8	70	27	54	41	14	5
	N[g]	107	75	256	125	261	117	198	85	232	112	262	167	146	74	175	127	252	80
	ACC[h]	0.78	0.84	0.63	0.70	0.58	0.62	0.63	0.69	0.65	0.79	0.53	0.51	0.66	0.75	0.73	0.78	0.67	0.87
	SE[i]	0.87	0.92	0.65	1.00	0.64	0.67	0.69	0.69	0.52	0.74	1.00	1.00	0.67	0.78	0.81	0.88	0.71	0.80
	SP[j]	0.67	0.73	0.63	0.69	0.57	0.62	0.62	0.69	0.67	0.79	0.52	0.49	0.65	0.74	0.70	0.75	0.67	0.88
	MCC[k]	0.55	0.68	0.14	0.32	0.10	0.12	0.21	0.29	0.13	0.42	0.19	0.20	0.30	0.47	0.45	0.55	0.18	0.43
	BA[l]	0.77	0.83	0.64	0.84	0.61	0.64	0.66	0.69	0.59	0.77	0.76	0.74	0.66	0.76	0.76	0.81	0.69	0.84
	AUC[m]	0.84	0.86	0.71	0.84	0.62	0.60	0.67	0.71	0.70	0.81	0.79	0.74	0.71	0.80	0.82	0.84	0.75	0.88
	%[n]	1.00	0.75	1.00	0.48	1.00	0.45	1.00	0.44	1.00	0.49	1.00	0.64	1.00	0.47	1.00	0.73	1.00	0.32

[a]Distance threshold. [b]Confidence threshold. [c]Number of neighbors. [d]Applicability domain . [e]Number of chemicals. [f]Number of positive chemicals. [g]Number of negative chemicals. [h] Accuracy. [i]Sensitivity. [j]Specificity. [k]Matthews Correlation Coefficient. [l]Balanced Accuracy. [m]Area Under ROC curve. [n]AD coverage. *Descriptors selected by VSURF.

Table 3. Performance of the best BRF classification models. For each MIE, the best model between the model based on all descriptors and the model based on a VSURF selection was selected by considering the mean MCC in internal and external validation within the AD.

		PXR_up		PXR_dn		LXR_up		LXR_dn		AhR_up		AhR_dn		NrF2_up		PPARγ_up		PPARα_up	
No. of descriptors		1095		18*		1134		1116		13*		1126		1112		937		1126	
No. of trees		501		101		201		201		151		501		201		501		201	
T_D[a]		0.65		0.70		0.60		0.70		0.70		0.60		0.65		0.60		0.65	
T_C[b]		100th		90th		95th		90th		90th		100th		95th		100th		90th	
NN[c]		1		1		5		1		1		1		1		1		5	
AD[d]		all	inAD	all	inAD	all	inAD	all	inAD	all	all	all	inAD	all	inAD	all	inAD	all	inAD
10-fold CV	**#[e]**	934	598	1079	482	1089	637	904	368	1045	463	1079	535	853	428	908	643	1057	546
	P[f]	512	341	66	26	54	30	112	37	129	51	40	18	276	130	212	123	50	23
	N[g]	422	257	1013	456	1035	607	792	331	916	412	1039	517	577	298	696	520	1007	523
	ACC[h]	0.73	0.80	0.63	0.77	0.83	0.93	0.64	0.78	0.77	0.89	0.55	0.56	0.67	0.76	0.76	0.83	0.87	0.95
	SE[i]	0.81	0.88	0.73	0.73	0.37	0.30	0.76	0.89	0.51	0.59	0.53	0.72	0.71	0.78	0.62	0.71	0.42	0.43
	SP[j]	0.64	0.70	0.63	0.77	0.86	0.96	0.62	0.77	0.81	0.93	0.55	0.56	0.66	0.76	0.81	0.86	0.90	0.97
	MCC[k]	0.46	0.59	0.17	0.26	0.14	0.26	0.25	0.43	0.25	0.49	0.03	0.10	0.34	0.50	0.40	0.52	0.21	0.40
	BA[l]	0.73	0.79	0.68	0.75	0.61	0.63	0.69	0.83	0.66	0.76	0.54	0.64	0.68	0.77	0.71	0.79	0.66	0.70
	AUC[m]	0.79	0.84	0.72	0.79	0.67	0.72	0.74	0.83	0.75	0.81	0.57	0.62	0.74	0.81	0.80	0.83	0.71	0.77
	%[n]	1.00	0.64	1.00	0.45	1.00	0.58	1.00	0.41	1.00	0.44	1.00	0.50	1.00	0.50	1.00	0.71	1.00	0.52
VS	**#[e]**	235	151	273	120	275	148	227	103	265	118	272	134	216	120	229	166	266	112
	P[f]	128	92	17	6	14	6	29	14	33	14	10	6	70	36	54	35	14	6
	N[g]	107	59	256	114	261	142	198	89	232	104	262	128	146	84	175	131	252	106
	ACC[h]	0.77	0.87	0.71	0.83	0.81	0.91	0.67	0.78	0.80	0.90	0.49	0.54	0.66	0.74	0.75	0.81	0.89	0.96
	SE[i]	0.88	0.93	0.76	0.83	0.36	0.33	0.69	0.64	0.52	0.43	0.90	1.00	0.64	0.75	0.76	0.83	0.71	0.50
	SP[i]	0.64	0.76	0.70	0.82	0.83	0.94	0.67	0.80	0.84	0.96	0.47	0.52	0.67	0.74	0.75	0.81	0.90	0.98
	MCC[k]	0.53	0.72	0.24	0.35	0.11	0.20	0.25	0.34	0.30	0.45	0.14	0.22	0.30	0.46	0.45	0.56	0.40	0.52
	BA[l]	0.76	0.85	0.73	0.83	0.59	0.63	0.68	0.72	0.68	0.70	0.69	0.76	0.66	0.74	0.75	0.82	0.81	0.74
	AUC[m]	0.85	0.87	0.79	0.83	0.63	0.54	0.71	0.82	0.70	0.76	0.75	0.65	0.72	0.81	0.83	0.88	0.77	0.73
	%[n]	1.00	0.64	1.00	0.44	1.00	0.54	1.00	0.45	1.00	0.45	1.00	0.49	1.00	0.56	1.00	0.72	1.00	0.42

[a]Distance threshold. [b]Confidence threshold. [c]Number of neighbors. [d]Applicability domain . [e]Number of chemicals. [f]Number of positive chemicals. [g]Number of negative chemicals. [h] Accuracy. [i]Sensitivity. [j]Specificity. [k]Matthews Correlation Coefficient. [l]Balanced Accuracy. [m]Area Under ROC curve. [n]AD coverage. [*]Descriptors selected by VSURF.

In some cases, MCC values were low in external validation for endpoints with highly unbalanced datasets, as in the case of undersampled datasets for LXR_up, LXR_dn and AhR_dn (MCC < 0.30 for chemicals in the AD). A possible explanation for this poor performance, can be found in the extreme degree of imbalance of some datasets (i.e. less than 10% of active chemicals) that seriously undermines the reliability of statistical indicators. This will be discussed in the next section.

BRF models were also associated with the lowest external MCC values (for chemicals within the AD) for PXR_dn LXR_up, LXR_dn and AhR_dn marked imbalance between classes.

Table 4 shows the results of consensus modeling in external validation, while Figure 3 compares the performance of consensus models with those of single BRF and undersampled models for each MIE endpoint. Because undersampled and BRF models were characterized by different TSs, consensus internal performance was not evaluated.

In the majority of cases, performance (i.e. MCC) of consensus models improved the results that were obtained with single models when predictions within AD were considered. An exception is the case of the endpoints PXR_up and LXR_dn, since a consensus approach gave results comparable to those obtained with the best single models. In the case of models for LXR the coverage was higher (60% vs. 42%) than the coverage of the single BRF model that was characterized by the same MCC value (i.e. MCC = 0.34).

The other cases were characterized by a gain in performance at the expense of a reduced coverage. This was not unexpected since the consensus strategy returned a prediction only if the query chemical was in the AD of at least three out of four single models.

Table 4. Performance of the consensus models in external validation. For each MIE, the consensus models was based on a majority vote among the four single BRF and undersampled models. Consensus predictions were generated only if the 75% of single predictions (i.e, three out of four) were concordant. When AD was considered, predictions were generated only if, for a given chemical, at least three single predictions were in the AD.

	PXR_up		**PXR_dn**		**LXR_up**		**LXR_dn**		**NrF2_up**		**AhR_up**		**AhR_dn**		**PPARγ_up**		**PPARα_up**	
AD[a]	all	in	all	in	all	in	all	in	all	in	all	in	all	in	all	in	all	in
#[b]	214	146	237	94	215	74	207	136	200	94	234	95	229	107	207	158	220	73
P[c]	118	85	12	3	10	5	29	18	64	26	30	16	10	6	48	35	13	5
N[d]	96	61	225	91	205	69	178	118	136	68	204	79	219	101	159	123	207	68
ACC[e]	0.79	0.86	0.79	0.91	0.76	0.86	0.65	0.74	0.67	0.79	0.71	0.86	0.60	0.66	0.78	0.82	0.90	0.96
SE[f]	0.89	0.94	0.67	1.00	0.50	0.40	0.69	0.72	0.64	0.81	0.53	0.75	0.90	1.00	0.85	0.89	0.69	0.80
SP[g]	0.68	0.75	0.80	0.91	0.78	0.90	0.65	0.75	0.68	0.78	0.74	0.89	0.58	0.64	0.75	0.80	0.91	0.97
MCC[h]	0.59	0.72	0.24	0.50	0.14	0.23	0.24	0.34	0.30	0.54	0.20	0.57	0.20	0.30	0.53	0.60	0.42	0.71
BA[i]	0.78	0.85	0.73	0.96	0.64	0.65	0.67	0.73	0.66	0.79	0.63	0.82	0.74	0.82	0.80	0.85	0.80	0.89
%[j]	0.91	0.62	0.87	0.34	0.78	0.27	0.91	0.60	0.93	0.44	0.88	0.36	0.84	0.39	0.90	0.69	0.83	0.27

[a]Applicability domain . [b]Number of chemicals. [c]Number of positive chemicals. [d]Number of negative chemicals. [e]Accuracy. [f]Sensitivity. [g]Specificity. [h]Matthews Correlation Coefficient. [i] Balanced Accuracy. [j]AD coverage.

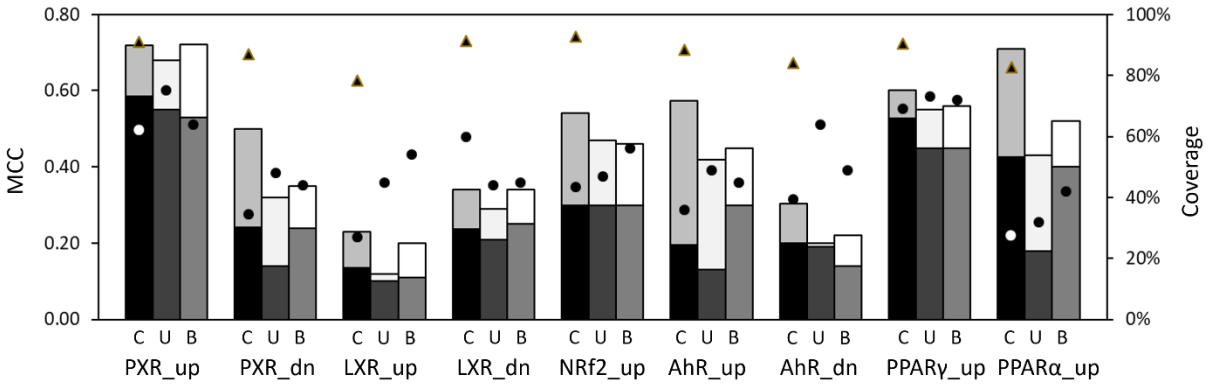

Figure 3. Comparison of consensus models with single BRF and undersampled models. For each MIE, MCC values (left x axis) were reported for consensus (C), undersampled (U) and BRF (B) models. For each staked column pair, the bottom bar shows the MCC calculated on the entire VS, while the top bar shows the MCC for chemicals in AD. Black circles indicates the coverage of the AD of each model (right y axis). Black triangles indicates the coverage of the consensus model without considering the AD.

Table 5 shows the results of the virtual screening analysis for the prioritization of steatotic chemical included in the datasets SS1 and SS2, while Figure 4 reports the ROC curves resulting from the screening procedures.

Table 5. Results of virtual screening of steatotic chemicals. The efficiency of the screening process is presented for the datasets SS1 and SS2. For SS1, results are reported for the use of experimental data on steatosis while considering the inclusion or exclusion of steatotic chemicals that are negative in all the MIE assays. Results based on predicted data from consensus QSAR models for SS1 and SS2 are also reported. For each case, the number of active and inactive chemicals and the value of the Area Under the ROC Curve (AUC) are reported

DBS	MIE data	Decoys	Actives	Area Under Curve (AUC)
SS1	Experimental	191	16	0.548
SS1	Experimental	191	10	0.762
SS1	Predicted	191	10	0.758
SS2	Predicted	84	6	0.720

As shown in Figure 4A, the screening for steatotic chemicals of the SS1 based on experimental ToxCast data led to poor results (AUC =0.548). This was due to the presence of six out of 16 active chemicals being experimentally inactive for all the considered ToxCast MIE assays. An explanation of the poor capability of the MIEs considered to identify some steatotic chemicals is the existence of additional mechanism underpinning steatosis not addressed by the MIEs modeled here. This aspect is addressed more in-depth in the discussion section.

The removal of the six false negative chemicals led to a sensible improvement of results (AUC =0.762). The 50% of active chemicals was ranked in the top 10% of the SS1, while the 90% was in the top 40% (Figure 4B, dashed line). The repetition of the screening procedure using the

prediction of the nine QSAR consensus models lead to analogous results (AUC = 0.758) (Figure 4B, dotted line), confirming the accuracy of the *in silico* models in predicting experimental data.

The same conclusion can be drawn from the second validation performed on SS2 (Figure 4C). The validation revealed an AUC value in line with those observed above (AUC =0.720), characterized by the inclusion of the 80% of active chemicals in the top 30% of the full SS2.

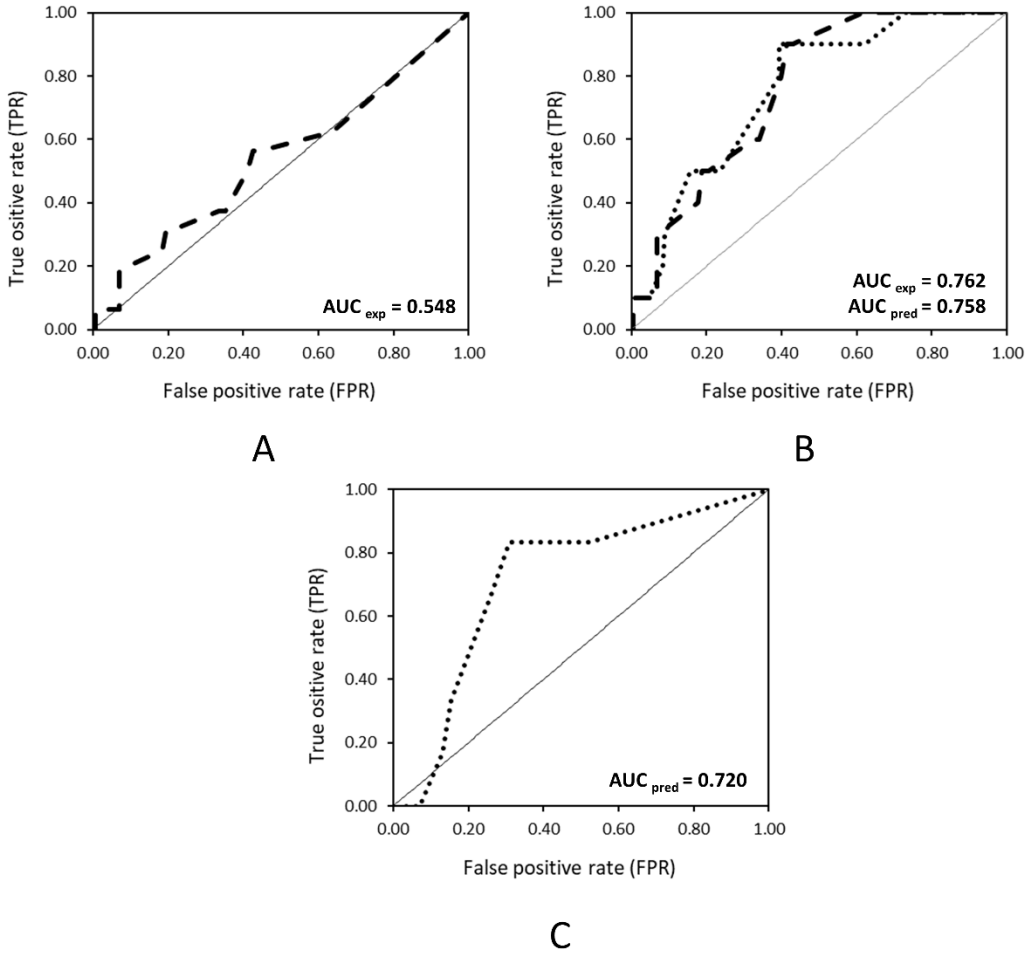

Figure 4. Receiver Operating Characteristic (ROC) curves resulting from virtual screening of steatotic chemicals. ROC curves are reported for: (A) Screening dataset 1 (SS1), including steatotic chemicals negative for all the considered ToxCast assays; (B) Screening dataset 1 (SS1), excluding steatotic chemicals that are negative for all the considered ToxCast assays; (C) Screening dataset 2 (SS2). Dashed lines refer to ROC curve obtained from ToxCast experimental data, while dotted lines refers to predictions from QSAR consensus models.

DISCUSSION

RF are a valuable ensemble method in the field of computational toxicology [8, 50, 51], and it has even been proposed that they can handle multiple mechanisms of action.[51] We observed that RFs performed relatively well without prior removal of irrelevant descriptors (such as in the case of PXR_dn and AhR_up undersampled models, and the majority of BRF models).

However, in some cases a previous feature selection had a positive influence[51] as we also observed for several models obtained with the undersampling approach. This gain in performance is not unexpected, since the R package VSURF used of RF for performing feature selection; for this reason selected descriptors are particularly suitable for the learning of models based on the same approach.[36] This also represents an advantage in terms of computation time with easier interpretation of relevant chemical descriptors.

When RFs are applied to unbalanced datasets, they can generate bootstrap samples of the training data containing few or no chemicals from the minority class in each tree.[51] The direct consequence of this biased sampling is that the predictive performance of each tree for the minority class tends to be weak.

To overcome this problem, we used two modeling strategies for learning from unbalanced data using RF: undersampling of the majority class and BRF. Several publications reported that undersampling is a suitable method for modelling unbalanced datasets [39, 52-54]. A major advantage is that it can be applied to every algorithm. However, this technique may cause some loss of information in terms of coverage of the chemical space of the majority class. We therefore also tested the combination of under-sampling with the BRF method. In this case, the main limitation was related to a loss of statistical performance in 10-fold-CV, which was significant for some 'difficult' cases (e.g. AhR_dn). However, a fair comparison of the two techniques cannot be based

on internal validation parameters, because of the different proportions of active and inactive chemicals in the TSs. On the other hand, the common unbalanced external validation set permits a fair comparison of the two modeling approaches that gave similar predictive performance.

We analyzed the relative importance of descriptors in determining predictions. This was done by measuring the difference in classification error rates between the original RF model and a model obtained by permuting each descriptor one at a time. The differences were then averaged over all trees, and normalized by the standard deviation of the differences. This revealed that the sets composed of the five most important descriptors associated with each model did not share any common descriptor. Nevertheless, it is interesting that five descriptor classes were over-represented within these sets and across all the models (undersampling and BRF):

1) "P_VSA-like descriptors". This class describes the van der Waals surface area (VSA) associated with the lipophilicity of pharmacophore points;
2) "2D autocorrelations". These spatial autocorrelations preferentially measure the level of spatial interdependence between ionization potentials and electronegativity;
3) "2D matrix-based descriptors" and "Burden eigenvalues". In our case information provided by Burden matrixes weighted by intrinsic state, volume, electronegativity and mass had a high discriminating power;
4) "Molecular properties". In our case this class was represented by the partition coefficient between octanol and water (Log P).

The majority of models described in this article predicts with a level of precision that is suitable for screening purposes and their predictive performance is comparable to that of the models of the Tox21 challenge [55] in terms of AU-ROC and BA.

Interestingly, our data selection for chemical purity and generalized cytotoxicity adds to the biological and chemical relevance of the models but does not always result into an enhanced performance than similar modelling exercises.[55]

While the z-score treatment mitigated the issue related to false positives (due to the "burst effect") no action was taken to detect the presence of false negatives. It has already been reported that that the volatility of particular chemical categories (e.g., solvent chemicals) included in ToxCast or the low solubility may explain their general lack of significant effect.[30]

A comparative analysis of Tables 2 and 3 indicates that the values of the statistical parameters (mainly MCC) based on external validation on the most unbalanced datasets (i.e. less than 20% of active chemicals) were unsatisfactory for both approaches and for the same datasets. This highlights the challenge of predicting highly unbalanced datasets regardless of the modeling approach.

Statistical analyses were done to identify critical MCC thresholds for reliably evaluating the performance of models on binary datasets with different degree of imbalance. These thresholds correspond to a reasonable minimum predictivity and we defined this by imposing a minimum percentage of correctly predicted positive and negative chemicals of 75% (i.e. SE = SP= 75%).

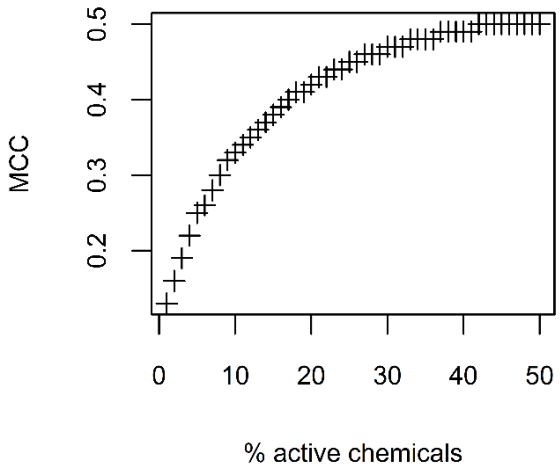

Figure 5. Example of MCC calculated for binary datasets with different degrees of imbalance. Values correspond to a SE and SP, calculated for various degrees of balance of a sample binary dataset (i.e. % active chemicals). Simulation are based on an artificial dataset of 1000 records.

Figure 5 shows how MCC values respecting the previously defined quality criterion vary as a function of the percentage of positive chemicals (from 5% to 50%). Very unbalanced datasets are linked to low MCC values.

Further analysis was also done varying the size of the dataset and keeping constant the degree of imbalance (data not shown). This indicated that the number of samples in the datasets may influence MCC in the case of less populated datasets. MCC values showed a less regular behavior when simulations were performed on artificial datasets including less than 100 samples.

Critical thresholds for MCC were calculated separately for undersampling and BRF approaches for internal (i.e. 10-fold-CV) and external validation (i.e. VS). The number of active and inactive chemicals in the AD of each model was considered to determine the critical MCC threshold for the specific dataset, according to an acceptance threshold of 75%. A MCC above its corresponding

threshold identifies a model that can be considered as valid from a statistical point of view since more than 75% of the chemicals are correctly predicted.

Figures 6A and 6B summarize the internal and external performance of models obtained by the undersampling and BRF approaches in terms of MCC calculated for chemicals within the AD. The figures identify models fulfilling the quality standard. As already mentioned, internal performance of BRF and undersampling-based models cannot be compared directly since the latter is calculated on an evenly distributed dataset. In this last case, critical MCC thresholds for internal validation of undersampling based models are always near 0.50, and this threshold is in line with commonly adopted quality standards.[56]

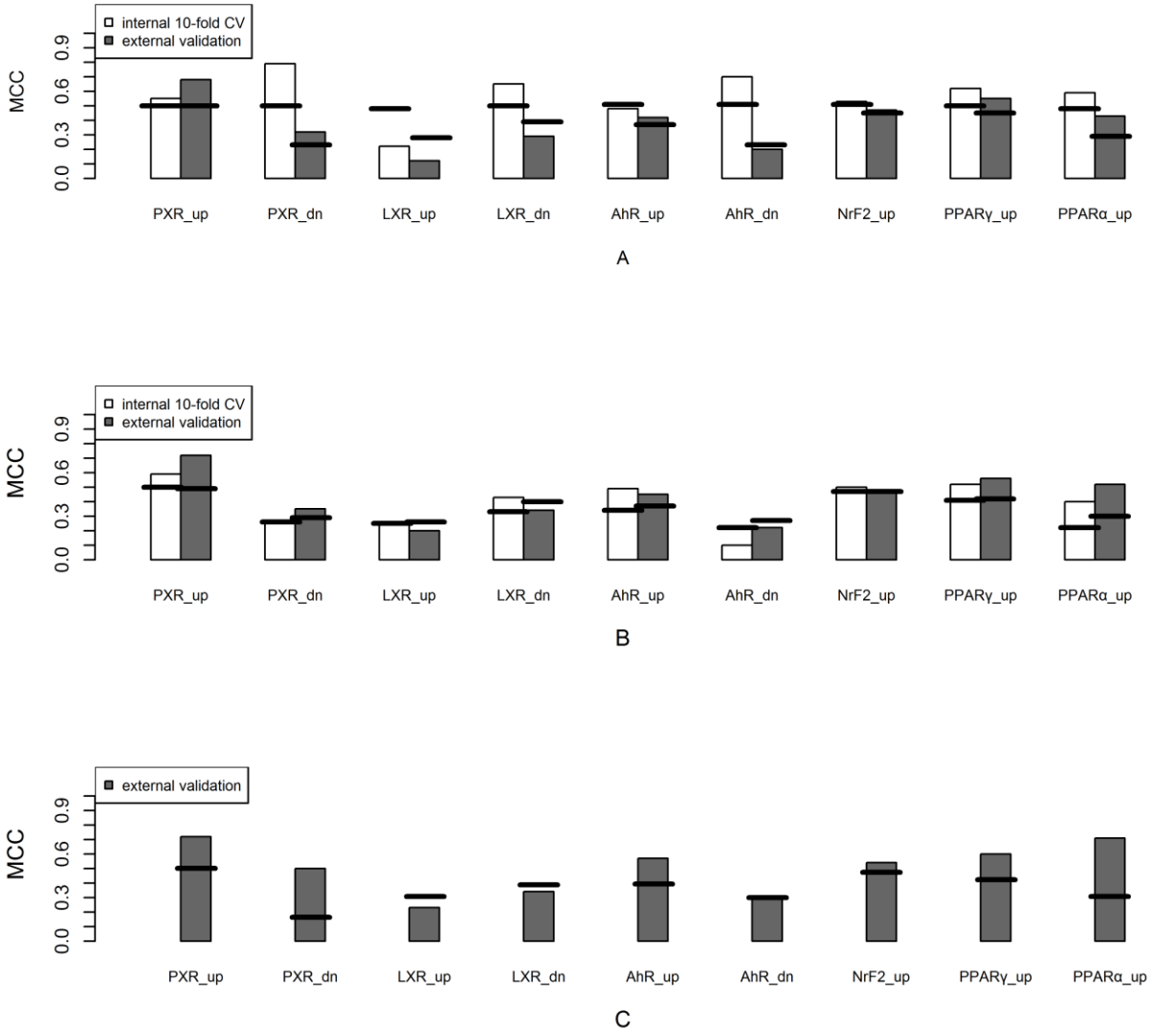

Figure 6. Performance (within the AD) of undersampling and BRF compared to critical MCC thresholds**.** Internal (white bars) and external (grey bars) performance were reported for A) models obtained by undersampling; B) BRF models and C) consensus models. Black lines indicate the critical MCC threshold corresponding to SE and SP of 75%.

PXR_up, PPARγ_up and NrF2_up datasets show a moderate degree of imbalance (i.e. more than 20% of positive chemicals) (Table 1). Models for PXR_up (55% of active chemicals) and

PPARγ_up (23% of active chemicals) are in general acceptable with all approaches or close to the threshold for MCC in some cases even considering the entire dataset (i.e. not only chemicals within the AD) (Figure 6). For these datasets the corresponding coverages within the AD can be regarded as adequate (around 70%) (Tables 2 and 3). Conversely, the reliability of the models obtained by undersampling for NrF2_up (32% of active chemicals) is just below the corresponding acceptability threshold for chemicals within the AD (VSURF selected descriptors) (Figure 6B). Moreover, the coverage of NrF2 models is sometimes not optimal, being near 50% for both internal and external validation (Tables 2 and 3).

It is interesting that both the undersampling and BRF approaches provided the highest MCC values for the least unbalanced data (i.e. Nrf2_up, PXR_up and PPARγ_up). The models for PPARγ_up and PXR_up gave a high MCC in internal and external validation regardless of the modeling approach indicating particularly reliable models. The predictive performance of these models can be regarded as especially robust since it was computed as a function of a VS containing a large number of chemicals (Table 2).

AhR_up models were valid only when the BRF approach was applied to chemicals in the AD with a coverage of about 50%. The model obtained by undersampling yielded a MCC in internal validation (0.48) that is slightly lower than the associated threshold (0.51) (Figure 6A).

As shown in Figure 6, MCCs of PXR_dn models were always above the critical thresholds when models were based on VSURF selected descriptors. The internal MCC for the BRF model barely reaches its critical threshold, but it still can be considered useful.

PPARα_up predicting models reached their critical thresholds for chemicals in the AD and can be considered valid, although the models obtained by undersampling (based on all descriptors) gave an unsatisfactory coverage (32%) in external validation (Table 2).

Models (undersampling and BRF approaches) predicting the endpoints LXR_up, LXR_dn and AhR_dn did not reach the MCC critical thresholds during internal or external validation or both. They cannot therefore be considered as valid (Figure 4). The unsatisfactory performance of models predicting some down-regulation endpoints (AhR_dn and LXR_dn) must also be appraised in relation to the poor quality of input data. In general, it can be observed that models predicting up-regulation endpoints performed better than those predicting down-regulation. This was to some extent expected since information on the ToxCast dashboard[46] or in the file describing the ToxCast endpoints clearly indicates that the assays were not developed or optimized to detect loss of signal.

Consensus modeling was also addressed in order to evaluate its impact on predictivity. Indeed, it has been largely demonstrated that combining different predictions from different sources for a given chemical in a weight-of-evidence approach enables in many cases to positively revaluing the role of *in silico* methods and increasing their relevance for real-life applications.[45, 57, 58] As shown in Figure 6C, consensus modeling allowed in some cases to improve performance of the best BRF and undersampled single models. This is the case of NrF2_up and AhR_up models, that were only barely able to reach the relative critical thresholds for chemicals in AD. The application of consensus modelling allowed to reach the critical thresholds for chemicals in AD for both endpoints (i.e., MCC = 0.54 and 0.57 on the VS, respectively). As for AhR_dn, consensus modeling allowed to reach the critical threshold in external validation, compared to single models that resulted always below the respective thresholds. On the other hand, the LXR_up and LXR_dn endpoints were still below the critical thresholds, even if the gap between the reached performance and the critical threshold was sensibly reduced (Figure 6).

Because of the overall enhanced robustness that is associated to consensus approaches we decided to adopt this strategy to for an *in silico* screening of steatotic chemicals. In this regard, a

virtual screening method was developed. Virtual screening[59] is a widely applied strategy that can identify large number of hits (i.e. positive chemicals) while screening only a portion of a database, because chemicals predicted to be inactive with high confidence are skipped. This approach is important with a view to prioritizing toxicity testing for the chemicals that are more likely to be hazardous. Moreover, the method is particularly suitable for endpoints characterized by a low percentage of active chemicals and a large portion of decoys (i.e. inactive chemicals). Steatosis data from ToxCast represent this situation, since steatotic chemicals are largely underrepresented (i.e. less than 10%).

The first phase of virtual screening evaluation used experimental results from MIE assays in ToxCast (Table 1). This served to confirm the existence of a correlation between incidence of steatosis with ToxCast assays. However, this first attempt led to poor results. This was due to the presence of some experimentally steatotic chemicals showing no interactions with the TF analyzed in this work. For these chemicals, the AOP in Figure 1 was not useful for explaining their steatotic potential. This was not fully unexpected. At present, a definitive and comprehensive mechanistic understanding of the molecular causes of the hepatic steatosis is still not complete. Other mechanisms of action exist leading to hepatic steatosis that were not included in the AOP network described in Figure 1. Consequently, possible additional MIEs were not addressed in the screening protocol here presented and, for this reason, it was impossible to identify chemicals generating steatosis due to different molecular causes.

Despite these limitations, the current screening scheme was able to recognize 10 out of 16 active samples included in the SS1 (i.e. about the 60%) and all the steatotic chemicals in the SS2. More importantly, QSAR predictions returned results comparable with those obtained using experimental assays. This confirmed the capability of the computational procedure here presented

to estimate experimental results and its suitability for screening purposes. In this regard, the virtual screening strategy demonstrated a fair applicability and the potential to support reliably other testing methods (e.g., *in vitro* and *in vivo*) for assessing the hazartd posed by chemicals.

The extension of the current knowledge on the AOP for hepatic statosis and the identification of new MIE may potentially lead to improvement of the whole screening strategy. Elucidating other MIE not covered by the current scheme may allow re-evaluating the toxicity of steatotic chemicals not addressed by the current scheme.

CONCLUSIONS

Computational methods such as QSARs for the prediction of MIE were already recognized by several authors[9, 11] as a valuable first step within a tiered strategy for IATA regulatory application, for their ability to screen large numbers of chemicals at low cost and in a relatively short time. The models for the endpoints PXR_up/dn, NrfF2_up, PPARγ_up and PPARα_up described in this article can be beneficial for prioritizing chemical lists with respect to their steatotic potential or for the toxicological profiling of chemicals of interest and can be regarded as reliable surrogates of HTS *in vitro* tests. These models fulfil the mandatory OECD validation principles for QSAR models [60]: they all have well-defined endpoints, transparent algorithms, defined AD and satisfactory measures of goodness-of–fit, robustness and predictivity.

Moreover, the AD definition we adopted provides information on the confidence that can be assigned to each prediction and this evidence will be important for the parametrization of quantitative AOPs that present a major scientific challenge of the EU-ToxRisk project[14] by providing useful information for the choice of priors in Bayesian contexts[61].

In addition, the developed QSAR models were successfully integrated in a virtual screening strategy for identifying chemicals causing hepatic steatosis. This finding confirms the general applicability of this *in silico* approach to real-life problems. Further improvements of the strategy here described may be the identification of additional MIE to be included in the current AOP scheme, and the development of new QSAR models to expand the domain of applicability of the screening tool.

An interesting additional perspective from a computational point of view will be to use QSARs to provide information for systems biology models aimed at modeling quantitative AOPs for hepatic steatosis or other AOPs involving the same targets.

More generally, the information will also be useful in weight-of-evidence approaches using Demptesr-Shafer theory to model uncertainty in toxicological decision-making.[62]

Acknowledgements

This study has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 681002 (EU-ToxRisk project). The information and views set out in this article reflect only the authors' view and the Commission is not responsible for any use that may be made of the information it contains.

We would also like to acknowledge Dr. Alberto Manganaro (Kode chemoinformatics) for his technical support, Dr. Fredéric Bois (INERIS) for input on the biological mechanisms of MIEs, Dr. Ann Richard, Dr Chris Grulke and Dr. Matthew Martin (US EPA) for answering our questions on ToxCast and Tox21 data, and Judith Baggot for proofreading the manuscript.

Supporting Information

Table S1: Nine datasets used for models derivation and validation with ID, CAS number, pAC50 values, z-score and classification (. i.e. 1 and 0) label (XLSX)

Original list of assays selected their classification into CIS- and TRANS-assays, and the number of active and inactive chemicals for each assay (Table S2).

Performance of undersampled models derived from all DRAGON descriptors (Table S3) and VSURF selected descriptors (Table S4).

Performance of BRF models derived from all DRAGON descriptors (Table S5) and VSURF selected descriptors (Table S6). Statistics refer to 10-fold-cross-validation and external validation. Results of y-scrambling analysis (Table S7) (Word document).

.

REFERENCES

(1) Nassir, F.; Rector, R. S.; Hammoud, G. M.; Ibdah, J. A., Pathogenesis and Prevention of Hepatic Steatosis. *Gastroenterol. Hepatol. (N Y)* **2015,** *11*, 167-175.
(2) Bedogni, G.; Nobili, V.; Tiribelli, C., Epidemiology of fatty liver: an update. *World J. Gastroenterol.* **2014,** *20*, 9050-9054.
(3) Foulds, C. E.; Trevino, L. S.; York, B.; Walker, C. L., Endocrine-disrupting chemicals and fatty liver disease. *Nat. Rev. Endocrinol.* **2017,** 445-457.
(4) Polyzos, S. A.; Kountouras, J.; Deretzi, G.; Zavos, C.; Mantzoros, C. S., The emerging role of endocrine disruptors in pathogenesis of insulin resistance: a concept implicating nonalcoholic fatty liver disease. *Curr. Mol. Med.* **2012,** *12*, 68-82.
(5) Yang, O.; Kim, H. L.; Weon, J. I.; Seo, Y. R., Endocrine-disrupting Chemicals: Review of Toxicological Mechanisms Using Molecular Pathway Analysis. *J. Cancer Prev.* **2015,** *20*, 12-24.
(6) Benigni, R.; Battistelli, C. L.; Bossa, C.; Giuliani, A.; Tcheremenskaia, O., Endocrine Disruptors: Data-based survey of in vivo tests, predictive models and the Adverse Outcome Pathway. *Regul. Toxicol. Pharmacol.* **2017,** *86*, 18-24.
(7) Dearden, J. C., The History and Development of Quantitative Structure-Activity Relationships (QSARs). *IJQSPR* **2016,** *1*, 43.
(8) Low, Y.; Uehara, T.; Minowa, Y.; Yamada, H.; Ohno, Y.; Urushidani, T.; Sedykh, A.; Muratov, E.; Kuz'min, V.; Fourches, D.; Zhu, H.; Rusyn, I.; Tropsha, A., Predicting drug-induced hepatotoxicity using QSAR and toxicogenomics approaches. *Chem. Res. Toxicol.* **2011,** *24*, 1251-1262.
(9) Tollefsen, K. E.; Scholz, S.; Cronin, M. T.; Edwards, S. W.; de Knecht, J.; Crofton, K.; Garcia-Reyero, N.; Hartung, T.; Worth, A.; Patlewicz, G., Applying Adverse Outcome Pathways (AOPs) to support Integrated Approaches to Testing and Assessment (IATA). *Regul. Toxicol. Pharmacol.* **2014,** *70*, 629-640.
(10) Leist, M.; Ghallab, A.; Graepel, R.; Marchan, R.; Hassan, R.; Bennekou, S. H.; Limonciel, A.; Vinken, M.; Schildknecht, S.; Waldmann, T.; Danen, E.; van Ravenzwaay, B.; Kamp, H.; Gardner, I.; Godoy, P.; Bois, F. Y.; Braeuning, A.; Reif, R.; Oesch, F.; Drasdo, D.; Hohme, S.; Schwarz, M.; Hartung, T.; Braunbeck, T.; Beltman, J.; Vrieling, H.; Sanz, F.; Forsby, A.; Gadaleta, D.; Fisher, C.; Kelm, J.; Fluri, D.; Ecker, G.; Zdrazil, B.; Terron, A.; Jennings, P.; van der Burg, B.; Dooley, S.; Meijer, A. H.; Willighagen, E.; Martens, M.; Evelo, C.; Mombelli, E.; Taboureau, O.; Mantovani, A.; Hardy, B.; Koch, B.; Escher, S.; van Thriel, C.; Cadenas, C.; Kroese, D.; van de Water, B.; Hengstler, J. G., Adverse outcome pathways: opportunities, limitations and open questions. *Arch. Toxicol.* **2017,** *91*, 3477-3505.
(11) Patlewicz, G.; Simon, T. W.; Rowlands, J. C.; Budinsky, R. A.; Becker, R. A., Proposing a scientific confidence framework to help support the application of adverse outcome pathways for regulatory purposes. *Regul. Toxicol. Pharmacol.* **2015,** *71*, 463-477.
(12) Ankley, G. T.; Bennett, R. S.; Erickson, R. J.; Hoff, D. J.; Hornung, M. W.; Johnson, R. D.; Mount, D. R.; Nichols, J. W.; Russom, C. L.; Schmieder, P. K.; Serrrano, J. A.; Tietge, J. E.; Villeneuve, D. L., Adverse Outcome Pathways: a Conceptual Framework to Support Ecotoxicology Research and Risk Assessment. *Environ. Toxicol. Chem.* **2010,** *29*, 730-741.

(13) Strickland, J.; Zang, Q.; Paris, M.; Lehmann, D. M.; Allen, D.; Choksi, N.; Matheson, J.; Jacobs, A.; Casey, W.; Kleinstreuer, N., Multivariate models for prediction of human skin sensitization hazard. *J. Appl. Toxicol.* **2017,** *37*, 347-360.
(14) Daneshian, M.; Kamp, H.; Hengstler, J.; Leist, M.; van de Water, B., Highlight report: Launch of a large integrated European in vitro toxicology project: EU-ToxRisk. *Arch. Toxicol.* **2016,** *90*, 1021-1024.
(15) Wittwehr, C.; Aladjov, H.; Ankley, G.; Byrne, H. J.; de Knecht, J.; Heinzle, E.; Klambauer, G.; Landesmann, B.; Luijten, M.; MacKay, C.; Maxwell, G.; Meek, M. E.; Paini, A.; Perkins, E.; Sobanski, T.; Villeneuve, D.; Waters, K. M.; Whelan, M., How Adverse Outcome Pathways Can Aid the Development and Use of Computational Prediction Models for Regulatory Toxicology. *Toxicol. Sci.* **2017,** *155*, 326-336.
(16) Kode: *DRAGON* 7.0.8, 2017; https://chm.kode-solutions.net/products_dragon.php.
(17) Dix, D. J.; Houck, K. A.; Martin, M. T.; Richard, A. M.; Setzer, R. W.; Kavlock, R. J., The ToxCast program for prioritizing toxicity testing of environmental chemicals. *Toxicol. Sci.* **2007,** *95*, 5-12.
(18) Mansouri, K.; Abdelaziz, A.; Rybacka, A.; Roncaglioni, A.; Tropsha, A.; Varnek, A.; Zakharov, A.; Worth, A.; Richard, A. M.; Grulke, C. M.; Trisciuzzi, D.; Fourches, D.; Horvath, D.; Benfenati, E.; Muratov, E.; Wedebye, E. B.; Grisoni, F.; Mangiatordi, G. F.; Incisivo, G. M.; Hong, H.; Ng, H. W.; Tetko, I. V.; Balabin, I.; Kancherla, J.; Shen, J.; Burton, J.; Nicklaus, M.; Cassotti, M.; Nikolov, N. G.; Nicolotti, O.; Andersson, P. L.; Zang, Q.; Politi, R.; Beger, R. D.; Todeschini, R.; Huang, R.; Farag, S.; Rosenberg, S. A.; Slavov, S.; Hu, X.; Judson, R. S., CERAPP: Collaborative Estrogen Receptor Activity Prediction Project. *Environ. Health Perspect.* **2016,** *124*, 1023-1033.
(19) Boughorbel, S.; Jarray, F.; El-Anbari, M., Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLoS One* **2017,** *12*, e0177678.
(20) Huang, R.; Xia, M., Editorial: Tox21 Challenge to Build Predictive Models of Nuclear Receptor and Stress Response Pathways As Mediated by Exposure to Environmental Toxicants and Drugs. *Front. Environ. Sci.* **2017,** *5*.
(21) AOPWiki. https://aopwiki.org/ (accessed Apr 4, 2018).
(22) Lee, L. Y.; Kohler, U. A.; Zhang, L.; Roenneburg, D.; Werner, S.; Johnson, J. A.; Foley, D. P., Activation of the Nrf2-ARE pathway in hepatocytes protects against steatosis in nutritionally induced non-alcoholic steatohepatitis in mice. *Toxicol. Sci.* **2014,** *142*, 361-374.
(23) EPA. Toxicity ForeCaster (ToxCast™) Data. https://www.epa.gov/chemical-research/toxicity-forecaster-toxcasttm-data.
(24) Romanov, S.; Medvedev, A.; Gambarian, M.; Poltoratskaya, N.; Moeser, M.; Medvedeva, L.; Gambarian, M.; Diatchenko, L.; Makarov, S., Homogeneous reporter system enables quantitative functional assessment of multiple transcription factors. *Nat. Methods* **2008,** *5*, 253-260.
(25) Martin, M. T.; Dix, D. J.; Judson, R. S.; Kavlock, R. J.; Reif, D. M.; Richard, A. M.; Rotroff, D. M.; Romanov, S.; Medvedev, A.; Poltoratskaya, N.; Gambarian, M.; Moeser, M.; Makarov, S. S.; Houck, K. A., Impact of environmental chemicals on key transcription regulators and correlation to toxicity end points within EPA's ToxCast program. *Chem. Res. Toxicol.* **2010,** *23*, 578-590.
(26) ChemSpider. (accessed Apr 4, 2018).
(27) ChemIDplus. https://chem.nlm.nih.gov/chemidplus/ (accessed Apr 4, 2018).

(28) Floris, M.; Manganaro, A.; Nicolotti, O.; Medda, R.; Mangiatordi, G. F.; Benfenati, E., A generalizable definition of chemical similarity for read-across. *J. Cheminform.* **2014,** *6*, 39.
(29) Kode: *IstMolBase*, Version 1.0.2, 2017; https://www.kode-solutions.net/.
(30) Judson, R.; Houck, K.; Martin, M.; Richard, A. M.; Knudsen, T. B.; Shah, I.; Little, S.; Wambaugh, J.; Woodrow Setzer, R.; Kothya, P.; Phuong, J.; Filer, D.; Smith, D.; Reif, D.; Rotroff, D.; Kleinstreuer, N.; Sipes, N.; Xia, M.; Huang, R.; Crofton, K.; Thomas, R. S., Editor's Highlight: Analysis of the Effects of Cell Stress and Cytotoxicity on In Vitro Assay Activity Across a Diverse Chemical and Assay Space. *Toxicol. Sci.* **2016,** *152*, 323-339.
(31) Matthews, B. W., Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **1975,** *405*, 442-451.
(32) Cooper, J. A., 2nd; Saracci, R.; Cole, P., Describing the validity of carcinogen screening tests. *Br. J. Cancer* **1979,** *39*, 87-89.
(33) Hanley, J. A.; McNeil, B. J., A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* **1983,** *148*, 839-43.
(34) Gramatica, P., Principles of QSAR models validation: internal and external. *QSAR Comb. Sci.* **2007,** *26*, 694-701.
(35) Eriksson, L.; Jaworska, J.; Worth, A. P.; Cronin, M. T. D.; McDowell, R. M.; Gramatica, P., Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. *Environ. Health Perspect.* **2003,** *111*, 1361-1375.
(36) Genuer, R.; Poggi, J. M.; Tuleau-Malot, C., VSURF: An R Package for Variable Selection Using Random Forests. *The R Journal* **2015,** *7*, 19-33.
(37) Breiman, L., Random Forests. *Machine Learning* **2001,** *45*, 5-32.
(38) Bryll, R.; Gutierrez-Osuna, R.; Quek, F., Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets. *Pattern Recognition* **2003,** *36*, 1291-1302.
(39) Zakharov, A. V.; Peach, M. L.; Sitzmann, M.; Nicklaus, M. C., QSAR modeling of imbalanced high-throughput screening data in PubChem. *J. Chem. Inf. Model.* **2014,** *54*, 705-712.
(40) Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B., KNIME: The Konstanz Information Miner. In *Data Analysis, Machine Learning and Applications*, Preisach, C.; Burkhardt, H.; Schmidt-Thieme, L.; Decker, R., Eds.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2008; pp 319-326.
(41) Chen, C.; Liaw, A. *Using Random Forest to Learn Imbalanced Data*; University of California: Berkeley, 2004; p 12.
(42) Liaw, A.; Wiener, M., Classification and Regression by RandomForest. *R News* **2002,** *2*, 18-22.
(43) Netzeva, T. I.; Worth, A.; Aldenberg, T.; Benigni, R.; Cronin, M. T.; Gramatica, P.; Jaworska, J. S.; Kahn, S.; Klopman, G.; Marchant, C. A.; Myatt, G.; Nikolova-Jeliazkova, N.; Patlewicz, G. Y.; Perkins, R.; Roberts, D.; Schultz, T.; Stanton, D. W.; van de Sandt, J. J.; Tong, W.; Veith, G.; Yang, C., Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. The report and recommendations of ECVAM Workshop 52. *ATLA, Altern. Lab. Anim.* **2005,** *33*, 155-173.
(44) Sheridan, R. P., Three useful dimensions for domain applicability in QSAR models using random forest. *J. Chem. Inf. Model.* **2012,** *52*, 814-823.
(45) Gadaleta, D.; Porta, N.; Vrontaki, E.; Manganelli, S.; Manganaro, A.; Sello, G.; Honma, M.; Benfenati, E., Integrating computational methods to predict mutagenicity of aromatic azo compounds. *J. Environ. Sci. Health C Environ. Carcinog. Ecotoxicol. Rev.* **2017,** *35*, 239-257.

(46) ToxCast Dashboard. https://www.epa.gov/chemical-research/toxcast-dashboard (accessed Apr 4, 2018).
(47) Li, Z.; Berk, M.; McIntyre, T. M.; Gores, G. J.; Feldstein, A. E., The Lysosomal-Mitochondrial Axis in Free Fatty Acid–Induced Hepatic Lipotoxicity. *Hepatology (Baltimore, Md.)* **2008,** *47*, 1495-1503.
(48) Truchon, J. F.; Bayly, C. I., Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. *J. Chem. Inf. Model.* **2007,** *47*, 488-508.
(49) Triballeau, N.; Acher, F.; Brabet, I.; Pin, J. P.; Bertrand, H. O., Virtual screening workflow development guided by the "receiver operating characteristic" curve approach. Application to high-throughput docking on metabotropic glutamate receptor subtype 4. *J. Med. Chem.* **2005,** *48*, 2534-2547.
(50) Polishchuk, P. G.; Muratov, E. N.; Artemenko, A. G.; Kolumbin, O. G.; Muratov, N. N.; Kuz'min, V. E., Application of random forest approach to QSAR prediction of aquatic toxicity. *J. Chem. Inf. Model.* **2009,** *49*, 2481-2488.
(51) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P., Random forest: a classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* **2003,** *43*, 1947-1958.
(52) Zhu, X. W.; Xin, Y. J.; Chen, Q. H., Chemical and in vitro biological information to predict mouse liver toxicity using recursive random forests. *SAR QSAR Environ. Res.* **2016,** *27*, 559-572.
(53) Martin, T. M., Prediction of in vitro and in vivo oestrogen receptor activity using hierarchical clustering. *SAR QSAR Environ. Res.* **2016,** *27*, 17-30.
(54) Soufan, O.; Ba-alawi, W.; Afeef, M.; Essack, M.; Rodionov, V.; Kalnis, P.; Bajic, V. B., Mining Chemical Activity Status from High-Throughput Screening Assays. *PLoS One* **2015,** *10*, e0144426.
(55) Huang, R.; Xia, M.; Nguyen, D.-T.; Zhao, T.; Sakamuru, S.; Zhao, J.; Shahane, S. A.; Rossoshek, A.; Simeonov, A., Tox21Challenge to Build Predictive Models of Nuclear Receptor and Stress Response Pathways as Mediated by Exposure to Environmental Chemicals and Drugs. *Front. Environ. Sci.* **2016,** *3*.
(56) Han, L.; Wang, Y.; Bryant, S. H., Developing and validating predictive decision tree models from mining chemical structural fingerprints and high-throughput screening data in PubChem. *BMC Bioinformatics* **2008,** *9*, 401.
(57) Mazzatorta, P.; Tran, L. A.; Schilter, B.; Grigorov, M., Integration of structure-activity relationship and artificial intelligence systems to improve in silico prediction of ames test mutagenicity. *J. Chem. Inf. Model.* **2007,** *47*, 34-8.
(58) Tropsha, A., Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol. Inf.* **2010,** *29*, 476-488.
(59) Walters, W. P.; Stahl, M. T.; Murcko, M. A., Virtual screening—an overview. *Drug Discov. Today* **1998,** *3*, 160-178.
(60) OECD. OECD Principles for the Validation, for Regulatory Purposes, of (Quantitative) Structure-Activity Relationship Models. http://www.oecd.org/chemicalsafety/risk-assessment/37849783.pdf (accessed Apr 9, 2018).
(61) Bois, F. Y., GNU MCSim: Bayesian statistical inference for SBML-coded systems biology models. *Bioinformatics* **2009,** *25*, 1453-1454.
(62) Park, S. J.; Ogunseitan, O. A.; Lejano, R. P., Dempster-Shafer theory applied to regulatory decision process for selecting safer alternatives to toxic chemicals in consumer products. *Integr. Environ. Assess. Manag.* **2014,** *10*, 12-21.